UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Economia

Corso di Laurea specialistica in Net Economy

Tecnologia e Management dell'informazione e della conoscenza

Tesi di Laurea specialistica

METODI E MODELLI PER L'AGGREGAZIONE, L'ANALISI E LA CERTIFICAZIONE DI SORGENTI DATI ETEROGENEE E DISTRIBUITE

IL CASO DI TRENTINO RISCOSSIONI S.P.A.

Relatore:
Prof. Paolo Giorgini

Correlatore:
Dott. Marco Combetto

Laureando:
Stefano Brida

Anno Accademico 2008-2009

INDICE

INTRODUZIONE

Nelle moderne realtà aziendali è inevitabile che differenti parti dell'organizzazione utilizzino sistemi diversi per produrre e memorizzare i dati necessari per lo svolgimento delle loro attività. Soltanto combinando le informazioni provenienti dai diversi sistemi l'organizzazione potrà accedere al pieno valore dei dati che possiede. La domanda di prodotti di integrazione dei dati consistenti, flessibili e performanti è oggi in continua crescita, la complessità e la rapidità di evoluzione dei business moderni rende vitale per l'azienda disporre di dati integrati, aggiornati e di qualità. Nel contesto attuale grandi quantità di dati sono memorizzate in database relazionali, tuttavia, esiste una pluralità di altre fonti di dati, più o meno strutturati (documenti di testo, fogli di calcolo, ecc.), che con l'avvento del Web si è ulteriormente allargata. Per poter competere nel mercato attuale le imprese hanno la necessità di accedere alle informazioni presenti in tutte le basi dati possedute internamente, attraverso una vista unica integrata e consolidata, inoltre, la necessità di accedere a dati esterni (dati sul Web) si fa sempre più pressante.

L'obiettivo di questa tesi è l'investigazione e lo studio dettagliato delle diverse metodologie e tecnologie esistenti a supporto dell'integrazione di sorgenti di dati eterogenee e distribuite in un ambiente unico, omogeneo e storicizzato. In uno scenario contraddistinto dalla presenza di diverse tecniche di integrazione dei dati sono state analizzate le principali metodologie di Data Integration, descrivendo funzionalità e caratteristiche delle diverse tecnologie presenti sul mercato. Definiti ed esaminati i principali prodotti di integrazione, viene analizzato l'aspetto economico di tali applicativi attraverso un'analisi delle diverse tipologie di costo generate da un progetto di Data Integration. La tesi, oltre a descrivere gli aspetti positivi e i punti di forza che caratterizzano le metodologie di integrazione consolidate, evidenzia gli aspetti più critici ed accenna al possibile impiego di metodologie di nuova generazione, in particolare di tecnologie semantiche, che, sebbene ancora immature, sembrano essere promettenti e si pongono l'obiettivo di superare i limiti tipici delle tecniche tradizionali.

In seguito all'esposizione dello stato dell'arte della disciplina della Data Integration, la tesi descrive e analizza l'impiego di una metodologia specifica in un caso reale nel settore della gestione delle entrate, con particolare riferimento agli aspetti di accertamento dei tributi, dominio applicativo caratterizzato da alta variabilità ed eterogeneità dei dati. Nello specifico, viene valutato l'utilizzo di alcuni strumenti a supporto della realizzazione di una base dati integrata, omogenea e storicizzata, con il fine di giungere ad una soluzione implementabile in tempi medio-brevi e con un chiaro valore aggiunto per l'utilizzatore finale (committente del progetto). In questo senso i risultati ottenuti dalle attività di

sperimentazione sono stati giudicati buoni dal committente e sono oggi il punto di partenza per la realizzazione di un sistema di integrazione che porterà in produzione ciò che si è appreso dall'esperienza sperimentale. Il lavoro pratico svolto nell'ambito di questa tesi da un lato ha permesso di riconoscere i punti di forza e gli aspetti positivi delle metodologie di integrazione tradizionali, dall'altro ha contribuito ad evidenziarne i limiti e le criticità, determinando l'avvio di una fase di valutazione di un possibile impiego di metodologie innovative che, attraverso l'applicazione di tecnologie semantiche, promettono di superare alcuni limiti delle tecniche di integrazione tradizionali.

La tesi è frutto del lavoro effettuato nell'ambito dello stage svolto presso Informatica Trentina S.p.A. nel periodo compreso tra marzo e agosto 2009. In particolare, nel corso dello stage sono state effettuate attività a supporto del progetto di sperimentazione che ha coinvolto Trentino Riscossioni S.p.A., Informatica Trentina S.p.A. e il Comune di Folgaria. Nell'ambito di questa sperimentazione è stata testata, su un ambito specifico dell'accertamento tributario, una soluzione di Data Warehousing dotata di funzionalità di integrazione automatica dei dati, tramite l'utilizzo di tecniche di inferenza statistica, e di strumenti a supporto della gestione della qualità del dato.

La tesi è organizzata in due sezioni composte da tre capitoli ciascuna. Nella sezione I vengono analizzate le principali metodologie di integrazione, proponendo uno stato dell'arte della disciplina che parte dalle origini dei sistemi di gestione dei dati, descrive le principali metodologie consolidate, per arrivare all'analisi dei metodi più innovativi e recenti che si propongono sul mercato. Nella sezione II vengono invece descritte nei dettagli le attività di sperimentazione relative all'integrazione dei dati riguardanti il Comune di Folgaria, mettendo in luce pregi e difetti della situazione sperimentata, i vantaggi ottenibili dalla realizzazione di una base dati integrata e le criticità che si presentano nel realizzare soluzioni di questo tipo, sia a livello generale sia in riferimento al contesto operativo di Trentino Riscossioni.

Il capitolo 1 apre la parte di analisi dello stato dell'arte della disciplina della Data Integration presentando una breve analisi storica dei sistemi di gestione dei dati. In questa prima parte del capitolo viene evidenziata l'evoluzione dei sistemi di Data Management con particolare riferimento alle tecnologie di memorizzazione, analisi e trasformazione dei dati che hanno permesso la realizzazione dei moderni prodotti di Data Integration. Il capitolo prosegue con l'analisi degli approcci di integrazione che si basano sulle tecnologie dei database relazionali, sulle tecniche ETL (Extract, Transform and Load) e su metodi matematico-statistici per il confronto dei record. Vengono in particolare descritti i tre approcci principali: quello che viene definito "Data Consolidation" ovvero l'approccio che ha portato alla definizione della disciplina del Data Warehousing, l'approccio definito "Data Replication" che prevede l'uso di più contenitori integrati (Data Mart) per soddisfare i bisogni specifici di applicazioni diverse. In contrapposizione a questi si colloca

l'approccio definito "Data Federation", che cerca di superare i limiti tipici dei Data Warehouse, proponendo soluzioni di integrazione più scalabili e flessibili. Il capitolo si conclude con l'analisi di un particolare sottoinsieme della Data Integration che nel tempo ha assunto un'importanza sempre maggiore: si tratta della disciplina della Data Quality.

Nel capitolo 2 è presentata la situazione del mercato relativo ai prodotti di Data Integration. Nella prima parte del capitolo trova spazio un'analisi del mercato attuale, dell'andamento negli ultimi anni e delle prospettive future. Vengono quindi descritti i requisiti che un prodotto di Data Integration deve soddisfare per risultare appetibile sul mercato e in seguito sono analizzate caratteristiche e funzionalità dei principali prodotti di Data Integration. Attraverso questa analisi si può capire in che direzione si muove lo sviluppo di tale tipologia di software, con particolare riferimento all'offerta di servizi di integrazione via web-service e allo sviluppo di soluzioni open source. Lo studio del mercato prosegue con un'analisi dei costi dei prodotti presentati, analisi che permette di capire quali sono i costi totali derivanti dall'implementazione di una soluzione di integrazione dei dati (costi di licenza, hardware, supporto, manutenzione, ecc.). Il capitolo si chiude con l'analisi dei limiti e degli aspetti critici che contraddistinguono le tradizionali tecnologie di integrazione.

Il capitolo 3 conclude la prima sezione della tesi esponendo le soluzioni di integrazione più recenti ed innovative, soluzioni che tramite l'utilizzo delle tecnologie proprie del Semantic Web cercano di superare i limiti degli approcci tradizionali. Nel capitolo vengono descritte brevemente le principali tecnologie e linguaggi del Semantic Web dando risalto a come possono essere sfruttate per realizzare metodologie di Data Integration innovative. L'analisi proposta cerca di evidenziare le principali esigenze che spingono verso la realizzazione e l'adozione di nuovi strumenti per l'integrazione di dati, strumenti in grado di sfruttare descrizioni semantiche del contesto di business attraverso l'implementazione di linguaggi ontologici e la realizzazione di archivi condivisi di metadati. Nel corso del capitolo sono descritti punti di forza e criticità dei nuovi approcci anche attraverso un confronto con le tecniche tradizionali. Infine, vengono presentati i primi prodotti di Semantic Data Integration presenti sul mercato.

Il capitolo 4 apre la seconda sezione della tesi con la descrizione del contesto di riferimento del progetto di integrazione sperimentale. Nel capitolo sono descritte caratteristiche e funzionalità di Trentino Riscossioni S.p.A., con particolare riferimento al progetto di sviluppo del sistema informativo a supporto delle attività di riscossione e gestione delle entrate. Viene inoltre descritto il contesto tributario oggetto della sperimentazione con riferimento specifico alle attività di accertamento di due tipologie di tributi (I.C.I. e T.A.R.S.U.). Inoltre, è presente una breve descrizione della società di sistema della Provincia Autonoma di Trento per lo sviluppo di soluzioni nel campo dell'informatica e delle telecomunicazioni, Informatica Trentina S.p.A., e in particolare

dell'area ricerca e innovazione (Trentino as a Lab, settore dell'azienda in cui è stata svolta l'attività di stage) che ha seguito in prima persona le attività di sperimentazione.

Nel capitolo 5 è presentato nei dettagli il caso di reale applicazione, essenzialmente basato sulla metodologia del Data Warehousing, ma con alcuni aspetti innovativi legati all'incrocio dei dati tra sorgenti multiple mediante inferenza statistica, sperimentato con mano e sul campo con i dati del Comune di Folgaria. La sperimentazione ha visto l'applicazione di tale metodologia di integrazione nel contesto tributario locale che coinvolge Trentino Riscossioni e i 223 comuni trentini (la sperimentazione si è focalizzata su un singolo Comune ma sono coinvolte anche basi dati provinciali che raffigurano l'intera realtà trentina). Nella prima parte del capitolo è esposto il lavoro eseguito sulle diverse sorgenti di dati, dal recupero dei dati alla realizzazione di una documentazione tecnico/descrittiva dei diversi flussi. Successivamente sono descritte nei dettagli le diverse fasi e attività che hanno portato alla realizzazione di una base dati integrata sperimentale, mettendo in luce problemi e criticità rilevate. Il capitolo si chiude con la valutazione dei risultati conseguiti in termini di qualità di integrazione dei dati, valutando per quanto possibile gli effetti benefici apportati dalla realizzazione di una base dati integrata nello specifico contesto di business di Trentino Riscossioni.

Il capitolo 6 chiude la seconda parte della tesi con l'analisi della possibile implementazione di tecnologie innovative che potrebbero risultare utili nel contesto di Trentino Riscossioni, contesto caratterizzato da alta variabilità ed eterogeneità di formati e schemi di dati. Si cercherà di capire quali nuove tecnologie possono portare alla realizzazione di un sistema di integrazione più flessibile ed espandibile rispetto alla tecnologia consolidata sperimentata, cercando di capire inoltre se tali tecnologie possono essere complementari e non sostitutive nei confronti della soluzione sperimentata. Nel capitolo sono quindi proposti alcuni scenari di innovazione ritenuti promettenti e che potrebbero essere oggetto di nuove sperimentazioni. In particolare, per una delle tecnologie innovative individuate è già stata intrapresa una nuova sperimentazione con l'obiettivo di verificare l'effettiva utilità dell'utilizzo di un approccio semantico nel contesto operativo di Trentino Riscossioni e più in generale nell'ambito del sistema informativo provinciale. Si tratta della tecnologia semantica sviluppata dal consorzio OKKAM nell'ambito di un progetto europeo di cui l'Università di Trento è coordinatore. Infine, è proposto uno scenario applicativo per la gestione delle diverse fonti dati disponibili nella realtà trentina ai fini di migliorare la fase di collezione, analisi e mappatura delle nuove fonti dati che si renderanno disponibili nelle fasi successive del progetto di integrazione.

Ringraziamenti

Giunto al termine di questo percorso desidero ringraziare tutte le persone che hanno permesso e supportato la realizzazione di questa tesi. In particolare, ringrazio il Prof. Paolo Giorgini per avermi seguito e supportato nel corso dello stage e nella stesura della tesi. Ringrazio inoltre il Dott. Marco Combetto per aver reso possibile la mia partecipazione a questo progetto, per la sua disponibilità e per il contributo dato nella realizzazione della tesi. Infine, ringrazio Informatica Trentina S.p.A., Trentino Riscossioni S.p.A., il Comune di Folgaria, Gruppo S Lab e tutti coloro che hanno partecipato al progetto, rendendo disponibili informazioni, documenti e dati indispensabili per la stesura della tesi.

SEZIONE I

DATA INTEGRATION: STATO DELL'ARTE

CAPITOLO 1

CONTESTO STORICO E METODOLOGIE CONSOLIDATE

L'obiettivo di questo primo capitolo è di illustrare le tecniche tradizionali di Data Integration. Tecniche che permettono la progettazione e l'implementazione di uno schema dati valido, consolidato, efficiente ed estendibile in grado di costituire la base per intraprendere attività di business, di analisi e di pianificazione. Nel capitolo sono quindi descritti modelli, metodologie e tecnologie disponibili per il disegno di una base di dati integrata composta da flussi di dati eterogenei e distribuiti.

Nella prima parte del capitolo è proposta una breve analisi del contesto storico relativo alla disciplina della gestione dei dati (Data Management). In particolare, sono illustrate sinteticamente le principali tappe evolutive che hanno portato alla definizione delle attuali metodologie di integrazione dei dati.

In seguito, vengono presentate le principali metodologie consolidate di Data Integration. Tali metodologie sono basate sulla tecnologia dei database relazionali, su tecniche di estrazione, trasformazione e caricamento dei dati (ETL), su metodi di inferenza statistica per il confronto delle entità e sulle tradizionali tecnologie di accesso ai dati (ODBC, JDBC, XQuery, ecc.). Sono metodologie che vantano parecchi anni di esperienza e sono oggi riconosciute dal mercato come affidabili e consolidate. L'analisi proposta ne determina vantaggi e svantaggi, cercando di evidenziarne le principali criticità.

Il capitolo si conclude con un accenno ad un particolare sottoinsieme della Data Integration: la disciplina della Data Quality. La qualità del dato riveste un'importanza fondamentale ai fini di ottenere un database integrato efficiente e funzionale che permetta di prendere le decisioni di business più corrette possibili. Sono presentate sinteticamente le caratteristiche principali della disciplina, accennando ad alcune metodologie di gestione e controllo della qualità del dato.

1.1 DATA MANAGEMENT

Il termine Data Management racchiude tutte le tecnologie che si occupano di controllo, accesso, protezione, distribuzione e analisi di dati ed informazioni. Vediamo brevemente le principali tappe storiche della disciplina [1] con particolare attenzione alle innovazioni che hanno portato alla realizzazione di metodologie di progettazione, gestione ed integrazione dei dati e in particolare alla nascita del concetto di Data Warehouse (in Figura 1 sono rappresentate le principali tappe evolutive della disciplina del Data Management).

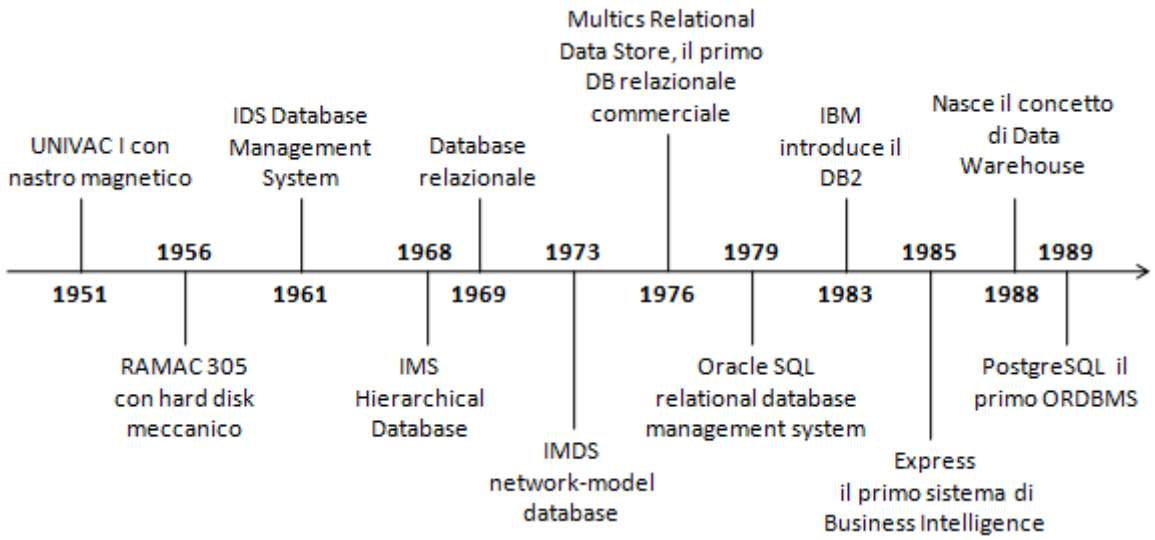

Figura 1 - Data Management: le principali tappe storiche.

Prima degli anni ’50 operazioni di elaborazione di dati comportavano ripetuti scambi di schede perforate tra i pochi centri di elaborazione presenti nel mondo. I risultati di tali elaborazioni potevano essere stampati su carta o convertiti in altre schede perforate e la gestione dei dati implicava una dispendiosa attività di archiviazione manuale delle schede.

Il contesto cominciò a mutare nel 1951 con il lancio del primo computer commerciale, l’UNIVAC I (UNIVersal Automatic Computer I), dotato di un nastro magnetico in grado di memorizzare centinaia di record al secondo. Fu progettato principalmente da John Presper Eckert (ingegnere elettronico) e da John Mauchly (fisico), gli inventori dell’ENIAC[1]. L’UNIVAC I fu lanciato sul mercato nel 1951 dalla Remington Rand[2] e il primo cliente fu l’ufficio del censimento americano che acquistò la macchina il 31 marzo 1951. Tra il 1951 e il 1954 furono prodotti 46 esemplari dell’UNIVAC, ma soltanto 18 di questi sono stati effettivamente installati in quanto si trattava di un sistema troppo costoso per la grande maggioranza delle imprese e Enti americani (alcuni esemplari invenduti sono stati donati alle maggiori università degli Stati Uniti).

Nel 1956, IBM realizza il primo computer con hard disk meccanico (RAMAC 305[3]), costituito da 50 piatti magnetici in grado di contenere 5 megabyte di dati. Con l’introduzione dei piatti rotanti i dati possono essere letti con accesso casuale, mentre con nastri e schede perforate era possibile soltanto una lettura sequenziale. Tuttavia, per

[1] *Electronic Numerical Integrator And Computer (ENIAC)*, fu il primo computer elettronico polivalente, la cui realizzazione fu finanziata dall’esercito statunitense nel corso della seconda guerra mondiale.
[2] Operativa dal 1927 al 1986, fu una delle prime società statunitensi a cimentarsi nella produzione di computer.
[3] *Random Access Method of Accounting and Control (RAMAC)*, sistema caratterizzato dall’accesso casuale ai dati memorizzati che garantisce una maggiore velocità.

decenni sono state utilizzate procedure di elaborazione di dati sequenziali di tipo batch[4] e sono stati necessari diversi anni perché si cambiasse completamente il modo di navigare i dati sfruttando a pieno le potenzialità dell'accesso casuale.

La nascita della disciplina del Data Management viene fatta risalire al 1961, quando Charles Bachman, dipendente della General Electric, sviluppa il primo Database Management System (DBMS): Integrated Data Store (IDS). L'IDS supportava la definizione dello schema dei dati e la registrazione di tutti gli eventi che avvengono su di essi. Il limite di questo sistema stava nel fatto che era possibile eseguirlo soltanto su mainframe della General Electric, inotre, la generazione di tutte le tabelle richiedeva un'intervento manuale da parte dei programmatori. L'IDS fu quindi un sistema innovativo ma scarsamente utilizzabile al di fuori della General Electric. Tuttavia, un cliente della GE, la Goodrich Chemical, decise di riscrivere l'intero sistema per renderlo più flessibile, giungendo alla realizzazione dell'Integrated Data Management System (IDMS). L'IDMS fu ripreso nel 1973 dalla Cullinane Corporation[5] che, dopo averlo perfezionato, lo lanciò sul mercato, divenendo di fatto una delle più grosse software house dell'epoca.

Nel 1969, IBM, dopo aver sviluppato nel 1968 un database gerarchico per i suoi mainframe, Information Management System (IMS), finanzia la ricerca di un sistema migliore in grado di gestire grandi quantità di dati. Il ricercatore Edgar Frank Codd lancia l'idea del database relazionale, dove i dati sono interamente organizzati in tabelle, tra loro relazionate, contenenti un record per riga. IBM avvia quindi un progetto, denominato System/R, incentrato interamente su questa idea. Tuttavia, la necessità di valorizzare il sistema IMS fa si che il System/R non venga lanciato sul mercato, se non dopo il 1980 quando ne viene sfruttata la tecnologia all'interno di un nuovo prodotto.

Nel frattempo, nel 1973 presso l'università della California di Berkeley, i professori Michael Stonebraker e Eugene Wong, facendo uso delle poche informazioni pubblicate sul System/R, sviluppano il proprio database relazionale, l'Ingres. L'Ingres venne commercializzato dalla Ingres Corp. e da altri venditori della Silicon Valley ed ha costituito la base per molte applicazioni successive, come ad esempio Sybase, Microsoft SQL Server, NonStop SQL. Nel 1976 Honeywell lancia sul mercato il Multics Relational Data Store, il primo Relational Database Management System (RDBMS) commerciale.

Verso la fine degli anni '60 venne sviluppata una nuova tecnologia in grado di sfruttare i database relazionali: si tratta dei Decision Support System (DSS) creati per dare la possibilità ai manager di utilizzare meglio i dati al fine di prendere decisioni ottimali. Il

[4] Il termine batch risale ai tempi delle schede perforate ed è utilizzato per indicare un insieme di comandi o programmi, tipicamente non interattivi, aggregati per un'esecuzione non immediata ma rimandata nel tempo.
[5] Fondata nel 1968 da John Cullinane e Larry English. A partire dal 1978 conosciuta come Cullinane Database Systems, nel 1983 cambia nuovamente nome e diviene Cullinet Software. Nel 1989 fu acquisita dalla Computer Associates.

primo prodotto commerciale di questo tipo fu l'Express, disponibile dal 1970. Altri software DSS furono sviluppati da singoli dipartimenti informatici ma non furono mai lanciati sul mercato.

Nel 1977 Larry Ellison fonda la Software Development Laboratories (SDL), divenuta successivamente Relational Software Inc. (RSI), con l'obiettivo di migliorare la tecnologia relazionale. Ellison e i suoi collaboratori sviluppano una prima versione del linguaggio SQL (Structured Query Language**)**, introducendo funzionalità di base di interrogazione e combinazione di tabelle (join). Nel 1979 RSI rilascia Oracle Version 2, il primo RDBMS commerciale con supporto a SQL. Nel 1982 la società cambia nome diventando la Oracle Corporation, ancora oggi azienda leader del mercato dei database relazionali.

Nel 1983 IBM introduce il DB2, un sistema DBMS che utilizza la tecnologia relazionale del System/R e il linguaggio SQL. Per diversi anni il DB2 resta vincolato ai mainframe IBM e soltanto a partire dagli anni '90 viene aperto ad altri sistemi (OS/2, UNIX, Linux, Windows). Il DB2 si è evoluto negli anni ed è oggi uno dei migliori prodotti DBMS presenti sul mercato, nel 2008 risulta secondo, come quota di mercato, soltanto ad Oracle [2].

Nel 1984 Teradata[6] crea un potente RDBMS finalizzato al supporto delle decisioni mentre nel 1985 nasce il primo software di Business Intelligence, sviluppato dalla Metaphor Computer System Inc.[7] per la Procter & Gamble Co. con l'obiettivo di effettuare analisi sui dati di vendita. Nel 1986 la Pilot Software Inc.[8] lancia Command Center, il primo Executive Information System[9] (EIS) commerciale, un prodotto in grado di fornire ai manager un supporto nelle decisioni.

Nel corso degli anni '80 il professor Michael Stonebraker torna a Berkeley ed avvia un progetto di evoluzione dell'Ingres. Il nuovo progetto prende il nome di Postgres e si pone l'obiettivo di superare i limiti dei sistemi DBMS dell'epoca. Postgres introduce la possibilità di definire i tipi di dati e tutte le relazioni presenti tra questi, fissando vincoli e regole. Nel 1988 è realizzato un primo prototipo del prodotto, il cui rilascio avviene nel luglio 1989 dando vita ad una nuova tipologia di prodotti denominati Object-Relational Database Management System (ORDBMS). Nel 1993, nonostante il grande successo e le numerose richieste dal mercato, il progetto viene interrotto e il codice sorgente viene rilasciato con licenza open source permettendo il successivo sviluppo di altri prodotti.

---

[6] Azienda statunitense fondata nel 1979 con l'obiettivo di realizzare un software RDBMS a supporto delle decisioni in grado di operare più operazioni in parallelo sfruttando sistemi multiprocessore. Il nome Teradata simbolizza la capacità del software di gestire terabyte di dati.

[7] Una compagnia derivata dalla divisione ricerca e sviluppo della Xerox, nel 1992 venne acquisita da IBM.

[8] Acquisita da SAP nel 2007. Il nuovo prodotto ha preso il nome di SAP Strategy Management.

[9] Considerati una forma specializzata di Decision Support System (DSS) nell'ambito business. In tempi recenti il termine Executive Information System ha perso popolarità in favore del concetto di Business Intelligence.

Nel 1988 i ricercatori di IBM, Barry Devlin e Paul Murphy pubblicano l'articolo "*An architecture for a business and information systems*[10]*"* dove per la prima volta compare il termine Information Warehouse e danno il via allo sviluppo di un Data Warehouse sperimentale. Nel periodo 1990-1991 si affacciano sul mercato i primi prodotti a supporto della realizzazione di Data Warehouse. La Red Brick Systems realizza Red Brick Warehouse, il primo RDBMS progettato specificatamente per la realizzazione di un Data Warehouse, mentre la Prism Solutions lancia Prism Warehouse Manager, un software dedicato allo sviluppo di Data Warehouse. Nel 1991 William Harvey Inmon[11] pubblica il libro "*Building the Data Warehouse*" rendendo di fatto noto all'intera comunità informatica il concetto di Data Warehouse.

A partire dai primi anni '90, con l'adozione di massa dei computer, si completa una prima fase di evoluzione del Data Management. Nel corso degli anni '90 la tecnologia dei database relazionali raggiunge la sua piena maturazione ed il mercato la riconosce come potente ed affidabile. Soltanto negli anni 2000, in seguito all'esplosione della rete Internet, viene dato un nuovo impulso alla ricerca di nuove vie per gestire e memorizzare i dati, investigando e sviluppando nuove tecnologie e metodologie promettenti ma che ad oggi non sono ancora affermate sul mercato.

1.2 L'EVOLUZIONE DELLA DATA INTEGRATION

Alla luce delle evoluzioni tecnologiche presentate nel paragrafo precedente, viene descritta l'evoluzione della disciplina della Data Integration dai suoi inizi ad oggi, accennando alle diverse metodologie di integrazione sviluppate negli ultimi decenni. Secondo quanto riportato in [3] si possono individuare le tappe che hanno portato alla realizzazione delle moderne metodologie di Data Integration.

Negli anni successivi all'avvento dei primi computer la gestione dei sistemi informativi era piuttosto semplice, al mondo vi erano pochi software che producevano una limitata quantità di dati facilmente organizzabili in una base dati unica. Questo fino a quando non si presentò la necessità di aggiungere nuove funzionalità ai sistemi esistenti, rendendoli più potenti ma complicandoli in modo rilevante. Infatti, la realizzazione di nuove funzionalità comportava l'aggiunta di nuovi dati, in quantità massive, e ci si trovò dinanzi ad un grande problema, che ancora oggi affligge i sistemi informativi: la scalabilità. Il problema della scalabilità può assumere diverse forme: all'epoca si trattava principalmente di scalabilità nella gestione di grandi volumi di dati e di scalabilità nella gestione di grandi volumi di

[10] http://www.research.ibm.com/journal/sj/271/ibmsj2701G.pdf

[11] W. H. Inmon è conosciuto come il padre del Data Warehouse. Ad oggi ha scritto oltre 650 articoli che trattano della progettazione, costruzione, utilizzo e mantenimento di un Data Warehouse, rendendolo di fatto uno dei maggiori esperti nel settore.

transazioni. Con la diffusione dei computer e il conseguente aumento della quantità di dati creati dai sistemi informativi si aggiunge la necessità di avere sistemi di integrazione di dati contraddistinti da un'elevata scalabilità. Per questo motivo, risulta evidente che i primi approcci, basati su una base dati unica per assolvere tutte le esigenze, non erano funzionali per servire un sistema con necessità differenti.

Si assiste così alla nascita dei primi sistemi in grado di sfruttare database multipli, con dati organizzati in contenitori diversi per rispondere a necessità differenti. Il problema di queste soluzioni stava nella ridondanza dei dati tra i diversi contenitori e nel fatto che quest'ultimi spesso non erano tra loro sincronizzati. La scomodità di avere dati ridondanti e non sincronizzati fu per gli utenti una delle prime evidenze della necessità di avere dati integrati. Prendere decisioni di business basandosi su dati di questo tipo si rivelò estremamente problematico e le grandi organizzazioni cominciarono a richiedere sistemi per integrare i dati.

Tale necessità si fece ancora più pressante con l'emergere di problemi nel trasferimento dei dati in un altro database. Spostare dati da un contenitore all'altro senza un sistema in grado di integrarli presentava diverse problematiche: allineare i dati allo stesso istante temporale, diversi livelli di granularità di sorgente e destinazione, ma soprattutto si doveva evitare che i dati relativi ad una stessa entità risultassero come appartenenti ad entità diverse creando ulteriore caos nel database di destinazione. Spostando semplicemente dati da un contenitore all'altro questi problemi non potevano essere governati efficacemente. C'era bisogno di un componente intermedio in grado di operare sui dati di provenienza adattandoli e conformandoli in base alla loro destinazione.

Si assiste così alla nascita dei primi software in grado di operare trasformazioni sui dati in modo da renderli compatibili con il database di destinazione. I primi software di questo tipo erano in grado di compiere trasformazioni di base, ad esempio: riformattare i dati, riordinarli, assegnarli una nuova struttura, aggregarli secondo livelli di granularità differenti, adattarli ad un altro DBMS o sistema operativo, ridefinire le logiche di collegamento tra i dati. Si trattava di funzionalità semplici prese singolarmente, tuttavia i primi sistemi di questo tipo avevano qualche problema quando si doveva fare uso di diverse trasformazioni contemporaneamente.

Con l'affermarsi di quest'ultimi software la domanda di nuove trasformazioni si fece da subito incalzante, nuove esigenze di business richiedevano software di integrazione più sofisticati. I sistemi di Data Integration evolvono in risposta alle seguenti esigenze:

- Maggior numero di programmi da integrare;
- Eterogeneità di funzionalità da supportare;
- Utilizzi sempre più diversificati degli stessi dati per coprire esigenze di business differenti;

- Numero di fonti di dati in continua crescita;
- Aumento dei tipi di dati gestibili;
- Eterogeneità di sistemi operativi e DBMS presenti sul mercato;
- Necessità di valutare e migliorare la qualità del dato.

L'ultimo punto segna un'evoluzione importante, valutare e, se possibile, aumentare la qualità dei dati nel momento della trasformazione e integrazione diviene sempre più importante per permettere di prendere decisioni corrette basate su tali dati. Le prime funzionalità di questo tipo si occupavano di verificare l'integrità dei dati e di controllare il rispetto di determinati vincoli. L'evoluzione successiva riguarda l'utilizzo di metadati, la fase di trasformazione costituisce un eccellente momento in cui raccogliere metadati in modo da aumentare le potenzialità conoscitive del database integrato. I software di Data Integration giungono così ad una maturazione dal punto di vista delle funzionalità offerte, ma il continuo aumento dei dati pone un problema di velocità di elaborazione. Le grandi organizzazioni hanno accumulato notevoli quantità di dati e la loro elaborazione diventa lenta, soprattutto quando più operazioni sono lanciate in parallelo. Per questo motivo, i software si devono adeguare alla richiesta di maggiori prestazioni e sono adattati per l'utilizzo in parallelo su più macchine, velocizzando l'elaborazione dei dati a discapito però di un maggiore numero di computer necessari con il conseguente aumento dei costi per l'hardware.

Risolto anche questo problema si assiste alla maturazione e adozione di massa del concetto di Data Warehouse, visto dalle aziende come il luogo dove memorizzare ed integrare i dati provenienti da sistemi legacy[12]. Con la diffusione dei Data Warehouse l'integrazione dei dati è ormai un concetto consolidato in molte realtà. I sistemi di Data Integration continuano comunque nella loro evoluzione aggiungendo funzionalità e tecnologie sofisticate e sempre più prestanti, arrivando al giorno d'oggi ad essere software consolidati, completi e ricchi di funzionalità. Tuttavia, questa evoluzione ha portato con se un grande problema: la crescita dei costi per adottare tali soluzioni. In tempi recenti il costo dei prodotti di Data Integration si è fatto talmente elevato che solo le organizzazioni più grandi riescono ad affrontare tale spesa. Il problema è che nel business moderno tutte le aziende grandi, piccole e medie hanno bisogno di queste tecnologie.

Secondo W. H. Inmon [3] qui si apre una nuova strada per il futuro: quella dell'open source, una strada costituita da soluzioni pervasive e scalabili, in grado di servire ogni tipo di esigenza e organizzazione senza le restrizioni imposte dai classici prodotti di Data Integration. Un'evoluzione di questo tipo permetterebbe di avere i seguenti benefici:

- Ridotte barriere nell'adozione della tecnologia;

[12] Sistemi operanti con tecnologie obsolete che spesso non vengono sostituiti con tecnologie più recenti a causa degli alti costi di implementazione e di migrazione.

- Costi ridotti;
- Prezzo basato sul reale utilizzo;
- Bassissimo costo di infrastruttura;
- Comunità di utenti che condividono la loro esperienza.

Una soluzione di questo tipo è quella proposta da Talend[13] i cui servizi e funzionalità sono offerti sulla base di un abbonamento. Talend ha aperto la strada per una nuova evoluzione dei software di Data Integration, un approccio nuovo, incentrato sull'utente, che offre le sue funzionalità come servizio, da pagare in base all'effettivo utilizzo (l'approccio open source di Talend è approfondito nel paragrafo 2.2.2).

Sintetizzando, dal punto di vista tecnico/funzionale, si possono individuare principalmente tre generazioni di software di Data Integration [4]:

- **Prima generazione**: software basati sui linguaggi di programmazione, integrare dati significava scrivere il codice necessario per integrare i linguaggi dei diversi database;
- **Seconda generazione**: software dotati di un'interfaccia utente grafica (GUI) con tool grafici a supporto delle fasi ETL;
- **Terza generazione**: Data Integration Suite, pacchetti software costituiti da una collezione di diversi strumenti che condividono la stessa base di metadati. Ogni componente si specializza su una determinata fase del processo di integrazione. Si tratta di pacchetti che includono una grande varietà di funzioni: ETL, ELT, data quality, data profiling, data modeling, data mapping, data federation, data replication, metadata management e metadata reporting.

1.3 METODOLOGIE DI INTEGRAZIONE

Il termine Data Integration racchiude tutte quelle attività e processi che permettono di unire due o più basi di dati provenienti da fonti diverse, in un database o in una vista unica facilmente consultabile dall'utente finale, al fine di effettuare attività di analisi sui dati integrati. Nella maggior parte dei casi l'utilizzo di dati integrati permette infatti di conseguire risultati di business migliori.

Gartner[14] da una definizione più articolata di Data Integration [5]:

> *"A discipline comprising the practices, architectural techniques, and tools for achieving the consistent access to, and delivery of, data across the spectrum of data*

[13] http://www.talend.com
[14] Gartner è una società statunitense che si offre servizi di consulenza, previsione e ricerca nel campo ICT.

subject areas and data structure types in the enterprise, in order to meet the data consumption requirements of all applications and business processes."

Secondo questo punto di vista gli applicativi di Data Integration si pongono al centro delle infrastrutture basate su dati e informazioni, garantendo la possibilità di superare i tipici problemi di condivisione dei dati e di rendere i dati fruibili per tutte le applicazioni e per i processi di business dell'azienda. Si possono individuare principalmente tre metodologie di integrazione [6] [7]: Data Consolidation, Data Propagation e Data Federation.

Data Consolidation

Il concetto di "consolidamento" presuppone l'integrazione di dati provenienti da fonti diverse in un database unico, persistente e storicizzato, attraverso l'applicazione di tecniche ETL (Figura 2). Tutte le integrazioni e trasformazioni necessarie sono svolte nella fase che precede il caricamento nel database unificato (Data Warehouse). Questo tipo di approccio è senza dubbio quello più utilizzato dalle aziende e quello con più anni di esperienza alle spalle. Una soluzione di questo tipo favorisce la realizzazione di applicativi di Business Intelligence rendendo possibili sofisticate analisi dei dati aziendali. Tuttavia, una soluzione di questo tipo comporta un periodo di latenza necessario per integrare i dati nel database unico. L'entità di questo tempo di latenza dipende essenzialmente dall'utilizzo di procedure di integrazione di tipo batch o real-time e da quanto spesso si aggiornano i dati nel database integrato. Data Warehouse e Data Mart sono soluzioni che utilizzano questa metodologia di integrazione.

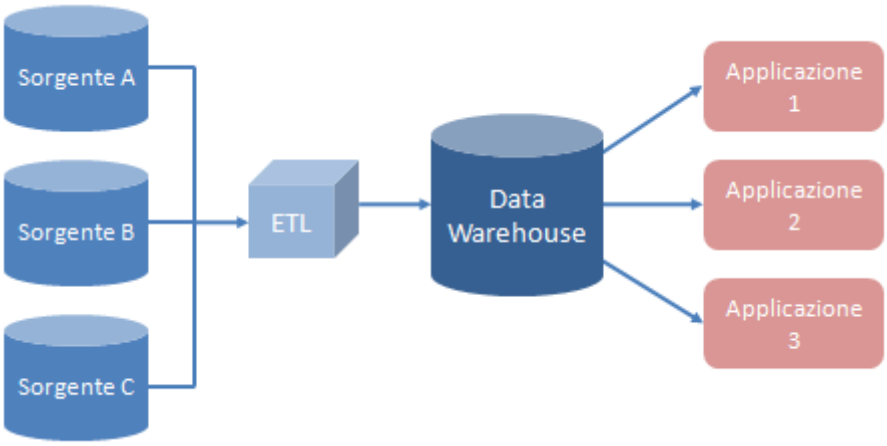

Figura 2 - Rappresentazione schematica di una soluzione di Data Consolidation.

Data Propagation

A differenza di un sistema consolidato, in cui il trasferimento dei dati nel Data Warehouse avviene ad intervalli di tempo prefissati, un sistema basato sulla propagazione è un sistema in continuo movimento. Questa tipologia di approccio è molto utilizzata in quanto facile da realizzare. In un sistema di questo tipo quando un'applicazione necessita di dati

memorizzati in un altro sistema si realizza una procedura automatica che copia i dati necessari in un database dedicato per l'applicazione in questione (Figura 3). In questo modo per ogni applicazione si viene a creare un Data Store dedicato in cui confluiscono i dati provenienti dai sistemi operativi/gestionali dell'azienda. Nei sistemi più sofisticati tale movimentazione di dati può essere anche bidirezionale, tuttavia i sistemi più diffusi risultano essere monodirezionali in quanto più semplici e meno costosi da realizzare. Il problema di un approccio di questo tipo sta nella difficoltà di mantenere sincronizzati i diversi contenitori di dati, più grande diventa il sistema e più è complicato garantire la consistenza dei dati.

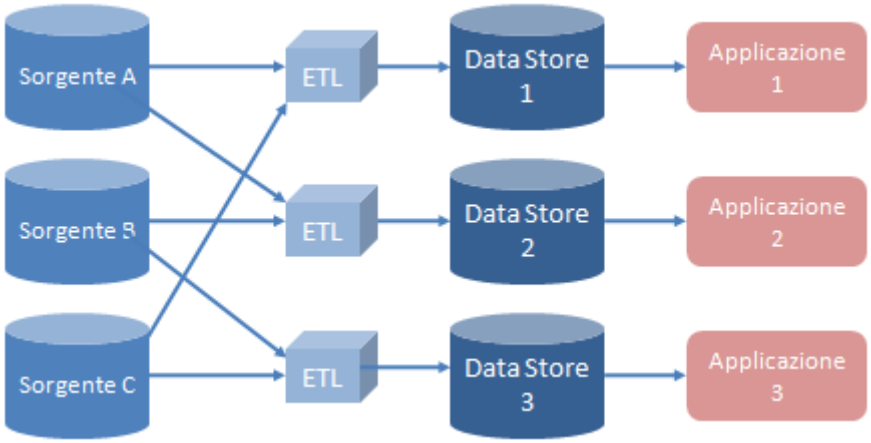

Figura 3 - Rappresentazione schematica di una soluzione di Data Propagation.

Data Federation

L'approccio federato consiste nel creare una vista virtuale unificata di più fonti di dati eterogenee e distribuite (Figura 4). Soluzioni di questo tipo implicano l'utilizzo di metadati o di schemi di collegamento per mettere in connessione le diverse sorgenti garantendo l'accesso ai dati attraverso una vista virtuale unificata. Il vantaggio di una metodologia di questo tipo sta nell'eliminazione dei tempi di latenza e nella maggior flessibilità rispetto a una soluzione di Data Consolidation. Tuttavia, l'approccio federato non permette la realizzazione di operazioni di consolidazione e ripulitura dei dati del livello garantito da una soluzione di Data Warehousing. L'approccio federato è solitamente utilizzato per dare un accesso unificato a più basi dati geograficamente dislocate, per la sincronizzazione di dati o in ambiti caratterizzati da un numero molto elevato di fonti di dati. L'implementazione di un database federato è un'operazione meno costosa rispetto alla realizzazione di un Data Warehouse ma non sempre risulta applicabile con efficacia. In determinate situazioni il consolidamento dei dati, seppur più costoso, rappresenta la soluzione migliore.

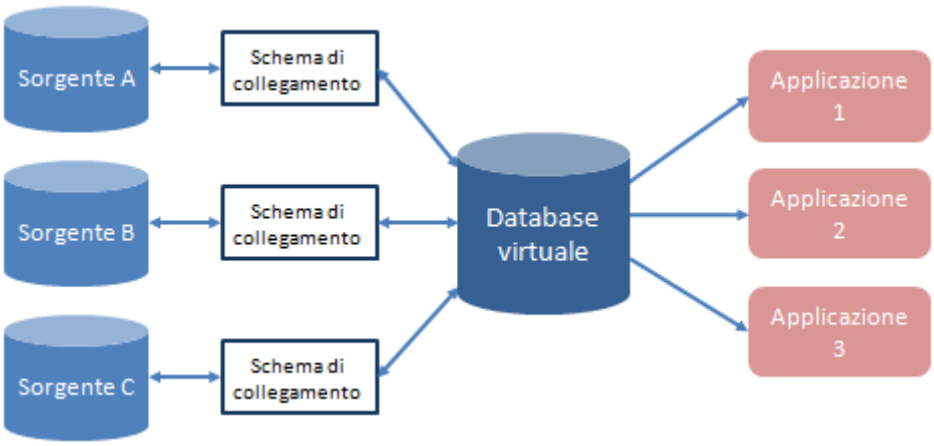

Figura 4 - Rappresentazione schematica di una soluzione di Data Federation.

1.3.1 DATA WAREHOUSING

Nel paragrafo 1.1 abbiamo visto le principali tappe storiche che hanno portato alla definizione del concetto di Data Warehouse. In particolare, in seguito alla pubblicazione del libro di W. H. Inmon *"Building the Data Warehouse"*, la disciplina inizia a diffondersi a livello mondiale tanto che nel 1995 prende vita il Data Warehousing Institute, un'organizzazione che si pone l'obiettivo di diffondere e promuovere nel mondo il concetto di Data Warehouse e di fornire assistenza e attività di ricerca nell'ambito della Data Integration e della Business Intelligence. Si arriva così al 1996 quando Ralph Kimball, altro pioniere del Data Warehousing, pubblica il libro *"The Data Warehouse Toolkit"* che, assieme al libro di W. H. Inmon, costituisce uno dei principali riferimenti della disciplina. A partire dal 1997 i principali prodotti RDBMS iniziano a supportare la realizzazione di Data Warehouse.

I due autori sopracitati danno la propria definizione di Data Warehouse, vediamole entrambe. W. H. Inmon riporta la seguente definizione [8]:

> *"Data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis."*

Inmon specifica inoltre che cosa intende con il termine warehouse:

> *"A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process."*

Secondo la visione di Inmon un Data Warehouse è quindi contraddistinto dai seguenti termini:

- **Subject-oriented**: i dati danno informazioni su un particolare soggetto interno al business dell'azienda;

- **Integrated**: i dati presenti del Data Warehouse provengono da una varietà di fonti e vengono integrati in modo coerente in un database unico;
- **Time-variant**: tutti i dati presenti nel Data Warehouse sono riferiti ad un determinato istante temporale;
- **Non-volatile**: i dati sono stabili nel Data Warehouse, si possono aggiungere nuovi dati in futuro ma nessun dato verrà mai rimosso. In questo modo si ottiene un'immagine consistente e storicizzata del business di riferimento.

Questa definizione risale al 1995 ma risulta ancora oggi abbastanza precisa e valida. Tuttavia, oggi un Data Warehouse riferito a un singolo soggetto viene identificato con il termine Data Mart, mentre con il termine Data Warehouse si fa riferimento all'intero business dell'azienda. Inoltre, i dati oggi possono essere volatili, visto che la quantità di spazio occupata da un Data Warehouse aumenta nel tempo, in alcuni casi si preferisce mantenere al suo interno soltanto determinati periodi storici.

R. Kimball fornisce una definizione di Data Warehouse meno profonda rispetto a quella di Inmon, ma per questo non meno significativa [9]:

> *"A Data Warehouse is a copy of transaction data specifically structured for query and analysis."*

Con il termine Data Warehousing viene identificata la disciplina che raccoglie tutto ciò che è necessario per costruire un Data Warehouse (Figura 5). In particolare, si riferisce ai processi di creazione, popolamento ed interrogazione del Data Warehouse e coinvolge un discreto numero di tecnologie, tra cui [10]:

- **Source System Identification**: prima di avviare la realizzazione del Data Warehouse occorre identificare le basi dati di partenza e definire le procedure per l'invio dei dati. Si tratta di un'operazione semplice se i database di partenza sono in formato relazionale, più complessa nel caso i dati siano organizzati secondo logiche differenti;
- **Data Warehouse Design and Creation**: prima dell'effettiva creazione un Data Warehouse va progettato. Solitamente si progetta la struttura del Data Warehouse in base alle interrogazioni per le quali verrà utilizzato. I processi di design sono iterativi e uno schema può essere modificato diverse volte prima di giungere alla soluzione definitiva. Su tale fase va posta la massima attenzione poiché quando il Data Warehouse sarà popolato da grandi quantità di dati risulterà molto difficile intervenire per modificare la sua struttura;
- **Data Acquisition**: si tratta di tutte le procedure che definiscono la movimentazione dei dati dalle sorgenti al Data Warehouse. È solitamente una delle attività più lunghe e costose da realizzare, in quanto si richiede lo sviluppo o l'acquisizione di

strumenti ETL specifici. Una volta consolidate le regole per la migrazione dei dati i processi saranno schedulati per essere eseguiti a determinati intervalli temporali;

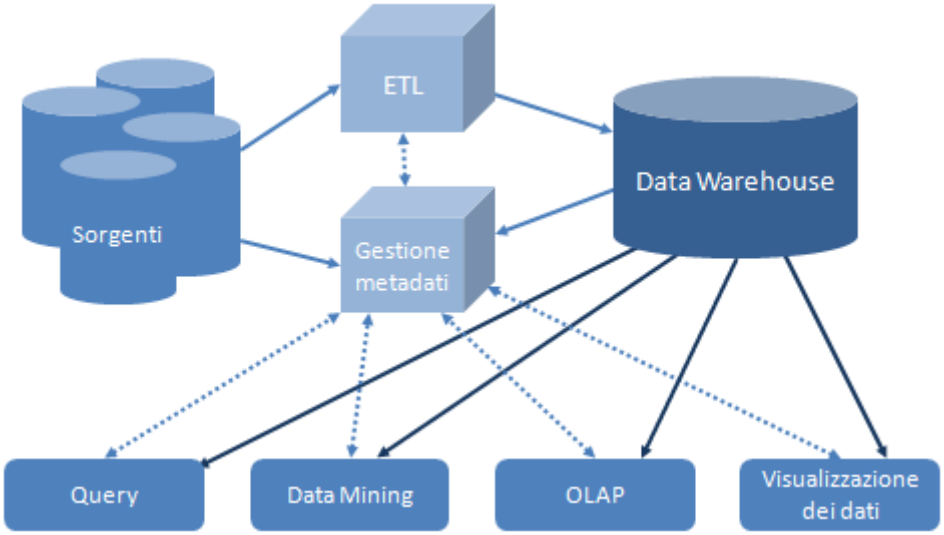

Figura 5 - Rappresentazione semplificata della struttura di un sistema di Data Warehousing.

- **Changed Data Capture**: si riferisce alle tecniche e agli strumenti in grado di capire quali record del Data Warehouse sono stati modificati nei database sorgenti. In questo modo vengono caricati nel Data Warehouse soltanto i record modificati. Quest'area si compone di diverse tecnologie: replication servers, publish/subscribe, triggers and stored procedures, database log analysis;
- **Data Cleansing**: è tipicamente una parte delle attività di *Transform* dell'ETL. I database sorgenti contengono spesso dati errati o in eccesso. Importare tali dati nel Data Warehouse non solo è inutile ma anche pericoloso in quanto l'idea alla base di un Data Warehouse è di supportare il processo decisionale. Prendere decisioni sulla base di dati errati potrebbe avere conseguenze disastrose per l'azienda. Assicurare la qualità del dato comporta procedure di verifica complesse, quella del Data Cleansing è un area molto vasta dove sono disponibili diverse tecnologie per la pulizia dei dati;
- **Data Aggregation**: come la precedente è un'attività che si compie all'interno della fase *Transform* dell'ETL. Un Data Warehouse può essere progettato per contenere dati a differenti livelli di dettaglio. Dati aggregati comportano tempi di interrogazione minori ma si perde il dettaglio a livello di singolo evento. Questo trade-off deve essere valutato attentamente in fase di progettazione in quanto ricostruire il Data Warehouse per la necessità di avere informazioni più dettagliate non è una strada facilmente percorribile;
- **Business Intelligence (BI)**: quando il Data Warehouse è stato creato e popolato è possibile estrarre informazioni significative, in grado di conferire all'azienda un

vantaggio competitivo. L'area della Business Intelligence contiene tecnologie diverse: Decision Support System (DSS), Executive Information Systems (EIS), On-Line Analytical Processing (OLAP), Relational OLAP (ROLAP), Multi-Dimensional OLAP (MOLAP), Hybrid OLAP (HOLAP, combinazione di MOLAP e ROLAP), ecc. more. L'area della BI può essere suddivisa in quattro aree principali:

  - **Multi-dimensional Analysis**: strumenti che permettono di analizzare i dati da un numero di angolazioni differente. Tali strumenti fanno uso di un database multidimensionale solitamente definito "cubo";
  - **Data Mining**: sono strumenti in grado di effettuare automaticamente ricerche di specifici modelli e schemi (pattern) nei dati. Tali strumenti utilizzano solitamente complesse tecniche statistiche. La differenza con gli strumenti OLAP sta nel fatto che con gli OLAP bisogna conoscere la domanda da porre sui dati, mentre con le tecniche di data mining non è necessario sapere cosa chiedere;
  - **Data Visualization**: sono strumenti in grado di dare una rappresentazione grafica dei dati, anche sottoforma di complessi modelli tridimensionali. Questi strumenti si basano sulla teoria che un utente può capire meglio i dati se li vede rappresentati graficamente;
  - **Query**: si tratta essenzialmente di strumenti che utilizzano il linguaggio SQL per effettuare interrogazioni sui dati;
- **Metadata Management**: durante i processi di identificazione, acquisizione e interrogazione dei dati si originano una molteplicità di metadati. Gestire tali metadati può risultare molto importante ai fine di ottenere analisi future migliori. La gestione dei metadati permette inoltre di comprendere al meglio la struttura del Data Warehouse. I moderni Data Warehouse sono progettati con grande attenzione alla gestione dei metadati e i maggiori prodotti di Data Integration prevedono un contenitore per i metadati condiviso tra tutti gli strumenti facenti parte del Data Warehousing.

1.3.1.1 OBIETTIVI

Prima di entrare nei dettagli delle fasi di progettazione e implementazione di un Data Warehouse è utile focalizzare quali sono i suoi principali obiettivi. Secondo R. Kimball un Data Warehouse deve essere in grado di soddisfare i seguenti requisiti [9]:

- *Rendere le informazioni facilmente accessibili*: i dati contenuti in un Data Warehouse devono essere facilmente comprensibili, intuitivi ed ovvi non solo per i progettisti ma anche per gli esperti di business che dovranno utilizzare tali dati;

- *Presentare le informazioni in maniera consistente*: i dati devono essere credibili. Dati provenienti da fonti differenti devono essere puliti e integrati al medesimo livello di aggregazione. Informazioni consistenti sono sinonimo di informazioni di qualità;
- *Adattarsi ai cambiamenti*: le necessità degli utenti, le condizioni di business e le tecnologie sono soggette a frequenti cambiamenti. Un Data Warehouse deve essere progettato tenendo conto delle possibili variazioni del contesto in cui si inserisce. I cambiamenti dovrebbero essere possibili senza conseguenze come la perdita di dati o l'incompatibilità con altre applicazioni. Nella realtà un Data Warehouse è difficilmente adattabile senza conseguenze e questo costituisce uno dei più grandi limiti di tale tecnologia;
- *Essere un contenitore solido e sicuro*: un Data Warehouse deve essere in grado di proteggere il patrimonio informativo dell'azienda. Al suo interno sono memorizzate informazioni cruciali, che determinano spesso il vantaggio competitivo dell'azienda sul mercato. Un Data Warehouse deve quindi possedere un sistema di accesso controllato;
- *Costituire la base su cui prendere decisioni di business*: il Data Warehouse deve contenere i dati più corretti, sui quali è possibile prendere le decisioni di business migliori;
- *Essere accettato dagli utenti*: non è importante quanto elegante è la soluzione sviluppata, se gli utenti non la utilizzano con costanza il progetto è da ritenersi fallito. Gli utenti dimostrano di preferire soluzioni semplici da utilizzare, chi progetta un Data Warehouse deve tenere conto delle preferenze dei futuri utilizzatori se vuole realizzare una soluzione di successo.

1.3.1.2 ARCHITETTURA

Con il termine architettura si intende una concettualizzazione di come il Data Warehouse viene costruito. Non vi sono architetture giuste o sbagliate, esistono architetture differenti ognuna adatta in un particolare ambiente e situazione. La scelta dell'architettura corretta per un determinato contesto è fondamentale, rende meno complicate le operazioni di costruzione e manutenzione del Data Warehouse e ne permette un utilizzo migliore da parte degli utenti finali. Sintetizzando ciò che riportano i principali autori [9] [11] [12] è possibile definire un'architettura generica per un Data Warehouse, da adattare di volta in volta in base al contesto di utilizzo, che consiste in una serie di strati interconnessi:

- **Operational source systems**: si tratta del livello base che garantisce al Data Warehouse il collegamento con le fonti dei dati. Il requisito principale dei sistemi sorgente è costituito da performance di elaborazione e disponibilità del sistema. È

importante che i sistemi sorgente mantengano tali requisiti su buoni livelli, in caso contrario i problemi si riverseranno sull'intero Data Warehouse;

- **Data staging area**: si tratta di un'area di memorizzazione dei dati intermedia. Una volta realizzata l'interfaccia di collegamento con i sistemi sorgente i dati vengono caricati in un ambiente temporaneo in attesa di essere sottoposti ad elaborazioni tramite tecniche ETL. Soltanto successivamente a tali elaborazioni i dati verranno inseriti nell'ambiente integrato finale;
- **Metadata layer**: è un livello che mantiene traccia di tutte le elaborazioni a cui sono sottoposti i dati con la finalità di migliorare tali operazioni in futuro o di tenere traccia di eventuali problemi o errori. Kimball identifica le seguenti categorie di metadati interne a un Data Warehouse [13]:
    - **Source system metadata**: metadati riguardanti le basi dati originarie. Si tratta di informazioni riguardanti gli schemi logici, gli attributi contenuti, il loro significato semantico, relazioni tra i dati e frequenza di aggiornamento;
    - **Data staging metadata**: sono informazioni relative alla fase di acquisizione dei dati originari, sulla trasmissione dei dati, sulle tabelle da importare, sulle chiavi da utilizzare. Questi metadati memorizzano informazioni necessarie ai processi di trasformazione e aggregazione dei dati;
    - **DBMS metadata**: si tratta di dati relativi al contenuto e alla struttura delle tabelle che costituiscono il Data Warehouse.
- **Data presentation area**: si tratta dell'area dove i dati sono organizzati, immagazzinati definitivamente e resi disponibili per gli utilizzi successivi. È la parte del Data Warehouse che interessa direttamente l'area di business che dovrà utilizzare i dati integrati per pianificare le proprie attività;
- **Data Access Tools**: sono gli applicativi che permettono di analizzare i dati ed utilizzarli per prendere decisioni di business. Tutti gli strumenti di Business Intelligence elencati in precedenza nel paragrafo 1.3.1 rientrano in questa parte. Sono strumenti che interrogano e analizzano i dati presenti nell'area definita data presentation.

1.3.1.3 MODELLO DEI DATI

Ci sono fondamentalmente due approcci per definire il modello dei di un Data Warehouse: l'approccio dimensionale e l'approccio normalizzato [12]. L'approccio dimensionale è proposto da Kimball ed è conosciuto anche come schema a stella, mentre Inmon suggerisce l'utilizzo dell'approccio normalizzato. Tuttavia, pur utilizzando schemi progettuali differenti, i due autori giungono ad una soluzione finale che opera sui medesimi principi. L'approccio dimensionale organizza i dati in dimensioni e fatti. I concetti di dimensione e fatto vengono coniati nel corso degli anni '60, in un progetto di ricerca nato dalla

collaborazione tra la General Mills[15] e il Dartmouth College e tutt'oggi vengono utilizzati per identificare le tabelle presenti in un Data Warehouse [9]. Kimball riprende i concetti di dimensione e fatto e propone lo schema dimensionale che prevede la presenza di una tabella fatto centrale (dove con fatto si intende un evento, una transazione, qualcosa che avviene in un determinato momento) alla quale sono collegate più tabelle definite dimensioni (una dimensione contiene informazioni su una determinata entità che è coinvolta nel fatto, può essere una persona, un luogo, un oggetto, ecc.) (Figura 6) [9]. Il vantaggio di questo tipo di approccio sta nel fatto che gli utenti riescono a capire meglio la struttura del Data Warehouse e, di conseguenza, sono in grado di usarlo con relativa facilità; questo garantisce tempi di operatività della soluzione molto brevi. Tuttavia, tale approccio presenta anche una serie di svantaggi:

- Per mantenere l'integrità delle tabelle fatti e dimensioni il caricamento di dati da sistemi gestionali risulta complicato;
- Se la società cambia il suo modello business è molto difficile modificare la struttura del Data Warehouse per adattarla alle nuove necessità.

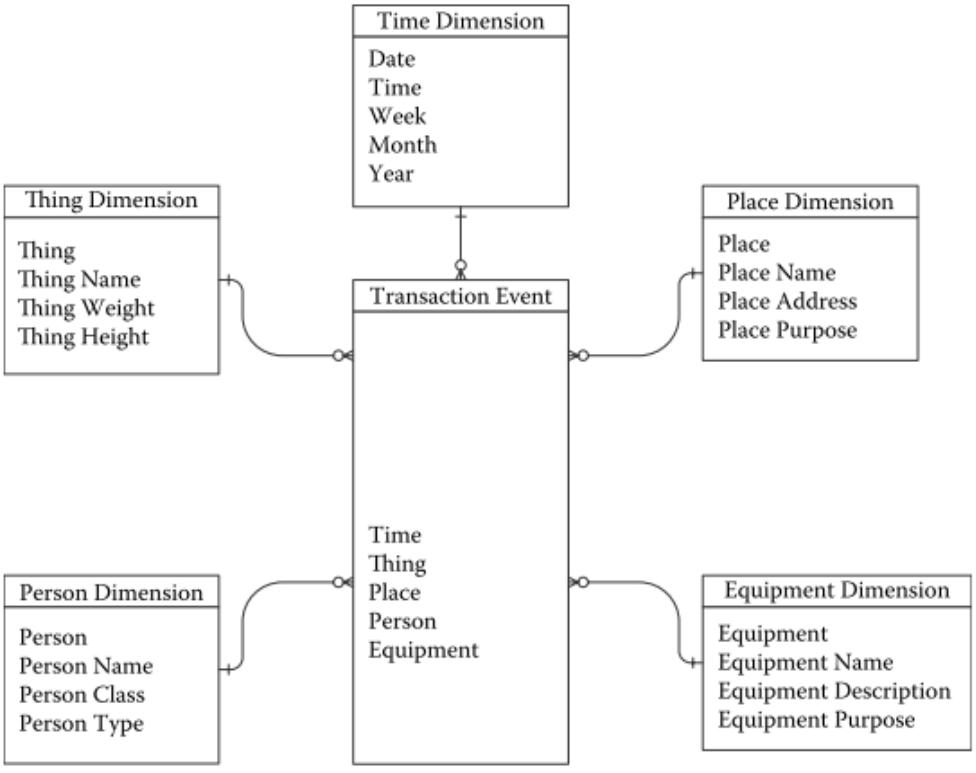

Figura 6 - Modello dei dati dimensionale (schema a stella) (Fonte: [12]).

L'approccio normalizzato (Figura 7) è proposto da W. H. Inmon [11] ed è conosciuto come *Third Normal Form Data Model*.

[15] Azienda statunitense, fondata nel 1856, operante nel campo dell'industria alimentare.

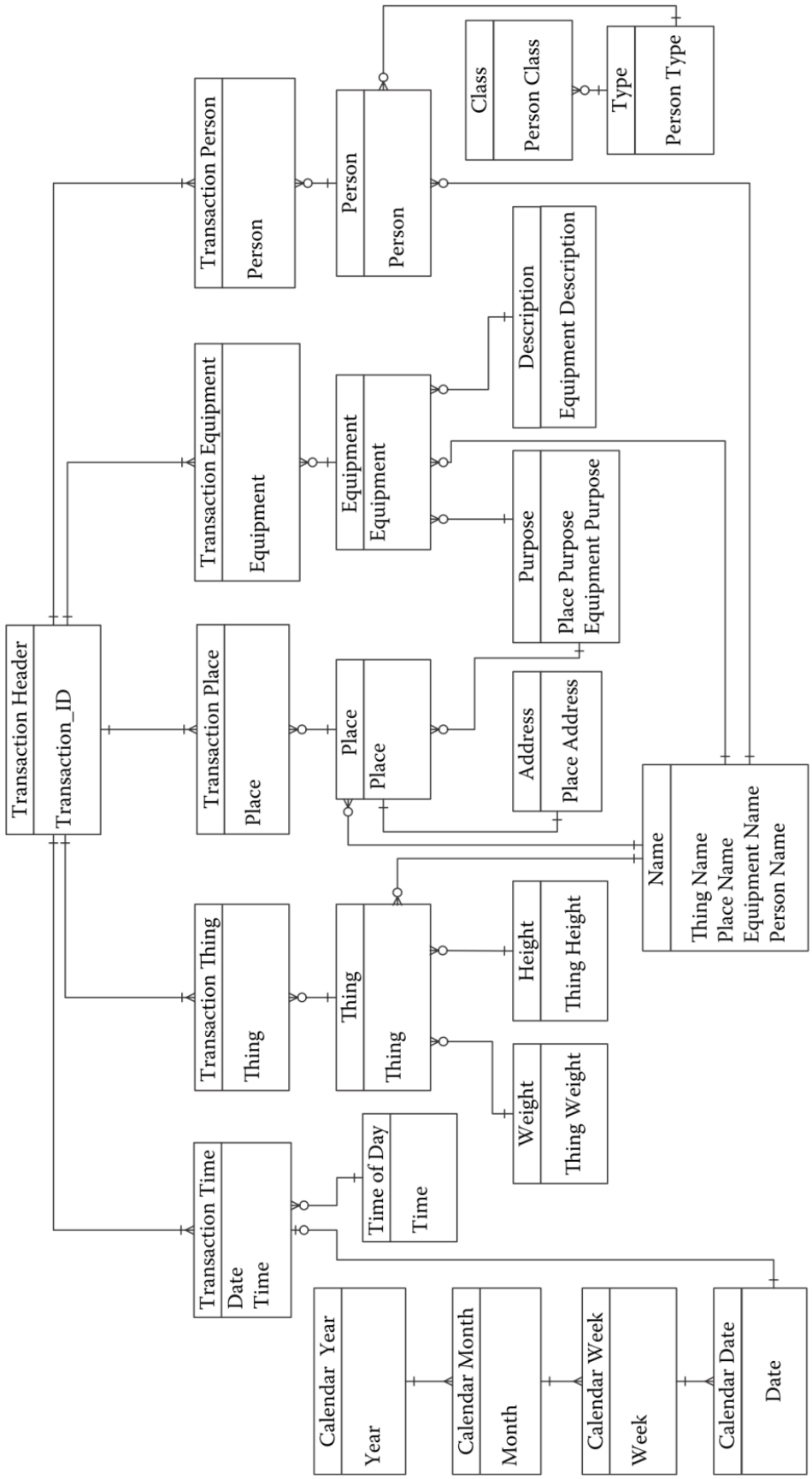

Figura 7 - Modello dei dati normalizzato (*Third Normal Form Data Model*) (Fonte: [12]).

Tale approccio, anziché organizzare i dati catturando gerarchie e relazioni in dimensioni e fatti, permette di rappresentare i dati seguendo lo schema di funzionamento dell'azienda. Le tabelle sono organizzate in aree che riflettono le principali categorie con cui opera l'azienda (clienti, prodotti, finanza, ecc.).Questo tipo di approccio garantisce quindi una maggiore flessibilità, permettendo al Data Warehouse di adattarsi al meglio alle necessità aziendali. La maggiore flessibilità permette inoltre di aggiungere facilmente informazioni al Data Warehouse. Tuttavia, la complessità con la quale sono strutturati i dati può creare dei problemi agli utenti nel ricavare le informazioni di cui necessitano per le analisi di business.

1.3.1.4 PROGETTAZIONE

Anche per quanto riguarda la metodologia di progettazione del Data Warehouse i due autori di riferimento, Kimball e Inmon, propongono due approcci differenti. La scelta di un determinato approccio metodologico dipende strettamente dall'organizzazione aziendale, dal contesto di business, dalla tipologia di utenti e dalle loro necessità e dall'architettura tecnica del sistema di Data Warehousing che si intende realizzare. L'elevato numero di insuccessi nelle prime implementazioni di soluzioni di Data Warehousing ha portato alla definizione di metodologie più incentrate sugli aspetti di business: si parla di approcci top-down, bottom-up e incrementale a cui corrispondono diverse tipologie di Data Warehouse.

Approccio top-down

W. H. Inmon ha definito il Data Warehouse come il deposito dei dati centralizzato che deve servire l'intera organizzazione; per questo è da sempre promotore dell'approccio top-down.

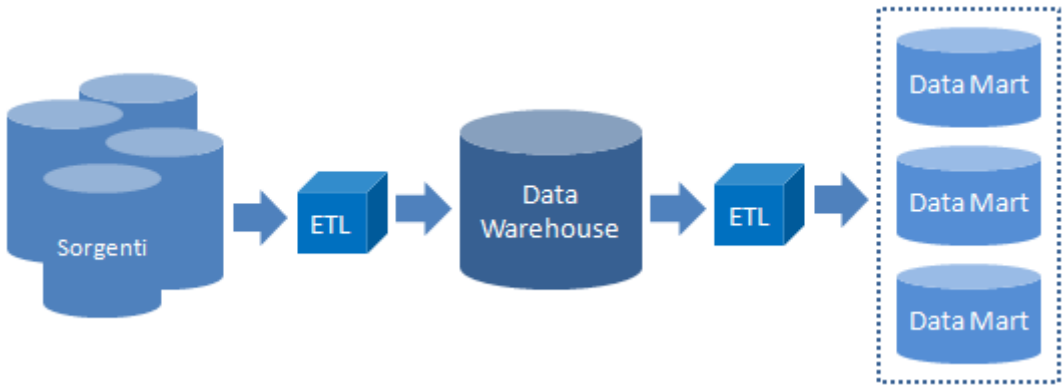

Figura 8 - Approccio top-down per la progettazione di un Data Warehouse.

La criticità dell'approccio di tipo top-down è data dalla difficoltà di gestione di un progetto in grado di tenere in considerazione le singole necessità di ogni area aziendale. Il rischio è che la fase di progettazione risulti troppo complicata, comportando un rallentamento delle attività aziendali con risultati concreti ottenibili troppo in avanti nel tempo. Il vantaggio principale di questo approccio sta nella semplicità con cui si può realizzare un Data Mart

specifico per soddisfare nuove esigenze di business non previste in un primo momento. La scelta di un approccio di questo tipo deve passare da un'attenta valutazione dello sforzo che il cliente vuole sostenere e dell'effettivo impatto strategico che si aspetta dalla realizzazione di un Data Warehouse. Nella visione di Inmon il Data Warehouse è quindi il centro informativo per l'intera azienda che costituisce la base su cui effettuare attività di Business Intelligence e pianificazione in grado di migliorare la gestione del business aziendale.

Approccio bottom-up

R. Kimball promuove un approccio inverso rispetto a quello proposto da Inmon, un approccio che parte dalla realizzazione non coordinata di una serie di Data Mart, specializzati su singoli fabbisogni informativi di una determinata area aziendale. Tali Data Mart contengono dati sia al massimo livello di dettaglio sia dati di tipo riassuntivo e, qualora fosse necessario, possono essere uniti in un Data Warehouse omnicomprensivo (Figura 9). Questo tipo di approccio ha il vantaggio di permettere il conseguimento di risultati concreti in tempi molto brevi con costi variabili, dipendenti dal numero di dati da elaborare e di Data Mart da realizzare. Tuttavia, a fronte di bassi costi iniziali e tempi relativamente brevi si potrebbero incontrare difficoltà in futuro qualora si volessero riunire tutti i dati sparsi nei vari Data Mart in un contenitore unico centralizzato. Un'operazione di questo tipo presenta solitamente costi elevati e non sempre è realizzabile. Infine, i costi di manutenzione di un sistema di questo tipo aumentano costantemente all'aumentare della complessità del sistema (numero elevato di Data Mart).

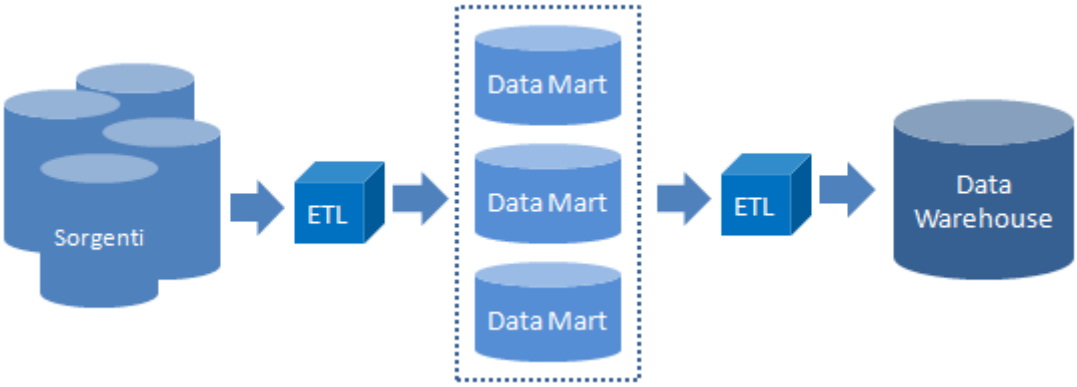

Figura 9 - Approccio bottom-up per la progettazione di un Data Warehouse.

Approccio incrementale

L'approccio incrementale (definito anche approccio ibrido o federato) cerca di combinare i vantaggi dei due approcci sopra descritti. Tale approccio prevede la realizzazione di un modello informativo comune sul quale vengono sviluppati i modelli di dati del Data Warehouse e dei singoli Data Mart. In seguito ai processi di acquisizione e trasformazione dei dati viene creato uno schema di collegamento comune che definisce le relazioni tra i

dati di origine con quelli presenti nel Data Warehouse e nei diversi Data Mart, permettendo un collegamento più semplice e immediato tra le basi di dati in gioco (Figura 10).

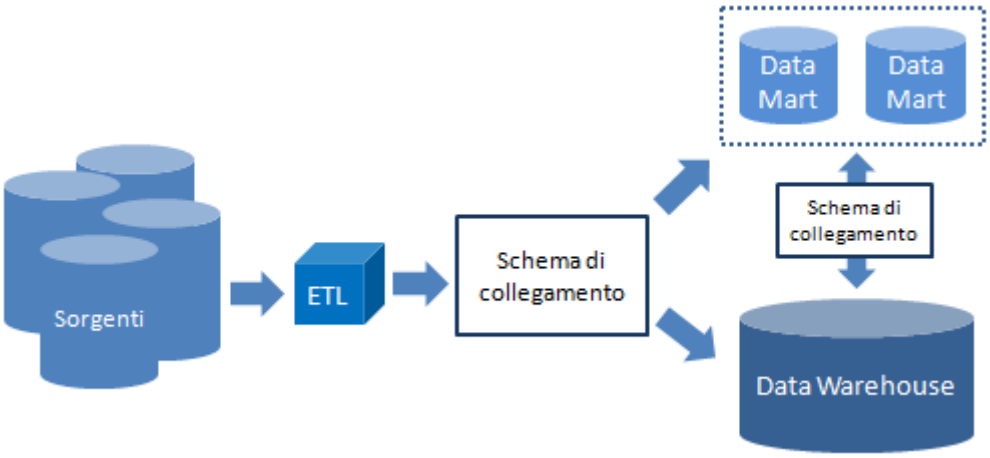

Figura 10 - Approccio incrementale per la progettazione di un Data Warehouse.

Un approccio di questo tipo minimizza i problemi di integrazione tra contenitori di dati differenti e consente di individuare e condividere dimensioni e fatti relativi ai processi aziendali. Questo tipo di approccio è sostenuto anche da Kimball e in molti casi consente di ottenere risultati simili all'approccio top-down ma in maniera più semplice e veloce.

L'approccio incrementale è la base di partenza per la definizione della metodologia di integrazione, conosciuta come Data Federation, che viene analizzata nel paragrafo 1.3.2.

1.3.1.5 DATA MART

Con l'aumento delle dimensioni e della complessità dei Data Warehouse i diversi dipartimenti aziendali hanno iniziato a manifestare la necessità di avere un contenitore di dati strutturato secondo le proprie esigenze, che contenesse soltanto informazioni utili a perseguire i propri obiettivi specifici. Per questo motivo si cominciò a pensare di creare dei contenitori a valle del Data Warehouse principale in grado di focalizzarsi ognuno sulle esigenze di una ristretta area aziendale. Un Data Mart e un Data Warehouse sono due strutture diverse che utilizzano architetture differenti, tuttavia, in una visione semplificata sono tra loro molto simili. Inmon da la seguente definizione di Data Mart [14]:

> *"A collection of subjects areas organized for decision support based on the needs of a given department."*

Un Data Mart è quindi un contenitore di dati specializzato su un particolare soggetto che, solitamente, si colloca a valle di un Data Warehouse, dal quale si alimenta. In termini tecnici un Data Mart è un sottoinsieme, logico o fisico, di un Data Warehouse e si pone il fine di servire le esigenze di una singola divisione aziendale (Figura 11).

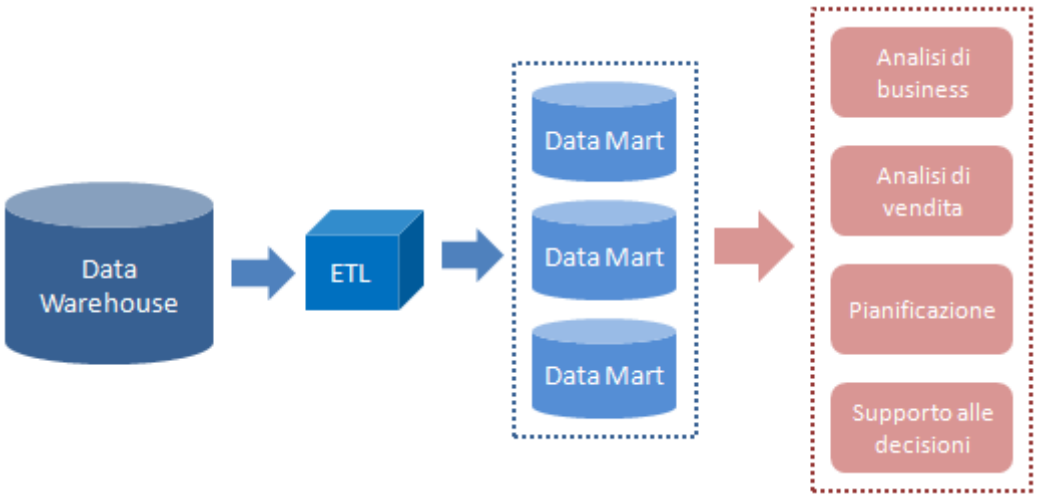

Figura 11 - Architettura di una soluzione di Data Warehousing con Data Mart.

La realizzazione di uno o più Data Mart a valle di un Data Warehouse comporta una serie di vantaggi:

- Possibilità di utilizzare uno schema di dati differente, mirato ad ottimizzare le analisi di business richieste da una particolare area aziendale, aumentando di fatto la flessibilità dell'intero sistema di Data Warehousing;
- Aumento delle performance potendo disporre di un hardware dedicato per singolo Data Mart;
- Permettere ad alcuni soggetti di accedere ad un insieme di dati più ristretto, garantendo una maggiore sicurezza del sistema.

Con l'affermarsi dell'approccio illustrato in Figura 11 si è creata una certa confusione interna alla disciplina del Data Warehousing. In molti si convinsero che fosse possibile costruire uno o più Data Mart e poi fonderli in un Data Warehouse unico, oppure trasformare un Data Mart in un Data Warehouse quando diventava troppo grande. Inmon sottolinea che si tratta di idee errate, le strutture di un Data Mart e di un Data Warehouse sono profondamente differenti, un passaggio da un'architettura all'altra non è possibile per una serie di motivi [14]:

- Un Data Mart è creato per soddisfare gli obiettivi di un singolo soggetto interno all'azienda mentre un Data Warehouse nasce con l'obiettivo di adattarsi ai bisogni dell'intera organizzazione. Un struttura di dati può essere funzionale per un singolo dipartimento o per l'intera azienda ma non per entrambi allo stesso tempo;
- Il livello di granularità dei dati è profondamente diverso tra un Data Mart e un Data Warehouse. Il Data Mart contiene solitamente dati aggregati e sintetizzati, il Data Warehouse contiene dati al maggior livello di dettaglio possibile;
- Un Data Warehouse contiene una grande quantità di dati storicizzati, mentre un Data Mart si riferisce ad intervalli temporali più brevi.

1.3.1.6 VANTAGGI E SVANTAGGI

La disciplina del Data Warehousing è oggi matura e consolidata e, nonostante sia sul mercato da diversi anni, ha saputo evolversi per non restare obsoleta dimostrando abilità nell'adattarsi a nuove esigenze. Questi anni di evoluzione hanno permesso a tale metodologia di maturare, presentando soluzioni di integrazione dei dati che garantiscono una serie di vantaggi:

- Un Data Warehouse garantisce un modello unico per tutti i dati di interesse per l'azienda indipendentemente dalla loro fonte. Questo facilita di molto le successive operazioni di analisi e utilizzo dei dati. L'esecuzione di query complesse è di molto facilitata potendo contare su un ambiente unico su cui operare;
- Il caricamento dei dati in un ambiente unico omogeneo tramite tecniche ETL permette di bonificare i dati rimuovendo inconsistenze, errori e duplicazioni, garantendo così la realizzazione di un database integrato contenente le singole entità con informazioni corrette;
- Il Data Warehouse può essere utilizzato come fonte per i sistemi di supporto alle decisioni per migliorare i processi decisionali;
- Un Data Warehouse è in grado di storicizzare una grande quantità di informazioni rendendo possibili complesse analisi sui dati storici con il fine di prevedere al meglio l'andamento futuro del business.

Tuttavia, si possono ravvisare anche degli svantaggi nell'utilizzo e nella realizzazione di un Data Warehouse:

- La progettazione, preparazione e realizzazione di un Data Warehouse può essere un'attività molto complessa che richiede molto tempo e ingenti risorse;
- Anche una volta realizzato, il sistema presenta dei costi di manutenzione che in determinate situazioni risultano molto elevati. I Data Warehouse moderni non sono statici e i costi per aggiornare la struttura sono spesso elevati;
- I processi ETL possono essere molto complessi e richiedere molto tempo per ottenere un buon risultato. Processi ETL complessi comportano inoltre la presenza di un tempo di latenza nel caricamento dei dati.

Alla luce di queste considerazioni la disciplina del Data Warehousing è chiamata ancora una volta ad evolversi per restare al passo con i tempi. In particolare, si possono identificare alcune strategie di evoluzione dei Data Warehouse [12]:

- **Scalabilità e prestazioni**: la quantità di dati processabili dai sistemi di Data Warehousing è in continuo aumento senza presentare decadimenti prestazionali. Inoltre, anche la scalabilità di questi sistemi è in continuo aumento per venire incontro alle esigenze dettate dai più recenti andamenti del mercato;

- **Real-time Data Warehousing**: le tecnologie di memorizzazione e trasferimento dei dati si sono evolute a tal punto che è oggi possibile supportate un trasferimento continuo di grandi quantità di dati per lunghi periodi di tempo. Questo garantisce ai moderni Data Warehouse di ottenere i dati in tempo reale appena vengono inseriti nei sistemi sorgente. Ottimizzando le operazioni di ETL tali dati saranno memorizzati nel Data Warehouse in tempi brevissimi riducendo o eliminando completamente i tempi di latenza. In alcuni settori disporre dei dati in tempo reale è fondamentale;
- **Qualità del dato**: garantire la qualità del dato è la nuova frontiera del Data Warehousing. Poter disporre di un dato corretto è fondamentale per prendere le giuste decisioni. Per questo motivo le moderne soluzioni di Data Integration presentano componenti avanzati per la gestione della qualità del dato, che si integrano anche con i sistemi sorgente per evitare a monte l'inserimento di dati errati.

1.3.2 DATA FEDERATION

Il concetto di "database federato" ha origine nel corso degli anni '80 grazie ai contributi di Dennis Heimbigner e Dennis McLeod, rispettivamente ricercatori all'Università del Colorado e della California. Il termine compare per la prima volta nel corso della AFIPS[16] National Computer Conference, tenutasi a Anaheim in California nel 1980, dove i due ricercatori abbozzano l'idea di un'architettura federata per i Database Systems [15]. Heimbigner e McLeod approfondiscono il concetto e cinque anni più tardi, 1985, pubblicano un articolo in cui descrivono il funzionamento di un'architettura federata per sistemi di Data Management, nel quale troviamo una definizione sintetica di architettura federata [16]:

> *"A collection of indipendent database systems united into a loosely coupled federation in order to share and exchange information."*

In altre parole un database federato è una tipologia di Database System che integra diverse basi di dati in una singola vista unificata, dove i database sorgente sono collegati per via telematica e di conseguenza possono essere geograficamente decentralizzati. Un database federato lascia i sistemi sorgente completamente autonomi, i dati non vengono trasferiti o replicati in un ambiente integrato centrale ma vengono letti per via telematica e integrati in un database federato (o virtuale). Un Federated Database System (FDBS) è quindi una particolare tipologia di database che Heimbigner e McLeod definiscono come "Physically

[16] American Federation of Information Processingg Societies, un'organizzazione statunitense fondata nel 1961 con l'obiettivo di sviluppare le tecnologie di elaborazione di dati e informazioni all'interno degli Stati Uniti.

Decentralized Databases". I due autori classificano le diverse tipologie di sistemi per la memorizzazione dei dati secondo due dimensioni: struttura/organizzazione logica e struttura/organizzazione fisica. Ogni dimensione può essere a sua volta di tipo centralizzato o decentralizzato. Combinando questi attributi si ottengono diverse classi di database:

- **Database integrati tradizionali**: caratterizzati da un'architettura centralizzata sia a livello logico che fisico;
- **Database distribuiti**: architettura logica centralizzata e fisica decentralizzata;
- **Database federati**: contraddistinti da un'architettura logica decentralizzata e da un'architettura fisica che può essere sia di tipo centralizzato che decentralizzato.

Il vantaggio di un database federato è costituito dal fatto che è possibile realizzare una singola interrogazione (query) in grado di recuperare i dati attraverso sistemi diversi collocati in ambienti e architetture differenti. Per fare questo il sistema federato deve dividere la singola interrogazione in una serie di query secondarie in grado di accedere ai dati nei singoli database sorgente, per poi ricomporre i risultati in un output unificato. L'esistenza di molteplici linguaggi di interrogazione implica che un sistema federato sia in grado di risolvere query attraverso una moltitudine di linguaggi differenti.

I primi prodotti basati sulla tecnologia dei database federati risalgono ai primi anni '90 e segnano la nascita di una nuovo segmento del mercato. Tali prodotti vengono identificati con l'appellativo di Enterprise Information Integration (EII), termine coniato per indicare tutti quei prodotti che, tramite l'utilizzo della tecnologia dei database federati, assicurano un accesso unificato ai dati senza il bisogno di realizzare un Data Warehouse [17]. La disciplina si pone con l'obiettivo di superare i limiti tipici delle tradizionali soluzioni di Data Warehousing per rispondere alle nuove esigenze del mercato. Con l'avvento della rete Internet cresce l'esigenza di realizzare servizi Web in grado di utilizzare, analizzare ed elaborare dati provenienti da una moltitudine di fonti differenti, situazione per la quale l'utilizzo di un Data Warehouse è ritenuto inadeguato. L'affermazione del Web comporta un notevole aumento della quantità di dati esistenti; duplicare grandi quantità di dati, spesso contraddistinti da un elevata variabilità, in un database integrato non è più una strada percorribile.

Uno dei primi sistemi commerciali ad utilizzare la tecnica federata è il DB2 DataJoiner di IBM [18] che garantisce la presenza di un motore di integrazione in grado di accedere ai dati di sorgenti differenti attraverso la realizzazione di una singola query. Tale tecnologia è stata sviluppata e migliorata negli anni ed oggi il Database Management System di IBM (DB2) fornisce all'utente funzionalità di Data Federation consolidate.

1.3.2.1 ARCHITETTURA

McLeod e Heimbigner affermano che un'architettura federata necessita di un corretto bilanciamento tra due requisiti contrastanti: autonomia e condivisione. Ogni componente del sistema federato deve avere quanta più autonomia possibile, ma, allo stesso tempo, ogni componente deve essere in grado di condividere la giusta quantità di informazioni con l'intero sistema. L'elemento di base di un sistema federato sono i componenti, dove un componente identifica una singola fonte informativa che vuole condividere e scambiare informazioni. Un componente può quindi essere visto come un database autonomo che entra in contatto con tre schemi che gestiscono il funzionamento del singolo componente. I tre schemi necessari sono:

- **Schema privato**: descrive le caratteristiche della parte di dati memorizzata in locale nel database (che tipo di dati sono contenuti, con quale struttura e quale è il loro significato). Contiene inoltre informazioni specifiche sul componente, come ad esempio configurazione dell'ambiente e della rete;
- **Schema di esportazione**: specifica la porzione di informazioni che il componente vuole condividere con l'intero sistema federato. È costituito da un insieme di mappe che definiscono come esportare le informazioni;
- **Schema di importazione**: specifica le informazioni che il componente vuole ottenere da un altro componente del sistema.

Tuttavia, le più moderne tecnologie di Data Federation prevedono l'utilizzo di un'architettura che utilizza altre tipologie di schemi in modo da rendere il sistema più efficiente e poter gestire situazioni più complesse. In particolare si riscontra l'utilizzo dei seguenti schemi aggiuntivi:

- **Schema federato**: si tratta dell'integrazione di più schemi di esportazione. Contiene informazioni sulla distribuzione e integrazione dei dati tra le varie sorgenti che si genera con l'esecuzione di schemi di esportazione;
- **Schema esterno**: uno schema per poter esporre i dati ad applicazioni/utenti esterni.

In una visione semplificata, i moderni strumenti di Data Federation utilizzano l'architettura rappresentata in Figura 12. Per ogni database sorgente sono previsti i corrispettivi schemi interno, concettuale e di esportazione che permettono di rendere disponibili al sistema le singole sorgenti di dati. Tra questi e il database federato si colloca uno schema intermedio, definito mediatore, che ha il compito di gestire l'interscambio di dati tra le diverse componenti architetturali. Il mediatore è costituito da agenti software in grado di tradurre le query da un formato globale (adottato dal sistema federato) al linguaggio specifico della singola base dati sorgente. Il mediatore scompone e traduce la query globale, la invia ai rispettivi database di origine per poi ricomporre i risultati in un singolo database federato. Il database federato è il contenitore unico per i dati inviati dal mediatore ed è gestito da un

apposito Federation Server in grado di recuperare i dati dalle sorgenti, di aggregarli e di fornirli in output per i sistemi di analisi.

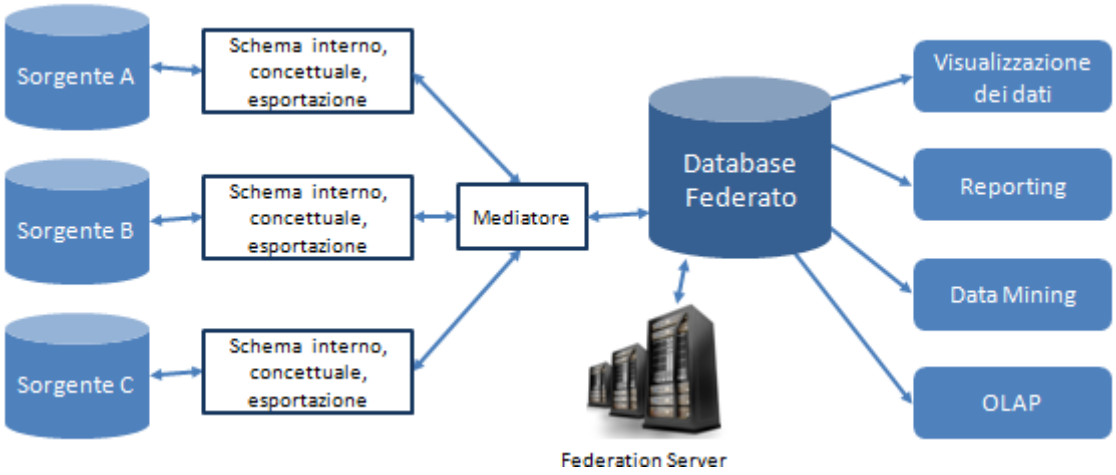

Figura 12 - Architettura generica di un sistema di Data Federation.

1.3.2.2 VANTAGGI E SVANTAGGI

Rispetto ad una soluzione basata su tecnologie di Data Warehousing un database federato aggiunge un altro componente (lo schema di collegamento) tra applicazione e dati, la cui introduzione comporta un compromesso in termini di performance. Tuttavia, i moderni prodotti di Data Federation offrono potenti strumenti di ottimizzazione del codice SQL che uniti alla potenza di calcolo degli elaboratori moderni fanno si che la presenza di un livello aggiuntivo non incida particolarmente sulle prestazioni dell'intero sistema federato.

Riassumendo è possibile elencare quelli che sono i vantaggi dati dalla realizzazione di una soluzione di integrazione dei dati tramite l'utilizzo di tecnologie di Data Federation:

- *Ottima scalabilità del sistema*: in un'architettura federata, a differenza di quanto accade in un Data Warehouse, è possibile aggiungere velocemente e con costi ridotti ulteriori fonti di dati. Tuttavia, aggiungere un elevato numero di fonti ulteriori rischia di creare problemi di performance della rete e dello schema di traduzione delle query. Inoltre, l'integrità dei dati rischierebbe di risultare compromessa;
- *Flessibilità*: un sistema federato risente minimamente di cambi di schema o di tecnologia dei sistemi sorgente, riuscendo ad adattarsi senza problemi a situazioni che prevedono evoluzioni molto frequenti;
- *Riduzione dei costi*: a differenza di un sistema di Data Warehousing che richiede ingenti investimenti per realizzare un'infrastruttura in grado di contenere i dati migrati dai sistemi di origine, un database federato lascia intatti i sistemi sorgenti limitandosi a collegare i loro dati in una vista unica virtuale. In generale, una

soluzione di Data Federation richiede meno infrastrutture rispetto a un Data Warehouse;

- *Nessun tempo di latenza*: ogni aggiornamento che viene inserito nei sistemi sorgente è da subito disponibile anche nel database federato;
- *Non è necessario replicare i dati*: l'approccio federato elimina o minimizza la replicazione dei dati. In alcune occasioni è comunque necessario replicare alcuni dati per incrementare le performance del sistema.

Nonostante presenti una discreta serie di vantaggi lo sviluppo di una soluzione di Data Federation comporta la presenza di alcuni svantaggi e criticità che in determinate situazioni fanno preferire lo sviluppo di una soluzione di Data Warehousing:

- *Dipendenza dai sistemi di origine*: un database federato dipende fortemente dai sistemi di origine, se tali sistemi si rivelano scarsamente affidabili i problemi si propagheranno all'intera architettura federata in misura dell'importanza della singola fonte di dati;
- *Performance*: nonostante i moderni prodotti di Data Federation garantiscano performance di integrazione dei dati pari o superiori rispetto a un Data Warehouse, tale situazione richiede lo sviluppo di query e algoritmi altamente ottimizzati e performance di rete di un determinato livello. In caso contrario le performance dell'intero sistema diminuiscono drasticamente;
- *Competenze*: lo sviluppo di un database federato richiede la conoscenza di una molteplicità di tecnologie, linguaggi, schemi e architetture differenti. La disponibilità di tecnici esperti e competenti è fondamentale;
- *Difficoltà di bonifica dei dati*: il fatto che i dati risiedano sui sistemi sorgenti comporta delle difficoltà nell'effettuare operazioni di bonifica e pulizia dei dati, rendendo complessa la gestione di situazioni nelle quali le fonti presentano una qualità del dato insufficiente.

1.3.3 CONCLUSIONI

Arrivati a questo punto è lecito chiedersi se, ed eventualmente quando, è da preferire una soluzione di Data Federation all'implementazione di un Data Warehouse. Alla luce di quanto visto finora non è possibile dire con certezza che una delle due soluzioni è migliore dell'altra, tuttavia si possono identificare alcuni scenari di utilizzo cercando di capire quale delle due metodologie si presta di più in base alle necessità riscontrabili in una situazione specifica. Vediamo quindi alcuni possibili scenari di applicazione in base a differenti necessità [19]:

- *Variabilità dei dati*: nell'approccio del Data Warehouse le trasformazioni dei dati avvengono tramite tecniche ETL nel momento in cui il dato viene estratto mentre

nell'approccio federato la trasformazione avviene ogni qual volta si esegue la query. Se una parte dei dati è interrogata molte volte è senz'altro preferibile effettuare una sola volta la trasformazione, di conseguenza il Data Warehouse appare avvantaggiato. Viceversa, quando il dato è interrogato una sola volta l'approccio federato è preferibile. Riassumendo, quando un dato è statico e interrogato raramente è preferibile l'approccio federato, quando invece il dato è variabile e frequentemente interrogato è preferibile l'approccio ETL tipico del Data Warehouse;

- *Indicizzazione*: in un Data Warehouse è possibile separare la fase di acquisizione dei dati dalla fase di ottimizzazione/indicizzazione che prepara i dati per analisi di tipo OLTP[17]. Nell'approccio federato entrambi i processi vanno eseguiti in un singolo database rendendo più complicata l'ottimizzazione delle due fasi. In situazioni caratterizzate da elevate quantità di dati un approccio federato potrebbe risultare difficilmente gestibile e le sue performance risulterebbero comunque minori;
- *Velocità di risposta*: in un Data Warehouse è solitamente presente un componente dedicato all'ottimizzazione dei dati per gli applicativi di Business Intelligence. In un database federato viene data priorità all'esecuzione dei processi OLTP determinando così un tempo di risposta maggiore per gli applicativi di Business Intelligence;
- *Complessità e variabilità dello schema*: in un Data Warehouse lo schema globale e le relazioni tra gli elementi del database relazionale sono definite una volta per tutte nel corso delle attività ETL di caricamento. In uno schema federato i collegamenti tra i diversi contenitori di dati vengono definiti ogni volta che viene eseguita una query. L'approccio del Data Warehouse è avvantaggiato nel caso in cui lo schema globale è complesso e non si presentano variazioni di schema delle singole basi dati di partenza. L'approccio federato risulta vantaggioso quando i database sorgenti cambiano spesso la loro struttura o quando lo schema globale non è troppo complesso;
- *Aggiornamento dei dati*: un Data Warehouse contiene fondamentalmente dati vecchi per buona parte del periodo che trascorre tra un caricamento e l'altro. D'altra parte, un approccio federato contiene dati costantemente aggiornati. Se c'è la necessità di avere dati costantemente aggiornati un database federato è senz'altro la soluzione da preferire. Tuttavia, i più moderni prodotti di Data Integration

[17] *On Line Transaction Processing*, è un insieme di tecniche utilizzate per l'analisi dei dati. A differenza delle analisi di tipo OLAP (*Online Analytical Processing*), la tecnologia OLTP non prevede la creazione di banche dati separate. Tale soluzione permette di avere dati sempre aggiornati ed evita fasi intermedie di trasformazione. Tuttavia, non è facilmente applicabile in situazioni caratterizzate da grandi quantità di dati, casi in cui è preferibile l'utilizzo di analisi OLAP.

garantiscono funzionalità avanzate di Data Warehousing che permettono di eseguire più operazioni in parallelo. In questo modo si possono caricare nuovi dati mentre sono in esecuzione delle query riducendo drasticamente la presenza di dati non aggiornati;

- *Trasformazione dei dati*: l'approccio ETL tipico di un Data Warehouse permette di effettuare complesse operazioni di trasformazione e bonifica di dati che permettono una corretta identificazione ed integrazione delle entità. I database federati supportano tali funzionalità ma ci sono alcune tipologie di trasformazioni difficilmente realizzabili in tempo reale.

In base a questa breve analisi l'approccio di Data Integration più completo ed affidabile appare essere quello del Data Warehouse. Tuttavia, le evoluzioni più recenti evidenziano come il Web e le sue tecnologie siano sempre più utilizzate per la gestione e la fornitura di dati tramite web-service rendendo di fatto sempre più numerose e varie le fonti di dati esistenti. In un contesto di questo tipo, con migliaia di fonti di dati differenti, una soluzione di Data Warehousing diventa irrealizzabile e l'approccio federato si dimostra senza dubbio preferibile. Nell'ambito dell'integrazione dei dati interni ad un'azienda resta comunque tutt'oggi preferibile la realizzazione di un Data Warehouse.

Nel capitolo 3 vedremo come l'approccio federato si sta trasformando in relazione all'evoluzione del Web e delle sue tecnologie, in particolare del Web semantico, abbracciando nuovi linguaggi di rappresentazione dei dati.

1.4 DATA QUALITY

La qualità del dato è da sempre una componente importante per le organizzazioni e oggi, in un mercato sempre più competitivo, assume un ruolo di fondamentale rilevanza per l'azienda e per il suo successo nei business in cui opera.

Una possibile definizione di Data Quality è quella riportata nello standard ISO 8402:

> *"The totality of characteristics of an entity that bear on its ability to satisfy stated or implied needs."*

Una definizione di questo tipo ci dice che la qualità del dato non dipende solo dalle caratteristiche del dato stesso ma anche dal contesto di business in cui è utilizzato. La qualità del dato è un componente critico dell'organizzazione, non implementare una strategia di valutazione e controllo della qualità dei dati che si possiedono può avere effetti disastrosi.

Jim Harris[18] sintetizza così gli effetti collaterali derivanti da una scarsa qualità dei dati [20]:

> *"Poor data quality is the path to the dark side. Poor data quality leads to bad business decisions. Bad business decisions leads to lost revenue. Lost revenue leads to suffering."*

La presenza di una scarsa qualità dei dati non è un problema teorico ma un reale problema di business che incide negativamente sull'efficacia delle decisioni critiche prese dall'azienda. Oggi una delle principali aree di investimento per un'azienda dovrebbe essere una strategia di supporto, verifica e miglioramento della qualità del dato. Nell'ultimo decennio diverse ricerche hanno confermato il costo di una scarsa qualità dei dati, vediamo alcune analisi significative [21]:

- Nel 1998 un sondaggio ha rivelato che una scarsa qualità delle informazioni implica dei costi per l'azienda che variano dall'1% al 20% dei profitti totali[19];
- Il Data Warehousing Institute (TDWI) ha stimato che una scarsa qualità dei dati relativi ai clienti costa nei soli Stati Uniti circa 661 miliardi di dollari l'anno, da suddividere in costi di stampa, spese di spedizione e retribuzione del personale[20];
- Una scarsa qualità dei dati costa alle aziende come minimo un 10% dei profitti. Tuttavia, il 20% appare per molti una stima più ragionevole[21];
- Oltre il 25% dei dati critici per il funzionamento di un'azienda sono inaccurati o incompleti[22].

Secondo Thomas C. Redman uno degli esempi più recenti degli effetti catastrofici di una scarsa qualità dei dati è rappresentato dalla crisi dei prestiti "*subprime*" che ha investito gli Stati Uniti [22]. La crisi pone le sue basi su innovazioni finanziarie che hanno consentito l'accesso al credito a quelle classi di popolazione che prima non aveva i requisiti per accedervi. Le innovazioni sono principalmente due:

- Nuovi mutui con basse rate iniziali (minori sono i primi pagamenti più gente è qualificata a richiedere il mutuo);

---

[18] Jim Harris è un consulente che opera da oltre 15 anni nei settori della Data Quality, Data Integration, Data Warehousing, Business Intelligence, Customer Data Integration e Master Data Management. Nella sua esperienza lavorativa ha offerto consulenza per oltre 500 aziende. Harris è oggi responsabile di due blog sulla Data Quality (Data Quality Pro e Obsessive-Compulsive Data Quality) ed è membro dell'International Association for Information and Data Quality (IAIDQ) e della Data Management Association International (DAMA). La sua esperienza lo rende uno dei maggiori esperti nel campo della qualità del dato.

[19] Andrea Malcolm, "Poor Data Quality Costs 10% of Revenues, Survey Reveals," ComputerWorld, July 1998.

[20] Wayne W. Eckerson, "Data Quality and the Bottom Line," TDWI Report Series, 2002.

[21] Thomas C. Redman, "Data: An Unfolding Quality Disaster," DM Review, August 2004.

[22] Rick Whiting, "Hamstrung By Defective Data," InformationWeek, May 2006.

- *Collateralized debt obligations* (CDOs) in pratica un'obbligazione che ha come garanzia (collaterale) un debito. Una CDO è formata da decine o centinaia di obbligazioni a loro volta garantite da centinaia di debiti individuali. In questo modo è possibile ridurre e condividere il rischio.

Dati di scarsa qualità e informazioni incomplete hanno contribuito ad amplificare il crollo di tali prodotti finanziari in ogni loro fase. Innanzitutto, i dati dei mutui sono risultati spesso incorretti (talvolta falsificati per avere accesso al credito senza avere nessuna garanzia reale di ripagare il prestito). Un altro fattore critico si è dimostrato essere la scarsa accuratezza dei sistemi di previsione della probabilità di un soggetto di ripagare il debito. Inoltre, molti sottoscrittori di tali prodotti non hanno capito che i tassi di interesse sarebbero aumentati significativamente. Infine, i CDOs sono prodotti finanziari molto complessi e molti investitori non hanno capito che prodotti hanno effettivamente acquistato. Tutti questi fattori sommati uno all'altro hanno contribuito ad innescare la crisi dei subprime che ha portato conseguenze drammatiche. Migliaia di famiglie hanno perso la propria casa, numerose aziende e banche sono fallite, le istituzioni finanziarie hanno perso centinaia di miliardi di dollari, il mercato azionario ha segnato crolli del 45%. Senza contare che il mercato finanziario ha perso completamente la fiducia dei cittadini, serviranno anni per recuperarla.

Secondo Redman si possono isolare sette diverse problematiche attinenti la qualità dei dati, problematiche che affliggono tutti i dipartimenti aziendali:

- *Difficoltà nel trovare i dati*: alcuni studi indicano che i lavoratori della conoscenza spendono oltre il 30% del loro tempo alla ricerca dei dati di cui hanno bisogno[23]. Altri studi ci dicono invece che tali ricerche nel 40% dei casi sono senza successo[24];
- *Dati incorretti*: il caso più semplice in cui rendersi conto dell'inesattezza dei dati è quando i loro valori sono in disaccordo con la loro controparte reale. Alcuni studi stimano in un 10-20% la percentuale di record che contiene errori all'interno dei database aziendali[25];
- *Contenuto dei dati descritto male*: un problema ricorrente riguarda la definizione di documentazione sui dati, sulla nomenclatura degli attributi che li descrivono e sul significato reale di tali attributi. Se questi aspetti sono descritti male il medesimo dato può essere interpretato con valenze differenti all'interno della stessa organizzazione, creando confusione e disallineamento tra reparti. La realizzazione di una definizione comune e chiara dei dati è fondamentale;

[23] Susan Feldman e Chris Sherman, *The High Cost of Not Finding Information*, 2001 e LexisNexis, *Workplace Productivity Study*, 2008.
[24] Susan Feldman, "The High Cost of Not Finding Information," KMWorld 13, no.3, March 2004.
[25] Gartner Study in Rick Whiting, "Hamstrung by Defective Data," Information Week, May 8, 2006.

- *Sicurezza e privacy dei dati*: sicurezza e privacy sono due problemi che soltanto di recente hanno attirato notevole attenzione e preoccupazione. Garantire la sicurezza e la privacy dei dati è un compito oneroso e costoso, che richiede la massima attenzione;
- *Inconsistenze tra fonti diverse*: la ridondanza contribuisce ad alimentare un altro problema, l'inconsistenza. Quanto due fonti informative dicono due cose contrastanti riguardo la stessa entità si crea confusione;
- *Eccesiva quantità di dati*: spesso le organizzazioni sono caricate di una dose eccessiva di dati che non utilizzeranno mai. Un sondaggio svolto da Accenture rivela che il 40% degli IT manager si lamentano dell'eccessivo carico di informazioni che il sistema deve sostenere[26]. Non è un problema di costo per memorizzare i dati ma piuttosto un problema di tempi e costi necessari per ricercare le informazioni utili all'interno di una quantità di dati eccessiva. Inoltre, maggiore è la quantità di dati maggiori sono i problemi di ridondanza;
- *Confusione interna all'organizzazione*: molte organizzazioni non si rendono effettivamente conto del reale valore dei dati che possiedono, spesso non sanno nemmeno determinare con precisione quanti e quali dati possiedono, come li utilizzano e che grado di correttezza hanno. Le organizzazioni moderne devono sapere dare una risposta a tali interrogativi, è necessario perseguire con convinzione una strategia di gestione della qualità dei dati.

Redman conclude la sua analisi preliminare sul concetto di Data Quality raggruppando in tre categorie gli effetti negativi che si ottengono in conseguenza di una scarsa qualità del dato:

- *Attività operative*: una scarsa qualità dei dati si riflette sulle attività operative determinando costi operativi più alti, una minore motivazione del personale e una minore soddisfazione dei clienti;
- *Definizione delle strategie aziendali*: dati di bassa qualità determinano difficoltà nel definire e nell'eseguire le strategie aziendali. Vi sono inoltre minori possibilità di conseguire valore dall'utilizzo dei dati, con difficoltà nel mantenere allineati i diversi reparti dell'organizzazione. Infine, il management viene ostacolato, non potendo sapere con certezza quali dati riflettono la reale situazione dell'azienda;
- *Decision Making*: dati scarsamente affidabili portano il management a prendere decisioni non ottimali che si riflettono in perdita di fiducia dell'azienda verso il mercato, perdita di vendite, aumento del rischio di investimento e difficoltà nel gestire il rischio che si è creato.

[26] Sondaggio riportato in: Marianne Kolbasuk McKee, "The Useless Hunt for Data," Information Week, January 1-8, 2007.

1.4.1 MISURARE LA QUALITÀ DEI DATI

Determinare il livello di qualità dei dati posseduti da un'azienda è un'operazione complessa. Per determinare la bontà dei dati è necessario definire delle metriche attraverso le quali misurare la qualità dei dati. Tuttavia, è molto difficile definire delle metriche universalmente valide in quanto la correttezza dei dati è profondamente legata ai singoli contesti operativi. La qualità del dato è un concetto multidimensionale la cui valutazione implica la definizione di metriche soggettive, adattabili ad un particolare contesto di business. È comunque possibile tentare di definire delle metriche universali indipendenti dal contesto di utilizzo dei dati. Pertanto si possono individuare due tipologie di valutazione [23]:

- *Indipendenti dal contesto o oggettive*: metriche che riflettono lo stato dei dati senza considerare come e dove vengono utilizzati;
- *Dipendenti dal contesto o soggettive*: misurazioni che tengono in considerazione il contesto di utilizzo, regole, caratteristiche e vincoli del business di riferimento.

Dei possibili indicatori per accertare la qualità dei dati indipendentemente dal contesto di utilizzo sono proposti da Thomas Redman in [22]. Redman propone due semplici indicatori in grado di determinare il livello di correttezza di un insieme di dati:

- Correttezza a livello di attributi $= 1 - \frac{numero\ di\ attributi\ errati}{numero\ totale\ di\ attributi}$
- Correttezza a livello di record $= 1 - \frac{numero\ di\ record\ errati}{numero\ totale\ di\ record}$

In un database contente 100 record, dove ogni record è composto da 12 attributi, ipotizzando la presenza di 20 errori in record diversi si otterrebbero i seguenti risultati:

- A livello di attributi $= 1 - 20/1200 = 98\ \%$
- A livello di record$= 1 - 20/100 = 80\ \%$

Secondo Redman il livello di correttezza a livello di record è un buon indicatore di qualità della base di dati in quanto permette di identificare la percentuale di record che contengono degli errori. Tuttavia, senza tenere conto del contesto di utilizzo dei dati tali misurazioni potrebbero risultare falsate. Nel capitolo 5 è proposta un'applicazione di tali misurazioni al fine di valutare la qualità generale delle basi dati oggetto del caso di studio di Trentino Riscossioni S.p.A. Altre tipologie di metriche oggettive fanno uso di tecniche matematico-statistiche per determinare il livello di completezza e correttezza dei dati. Ad esempio è possibile utilizzare l'analisi dell'andamento temporale dei dati permette di determinare gli scostamenti dal valore atteso e di identificare eventuali problematiche.

La definizione di metriche in grado di considerare il contesto passa dalla definizione delle dimensioni attraverso cui valutare la qualità dei dati. Per determinare quali sono i criteri più rilevanti rispetto a cui misurare la qualità dei dati in un determinato contesto molte

organizzazioni fanno compilare dei questionari agli utenti operanti nel contesto in oggetto. Le principali dimensioni da tenere in considerazione sono le seguenti [21] [23]:

- **Accessibilità**: indica la facilità con cui un utente può identificare, ottenere ed utilizzare i dati;
- **Comprensibilità**: determina quanto i dati sono facili da comprendere;
- **Accuratezza**: si riferisce alla differenza tra una stima di come dovrebbe essere valorizzato un attributo e il valore effettivo riportato dai dati;
- **Attendibilità**: indica il grado di credibilità e affidabilità dei dati, dipende dall'attendibilità della fonte di origine;
- **Completezza**: è una misura di corrispondenza tra il mondo reale e il dataset specifico. Indica quanti e quali dati mancano nel dataset per offrire una rappresentazione completa al 100% del contesto reale;
- **Consistenza**: il grado di consistenza dei dati, per ottenere una rappresentazione consistente i dati all'interno di un dataset devono essere strutturati nello stesso modo;
- **Correttezza**: il grado di esattezza e affidabilità dei dati;
- **Interpretabilità**: si riferisce alla disponibilità di una documentazione della base dati chiara e precisa che indichi agli utenti che tipologie di dati sono contenute nel database, come utilizzare e analizzare i dati;
- **Manipolabilità**: indica il grado di facilità con cui i dati possono essere elaborati per scopi differenti;
- **Oggettività**: indica l'imparzialità, l'obiettività dei dati;
- **Puntualità**: sta ad indicare quanto i dati sono aggiornati rispetto al contesto reale. È una misura di allineamento temporale della base dati rispetto al mondo reale e costituisce un indicatore di fondamentale importanza. Lavorare su dati obsoleti può portare a prendere decisioni critiche errate;
- **Quantità**: indica quanto è appropriato il volume di dati posseduti in riferimento ad una determinata attività. Lavorare con più o meno dati del necessario può rivelarsi controproducente e difficile da gestire;
- **Rilevanza**: indica quanto i dati sono appropriati e di aiuto in un determinato contesto applicativo;
- **Utilità**: indica quanti e quali benefici l'utilizzo dei dati apporta all'azienda. È una misura del valore aggiunto portato dall'utilizzo dei dati.

A partire da tali dimensioni un'organizzazione deve definire delle metriche ad hoc in grado di determinare la qualità dei dati nel proprio contesto di business. Accertare la qualità dei dati implica solitamente un processo sintetizzabile in tre fasi [23]:

1. Applicare metodi di misurazione oggettivi e soggettivi;

2. Comparare i risultati per identificare le possibili problematiche;
3. Determinare e intraprendere le azioni correttive necessarie.

1.4.2 DATA QUALITY MANAGEMENT

Nel contesto economico attuale i dati e le informazioni sono una risorsa aziendale di fondamentale importanza; la gestione della qualità dei dati assume oggi un'importanza critica. Secondo Thomas Redman, nonostante l'esistenza di decine di strumenti e metodologie a supporto della qualità dei dati, si possono sintetizzare essenzialmente due approcci [22]:

1. Cercare e correggere gli errori nei dati;
2. Prevenire gli errori alla fonte.

Molte aziende perseguono erroneamente la prima via, intraprendendo una strada spesso senza via d'uscita. L'approccio più corretto è il secondo, il che implica che è necessario intervenire sui sistemi che producono i dati (processi di business) e che li immettono nei database. Ridurre la quantità di dati errati alla fonte è molto più efficace che consolidare la qualità dei dati nel database finale. Infatti, intervenendo sui processi di origine si prevengono anche gli errori futuri riducendo il carico di lavoro necessario per sistemare i dati errati nel database.

Tuttavia, gestire la qualità dei dati richiede unità e programmi dedicati, iniziative a supporto della Data Quality sono sempre più frequenti nelle moderne realtà aziendali. SAP[27] propone un modello a cinque fasi per la gestione della Data Quality [21]:

1. **Data Assessment**: misurare ed analizzare (profilare) i dati per determinare il loro livello di correttezza;
2. **Data Cleaning**: pulire, correggere, standardizzare i dati anomali;
3. **Data Enhancement**: migliorare il contenuto del dataset aggiungendo dati che possono aumentare il valore informativo della base dati;
4. **Match & Consolidate**: risolvere e rimuovere dati duplicati, inconsistenti e ridondanti;
5. **Monitoring**: la qualità dei dati va continuamente monitorata per identificare e prevenire eventuali problemi.

Per concludere, è importante sottolineare che sviluppare un programma efficiente per la gestione della qualità dei dati non è una cosa facile. Sono richiesti investimenti in tecnologie e personale dedicato che rappresentano un costo notevole per le aziende, ma i ritorni possono essere veramente elevati. Purtroppo molte organizzazioni non hanno ancora

[27] Multinazionale tedesca tra i leader mondiali nel settore degli ERP che in seguito all'acquisizione di Business Objects ottiene notevoli competenze anche nel campo della Data Quality e Business Intelligence.

compreso l'importanza di gestire la qualità dei propri dati sviluppando strategie mirate, limitandosi spesso a singoli interventi correttivi in caso di problemi ma senza fare nulla per sistemare il problema alla fonte. Nel mercato odierno chi si è mosso in anticipo ed ha sviluppato delle strategie per migliorare la qualità dei propri dati si trova oggi in una posizione di vantaggio competitivo rispetto ai concorrenti [24].

1.5 CONCLUSIONI

L'evoluzione delle tecnologie di Data Management ha permesso la realizzazione di strumenti a supporto dell'integrazione dei dati sempre più sofisticati e ricchi di funzionalità. Fin dalla comparsa dei primi sistemi in grado di produrre e gestire dati si è resa evidente la necessità di disporre di una visione unificata ed integrata dei dati posseduti dall'azienda. In particolare, la necessità di metodologie e tecnologie di Data Integration diventa un bisogno primario a partire dagli anni '90, durante i quali si assiste ad una crescita esponenziale dei sistemi informatici e conseguentemente della quantità di dati disponibile.

In questo contesto di espansione esponenziale si affermano principalmente tre metodologie di integrazione dei dati, sulle quali si basano le prime tecnologie di Data Integration. Abbiamo innanzitutto la metodologia definita "Data Consolidation" che prevede il caricamento di tutti i dati di origine in un database unico; su tale metodologia si basa la tecnologia del Data Warehouse. Un altro metodo di integrazione è quello che viene definito "Data Propagation", ovvero un sistema che prevede la realizzazione più database integrati ognuno dedicato ad una specifica applicazione; questa metodologia trova applicazione nei sistemi di integrazione basati su Data Mart. Infine, troviamo la metodologia definita "Data Federation" che non prevede lo spostamento fisico dei dati in un database unificato, ma la realizzazione di una vista unificata virtuale, recuperando i dati dai singoli database di origine per via telematica.

Su queste tre metodologie di integrazione si sviluppano quindi i primi sistemi di Data Integration (data warehouse, data mart e database federati) ognuno con i propri vantaggi e svantaggi. Tuttavia, pur essendo l'approcio del Data Warehouse quello riconosciuto dal mercato come più completo e funzionale, esistono contesti applicativi nei quali lo sviluppo di soluzioni alternative appare preferibile. Come vedremo nel capitolo successivo, i moderni strumenti di Data Integration raggruppano più metodologie di integrazione in un prodotto unico, al fine di ottenere uno strumento in grado di rispondere ad ogni tipologia di necessità e che possa minimizzare gli svantaggi della singola metodologia. Inoltre, i moderni prodotti di integrazione garantiscono la presenza di una lunga serie di funzionalità aggiuntive, tra cui la gestione della qualità del dato che riveste un aspetto di fondamentale importanza negli attuali contesti di business.

CAPITOLO 2

IL MERCATO DELLA DATA INTEGRATION

In questo capitolo è riportata un'analisi della situazione attuale del mercato relativo ai prodotti di Data Integration. Sono descritte brevemente le caratteristiche che oggi il mercato richiede a un software di integrazione dei dati e successivamente sono presentati i principali prodotti disponibili, evidenziando per ognuno pregi, difetti e punti di forza. In seguito, è proposta un'analisi dell'approccio open source, approccio che promette di superare i limiti dei software proprietari garantendo l'accesso a soluzioni di integrazione dei dati anche alle organizzazioni con risorse ed esigenze limitate. Vengono quindi descritti i principali prodotti open source, mettendo in luce vantaggi e svantaggi di un approccio di questo tipo.

Il capitolo prosegue con un'analisi dei costi della Data Integration prendendo come esempio alcuni dei prodotti presentati. Attraverso questa analisi si cerca di capire quanto un progetto di integrazione dei dati impatta sull'economia di un'azienda, valutando quanto tali soluzioni sono alla portata di imprese di piccole-medie dimensioni con risorse limitate.

Il capitolo si chiude con un'analisi della situazione attuale del settore della Data Integration, mettendo in luce, dopo oltre 30 anni di esperienza, a che punto si trovano i software di Data Integration, quali problemi restano ancora oggi irrisolti e quali sono i possibili sviluppi futuri.

2.1 ANALISI DEL MERCATO

Nel paragrafo 1.3 abbiamo visto che secondo il punto di vista di Gartner gli applicativi di Data Integration si pongono al centro delle infrastrutture basate su dati e informazioni, garantendo la possibilità di superare i tipici problemi di condivisione dei dati [25]. Per questo tipo di prodotti il mercato è oggi quanto mai fertile e da più direzioni arrivano pressioni per aumentare gli investimenti in soluzioni di integrazione dei dati. Negli ultimi anni il mercato ha subito una forte evoluzione che ha portato ad una netta riduzione dei tempi richiesti per svolgere determinate attività (Figura 13). Tempi di risposta minori implicano la realizzazione di soluzioni di integrazione sempre più veloci ed efficienti; per questo l'adozione di prodotti di Data Integration è vista dalle imprese sempre più come una base strategica al fine di mantenere la competitività nel business di riferimento. Oggi assistiamo inoltre ad un'evoluzione dei processi di business, dove è richiesta una semplificazione dei processi e delle infrastrutture informatiche necessaria per garantire

maggiore trasparenza. Per raggiungere questo obiettivo è essenziale disporre di una vista completa e consistente dei dati che si possiedono. Anche in questo contesto i prodotti di Data Integration diventano una componente chiave per l'intera infrastruttura informativa. I fornitori di strumenti di integrazione in grado di rispondere al meglio alle necessità delle aziende hanno oggi grandi possibilità di espandere il proprio mercato.

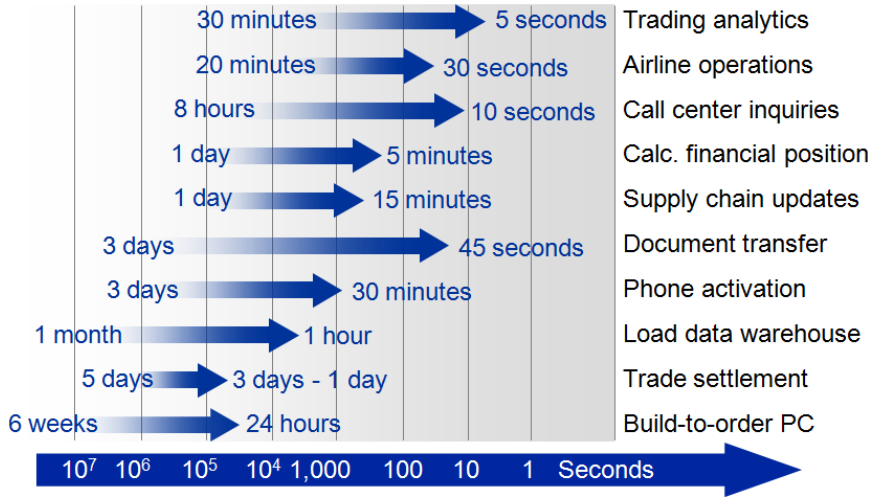

Figura 13 – Diminuzione dei tempi di risposta in alcuni contesti di business (Fonte: [25]).

Da un punto di vista tecnologico i primi sistemi di Data Integration venivano inclusi tradizionalmente in prodotti per mercati collegati. Ad esempio, i prodotti ETL hanno fatto per anni la parte del leone e hanno contribuito a creare un punto d'incontro per la consolidazione del mercato. Tuttavia, anche i prodotti di Business Intelligence vengono forniti con componenti di Data Integration con particolare enfasi sui metadati. Vi sono poi altri mercati influenzati dall'integrazione dei dati che in qualche si sovrappongono allo spazio destinato ai prodotti di integrazione (data quality, data modeling, ecc.). Tutto questo ha portato ad una eccessiva frammentazione del mercato della Data Integration e ha contribuito a creare una situazione complessa all'interno delle aziende, dove divisioni differenti utilizzano prodotti diversi dando vita a diversi problemi: scarsa consistenza, molte sovrapposizioni, ridondanze e sistemi differenti per gestire i metadati tra loro non sincronizzati.

Quella appena descritta era la situazione che si presentava fino ad un paio di anni fa, da allora il mercato è maturato ma non si è ancora giunti al punto in cui per integrare i propri dati è possibile affidarsi ad una singola piattaforma con risultati consolidati. A fine 2007 Gartner stimava un mercato per i software di Data Integration di circa 1,4 miliardi di dollari, con una crescita annua superiore al 17% [5], prevedendo per il 2011 un giro d'affari di circa 2,6 miliardi di dollari [26]. Il mercato degli strumenti di Data Integration comprende oggi prodotti diversi, che offrono soluzioni per una varietà di scenari:

- Acquisizione di dati per la realizzazione di Data Warehouse e analisi di Business Intelligence;
- Creazione di Master Data Stores integrati;
- Trasferimenti e conversioni/trasformazioni di dati;
- Sincronizzazione di dati tra applicazioni differenti;
- Creazione di database federati partendo da sorgenti multiple e distribuite;
- Fornitura di servizi sui dati tramite architetture Service-Oriented (SOA[1]);
- Fornitura di servizi secondo il modello Software-as-a-Service (SaaS[2]);
- Aggregazione di dati strutturati con dati non strutturati.

Gartner afferma che Business Intelligence e Data Warehousing restano, ad oggi, le funzionalità principali richieste dalla domanda di prodotti di Data Integration [5].

2.1.1 FUNZIONALITÀ RICHIESTE DAL MERCATO

Al fine di poter dare una valutazione dei prodotti di Data Integration presenti sul mercato è importante capire quali funzionalità il mercato stesso richiede a tali tipologie di software. Gartner individua diverse classi di funzionalità [5]:

- **Connessione/adattamento ai dati**
  - Il prodotto deve essere in grado in interagire con strutture di dati differenti (database relazionali, formati di file testuali, tracciati XML, formati industriali, EDI (Electronic Data Interchange), ecc.);
  - Capacità di trattare formati di dati emergenti e non sempre strutturati: e-mail, dati Web, dati provenienti da suite di office automation, ecc.;
  - Supportare diverse modalità di interazione con i dati: acquisizione di grandi volumi di dati in tempi ragionevoli, integrazione automatica di dati modificati, acquisizione di dati in base a eventi (in base al tempo o al valore dei dati);
- **Invio/trasferimento dei dati**
  - Abilità di rendere i dati disponibili per altre applicazioni secondo modalità differenti: spostamento fisico del dato, creazione di viste sui dati, replicazione dei dati su altri sistemi DBMS;
  - Trasferimento di dati in tempo reale;

[1] Con il termine *Service-Oriented Architecture (SOA)* si indica un'architettura software ottimizzata per supportare l'uso di servizi Web e in grado di garantire l'interoperabilità tra sistemi diversi. In questo modo le singole applicazioni sono viste come componenti di un unico processo di business e soddisfano le richieste degli utenti in modo integrato e trasparente.

[2] *Software as a service (SaaS)*, è un modello di distribuzione del software dove un produttore sviluppa e gestisce un'applicazione web che mette a disposizione dei propri clienti. Con un modello di questo tipo i clienti non pagano per il possesso del software ma per l'utilizzo dello stesso.

  - Capacità di trasferire i dati in base a specifici intervalli temporali o determinate tipologie di eventi;

- **Trasformazione dei dati**
  - Capacità di compiere operazioni di trasformazione dei dati di complessità differente: trasformazioni di base (conversione del tipo di dato, manipolazione delle stringhe e calcoli di base), trasformazioni intermedie (aggregazioni di dati), trasformazioni complesse (parsing di dati non strutturati o di elementi multimediali);
  - Garantire la possibilità di creare trasformazioni personalizzate;

- **Gestione dei metadati**

  È la parte dei sistemi di Data Integration che sta assumendo sempre più importanza, include la gestione dei metadati e la gestione del modello dei dati, ai fini di:
  - Scoprire e acquisire automaticamente metadati dalle basi di dati originarie, da applicazioni e altre fonti;
  - Creazione e manutenzione del modello dei dati;
  - Mapping da modello fisico a logico;
  - Definire le relazioni tra modelli di dati diversi;
  - Effettuare analisi sui dati e creare report in formato grafico o tabellare;
  - Capacità di gestire un database dei metadati condivisibile con altri applicativi;
  - Sincronizzazione automatica di metadati tra istanze diverse dell'applicativo;
  - Possibilità di estendere manualmente i metadati attraverso un'interfaccia utente;

- **Ambiente di design e sviluppo**

  Disporre di componenti con il compito di facilitare la definizione delle specifiche e la costruzione del processo di integrazione dei dati:
  - Rappresentazione grafica di oggetti, modelli e flussi di dati;
  - Gestione dello stato di avanzamento del processo;
  - Capacità di sviluppo in team: collaborazione e controllo delle versioni;
  - Supporto al riuso tra sviluppatori e progetti diversi, facilitare l'identificazione di ridondanze;
  - Fasi di test e debugging;

- **Strumenti di Data Governance**
  - Meccanismi in grado di assicurare la qualità dei dati nel tempo. Includono strumenti come: Data Profiling, Data Mining, Data Quality Assessment;

- **Supporto di piattaforme differenti**
  - Il prodotto deve supportare hardware e sistemi operativi differenti;

- **Amministrazione del sistema**
  Funzionalità che permettano di seguire, supportare, monitorare e correggere il processo di integrazione:
  - Segnalazione di errori;
  - Monitoraggio e controllo del processo in fase di esecuzione;
  - Raccolta di statistiche nella fase di esecuzione per determinare l'efficienza del processo;
  - Architettura in grado di assicurare performance e scalabilità;
- **Architettura e integrazione**
  Livello di consistenza e interoperabilità tra le varie componenti del prodotto:
  - Minor numero possibile di componenti per la gestione dei diversi modelli di dati (preferibilmente un componente unico);
  - Gestione uniforme dei metadati, con possibilità di condivisione;
  - Ambiente di design comune;
  - Abilità di spostarsi da un componente all'altro senza problemi di rilievo;
  - Interoperabilità con altri prodotti di Data Integration, tramite interfacce certificate e interfacce di programma consolidate;
  - Grado di efficienza nella gestione di tutti i modelli di dati supportati;
- **Servizi**
  - Dato che i servizi basati sui dati sono in continua crescita, i prodotti di Data Integration devono avere delle caratteristiche service-oriented e garantire la presenza di un'architettura software adatta a supportare l'uso di servizi SOA.

2.1.2 CRITERI PER LA VALUTAZIONE DEI PRODOTTI

Definite le funzionalità che il mercato richiede ad un applicativo di Data Integration, è necessario stabilire quali sono i criteri per valutare l'effettiva validità di un prodotto. Per una prima scrematura dei prodotti Gartner considera i seguenti requisiti funzionali [5]:

- Capacità di connessione a diversi formati di database: accesso a tutti i prodotti relazionali, a forme di dati non relazionali, a dati organizzati in tabelle (testo con separatori, ad esempio file .txt e .csv) e in formato XML;
- Modalità di connessione a tali tipologie di database;
- Modalità di invio/trasferimento dei dati (ETL, viste federate, ecc.);
- Supporto di operazioni di trasformazione dei dati: come minimo devono supportare trasformazioni di base come cambio del tipo di dato, manipolazione delle stringhe e calcoli;

- Livello di supporto ai metadati: scoperta automatica di metadati, analisi automatiche, sistemi di gestione dei metadati (possibilmente aperti per lo scambio di metadati con altri sistemi);
- Design e supporto nell'utilizzo del prodotto: ambiente di progettazione grafico e usabile e possibilità di lavoro in gruppo (collaborazione e controllo delle versioni);
- Supporto alla gestione dei dati: possibilità di operare a livello di metadati con strumenti di data quality o data profiling;
- Supporto di sistemi operativi differenti: Windows, Unix o Linux;
- Possibilità di mettere a disposizione funzionalità come servizi secondo i principi SOA.

Inoltre, i prodotti analizzati da Gartner devono soddisfare i seguenti requisiti non funzionali:

- Il prodotto deve generare almeno 20 milioni di dollari di ricavi annui ed avere un mercato di almeno 300 clienti;
- Il produttore deve offrire un servizio di supporto ai clienti in almeno due delle maggiori regioni geografiche (Nord America, America Latina, Europa, Asia);
- Il prodotto deve essere utilizzato dai clienti costantemente e a livello di impresa, non occasionalmente, non da una singola unità operativa o a livello di sperimentazione.

Applicando tali criteri Gartner seleziona 36 aziende delle quali valuta quanto il prodotto proposto soddisfa i requisiti individuati. Per determinare la posizione del prodotto sul mercato Gartner tiene conto di:

- *Abilità di svolgere le operazioni individuate dai requisiti funzionali*: in particolare viene data molta rilevanza alle capacità di gestione dei metadati e al supporto di strumenti a supporto della qualità del dato;
- *Completezza della visione***:** pone l'attenzione sulla capacità del produttore di capire il mercato, di comprendere i bisogni e le necessità che il prodotto deve soddisfare. Vengono fissati degli indicatori di valutazione per giudicare il grado di comprensione del mercato da parte del produttore. Sapere cogliere le esigenze del mercato è un fattore fondamentale, Gartner valuta positivamente quei prodotti che si dimostrano indipendenti da specifici modelli di dati e architetture di sistema e che sono in grado di interoperare con gli altri prodotti presenti sul mercato. Altri fattori considerati sono le strategie di vendita (in diverse aree geografiche), la velocità di espansione in nuovi mercati e lo sviluppo di funzioni innovative. Inoltre, per capire quanto il prodotto è valido e completo Gartner ha svolto un sondaggio su oltre 300 clienti per capire quanto sono soddisfatti del prodotto che stanno utilizzando.

Questa analisi permette a Gartner di individuare i migliori prodotti presenti sul mercato suddividendoli in quattro categorie:

- **Leaders**: sono coloro che si trovano in una posizione di vantaggio sul mercato offrendo un prodotto che supporta ottimamente diverse funzionalità. Sono solitamente prodotti che si dimostrano forti e consolidati nelle attività classiche di Data Integration come l'ETL, ma non tralasciano il supporto a modelli nuovi (Data Federation) e supportano la realizzazione di servizi in contesto SOA. Tali produttori sono collocati sul mercato da molto tempo e ne definiscono l'andamento lanciando nuove funzionalità o identificando nuove aree di business per i propri prodotti;
- **Challengers**: sono coloro che seguono da vicino i leaders. Posizionati nel cuore del mercato, offrono prodotti molto validi ma che non presentano un ventaglio di funzionalità così ampio come quello dei prodotti leader, oppure sono prodotti limitati ad ambienti specifici. Hanno una ottima visione del mercato ma spesso ostacolata da una strategia di gestione dei propri prodotti non ottimamente coordinata. Si può dire che mancano della maturità e dell'esperienza necessarie per divenire leader del mercato;
- **Visionaries**: hanno una buona visione dei settori principali del mercato e si dimostrano correttamente allineati alla domanda. Peccano però di credibilità verso i clienti in quanto il loro prodotto supporta spesso domini di applicazione limitati o non è in grado di fornire una gamma di funzionalità sufficientemente ampia e completa. Chi appartiene a questa categoria solitamente è entrato da poco nel mercato, possiede una base installata ristretta e non è ancora riconosciuto come prodotto affidabile e consolidato dai grandi rivenditori o consulenti;
- **Niche Players**: sono coloro che hanno delle lacune sia nella completezza della visione del mercato sia nell'abilità di esecuzione delle funzionalità richieste. Tali prodotti mancano spesso di funzioni chiave, indispensabili per avere successo sul mercato. Inoltre, non sono riconosciuti dai clienti come prodotti collaudati. Per avere maggior successo sul mercato devono costruirsi una nuova reputazione, oltre a realizzare un prodotto più completo e funzionale. Alcuni di questi produttori offrono comunque soluzioni valide per domini applicativi di nicchia.

Il risultato finale dell'analisi di Gartner è sintetizzato in un grafico (quadrante), redatto solitamente con frequenza annuale. Il quadrante riporta le aziende, che rispettano i requisiti prefissati, suddivise nelle quattro categorie individuate (in Figura 14 sono riportati i quadranti rispettivamente degli anni 2007 e 2008).

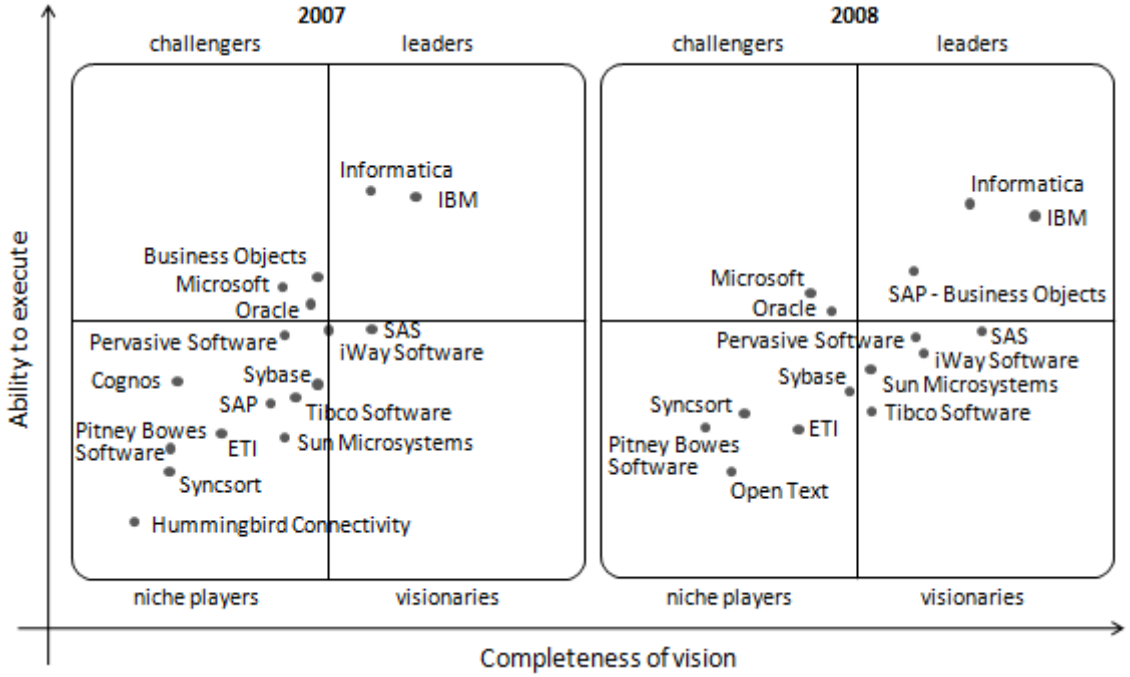

Figura 14 - Magic Quadrant for Data Integration Tools (Fonti: [5] e [28]. Rielaborazione dell'autore).

Dalla rappresentazione grafica di Gartner si può notare come nel 2007 la maggior parte dei produttori si colloca nel segmento "Niche Players", mentre nel 2008 si assiste ad uno spostamento verso il settore "Visionaries", sintomo di maggiore maturità dell'offerta. Si può inoltre notare come il mercato, nel 2007, sia dominato da due produttori, Informatica e IBM, con un distacco abbastanza netto dagli inseguitori che si collocano in prossimità del centro del quadrante. Nel 2008 la situazione muta soprattutto in conseguenza dell'acquisizione da parte di SAP di Business Objects [27], acquisizione che permette a SAP di fare un notevole balzo avanti insidiando di fatto la leadership di Informatica e IBM. Altri due colossi come Microsoft e Oracle non mutano praticamente la loro posizione nel periodo 2007-2008, mentre si può notare una maggiore completezza di visione di SAS e iWay Software che nel 2008 si spostano verso la parte destra del quadrante, non riuscendo però a collocarsi nel quadrante superiore a causa di un ancora troppo scarso supporto ad alcune funzionalità supportate invece dai prodotti leader. Inoltre, è da sottolineare il passaggio alla categoria "Visionaries" di Sun Microsystems e Tibco Software, sintomo di un maggiore investimento delle due compagnie in questo tipo di mercato.

Dal confronto della situazione in questi due anni si comprende come diverse compagnie stanno investendo in questo tipo di mercato, con l'intento di migliorare il loro prodotto in modo da contrastare la leadership di Informatica e IBM. In ottica futura un concorrente insidioso per il prodotti leader potrebbe essere Oracle vista l'intenzione della software

house californiana di acquisire Sun Microsystems [29] [30] [31], superando proprio la concorrenza di uno dei leader del settore, IBM[3].

2.2 ANALISI DEI PRODOTTI SUL MERCATO

Il manifestarsi delle prime necessità di dati integrati ha portato le aziende ad affrontare il problema internamente in quanto il mercato non sapeva offrire soluzioni sufficientemente flessibili ed affidabili. Per questo il primo approccio per rispondere alle necessità di avere dati integrati fu quello di sviluppare internamente all'azienda software ad hoc soprattutto per eseguire le fasi di estrazione, trasformazione e caricamento dei dati in un ambiente unico e integrato. Nonostante i recenti progressi dei prodotti di Data Integration ancora oggi la maggior parte delle aziende utilizza soluzioni ETL personalizzate per rispondere alle necessità di integrazione.

Tuttavia, le più recenti evoluzioni del mercato hanno portato ad un aumento della domanda di prodotti completi di Data Integration, portando al 60% la percentuale delle imprese che utilizza uno dei pacchetti di prodotti di integrazione offerti sul mercato, con lo scopo di effettuare attività di Business Intelligence [7]. La recente crisi economica ha portato inoltre ad una diminuzione dei budget assegnati allo sviluppo dell'Information Technology nelle aziende, determinando un incremento dell'adozione di soluzioni di integrazione open source.

Si può quindi affermare che il mercato della Data Integration è oggi caratterizzato dalla convivenza di tre tipologie di prodotti [7]:

1. **Software personalizzati**: con l'emergere delle prime necessità di integrazione dei dati molte imprese svilupparono internamente prodotti ad hoc in grado di rispondere alle esigenze specifiche del proprio ambito di business. Con la maturazione del mercato dei prodotti di Data Integration questo tipo di approccio è divenuto sempre meno conveniente. Inoltre, l'emergere di architetture SOA e applicazioni SaaS sta decretando la fine dei prodotti sviluppati in casa. Oggi le suite di Data Integration presenti sul mercato offrono sicuramente funzionalità e affidabilità migliori;
2. **Software proprietari**: lo sviluppo di applicativi di Data Integration ha contribuito ad aumentare la produttività delle attività collegate all'integrazione dei dati. I prodotti di integrazione dei dati sono maturati costantemente negli anni garantendo

[3] Nel momento in cui scrivo (1 novembre 2009) il processo di acquisizione di Sun da parte di Oracle non è ancora terminato. In seguito al parere favorevole del Dipartimento di Giustizia americano [30], l'Unione Europea ha espresso dei dubbi sul completamento dell'acquisizione [31], dubbi relativi soprattutto al destino del DBMS open source MySQL. L'Unione Europea è tenuta a pronunciarsi definitivamente entro il 19 gennaio 2010, soltanto dopo tale data sapremo se l'acquisizione potrà divenire realtà.

un ventaglio di funzionalità sempre più ricco e variegato, rendendo tali applicativi idonei a supportare la grande maggioranza degli scenari di business che richiedono l'utilizzo di dati integrati. Il numero di applicativi sul mercato è oggi elevato, si va dalle suite di prodotti in grado di coprire la quasi totalità delle necessità aziendali a prodotti specializzati in particolari contesti di business o specifiche problematiche;

3. **Software open source**: il limite dei maggiori prodotti proprietari presenti sul mercato sono i costi necessari per la loro implementazione. Per venire in contro alle necessità delle aziende più piccole e con risorse limitate si sono da poco affacciate sul mercato i primi prodotti open source, prodotti in grado di supportare una discreta quantità di funzioni ma con un costo decisamente minore rispetto ai prodotti proprietari (costi di licenza nulli, costi di infrastruttura ridotti, servizi pagati in base all'utilizzo).

Inoltre, come vedremo nel paragrafo 3.4, si stanno affacciando sul mercato i primi prodotti che implementano tecnologie semantiche. La tabella che segue (Tabella 1) presenta un confronto schematico dei tre approcci evidenziandone i punti di forza e le criticità.

	Software Open source	**Software Commerciali**	**Software personalizzati**
Funzionalità			
Costi di licenza			
Costi totali			
Disponibilità di esperienza, pratica, conoscenza			
Collaborazione con i partner			
Supporto ai sistemi critici			
Servizi e supporto			
Flessibilità e indipendenza			

= forte/consolidato = moderato/cautela = debole/rischio

Tabella 1 - Confronto tra le tre tipologie di prodotti per l'integrazione dei dati.
(Fonte: [25]. Rielaborazione dell'autore).

2.2.1 SOLUZIONI PROPRIETARIE

Vediamo quali sono i punti di forza e le criticità dei principali prodotti consolidati disponibili sul mercato. L'analisi parte dai prodotti leader (IBM, SAP-Business Objects e Informatica) per poi valutare caratteristiche e funzionalità dei principali prodotti concorrenti che aspirano a conquistare una posizione di leadership nel mercato (Microsoft, Oracle e Pervasive Software).

2.2.1.1 IBM

IBM dimostra un'ottima visione del mercato, la sua offerta è costituita da un pacchetto, WebSphere Data Integration Suite, contenente diversi prodotti: Information Server, DataStage, Change Data Capture, Federation Server, QualityStage, Data Architect, Replication Server e Information Analyzer. Secondo l'analisi di Gartner la suite continua a migliorare garantendo un ambiente di lavoro sempre più consistente e funzionale. L'applicativo si dimostra perfettamente adattabile a campi applicativi diversi rispondendo ad un'ampia gamma di esigenze. I suoi clienti si dimostrano molto soddisfatti e in molti utilizzano il prodotto da diverso tempo. Un punto di forza notevole è lo stretto rapporto che IBM mantiene con i propri clienti. I due soggetti collaborano infatti attivamente attraverso varie iniziative, ad esempio: integrazione di dati di tipo B2B, Business Intelligence e Data Warehousing, gestione dei dati tramite architetture SOA. L'integrazione da parte di IBM di Cognos [32] costituisce un valore aggiunto di grande importanza potendo abbinare all'integrazione dei dati elementi di analisi tipici della Business Intelligence[4]. Infine, il tool di IBM può essere utilizzato anche per la Data Federation. Tuttavia, dall'analisi risulta che i clienti che utilizzano le funzionalità di Data Federation sono ancora una piccola minoranza e non è quindi possibile esprimere una valutazione completa sull'effettiva bontà di questa parte del prodotto.

I punti critici della suite IBM sono essenzialmente due: innanzitutto molti clienti, pur ritenendosi soddisfatti delle funzionalità del prodotto, lamentano una curva di apprendimento troppo difficile, che spesso si traduce in tempi di implementazione molto lunghi (superiori ai 6 mesi). IBM ha rilevato che questo problema non è manifestato dai clienti che sono soliti usare solo le funzioni base del prodotto (trasformazioni preimpostate), ma da coloro che ne fanno un utilizzo più completo e approfondito (trasformazioni ad hoc). In secondo luogo IBM si è dimostrata molto attiva nell'acquisire una grande varietà di tecnologie sul mercato ma in molti lamentano una scarsa compatibilità tra i diversi prodotti offerti. Se in futuro vuole mantenere la leadership IMB deve migliorare la compatibilità tra le diverse funzionalità offerte dal suo pacchetto di prodotti, garantendo maggiore interoperabilità.

2.2.1.2 INFORMATICA

Informatica è probabilmente il produttore più affermato sul mercato, dove la sua presenza è in continua espansione, con una base installata di oltre 3.000 clienti e profitti in crescita costante. Informatica ha dimostrato di reggere la concorrenza nel settore attraverso l'acquisizione di nuovi importanti clienti e con l'espansione dei servizi offerti ai clienti di lunga data. I ricavi annuali si aggirano attorno ai 400 milioni di dollari e l'azienda può

[4] Cognos era un'azienda canadese leader nel settore della Business Intelligence e Performance Management.

contare sul lavoro dei migliori specialisti nel campo della Data Integration. Informatica ha costruito la sua ottima reputazione con l'offerta di una tecnologia solida e completa, aggiornata regolarmente e con la garanzia di un ottimo supporto post-vendita. In molti ambiti il suo prodotto è di fatto riconosciuto come standard del mercato. Rispetto al prodotto IBM quello di Informatica presenta una curva di apprendimento più graduale, che richiede conoscenze approfondite soltanto negli utilizzi più avanzati del prodotto. Con il lancio di PowerCenter 8.6, nel corso del 2008, Informatica ha implementato l'integrazione di dati in tempo reale e ha aggiunto all'applicativo una componente specifica per la gestione della qualità dei dati, rendendo il prodotto ancora più completo e rispondendo a quelle che erano le più recenti esigenze del mercato. Tuttavia, nonostante l'elaborazione in tempo reale molti clienti continuano a preferire quella di tipo batch, Informatica deve quindi stare attenta a non trascurare le volontà di coloro che operano ancora con la vecchia metodologia di integrazione. Informatica non deve dimenticarsi dell'esistenza di diverse tipologie di clienti evitando di concentrare risorse eccessive sullo sviluppo di funzioni innovative a discapito delle funzionalità consolidate. D'altra parte Informatica sta cercando di spingere i clienti che utilizzano ancora le metodologie tradizionali all'adozione dei nuovi strumenti.

La preoccupazione principale per Informatica è data dalla sua particolare posizione sul mercato: l'azienda, produttore indipendente che fa della Data Integration il suo core business, si trova ad operare in un mercato molto competitivo nel quale operano società molto grandi e potenti (SAP, Oracle, Microsoft, IBM) la cui offerta, al di la dei prodotti di Data Integration, può essere molto completa (DBMS, ERP, CRM, ecc.). Questo da una parte crea ad Informatica delle opportunità, potendosi focalizzare sullo sviluppo di un unico prodotto garantendo una soluzione migliore rispetto ai concorrenti, ma è anche fonte di grandi sfide, infatti, i concorrenti possono offrire una piattaforma di software completa a prezzi vantaggiosi ponendo una barriera all'ingresso per il prodotto di Informatica che offre soltanto funzionalità di Data Integration. Infine, l'analisi ha rivelato che alcuni clienti lamentano una competenza non eccelsa nei campi non strettamente collegati all'ETL, come la Data Federation e la Data Replication.

2.2.1.3 SAP BUSINESS OBJECTS

Con l'acquisizione di Business Objects, SAP ha saputo sfruttare la sua grande base installata di applicativi gestionali per promuovere l'adozione del nuovo prodotto di Data Integration e Business Intelligence presso i propri clienti. Questo ha consentito a SAP di conquistare una buona fetta di mercato in tempi relativamente brevi. Il nuovo prodotto offerto da SAP garantisce funzionalità avanzate di modellazione dei dati e di gestione dei metadati in scenari di integrazione differenti. Il prodotto include il supporto alla federazione dei dati (Business Objects Data Federator) e una nuova piattaforma (Data

Services) che combina le funzionalità di integrazione con quelle per la gestione della qualità del dato. È soprattutto in queste due componenti che si nota l'esperienza di Business Objects, esperienza che permette al nuovo prodotto di SAP di competere con i leader del mercato. Il prodotto offre inoltre funzionalità avanzate di Data Quality, Data Mining e Text Mining, derivanti dalla precedente acquisizione, da parte di Business Objects, di Firstlogic[5] nel 2006 [33]. La fusione di questi prodotti in un pacchetto unico ha permesso quindi la realizzazione di un prodotto completo con funzionalità di gestione e manipolazione dei dati consolidate e avanzate.

Tuttavia, l'acquisizione di Business Objects ha portato anche qualche problema a SAP. I due prodotti sono ancora percepiti dal mercato come due soluzioni distinte e SAP ha dovuto sviluppare una strategia di allineamento per coloro che avevano iniziato a lavorare con uno dei due prodotti precedenti. Inoltre, il prodotto non è ancora in grado di processare i dati in tempo reale e questo fa si che le aziende con questa necessità si rivolgano ad altri prodotti. Ad oggi, si può affermare che il prodotto offerto da SAP-Business Objects è senz'altro buono ma non è ancora del tutto maturo e non del tutto completo, per conquistare altre quote di mercato dovrà migliorare alcuni aspetti che per molte aziende risultano fondamentali.

2.2.1.4 ORACLE

Oracle offre funzionalità di Data Integration tramite due prodotti distinti: Oracle Warehouse Builder (OWB), prodotto incluso nella licenza del DBMS, e Oracle Data Integrator (ODI), prodotto indipendente. Nel corso del 2008 Oracle ha lanciato sul mercato Data Integration Suite, basata essenzialmente su ODI con l'aggiunta di un'altra serie di strumenti per offrire una gamma più ampia di funzionalità. Inoltre, l'acquisizione di BEA Systems[6] ha garantito ad Oracle la possibilità di migliorare le funzionalità di data federation [34]. L'adozione dei due prodotti Oracle è in continua crescita, soprattutto nelle tradizionali implementazioni ETL a supporto di Business Intelligence e Data Warehousing. La perfetta interoperabilità tra i prodotti di integrazione, il Database Management System, e i componenti middleware costituisce uno dei principali punti di forza dei prodotti Oracle. In tempi recenti Oracle ha formalizzato i primi passi per l'integrazione dei due prodotti, OWB e ODI, in una soluzione unica, l'obiettivo, comunque di medio-lungo periodo, è di proporsi sul mercato con un singolo pacchetto di strumenti.

I prodotti Oracle si sono dimostrati in grado di rispondere in tempi molto brevi alle richieste del mercato con particolare enfasi alle elaborazioni ETL ma senza ignorare

[5] Firstlogic era una società americana leader nel campo della Data Quality e della Data Cleansing.
[6] BEA Systems era una società che si occupava principalmente dello sviluppo di software di infrastruttura conosciuti col nome di "middleware", software che permettono ad un'applicazione di connettersi a un database.

funzionalità innovative (nel capitolo 3 vedremo come Oracle sia stata la prima software house di un certo peso a supportare l'utilizzo di linguaggi semantici nel proprio DBMS). La grande base installata di DBMS e di applicativi di Business Intelligence permettono a Oracle di mantenersi in una posizione di forza sul mercato con la possibilità di aumentare il tasso di adozione dei propri prodotti di integrazione. Nonostante l'intenzione dichiarata di unire i propri prodotti in un pacchetto unico l'attuale portafoglio di applicativi risulta troppo vario e frammentato, generando una confusione che spesso confonde i clienti e aumenta la complessità nell'implementare una soluzione efficace adatta alle proprie esigenze. Per poter insidiare i leader del mercato Oracle dovrà lavorare molto su questo aspetto, anche in prospettiva dell'acquisizione di Sun Microsystems, Oracle deve mettere ordine nella propria offerta e garantire la presenza un prodotto unico e consistente.

2.2.1.5 MICROSOFT

L'offerta di Microsoft nel settore della Data Integration è rappresentata da SQL Server Integration Services (SSIS). Con questo prodotto Microsoft garantisce il supporto a trasferimenti di dati tramite tecniche ETL e all'integrazione in tempo reale grazie all'interazione con BizTalk Server[7]. Microsoft offre inoltre funzionalità di replicazione e sincronizzazione di dati e un supporto di base a funzionalità di data federation tramite il proprio DBMS (SQL Server). L'ultima versione di SSIS espande la connettività ad una moltitudine di fonti dati differenti grazie anche alla possibilità di connessione a prodotti come SAP Business Information Warehouse (SAP BW), SAP ERP, Oracle, Teradata e IBM DB2. Il prodotto SSIS supporta ottimamente funzionalità di Data Warehousing e analisi dei dati tramite la costituzione di Data Mart o attraverso l'interazione con SQL Server Analysis Services (SSAS). I clienti del prodotto Microsoft si dicono sostanzialmente soddisfatti citando in particolare bassi costi di possesso, semplicità e velocità di implementazione, facilità di utilizzo, buone performance e ottima integrazione con gli altri prodotti della famiglia SQL Server. La recente acquisizione di Zoomix[8] ha permesso a Microsoft di rendere il suo prodotto completo anche per quanto riguarda le funzionalità di data quality e data cleansing [35]. La presenza e la rilevanza a livello globale di Microsoft facilita inoltre la promozione del proprio prodotto, anche attraverso l'offerta di pacchetti software completi a prezzi molto vantaggiosi.

Nonostante il prodotto Microsoft offra un'ampia gamma di funzionalità sono pochi i clienti che lo utilizzano in maniera approfondita, soprattutto la possibilità di interagire con altri software Microsoft (BizTalk Server e SQL Server) è scarsamente sfruttata. Probabilmente Microsoft deve rivedere la sua strategia di marketing puntando maggiormente alla

[7] BizTalk Server è un prodotto Microsoft che fornisce funzionalità di middleware e di Business Process Management (BPM). Esso permette alle aziende di automatizzare e integrare i processi aziendali.
[8] Zoomix era una start-up israeliana focalizzata sullo sviluppo di software a supporto della data quality.

valorizzazione di quei componenti che permettono al suo prodotto di distinguersi dalle altre soluzioni presenti sul mercato. Soltanto sfruttando i punti di forza del proprio prodotto Microsoft può provare ad impensierire i tre leader del mercato. Altri punti deboli di SSIS sono un ancora non perfetto supporto verso alcune tipologie di dati e la scarsa consistenza nella gestione dei metadati. Infine, un altro grande limite del prodotto è l'incapacità di operare al di fuori del sistema operativo Microsoft (Windows).

2.2.1.6 PERVASIVE SOFTWARE

Pervasive offre un prodotto solido (Pervasive Data Integrator), con un buon ventaglio di funzionalità e ad un prezzo piuttosto contenuto rispetto ai principali concorrenti. L'azienda si dimostra umile e modesta nel rapporto con i clienti e si propone sul mercato con abilità ed esperienza, doti che gli hanno permesso di superare la quota di 3500 clienti. Il prodotto di Pervasive risulta molto utilizzato nei classici ambiti ETL ma anche a supporto dell'integrazione di applicazioni (EAI) e in sistemi di tipo real-time. I punti di forza del prodotto sono l'ottimo supporto ai metadati e l'implementazione di funzionalità SOA. Il prodotto supporta la creazione di un deposito dei metadati in grado di supportare l'estrazione automatica di informazioni da formati di dati strutturati, semi-strutturati e non strutturati. Dal punto di vista SOA il prodotto supporta funzionalità web service che permettono di esporre facilmente sul Web i dati integrati. I clienti di Pervasive sono solitamente di lunga data ma utilizzano quasi tutti l'ultima versione del prodotto grazie alla presenza di un semplice ed efficace sistema di aggiornamento alle nuove versioni. Inoltre, Pervasive ha recentemente sviluppato una componente SaaS in grado di offrire funzionalità di integrazione dei dati come servizio, mostrando grande attenzione ai recenti andamenti del mercato, soprattutto nell'ottica di contrastare l'espansione dei servizi di integrazione open source.

Tuttavia, il basso costo di acquisizione del prodotto ha fatto si che molti clienti lo utilizzino in modi diversi in una varietà di situazioni differenti anziché averne fatto uno standard unico per l'intera azienda. Implementazioni di questo tipo annullano di fatto il vantaggio di possedere un deposito dei dati centralizzato e consolidato in quanto si originino depositi multipli che creano confusione e disallineamento all'interno della stessa impresa. Inoltre, il prodotto non presenta funzionalità di data federation e i servizi di data cleansing sono implementati tramite prodotti partner, creando talvolta problemi di consistenza. Tutto sommato, Pervasive offre un prodotto economico con funzionalità sufficientemente avanzate per la maggior parte dei contesti di applicazione.

2.2.2 SOLUZIONI OPEN SOURCE

L'acquisizione e l'implementazione dei prodotti proprietari descritti nel paragrafo precedente comporta ingenti investimenti in termini di tempo e risorse, investimenti che

non tutte le aziende possono o sono disposte a sostenere. Per questo motivo, in tempi recenti, prodotti di Data Integration open source hanno cominciato ad affermarsi sul mercato. Attraverso la realizzazione di prodotti open source è possibile distribuire sviluppi e costi tra la rete di sviluppatori ed utilizzatori del prodotto, diminuendo il costo totale della soluzione e facilitando lo sviluppo del prodotto. Si possono isolare sostanzialmente due modelli di open source [7]:

1. **Approccio basato su progetti o comunità**: tipicamente implica la presenza di una fondazione non-profit che coordina lo sviluppo del prodotto e ne mantiene i diritti e di persone che contribuiscono allo sviluppo e alla manutenzione;
2. **Approccio commerciale**: comporta la presenza di una compagnia che vende supporto e servizi sul prodotto open source sviluppato. Un'azienda di questo tipo opera sostanzialmente come una software house tradizionale ad eccezione del fatto che il codice sorgente del proprio prodotto non è segreto. Questo permette una maggiore interazione tra il venditore, gli sviluppatori e la comunità di clienti che si allarga man mano che il prodotto viene adottato dal mercato. Si può dire che un modello di questo tipo si focalizza maggiormente sull'utente rispetto a un software proprietario.

Nel campo della Data Integration si sta affermando soprattutto il secondo tipo di approccio e i venditori guadagnano dall'offerta di consulenza, assistenza e funzionalità aggiuntive proprietarie sviluppate sulla piattaforma open source di base. Un modello di questo tipo porta a minori costi di possesso, permette una maggiore flessibilità del sistema e coinvolge i clienti nell'ottimizzazione delle funzionalità del prodotto e nello sviluppo di nuove funzioni. Prodotti di questo tipo offrono le loro funzionalità come servizio con costi proporzionati all'effettivo utilizzo del sistema, rendendo di fatto accessibili le moderne tecnologie di integrazione anche alle aziende più piccole e con meno risorse.

L'esclusione di questa categoria di prodotti dal quadrante (Figura 14) ha suscitato qualche critica nei confronti di Gartner [36]. In particolare, ha sollevato delle perplessità l'esclusione di Talend in quanto il suo prodotto, Talend Integration Suite, rientra nei requisiti definiti da Gartner per l'immissione nel quadrante. Gartner si sta dimostrando eccessivamente conservativa giudicando i prodotti di integrazione open source ancora immaturi.

2.2.2.1 TALEND INTEGRATION SUITE

L'approccio open source di Talend prevede la disponibilità di due prodotti:

1. **Talend Open Studio**: suite gratuita scaricabile gratuitamente con licenza open source (GPL). Talend Open Studio si presenta come prodotto di Data Integration completo e contraddistinto da un ampia gamma di funzionalità, sufficienti per la maggior parte delle necessità;

2. **Talend Integration Suite**: è una versione potenziata del prodotto gratuito che aggiunge funzionalità avanzate come lo sviluppo collaborativo e monitoraggio avanzato del progetto.

Per chi non possiede l'hardware necessario per supportare il sistema c'è una terza opzione costituita da Talend On Demand, ovvero un offerta di tipo Software ad a Service (SaaS). I prodotti di Talend offrono ad oggi le seguenti funzionalità:

- Ambiente di sviluppo user-friendly (basato sulla piattaforma Eclipse)
- Elevato numero di connessioni preimpostate
- Deposito comune dei metadati
- Supporto allo sviluppo collaborativo
- Servizi di trasformazione di dati
- Funzionalità di monitoraggio dell'andamento dell'integrazione
- Data Profiling e Data Quality

Vediamo quindi quali sono i punti di forza dell'approccio di open source di Talend [37]:

- *Nessuna barriera all'adozione*: la disponibilità gratuita del prodotto di base rende praticamente immediata l'installazione del software. Talend supporta il cliente attraverso tutorial sull'utilizzo di base, inoltre è possibile fare affidamento ad una vasta comunità di utilizzatori;
- *Curva di apprendimento veloce*: il prodotto si presenta graficamente amichevole e facile da utilizzare. L'interfaccia grafica è intuitiva e l'utilizzo delle funzionalità di base non richiede particolari addestramenti;
- *Modello di prezzi stabile e prevedibile*: i prodotti proprietari prevedono spesso costi elevati man mano che si espandono le funzionalità e le capacità del prodotto, con costi di licenza che aumentano all'aumentare delle macchine installate. Questo rende spesso difficile una corretta previsione dei costi nelle fasi iniziali del progetto, soltanto a lavoro ultimato è possibile rendersi conto del costo effettivo della soluzione adottata. Talend prevede un modello di costo basato sul numero di sviluppatori e sull'utilizzo del servizio, indipendente da licenze, hardware e quantità di dati da integrare;
- *L'importanza di una comunità a supporto*: la comunità online di esperti ed utilizzatori del prodotto è già oggi molto vasta ed è un fattore di grande importanza per facilitare l'implementazione e il mantenimento delle soluzioni offerte da Talend. Forum, wiki, guide e contributi gratuiti degli utenti rappresentano un valore aggiunto che solo un prodotto di questo tipo può offrire;
- *Ampio supporto a tipologie di dati differenti*: con oltre 400 connessioni preimpostate la soluzione di Talend garantisce la compatibilità con un grande numero di sistemi, database, pacchetti di software, applicazioni gestionali, servizi

web, ecc. Nessun'altra soluzione sul mercato vanta un numero di possibili connessioni così elevato;

- *Flessibilità, versatilità e riuso del prodotto*: Talend non si limita ad un supporto alle tecniche standard di ETL ma permette l'implementazione di diverse strategie di integrazione. La possibilità di riuso di progetti già perfezionati costituisce inoltre un altro punto di forza dell'approccio open source;
- *Funzionalità e performance*: il livello di funzionalità offerto è paragonabile a quello dei prodotti proprietari. Tuttavia, si registrano alcune lacune nel campo della modellazione dei dati, data quality e data mining. Un team di ricerca e sviluppo dedicato permette al prodotto di essere sempre aggiornato alle ultime esigenze del mercato e di proporre funzionalità innovative;
- *Costi e tempi ottimizzati*: le soluzioni offerte da Talend risultano da un 50% a un 80% più economiche rispetto ai prodotti tradizionali, essendo meno costose da acquisire e mantenere e permettono uno sviluppo più rapido del sistema di integrazione.

2.2.2.2 ALTRI PRODOTTI

Il successo di Talend ha spinto altri produttori a lanciarsi nel mercato della Data Integration open source. Ad oggi, altri prodotti di un certo rilievo sono:

- **SnapLogic DataFlow**: il prodotto si pone l'obiettivo di ridurre la complessità dei progetti di Data Integration tramite l'introduzione di uno standard per la consistenza, implementato attraverso l'utilizzo di una rete di integrazione via Web in grado di collegarsi ad una moltitudine di tipologie di archivi garantendo un accesso unico e consolidato ai dati;
- **CloverETL**: suite open source che supporta diverse funzionalità: Master Data Management, Data Integration, Data Warehouse, Data Migration, Data Cleansing, Data Synchronization e Data Consolidation;
- **XAware**: come Talend sfrutta la piattaforma Ecplise garantendo la presenza di un ambiente di sviluppo familiare a molti sviluppatori. Permette il collegamento a svariate fonti di dati e garantisce integrazione in tempo reale. Il flusso di dati può essere bidirezionale grazie al supporto di servizi SOA;
- **Apatar**: fornisce principalmente funzionalità di integrazione tra sorgenti di dati aziendali, con connessioni preimpostate a una discreta quantità di formati di dati e applicazioni. Inoltre, garantisce la popolazione di Data Warehouse e Data Mart. Anche questo prodotto utilizza la piattaforma Eclipse.

2.3 ANALISI DEI COSTI

Con l'affermarsi delle necessità di integrazione dei dati le aziende hanno cominciato ad analizzare nei dettagli costi e risorse richieste da progetti di questo tipo. Calcolare il Total Cost of Ownership (TCO) di una tecnologia di integrazione richiede la definizione di metriche complesse, in grado di tenere in considerazione una moltitudine di fonti di costo differenti. Una possibile metodologia, utilizzata nelle analisi riportate in [38] e [39], è quella di analizzare una base consistente di installazioni di prodotti di integrazione (nell'ordine di alcune centinaia di implementazioni) in imprese che fanno uso di almeno 20 prodotti software differenti. Nei due report analizzati lo studio di tempi di adozione, costi e risorse impiegate nei progetti di integrazione si basa sui risultati di un sondaggio sottoposto a centinaia di imprese (212 in [39] e oltre 400 in [38]) geograficamente distribuite e operanti in settori diversi.

Prima di valutare il TCO di alcuni dei prodotti analizzati nel paragrafo 2.2 è bene esaminare alcuni indicatori di tendenza del mercato. In particolare, l'analisi proposta in [38] evidenzia quali sono i criteri che influenzano le aziende nella scelta di un prodotto di Data Integration, mettendo in luce come il TCO sia il criterio che maggiormente influenza la decisione dell'azienda (Figura 15).

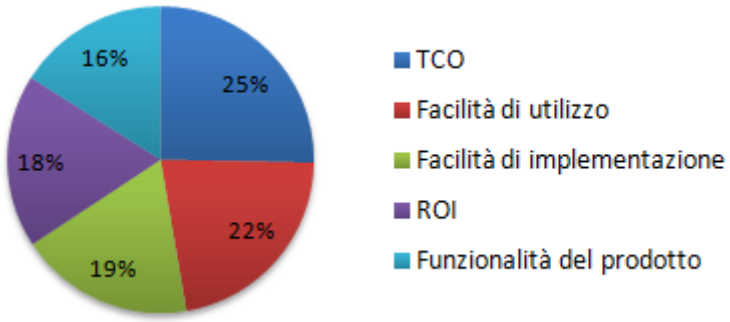

Figura 15 - Criteri utilizzati nella scelta del prodotto di integrazione (Fonte: [38]. Rielaborazione dell'autore).

In secondo luogo è utile analizzare in quali contesti applicativi vengono utilizzati i prodotti di integrazione. Secondo l'analisi di Bloor[9] [39] l'utilizzo più frequente dei prodotti di Data Integration è in scenari di conversione e migrazione di dati, seguito dalle classiche implementazioni ETL con Data Warehouse (Figura 16). Altre necessità sono lo scambio di dati tra imprese e la sincronizzazione di applicazioni. Da sottolineare il fatto che l'utilizzo di prodotti di Data Integration per l'implementazione di soluzioni SOA è ancora allo stadio iniziale, dovuto probabilmente al fatto che solo in tempi recenti molte aziende hanno capito l'importanza di sviluppare servizi sui dati di tipo SOA.

[9] Bloor Research è una società europea che offre servizi di ricerca, analisi e consulenza in campo IT.

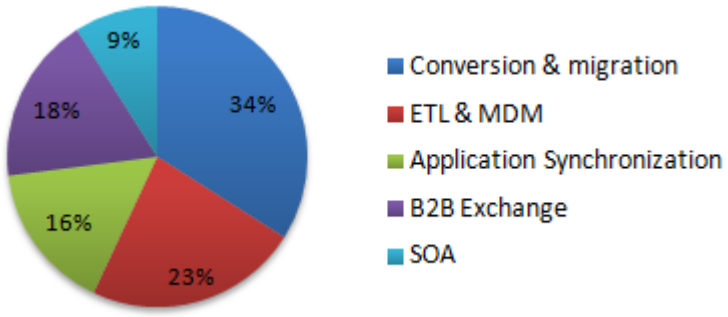

Figura 16 - Tipologie di utilizzo degli applicativi di Data Integration (Fonte: [39]. Rielaborazione dell'autore).

Per analizzare il TCO di una soluzione di Data Integration è necessario individuare tutte le tipologie di costi che influenzano il totale. L'analisi proposta in [39] analizza i costi sia in termini di tempo che di denaro speso. In particolare, considera tutte le tipologie di costo derivanti dall'acquisizione e dall'utilizzo di prodotti di integrazione in un arco temporale di 5 anni. L'analisi di Bloor prende in considerazione i prodotti proprietari descritti nel paragrafo 2.2.1, le soluzioni che fanno uso di software personalizzato e i software open source. Per calcolare il TCO di una soluzione di Data Integration nel modo più completo e corretto possibile è necessario considerare i costi che si originano in ogni fase del ciclo di vita del prodotto (Figura 17). Per questo l'analisi di Bloor parte dalla valutazione dei tempi necessari per svolgere le prime fasi del ciclo di vita (analisi dei requisiti e valutazione dei prodotti).

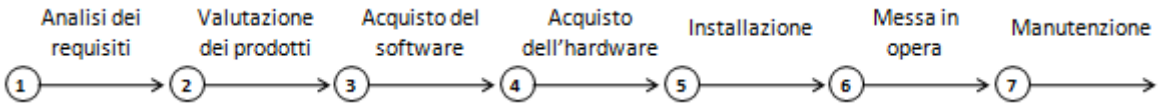

Figura 17 - Il ciclo di vita di un prodotto di Data Integration (Fonte: [38]. Rielaborazione dell'autore).

In Figura 18 è riportato il tempo (in settimane/uomo) utilizzato dalle aziende per completare la fase di analisi dei requisiti precedente all'acquisizione del prodotto.

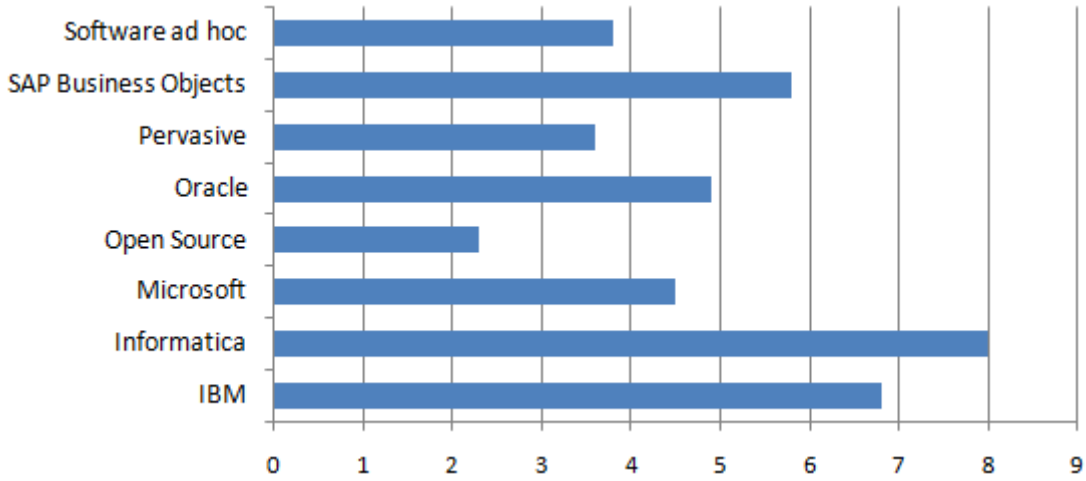

Figura 18 - Tempo necessario per condurre l'analisi dei requisiti (Fonte: [39]. Rielaborazione dell'autore).

A prima vista risaltano i tempi brevi richiesti dai software open source, probabilmente dettati dal fatto che tali software sono gratuitamente disponibili per il download e di conseguenza si possono provare ancora prima di avviare una vera e propria fase di analisi dei requisiti. Come vedremo nei grafici successivi, si può identificare una relazione tra costo del prodotto e tempo necessario per l'analisi dei requisiti (più alto è il costo del prodotto e più lunga è la fase di analisi).

Nella fase successiva l'analisi rimane incentrata sulla variabile temporale, cercando valutare i tempi necessari per l'apprendimento delle funzionalità del prodotto acquistato e il tempo necessario per mettere in produzione la prima soluzione. In Figura 19 vediamo i risultati di questa indagine espressi in settimane. I tempi di entrata in produzione ci danno un'indicazione generale delle tempistiche per la messa in opera del prodotto ma è un indicatore da utilizzare con cautela in quanto non possiamo conoscere con precisione la complessità dei progetti (chiaramente più il progetto è complesso più il tempo di messa in opera si dilata). Tempi lunghi sono probabilmente dovuti alla complessità del progetto sviluppato; questo può significare che alcuni prodotti si prestano meglio di altri all'utilizzo in contesti complessi. Dall'analisi del grafico in Figura 19 si nota come i prodotti open source si contraddistinguano per il breve periodo di apprendimento, mentre i prodotti più completi, e anche più complessi (IBM e Informatica), sono quelli che richiedono tempi di apprendimento più lunghi, determinando un maggiore TCO della soluzione. Tuttavia, il tempo per la messa in opera della prima soluzione non differisce di molto nei prodotti analizzati, ad eccezione di Microsoft che permette un'implementazione in tempi relativamente brevi.

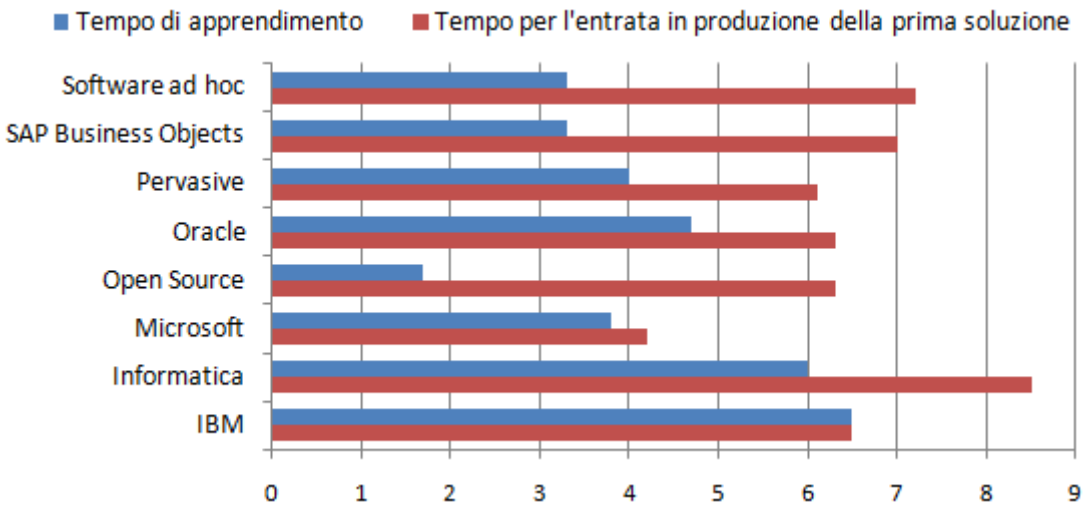

Figura 19 - Tempi per l'apprendimento e la messa in opera del prodotto.
(Fonte:[39]. Rielaborazione dell'autore).

La fase successiva dell'analisi si focalizza sulla valutazione dei costi iniziali di hardware e software che l'azienda deve sostenere per installare i prodotti in questione (Figura 20).

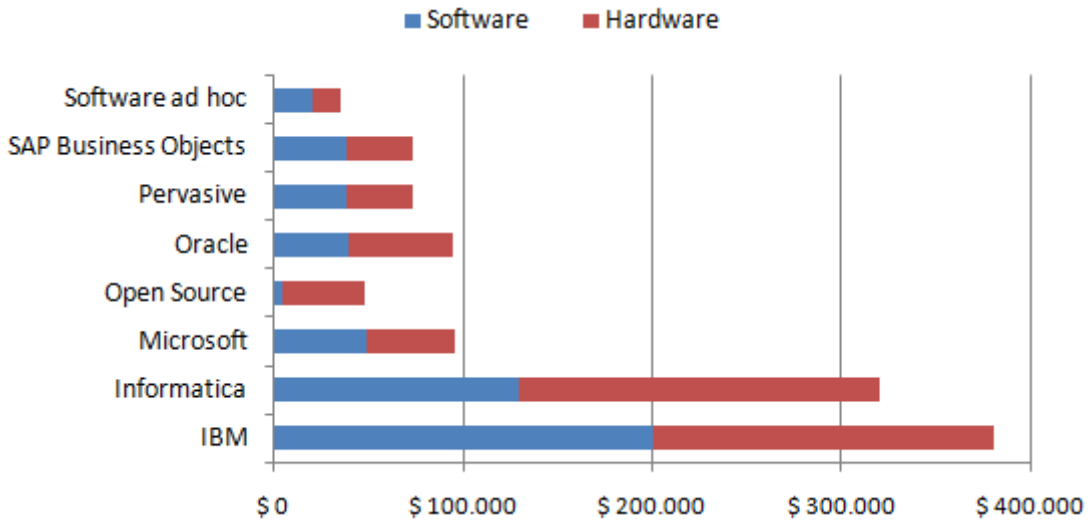

Figura 20 - Costi iniziali di hardware e software (Fonte:[39]. Rielaborazione dell'autore).

Come preannunciato i prodotti di IBM e Informatica sono di gran lunga le soluzioni più costose sul mercato. I due prodotti in questione sono pensati per operare in ambiti complessi contraddistinti da elevate moli di dati; la loro implementazione ha senso soltanto in grandi progetti di integrazione sviluppati solitamente da imprese di grandi dimensioni. Come è lecito aspettarsi i software open source sono contraddistinti da costi software praticamente nulli e da costi hardware limitati, variabili in base alla quantità di dati da processare. Le soluzioni proprietarie più economiche si dimostrano i prodotti di SAP Business Objects e Pervasive. I software personalizzati si caratterizzano per bassi costi sia per l'hardware che per il software, questo si spiega con il fatto che si tratta di prodotti relativamente semplici ed operanti in domini contraddistinti da una complessità e da una quantità di dati limitata.

L'analisi dei costi prosegue con la valutazione dei costi amministrativi, dei costi di manutenzione e dei costi associati al personale tecnico che lavora al progetto, valutati considerando un orizzonte temporale di un anno (Figura 21). Il risultato riflette la situazione vista per i costi iniziali: i prodotti con il costo iniziale maggiore presentano anche costi di gestione più elevati. L'unico prodotto proprietario in grado di rivaleggiare a livello di costi con i software personalizzati e i prodotti open source è la soluzione proposta da Pervasive. Per quanto riguarda i software personalizzati, che in base all'analisi finora condotta appaiono molto competitivi, c'è da dire che un'analisi di questo tipo non tiene tuttavia conto di alcuni fattori: mancano infatti i costi di test e debugging, i costi hardware sono bassi in quanto si riutilizza spesso l'hardware già presente in azienda, inoltre spesso il personale tecnico non si occupa di integrazione a tempo pieno e probabilmente non sempre viene inserito nel computo dei costi delle risorse umane. Nella realtà lo sviluppo di software personalizzato è una soluzione meno competitiva rispetto a quanto rivelato da questa analisi.

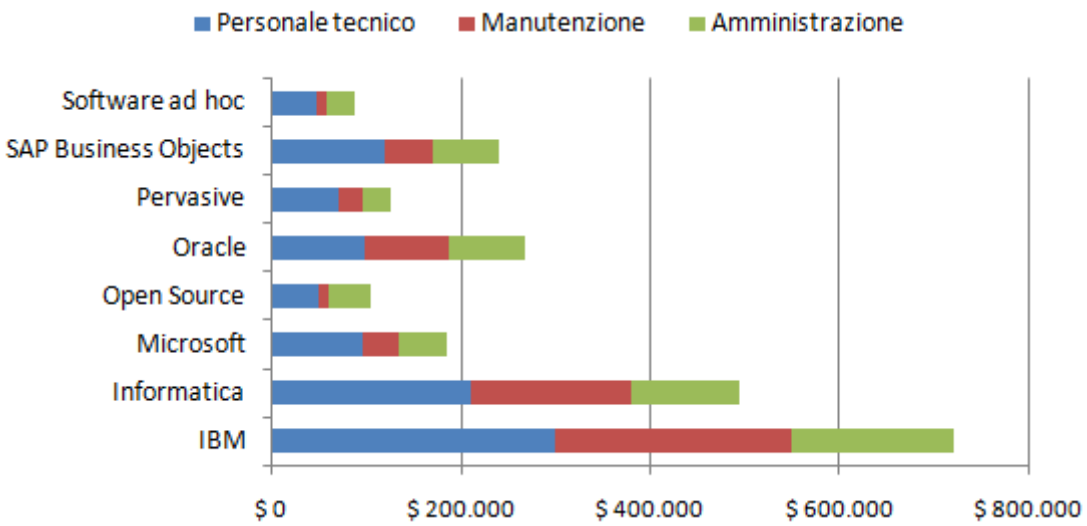

Figura 21 - Costi annuali di un prodotto di Data Integration (Fonte:[39]. Rielaborazione dell'autore).

L'analisi del TCO dei prodotti in questione si conclude con due prospetti riassuntivi dei TCO individuati: il primo evidenzia il TCO su un arco temporale di 5 anni (Figura 22) mentre il secondo rappresenta il TCO medio per lo sviluppo di un singolo progetto (Figura 23). Nel calcolo del TCO nel periodo di 5 anni non sono considerati i costi di analisi e valutazione dei prodotti. Dato che un prodotto di Data Integration è solitamente utilizzato per periodi anche più lunghi di 5 anni, l'analisi proposta in Figura 22 costituisce una possibile guida per indirizzare le aziende nella scelta di un prodotto di integrazione. Da questi ultimi due grafici sono esclusi i software personalizzati poiché, per quanto si è detto in precedenza, l'analisi dei loro costi appare falsata. Da un punto di vista meramente di costo le soluzioni vincenti sono quindi il prodotto di Pervasive e i software open source, seguiti dall'offerta di Microsoft.

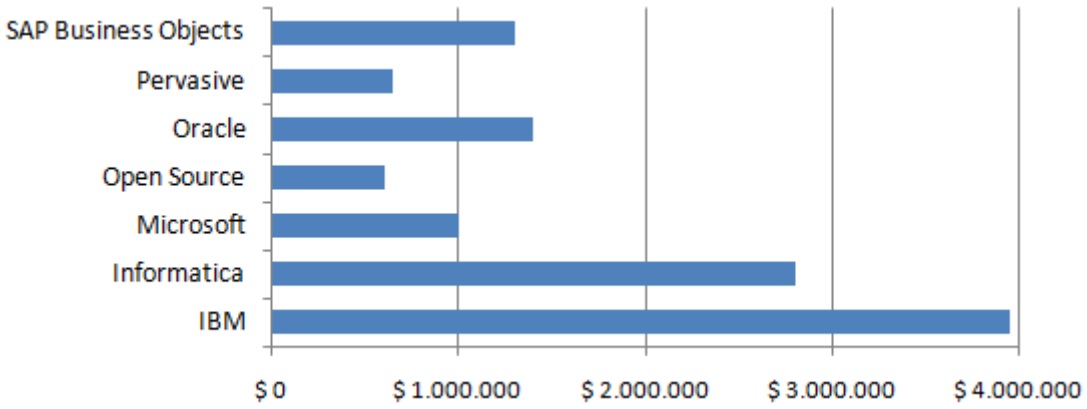

Figura 22 - Total cost of ownership in un periodo di 5 anni (Fonte:[39]. Rielaborazione dell'autore).

Il grafico proposto in Figura 23 presenta i costi medi per progetto delle soluzioni analizzate. Soprattutto nell'ambito della Data Integration la valutazione del costo per progetto è un indicatore molto importante; il solo TCO non sempre è un indicatore

significativo in quanto la realizzazione di un maggior numero di progetti permette di ammortizzare meglio il costo del prodotto. In base ai risultati riportati in Figura 23 possiamo notare come Microsoft e Pervasive presentino il miglior valore sulla base di costo per progetto.

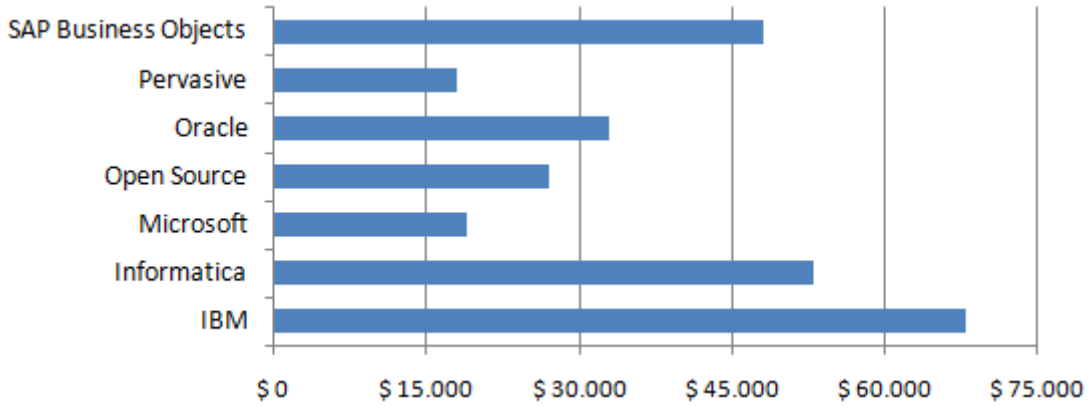

Figura 23 - Total cost of ownership per singolo progetto (Fonte:[39]. Rielaborazione dell'autore).

In conclusione, l'analisi proposta permette di identificare il grado di idoneità delle soluzioni analizzate ad essere implementate in progetti di tipologia, dimensione e complessità differenti. Si può dire che Pervasive è il prodotto da preferire in progetti di piccola dimensione, soprattutto per i costi contenuti. In scenari di media dimensione i prodotti migliori appaiono essere i software open source e la soluzione offerta da Microsoft. I prodotti proposti da SAP Business Objects, Informatica, IBM e Oracle sono invece indirizzati a scenari di medie/grandi dimensioni contraddistinti da complessità e quantità di dati elevate.

2.4 CONCLUSIONI

Per molti anni i prodotti di Data Integration sono stati considerati come soluzioni costose e impegnative che richiedevano ingenti investimenti per ottenere risultati appena discreti. Tuttavia, le recenti evoluzioni del contesto economico globale rendono sempre più chiara ed evidente la necessità di disporre di dati integrati. Circa 15 anni fa, successivamente all'introduzione del concetto di Data Warehouse e con lo sviluppo delle tecniche ETL, i prodotti di integrazione promettevano di rendere disponibili le seguenti funzionalità [40]:

- Sviluppo e implementazione più veloce ed efficiente
- Riuso del codice
- Maggiore standardizzazione
- Aumento della qualità dei dati
- Riduzione delle operazioni di mappatura manuale dei dati
- Maggiore trasparenza del processo di integrazione

- Controllo e gestione dei cambiamenti
- Migrazioni di dati facili e veloci
- Flessibilità interna all'applicazione
- Interoperabilità con altri applicativi
- Facilità di utilizzo

In realtà tali promesse non sono state completamente realizzate. Per questo motivo è oggi necessaria un'evoluzione dei prodotti di Data Integration, un'evoluzione capace di sfruttare nuove opportunità e superare alcune sfide, come ad esempio [40]:

- *Nuove architetture*: la maggior parte dei prodotti sul mercato si basa ancora su architetture vecchie di anni. L'apertura verso architetture innovative in grado di far fronte alle necessità del mercato attuale è senza dubbio una strada da percorrere;
- *Modello dei prezzi innovativo*: molti prodotti si basano ancora su un modello di prezzi datato, basato su licenze per macchina, piuttosto che su un modello innovativo, basato sull'effettivo utilizzo (throughput) del sistema. Il primo passo verso l'adozione di questo modello di prezzi è stato fatto dai primi prodotti di integrazione open source che hanno scelto questo tipo di approccio per venire incontro alle necessità delle aziende con budget limitati;
- *Riduzione dei tempi di progettazione*: nella maggior parte dei progetti di integrazione troppo tempo viene consumato per la fase di progettazione del sistema. È necessario ripensare questa fase spostando la fase di design a un livello superiore dove termini di business universalmente riconosciuti possano essere usati per descrivere il contesto di riferimento codificandolo in un modello astratto della realtà;
- *Potenziamento di soluzioni SOA e Saas*: l'outsourcing di dati e servizi ha reso fondamentale esporre i dati verso l'esterno. In questo campo un prodotto di Data Integration potrebbe fungere da servizio in grado di standardizzare lo scambio e la rappresentazione dei dati in modo da facilitarne l'integrazione in viste virtuali unificate.

Tuttavia, in ottica futura, rispondere a queste sfide può non essere sufficiente. Le recenti evoluzioni del contesto informatico hanno infatti amplificato il problema dell'integrazione dei dati. La riduzione dei prezzi dell'hardware, l'aumento della capacità di memorizzazione, l'esplosione della rete internet, l'utilizzo massivo di sensori RFID[10] e l'esternalizzazione dei servizi hanno portato ad una crescita esponenziale della quantità di dati disponibili. Di conseguenza i prodotti di Data Integration si trovano a fronteggiare un

[10] *Radio Frequency Identification (RFID)*, è una tecnologia per l'identificazione automatica di oggetti, animali o persone. Si basa sulla capacità di accedere a distanza a dati memorizzati su dispositivi che sono in grado di rispondere comunicando le informazioni in essi contenute.

contesto nuovo contraddistinto da grandi quantità dati, memorizzati in strutture caratterizzate da schemi e linguaggi di rappresentazione differenti. Tutto questo determina la necessità di funzionalità innovative che garantiscano la realizzazione di sistemi in grado di supportare [40] [41] [42]:

- **Razionalizzazione semantica**: la definizione di una semantica condivisa di terminologie, definizioni e regole permetterebbe di creare dei modelli di dati standardizzabili, flessibili e condivisi e di conseguenza facilmente integrabili;
- **Astrazione**: nel senso di separare il significato delle cose dalla loro implementazione fisica. Sulla base della semantica definita è possibile creare diversi strati di astrazione a supporto dell'integrazione dei dati per scopi differenti;
- **Scalabilità**: i contesti nei quali si originano i dati sono oggi meno prevedibili e controllabili che in passato (RFID, Web, SOA, ecc.). Di conseguenza risultano meno prevedibili anche i tempi e i modi con i quali procedere al caricamento e alla trasformazione dei dati. Fare affidamento ad uno schema dei dati sostanzialmente fisso (Data Warehouse) può risultare controproducente e costoso. Modelli e schemi di dati più flessibili sono senza dubbio una necessità in molti contesti;
- **Velocità e performance**: l'aumento della velocità del business moderno richiede applicativi in grado di fornire dati di sintesi e indicatori aggiornati pressoché in tempo reale. Inoltre, il costante aumento della quantità di dati disponibili pone un problema di performance. Le nuove architetture dovranno riuscire a gestire quantità di dati fino a pochi anni fa inimmaginabili, con fattori di crescita stimati nell'ordine di 10–100 volte nell'arco temporale di 5 anni;
- **Produttività e riuso**: negli strumenti attuali si denota una carenza nella gestione del ciclo di vita del progetto. Carenza che spesso limita la produttività della soluzione sviluppata e ne impedisce un eventuale riuso. La realizzazione di una repository dei metadati condivisa potrebbe fungere come punto di raccolta e di contatto tra le informazioni e i dati che si generano nelle varie fasi del progetto, permettendo una migliore gestione del ciclo di vita.

Nel capitolo seguente vedremo come l'implementazione di tecnologie semantiche possa contribuire a migliorare alcuni degli aspetti critici degli attuali prodotti di integrazione.

CAPITOLO 3

METODOLOGIE SEMANTICHE

Durante lo svolgimento dello stage è stata condotta un'attività di ricerca sullo stato dell'arte della disciplina della Data Integration che ha condotto all'analisi delle più moderne e innovative tecnologie a supporto dell'integrazione dei dati. In questo capitolo vengono presentate le principali e più promettenti metodologie innovative nell'ambito della Data Integration, con particolare riferimento al ruolo sempre più rilevante della semantica. Per questo viene innanzitutto analizzato il concetto di semantica cercando di capire perché sta assumendo sempre maggiore importanza nel settore della gestione dei dati. Segue quindi una breve esposizione del concetto di Semantic Web e dei linguaggi che lo compongono in quanto tecnologie utilizzate dalle metodologie innovative di Data Integration. In seguito, vengono esposte le principali metodologie e tecniche semantiche a supporto dell'integrazione dei dati, cercando di evidenziare quali sono i vantaggi che apportano rispetto al modello relazionale e quali possono essere eventuali problemi e criticità nell'adottare tali metodologie in un contesto di reale applicazione.

3.1 IL RUOLO EMERGENTE DELLA SEMANTICA

Le tecnologie informatiche tradizionali si trovano oggi a fronteggiare una serie di problematiche: conversione in web-service, esternalizzazione dei business, esplosione della quantità di dati, assenza di strumenti in grado di usare proficuamente l'abbondanza di informazioni, ingente peso dei costi di manutenzione e necessità di sistemi operanti in tempo reale con funzionalità di analisi avanzate [43]. Non è facile rispondere a tali necessità continuando ad utilizzare rappresentazioni astratte ed arbitrarie dei dati, è necessario codificare il significato dei dati, la loro semantica.

Il termine semantica individua la disciplina che studia il significato di parole, frasi e testi. Nel settore dell'Information and Communication Technology (ICT) l'utilizzo di tecnologie semantiche garantisce la possibilità di inferire il significato dei dati e di determinare la loro utilità in un determinato contesto. Le tecnologie innovative di Data Integration fanno uso delle tecnologie semantiche sviluppate con l'obiettivo realizzare una rete fondata sulla conoscenza che Tim Berners-Lee ha definito come Semantic Web[1]. Attraverso l'utilizzo di

[1] Tim Berners-Lee, informatico britannico noto per l'invenzione del World Wide Web (1991), nel 1999 pubblica il libro *Weaving the Web* nel quale esprime la sua visione sul Web del futuro, un Web intelligente in grado di dare un significato ai dati che contiene, coniando di fatto la definizione di Semantic Web.

tecnologie semantiche si può cercare di rappresentare la realtà in tutta la sua complessità, superando di fatto i limiti imposti dalla tecnologia relazionale. L'utilizzo di tecnologie semantiche garantisce l'interoperabilità tra dati, informazioni e sistemi andando oltre una normale integrazione. Applicazioni semplici e operanti in domini limitati possono continuare ad essere produttive senza l'applicazione di tecnologie semantiche. Tuttavia, l'ICT deve oggi affrontare sfide spesso non localizzate: l'esternalizzazione di business, la necessità di collaborare con clienti e fornitori rende il contesto ampio, complesso e spesso caratterizzato dalla presenza di dati scarsamente strutturati e organizzati. La necessità di prodotti in grado di integrare, memorizzare ed elaborare informazioni più ricche e dettagliate in tempo reale è in costante aumento.

3.1.1 I LIMITI DEL MODELLO RELAZIONALE

All'interno delle organizzazioni si presenta spesso un problema: i manager si trovano ad analizzare sofisticati report che talvolta dicono l'uno il contrario dell'altro in relaziona alla medesima problematica. La domanda corretta da porsi non è quale di questi sia quello giusto ma se ci troviamo di fronte ad un problema dei report o ad un problema di conoscenza. Pochissimi contesti di business mantengono le loro caratteristiche inalterate nel tempo, la maggior parte ha caratteristiche diverse, in costante evoluzione e spesso in conflitto tra loro anche se osservate dallo stesso punto di vista. Il problema dei tradizionali sistemi di Data Integration è dato dal fatto che cercano di operare in un mondo fatto di significati variegati e in costante evoluzione utilizzando una singola rappresentazione della realtà, una realtà con un significato univoco.

L'origine di questo problema è da ricercare agli albori della disciplina del Data Management. Quando vennero sviluppati i primi sistemi di gestione dei dati ci si trovava in un contesto caratterizzato da computer poco potenti con quantità di memoria limitate. L'unico approccio possibile era quindi rappresentare il singolo contesto operativo nel modo più semplice possibile limitando il carico di lavoro per l'elaboratore e il consumo di memoria. La conseguenza fu la proliferazione di sistemi incentrati su singole attività, spesso tra loro incompatibili e scarsamente inclini all'integrazione. Inoltre, tale metodologia è stata utilizzata anche per rispondere alle necessità di integrazione dei dati determinando lo sviluppo di prodotti di Data Integration caratterizzati da semplicità e povertà di rappresentazione. In particolare, il concetto di Data Warehouse si è basato per anni sull'assunzione che tutti i dati memorizzati avessero un significato univoco. Nella comunità informatica tale fenomeno è conosciuto con il nome di "*Single Version of the Truth*" e costituisce un forte limite dei sistemi di Data Management e Data Integration tradizionali.

In seguito all'esternalizzazione dei business e al successo della rete Internet la complessità del contesto e le necessità di dati integrati sono aumentate costantemente. In particolare, il

numero di possibili punti di integrazione (P) di *n* sistemi è dato da [43]: $P = n * (n - 1)/2$, il che significa che tra 20 sistemi differenti vi sono 190 punti di connessione. Se pensiamo a quanti sistemi operano nella rete Web ci rendiamo conto della complessità che le tecnologie ICT si trovano ad affrontare. Soltanto metodi di integrazione semantici possono garantire una soluzione praticabile, la realizzazione di un modello semantico dei dati permetterebbe la realizzazione di un sistema di collegamento uniforme, in grado di cogliere nei dettagli tutte le relazioni tra gli elementi di sistemi diversi.

Nonostante le evidenti necessità di implementare strumenti semantici per la gestione dei dati il mondo ICT presenta delle barriere nell'adozione di tali tecnologie. Una possibile causa di tale comportamento è data dal fatto che 30 anni di esperienza con la tecnologia relazionale hanno contribuito a creare negli individui una mentalità ristretta dalla realtà semplificata che hanno creato, il che porta ad accettare le limitazioni tipiche del modello relazionale. Inoltre, in molti si chiedono perché sia necessario sviluppare una nuova tecnologia quando con l'attuale è possibile gestire i medesimi contesti applicativi, seppur con dei compromessi più o meno evidenti.

La tecnologia relazionale ha raggiunto oggi ottimi livelli di affidabilità e consistenza nella memorizzazione dei dati ma la parte di intelligenza, necessaria per dare un senso a ciò che si è memorizzato, va aggiunta manualmente ogni qual volta si presenta la necessità di ottenere qualche informazione aggiuntiva dai dati. Nelle tecnologie semantiche una parte di intelligenza è codificata direttamente nel modello dei dati e questo permette al sistema di inferire automaticamente nuove informazioni basandosi su relazioni, regole e vincoli definiti sui dati stessi. La tecnologia semantica fornisce inoltre strumenti per modellare domini complessi, come ad esempio le ontologie. Un'ontologia descrive i concetti e le relazioni che contraddistinguono un determinato dominio ma senza focalizzarsi su un singolo ambito di applicazione, può essere utilizzata per scopi differenti ed ampliata in base a nuove esigenze, originate da variazioni del contesto di riferimento.

Un sistema semantico, a differenza di un database relazionale, non è solo un contenitore di dati ma è un sistema software attivo e dichiarativo. Un sistema relazionale diventa una cosa di questo tipo solo dopo essere stato completato con viste, queries, script di caricamento, stored procedures (programmi interni al database). La differenza con un modello semantico è data dal fatto che, in quest'ultimo, dati, metadati, vincoli e costrutti logici sono contenuti tutti nella medesima struttura e le relazioni tra tali elementi sono definite esplicitamente. In un sistema relazionale questi elementi sono dispersi e spesso nascosti all'utilizzatore. Inoltre, grazie all'utilizzo della struttura a grafo, il modello semantico risulta dinamico ed è in grado di aumentare la propria conoscenza attraverso la ricerca di conoscenza implicita, sfruttando la struttura del grafo. In un modello semantico le informazioni relative ad una relazione tra elementi possono essere inferite direttamente

dai dati, mentre in un database relazionale il contesto va specificato di volta in volta in ogni query (ad esempio tramite la clausola SQL WHERE).

3.2 LE TECNOLOGIE DEL SEMANTIC WEB

Con la definizione del Semantic Web Tim Berners-Lee concretizza l'idea di dare un valore aggiunto al Web tradizionale. Un valore aggiunto che permette di processare in maniera automatica parte dei dati e delle informazioni attraverso l'applicazione della logica e di linguaggi studiati per rappresentare ed esporre sul Web la conoscenza. Il Semantic Web non cambia i principi alla base del Web originario ma ne rappresenta il giusto completamento. Il World Wide Web Consortium (W3C) sta sostenendo attivamente le tecnologie del Semantic Web per agevolare il Web ad esprimere tutto il suo potenziale, il suo motto recita infatti:

> *"Leading the Web to its Full Potential."*

Tim Berners-Lee da questa definizione di Semantic Web [44]:

> *"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."*

Tale definizione focalizza i tre punti chiave del Semantic Web:

- È un'estensione del Web attuale;
- Ha come obiettivo la cooperazione tra computer e individui;
- È realizzabile solo assegnando un significato preciso ai dati.

Attraverso l'implementazione di un Web semantico è possibile permettere alle macchine di elaborare automaticamente le informazioni attraverso l'utilizzo della logica. Tuttavia, per fare questo è necessario trovare un linguaggio logico in grado di gestire sia i dati che le regole per il ragionamento logico. Un linguaggio che deve risultare abbastanza espressivo per descrivere contesti complessi ma allo stesso tempo deve essere semplice e leggero per non rischiare di incorrere in paradossi o inconsistenze.

Per spiegare meglio il funzionamento del Semantic Web e di quali tecnologie necessita, Tim Berners-Lee e i ricercatori del W3C hanno realizzato una rappresentazione schematica dei livelli necessari per concretizzare il Semantic Web (la cosiddetta Semantic Web Stack, Figura 24). Si tratta di una struttura modulare dove ad ogni livello corrisponde un determinato linguaggio, la figura rappresenta infatti una gerarchia di linguaggi dove ogni livello sfrutta le capacità del livello inferiore. Le tecnologie che vanno dalla base (URI/IRI) al linguaggio OWL sono già state standardizzate dal W3C e sono già oggi pienamente utilizzabili, tuttavia non è ancora del tutto chiaro come la parte superiore della gerarchia

verrà implementata. Per ottenere la piena realizzazione del Semantic Web sarà infatti necessario sviluppare tutti i livelli rappresentati [45].

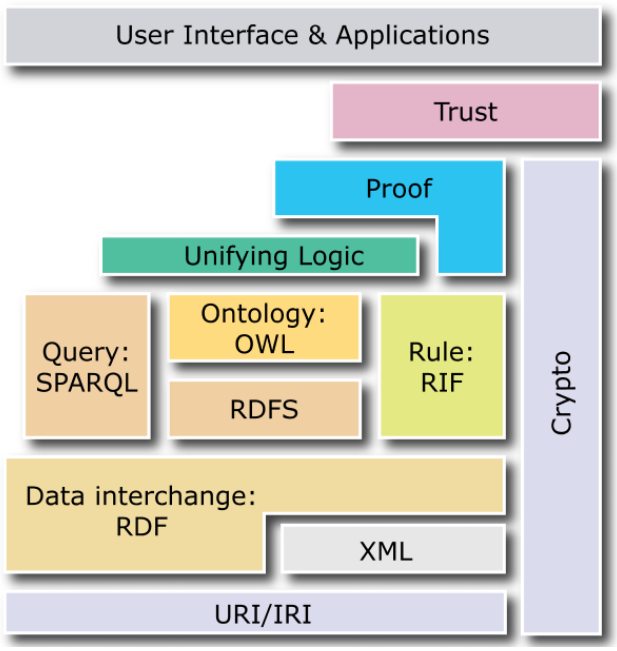

Figura 24 – I livelli del Semantic Web (Fonte: W3C. Semantic Web Layer Cake[2], marzo 2007).

Vediamo quindi sinteticamente quale è il ruolo di ogni singolo livello della gerarchia. Partendo dal basso possiamo identificare quel gruppo di tecnologie che sono alla base del Web tradizionale e che rimanendo immutate forniscono le basi anche per il Web semantico:

Uniform Resource Identifier (URI)

Sono lo strumento utilizzato dal Web tradizionale per identificare univocamente una risorsa generica (indirizzo Web, documento, immagine, file, ecc.). Consistono in una stringa di caratteri (ASCII[3]) che identificano un nome o una risorsa nella rete Internet.

Internationalized Resource Identifier (IRI)

Sono una generalizzazione delle URI, sono lo strumento che garantisce la possibilità di identificare univocamente le risorse del Web semantico. A differenza delle URI, che utilizzano i caratteri previsti dal sistema ASCII, possono contenere caratteri appartenenti all'Universal Character Set[4].

Al di sopra di queste tecnologie di base troviamo le tecnologie che definiscono formati di memorizzazione e interscambio di dati:

---

[2] http://www.w3.org/2007/03/layerCake.png
[3] Sistema di codifica dei caratteri a 7 bit, accettato come standard dall'ISO (ISO 646).
[4] Standard per la rappresentazione di caratteri a 16 o 32 bit definito dall'ISO/IEC 10646, può contenere circa un milione di caratteri diversi.

eXtensible Markup Language (XML)

È un linguaggio che permette agli utenti di creare etichette personalizzate (tag). Più precisamente, XML è un metalinguaggio che permette di creare nuovi linguaggi con lo scopo di descrivere dati strutturati. Permette di ottenere un'interoperabilità sintattica completa (capacità di interpretare la sintassi dei file scambiati) e un'interoperabilità strutturale parziale (interpretare la struttura di schemi logici differenti). XML non include una descrizione semantica del significato dei dati e non garantisce quindi nessuna interoperabilità semantica.

Resource Description Framework (RDF)

È un linguaggio creato per realizzare modelli di dati in grado di descrivere gli oggetti e le loro relazioni. In sostanza è il modello dei dati definito e promosso dal W3C con l'obiettivo di favorire un'interoperabilità semantica, una capacità di scambio basata sul significato dell'informazione invece che sulla sua struttura. Ogni elemento descritto tramite RDF è definito risorsa ed ogni risorsa è identificata da una URI. Nel modello RDF i dati vengono memorizzati in triple costituite dai seguenti elementi: *soggetto – predicato – oggetto*. In questo modo è possibile codificare asserzioni sulle risorse affermando che un dato elemento (soggetto) ha delle proprietà (predicato) con determinati valori (oggetto). Il linguaggio RDF permette quindi di definire relazioni tra risorse tramite proprietà e valori, tuttavia RDF non fornisce alcun meccanismo per dichiarare tali proprietà, ne per definire le relazioni tra proprietà e risorse. Inoltre, URI diverse possono identificare la stessa risorsa, per garantire un'interoperabilità semantica completa è necessario uno strumento in grado di determinare quando due identificatori hanno lo stesso significato.

Per superare i limiti del linguaggio RDF si è reso necessario lo sviluppo di un livello superiore della gerarchia del Semantic Web. Tale livello prevede l'utilizzo di linguaggi ontologici:

RDF Schema (RDFS)

Il linguaggio RDF Schema ha il compito di aggiungere un primo livello di logica a RDF. RDFS permette di definire nuovi tipi di classi, introduce inoltre il concetto di sottoclasse permettendo di definire gerarchie tra classi. Infine, RDFS permette di creare proprietà per descrivere un concetto, consentendo di esprimere in maniera sistematica le proprietà di elementi simili.

Web Ontology Language (OWL)

Nonostante RDFS presenti alcuni strumenti per definire modelli concettuali basati su RDF, il mondo della logica è molto più complesso e per ottenere un linguaggio in grado abilitare il ragionamento automatico è necessario utilizzare un linguaggio più completo. Tale necessità è stata standardizzata in OWL, un linguaggio ontologico che utilizza

alcune parti della First Order Logic (FOL[5]) per garantire una maggiore espressività rispetto a RDFS. OWL definisce costrutti avanzati per descrivere la semantica di risorse RDF. OWL supporta l'aggiunta di vincoli, restrizioni e caratteristiche delle proprietà (transitiva, ecc.) che permettono di abilitare il ragionamento logico automatico nel Semantic Web.

Accanto ai linguaggi ontologici si posizionano due linguaggi con il ruolo di assolvere compiti specifici:

SPARQL Protocol and RDF Query Language (SPARQL)

SPARQL è un linguaggio di interrogazione per RDF. SPARQL fornisce le funzionalità per interrogare insiemi di triple RDF e anche un protocollo di comunicazione che permette di effettuare le richieste in ambiente Web.

Rule Interchange Format (RIF)

RIF è un linguaggio che si pone l'obiettivo di fornire un formato di interscambio per differenti linguaggi di regole e motori di inferenza. Le regole sono utili perché ampliano la capacità delle macchine di elaborare conoscenza rappresentando le relazioni che non possono essere descritte direttamente in un'ontologia.

Infine, nella parte alta della gerarchia dei linguaggi del Semantic Web vi sono quelle tecnologie che tutt'oggi non sono ancora completamente standardizzate, ma che sono necessarie per la completa affermazione del Semantic Web:

Trust

Tali tecnologie dovrebbero avere il compito di verificare l'affidabilità delle informazioni contenute nel Semantic Web attraverso due modalità:

1. Verifica dell'affidabilità della fonte informativa;
2. Affidandosi alla logica per derivare nuove informazioni.

Proof

Il livello della prova dovrebbe garantire all'utente la possibilità di capire come il trust è stato calcolato. Dovrebbe quindi rivelare all'utente la logica utilizzata dal sistema nel ragionamento automatico.

Unifying Logic

Il livello della prova è raggiungibile soltanto se tutto si basa su una logica solida, quindi il livello sottostante alla prova deve presentare una logica unificante in grado di abilitare la prova ma al tempo stesso di descrivere l'informazione attraverso i linguaggi ontologici sottostanti. Tuttavia, esistono diversi gruppi di logiche, con proprietà ed espressività

[5] La First Order Logic permette di definire un linguaggio simbolico per tradurre gli enunciati del linguaggio naturale in formule atomiche interpretabili dalle macchine in maniera automatica.

differenti, la comunità del Semantic Web non ha ancora stabilito con precisione quale dovrà essere questo linguaggio unificante.

Infine, la parte più alta della gerarchia è occupata da una tecnologia non propriamente specifica del Semantic Web, il livello delle applicazioni e delle interfacce utente (***User Interface & Application***) che possono essere realizzate sulla base dell'interoperabilità tra le tecnologie sottostanti. Anche la parte laterale della gerarchia, occupata dalla dicitura ***Crypto***, non individua una tecnologia tipica del Semantic Web. Il termine "*Crypto*" sta infatti ad indicare quell'insieme di tecnologie di crittografia e firma digitale che consentono di assicurare il raggiungimento del trust, ovvero che permettono agli agenti software di verificare il livello di affidabilità delle informazioni contenute nelle asserzioni RDF.

3.2.1 IL MODELLO A GRAFO

Con l'avvento del Web si è iniziato a pensare a nuove modalità per la memorizzazione e lo scambio di dati, tecnologie che potessero garantire una maggiore interoperabilità tra dati rispetto a quanto possibile con il modello relazionale. In sostanza, tale necessità si è concretizzata in un primo momento con la nascita e l'affermazione dei linguaggi basati su XML. Nel corso degli anni 2000 il linguaggio XML si è affermato come formato di esportazione ed importazione di dati tra sistemi ed applicazioni differenti. Il successo di tale linguaggio si spiega con il fatto che permette lo scambio e l'integrazione di dati tra applicazioni differenti che utilizzano il Web come canale di comunicazione. Tuttavia, XML presenta dei limiti che non lo rendono adatto ad essere il modello dei dati per il Web.

Attraverso XML è possibile scambiare informazioni tra due sistemi diversi con schemi differenti grazie alla realizzazione di un codice che trasforma le informazioni da un formato all'altro. Per realizzare le trasformazioni è stato definito un apposito linguaggio: l'eXtensible Stylesheet Language Transformations (XSLT). Tuttavia, tali trasformazioni risultano spesso molto complesse e in uno scenario che considera l'intero Web il numero di trasformate XSLT necessarie cresce esponenzialmente in funzione del numero di schemi differenti da integrare.

Il linguaggio XML modella i dati attraverso una struttura ad albero (Figura 25b), formato estendibile, flessibile e particolarmente adatto a veicolare informazioni sul Web. Tale modello presenta però un grande limite: scarsa espressività nella definizione di relazioni complesse tra i dati. D'altra parte il modello relazionale (Figura 25a) garantisce la possibilità di definire con maggior dettaglio le relazioni tra i dati ma allo stesso tempo non supporta la definizione di schemi di dati estensibili e flessibili. Per questo la comunità del Semantic Web si è preoccupata di ricercare un modello dei dati in grado di racchiudere i

vantaggi sia del modello relazionale che del modello ad albero. Tale modello è il grafo orientato etichettato (Figura 25c), in cui dati e relazioni vengono rappresentati tramite un insieme di nodi collegati da relazioni etichettate e orientate [45]. Con il modello a grafo è quindi possibile modellare i dati secondo lo schema relazionale, il modello ad albero o un misto dei due. Questo conferisce al modello a grafo la giusta espressività nel modellare le relazioni e al tempo stesso una notevole flessibilità ed estendibilità dello schema, il che lo rende il modello ideale per rappresentare i dati sul Web. Il linguaggio RDF rappresenta i dati secondo il modello a grafo ed è oggi riconosciuto come il formato standard dei dati per il Semantic Web.

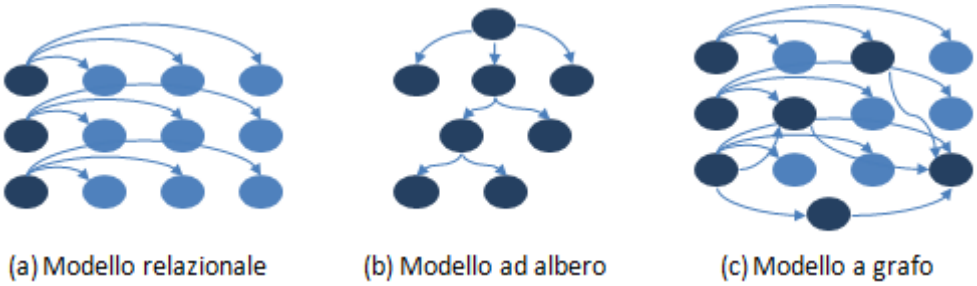

Figura 25 – Il modello del grafo etichettato e orientato (Fonte: [45]. Rielaborazione dell'autore).

In particolare, il linguaggio RDF si pone l'obiettivo di definire un sistema per la descrizione delle risorse, dove ognuna di queste sia identificabile univocamente. Per fare questo RDF prevede come struttura di base l'enunciato, detto anche *tripla*, formato da tre componenti rappresentabili attraverso il grafo (Figura 26):

- **Soggetto**: l'identificativo della risorsa;
- **Predicato**: la proprietà che si vuole descrivere;
- **Oggetto**: il valore che assume il predicato.

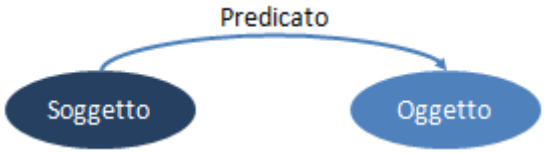

Figura 26 - Rappresentazione di una tripla RDF con il modello a grafo.

RDF prevede che ogni parte dell'enunciato sia identificata univocamente attraverso le URI o le IRI. Ad esempio, l'enunciato "*Dante Alighieri è l'autore della Divina Commedia*" si traduce in RDF con la seguente sintassi:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description about="http://www.book.it/DivinaCommedia">
      <dc:creator>Dante Alighieri</dc:creator>
</rdf:Description>
</rdf:RDF>
```

Un modello di questo tipo risulta flessibile ed efficiente quando è necessario integrare dati provenienti da una grande varietà di fonti differenti, ognuna con un proprio punto di vista su una determinata risorsa. La flessibilità del modello a grafo permette inoltre di rappresentare facilmente dati relazionali, dati in forma tabellare e gerarchica garantendo quindi un'ottima compatibilità con i tradizionali modelli di dati. In Tabella 2 sono riportate le principali differenze tra lo schema dati relazionale e il modello semantico.

MODELLO SEMANTICO (grafo)	**MODELLO RELAZIONALE**
Dati rappresentati a livello concettuale.	Dati rappresentati a livello fisico.
Modello dei dati flessibile e in grado di esprimere ogni tipo di relazione tra i dati. Lo schema è realizzato attraverso l'utilizzo di linguaggi ontologici.	I dati sono contenuti in uno schema rigido. Il linguaggio dello schema ha un'espressività molto limitata se confrontata con i linguaggi ontologici.
Memorizza informazioni: dati con un contesto. Il significato dei dati è formalmente espresso ed esplicitato.	Memorizza dati: il loro significato è implicito. Le relazioni sono tradotte in un insieme di colonne e restrizioni
Permette la classificazione gerarchica delle relazioni tra oggetti, concetti e relazioni principali.	Non supporta la classificazione gerarchica di relazioni e concetti.
Linguaggio di interrogazione flessibile, non dipendente dallo schema dei dati. Tutto è esplicito.	Le queries sono fortemente legate allo schema dei dati. Richiede l'esplicitazione di alcune relazioni di base (joins).
Supporto di sistemi a regole che permettono di esprimere nuovi concetti e relazioni in aggiunta a quelle esistenti.	Concetti e relazioni sono limitati a ciò che è descritto attraverso lo schema. Non sono supportati linguaggi in grado di derivare nuove relazioni e concetti.
Capacità di inferire automaticamente nuove informazioni a partire dai dati di origine attraverso l'implementazione della logica.	Nessuna capacità di inferenza automatica.
Accesso alle informazioni indipendente dallo schema e dipendente dal dominio di applicazione. Situazione ideale per integrare informazioni da fonti diverse.	Accesso alle informazioni fortemente dipendente dallo schema. Difficoltà nell'integrare e riconciliare dati da applicazioni e fonti diverse.

Tabella 2 - Principali differenze tra modello dei dati semantico e modello relazionale (Fonte: [46]).

3.2.2 ONTOLOGIE

Nella descrizione dei linguaggi che compongono la gerarchia del Semantic Web è stata evidenziata l'importanza dei linguaggi ontologici per conferire al Semantic Web la piena realizzazione. Le ontologie rappresentano infatti una delle componenti più importanti della gerarchia, attraverso tali artefatti è possibile rappresentare contesti reali ad un livello di complessità notevole. L'ontologia è uno dei principali rami della filosofia e si occupa dello studio dell'essere in quanto tale e delle sue categorie fondamentali. Nel passaggio dalla

filosofia all'informatica l'ontologia diviene lo strumento per descrivere, rappresentare, concettualizzare quella parte del mondo che ci interessa modellare in un'applicazione. L'utilizzo del termine ontologia in ambito informatico risale ai tempi dei primi sistemi di Intelligenza Artificiale (IA), in particolare il termine diventa di uso comune nella comunità della IA nel corso degli anni '80. Tuttavia, solo con la definizione del concetto di Semantic Web il termine ontologia assume grande popolarità anche al di fuori della ristretta cerchia di ricercatori. Con riferimento alle tecnologie del Semantic Web il termine ontologia viene usato per definire modelli generici di dati. Un'ontologia indica una struttura di dati che contiene le entità rilevanti, le relazioni tra di esse, regole e vincoli specifici del dominio che si vuole rappresentare. Il tutto viene formalizzato tramite l'utilizzo di linguaggi ontologici (OWL in particolare) che rispondono alle leggi della logica formale.

Negli anni sono state proposte diverse definizioni di ontologia, in un primo momento da parte dei ricercatori dell'Intelligenza Artificiale e successivamente dalla comunità del Semantic Web. La definizione più conosciuta è quella che Thomas Gruber ha dato nel 1993 [47]:

"An ontology is an explicit specification of a conceptualization."

La definizione di Gruber pone l'accento su due caratteristiche fondamentali di un'ontologia:

- Un'ontologia è una **concettualizzazione**: definisce i concetti rilevanti per tradurre il dominio in esame in un modello astratto, talvolta semplificato, della realtà;
- Un'ontologia è una **specificazione**: la concettualizzazione si ottiene attraverso la rappresentazione minuziosa di concetti e relazioni che sussistono nel dominio.

Nel 1998, partendo dalla descrizione di Gruber, i ricercatori Rudi Studer, Richard Benjamins e Dieter Fensel elaborano una definizione più completa del concetto di ontologia [48]:

"An ontology is a formal, explicit specification of a shared conceptualization."

I tre ricercatori, con l'obiettivo di dettagliare maggiormente il significato del termine, aggiungono alla definizione originale tre aggettivi asserendo che un'ontologia deve essere:

- **Formale**: espressa in un linguaggio o formato interpretabile dalle macchine;
- **Esplicita**: i concetti utilizzati e le restrizioni sul loro utilizzo devono essere definiti esplicitamente senza lasciare spazio ad ambiguità;
- **Condivisa**: un'ontologia non si riferisce al pensiero di un singolo ma cattura la conoscenza di un gruppo. Pertanto, tale conoscenza deve essere condivisibile e utilizzabile da tutta la comunità.

Nel campo dell'informatica le ontologie sono oggi utilizzate come sistemi di rappresentazione e condivisione della conoscenza. Grazie alla loro flessibilità e potenza di

rappresentazione le ontologie possono essere utilizzate per scopi differenti: classificazione e rappresentazione di contesti reali, ragionamento deduttivo e inferenza automatica di nuove informazioni, comunicazione, scambio e integrazione di informazioni tra sistemi diversi.

Le ontologie rappresentate tramite i linguaggi ontologici moderni presentano una struttura comune, che prevede la presenza dei seguenti componenti di base[6]:

- **Individui**: sono il livello base dell'ontologia, si tratta di oggetti concreti (animali, persone, case, automobili, ecc.) o di rappresentazioni astratte (parole, numeri, ecc.).
- **Classi**: sono insiemi, collezioni di determinate tipologie di oggetti che condividono le medesime caratteristiche. Esempi di classi sono: persone, veicoli, automobili. Le ontologie prevedono la realizzazione di una gerarchia delle classi, in modo che automobili sia una sottoclasse di veicoli;
- **Attributi**: gli oggetti rappresentati in un'ontologia possono essere descritti in relazione tra loro o rispetto ad altri elementi, definiti attributi. Ogni attributo può essere una classe o un individuo e il suo valore può essere un tipo di dato complesso;
- **Relazioni**: specificano come gli oggetti che compongono l'ontologia sono tra loro collegati. Un dei punti di forza delle ontologie è la loro capacità di espressione nel definire relazioni complesse tra gli elementi che la compongono;
- **Restrizioni**: stabiliscono i vincoli per l'accettazione delle informazioni immesse in input nell'ontologia;
- **Regole**: si tratta di espressioni di tipo if-then che descrivono le logiche per inferire nuove asserzioni;
- **Assiomi**: asserzioni rappresentate in una forma logica che racchiudono tutte le teorie del dominio concettualizzato nell'ontologia;
- **Eventi**: cambiamenti di attributi o relazioni.

3.3 SEMANTIC INTEGRATION

L'implementazione di sistemi di integrazione e matching semantici richiede prima di tutto la realizzazione di tassonomie e la classificazione degli oggetti presenti nel dominio di applicazione. Per fare questo si possono utilizzare principalmente due approcci [49]:

- **Top-down**: utenti esperti del dominio classificano tutti gli oggetti e le entità secondo una rigida tassonomia;

[6] http://en.wikipedia.org/wiki/Ontology_components & http://www.w3.org/TR/owl-guide/

- **Bottom-up**: si cerca di far costruire al computer ontologie e informazioni semantiche in maniera automatica dall'analisi di dati. La difficoltà principale sta nel riuscire a dare al computer la giusta chiave interpretativa per identificare pattern semantici nei dati.

L'obiettivo di un sistema di integrazione semantico è di pulire ed aggregare i dati attraverso l'apprendimento del significato dei singoli elementi e non solo dalla loro sintassi o da pattern ricorrenti. In sistemi semantici il contesto diviene una componente fondamentale, solo considerando il contesto si può riuscire a dare un valore a dei dati apparentemente senza significato. Le tecnologie semantiche si pongono l'obiettivo di applicare strumenti e algoritmi in grado di identificare e comprendere il significato delle cose in maniera automatica, in modo che contesto e significato siano interpretabili e processabili dalle macchine e non solo dalla mente umana. Le tecnologie di integrazione semantica si preoccupano di identificare il significato di una cosa in un determinato contesto e sono in grado di ricordare ciò che hanno imparato. Attraverso l'applicazione di tali tecnologie è possibile superare alcuni dei limiti tradizionali della Data Integration.

In questa parte del capitolo sono esposte le principali tecniche, tecnologie e metodologie che hanno saputo portare delle innovazioni nel campo della Data Integration attraverso l'applicazione di tecnologie semantiche e di architetture service-oriented (SOA), che hanno favorito la migrazione da applicazione a servizio (SaaS), cercando di capire se e come l'applicazione delle tecnologie del Semantic Web può semplificare l'integrazione di dati nelle moderne realtà aziendali.

3.3.1 ENTITY RESOLUTION E SEMANTIC MATCHING

Quando si effettuano operazioni di integrazione di dati è ricorrente trovare informazioni sui medesimi soggetti o oggetti in basi dati differenti. Per poter procedere all'integrazione di tali informazioni è necessario determinare quando due o più rappresentazioni di dati si riferiscono al medesimo soggetto o oggetto che chiameremo entità. Operazioni di questo tipo sono note nel mondo della Data Integration con il termine di "*Entity Resolution*". La risoluzione delle entità è un processo in cui si cerca di applicare algoritmi intelligenti in grado di identificare la stessa entità in basi dati differenti attraverso la comparazione di attributi e di relazioni complesse tra entità. Algoritmi di questo tipo si basano essenzialmente su metodologie statistiche attraverso le quali il programma attribuisce dei punteggi alle entità che indicano la probabilità che si parli della stessa cosa. Starà poi all'utente umano fissare dei vincoli in base ai quali il sistema decide autonomamente se due o più rappresentazioni si riferiscono alla stessa entità, e vanno quindi integrate in una vista unica, oppure sono due entità distinte e vanno integrate come due oggetti separati.

La disciplina dell'Entity Resolution è un'area di ricerca tutto sommato recente, dove si fa ancora confusione sull'utilizzo di determinati termini. John Talburt[7] ha provato a chiarire la situazione elaborando una breve definizione dei termini ricorrenti nel settore dell'Entity Resolution [50]:

- **Entity Resolution**: è un termine generico che indica il processo che va dall'estrazione delle entità dalle fonti di origine, al collegamento delle stesse entità, per arrivare all'esplorazione delle relazioni complesse che intercorrono tra di esse;
- **Entity Identification**: indica un caso particolare di risoluzione di entità, ovvero il caso in cui siamo già in possesso di un database di entità conosciute a cui collegare le entità di una nuova base dati da integrare;
- **Entity Disambiguation**: si riferisce alle operazioni che permettono di determinare che i dati sono relativi a entità diverse;
- **Entity Extraction**: si riferisce a quell'insieme di attività che permettono di estrarre tutte le informazioni relative a una singola entità da un database. L'estrazione di entità può avvenire a partire da dati strutturati, operazione piuttosto semplice, o da dati non strutturati, operazione decisamente più complessa.

La risoluzione delle entità è un processo di fondamentale importanza in operazioni di Data Integration attraverso le quali l'organizzazione cerca di ottenere una vista unica e integrata dei dati sui propri clienti, prodotti, beni, materie prime e altri dati di primaria importanza per il business aziendale (dati sui costi, dati finanziari, ecc.). Spesso infatti informazioni su tali tipologie di entità sono presenti in più database dislocati in parti e applicazioni diverse all'interno della stessa azienda. Riuscire ad ottenere una vista unica e integrata su tali entità è fondamentale per minimizzare errori e costi.

Secondo John Talburt [51] vi è una credenza comune che identifica come coincidenti l'Entity Resolution e le tecniche di Data Matching. Le tecniche di Data Matching sono una parte fondamentale dell'Entity Resolution ma spesso non sono sufficienti. Le tecniche di Data Matching si basano sull'assunzione che quando due record condividono circa le stesse informazioni o attributi rappresentano la stessa entità. Il problema delle tecniche di Data Matching è dato dalla produzione di risultati falsati, in particolare dalla presenza di:

- **Falsi negativi**: record che non vengono abbinati dal sistema ma che in realtà dovrebbero essere uniti;
- **Falsi positivi**: record che vengono uniti dal sistema ma che in realtà non dovrebbero essere abbinati.

[7] Professore di Scienze Informatiche all'Università dell'Arkansas e direttore del Laboratory for Advanced Research in Entity Resolution and Information Quality.

Il problema maggiore è senza dubbio quello dei falsi negativi, infatti, mentre per i falsi positivi è possibile comunque intervenire in un secondo momento definendo nuove regole di matching oppure aggregandoli manualmente, sistemare i falsi negativi è un'operazione decisamente più complicata che comporta la separazione di un record unico in due record distinti con le informazioni corrette. Gli strumenti di Entity Resolution cercano di evitare il problema dei falsi allargando l'orizzonte di analisi ricercando relazioni nascoste tra i dati relativi ad una stessa entità. Un esempio classico è quello di una persona sposata memorizzata in archivio con cognomi e indirizzi diversi (rispettivamente il suo e quello del marito), un sistema di Data Matching tradizionale non sarà in grado di capire che si tratta della stessa persona poiché il confronto sui nomi e sugli indirizzi rende le due entità separate. Al contrario, un sistema di Entity Integration allargando l'orizzonte di analisi su altri attributi (data di nascita, patente di guida, ecc.) o su altri archivi (matrimoni, servizi attivati dalla persona, ecc.) sarà in grado di capire che i due record si riferiscono alla stessa entità. Il problema di un approccio di questo tipo sta nell'effettiva disponibilità di dati aggiuntivi, non sempre infatti è possibile avere accesso a tali informazioni.

Le tecniche di Entity Resolution forniscono gli strumenti per aggregare e collegare i record che si riferiscono alla stessa entità ed sono una risorsa preziosa per le tecniche di Data Integration, in particolare per il Data Warehousing. Tuttavia, le tecniche di Entity Resolution possono operare indipendentemente dai sistemi di Data Integration identificando le entità in una moltitudine di fonti differenti ma senza necessariamente procedere alla loro integrazione. È possibile ad esempio realizzare un motore di ricerca in grado di estrarre le informazioni su una determinata entità creando di fatto una vista unificata ma solo temporanea. La disciplina dell'Entity Resolution costituisce la base di partenza per l'applicazione di tecnologie e metodologie di matching semantico in grado di superare i limiti tipici delle tradizionali tecnologie di pattern matching che si basano soltanto sull'analisi sintattica sintassi dei record. Tuttavia, le tecniche di analisi dei pattern basate sulla sintassi non sono in grado di catturare il significato e il contesto dei dati. Le metodologie semantiche sfruttano le informazioni già presenti nel contesto di business per identificare contesto e significato di dati nuovi, privi di significato se analizzati con tecniche tradizionali. L'implementazione di una semantica dei dati permetterebbe quindi di arginare alcuni problemi di matching delle entità tipici delle tecniche di Data Integration tradizionali andando a costituire un rilevante valore aggiunto per il business attraverso una riduzione dei tempi e delle risorse necessarie per portare a termine progetti di integrazione di dati.

3.3.2 SCHEMA MATCHING

L'integrazione dello schema dei dati è un aspetto molto importante in ogni progetto di Data Integration. L'integrazione degli schemi delle diverse sorgenti costituisce di fatto in primo

passo per realizzare un database integrato [52]. L'integrazione di schemi richiede innanzitutto l'identificazione delle corrispondenze semantiche tra due o più schemi differenti. Questa prima fase è solitamente identificata con il nome di *Schema Matching*. Le corrispondenze identificate nella prima fase vanno quindi utilizzate per sviluppare precise trasformazioni per mappare i dati dalla fonte alla sorgente. Questa seconda fase prende il nome di *Schema Mapping.*

Tradizionalmente la fase di schema matching si focalizza sull'analisi delle istanze dei dati in quanto la documentazione degli schemi è solitamente incompleta o non aggiornata agli ultimi sviluppi. Tuttavia, da una prospettiva esterna all'organizzazione le istanze dei dati sono difficilmente ottenibili, per motivi di sicurezza e sensibilità dei dati. Per questo uno strumento di schema integration dovrebbe utilizzare qualsiasi documento disponibile in grado di dare informazioni sullo schema dei dati.

In [52] viene presentato uno strumento di schema matching (Harmony) che combina diversi algoritmi di matching in un'interfaccia utente unificata che permette di vedere e modificare le corrispondenze rilevate nei dati. Lo strumento si basa su un modello di integrazione a 5 fasi:

1. **Preparazione dello schema**: in questa fase va catturata tutta la conoscenza esistente riguardo lo schema sorgente e lo schema di destinazione in modo da facilitare l'esecuzione delle fasi successive;
2. **Schema Matching**: si tratta di stabilire le corrispondenze a livello semantico tra due o più schemi di dati. La determinazione di tali corrispondenze viene formalizzata attraverso la creazione di link semantici utilizzabili per integrare nuove istanze dei dati;
3. **Schema Mapping**: in questa fase vengono definite le regole, a livello logico, necessarie per trasformare i dati da uno schema all'altro;
4. **Integrazione dei dati**: è la fase in cui le istanze dei dati vengono integrate. È una fase delicata in quanto implica il riconoscimento delle istanze che si riferiscono ad una medesima entità. Inoltre, in questa fase è possibile rimuovere e pulire i valori errati ovvero quei dati che sono in contrasto con i vincoli definiti dallo schema dei dati;
5. **Implementazione del sistema**: si tratta di implementare il sistema che automaticamente integra i dati in base alle configurazioni precedentemente effettuate.

Lo strumento sviluppato (Harmony Integration Workbench [52]) è in grado, attraverso l'analisi delle fonti di dati, di proporre automaticamente i collegamenti tra gli elementi che le compongono. Per fare questo Harmony utilizza degli algoritmi di confronto e analisi dei dati (strutturati e non strutturati) che permettono di attribuire un punteggio di match tra i

vari attributi degli schemi interessati. Il workbench propone quindi una mappa delle relazioni automaticamente identificate, attraverso l'ambiente grafico l'utente può quindi validare le relazioni corrette, eliminare quelle errate o definirne di nuove. Risulta chiaro che l'efficacia di un sistema di matching automatico di questo tipo dipende fortemente dalla ricchezza di rappresentazione del dominio nelle diverse basi dati. Utilizzando come fonti ontologie o schemi di dati ricchi di metadati che descrivono a nei dettagli relazioni e significato degli attributi si otterrà sicuramente un risultato migliore rispetto a quanto si otterrebbe partendo da schemi di dati poveri come può essere una semplice tabella excel, un file XML o un database relazionale. L'utilità di uno strumento di integrazione automatica di schemi di dati viene riassunta dagli autori del workbench in seguito ad una fase di prova del prodotto in collaborazione con il dipartimento della difesa americano. Vengono identificati alcuni scenari in cui uno strumento di Schema Matching può risultare utile [53]:

- *Valutare la praticabilità di un progetto*: uno strumento di schema matching può essere utile per valutare l'ammontare di lavoro necessario per integrare due o più sistemi. Attraverso un'integrazione automatica degli schemi è possibile identificare la numerosità e la complessità dei collegamenti. Ciò rende possibile la stima del lavoro manuale da compiere e dei costi di un progetto di integrazione;
- *Identificare gli oggetti da integrazione*: uno strumento di schema matching è utile per identificare le entità e i relativi attributi rappresentati nei database e permette di definire velocemente le informazioni in comune tra più schemi di dati;
- *Creare uno schema di scambio*: in seguito all'identificazione di entità e attributi è possibile utilizzare lo strumento di schema matching per realizzare uno schema di collegamento per poter trasportare o scambiare dati da un sistema all'altro;
- *Sintetizzare l'ammontare di conoscenza posseduto dall'impresa*: la realizzazione di uno schema di collegamento tra le basi dati aziendali permette di codificare un vocabolario della conoscenza posseduta, dove sono memorizzate informazioni su entità, oggetti e collegamenti tra basi dati differenti;
- *Identificare relazioni nascoste*: in imprese con notevoli quantità di dati spesso non ci si rende conto di cosa effettivamente contengono tutte le fonti informative. Attraverso un'analisi automatica dei collegamenti è possibile identificare relazioni nascoste o non codificate, di cui l'azienda ignorava l'esistenza.

3.3.3 UTILIZZO DI ONTOLOGIE

Attraverso l'implementazione di un'ontologia è possibile descrivere un determinato contesto attraverso l'esplicitazione di caratteristiche, relazioni e vincoli che intercorrono tra gli elementi che contraddistinguono lo specifico dominio. L'implementazione del

modello a grafo permette una rappresentazione della realtà molto più ricca e completa di quanto si possa ottenere con il modello relazionale. In questo paragrafo vediamo alcune possibili applicazioni di ontologie a supporto dell'integrazione di dati e informazioni.

3.3.3.1 DA APPLICAZIONE A SERVIZIO

Uno dei maggiori problemi dei sistemi di Data Integration tradizionali è costituito dal fatto che ci si trova ad operare con software di tipologie differenti, più o meno complessi, realizzati nel corso degli ultimi decenni. Un contesto sicuramente non ottimale nell'ottica di un'integrazione efficace e veloce dei dati prodotti da tali applicativi. Lo scenario ideale per soddisfare a pieno le necessità di integrazione è rappresentato da un insieme di servizi, operanti via Web, che comunicano attraverso un protocollo standard [43].

Tuttavia, uno scenario di questo tipo comporta una serie di ostacoli. Innanzitutto è necessario chiedersi come, quali e quante delle applicazioni attuali trasformare in Web Service. Inoltre, i problemi di integrazione non sarebbero comunque completamente risolti. Soltanto utilizzando un'ontologia in grado di descrivere nei minimi dettagli dati, relazioni e applicazioni si può pensare che uno scenario di questo tipo costituisca un vero valore aggiunto per l'integrazione dei dati. Un'ontologia supporta infatti delle caratteristiche che non sono riscontrabili negli approcci classici fino ad oggi utilizzati [43]:

- Un'ontologia può descrivere allo stesso tempo dati, metadati, schemi, applicazioni e interfacce dettagliando a livello semantico relazioni e vincoli tra tali oggetti;
- La rappresentazione di dati e schema dei dati permette di rappresentare i dati in un formato facilmente comprensibile dalle macchine che ben si presta a operazioni di trasmissione, trasformazione, interrogazione e memorizzazione dei dati;
- Un'ontologia può descrivere nei dettagli il comportamento dei servizi e di tutte le applicazioni collegate in una maniera non possibile con un modello relazionale per la mancanza di una logica applicata al sistema;
- Un'ontologia non si appoggia a indici e chiavi per descrivere i propri elementi. In caso di cambiamenti del contesto l'ontologia è molto più flessibile rispetto a un sistema relazionale;
- Un'ontologia è progettata per essere altamente scalabile e per condividere facilmente le informazioni che contiene.

3.3.3.2 ONTOLOGY-BASED DATA INTEGRATION

Nella moderna società dell'informazione la domanda per l'accesso a grandi quantità di informazioni è in continua crescita, tuttavia, tali informazioni sono spesso eterogeneamente distribuite. Per garantire un'efficiente piattaforma di condivisione delle informazioni è necessario superare alcuni problemi. Il problema principale nell'aggregare fonti di dati eterogenee e distribuite è identificabile come un problema di interoperabilità. Spesso le

fonti dati che si vogliono integrare presentano strutture e schemi di dati differenti e non è possibile interrogarle con un linguaggio di query universale. Il problema dell'interoperabilità si può ricondurre alla presenza di due livelli di eterogeneità [54]:

- **Eterogeneità strutturale**: sistemi diversi memorizzano i dati secondo schemi e strutture differenti;
- **Eterogeneità semantica**: si riferisce al contenuto e al significato dei dati contenuti in un database.

Per raggiungere l'interoperabilità semantica è necessario esplicitare il significato dell'informazione attraverso tutti i sistemi identificati. In particolare l'eterogeneità semantica si può ricondurre a tre cause principali [54]:

- **Conflitti di significato**: quando i dati sembrano avere lo stesso significato ma in realtà differiscono, ad esempio perché relativi a istanti temporali diversi;
- **Conflitti di misurazione**: quando sistemi diversi misurano la stessa cosa con unità di misura differenti;
- **Conflitti di denominazione**: quando le denominazioni degli schemi dei dati differiscono significativamente. Un problema molto frequente è dato dalla presenza di sinonimi ed omonimi.

L'utilizzo di ontologie per l'esplicitazione della conoscenza implicita e nascosta è senza dubbio un approccio percorribile per superare il problema dell'interoperabilità semantica. È possibile individuare diversi approcci per l'utilizzo di ontologie per l'esplicitazione del contesto (Figura 27):

a) *Realizzazione di un'ontologia singola*: si tratta di implementare un vocabolario semantico condiviso tra tutte le basi di dati individuate. In questo modo è possibile recuperare dati da tutte le fonti semplicemente interrogando con un unico linguaggio l'ontologia;
b) *Utilizzo di più ontologie*: ogni base di dati è descritta da un'ontologia. Le singole ontologie dovrebbero essere sviluppate a partire da un vocabolario comune in maniera da garantire l'interoperabilità semantica all'interno del sistema. Il vantaggio di un approccio di questo tipo sta nel fatto che la realizzazione di un'ontologia unica globale è spesso un'operazione dispendiosa e complicata. Inoltre, tale approccio garantisce una maggiore espandibilità e flessibilità del sistema. Tuttavia, la mancanza di un vocabolario semantico esplicitamente codificato potrebbe, a lungo andare, portare a problemi di incompatibilità e difficoltà nello scambio di informazioni tra le diverse basi di dati;
c) *Approccio ibrido*: colma la lacuna dell'approccio a ontologie multiple prevedendo la codificazione esplicita di un vocabolario comune in grado di collegare tutte le

ontologie e di garantire nel tempo la piena interoperabilità del sistema. Nei sistemi più complessi il vocabolario condiviso può essere un'ontologia vera e propria.

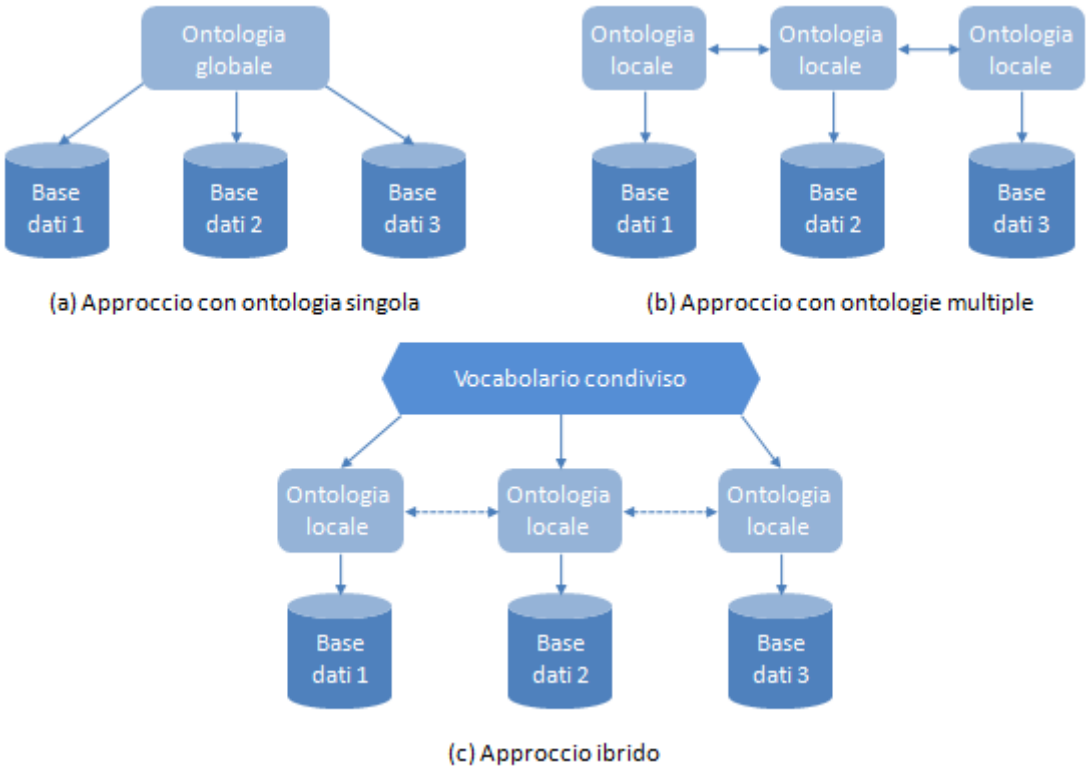

Figura 27 - Tre possibili approcci per l'utilizzo di ontologie per l'esplicitazione del contesto. (Fonte: [54]. Rielaborazione dell'autore).

La realizzazione di un sistema di integrazione di questo tipo permette di creare un livello concettuale, contenente il significato dei dati, separato dal livello di memorizzazione dei dati stessi. Le applicazioni accederanno soltanto al livello concettuale e il livello dei dati rimarrà nascosto all'utente. L'implementazione di ontologie permette quindi di raggiungere l'interoperabilità semantica attraverso la creazione di collegamenti tra le entità contenute nelle diverse basi dati.

L'implementazione successiva consiste nell'esporre via web-service le URI dei dati mappati nell'ontologia (attraverso l'utilizzo del linguaggio RDF) in modo da rendere possibile l'interrogazione dei dati tramite il linguaggio di query standard per il Semantic Web: SPARQL. La realizzazione di un sistema di rappresentazione e interrogazione dei dati di questo tipo permette lo sviluppo di applicazioni Web per la consultazione, visualizzazione, analisi ed elaborazione dei dati per una moltitudine di obiettivi differenti (servizi per clienti, fornitori, manager, analisti, ecc.). Maggiore sarà la quantità di dati esposti e collegati maggiori saranno i benefici derivanti dall'effetto rete che andranno a tradursi in notevole valore aggiunto informativo per l'azienda [46].

3.3.3.3 ONTOLOGY MATCHING

Come abbiamo visto nei paragrafi precedenti l'utilizzo di ontologie è visto come la soluzione per una moltitudine di applicazioni (integrazione dei dati, fornitura di servizi via Web, ecc.). tuttavia, in un sistema aperto come è il Web si potrebbe generare un grande numero di ontologie riproponendo di fatto un problema di interoperabilità. In pratica, l'eterogeneità del sistema non risulterebbe ridotta ma soltanto spostata a un livello superiore.

Come affermano Euzenat e Shvaiko [55], per porre fine al problema dell'eterogeneità semantica e garantire la piena interoperabilità del sistema è quindi necessario un sistema in grado di collegare le ontologie esistenti. La disciplina dell'Ontology Matching mira proprio a sviluppare sistemi automatici in grado di trovare le corrispondenze tra entità semanticamente rappresentate in ontologie diverse. Le corrispondenze tra ontologie potranno essere utilizzate per attività di vario tipo: unione di ontologie, interrogazione di grandi quantità di dati, trasformazione di dati, ecc.

Tuttavia, la ricerca su metodologie di Ontology Matching è ancora agli albori e saranno necessari dai 5 ai 10 anni per vedere tali tecnologie affermarsi sul mercato e quindi garantire la piena interoperabilità tra sistemi basati su ontologie [55]. Alcuni sistemi di Ontology Matching sono comunque già stati realizzati ma, al di la della qualità del risultato, hanno evidenziato un problema di prestazioni che diviene molto rilevante visto che i sistemi basati su ontologie sono pensati per operare in contesti molto dinamici che non possono tollerare lunghi tempi di risposta. Per permette l'adozione di massa di sistemi basati su ontologie è quindi necessario trovare in tempi brevi una soluzione efficace alla problematica dell'interoperabilità tra ontologie.

3.3.4 UN'ARCHITETTURA SEMANTICA PER L'IMPRESA

L'applicazione in ambito aziendale delle tecnologie semantiche permette di ridefinire l'architettura informatica dell'azienda mettendo al centro del sistema una piattaforma di gestione della conoscenza in grado di inferire automaticamente nuove informazioni e di fungere come base per le interrogazioni da pare degli utenti o da altre applicazioni aziendali.

In Figura 28 vediamo un esempio di quella che potrebbe essere un'architettura di questo tipo. Alla base del sistema sono collocate le fonti di dati, siano essi dati strutturati (database relazionali RDBMS o dati in formato XML) o dati non strutturati (documenti word, e-mail, fogli di calcolo, annotazioni, ecc.). Tutti i dati in possesso dell'azienda vengono mappati in formato semantico (RDF) e descritti tramite un'ontologia di dominio. Aggiungendo a questo linguaggi di inferenza, logica, regole e un protocollo di

interrogazione (SPARQL) si ottiene la base di conoscenza che fungerà da perno per tutte le applicazioni aziendali.

Un'architettura di questo tipo garantirebbe la risoluzione dei tipici problemi di integrazione prestandosi inoltre ad essere la base per lo sviluppo di tutta una serie di applicazioni che possono essere [46]:

- *Ricerca semantica di informazioni*: essendo tutti i dati mappati semanticamente la realizzazione di un sistema di ricerca semantico garantisce l'opportunità di estrapolare tutte le informazioni su una determinata entità in modo pratico e veloce, senza ricorrere a sofisticate tecniche di integrazione e matching;
- *Business Intelligence avanzata*: si possono sfruttare la logica e le regole esplicitate nel modello semantico per inferire con relativa facilità nuove informazioni e pattern nascosti nei dati. Un'architettura di questo tipo facilita le attività di analisi dei dati e permette all'azienda di ottenere un vantaggio competitivo nel business;
- *Realizzazione di servizi web-oriented*: grazie alla base dati semantica è possibile esporre con grande facilità dati sul Web per realizzare applicazioni per clienti, fornitori, ecc.

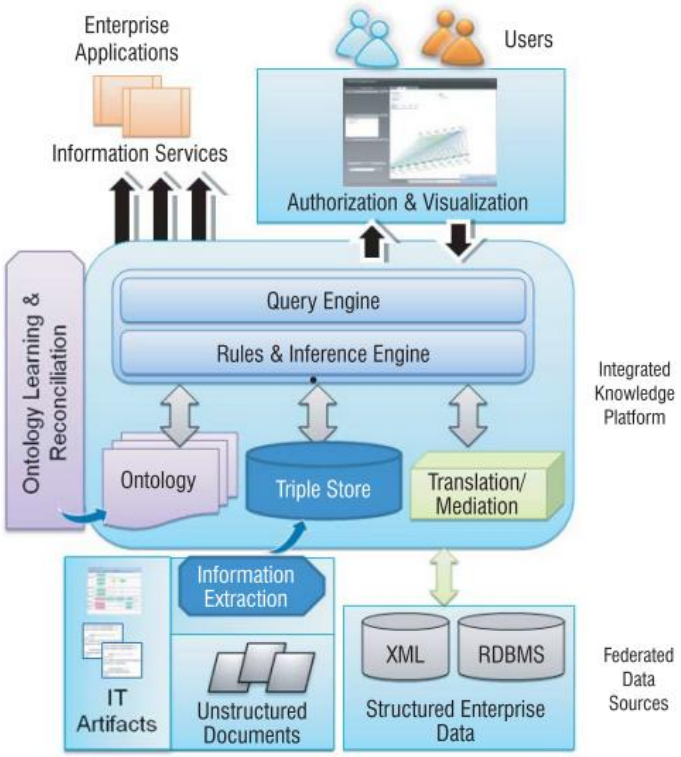

Figura 28 – Un esempio di architettura semantica per l'impresa (Fonte: [46]).

3.4 SEMANTIC DATA INTEGRATION SYSTEM

Come si è visto nel capitolo precedente l'approccio tradizionale al problema dell'integrazione dei dati ha portato alla realizzazione di strumenti complessi e costosi la cui implementazione è un'attività che spesso richiede lunghi tempi di operatività. La realizzazione di un modello di impresa semantico ha l'obiettivo di garantire una via più facile e meno costosa rispetto all'approccio tradizionale di Data Integration. Tuttavia, la transizione verso un modello semantico di impresa richiede significativi passi avanti nella gestione dei metadati e nella realizzazione di un modello strutturato del contesto di business aziendale [56]. Con l'affermazione delle tecnologie del Semantic Web sono recentemente apparsi sul mercato i primi software con pieno supporto alla Semantic Data Integration: Expressor Integrator, Progess Software, DataLens Sytem. Per quanto riguarda le aziende leader del mercato soltanto Oracle ha dimostrato di investire significativamente sulle tecnologie semantiche, introducendo, nell'ultima versione del proprio DBMS (Oracle 11g), il pieno supporto ai linguaggi RDF e OWL [57]. Attraverso tale implementazione il DBMS di Oracle è oggi in grado di supportare nuove funzionalità, distinguendosi dagli altri prodotti tradizionali:

- Memorizzazione, caricamento e trasformazione di dati in formato RDF, OWL e di ontologie;
- Inferenza automatica sfruttando le proprietà di OWL, la semantica di RDFS e le regole definite dagli utenti tramite appositi linguaggi;
- Interrogazione e mappatura di dati in formato relazionale sfruttando le ontologie;
- Interrogazione di dati RDF, OWL e di ontologie attraverso l'utilizzo di schemi SPARQL incorporati in SQL.

3.4.1 EXPRESSOR INTEGRATOR

Expressor Software è una società statunitense, con sede a Burlington, Massachussets, fondata nel 2003 da esperti di Data Integration e Data Warehousing con l'obiettivo di affermare metodologie innovative di integrazione dei dati, in particolare tramite l'implementazione di tecnologie semantiche. Il prodotto, expressor semantic data integration system, è stato annunciato nel maggio 2008 e reso disponibile sul mercato (versione 1.0) nel giugno dello stesso anno. La versione successiva (1.2) venne lanciata nel novembre 2008 mentre nel corso del 2009 il prodotto raggiunge una discreta maturità con il lancio della versione 2.0 rilasciata nel mese di maggio. Nel frattempo l'azienda ha saputo mettersi in mostra sul mercato risultando tra le migliori quattro aziende innovatrici nel settore della Data Integration nell'annuale report di Gartner [58].

Secondo i fondatori di expressor il modello di prezzi sul quale si basano i tradizionali prodotti di Data Integration (licenze legate al numero di macchine su cui viene installato il prodotto) è oggi obsoleto e superato [59]. È necessario semplificare tale modello offrendo una soluzione caratterizzata da un design semplificato, che abiliti il riuso attraverso un sistema online di gestione condivisa dei metadati. Il prodotto expressor si basa infatti su un modello dei metadati che razionalizza i termini di business, applicando specifiche trasformazioni di dati e regole di business che rendono possibile il riutilizzo del modello tra imprese differenti, riducendo sostanzialmente i tempi di sviluppo e implementazione della soluzione. Il prodotto offre quindi un ambiente grafico di gestione dei metadati condiviso attraverso il quale è possibile sviluppare, con un approccio collaborativo, specifiche soluzioni di integrazione. La base del sistema è costituita da un potente motore di elaborazione di dati che supporta operazioni sia di tipo batch che in tempo reale ed è in grado di processare terabytes di dati all'ora. Il prodotto di expressor garantisce alle aziende i seguenti vantaggi [59]:

- **Semantica intelligente**: il sofisticato sistema di gestione dei metadati permette di costruire modelli semantici flessibili e riutilizzabili;
- **Gestione del ciclo di vita**: la presenza di un deposito centralizzato dei metadati offre la possibilità di monitorare e gestire lo sviluppo del progetto in ogni fase del suo ciclo vitale;
- **Riduzione dei costi**: la condivisione del deposito dei metadati rende riutilizzabili efficacemente e velocemente trasformazioni sui dati, regole e modelli di business, contribuendo a ridurre sostanzialmente i costi totali di possesso del prodotto;
- **Scalabilità**: il potente motore di elaborazione dati permette di gestire senza problemi di performance progetti dalle più disparate necessità indipendentemente dal tipo di elaborazione richiesta (batch o real time). Il sistema è inoltre in grado di connettersi automaticamente ad una grande quantità di strutture dati differenti garantendo un'ottima interoperabilità tra i sistemi esistenti.

Nella sua ultima versione (2.0) il prodotto è costituito da tre elementi principali [60]:

1. **Integrator**: è lo strumento che permette di sviluppare un determinato progetto di integrazione (ETL, sincronizzazione o migrazione di dati, ecc.). È a sua volta costituito da una serie di strumenti con finalità differenti (connessione ai dati di origine, sviluppo dello schema integrato, trasformazioni di dati, ecc.);
2. **Repositor**: è la componente di gestione dei metadati che abilita il riuso dei progetti. Include un dizionario semantico, regole di business, dettagli di implementazione, ecc.;

3. **Processor**: è il motore di elaborazione dati. Si tratta di un componente complesso e altamente performante, ottimizzato per elaborare più processi contemporaneamente e in parallelo su più hardware.

In conclusione, expressor ha saputo introdurre sapientemente sul mercato una tecnologia con ottime performance a costi ridotti e potenzialmente migliore delle tradizionali. Tuttavia, il successo sul mercato di un prodotto potenzialmente dirompente non è legato solamente all'introduzione di una tecnologia superiore a un prezzo ridotto ma è anche questione di marketing. Expressor ha inizialmente sviluppato una strategia commerciale che le permettesse di emergere fra le imprese innovatrici senza pensare di entrare in competizione con colossi della Data Integration quali IBM e Informatica. Inoltre, sta cercando di espandere gradualmente il suo mercato al di fuori degli Stati Uniti (principalmente in Europa e in particolare nel Regno Unito). Recenti accordi di partnership con aziende consolidate sul mercato della Data Warehousing (ad esempio con Netezza[8]) hanno permesso ad expressor di combinare il proprio sistema semantico con un potente sistema di gestione di Data Warehouse, confermando l'intenzione dell'azienda di competere, in un prossimo futuro, con i tradizionali prodotti di Data Warehousing da anni fortemente consolidati sul mercato [61].

3.4.2 ALTRI PRODOTTI

L'affermarsi sul mercato di tecnologie semantiche ha spinto anche altre aziende a lanciarsi nel settore della Semantic Data Integration. Oltre ad expressor software altre due soluzioni degne di nota si sono recentemente affacciate sul mercato:

- **DataLens System**: è un prodotto sviluppato da Silver Creek Systems con l'obiettivo di superare i limiti prestazionali e di costo delle soluzioni tradizionali. Attraverso l'implementazione di tecnologie semantiche la società offre una soluzione facile da sviluppare e gestire, riusabile, con supporto a web service e standard di integrazione SOA. Il prodotto offre funzionalità di Data Integration, Data Quality, Data Governance e Content Enrichment. DataLens System utilizza una tecnologia proprietaria in grado di riconoscere, analizzare e gestire pattern semantici all'interno dei dati aziendali, garantendo una ricostruzione automatica del modello di business con supporto di funzionalità avanzate di gestione della qualità del dato. L'azienda vanta partnership con diverse società, alcune delle quali di grande rilievo (Oracle per la parte di Data Quality, Data Cleaning e Data Matching, IBM per la standardizzazione e la gestione di formati dati complessi);

[8] Netezza offre una tecnologia, che include un DBMS e un hardware dedicato, ottimizzata per Data Warehousing e Business Intelligence.

- **Progress DataXtend Semantic Integrator (SI)**: si tratta di un prodotto sviluppato da Progress Software con l'obiettivo di sfruttare le tecnologie semantiche per arricchire il modello di business aziendale attraverso l'inserimento di una semantica del dato. Il prodotto garantisce lo sviluppo di un modello dei dati semantico che abilita l'integrazione automatica da fonti differenti, con particolare attenzione alla qualità e all'integrità dei dati aggregati. Il prodotto offre funzionalità di trasformazione, aggregazione e mappatura di dati e schemi, presenta funzionalità avanzate per la validazione dei dati per assicurarne la consistenza, offre la possibilità di descrivere contesti di business attraverso l'esplicitazione delle regole che li governano con pieno supporto a contesti SOA.

3.5 CONCLUSIONI

Negli ultimi anni si è assistito a un cambiamento del modello industriale soprattutto in riferimento all'architettura informatica delle aziende. Il nuovo contesto di business pone l'enfasi sulla disponibilità in tempo reale delle informazioni e sulla disponibilità di dati riassuntivi, report e analisi sui dati. Per rispondere a tali bisogni è necessario disporre di metodologie di integrazione dei dati semplici ed efficienti. Per questo motivo per le tecnologie del Semantic Web c'è oggi la possibilità affermarsi sul mercato al fine di promuovere un modello informativo per l'impresa più efficace, più ricco e dettagliato che permetta di rispondere in maniera adeguata alle richieste del business attuale.

In particolare, le potenzialità del modello semantico sono le seguenti [46]:

- *Modello dei dati flessibile*: il modello a grafo permette di superare i limiti tipici del modello relazionale e del modello ad albero, garantendo la possibilità di realizzare schemi di dati complessi ma allo stesso tempo flessibili ed espandibili con facilità;
- *Ragionamento automatico*: le tecnologie semantiche incorporano i costrutti della logica rendendo possibili inferenze e ragionamento automatico sui dati di partenza;
- *Integrazione di informazioni facilitata*: grazie alla descrizione del contesto, del contenuto e del significato dei dati è possibile integrare in maniera più semplice, efficace e veloce diversi domini informativi sfruttando le relazioni implicite ed esplicite descritte nel modello dei dati. L'implementazione di ontologie facilita ulteriormente operazioni di questo tipo;
- *Interrogazione dei dati*: il modello semantico prevede un linguaggio appositamente studiato per interrogare ontologie e dati rappresentati con il modello a grafo. L'esplicitazione delle caratteristiche del dominio applicativo

permette la realizzazione di query più profonde in grado di sfruttare i concetti e le relazioni dichiarate nel modello semantico. Il linguaggio di query si focalizza sul livello concettuale (la descrizione del dominio, di entità e relazioni) anziché su quello fisico (come e dove sono memorizzati i dati).

La tabella sottostante riporta un confronto sintetico tra le caratteristiche dei sistemi di Data Integration tradizionale e dei sistemi di integrazione semantici.

	Data Integration tradizionale	**Data Integration semantica**
Struttura dei dati	Per lo più relazionale: rigida e focalizzata sui dati	Modello a grafo: flessibile, espandibile e focalizzato sulle relazioni tra i dati
Metodologia di integrazione	ETL: estrazione dei dati dalla fonte di origine, trasformazione e standardizzazione, caricamento in ambiente omogeneo	Collegamenti tra le fonti di dati utilizzando definizioni e relazioni sui dati condivise in un'ontologia di dominio
Scalabilità del sistema	Ogni fonte di dati aggiuntiva aumenta i costi esponenzialmente	Nuove fonti di dati integrabili facilmente e con costi contenuti
Ricchezza contestuale	Limitata dai costi e dal carico di lavoro e manutenzione dello staff di progetto	Benefici dall'effetto di rete: l'aggiunta di nuovi dati migliora la rappresentazione del contesto
Fonti informative	Prevalentemente interne	Interne ed esterne, grande facilità nell'esporre i dati sul Web e di realizzare servizi
Coinvolgimento delle unità di business	Personale che richiede report sui dati	Manager e progettisti delle ontologie e attività di data linking esterne
Metodologia di standardizzazione	Standard unico, senza eccezioni, possibilità di perdita di potenziale informativo	Permette di standardizzare i dati e le informazioni contestuali senza alcuna perdita di informazioni

Tabella 3 – Data Integration tradizionale vs. Data Integration semantica (Fonte: [46]).

Tuttavia, nonostante i diversi benefici derivanti dall'implementazione di tecnologie semantiche restano ancora alcune problematiche che frenano lo sviluppo di soluzioni basate su modelli semantici all'interno delle imprese [46]:

- C'è ancora oggi un certo disallineamento tra quello che le tecnologie semantiche promettono di fare e quello che nella pratica riescono a supportare. In particolare, non è ancora del tutto chiaro quale linguaggio della logica verrà utilizzato e come verrà implementato il livello delle regole. Inoltre, i linguaggi ontologici sono nati per operare in un ambiente aperto con caratteristiche talvolta profondamente differenti rispetto a un ambiente controllato e chiuso presente in determinati ambiti

aziendali. L'affidabilità e l'utilità di modelli semantici in ambiti chiusi e controllati non è ancora del tutto chiara e comprovata;

- L'efficacia delle soluzioni semantiche dipende largamente dalla qualità delle ontologie: migliore è l'esplicitazione delle caratteristiche di un determinato contesto in un ontologia e più evidenti sono i vantaggi di un sistema semantico. Tuttavia, la traduzione di interi domini applicativi, contraddistinti da esperienza e conoscenza, in un linguaggio interpretabile dalle macchine è un'attività del tutto manuale non sempre facile e a prova di errore;
- La creazione di ontologie può rappresentare una barriera all'entrata a causa dei costi di apprendimento di linguaggi ontologici. Nonostante non sia richiesta una profonda conoscenza del modello a grafo e dei costrutti logici implementati nei linguaggi ontologici una certa esperienza minima in materia è necessaria. Se le grandi industrie non percepiscono a pieno il valore aggiunto del modello semantico difficilmente investiranno nella formazione di risorse in grado di sviluppare soluzioni basate su tali tecnologie;
- Garantire l'interoperabilità tra ontologie diverse non è un compito facile e, come abbiamo visto nel paragrafo 3.3.3.3, non è ancora chiara la strada da intraprendere per risolvere il problema dell'eterogeneità semantica. Solo risolvendo tale problematica è possibile garantire la completa interoperabilità tra modelli semantici differenti;
- L'interazione con queste nuove basi di conoscenza richiede lo sviluppo di migliori tecnologie di navigazione e visualizzazione del modello a grafo sottostante. La grandezza e la complessità del modello dei dati semantico richiedono lo sviluppo di tecnologie per la visualizzazione delle relazioni tra i dati a determinati livelli di granularità.

Nonostante la presenza di alcune problematiche le tecnologie semantiche sono oggi pronte per uscire dalla fase di ricerca e per essere adottate dal mercato. Il W3C stesso offre una lista di casi di studio, casi d'uso fornendo le linee guida per l'implementazione pratica delle tecnologie del Semantic Web[9].

[9] http://www.w3.org/2001/sw/sweo/public/UseCases/

SEZIONE II

IL CASO DI TRENTINO RISCOSSIONI S.P.A.

CAPITOLO 4

IL CONTESTO DI RIFERIMENTO

In questo capitolo è descritto il contesto di riferimento del caso di studio presentato, progetto al quale ho partecipato attivamente nel corso dello stage svolto presso Informatica Trentina S.p.A. nel periodo compreso tra marzo e agosto 2009. Il capitolo si apre con una descrizione dell'azienda Informatica Trentina S.p.A., società di sistema della Provincia Autonoma di Trento con la funzione di fornire soluzioni nel campo ICT per il settore pubblico provinciale. In particolare, lo stage è stato effettuato presso il settore ricerca e innovazione, denominato Trentino as a Lab, ed è pertanto presente una breve descrizione del ruolo e delle funzionalità di questa specifica area dell'azienda.

Il capitolo prosegue con l'esposizione del contesto del caso di studio, presentando la società di sistema per la riscossione e la gestione delle entrate, Trentino Riscossioni S.p.A., descrivendone gli obiettivi e le funzioni che svolge per gli Enti e per i cittadini. Più nello specifico vengono analizzate caratteristiche e problematiche del sistema informativo a supporto delle attività dell'ente, evidenziando quali sono gli sviluppi necessari per migliorare le attività lavorative, con un breve accenno al piano industriale della società nel breve e nel medio/lungo periodo. Infine, è descritto il contesto tributario di riferimento, con particolare attenzione alle attività di accertamento che costituiscono l'oggetto del progetto di integrazione sperimentale descritto nel capitolo 5.

4.1 INFORMATICA TRENTINA S.P.A.

Informatica Trentina è stata costituita nel 1983 su iniziativa della Provincia Autonoma di Trento e di altri Enti del Trentino, ai quali si aggiunse Finsiel S.p.A., ed ha iniziato la propria attività nel novembre 1984. Il 21 novembre 2002 Finsiel ha trasferito la propria quota azionaria, pari al 40,41%, a DeDa S.r.l., società controllata da Dedagroup S.p.A.. Il 29 dicembre 2006 la quota di DeDa è stata venduta a Tecnofin Trentina S.p.A., trasformando così Informatica Trentina in una società propriamente pubblica, vale a dire una società "in house"[1]. L'operazione si è resa necessaria per adeguare l'assetto societario

[1] Si ha gestione in house quando le pubbliche amministrazioni realizzano le loro attività di competenza attraverso i propri organismi, senza ricorrere al mercato per procurarsi (mediante appalti) i servizi e le forniture occorrenti o per erogare alla collettività prestazioni di pubblico servizio. Gli organismi in house possono essere dotati di una propria personalità giuridica, distinta da quella dell'amministrazione di appartenenza.

ai vincoli previsti dalla vigente normativa comunitaria e nazionale per l'affidamento dei servizi da parte delle pubbliche amministrazioni in favore delle loro società strumentali. Nata con l'obiettivo di progettare, realizzare e gestire il Sistema Informativo Elettronico della Provincia Autonoma di Trento[2], oggi Informatica Trentina fornisce servizi di consulenza, progettazione, sviluppo e gestione di sistemi informativi e reti telematiche per la Pubblica Amministrazione Locale. Nel 2008 ha realizzato un fatturato di oltre 50 milioni di Euro, con un organico di 269 dipendenti.

4.1.1 MISSION E SERVIZI

Informatica Trentina fornisce soluzioni nel campo dell'Information and Communication Technology ed è il punto di riferimento delle Amministrazioni e degli Enti del territorio per lo sviluppo dell'ICT, in qualità di:

- **Strumento di sistema**: in quanto opera per contribuire allo sviluppo del sistema pubblico trentino;
- **Strumento di innovazione**: in quanto promuove l'innovazione nella Pubblica Amministrazione Locale;
- **Strumento di collaborazione**: in quanto coopera con le imprese ICT del territorio per l'ammodernamento della Pubblica Amministrazione.

Informatica Trentina offre agli Enti i seguenti servizi:

Customer Service Desk

Il Customer Service Desk rappresenta il "singolo punto di contatto" per tutte le richieste di supporto ed intervento degli utenti connesse alla fruizione dei servizi di assistenza, attraverso il coordinamento delle attività concorrenti alla soluzione del problema, integrando i processi attivati dall'utente con i servizi erogati dalla Società. Informatica Trentina gestisce complessivamente circa 165.000 contatti all'anno, riferiti a richieste di supporto ed assistenza effettuate dall'utenza (Provincia Autonoma di Trento, Azienda Provinciale per i Servizi Sanitari, Comuni, Altri Enti) di cui 150.000 in entrata e 15.000 in uscita. Il supporto applicativo all'utenza garantisce un efficace utilizzo dei sistemi e delle applicazioni informatiche, ivi compresi gli strumenti di automazione d'ufficio, ed è rivolto a tutti gli utenti delle applicazioni il cui esercizio è in carico alla Società, con l'obiettivo di agevolare i medesimi nella fruizione delle applicazioni. Sono stati gestiti 26.000 ticket nel 2008. Il DTM Fleet Management, integrando il servizio di Desktop Management (DTM), ovvero la gestione delle apparecchiature informatiche relative alle postazioni di lavoro installate presso gli utenti, con il servizio di Noleggio operativo, per la fornitura di un posto di lavoro (personal computer e periferiche accessorie), permette al

[2] Legge provinciale 6 maggio 1980, n. 10.

cliente di sgravarsi totalmente dalla gestione degli asset. Sono serviti in circa 11.260 posti di lavoro di cui: 6.170 PC della Provincia (5.080 sedi provinciali, 540 Scuole obbligo, 550 Biblioteche), 4.960 PC dell'Azienda Provinciale per i Servizi Sanitari, 130 PC di altri Enti.

eProcurement

L'eProcurement, inteso quale insieme di strategie d'acquisto abbinate a strumenti tecnologici, aspetti procedurali e organizzativi per l'acquisizione di beni e servizi on-line, sfruttando le possibilità offerte da Internet, contribuisce alla razionalizzazione della spesa di beni e servizi degli Enti Pubblici della Provincia Autonoma di Trento, offrendo ai fornitori locali di piccole e medie dimensioni l'accesso al mercato della Pubblica Amministrazione, mettendo a disposizione canali digitali di vendita con un possibile ampliamento della base clienti. Nel corso del 2008 sono state realizzate 134 gare telematiche, 5 gare soprasoglia comunitaria per un valore complessivo di spesa affrontata di circa 35.300.000 di Euro ed un risparmio di circa 5.900.000 di Euro, pari a circa il 16,7% della spesa affrontata. Sono stati attivati, inoltre, 2 negozi elettronici: "fotocopiatori" e "apparati radio terminali" che vanno ad aggiungersi ai negozi attivati nel 2007 (carburanti per autotrazione mediante fuel card, sale ad uso stradale, carburanti per autotrazione con consegna).

Data Center

Il Data Center offre, accanto ai tradizionali servizi centralizzati, sistemi e servizi per sofisticate e sicure applicazioni del mondo WEB. Sono presenti nel Data Center di Informatica Trentina 250 server a supporto di soluzioni gestionali e applicazioni evolute del mondo web.

Formazione

I servizi di Formazione, la società elabora percorsi formativi su misura che hanno l'obiettivo di trasferire ai partecipanti le conoscenze necessarie, teoriche e pratiche, per diventare autosufficienti nell'utilizzo di prodotti informatici. Nel triennio 2006-2008 Informatica Trentina ha erogato oltre 15.000 giornate a persona di formazione, sia sui programmi sviluppati dalla Società, che su prodotti di automazione d'ufficio. Il 2008 ha visto l'erogazione di oltre 4.000 giornate/persona di formazione, con l'impegno di circa 350 giorni/docente.

Telecomunicazioni

Il servizio di Telecomunicazione è finalizzato a garantire l'utilizzo dell'infrastruttura della rete telematica provinciale come supporto di base per consentire lo scambio di informazioni elettroniche. La TELPAT, moderna rete per le telecomunicazioni, potente, sicura e diffusa capillarmente sul territorio provinciale, garantisce, tramite collegamenti in larga banda su fibra ottica e wireless, un efficiente punto di accesso ad Internet e ad

altri servizi telematici per gli utenti del territorio provinciale. Sono collegati in rete tramite TELPAT oltre 970 uffici degli Enti del territorio (di cui 150 sedi PAT, 260 sedi APSS e dei medici di base, 108 Scuole, 139 Biblioteche, 11 sedi dell'Agenzia del Lavoro, 3 APT di ambito, 13 Comunità di Valle, 223 Comuni, 24 sedi del Catasto e Libro Fondiario, 32 sedi dell'Università di Trento, 4 sedi dell'Opera Universitaria, 2 sedi della Regione Autonoma Trentino-Alto Adige, l'Istituto Agrario di San Michele, la Fondazione Bruno Kessler) e oltre 100 siti web.

4.1.2 TRENTINO AS A LAB

Trentino as a Lab (TasLab) è una rete dedicata all'innovazione, il cui obiettivo è di sviluppare un approccio che pone l'utente al centro del processo d'innovazione; esso coinvolge i tre principali attori del processo d'innovazione, ossia i centri di ricerca, le aziende ICT e gli utenti. I centri di ricerca coinvolti sono: il Dipartimento di Ingegneria e Scienze dell'Informazione dell'Università di Trento, la Fondazione Bruno Kessler, GraphiTech, CREATE-NET, ISTC-CNR, il Laboratorio di Interoperabilità ed e-Government (LEGO), varie imprese tra le quali I&S Informatica e Servizi S.r.l., COGITO S.r.l., Sinergis S.r.l., Trentino Network S.r.l., GPI S.p.A., DedaGroup S.p.A., ALGORAB S.r.l., HEIDI S.p.A., Centro Ricerche Fiat (CRF), Siemens S.p.A. e varie organizzazioni in qualità di utenti finali, tra le quali la Provincia Autonoma di Trento, l'Azienda Provinciale per i Servizi Sanitari, il Consorzio dei Comuni Trentini e Trentino Riscossioni S.p.A.. L'iniziativa TasLab è patrocinata dalla Provincia Autonoma di Trento. In questo senso, essendo il team di Innovation Managers localizzato in Informatica Trentina, Informatica Trentina agisce in qualità di catalizzatore e coordinatore del Trentino as a Lab.

4.1.2.1 MISSION E OBIETTIVI

Mission del Trentino as a Lab è la creazione di un'infrastruttura di innovazione avanzata capace di rispondere alle esigenze degli utenti attuali e futuri, non solamente dal punto di vista dell'ICT, ma anche in una prospettiva culturale e socio-economica. Lo scopo è di ridurre il divario digitale e allo stesso tempo di sperimentare nuove soluzioni ICT con un forte coinvolgimento dell'utente su tutto il territorio montano trentino.

I principali obiettivi del TasLab sono i seguenti:

- Lo sviluppo di un ambiente incentrato sull'utente, nel quale l'innovazione è il modo di essere, pensare ed evolvere dei cittadini trentini;
- Lo sviluppo di un ambiente naturalmente predisposto ad istanziare l'innovazione nel suo ciclo complessivo, dalla ricerca di base ai prodotti di mercato;
- Lo sfruttamento dei risultati di progetti già attivi per uno sviluppo del Trentino, sostenibile ed attento all'ambiente ed alle esigenze delle persone;

- Lo sviluppo di varie partnership con altri territori europei ed internazionali, derivanti anche da collaborazioni di eccellenza già attivate dagli stakeholders trentini sia a livello nazionale, sia a livello internazionale.

L'approccio si sviluppa su una dimensione orizzontale e verticale. Per quanto riguarda la dimensione orizzontale, l'approccio adotta una prospettiva eco-sistemica (socio-economica), nella quale i diversi attori (cittadini, pubbliche amministrazioni, aziende e centri di ricerca), quali organismi dell'ecosistema, interagiscono gli uni con gli altri ed evolvono in base alle condizioni locali/globali. L'approccio eco-sistemico orizzontale è associato ad una dimensione verticale, che concentra l'attenzione su alcune aree vocazionali del Trentino: le aree centrali del sistema dei valori del territorio trentino (ad esempio: eInclusion, eMobility, eBusiness ed eTourism, qualità della vita ed eEnviroment).

4.1.2.2 REFERENZE ED ESPERIENZE

L'iniziativa Trentino as a Lab è nata nel 2005 e si trova oggi in una fase di ampliamento del team di Innovation Managers, sintomo del successo del progetto. Gli attori che lo compongono possiedono una notevole esperienza di coinvolgimento nei progetti R&D europei e nei progetti focalizzati sull'utente. Vari casi pilota e strutture sperimentali sono state o si stanno costruendo sul territorio trentino.

Trentino as a Lab è membro di ENOLL (European Network of Living Labs) come membro coordinatore dell'area *eResponse* ed è coinvolto nello sviluppo del centro internazionale di addestramento della Protezione Civile, Trentino Training Center. Nell'area *eEnvironment* Trentino as a Lab è coinvolto nella definizione di servizi innovativi basati sulla tecnologia Cosmo-Skymed con il diretto coinvolgimento di Telespazio, delle aziende e i centri di ricerca trentini. Nell'area *eInclusion* partecipa al progetto CSS (Cartella Socio Sanitaria) volto all'integrazione della cartella sociale con la cartella clinica del cittadino al fine di creare un patrimonio di dati integrati. Nell'area di *eGovernment* è coinvolto nel progetto di Trentino Riscossioni, società a totale partecipazione pubblica che svolge il servizio di riscossione e accertamento delle entrate degli enti pubblici del Trentino, volto alla realizzazione del principio di equità fiscale attraverso il supporto delle tecnologie ICT e nel progetto Pro.De, progetto interregionale sulla de-materializzazione. Nella stessa area ha avviato una collaborazione con il Mozambico per la creazione e avviamento di un Living Lab e per lo sviluppo del piano strategico per l'interoperabilità eGovernment per il Mozambico (progetto eGIF). Inoltre, nell'ambito delle attività finanziate dal Fondo Sociale Europeo, stanno proseguendo le attività di analisi e progettazione della rete finalizzata all'innovazione e trasferimento delle conoscenze dai centri di ricerca alle aziende. Sono anche in fase avanzata di attivazione quattro progetti di innovazione pilota con le aziende del territorio.

4.2 TRENTINO RISCOSSIONI S.P.A.

Trentino Riscossioni S.p.A. è stata costituita l'1 Dicembre 2006 con capitale interamente della Provincia di Trento e la previsione di una successiva partecipazione degli Enti locali e di altri Enti pubblici[3]. La Provincia autonoma di Trento ne ha deliberato la costituzione in seguito al d.l. 203 del 30 settembre 2005, il quale stabilisce che[4]: "*dal 1 ottobre 2006 Riscossione S.p.A.*[5] *subentra ai contratti in essere tra concessionari ed Enti Locali con l'obiettivo di gestire in via esclusiva la riscossione mediante ruolo per tutti gli Enti pubblici, salvo diversa decisione di quest'ultimi*". Nel timore che criteri, modalità e tempi di restituzione delle somme venissero a dipendere da scelte statali, con conseguenti minacce per la propria autonomia finanziaria, la Provincia autonoma di Trento ha deciso di mantenere il pieno controllo in materia tributaria, costituendo un'apposita società da lei controllata. Trentino Riscossioni nasce quindi con le seguenti connotazioni:

- A prevalente o esclusivo capitale pubblico (partecipata da Provincia e Comuni);
- Affidataria delle attività di accertamento, liquidazione, riscossione spontanea e riscossione coattiva delle entrate;
- Come strumento di sistema attraverso cui la Provincia attua la propria strategia finanziaria;
- Con funzioni di organismo pagatore dei benefici previsti per l'agricoltura dai regolamenti comunitari (funzione svolta a livello nazionale dall'Agea[6]) e degli aiuti previsti dalla normativa provinciale.

4.2.1 OBIETTIVI

L'obiettivo di Trentino Riscossioni è di rappresentare, nel settore della gestione delle entrate tributarie e patrimoniali, un referente unico per i cittadini e per gli Enti pubblici del Trentino. La situazione attuale prevede l'utilizzo di diversi sistemi per il pagamento dei tributi, sistemi tra loro non integrati. Ad esempio, per pagare la tassa automobilistica si utilizza un sistema, per versare l'I.C.I. (Imposta Comunale sugli Immobili) un altro, per le tariffe sui rifiuti un altro ancora. Per chi possiede più immobili la situazione diventa ancora più complicata. Trentino Riscossioni si pone quindi l'obiettivo di essere un punto di riferimento unitario a cui i vari Enti affidano la riscossione dei tributi (anziché avere una

[3] Art. 34 della Legge Provinciale n. 3 del 16 giugno 2006 e, successivamente, delibera della Giunta Provinciale n. 1658 del 18 agosto 2006.
[4] D.l. 203, 30 settembre 2005 "Misure di contrasto all'evasione fiscale e disposizioni urgenti in materia tributaria e finanziaria".
[5] Nel 2007 ha cambiato nome in Equitalia S.p.A.
[6] Agenzia per le Erogazioni in Agricoltura. L'Unione Europea sostiene la produzione agricola dei Paesi della Comunità attraverso l'erogazione di aiuti, contributi e premi. Tali erogazioni sono gestite dagli Stati Membri tramite Organismi Pagatori (l'Agea per lo Stato italiano).

società diversa per ogni Ente) con l'intento di ottenere procedure tributarie più semplici, veloci e omogenee. Far confluire tutti gli enti in un unico soggetto garantirà una serie di vantaggi, come ad esempio:

- Riduzione di costi (e quindi delle tariffe al cittadino);
- Rapporto più diretto dell'Ente con il cittadino;
- Rispetto dall'autonomia di ciascun soggetto coinvolto.

La *mission* di Trentino Riscossioni vuole comunque andare oltre la mera attività di riscossione. La società si configura come strumento di sistema della Provincia, con un modello operativo che prevede:

- La partecipazione societaria degli Enti (per facilitare l'interscambio di dati favorendo la creazione di uno strumento concreto per la gestione delle entrate);
- L'armonizzazione di leggi, regolamenti e procedure applicative tra tutti gli Enti provinciali (circa 250 soggetti pubblici);
- La creazione di un unico sistema informativo che collega in rete i soggetti e gli enti interessati;
- Il superamento delle carenze di conoscenze e professionalità.

L'obiettivo finale di Trentino Riscossioni è quindi quello di innescare un circolo virtuoso che porti vantaggi non solo agli Enti pubblici partecipanti ma in particolar modo ai contribuenti.

4.2.2 OPERATIVITÀ E GOVERNANCE

Il primo passo per l'avvio delle attività di Trentino Riscossioni è stato l'affidamento alla società della gestione dei tributi provinciali: dal 1° marzo 2007 il referente della Provincia in materia tributaria è diventato Trentino Riscossioni, subentrando al Servizio Tributi della Provincia autonoma. Trentino Riscossioni diviene pienamente operativa e inizia l'attività di riscossione nell'ottobre 2007 e ad oggi è composta da circa 40 dipendenti. Dal dicembre 2007 ad oggi hanno aderito a Trentino Riscossioni, oltre alla Provincia, 96 Enti che rappresentano circa 306.000 cittadini. I 96 Enti aderenti hanno affidato servizi per conto di 117 comuni per un totale di circa 290.000 abitanti. Trentino Riscossioni gestisce quindi servizi per il 53% dei comuni trentini (117 comuni su un totale di 223), coprendo il 56% della popolazione (290.000 abitanti su un totale di 519.800)[7]. Nel dettaglio, Trentino Riscossioni gestisce[8]:

- Riscossione ordinaria tariffa di igiene ambientale (TIA) per 64 comuni;
- Riscossione ordinaria canone idrico per 7 comuni;

[7] Dati demografici aggiornati al 1 gennaio 2009.
[8] Dati aggiornati al 15 novembre 2009.

- Riscossione violazioni codice della strada per 47 comuni;
- Attività di accertamento per 9 comuni;
- Riscossione coattiva delle entrate per 47 comuni;

La norma istitutiva di Trentino Riscossioni prevede le modalità di partecipazione alla società, ovvero le funzioni di governo, direttiva, indirizzo e controllo [62]. Trentino Riscossioni si configura come società "in-house" e l'affidamento dei servizi è diretto, senza gara d'appalto (l'Ente non si rivolge al mercato ma ad un soggetto che è sostanzialmente equiparabile ad un suo servizio interno). Gli Enti che affidano i propri servizi a Trentino Riscossioni entrano quindi nel capitale sociale e ne divengono soci. Questo fatto costituisce un fattore strategico per la piena realizzazione del progetto. La partecipazione diretta degli Enti permette di semplificare alcune attività (ad esempio evita problemi nel passaggio dei dati dall'Ente a Trentino Riscossioni e viceversa, innescando un circolo virtuoso che contribuirà al miglioramento della qualità dati in possesso degli Enti) e permette agli Enti soci di mantenere il controllo sulla società. Infatti, il modello di Governance adottato da Trentino Riscossioni prevede strumenti di controllo specifici che rafforzano il controllo esercitato dai soci sull'operato della società. La società è dotata di un modello di Governance atipico, non previsto dal codice civile, ma nel pieno rispetto dei requisiti comunitari che prevedono:

- La realizzazione dell'attività nei confronti dei soci;
- Il controllo analogo: i soci devono poter esercitare sulla società un controllo analogo a quello che operano sui loro servizi.

Dato l'alto numero di soci e l'importanza delle attività di riscossione è stato conferito a Trentino Riscossioni un modello di Governance (Figura 29) in grado di rafforzare il controllo analogo grazie alla presenza di due organi:

- **Assemblea di coordinamento**: a cui partecipano tutti gli Enti titolari del capitale sociale. Ha il potere di nominare il comitato di indirizzo e approvare le linee guida della società. Le sue decisioni sono valide quando vi è l'assenso del rappresentate della Provincia Autonoma di Trento e della maggioranza degli atri soci.
- **Comitato di indirizzo**: costituito da sei membri nominati dall'assemblea di coordinamento (3 rappresentanti della Provincia Autonoma di Trento e 3 degli altri soci). L'organo esercita il controllo analogo attraverso poteri incisivi:
    - Designa i membri del consiglio di amministrazione e del collegio sindacale;
    - Approva i livelli di prestazione e le tariffe (il corrispettivo che i singoli Enti devono versare a Trentino Riscossioni come pagamento dei servizi resi);
    - Approva i piani strategici/industriali.

Inoltre, il controllo analogo prevede altri strumenti di controllo, più stringenti rispetto a quanto previsto dal codice civile, ad esempio:

- Il divieto di alienare azioni ai privati;
- Il Consiglio di Amministrazione (CdA) non deve avere rilevanti poteri gestionali;
- Le decisioni più importanti devono essere sottoposte al vaglio degli Enti affidanti.

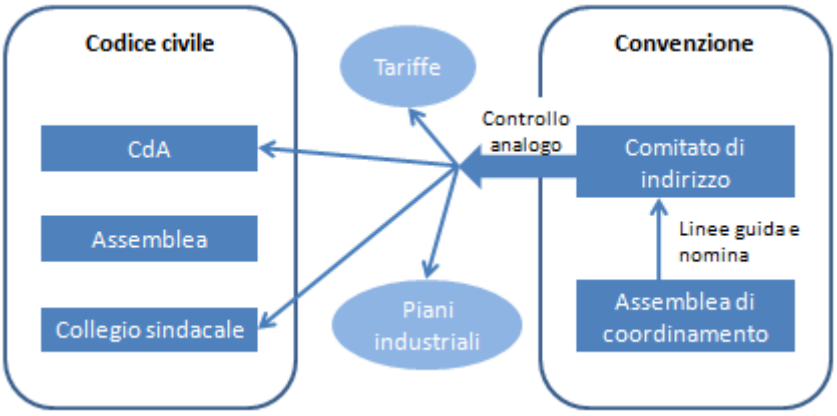

Figura 29 - Trentino Riscossioni: il modello di Governance.

4.2.3 FUNZIONALITÀ

La costituzione di Trentino Riscossioni è una mossa strategica nel panorama nazionale che va verso l'istituzione del federalismo fiscale. Attraverso tale società la Provincia Autonoma di Trento intende mantenere in casa il controllo sulle entrate tributarie, migliorando efficacia ed efficienza dell'intero processo di riscossione. Vediamo nel dettaglio quali sono i servizi offerti da Trentino Riscossioni agli Enti affidatari, servizi sui quali è modellata la struttura della società (Figura 30):

- Riscossione ordinaria di oneri e tributi;
- Riscossione coattiva (recupero dei tributi non versati dai contribuenti per via bonaria o tramite ingiunzione fiscale);
- Rapporto con il contribuente gestito su richiesta da Trentino Riscossioni;
- Supporto dell'eventuale contenzioso;
- Accertamenti tributari su due tipologie di entrate comunali (ICI e TARSU/TIA[9]);
- Acquisizione dei dati dagli Enti per la gestione a regime del sistema;
- Supporto nell'elaborazione della politica fiscale;
- Consulenza sull'organizzazione dell'Ufficio Tributi.

Un'attività di particolare importanza effettuata da Trentino Riscossioni è l'attività di accertamento del corretto importo versato dal contribuente. L'attività di accertamento mira a far emergere evasione ed elusione in materia di tributi locali, in modo da garantire alla finanza locale risorse proprie. Si tratta di un'attività che richiede l'utilizzo di specifici

[9] La TARSU è tassa per lo smaltimento dei rifiuti solidi urbani. A partire dal 1997 è stata disposta la progressiva sostituzione con la TIA (tariffa di igiene ambientale).

strumenti informatici e l'impiego di personale specializzato, dove il recupero di gettito contribuirà a realizzare una politica di riduzione delle tariffe e il conseguimento di una maggiore equità sociale. Trentino Riscossioni si pone l'obiettivo di superare la logica formale (controllo della dichiarazione) per privilegiare la sostanza del rapporto tributario: il controllo della correttezza dei versamenti. La strategia di Trentino Riscossioni è di rappresentare il profilo tributario del singolo contribuente (sulla base dei dati disponibili) per guardare alla sostanza del rapporto tributario.

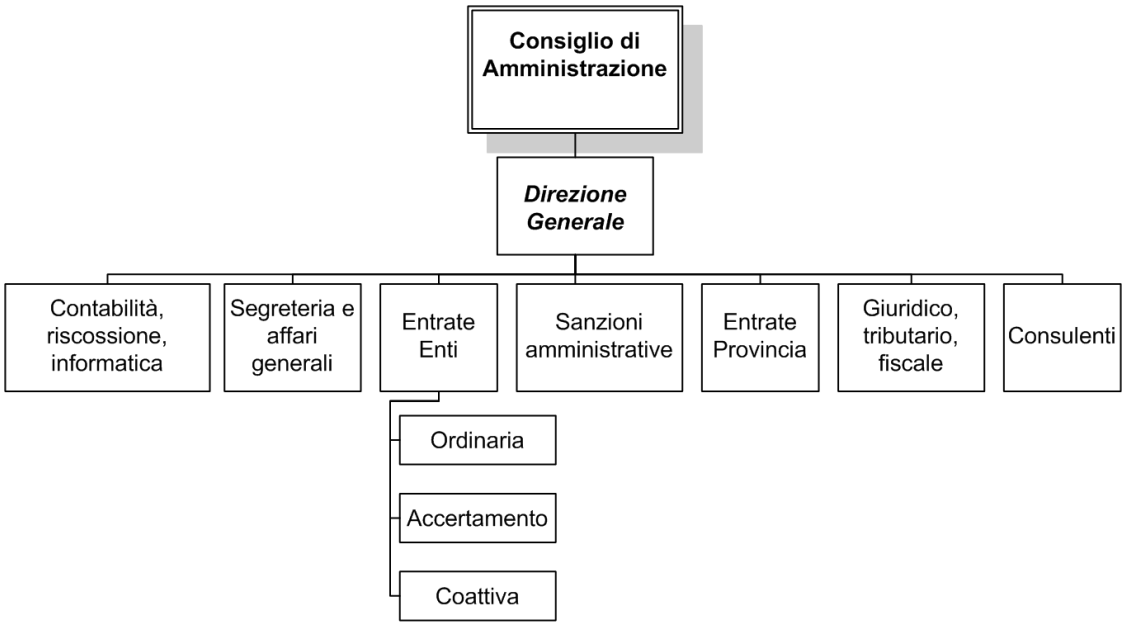

Figura 30 - Trentino Riscossioni S.p.A. - Organigramma Aziendale.

Vediamo quindi quali sono le tipologie di entrate gestite da Trentino Riscossioni, a tale scopo è opportuno dividerle in base agli Enti affidatari. Per la Provincia Autonoma di Trento, Trentino Riscossioni compie attività di riscossione e gestione dei seguenti contributi:

- Tassa Automobilistica Provinciale (bollo auto);
- Tassa per l'Abilitazione all'Esercizio Professionale;
- Imposta Provinciale sulle formalità di trascrizione, iscrizione ed annotazione dei veicoli richieste al pubblico registro automobilistico (IPT);
- Tributo Speciale per il deposito in discarica e per le altre forme di smaltimento dei rifiuti solidi;
- Addizionale Provinciale all'Accisa sul consumo di Energia Elettrica;
- Canone per l'occupazione di spazi ed aree pubbliche (COSAP).

Per i comuni ed altri enti (Comprensori, Unità Sovracomunali) Trentino Riscossioni offre servizi di riscossione/gestione delle seguenti entrate tributarie/patrimoniali:

- Imposta comunale sugli immobili (ICI);

- Tassa sui rifiuti solidi urbani (TARSU), a partire dal 2007 sostituita progressivamente con la Tariffa di igiene ambientale (TIA);
- Tariffe idriche (3 tipologie di tariffe: acquedotto, fognatura e depurazione);
- Canoni di locazione, di uso dei beni pubblici, ed in generale le entrate ed i corrispettivi dovuti dagli utenti per la fruizione dei servizi o dei beni comunali (asilo nido, mense scolastiche, legna, ecc.);
- Violazioni amministrative per conto dei Comandi di Polizia Locale (che hanno affidato il servizio);

Infine, Trentino Riscossioni gestisce la riscossione degli affitti dei posti letto per conto dell'Opera Universitaria.

4.2.4 IL SISTEMA INFORMATIVO

Da quanto visto finora si comprende come la Provincia si aspetti significativi benefici per i soggetti coinvolti (Enti e cittadini in particolare) dalla costituzione di Trentino Riscossioni. Nello specifico, mediante la costituzione di questa società di sistema, la Provincia si aspetta che gli enti saranno in grado di:

- Disporre di una visione globale e integrata della base dati impositiva;
- Aumentare efficacia e efficienza del processo di gestione delle entrate;
- Perseguire obiettivi di equità fiscale;
- Avere un miglior rapporto col cittadino.

Per quanto riguarda i cittadini, ci si aspetta che questi invece potranno in futuro:

- Contare su un referente unico per gestire le problematiche tributarie;
- Accedere via Web a una visione complessiva della propria posizione contributiva (posizioni assolte, scadenze, estratto conto,);
- Utilizzare strumenti di pagamento diversificati e innovativi (fra cui pagamento online).

E' evidente che per far fronte a questi obiettivi ambiziosi, Trentino Riscossioni dovrà dotarsi di un sistema informativo integrato ed evoluto per la gestione di tutti i processi di gestione delle Entrate. In tale senso, il mercato offre diverse soluzioni in tema di riscossione e gestione delle entrate e tutti i prodotti tendono a coprire l'intero ciclo di vita dell'entrata. In particolare, gli applicativi a supporto dell'attività di riscossione hanno raggiunto una buona maturità e sono oggi sistemi completi che prevedono un motore di riscossione capace di trattare non solo la riscossione degli incassi ma anche le problematiche di riversamento e rendicontazione agli Enti affidatari (funzione indispensabile nell'ottica di Trentino Riscossioni). Aggiungendo agli applicativi gestionali

e di riscossione portali rivolti alle comunicazioni fra Enti e Cittadini e includendo servizi come il pagamento online, si ottiene una soluzione completa, consolidata ed affidabile.

Tuttavia, da tali applicativi resta esclusa una funzione di fondamentale importanza: l'accertamento tributario. È bene mettere in risalto che non è una componente sottovalutata, ma al più viene proposta come servizio aggiuntivo e mai come componente integrata nell'applicativo principale. La ragione di questo comportamento risiede nel fatto che l'accertamento tributario è un tema molto complesso, che risente delle caratteristiche dei dati di partenza e degli assetti organizzativi degli Enti; risulta quindi una funzione non completamente automatizzabile dal punto di vista informatico. È evidente che all'interno del progetto di sviluppo di Trentino Riscossioni, la disponibilità di una componente di questo tipo è un fattore di grande rilevanza strategica, ed è per questo che la società, tramite la collaborazione con Informatica Trentina ed altre società esterne, intende costruire una soluzione di accertamento *ad hoc*, che si integri al meglio con la realtà tributaria locale.

4.2.4.1 LA BASE DATI INTEGRATA

Secondo il piano di sviluppo dell'Ente l'intero sistema informativo dovrà essere fondato su una base dati integrata e storicizzata, in grado di contenere tutti i dati utili ai fini del raggiungimento degli obiettivi della PAT e di Trentino Riscossioni. Tale base dati si pone l'obiettivo di integrare non solo tutti i dati utili ai fini tributari ma anche tutte le informazioni utili ai fini della programmazione dell'attività degli Enti. La base dati dovrà essere quindi funzionale per una serie di attività, tra cui:

- Accertamento tributario;
- Gestione del tributo;
- Riscossione e gestione degli incassi;
- Rendicontazione e reporting agli Enti;
- Riversamento dei flussi finanziari;
- Comunicazione e gestione del rapporto con il contribuente.

Le basi dati necessarie per la costituzione di questo database integrato (Data Warehouse) sono di tipologia diversa e sono gestite da soggetti differenti (Provincia, Enti, Comuni). Ad oggi, gli Enti coinvolti nel progetto hanno messo a disposizione le seguenti banche dati:

- Catasto urbano e fondiario;
- Libro Fondiario;
- Planimetrie degli edifici;
- Utenze elettriche e relative a servizi idrici;
- Dati urbanistici (Piani Regolatori);
- Registro Imprese Locali;

- Banche dati tributarie (dichiarazioni e versamenti);
- Banche dati acquisite presso i comuni (anagrafe e toponomastica).

Occorre sottolineare che il progetto non vuole limitarsi alla sola integrazione delle basi dati finora individuate, in quanto futuro ci sarà quasi sicuramente l'esigenza di aggiungere altre banche dati al sistema (vista la costante evoluzione del sistema tributario non è possibile sapere oggi quali dati potranno servire in futuro). Per questo l'architettura del Data Warehouse dovrà essere molto flessibile, garantendo la possibilità di integrare banche dati non previste al momento della progettazione iniziale. Vi è poi la necessità di gestire uno storico dei dati (funzione che non tutte le banche dati provinciali/comunali assolvono), in quanto per i tributi è previsto il recupero del dovuto non incassato fino a 5 anni indietro nel tempo. Sarà quindi di fondamentale importanza avere un sistema in grado di "fotografare" la situazione di un determinato contribuente in un determinato istante temporale.

La base dati integrata sarà quindi il perno su cui innestare i servizi di Trentino Riscossioni, pertanto sulla sua progettazione deve essere posta la massima attenzione. Particolare attenzione andrà posta nella verifica della qualità e della correttezza dei dati, in quanto alcune azioni, che si genereranno sulla base di tali dati, comportano conseguenze di un certo peso sia nei confronti del contribuente sia nei confronti di Trentino Riscossioni, qualora siano intraprese in modo errato, con implicazioni dal punto di vista legale. Lo schema sottostante (Figura 31) rappresenta una visione dell'architettura logica del sistema.

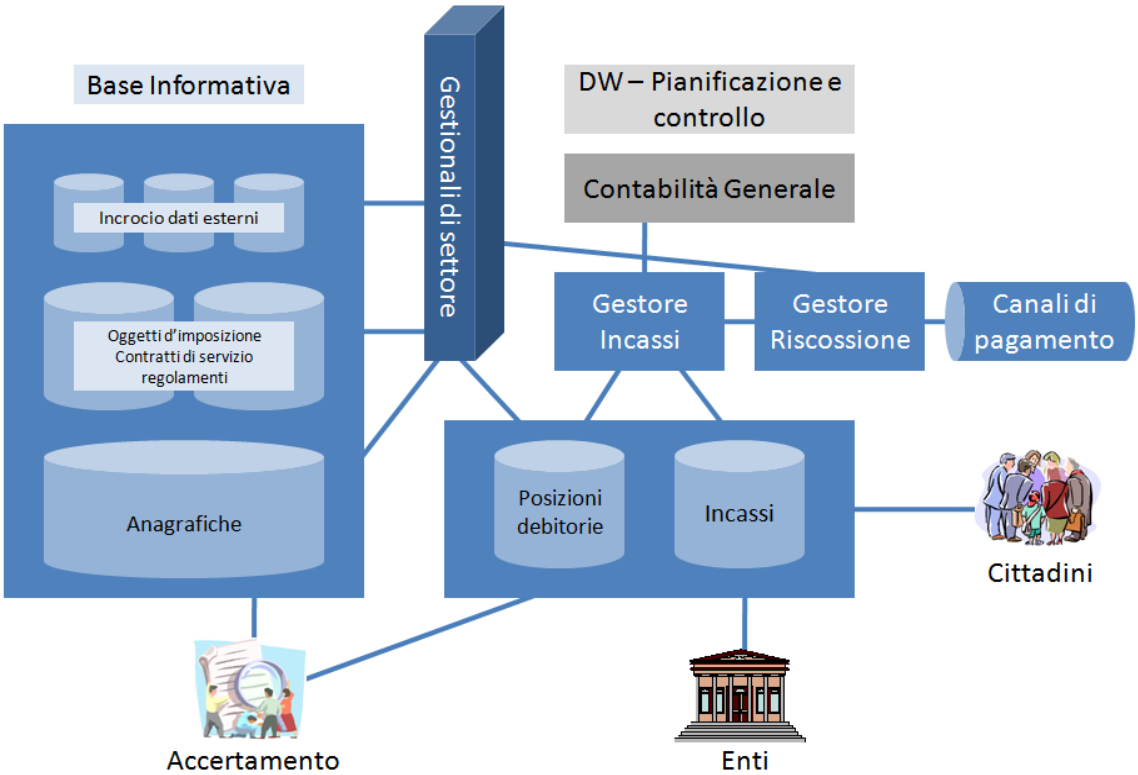

Figura 31 - Architettura logica del sistema informativo.

Per quanto riguarda la base dati integrata si sono individuati tre livelli logici:

1. *Anagrafica dei soggetti contribuenti*: è il livello base, comune a tutte le entrate. La sua costruzione parte dall'anagrafica dello stato civile dei Comuni o del Registro Imprese per i soggetti produttivi e va integrata con altre fonti in funzione del tipo di entrata;
2. *Oggetti/entità su cui è applicato il tributo*: immobili, terreni edificabili, superfici, nuclei familiari, rifiuti, contratti utenze, ecc.);
3. *Basi dati esterne*: da utilizzare per compiere verifiche incrociate sull'attendibilità delle dichiarazioni (elusione) e scoprire evasioni totali.

La realizzazione di questa base dati integrata è un progetto ambizioso, la cui realizzazione darà valenza al disegno di Trentino Riscossioni, permettendo una rilevazione più efficace e veloce di una grande fetta di evasione ed elusione, innescando un processo di recupero del gettito che garantirà una maggiore equità sociale all'interno del territorio trentino.

4.2.4.2 I GESTIONALI DI SETTORE

In Figura 31 si vede come la base dati informativa è in relazione con il complesso delle applicazioni gestionali di settore: gestore di incassi e riscossione, funzione di accertamento, interfaccia con gli Enti, comunicazioni con i contribuenti. Vediamo più nel dettaglio quali sono le funzioni di ognuna di queste componenti.

I gestionali di settore individuano i software di gestione di ogni singolo tributo trattato da Trentino Riscossioni e sono così rappresentati per indicare la possibilità di implementarli in un unico ambiente tecnologico. Rappresentano la chiave di tutto il sistema ed costituiscono un asset strategico per perseguire la strategia finanziaria della Provincia Autonoma di Trento, indirizzata a:

- Uniformare la gestione delle entrate;
- Ridurre l'eterogeneità delle soluzioni software;
- Razionalizzare spese e investimenti in sistemi informatici;
- Realizzare un sistema informativo integrato in grado di collegare in una rete unica tutti i soggetti pubblici coinvolti nella gestione delle entrate tributarie.

Il gestore degli incassi si occupa della rilevazione di pagamenti ritardati o omessi ed innesca i processi di recupero del gettito (dalla emissione di solleciti ad azioni di recupero coattivo). Tale componente alimenta inoltre la contabilità di Trentino Riscossioni registrando le informazioni opportune per permettere di riversare tali somme all'Ente di competenza.

Il gestore della riscossione amministra le liste di carico dei tributi, le stampe e le postalizzazioni, i resi dovuti e la rendicontazione per gli Enti. Tale componente viene alimentata con i flussi in uscita dai gestionali di settore e con i pagamenti acquisiti dal

sistema bancario e postale. Riguardo a questa componente si è scelto di acquisire un sistema di riscossione già pronto e disponibile sul mercato e concentrare gli investimenti sullo sviluppo ad hoc di altre componenti dell'architettura.

Sul gestore degli incassi si articola il sistema di contabilità di Trentino Riscossioni, un sistema che deve gestire la contabilità economico-patrimoniale e la specificità delle operazioni derivanti dalle attività di riscossione. Visti i rapporti della società con la Provincia Autonoma di Trento e con la società di sistema Informatica Trentina, si è deciso di sfruttare il know-how di quest'ultima individuando in SAP l'ambiente ideale su cui sviluppare la componente contabile.

Particolare importanza assumo i canali di comunicazione con gli Enti e il cittadino. Agli Enti va data la visibilità delle entrate di loro competenza a diversi livelli di aggregazione (arrivando fino al dettaglio massimo, ovvero fattura del singolo contribuente). A questo si aggiunge la visibilità di altre tipologie di dati, come ad esempio: solleciti, incassi non esigibili, perdite, procedure coattive attivate, ecc. L'Ente deve poter quindi monitorare secondo diversi livelli di aggruppamento tutta una serie di dati utili per la programmazione delle proprie attività.

Infine, per quanto riguarda il rapporto con il cittadino Trentino Riscossioni lo ritiene un aspetto di fondamentale importanza, sia per un fatto di immagine per la società sia per instaurare un rapporto di fiducia con l'utenza. Il sistema dovrà garantire al cittadino la possibilità di:

- Visualizzare il proprio estratto conto (obblighi assolti e scadenze);
- Facilitare il calcolo del dovuto;
- Effettuare pagamenti online;
- Offrire uno spazio di interazione con l'Ente per la formulazione di domande e risposte.

4.2.5 IL PIANO INDUSTRIALE

Nell'agosto 2006 sono state presentate le conclusioni di un primo studio con l'obiettivo di verificare la sostenibilità del primo piano industriale di Trentino Riscossioni. Lo studio, dopo aver analizzato il mercato di riferimento, ha stabilito che l'offerta di Trentino Riscossioni doveva svilupparsi secondo un approccio concorrenziale rispetto agli operatori di riscossione esterni (ad esempio Equitalia) e di convivenza con gli attori locali che operano nel campo della gestione delle entrate. Sono stati quindi definiti degli obiettivi di informatizzazione del sistema in base all'orizzonte temporale considerato.

Nel breve periodo Trentino Riscossioni punta all'unificazione dei sistemi di riscossione, tramite l'adozione di una soluzione unica in grado di interfacciarsi con i sistemi di gestione dei tributi esistenti. In questo modo si possono mantenere le funzionalità dei gestionali

esistenti, conservando inoltre gli stessi flussi di dati. Soltanto a valle, nella fase di acquisizione da parte del sistema di riscossione, si sviluppano gli adattamenti dei flussi. In questo modo Trentino Riscossioni ha potuto essere operativa in tempi brevi dedicando le proprie risorse ai processi di riscossione che costituiscono il core business della società. La soluzione applicativa di breve periodo è schematizzata in Figura 32. Si prevede quindi di sostituire la molteplicità dei sistemi di riscossione con una proposta unica in grado di acquisire i flussi in uscita dai gestori dei tributi.

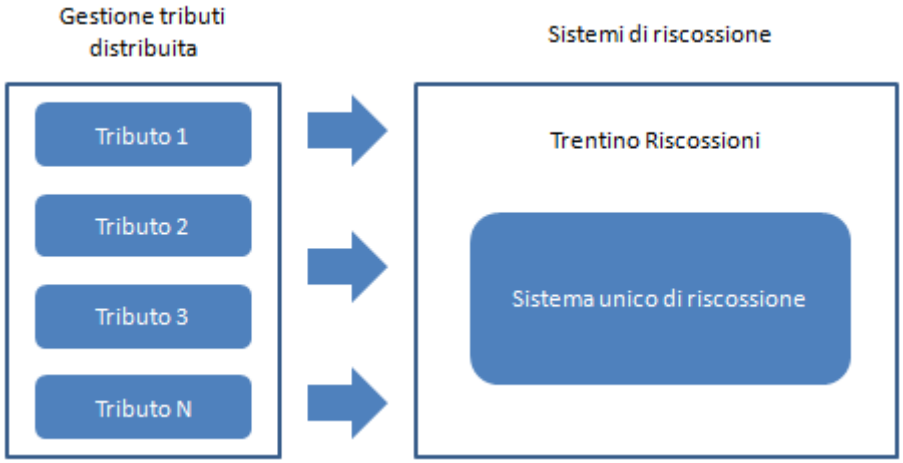

Figura 32 - Soluzione applicativa di breve periodo.

Nel medio-lungo periodo l'obiettivo diviene più ambizioso, mirando ad omogeneizzare le procedure applicative con l'idea di realizzare un unico sistema informativo integrato in grado di mettere in rete tutti i soggetti coinvolti, con una conseguente razionalizzazione dei costi di gestione[10]. In Figura 33 vediamo quella che sarà l'evoluzione del sistema informativo, passando dal sistema unico di riscossione (breve periodo) ad un sistema basato su una base dati integrata e in grado di coprire i processi di calcolo e gestione di ogni singola tipologia di entrata/tributo (lungo periodo).

Nel piano industriale 2008-2010 viene delineato il quadro strategico per lo sviluppo del sistema informativo a supporto dell'attività di Trentino Riscossioni. Dallo schema riportato in Figura 34 risulta evidente il ruolo centrale ricoperto dalla base dati integrata, che nell'ottica di Trentino Riscossioni è vista come il deposito dei dati certificati, con rappresentazione storica e aggiornata della realtà tributaria locale. Una base dati che si presta a supportare tutta una serie di funzionalità, tra le quali la funzione di accertamento per perseguire la lotta all'evasione ed all'elusione fiscale. Sugli stessi dati sarà possibile costruire il sistema per i rapporti con il cittadino, dove quest'ultimo può conoscere il proprio stato impositivo, con evidenza di obblighi assolti e obblighi in scadenza, con la possibilità di pagare online. Inoltre, tale base dati potrà essere la fonte da cui recuperare i

[10] Attuando quanto previsto dalla delibera della Giunta Provinciale n. 1658 del 18 agosto 2006.

dati di proprietà di beni mobili e immobili del contribuente per dare il via ai processi di riscossione coattiva delle entrate. Infine, potrà essere utilizzata come supporto decisionale dell'Ente nella fase di programmazione dei servizi, agevolando la definizione delle tariffe e la ripartizione dei costi.

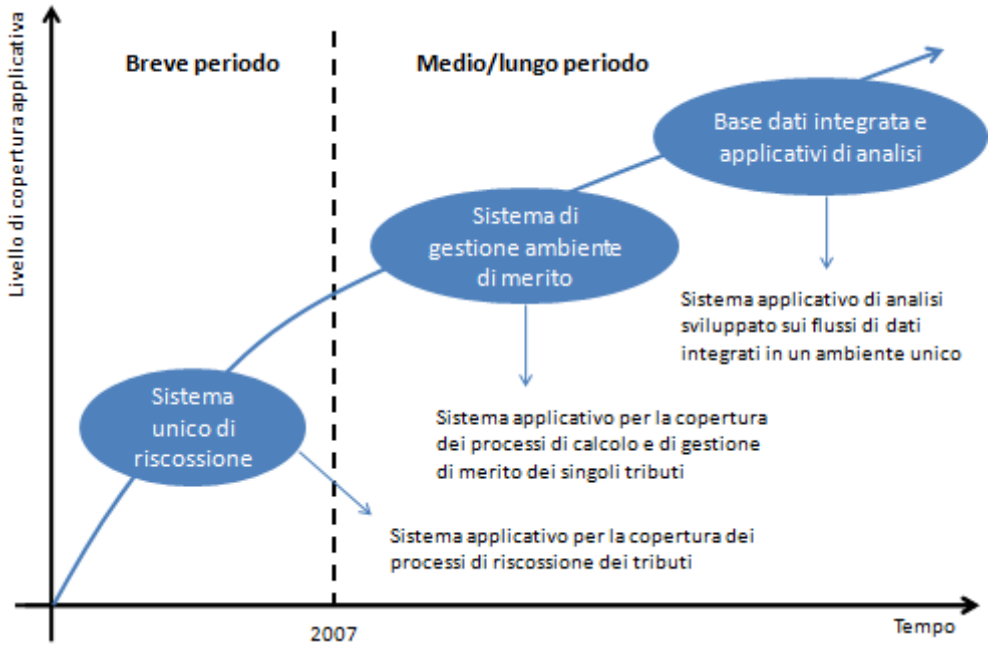

Figura 33 - Evoluzione del sistema informativo dal breve al lungo periodo.

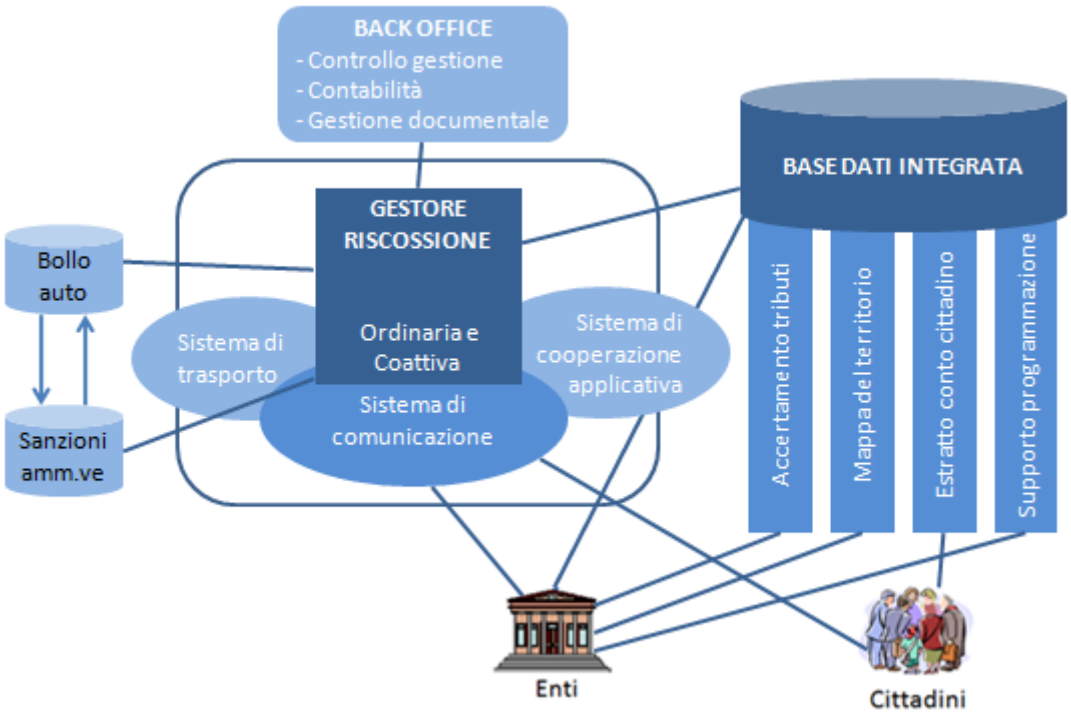

Figura 34 - Lo schema di sviluppo del sistema informativo a supporto delle attività di Trentino Riscossioni.

Nel giugno 2008 è stato definito un accordo con il Comune di Firenze con l'obiettivo di osservare più da vicino la loro attività sul tema dell'integrazione dei dati, iniziata nel 2004. Il lavoro sviluppato a Firenze è molto interessante e tocca una problematica molto simile a quella di Trentino Riscossioni, sebbene non limitata al solo contesto tributario. L'esperienza del Comune di Firenze ha favorevolmente impressionato i vertici di Trentino Riscossioni ed è stato deciso di avviare una collaborazione tra i due Enti su tale tematica. Grazie all'esperienza del Comune di Firenze e alla norma sul riuso delle soluzioni informatiche nell'ambito della Pubblica Amministrazione si è potuto dare il via in tempi molto brevi ad una fase di sperimentazione sui dati reali di un comune trentino (Comune di Folgaria)[11]. L'esperienza sull'integrazione dei dati sarà approfondita nel Capitolo 5.

4.3 IL CONTESTO TRIBUTARIO

In quest'ultima parte del capitolo viene descritto brevemente il contesto tributario nel quale opera Trentino Riscossioni, con particolare riferimento alle attività che interessano lo sviluppo della base dati integrata nella fase di sperimentazione. Una buona comprensione della realtà tributaria in cui opera la società è fondamentale per definire correttamente i requisiti del sistema informativo e per poter strutturare la base dati integrata in modo da facilitare le elaborazioni sui dati che andrà a contenere. Vediamo alcune nozioni tributarie di base, cominciando dalla definizione di tributo.

Nel nostro ordinamento non esiste una definizione puntuale di tributo, una possibile definizione è quella riportata in [63]: *"Si definisce tributaria l'entrata caratterizzata dalla coattività della prestazione. Tale coattività costituisce l'elemento essenziale ed incontestabile per l'individuazione del tributo e per la sua differenziazioni dalle altre entrate. Sono pertanto tributi le imposte, le tasse, i monopoli fiscali e i cosiddetti contributi"*. I cittadini sono tenuti al pagamento dei tributi in base all'obbligo imposto dall'articolo 53 della Costituzione, il quale recita: *"Tutti sono tenuti a concorrere alle spese pubbliche in ragione della loro capacità contributiva. Il sistema tributario è informato a criteri di progressività"*. L'articolo 23 prevede inoltre che: *"Nessuna prestazione personale o patrimoniale può essere imposta se non in base alla legge"*. Da questo risulta che gli Enti locali possono disciplinare con regolamento le proprie entrate tributarie entro i limiti posti dalla legge statale.

[11] Il Codice dell'Amministrazione Digitale (CAD) è stato emanato con il d.lgs. n. 82 del 7 marzo 2005 ed è entrato in vigore il 1gennaio 2006, subendo una serie di correttivi con il decreto legislativo n. 15 del 4 aprile 2006. Il CAD ha lo scopo di regolamentare disponibilità, gestione, accesso, trasferimento, conservazione e fruibilità dell'informazione digitale all'interno della Pubblica Amministrazione e nei suoi rapporti con i privati.

Possiamo quindi focalizzare l'attenzione sulle entrate tributarie degli Enti locali, in particolare su tasse e imposte. In [64] le imposte vengono definite come: *"Il tributo di più facile identificazione, in quanto riferito a una manifestazione di capacità economica, indipendentemente da una specifica attività di soggetti pubblici nei confronti del contribuente"*. Lo stesso autore chiarisce come le tasse si differenziano dalle imposte, in quanto: *"sono prelevate per lo svolgimento di una funzione pubblica nei confronti del contribuente"*. A differenza dell'imposta la tassa è quindi legata alla prestazione di un servizio da parte di un ente pubblico. Le tasse si pongono talvolta al confine con le tariffe, un confine non sempre ben definito ma contraddistinto da diverse sfumature (come vedremo nel paragrafo 4.3.2.2 relativamente ai tributi TARSU e TIA).

Nell'ambito del rapporto tributario si distinguono due categorie di soggetti:

- **Soggetto passivo**: colui che realizza il presupposto, ovvero il fatto al verificarsi del quale si origina l'obbligazione tributaria, alla quale deve adempiere;
- **Soggetto attivo**: l'Ente creditore del rapporto tributario.

Mentre il soggetto attivo è unico il soggetto passivo è costituito da un pluralità individui che coincidono con tutti i soggetti tenuti al pagamento del tributo.

La gestione delle entrate tributarie, in una visione di molto semplificata, è quindi un processo che implica i seguenti passaggi:

- Per ogni tipologia di imposta si sviluppa l'attività di calcolo e liquidazione. Operando sulle basi dati disponibili si determina l'imposta dovuta da parte di ogni contribuente (per le imposte che prevedono l'autoliquidazione è il contribuente stesso che, sulla base degli elementi forniti con la dichiarazione dei redditi, calcola l'imposta dovuta);
- Il contribuente, informato sui propri obblighi, effettua i pagamenti;
- I pagamenti confluiscono sul conto di Tesoreria e vengono rendicontati.

A questo punto è possibile avviare delle procedure di controllo sulle somme pervenute all'Ente rispetto all'imposta effettivamente dovuta dal contribuente. Si tratta dell'attività di accertamento dei tributi che assume fondamentale importanza nell'ottica del perseguimento di obiettivi di equità fiscale. Nel paragrafo che segue vediamo questa attività più nel dettaglio.

4.3.1 L'ACCERTAMENTO TRIBUTARIO

L'attività di accertamento si pone l'obiettivo di contrastare evasione ed elusione dei tributi. Lo sviluppo di una componente informatica in grado di automatizzare, per quanto possibile, i processi di accertamento, o di fornire un supporto all'operatore umano, costituirà un notevole valore aggiunto per il sistema informativo di Trentino Riscossioni.

Prima di approfondire le procedure di accertamento e tributi oggetto di quest'ultime, è opportuno definire le modalità secondo le quali il contribuente cerca di non assolvere i propri obblighi tributari, prassi che prendono il nome di evasione ed elusione fiscale. L'evasione fiscale può essere parziale o totale e consiste in tutte quelle pratiche volte a ridurre o eliminare il prelievo fiscale. Nel caso di beni immobili l'evasione coincide spesso con una delle seguenti situazioni:

- Aree edificabili non dichiarate;
- Edifici esistenti non censiti, ad esempio in seguito a ristrutturazioni, e su cui gravano utenze;
- Edifici accatastati ma privi di rendita catastale (sulla quale è calcolato l'importo dovuto per l'ICI);
- Ristrutturazioni o lavori senza licenza edilizia.

L'evasione fiscale comporta la violazione di specifiche norme ed è punita con sanzioni amministrative, pecuniarie e, oltre certe soglie, anche penalmente[12].

L'elusione fiscale è un fenomeno diverso, non assimilabile all'evasione, e consiste nel falsificare la natura dell'operazione con lo scopo di beneficiare di imposte di minor valore. È la situazione in cui il contribuente denuncia tutti i beni posseduti soggetti a imposta ma applica sistemi di calcolo dell'importo dovuto non corretti o ricorre alla creazione di situazioni artificiose. Riguardo ai beni immobili possono essere situazioni in cui:

- Si utilizzano rendite presunte anziché rendite effettive;
- Si attribuiscono residenze fittizie ai figli;
- Vi è una moltiplicazione di abitazioni principali (solitamente una assegnata alla moglie e una al marito).

A differenza dell'evasione l'elusione non si configura come illegale; tale pratica non comporta la violazione delle leggi vigenti ma le aggira nel loro aspetto sostanziale eludendo il motivo per il quale sono state approvate. In Italia esiste una norma antielusione[13] secondo la quale: *"sono inopponibili all'amministrazione finanziaria gli atti, fatti e negozi, anche collegati tra di loro, che siano contemporaneamente: (1) privi di valide ragioni economiche, (2) diretti ad aggirare norme tributarie e (3) volti ad ottenere una riduzione del carico fiscale altrimenti indebita".* La norma prevede tre condizioni che devono presentarsi contemporaneamente affinché il soggetto pubblico possa richiedere al contribuente le maggiori imposte dovute. Per rilevare evasione ed elusione è necessario ricorrere all'incrocio di dati anagrafici con dati delle utenze e delle proprietà di beni mobili e immobili. Ricostruita la posizione corretta del contribuente è necessario incrociare tale

[12] D.Lgs. n. 74 del 10 marzo 2000.
[13] Art. 37-bis del D.P.R. n. 600/1973.

situazione con i dati forniti dal contribuente stesso tramite dichiarazioni o versamenti tributari. L'eterogeneità di tali basi dati, unita alla complessità del contesto tributario, rende le operazioni di accertamento lunghe e laboriose; senza l'ausilio di uno strumento informatico dotato di funzionalità specifiche a supporto dell'accertamento non è possibile rilevare in tempi brevi tutti i casi di evasione/elusione fiscale.

4.3.2 I TRIBUTI OGGETTO DI ACCERTAMENTO

I tributi su cui si concentra l'attività di accertamento svolta da Trentino Riscossioni sono l'Imposta comunale sugli immobili (I.C.I.) e la Tassa per lo smaltimento dei rifiuti solidi urbani (T.A.R.S.U.). Vediamoli nello specifico, con particolare attenzione all'ICI, in quanto per valutare i risultati della sperimentazione ci si è focalizzati sullo sviluppo di strumenti a sostegno dell'accertamento di tale imposta.

4.3.2.1 I.C.I.

L'Imposta comunale sugli immobili è stata istituita con il decreto legislativo n. 504 del 1992 e costituisce una delle principali fonti di gettito dei comuni [63] [65]. L'articolo 1 del d.lgs. individua il presupposto dell'imposta che consiste nel possesso di fabbricati, aree fabbricabili e terreni agricoli situati nel territorio dello Stato italiano, indipendentemente dalla loro destinazione economica[14]. L'oggetto dell'obbligazione è quindi rappresentato dall'immobile. In particolare, si possono delineare le seguenti categorie di beni immobili soggetti a ICI:

- Fabbricati, aree edificabili e terreni agricoli posseduti come bene di consumo finale o a titolo di investimento;
- Le stesse tipologie di beni posseduti come fattori produttivi dell'azienda;
- Beni prodotto: il risultato finale di attività industriale diretta alla produzione (ad esempio un immobile costruito da un'impresa immobiliare e destinato alla vendita)[15];
- Beni merce: beni oggetto dell'attività di intermediazione (ad esempio acquistati da una società immobiliare per poi essere rivenduti).

L'articolo 3, comma 1 e 2, del d.lgs. 504/1992 individua come soggetti passivi dell'ICI il proprietario degli immobili e il titolare dei diritti reali di usufrutto, uso, abitazione, enfiteusi e superficie che gravano sui beni immobili soggetti all'imposta. Il comma 2 del medesimo articolo definisce che per i beni immobili concessi in locazione finanziaria il soggetto passivo è il locatario. Infine, rientra tra i soggetti passivi il concessionario che ha

[14] Ai sensi dell'articolo 15 della Legge 27 dicembre 1977, n. 984 i terreni agricoli ricadenti in aree montane o di collina sono esenti dall'ICI. In Trentino i terreni agricoli sono esenti dall'ICI.
[15] Si fa riferimento all'esercizio di impresa commerciale (art. 2195, n. 1 del codice civile)

ottenuto in concessione aree demaniali. L'articolo 4 del d.lgs 504/1992 indica come soggetto attivo il Comune sulla cui superficie gli immobili oggetto dell'imposta insistono interamente o prevalentemente. Il d.lgs. 504/1992 specifica inoltre che cosa si intende con il termine fabbricato[16]: *"l'unità immobiliare già iscritta o che comunque deve essere iscritta nel catasto edilizio urbano"*. Secondo la vigente normativa catastale[17] sono oggetto di imposta le costruzioni che rientrano nelle seguenti categorie catastali:

- Immobili a destinazione ordinaria:
  - Gruppo A: abitazioni e uffici;
  - Gruppo B: edifici di uso collettivo (scuole, ospedali, biblioteche, ecc.);
  - Gruppo C: immobili a destinazione commerciale;
- Immobili a destinazione speciale:
  - Gruppo D: alberghi, opifici, teatri, ecc.;
- Immobili a destinazione particolare:
  - Gruppo E: stazioni per servizi di trasporto, ponti comunali, costruzioni e fabbricati speciali per esigenze pubbliche, ecc.

Il calcolo dell'imposta

Il calcolo dell'imposta dovuta è funzione di diversi attributi. Per quanto riguarda i fabbricati il calcolo del dovuto è funzione di:

- **Base imponibile**, ottenuta dal prodotto di:
  - *Rendita catastale* vigente al 1° gennaio dell'anno di imposizione e rivaluta del 5%;
  - *Moltiplicatore*: assume valori diversi in base alla categoria[18];
- **Quota di possesso**: espressa in frazioni (1/2, 2/3, ecc.) è la quota di fabbricato posseduta da un singolo soggetto (persona fisica o persona giuridica);
- **Periodo di imposta**: il periodo coincide con l'anno solare e l'imposta è dovuta proporzionalmente ai mesi dell'anno per cui si è protratto il possesso;
- **Aliquota di imposta**: determinata da ogni singolo Comune entro i limiti imposti dalla legge[19];
- **Detrazioni**: determinate dal singolo Comune entro i limiti fissati dalla legge.

Relativamente alle aree edificabili la base imponibile coincide con il valore venale in comune commercio al primo gennaio dell'anno di imposizione. Il Comune può determinare dei valori minimi al metro quadro ai quali i contribuenti si devono uniformare;

[16] D.lgs. 504/1992, art. 2, comma 1, lett. a).
[17] R.d.l. 13 aprile 1939, n. 652.
[18] 100 per gli immobili del gruppo A e C, 140 per il gruppo B, 50 per gli A/10 e gruppo D, 34 per i C/1.
[19] Da un minimo del 4 per mille a un massimo del 7 per mille. Può variare in base alla categoria del fabbricato. In assenza di apposita delibera comunale si applica l'aliquota minima del 4 per mille.

tuttavia, ogni Comune prevede una serie di casistiche in cui si può beneficiare di un valore ridotto. Per il calcolo del dovuto si utilizzano aliquote e detrazioni in vigore l'anno precedente.

Vediamo un semplice esempio di calcolo dell'imposta: fabbricato di categoria A/02 (abitazioni di tipo civile) con rendita catastale di 300 euro posseduto al 100% dalla stessa persona per tutti i 12 mesi dell'anno, dove il Comune ha deliberato un'aliquota del 5 per mille e una detrazione di 130 euro. Otteniamo:

- Base imponibile: $300 \times 105 = 31500$ (dove 105 comprende il moltiplicatore (100 per il gruppo A) e la rivalutazione del 5%);
- Imposta lorda: $31500 \times 0{,}005 = 157{,}5$;
- Imposta netta: $157{,}5 - 130 = 27{,}5$;

Il pagamento dell'imposta può essere effettuato secondo due soluzioni:

- Versamento in saldo e in acconto: dal 1 al 16 giugno si versa in acconto il 50% del dovuto, dal 1 al 16 dicembre si versa il saldo della restante parte dovuta;
- Versamento unico entro la scadenza della prima rata (alcuni comuni permettono un versamento unico alla scadenza della seconda data).

A partire dal 2008 è stata abolita l'ICI sulla prima casa e sulle abitazioni equiparate dai comuni a quelle principali[20].

4.3.2.2 T.A.R.S.U. E T.I.A.

La tassa per lo smaltimento dei rifiuti solidi urbani (TARSU) è stata istituita con il Decreto Legislativo n. 507 del 15 novembre 1993[21]. L'articolo 62 del d.lgs. n. 507/1993 ne individua il presupposto: *"La tassa è dovuta per l'occupazione o la detenzione di locali ed aree scoperte, a qualsiasi uso adibiti, esistenti nelle zone del territorio comunale in cui il servizio è istituito ed attivato o comunque reso in via continuativa"*. L'importo non è quindi commisurato alla quantità di rifiuti prodotti ma alla quantità di spazio occupato. Soggetti passivi della TARSU sono: "*coloro che occupano o detengono i locali o le aree scoperte di cui all'art. 62 con vincolo di solidarietà tra i componenti del nucleo familiare o tra coloro che usano in comune i locali o le aree stesse*". Il soggetto passivo deve presentare al Comune, entro il 20 gennaio successivo all'inizio dell'occupazione, una denuncia dei locali e delle aree tassabili presenti nel territorio comunale. La denuncia deve contenere: codice fiscale dei componenti del nucleo familiare che dimorano nell'immobile,

[20] D.l. n. 93 del 27 maggio 2008 e decreto fiscale n. 93/2008. L'abitazione principale è dove il contribuente ha la residenza anagrafica o l'abituale dimora.

[21] D.lgs. n. 507 del 15 novembre 1993, revisione ed armonizzazione dell'imposta comunale sulla pubblicità e del diritto sulle pubbliche affissioni, della tassa per l'occupazione di spazi ed aree pubbliche dei comuni e delle province nonché della tassa per lo smaltimento dei rifiuti solidi urbani a norma dell'art. 4 della legge 23 ottobre 1992, n. 421, concernente il riordino della finanza territoriale.

superficie e destinazione d'uso dei singoli locali e la data di inizio dell'occupazione o detenzione.

L'articolo 65 del d.lgs n. 507/1993 contiene le indicazioni per il calcolo della tariffa: "*La tassa è commisurata alle quantità e qualità medie ordinarie per unità di superficie imponibile dei rifiuti solidi urbani interni ed equiparati producibili nei locali ed aree per il tipo di uso, cui i medesimi sono destinati, nonché al costo dello smaltimento. Le tariffe per ogni categoria o sottocategoria omogenea sono determinate dal comune, secondo il rapporto di copertura del costo prescelto entro i limiti di legge, moltiplicando il costo di smaltimento per unità di superficie imponibile accertata, previsto per l'anno successivo, per uno o più coefficienti di produttività quantitativa e qualitativa di rifiuti".* Concretamente, la tassa non è legata all'effettiva produzione di rifiuti ma alla superficie netta calpestabile dell'immobile[22].

Il Decreto Legislativo n. 22/97, il cosiddetto Decreto Ronchi, ha istituito la progressiva sostituzione della TARSU con la nuova Tariffa di Igiene Ambientale (TIA). Con il passaggio da tassa a tariffa l'idea del legislatore era di far pagare ai cittadini esattamente in base a quanto usufruiscono del servizio. Il Decreto individua il presupposto e soggetto passivo della tariffa[23]: "*La tariffa deve essere applicata nei confronti di chiunque occupi oppure conduca locali, o aree scoperte ad uso privato non costituenti accessorio o pertinenza dei locali medesimi, a qualsiasi uso adibiti, esistenti nelle zone del territorio comunale*".

Il d.p.R. n. 158 del 27 aprile 1999 determina il metodo di calcolo della tariffa, in particolare prevede che[24]: "*La tariffa è composta da una parte fissa, determinata in relazione alle componenti essenziali del costo del servizio, riferite in particolare agli investimenti per le opere e dai relativi ammortamenti, e da una parte variabile, rapportata alle quantità di rifiuti conferiti, al servizio fornito e all'entità dei costi di gestione*".

Fin dalla sua definizione la TIA ha originato molteplici dibattiti sulla sua natura non di carattere tributario, ma civilistico (tariffa) che la rende quindi soggetta a IVA. La Provincia Autonoma di Trento dispone di autonomia legislativa sulle tariffe ma non sui tributi, pertanto con la definizione della TIA è stato emanato un regolamento provinciale che, in alcuni aspetti, differenzia la TIA trentina da quella italiana[25]. In alcuni casi il passaggio da tassa a tariffa è stato ostacolato dai cittadini al punto che molti comuni continuano tutt'oggi

[22] La superficie netta calpestabile è la superficie dell'unità immobiliare al netto dei muri interni, pilastri e perimetrali.
[23] Titolo IV del decreto Ronchi.
[24] D.p.r. n. 158/99, Regolamento recante norme per la elaborazione del metodo normalizzato per la definizione della tariffa del servizio di gestione del ciclo dei rifiuti urbani. Modificato successivamente in alcune sue parti dalla Legge n. 488 del 23 dicembre 1999 (Legge Finanziaria 2000) e dalla Legge n. 289 del 27 dicembre 2002.
[25] Legge provinciale n. 5 del 14 aprile 1998.

ad applicare la TARSU. L'Agenzia delle Entrate si è pronunciata in materia due volte confermando la natura civilistica della TIA e la sua assoggettabilità all'IVA[26]. Tuttavia, tale interpretazione non è stata condivisa da parte della giurisprudenza, in particolare la Corte di Cassazione ne ha ravvisato la natura tributaria, escludendo la TIA dal campo di applicazione dell'IVA[27]. Il dibattito è comunque rimasto aperto fino alla pronuncia della Corte Costituzionale, le cui sentenze costituiscono fonte legislativa, la quale ha stabilito che[28]: *"La Tariffa di Igiene Ambientale non è inquadrabile tra le entrate non tributarie, ma costituisce una mera variante della TARSU*". La Corte stabilisce quindi che la TIA è un tassa e non una tariffa, escludendola dall'applicazione dell'IVA. La sentenza ha inoltre effetto retroattivo con la conseguenza che circa 15 milioni di famiglie potrebbero teoricamente chiedere il rimborso dell'IVA versata negli ultimi dieci anni[29]. In aggiunta, tale sentenza priva la Provincia Autonoma di Trento del potere legislativo in materia, rendendo di fatto illegali i criteri di applicazione della TIA nei comuni che applicano la tariffa secondo la normativa provinciale. Di fatto, si è in attesa di una pronuncia a livello nazionale che chiarisca una volta per tutte l'effettiva natura del tributo. Nel frattempo, la Provincia Autonoma di Trento suggerisce agli Enti di continuare a riscuotere la TIA secondo la normativa provinciale.

4.4 CONCLUSIONI

Trentino Riscossioni si presenta come una società giovane e in fase di espansione, operativa da poco più di due anni si aspetta di raggiungere la piena operatività entro fine 2010. Trentino Riscossioni si configura come una società "in-house" della Provincia Autonoma di Trento ed è dotata di un modello di governance particolare che rafforza il controllo dei soci (gli Enti che affidano servizi) sull'operato dell'azienda.

Il contesto di business nel quale opera Trentino Riscossioni è complesso e altamente variabile; un contesto che obbliga la società ad adattare i propri processi di funzionamento alle costanti evoluzioni in tema di tributi e gestione delle entrate. Per questo motivo, Trentino Riscossioni si pone come obiettivo di medio-lungo periodo quello di dotarsi di un sistema informativo sofisticato e innovativo, che sia in grado di raccogliere in una base dati integrata tutti i dati e le informazioni sulle quali si originano processi e attività dell'azienda. Tale sistema informativo, in particolare la base dati integrata, dovrà essere sufficientemente flessibile per adattarsi senza difficoltà alle frequenti variazioni del contesto di business in cui l'azienda opera. La realizzazione di una base dati integrata sulla

[26] R.M. 25/E del 5 febbraio 2003 e R.M. 250/E del 17 giugno 2008.
[27] Sentenza n. 17526/2007.
[28] Sentenza n. 238 del 24 luglio 2009.
[29] Con un costo per lo Stato stimato in 5 miliardi di euro.

quale sviluppare le diverse attività di Trentino Riscossioni (ma anche di altri enti e comuni trentini) permetterà di ottenere una serie di vantaggi. In particolare, Trentino Riscossioni si aspetta di aumentare efficacia ed efficienza dei processi di gestione delle entrate, di garantire maggiore equità fiscale nel territorio trentino (e di conseguenza una riduzione delle tariffe) e di poter contare su un sistema in grado di facilitare i rapporti con il cittadino. Nel capitolo successivo vengono analizzate le fasi di sperimentazione per la realizzazione di una base dati integrata a supporto delle attività dell'azienda.

CAPITOLO 5

ATTIVITÀ DI SPERIMENTAZIONE

In questo capitolo sono esposte le attività sperimentali che hanno consentito di testare sul campo una metodologia di integrazione consolidata. La sperimentazione con uno dei comuni che realmente ha dato in carico a Trentino Riscossioni le attività di accertamento (Comune di Folgaria) ha permesso di testare l'integrazione di una un decina di basi dati eterogenee e distribuite. Nel capitolo sono quindi descritte le attività e i processi necessari per arrivare alla creazione di una base di dati integrata e storicizzata mediante la soluzione tecnologica scelta. Inoltre, saranno presentati e analizzati gli esiti della sperimentazione, indicando pregi e difetti della soluzione testata.

Nel capitolo sono descritte le attività a cui ho collaborato nell'ambito della sperimentazione di integrazione dei dati. In particolar modo sono state svolte attività a supporto dello sviluppo della sperimentazione e dell'analisi dei dati ricavati fatta in collaborazione con il Comune di Folgaria, iniziata nel febbraio 2009 e conclusa nel novembre 2009.

Le attività svolte presso Informatica Trentina hanno comportato il recupero di alcune basi dati presso il fornitore o servizio Web specifico, descrivendo e documentando i vari attributi anche da un punto di vista semantico. Alcune basi dati sono state elaborate prima di avviare l'attività di integrazione attraverso operazioni di calcolo dei dati variati e la separazione dei tracciati multi record.

In un secondo momento, al fine di raccogliere l'esperienza degli operatori tributari, una parte dello stage (tre mesi) è stata spesa direttamente presso gli uffici di Trentino Riscossioni. In questo periodo sono stati approfonditi i processi di accertamento dei tributi ICI e TARSU. La collaborazione con alcuni impiegati del settore ha permesso di comprendere con maggiore dettaglio la metodologia di lavoro e l'uso che viene fatto del dato. Questa attività ha reso possibile una migliore definizione degli obiettivi da perseguire tramite l'utilizzo dei dati integrati, permettendo di definire le procedure di indagine e i casi d'uso più significativi riguardanti le attività di accertamento. Le procedure prodotte sono servite per definire dei criteri di analisi dei dati integrati con il fine di verificare l'entità del surplus informativo reso disponibile dalla base dati integrata. Tale surplus è stato misurato confrontando le operazioni di accertamento manuale del tributo ICI con i risultati ottenibili automaticamente dall'esecuzione di una procedura di calcolo, sviluppata ad hoc, operante sui dati integrati.

5.1 LA SPERIMENTAZIONE SUI DATI DEL COMUNE DI FOLGARIA

La prima fase verso la costituzione di una banca dati integrata ha preso avvio a Marzo 2009. Si tratta di una sperimentazione che coinvolge il Comune di Folgaria (3.142 abitanti e 4.500 posizioni soggette a verifica). La sperimentazione si concentra inizialmente su un singolo ambito di accertamento: l'imposta comunale sugli immobili (I.C.I.). Da questo punto di vista la scelta del Comune di Folgaria è strategica, si tratta infatti di una realtà fortemente turistica con un'alta concentrazione di categorie immobiliari contraddistinte da elevati tassi di evasione/elusione. Per questo Trentino Riscossioni si aspetta un notevole recupero di gettito operando un'analisi più approfondita dei dati relativi a questa località.

La sperimentazione coinvolge diversi attori:

- Provincia Autonoma di Trento:
- Servizio per la Semplificazione Informatica;
- Trentino Riscossioni;
- Informatica Trentina;
- Trentino as a Lab (TASLAB);
- Comune di Firenze: Direzione Sistemi Informativi;
- Aziende private: Gruppo-S.

L'aggregazione dei dati riguardanti il Comune di Folgaria è stata affidata a una società privata (Gruppo-S) che ha già maturato un'esperienza di questo tipo con il Comune di Firenze. La società utilizza un prodotto di integrazione basato su tecnologie Oracle, a cui ha aggiunto un modulo proprietario (MIND) che attraverso tecniche di inferenza statistica è in grado di valutare la qualità dei dati e di aggregarli in un contenitore unico (Data Warehouse) tramite l'assegnazione di pesi in base alla qualità/affidabilità della fonte da cui i dati provengono[1].

Nel sistema attuale la qualità dei dati raccolti dagli Enti è relativamente bassa; risulta evidente che la necessità di un dato corretto è un requisito essenziale per costruire validi processi di gestione ed accertamento delle entrate. Per questo il sistema dovrà garantire misurazioni analitiche sulla qualità del dato, instaurando un circolo vizioso tra Trentino Riscossioni e gli Enti per la convergenza e il miglioramento dei dati.

Da questa sperimentazione Trentino Riscossioni si attende quindi i seguenti risultati:

- Una prima versione di una base di dati centralizzata ed estendibile, navigabile secondo intervalli di tempo utili ai fini dell'accertamento[2];

[1] Le tecnologie Oracle utilizzate sono: Oracle 10/11g Application Server e Oracle BPEL PM.
[2] La sperimentazione si concentra sul periodo 2004-2008.

- Un risultato in termini di accertamento almeno pari a quanto si consegue attualmente con metodi manuali.

Quando il sistema sarà a pieno regime Trentino Riscossioni si aspetta inoltre:

- Una gestione delle entrate integrata e automatizzata;
- Una riduzione del personale di front-office e back-office;
- Un miglioramento generale della qualità del dato (circolo virtuoso fra Enti e Trentino Riscossioni);
- Un aumento di velocità ed efficienza delle fasi operative (accertamento, liquidazione, riscossione ordinaria e coattiva), con un conseguente e sostanzioso recupero di gettito;
- Una conoscenza estesa delle caratteristiche del contribuente a livello provinciale.

5.1.1 LE BASI DATI UTILIZZATE

All'interno della sperimentazione si utilizzano banche dati di due tipologie:

- **Database provinciali**: caratterizzati da flussi strutturati, interfacce Web e disponibilità di una buona documentazione (ma non sempre aggiornata alle ultime evoluzioni);
- **Database comunali**: generalmente sistemi eterogenei, di piccola dimensione e scarsamente documentati.

Questo fa si che i dati a disposizione di Trentino Riscossioni siano altamente eterogenei: formati diversi (txt, csv, mdb, ecc.), struttura e significato del dato differente (lo stesso attributo è usato con valenze diverse), qualità del dato incerta, molti campi nulli, mancanza di identificativi, ecc. Analizzare questi dati con il solo supporto di strumenti di base (Excel, Access) diventa un'operazione molto complicata e costosa (specialmente in termini di tempo e di risorse umane impiegate). Eseguire l'accertamento delle entrate in questo modo non è assolutamente una strada percorribile nel medio/lungo periodo considerando gli obiettivi di equità sociale che si è posta Trentino Riscossioni e il progressivo aumento degli affidamenti di attività di accertamento che determineranno una crescita notevole del numero di posizioni contributive da controllare.

Le basi dati utilizzate all'interno della sperimentazione sono le seguenti:

- **Catasto Urbano**[3]: è la banca dati catastale che contiene informazioni relative agli immobili di tipo fabbricato (particelle edificiali) presenti in un determinato comune[4]. Il dettaglio è a livello di particella (identificata da numero, denominatore

[3] Reperibile sul sistema informative OPENKAT (http://www.catastotn.it/).
[4] Il termine immobile identifica due tipologie di beni: fabbricati (particelle edificiali) e terreni (particelle fondiarie).

e subalterno). Le particelle sono collegate tramite la tabella *titolarità* con persone fisiche e giuridiche che hanno un certo tipo di rapporto con l'immobile (diritti reali di godimento). Il catasto urbano contiene gli attributi necessari per il calcolo dell'importo dovuto ICI: quota e mesi di possesso, rendita e categoria del fabbricato;

- **Catasto Fondiario**[3], gestito tramite database relazionale solo a partire dall'ottobre 2005 e completamente informatizzato solo a partire dal 2009 (il precedente sistema cartaceo non prevedeva l'utilizzo del codice fiscale). Tale situazione, unità alla rigidità del sistema normativo (aggiornamenti possibili solo in seguito ad un decreto del giudice tavolare), fanno si che la qualità dei dati sia spesso inadeguata. La normativa tavolare che regola il catasto fondiario non rende quindi possibili corpose attività di bonifica dei dati e la qualità di questi potrà migliorare soltanto con la movimentazione degli atti. Questi aspetti rendono tale base dati praticamente inutilizzabile nella sperimentazione almeno per la parte riguardante i soggetti fisici e giuridici. Il catasto fondiario contiene:
    - *Particelle*: tutte le particelle (edificiali e fondiarie) presenti in un determinato comune;
    - *Partite tavolari*: dati identificativi delle particelle;
    - *Soggetti*: soggetti fisici e giuridici che hanno una o più relazioni con le particelle;
    - *Diritti ed oneri reali*: contiene i titoli di possesso delle particelle;

  Inoltre, va evidenziato che catasto urbano e catasto fondiario non condividono la stessa base dati ed esistono pertanto due anagrafiche di soggetti titolari. Tali anagrafiche possono essere discordanti e fornire informazioni differenti a causa delle differenti normative che regolano i due catasti (normativa tavolare per il fondiario e normativa catastale per l'urbano). Dal punto di vista normativo il catasto urbano fa da riferimento ai fini fiscali (rendita e valore del bene censito) mentre il sistema fondiario/tavolare certifica la proprietà dell'immobile.
- **Catasto Metrico**[3]: contiene i dati metrici (dimensioni) relativi agli immobili. Contiene inoltre informazioni su persone fisiche e giuridiche, identificativi catastali e indirizzi;
- **Piano Regolatore Generale (PRG)**: il PRG è lo strumento che regola l'attività edificatoria in un territorio comunale, con la finalità di gestire l'espansione urbana. È redatto da un singolo comune o da più comuni adiacenti (nel secondo caso viene definito PRGI, piano regolatore generale intercomunale). La base dati è estratta all'anno 2004 e integrata con le variazioni per gli anni successivi. Il PRG contiene le seguenti informazioni:

  - *Particelle*: tutte le particelle fondiarie ed edificiali (contraddistinte dagli attributi: foglio, numero, denominatore e subalterno) presenti nel comune;
  - *Destinazione urbanistica*: per ogni particella viene riportata la destinazione urbanistica che gli è stata attribuita;
  - *Vincoli*: i vincoli edilizi fissati per ogni singola particella o gruppo di particelle;
  - *Zona di censimento*: attributo che identifica il valore al metro della particella, necessario per il calcolo dell'ICI dovuta;
- **Registro Imprese Locale (Parix)** [5]: Parix è l'applicazione realizzata da InfoCamere che permette ad Enti Pubblici e Uffici provinciali di consultare gratuitamente l'archivio delle imprese iscritte presso la Camera di Commercio di Trento. I dati sono forniti e certificati da InfoCamere e vengono aggiornati quotidianamente. Dal database sono state estratte tutte le imprese iscritte alla Camera di Commercio di Trento alla data del 27 marzo 2007 (il sistema è attivo da tale data e non contiene dati precedenti) e le imprese variate fino al giugno 2009. L'estrazione ha prodotto i seguenti flussi:
  - *Imprese*: tutte le imprese con sede nella provincia di Trento;
  - *Unità locali*[6]: tutte le unità locali presenti nella provincia di Trento, anche quelle relative ad imprese con sede al di fuori del territorio trentino;
  - *Persone fisiche e giuridiche*: tutti i soggetti con un ruolo nelle imprese o unità locali presenti sul territorio;
- **Database I.C.I.**: le basi dati relative al tributo ICI sono di due tipologie:
  - *Dichiarazioni*: la dichiarazione del cittadino a livello di singolo immobile;
  - *Versamenti*: i versamenti del cittadino pervenuti al Comune (aggregati a livello di contribuente e non di singolo immobile);
- **Database T.A.R.S.U.**: banca dati resa disponibile dal comune, contiene le posizioni di pagamento collegate ad immobili occupati;
- **Catasto Elettrico**: contiene tutti i contratti di fornitura elettrica, con indicazione della tipologia di contratto e i consumi di energia. Nonostante la banca dati preveda un collegamento tra utenza elettrica e particella, i record che contengono tale informazione sono un numero estremamente limitato. La banca dati è resa disponibile dalla società fornitrice del servizio elettrico (Trenta S.p.A.);
- **Catasto acquedotto**: contiene i contratti di fornitura idrica collegati ad una particella e ad un soggetto;

[5] Accessibile via Web all'indirizzo: https://www.parix.provincia.tn.it
[6] Definite dal Ministero come: *"l'impianto o corpo di impianti con ubicazione diversa da quella della sede principale o della sede legale, in cui si esercitano una o più attività dell'impresa"*.

- **Anagrafe Residenti**: è il registro della popolazione mantenuto dal comune con il fine di documentare la situazione e il numero delle persone residenti o di quelle che lo sono state in passato. Non è una banca dati storicizzata, questo comporta che per chi è stato residente in passato ed è tornato ad esserlo oggi, dopo un periodo di non residenza nel comune, troviamo soltanto i dati relativi all'ultima posizione;
- **Toponomastica**: è l'insieme dei nomi attribuiti alle entità geografiche (toponimi). Non è un flusso ricavato dall'anagrafe comunale ma fornito da Trentino Servizi, questo perché la toponomastica comunale contiene informazioni soltanto sui residenti e non mappa tutto il territorio.

5.2 IL FORNITORE: GRUPPO S LAB

Trentino Riscossioni, per una prima sperimentazione, ha scelto di non utilizzare i prodotti di Data Integration disponibili sul mercato ma di affidarsi alla soluzione e all'esperienza di una società italiana, Gruppo-S Lab S.r.l., che ha già maturato esperienze significative nell'aggregazione di tali tipologie di dati (in particolare l'aggregazione delle banche dati possedute dal Comune di Firenze e un progetto di equità fiscale per il Comune di Roma). Gruppo-S Lab è una società di ricerca, sviluppo e innovazione con sede a Reggio Emilia e che collabora in maniera intensiva con le alcune Università (Milano, Roma e Pisa). La società è formata da personale interno, da manager alla guida delle divisioni produttive e da professionisti esterni ricercati tra le migliori figure del mondo accademico e scientifico. Gruppo-S ha realizzato progetti per diversi Enti Pubblici: Regione Veneto, Regione Emilia Romagna, Comune di Roma, Comune di Firenze, Ministero della Sanità e Provincia di Reggio Emilia. Inoltre, partecipa attivamente a un progetto di riscossione coattiva che coinvolge società di riscossione come SORIS (Torino), GEMA (Foggia), AST (Imperia), SORI (Prato) e SAT (Settimo Torinese). Infine, la società collabora attivamente con la Comunità Europea all'interno del progetto inerente la farmacovigilanza (MED-ePHV).

L'esperienza di Gruppo-S nel settore pubblico, in particolare l'impegno in progetti fiscali che coinvolgono Enti pubblici e società di riscossione, ha fatto si che il prodotto offerto dalla società fosse il candidato ideale per una prima sperimentazione di integrazione dei dati ai fini fiscali.

5.2.1 IL PRODOTTO: RISORSA DATI

Il prodotto offerto da Gruppo-S Lab, chiamato Risorsa Dati, è in realtà un insieme di tecnologie consolidate (Oracle Database 10g, Oracle Application Server, Business Engine BPEL) sulle quali la società ha costruito un modulo proprietario (MIND) che attraverso tecniche di inferenza statistica permette l'integrazione dei dati. La gestione del processo di

integrazione dei dati è controllata attraverso tre interfacce con l'utente (cruscotto, console e esiti qualità) che in seguito verranno analizzate più nello specifico. In figura 4 è rappresentata l'architettura del prodotto Risorsa Dati. Il server che ospita il sistema è collocato presso il Comune di Firenze e gli attori coinvolti nel processo (Informatica Trentina e Gruppo-S Lab) vi possono accedere tramite una rete VPN messa a disposizione dal Comune stesso in seguito ad un accordo con il Comune di Trento.

Il processo di integrazione dei dati è costituito da una serie di fasi collegate tra loro (Figura 35). Il punto di partenza è costituito dalle basi dati di origine (BDS) che, a seguito di un'analisi della qualità dei dati ed a una normalizzazione dei tracciati, vengono trasferite in un ambiente temporaneo (TMP) per poi essere sottoposte ad integrazione (tramite il modulo MIND), per poi confluire nel database integrato finale (BDPI o Risorsa Dati).

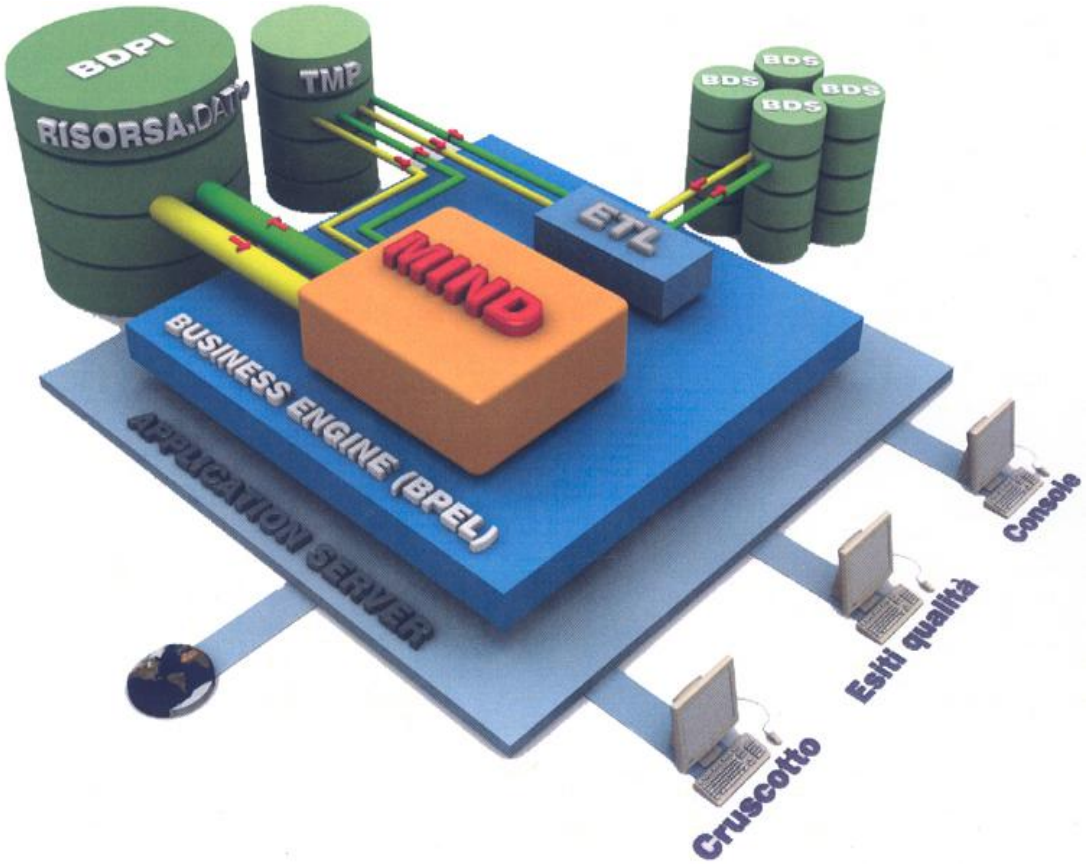

Figura 35 - Architettura del prodotto Risorsa Dati (Gruppo S Lab).

Come si vede nell'immagine il processo è bidirezionale, ovvero il dato integrato e pulito può percorrere a ritroso il percorso andando, qualora lo si desideri, a correggere l'informazione, errata o mancante, presente nella banca dati originale. In questo modo è possibile innescare un circolo virtuoso che porta al miglioramento della qualità dei dati sia nel database integrato a disposizione di Trentino Riscossioni sia all'Ente fornitore della banca dati originale. Nei paragrafi successivi vediamo nel dettaglio le singole fasi del processo di integrazione.

5.3 LE FASI DEL PROCESSO DI INTEGRAZIONE

In questo paragrafo viene analizzato il prodotto che Trentino Riscossioni, in collaborazione con Informatica Trentina, ha scelto di sperimentare per l'aggregazione dei dati relativi al Comune di Folgaria (i database elencati al paragrafo 5.1.1) mettendo in evidenza le fasi e i processi che sono stati necessari per realizzare una prima versione della base dati integrata.

5.3.1 OPERAZIONI PRELIMINARI

Prima di avviare il processo di integrazione dei dati, è stato necessario compiere delle operazioni preliminari. Innanzitutto è stato predisposto l'hardware per ospitare il sistema (server presso il Comune di Firenze, accessibile da remoto tramite rete VPN) ed un sistema di interscambio dei dati tra gli attori coinvolti (area di scambio dati sul portale Web di Trentino Riscossioni). In secondo luogo si è proceduto al recupero delle basi dati di partenza, da parte di Trentino Riscossioni e Informatica Trentina. Fase delicata in quanto è stato necessario il coinvolgimento di diversi Enti e società, in alcuni casi decisamente restie a fornire i propri dati provocando una dilatazione dei tempi. Recuperati i dati si è resa necessaria la realizzazione di una documentazione tecnica delle basi dati definendone i tracciati e le caratteristiche dei vari attributi che le compongono. In questa fase ho collaborato attivamente alla documentazione di alcune delle basi dati disponibili, elaborando, tramite query SQL, i flussi di variazioni per i vari anni da inviare a Gruppo-S secondo il formato concordato (record contraddistinti da: INSERT, nuovi dati, DELETE, dati da eliminare, UPDATE, dati variati). Eseguite queste operazioni preliminari i database, opportunamente elaborati, e i relativi tracciati sono stati caricati sul portale di Trentino Riscossioni e messi a disposizione di Gruppo-S per le fasi successive del processo. In Tabella 4 vediamo l'elenco delle basi dati recuperate e segnati in grigio gli anni per i quali sono disponibili (rappresentate dai contenitori BDS in Figura 35).

Banca dati	2004	2005	2006	2007	2008	2009
Anagrafe residenti	■	■	■	■	■	■
Toponomastica	■	■	■	■	■	■
Catasto Urbano	■	■	■	■	■	
Catasto Metrico		■	■	■	■	
Catasto Fondiario		■	■	■	■	
Piano Regolatore	■	■	■	■	■	■
Registro Imprese				■	■	■
Catasto Elettrico			■	■	■	
Catasto Acquedotto				■		
TARSU	■					
Versamenti ICI	■	■	■	■	■	

Tabella 4 - Disponibilità temporale dei database a disposizione di Trentino Riscossioni.

Per quanto riguarda la base dati "dichiarazioni ICI" è disponibile il censimento del 2003 che riporta la situazione agli anni 2001 e 2002.

5.3.2 OMOGENEIZZAZIONE E ANALISI DI QUALITÀ

Acquisiti gli archivi da Folgaria e da altre fonti (Catasto, Parix, Trenta, ecc.) si è proceduto alla loro normalizzazione e caricamento in ambiente omogeneo (TMP in Figura 35) attraverso tecniche ETL sviluppate da Gruppo-S tramite il linguaggio Business Process Execution Language (BPEL). Attraverso le operazioni di ETL i dati di origine sono quindi stati consolidati e inseriti in un ambiente unico omogeneo, in modo da renderli adatti alla fase di analisi di qualità e alla successiva fase di integrazione.

Una volta resi omogenei i dati si è proseguito con una disamina qualitativa dei contenuti per valutare il grado di correttezza, completezza e bontà dei dati a disposizione. Per ogni banca dati è stato prodotto un report dettagliato della situazione qualitativa e della completezza dei dati contenuti (con particolare enfasi agli attributi più significativi ai fini dell'integrazione). A titolo di esempio vediamo quale è la situazione di uno degli attributi più significativi ai fini del confronto di più banche dati: il codice fiscale (fondamentale ad esempio per confrontare i versamenti ICI con l'importo realmente dovuto determinabile dal catasto). In Tabella 5 è riportata la situazione dei codici fiscali all'interno delle banche dati che contengono informazioni sulle persone fisiche. Come si può vedere ci sono situazioni molto buone dove il codice fiscale è presente nella quasi totalità dei record, altre meno buone ma accettabili ed altre abbastanza problematiche (in particolare il fondiario non presenta quasi mai il codice fiscale dei soggetti fisici e non contiene nessuna partita IVA relativamente ai soggetti giuridici, inoltre il 17,5% dei record ha valorizzato soltanto nome e cognome, pertanto tali dati risultano praticamente non integrabili).

Banca dati	Totale record	Codice Fiscale Nullo	% di nulli
Anagrafe	5.273	1.495	28,35 %
Catasto Urbano	5.790	717	12,38 %
Catasto Metrico	14.921	1.386	9,29 %
Versamenti ICI	9.561	13	0,13 %
TARSU	6.804	0	0,00 %
Catasto Acquedotto	3.897	19	0,48 %
Fondiario	6.356	6.343	99,79 %

Tabella 5 - Analisi di qualità relativa ai codici fiscali.

In aggiunta, applicando gli indicatori di qualità descritti nel paragrafo 1.4.1, si è cercato di misurare la qualità delle diverse basi dati in modo puntuale per poter fornire una rappresentazione grafica dell'effettiva qualità dei dati (Figura 36). L'analisi si è basata

sulla misurazione di tipo "record-level" tenendo in considerazione gli attributi giudicati più importanti nel contesto applicativo della sperimentazione. Un'analisi di questo tipo ci da un'idea preliminare della qualità dei dati globale, ponendo l'attenzione sulle situazioni più problematiche dal punto di vista della qualità. In particolare, sono le basi dati catastali a presentare le situazioni qualitativamente peggiori, soprattutto per la ridotta presenza di codici fiscali e partite IVA nei flussi riguardanti persone fisiche e giuridiche. Nonostante alcune eccezioni, la situazione globale dei dati si è dimostrata sufficientemente buona ai fini di garantire un buon livello di precisione nell'integrazione automatica dei dati.

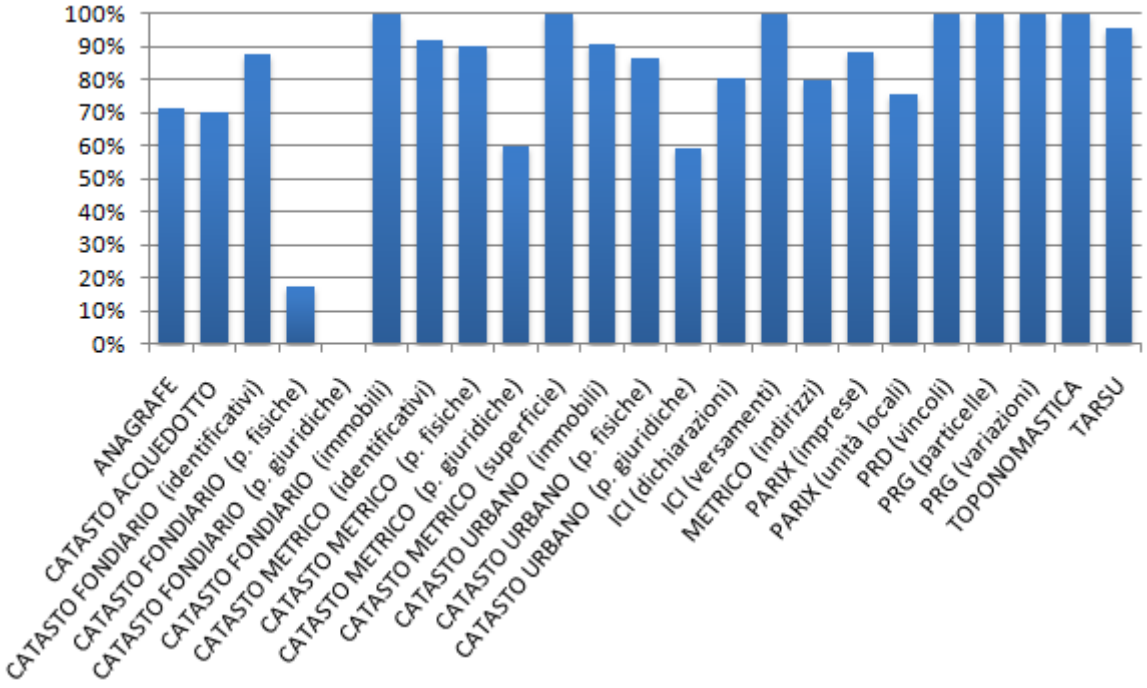

Figura 36 - Analisi qualitativa delle basi dati disponibili con riferimento all'anno 2004.

5.3.3 INTEGRAZIONE DEI DATI

Prima di procedere ad integrare i dati nell'ambiente finale è stato realizzato un primo schema logico della struttura del database integrato. In base alle esperienze precedenti di Gruppo-S si è definita una prima versione dello schema (Figura 37) che, pur non essendo quella definitiva, ha permesso di dare il via al caricamento delle prime basi dati nell'ambiente integrato. Lo schema presenta una serie di tabelle distinte secondo due tipologie: dimensioni e fatti. Le tabelle denominate "*dimensioni*" rappresentano le entità principali contenute nei database, ovvero persone, società e immobili. Le tabelle identificate come "*fatti*" contengono invece tutti gli atti, gli eventi, i rapporti che riguardano le dimensioni; ad esempio un versamento ICI è un fatto che collega persone e società con gli immobili, così come gli immobili sono collegati con persone e società da un certo tipo di rapporto (proprietario, usufruttuario, ecc.). In questo schema confluiscono tutti i dati relativi ad una determinata dimensione o fatto provenienti dai database di origine; ad

esempio la tabella *Persone* integra tutte le informazioni disponibili sulle persone fisiche provenienti dai diversi database (in questo caso anagrafe, catasto urbano, catasto metrico, versamenti e dichiarazioni ICI, TARSU, catasto acquedotto e catasto elettrico).

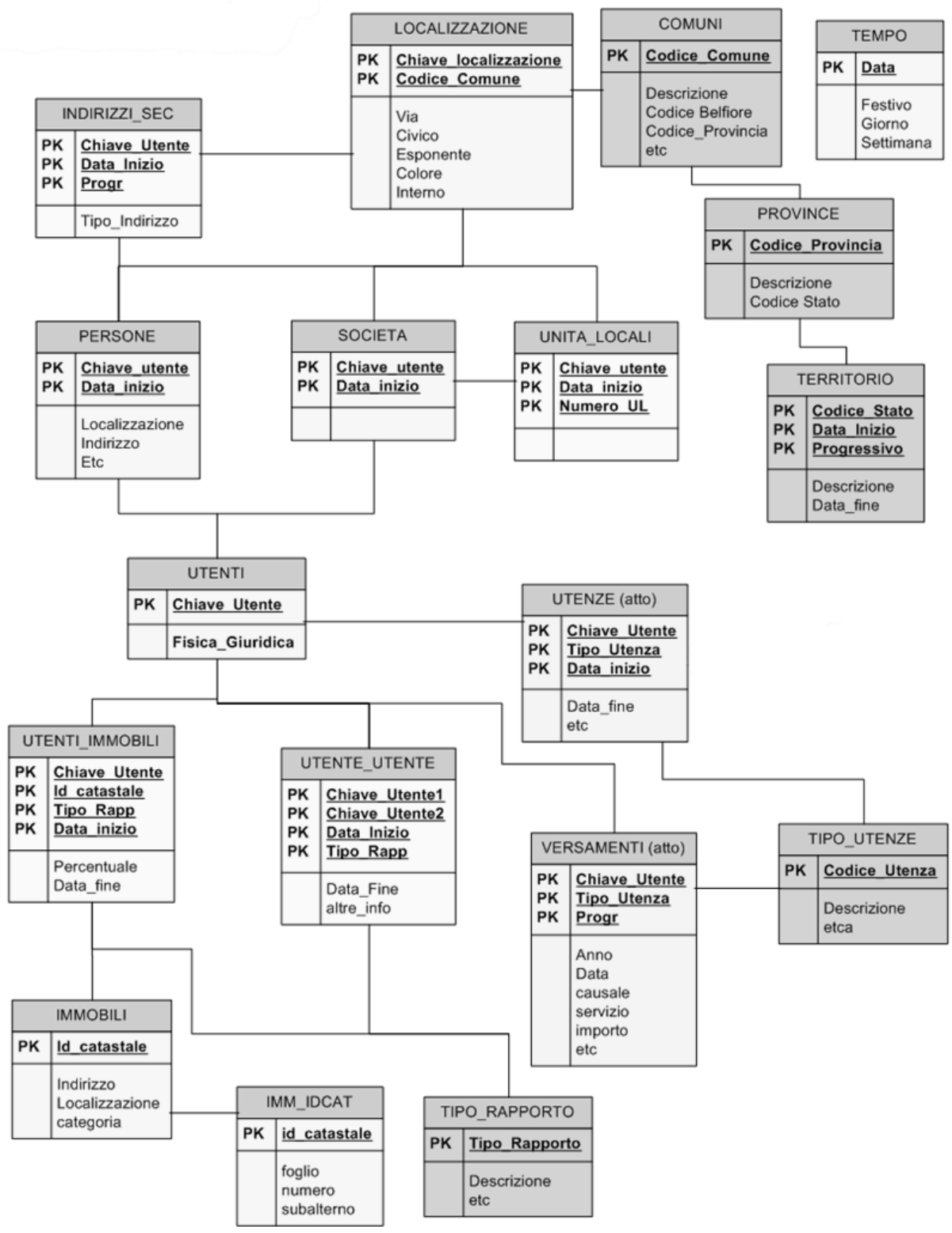

Figura 37 - Prima bozza dello schema del database integrato (BDPI).

L'integrazione dei dati viene effettuata dalla componente MIND, sviluppata da Gruppo-S, che utilizza tecniche di inferenza statistica per effettuare confronti tra informazioni presenti in database diversi. I criteri per eseguire l'integrazione vanno configurati manualmente tramite l'interfaccia utente definita *Console*. La Console è uno strumento applicativo destinato agli amministratori del sistema che permette di automatizzare e velocizzare tutta una serie di attività comunque gestibili attraverso l'accesso diretto al database e l'esecuzione di query. L'automazione delle attività più complesse e ripetitive permette di semplificare l'utilizzo dello strumento di integrazione. La Console è in realtà un insieme di più applicazioni, aggregate in modo da poter svolgere tramite un'interfaccia unica tutte le attività di gestione del sistema. Tali applicazioni sono:

- **BPEL Console**: consente di definire e gestire le regole per il caricamento e l'omogeneizzazione dei database durante la fase di ETL;
- **Oracle 10g Enterprise Manager**: è la console di gestione del database Oracle;
- **Mind Configurer**: sviluppato da Gruppo S Lab, è la componente chiave del processo di aggregazione. Permette di definire le regole di integrazione dei dati provenienti da database differenti;
- **Scheduler**: per di pianificare le attività di gestione, trasformazione, integrazione dei dati;
- **Application Server Manager**: permette di configurare l'application server Oracle;
- **Gestione Utenti e Gestione Indicatori**: permettono di gestire gli accessi al sistema e di definire gli indicatori di integrazione (interfaccia *Cruscotto*).

Ai fini dell'integrazione è interessante valutare più nello specifico la componente *Mind Configurer* per capire come vengono definite le regole per integrare i dati. Prima del caricamento di una qualsiasi base dati vanno definite le regole che governano il processo di popolamento del database integrato finale.

Le prime banche dati ad essere caricate nel database sono solitamente la Toponomastica e l'Anagrafe residenti. Essendo la toponomastica il primo flusso di dati da inserire nel database integrato non è stato necessario definire delle regole di integrazione, in quanto il database di destinazione era ancora vuoto. Il caricamento della toponomastica ha quindi determinato il popolamento della dimensione *Localizzazione* in BDPI. Per il caricamento successivo, Anagrafe, è stato necessario definire specifiche regole di integrazione in quanto viene eseguito un confronto degli indirizzi presenti in anagrafe con quelli caricati in Localizzazione (derivanti dalla toponomastica). Inoltre, il caricamento dell'anagrafe ha determinato il popolamento della dimensione *Persone*.

Vediamo come si presenta l'interfaccia di *Mind Configurer* per la gestione delle regole di popolamento del database integrato. Innanzitutto è possibile definire delle "distanze" in base ai quali il sistema identificherà ogni record come "match" o come "non match",

ovvero come uguale o diverso da un record già esistente. Il sistema opera su due passaggi quindi vengono definite le soglie sia per il primo che per il secondo passaggio (Figura 38).

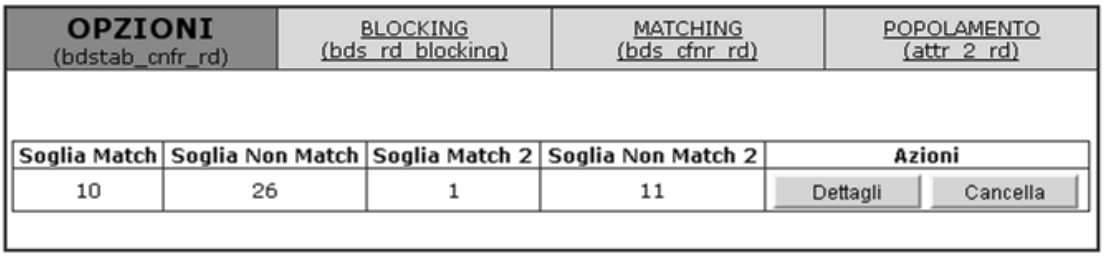

Figura 38 - *Interfaccia Mind Configurer*: gestione soglie di match.

A questo punto si selezionano gli attributi su cui eseguire il confronto tra le due banche dati In Figura 39 sono riportati gli attributi utilizzati per confrontare le persone fisiche presenti nel catasto urbano con le informazioni presenti nella dimensione *Persone* del database integrato.

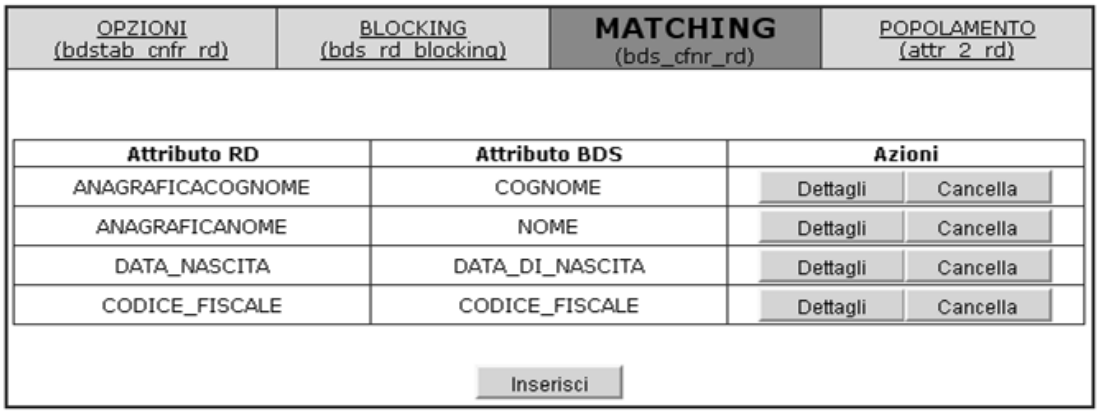

Figura 39 - *Interfaccia Mind Configurer*: selezione attributi su cui effettuare il confronto tra due banche dati.

Per ogni attributo è quindi possibile definire le regole di integrazione (Figura 40), che permettono al sistema di calcolare automaticamente la distanza tra due record e di prendere una decisione in base ai valori impostati in Figura 38.

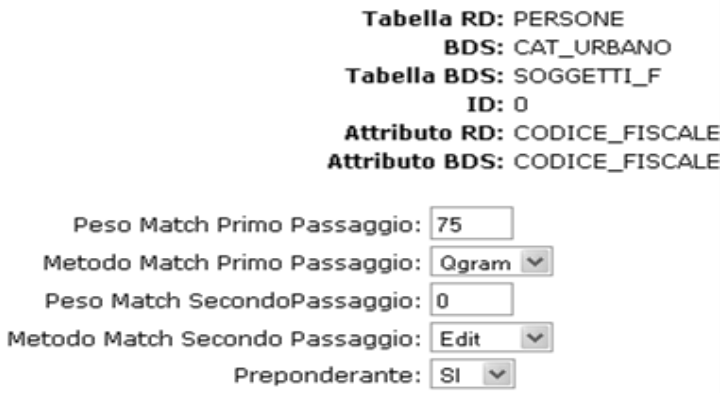

Figura 40 - *Interfaccia Mind Configurer*: definizione delle regole di integrazione per un singolo attributo.

Se la distanza tra due record di persone fisiche risulta minore di 10 i due record vengono identificati come la stessa persona e integrati in un'entità unica, in caso contrario vengono identificati come due persone diverse e restano due entità separate. L'interfaccia per la gestione delle regole di integrazione permette di definire i pesi dei singoli attributi (per i due passaggi) e l'algoritmo di match da utilizzare[7]. Il peso di match è un parametro fondamentale in quanto specifica in che misura si ritiene attendibile un dato presente in una base dati di origine. Per uno stesso attributo possono esistere infatti più banche dati che contengono tale informazione e il record risulta quindi "conteso" da più BDS. Quando il valore è lo stesso non si pone nessun problema; quando invece i valori sono differenti deve essere applicato un criterio di scelta per permettere al sistema di decidere automaticamente quale valore utilizzare. Nelle banche dati utilizzate nella sperimentazione, ad esempio, ci sono molti record contesi relativi alle persone fisiche. Solitamente si utilizza l'approccio che prevede la presenza di una banca dati dominante (in questo caso l'anagrafe in quanto i suoi dati sono certificati) le cui informazioni hanno un peso più elevato rispetto a quelle contenute in altre basi dati. Se non fosse possibile definire una banca dati dominante occorre valutare attentamente, caso per caso, a quale informazione assegnare un peso maggiore. In Figura 40 sono riportate le regole di integrazione riguardanti il codice fiscale, in questo caso viene dato maggior peso al codice fiscale presente nella dimensione *Persone* rispetto a quanto riportato in catasto, in quanto la dimensione *Persone* contiene i dati riportati in Anagrafe. Nel corso della sperimentazione la definizione delle regole di integrazione si è basata principalmente sull'esperienza maturata nel settore da Gruppo-S, tenendo chiaramente in considerazione anche la qualità delle diverse banche dati a disposizione di Trentino Riscossioni.

Seguendo questo procedimento sono state man mano caricate tutte le banche dati di origine nell'ambiente finale integrato (BDPI) andando di volta in volta a confrontare e integrare le informazioni contenute in una banca dati con quelle già presenti in ambiente BDPI. Queste serie di confronti automatici originano tre tipologie di record:

- **Record risolti**: ovvero quei record per cui il sistema è in grado di prendere automaticamente una decisione. Il sistema distingue tali record come:
    - **Match**: riconosce che due o più record si riferiscono alla stessa persona, oggetto o evento e li integra quindi come un'entità unica;
    - **Non-match**: riconosce che due o più record non si riferiscono alla stessa persona, oggetto o evento e li riporta come entità diverse;
- **Record con problemi:** sono quei record che non hanno valorizzati gli attributi di match (Figura 39) o che contengono errori grossolani su tali attributi, pertanto il

[7] È possibile scegliere diversi algoritmi che utilizzano metodi matematico/statistici differenti. Alcuni sono più indicati per valori testuali, altri per campi numerici.

sistema non è in grado di confrontarli e vengono automaticamente scartati escludendoli dall'ambiente finale integrato;

- **Record indecisi:** sono quei record per i quali il sistema non è in grado di prendere automaticamente una decisione certa ed è quindi necessario risolverli manualmente.

Queste tre tipologie di risultati sono osservabili e gestibili tramite l'interfaccia utente "Esiti Qualità". In particolare attraverso tale interfaccia è possibile avere una lista dei *record scartati* dal sistema per capire quali problemi presentano e se tali problemi sono eventualmente risolvibili manualmente, oppure se si tratta di record da scartare definitivamente data la scarsa qualità dei dati contenuti. Possiamo inoltre avere visione dei *record risolti*, per capire come sono stati integrati e quali informazioni sono andati a ricevere o a sovrascrivere (in base al peso attribuitogli in precedenza). Infine, è possibile visionare i *record indecisi*, per i quali il sistema offre la possibilità di prendere manualmente una decisione (Figura 41). Per ogni record indeciso il sistema propone la distanza calcolata tra i due record, maggiore è la distanza più alta è la probabilità che siano effettivamente due persone diverse, si tratta comunque di casi da valutare singolarmente attraverso verifiche manuali.

	ANAGRAFE					PERSONE			
match	NOME	DATA_NASCITA	COGNOME	CODICE_FISCALE	distanza	CODICE_FISCALE	ANAGRAFICACOGNOME	DATA_NASCITA	ANAGRAFICANOME
Indeciso Conferma	SERGIO	22/10/1938	CUEL	null	8	[illegible]	CUEL	22-MAY-51	SERGIO
No Match Conferma	LAURA	26/09/1967	CAPPELLETTI	null	10	[illegible]	CAPPELLETTI	18-MAR-20	LAURA
Match Conferma	MARIA	01/11/1919	BUZZI	null	6	[illegible]	BUZZI	11-NOV-07	MARIA

Figura 41 - *Interfaccia Esiti Qualità*: gestione dei record indecisi.

I record indecisi si sono rivelati per la maggior parte persone fisiche (circa 500 indecisioni) ed in numero minore persone giuridiche (circa 20 indecisioni). Considerando la quantità di record integrati le posizioni indecise si sono rivelate un numero decisamente contenuto. Tuttavia, la risoluzione manuale dei record indecisi è un'attività che consente di migliorare ulteriormente la qualità del dato integrato e per questo si è deciso di intervenire per risolvere tali indecisioni. Nel corso dello stage mi sono occupato personalmente della risoluzione dei record indecisi effettuando delle interrogazioni puntuali sulla base dati dell'Anagrafe Tributaria[8]. Il completamento di questo processo ha contribuito ad arricchire ulteriormente la base dati integrata, riducendo ulteriormente il numero di entità presenti (vedi Tabella 6 nel paragrafo 5.4.1).

[8] L'Anagrafe Tributaria è consultabile attraverso il sistema informativo SIATEL del Ministero delle Finanze (https://siatel.finanze.it). Il sistema permette il controllo dei dati fiscali presenti nelle dichiarazioni di autocertificazione dei contribuenti. Il servizio è stato utilizzato per effettuare controlli sui codici fiscali ed interrogazioni di anagrafiche.

5.4 RISULTATI

I risultati del processo di integrazione si possono valutare secondo punti di vista differenti. Innanzitutto, il sistema Risorsa Dati presenta delle interfacce grafiche che sintetizzano la qualità del processo di integrazione attraverso alcuni indicatori (percentuale di record risolti automaticamente, riduzione della numerosità dei record, ecc.). In secondo luogo, il sistema è strutturato per tenere traccia di tutte le bonifiche automatiche e manuali effettuate sui dati, in modo da poter estrarre facilmente dei flussi di dati bonificati da restituire all'ente come ritorno del processo di integrazione. Questo innesca il circolo virtuoso tra Enti e Trentino Riscossioni con la conseguenza di portare ad una migliore qualità del dato sia nei database degli Enti che nel Data Warehouse di Trentino Riscossioni. Infine, la bontà del processo di integrazione è stata valutata in un particolare contesto operativo: l'accertamento ICI. In seguito alla definizione delle variabili e delle procedure che governano tale dominio è stata realizzata una procedura di analisi automatica dei dati in grado di restituire una lista di contribuenti identificabili come potenziali evasori. Il confronto tra procedura automatica e attività manuali di verifica ha permesso di affermare la bontà del lavoro di integrazione svolto. Vediamo quindi nei dettagli come sono stati valutati i risultati della sperimentazione.

5.4.1 INDICATORI DI QUALITÀ DEL PROCESSO DI INTEGRAZIONE

Il sistema Risorsa Dati dispone di un'interfaccia utente (*Cruscotto*) in grado di sintetizzare, attraverso indicatori grafici, l'esito del processo di integrazione. In Figura 42 vediamo un esempio dell'interfaccia che sintetizza, relativamente al Catasto Urbano, la percentuale dei record integrati automaticamente dal sistema sul totale dei record presenti.

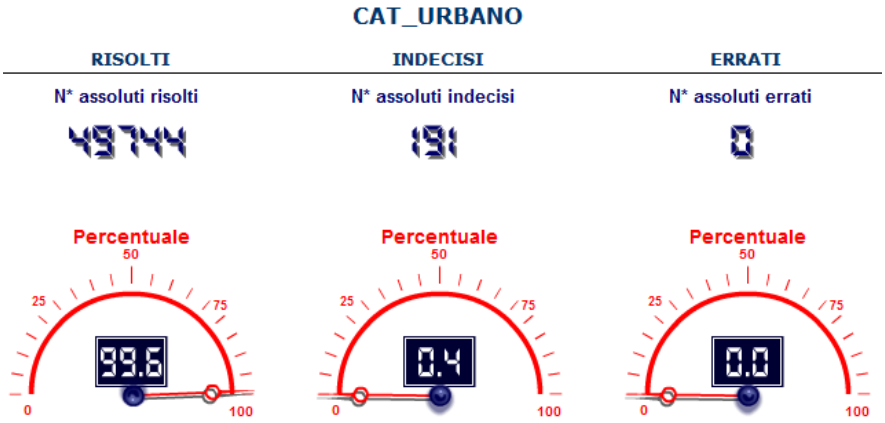

Figura 42 - *Interfaccia Cruscotto*: visualizzazione grafica dei risultati dell'integrazione

In questo caso si è raggiunto un livello di integrazione del 99,6%, ovvero il 99,6% dei record è stato integrato in maniera automatica dal sistema, mentre il restante 0,4% è la

parte di record con problemi o indecisi (in questo caso tutti indecisi). Secondo la valutazione di Gruppo-S, basata sulle loro esperienze precedenti, il livello di integrazione raggiunto con i dati del comune di Folgaria è molto buono; questo sta a significare che la qualità dei dati nelle basi dati di origine era sufficiente per garantire un processo di integrazione automatica dei dati soddisfacente. Un'ulteriore conferma dell'efficacia del processo di integrazione arriva anche dall'analisi del numero di record integrati in una entità unica. La Tabella 6 riporta la situazione relativa a persone fisiche, persone giuridiche e immobili. La colonna "*basi dati sorgenti*" rappresenta la somma di tutti record, relativi a persone fisiche, persone giuridiche e immobili contenuti nelle basi dati di origine. La colonna "*base dati integrata*" rappresenta invece il totale di record ottenuto a seguito dell'integrazione dei dati.

Esiti Integrazione	**Numerosità**	
	Basi dati sorgenti	*Base dati integrata*
Persone fisiche	173.328	14.266
Persone giuridiche	134.067	63.417
Immobili	78.936	10.006

Tabella 6 – Numero di complessivo di record e numero di entità ottenute dopo l'integrazione.

Come si vede in Tabella 6 per le persone giuridiche la riduzione di numerosità è minore; tale dato si spiega con il fatto che sono presenti le società di tutto il Trentino, mentre persone fisiche e immobili sono relativi al solo Comune di Folgaria.

Questi dati testimoniano la bontà del processo di integrazione, che ha permesso di inserire tutte le informazioni su una determinata persona, società o immobile in una sola anagrafica. Questo facilita enormemente l'attività di accertamento in quanto permette al personale di evitare una laboriosa e dispendiosa operazione manuale di ricostruzione dell'anagrafica raccogliendo le informazioni necessarie dai diversi contenitori. Grazie alla base dati integrata il personale ha quindi a disposizione un'anagrafica unica e aggiornata della persona, società o immobile che intende indagare e può concentrare le proprie energie su fasi più importanti del processo di accertamento.

5.4.2 BONIFICA DEI DATI E RITORNO AGLI ENTI

Un altro indicatore utile per valutare i risultati del processo di integrazione è costituito dal miglioramento della qualità dei dati, processo che permette di restituire agli Enti, fornitori delle diverse basi dati, un flusso di ritorno con dati aggiornati e bonificati. Ad esempio, in Tabella 7 è riportata la situazione relativa ai codici fiscali. Come si può vedere il processo di integrazione ha permesso il recupero automatico di una discreta quantità di codici fiscali

(globalmente circa il 37% dei codici fiscali nulli sono stati recuperati in seguito all'integrazione, nel caso del Catasto Acquedotto il recupero è stato quasi totale). Risorsa Dati permette di ricostruire questi flussi di ritorno garantendo così all'ente fornitore la possibilità di procedere ad eventuali bonifiche della base dati ad un costo praticamente nullo. Oltre al recupero di codici fiscali è stato possibile bonificare in maniera automatica anche altre informazioni, in particolare sono stati prodotti i seguenti ritorni: dati errati nella toponomastica (vie doppie e civici multipli segnalati al Comune), aziende riportate in Parix con dati del tutto errati o mancanti (segnalate alla Camera di Commercio di Trento), dati catastali non consistenti (segnalati al Catasto anche se una bonifica dei dati catastali è un'operazione che presenta delle complicazioni come si è accennato nel paragrafo 5.1.1).

Codici Fiscali recuperati automaticamente grazie all'integrazione				
Banca Dati	*Nulli Prima*	*Nulli Dopo*	*Differenza*	*% Recupero*
ICI	13	8	5	38,46 %
Catasto Urbano	717	329	388	54,11 %
Anagrafe	1495	1061	434	29,03 %
Catasto Acquedotto	19	1	18	94,74 %
TOTALE	**2244**	**1399**	**845**	**37,66 %**

Tabella 7 - Codici fiscali recuperati in seguito al processo di integrazione.

5.4.3 ANALISI DEI DATI AI FINI DELL'ACCERTAMENTO I.C.I.

Una volta determinata la qualità dell'integrazione tramite gli indicatori precedenti si è deciso si valutare il valore aggiunto del database integrato attraverso l'analisi di uno specifico contesto tributario: l'accertamento dell'imposta comunale sugli immobili (ICI). Con l'affidamento, da parte del Comune di Folgaria, delle attività di accertamento ICI il personale addetto di Trentino Riscossioni ha iniziato una complessa attività di analisi manuale dei dati a disposizione con l'obiettivo di identificare i maggiori casi di evasione. La trasposizione di questo tipo di attività in una procedura automatica ha richiesto una fase di studio delle metodologie di lavoro tradizionali, fase della quale mi sono occupato personalmente nel periodo dello stage trascorso presso Trentino Riscossioni. Lo studio e l'esplicitazione della metodologia di accertamento ha richiesto circa un mese di lavoro a stretto contatto con il personale che si occupa dell'accertamento tributario. L'output di tale attività è stato una documentazione schematica delle varie fasi del processo, mettendo particolare evidenza sugli aspetti più critici ovvero quella parte di informazioni frutto del ragionamento umano molto difficili da replicare in una procedura di elaborazione automatica.

Le procedure manuali presentano una serie di limiti che rendono l'attività di accertamento complicata e talvolta addirittura controproducente (in un caso specifico in un mese di lavoro non è stata trovata nemmeno una posizione di evasione, rendendo di fatto l'attività un ulteriore costo per l'azienda e per l'ente affidante anziché una fonte di guadagno). L'obiettivo di una procedura di analisi automatica dei dati integrati è di segnalare all'accertatore quali sono le posizioni da controllare e quali invece risultano corrette, in questo modo il personale di Trentino Riscossioni può concentrare la propria attività sull'analisi dei casi più sospetti senza perdere tempo ad analizzare le posizioni dei contribuenti che pagano il giusto.

L'algoritmo di calcolo sviluppato è piuttosto complesso e la logica della procedura è frutto della collaborazione tra tutti i soggetti coinvolti nel progetto e prevede l'elaborazione delle informazioni derivanti dalla quasi totalità delle banche dati integrate. Quello che si vuole dimostrare attraverso tale algoritmo è come l'utilizzo di dati integrati permetta di raggiungere un certo tipo di risultato, un risultato ben diverso e migliore rispetto all'analisi della singola base dati. L'algoritmo di calcolo si sviluppa secondo la seguente logica:

1. Elaborazione dei dati catastali relativi ai fabbricati e alle titolarità secondo una logica giornaliera, aggregata poi in una logica mensile (il calcolo dell'ICI va fatto su base mensile). Tale logica permette di ricostruire una sequenza di 12 fotogrammi che descrivono le caratteristiche dell'immobile nel corso dell'anno preso in esame[9];
2. Elaborazione dei dati relativi alle persone fisiche e giuridiche che hanno un rapporto con il fabbricato valido ai fini ICI. Mentre nel primo punto la base dati coinvolta era solo quella catastale, in questo caso le informazioni provengono da una moltitudine di fonti differenti: Anagrafe, Parix, Catasto Acquedotto, Catasto Urbano, Catasto Elettrico, Archivi ICI e TARSU, Catasto Metrico. Un'operazione di questo tipo ha permesso di ricostruire il profilo di persone e aziende che anche quando l'informazione presente in un determinato archivio era solo parziale;
3. Esecuzione di una serie di controlli di consistenza delle informazioni che portano alla validazione della singola posizione o alla generazione di eccezioni non gestibili dall'algoritmo;
4. Determinazione del fabbricato identificabile come abitazione principale. La procedura, grazie alle informazioni sulla residenza del soggetto fisico, ricavate al punto 2, confronta tale informazione con l'indirizzo dell'immobile da catasto per determinare quale fabbricato attribuire al contribuente come abitazione principale.

[9] È possibile accertare l'ICI fino a 5 anni indietro nel tempo, pertanto la sperimentazione si focalizza sull'analisi del periodo 2004-2008.

Una volta determinata l'abitazione principale l'algoritmo passa ad individuare le eventuali pertinenze[10];

5. Calcolo del dovuto per immobile su base mensile e aggregazione per soggetto. La posizione mensile viene quindi ricondotta ad un dovuto annuale tenendo conto di eventuali detrazioni;
6. Confronto del dovuto annuale con l'importo effettivamente versato dal contribuente.

La procedura termina quindi restituendo tre possibili risultati:

1. *Messaggio di errore*: a causa della mancanza di dati essenziali per il calcolo della posizione contributiva la procedura non è in grado di fornire un output per tale posizione;
2. *Un importo da versare con avvertimento*: la procedura riesce a calcolare la posizione ma ci sono alcuni dati non congruenti che vengono segnalati all'operatore (ad esempio non è stato possibile determinare l'abitazione principale del contribuente);
3. *Un importo da versare senza avvertimento*: la procedura è andata a buon fine e segnala all'operatore la differenza tra importo dovuto ed importo versato per ogni posizione.

In definitiva, la procedura è in grado di restituire un valore esatto del dovuto nei seguenti casi:

- Aziende proprietarie di immobili con rendita e categoria catastale presente in catasto;
- Non residenti proprietari di immobili con rendita e categoria attribuite in catasto;
- Residenti che non hanno abitazioni principali;
- Residenti con un solo fabbricato situato nell'indirizzo di residenza.

In tutti gli altri casi la procedura è comunque in grado di calcolare l'importo dovuto ma non sempre restituisce un valore esatto, i casi sospetti vengono comunque segnalati all'operatore dell'accertamento, starà a lui determinare l'esatto importo da accertare.

L'esecuzione dell'algoritmo ha restituito un elenco di posizioni contributive per le quali evidenzia eventuali differenze tra l'importo dovuto (determinato in base all'elaborazione dei dati catastali) e l'imposta effettivamente versata. In Tabella 8 vediamo un riepilogo dei

[10] L'articolo 817 del codice civile definisce le pertinenze come le "*cose destinate in modo durevole al servizio o ad ornamento di un'altra cosa*". Nel contesto ICI alla pertinenza deve essere riservato lo stesso trattamento fiscale dell'abitazione principale. Pertanto la pertinenza è assoggettabile all'aliquota applicata all'abitazione principale e beneficia di una detrazione d'imposta qualora l'importo della detrazione (130 euro per il Comune di Folgaria) non si esaurisca nell'abitazione principale. Possono essere identificate come pertinenze le seguenti soltanto le seguenti categorie catastali: C/2 (magazzini e locali deposito), C/6 (autorimesse), C/7 (tettoie chiuse o aperte).

risultati dell'elaborazione, dove sono stati definiti degli intervalli di differenza dovuto-versato per i quali è riportato il numero di posizioni riscontrate. In totale le posizioni soggette all'imposta nel Comune di Folgaria sono 5.579. Per 364 di queste posizioni la procedura non è stata in grado di calcolare un dovuto per una serie di cause: mancata consistenza dei dati delle quote o dei mesi di possesso, fabbricati senza rendita attribuita, posizioni con più di un'abitazione principale attribuita. Le restanti 5.215 posizioni sono state elaborate dalla procedura, in particolare 4.049 posizioni sono state calcolate correttamente e senza segnalazioni di avvertimento mentre le restanti 1.166 sono posizioni per le quali è stato calcolato un importo dovuto ma la procedura non è stata in grado di determinare la presenza o meno di un fabbricato di tipo prima casa.

Numerosità	**Range**
1661	da 0 a 10
186	da 10 a 20
116	da 20 a 30
93	da 30 a 40
87	da 40 a 50
88	da 50 a 60
79	da 60 a 70
67	da 70 a 80
54	da 80 a 90
75	da 90 a 100
652	da 100 a 200
394	da 200 a 300
135	da 300 a 400
75	da 400 a 500
39	da 500 a 600
38	da 600 a 700
24	da 700 a 800
16	da 800 a 900
17	da 900 a 1000
55	da 1000 a 2000
26	Oltre 2000
3977	**Totale ≥ 0**

Numerosità	**Range**
716	da -10 a 0
71	da -20 a -10
67	da -30 a -20
48	da -40 a -30
33	da -50 a -40
20	da -60 a -50
41	da -70 a -60
16	da -80 a -70
17	da -90 a -80
17	da -100 a -90
92	da -200 a -100
36	da -300 a -200
17	da -400 a -300
13	da -500 a -400
6	da -600 a -500
3	da -700 a -600
1	da -800 a -700
3	da -900 a -800
4	da -1000 a -900
10	da -2000 a -1000
7	meno di -2000
1238	**Totale < 0**

Tabella 8 - Risultati della procedura di accertamento automatico.

Dall'analisi dei risultati riportati in Tabella 8 si può osservare come il 55% dei contribuenti (2.377 posizioni) sia compreso nell'intervallo tra -10 e +10 euro; si può quindi affermare che metà dei contribuenti pagano l'importo corretto (l'accertamento si può inviare solo per posizioni di evasione/elusione maggiori di 12 euro). Questo è già un risultato importante in quanto consente al personale che si occupa dell'attività di accertamento di tralasciare il controllo di tali posizioni per concentrarsi su coloro che non pagano il giusto. Tuttavia, dall'analisi dei dati emerge un altro risultato, come si vede in Tabella 8 il range tra 100 e 400 euro presenta infatti una numerosità elevata (1.181 posizioni). Secondo gli esperti di

accertamento buona parte di tali posizioni si può ricondurre a quei casi in cui la procedura non è stata in grado di determinare l'abitazione principale. In particolare, 682 dei 1.181 sono residenti a cui non è stata attribuita un'abitazione principale, è quindi altamente probabile che tali posizioni siano quasi tutte corrette. Una procedura di determinazione dell'abitazione principale più precisa permetterebbe quindi di catalogare potenzialmente come corrette circa il 59% delle posizioni contributive del comune (3.558 posizioni), alleggerendo notevolmente il carico del lavoro di accertamento manuale. Le medesime considerazioni sono valide per gli importi negativi, si tratta probabilmente di posizioni a cui è stata attribuita erroneamente un'abitazione principale.

In seguito all'analisi dell'output della procedura il responsabile dell'attività di accertamento ha confermato la validità e l'utilità del risultato ottenuto, inoltre, il raggiungimento di tale traguardo è stato giudicato come un risultato rilevante per la sperimentazione da parte del consiglio di amministrazione di Trentino Riscossioni.

5.5 PUNTI DEBOLI E CRITICITÀ

Le fasi del processo di integrazione, descritte nel paragrafo 5.3, sono terminate recentemente (di fatto la sperimentazione si è conclusa nel novembre 2009) e tutti i dati recuperati da Trentino Riscossioni sono stati integrati in un ambiente unico e storicizzato (per quanto possibile) il cui schema finale è riportato in Figura 43.

Arrivati a questo punto è opportuno soffermarsi su quelli che appaiono i punti deboli del prodotto Risorsa Dati e su quelle che saranno le criticità future legate all'utilizzo di un sistema di questo tipo (Data Warehouse). Dopo alcuni mesi di lavoro sono stati infatti individuati i principali limiti del prodotto Risorsa Dati e le criticità che la realizzazione di un Data Warehouse tradizionale implicherebbe nel contesto di Trentino Riscossioni.

Riguardo Risorsa Dati si può dire che il sistema presenta delle interfacce utente davvero basilari e scarsamente personalizzabili, se non manualmente tramite la modifica del codice sottostante. Questo rappresenta un problema in quanto è necessario avere sempre a disposizione una figura esperta di Gruppo-S che sia in grado di risolvere eventuali problematiche o di fornire determinate elaborazioni. D'altra parte Gruppo-S sta lavorando sulla nuova interfaccia di gestione del prodotto disponibile a partire dal 2010, un prima dimostrazione della nuova interfaccia ha dato prova delle maggiori potenzialità del nuovo sistema di navigazione e gestione dei dati. Un altro punto debole è la scarsa conoscenza da parte di Trentino Riscossioni dello schema finale della base dati integrata, soprattutto delle modalità di gestione dello storico dei dati. Tuttavia, si tratta di una problematica minore, risolvibile grazie ad un maggiore supporto di Gruppo-S nella fase di creazione delle prime query.

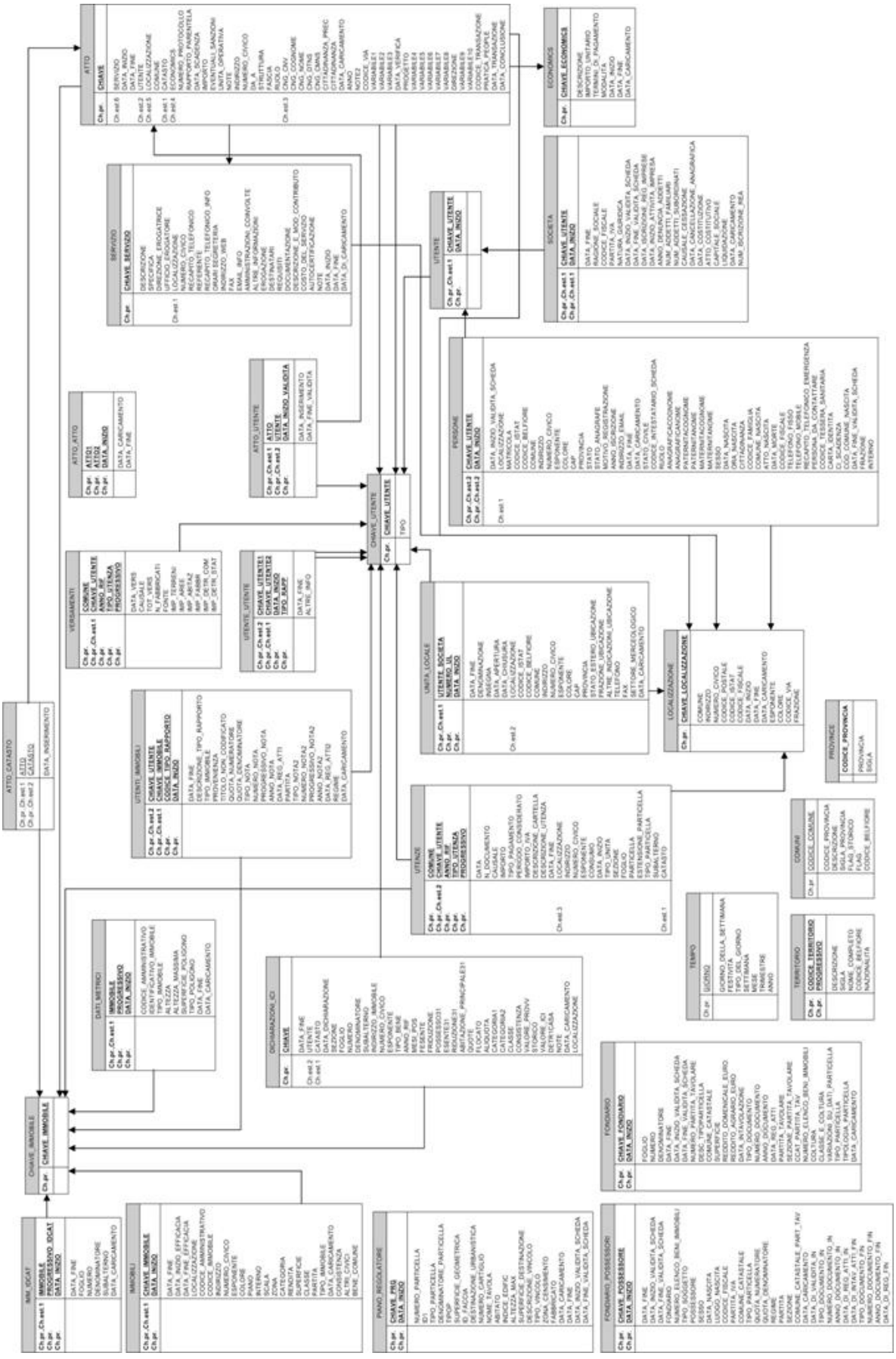

Figura 43 - Lo schema finale della base dati integrata.

Riguardo alle criticità che si presenterebbero in caso di un'espansione futura del sistema ad altri comuni o ad altre basi dati va senza dubbio segnalata la scarsa scalabilità del sistema. Con i processi e le metodologie attuali gestire i 223 comuni trentini sarebbe decisamente improponibile. Occorre definire una nuova componente del sistema che costituisca un'interfaccia tra il sistema e i comuni, il processo attuale di caricamento dei dati non è infatti sostenibile con oltre 200 comuni. All'interno della sperimentazione il caricamento dei dati è stata un'attività manuale di trasferimento delle singole basi dati tramite l'area scambio dati del portale di Trentino Riscossioni, successivamente il personale di Informatica Trentina o Gruppo-S provvede alla normalizzazione del tracciato, al calcolo delle variazioni, ecc. In un sistema a regime con oltre 200 comuni non sarebbe sicuramente una situazione sostenibile, va pensato un processo di caricamento dei nuovi flussi di dati almeno semi-automatizzato.

Infine, ci si scontra con le problematiche tipiche dei database relazionali e dei Data Warehouse: principalmente la scarsa flessibilità del sistema, aggravata in questo caso dal contesto altamente variabile dei tributi. Un progetto di questo tipo necessita di un sistema molto flessibile che permetta di adattare la propria struttura in seguito ad eventuali riforme del contesto tributario di riferimento o all'entrata di nuovi flussi informativi. Inoltre, il coinvolgimento di oltre 200 comuni determinerà la presenza di un'altissima eterogeneità di formati e strutture di dati. Ogni comune possiede sistemi informativi differenti e gli stessi dati possono risultare gestiti in modi diversi, con flussi e attributi differenti. L'eterogeneità dei formati dei dati costituisce una componente rilevante in termini di tempo e di costo, in quanto richiede la definizione di trasformazioni dei dati ad hoc per ogni comune, richiedendo l'intervento di un esperto delle procedure ETL. Una soluzione più flessibile dovrebbe garantire la possibilità di mappare i dati dalla sorgente alla fonte anche agli utenti meno esperti del sistema.

A fronte di queste problematiche Trentino Riscossioni ha deciso di intraprendere una sperimentazione parallela all'attuale, sperimentando l'utilizzo di tecnologie più recenti e innovative che si pongono l'obiettivo di superare i limiti tipici dei sistemi relazionali, garantendo al sistema maggiore flessibilità e adattabilità a situazioni nuove che rappresenterebbero un notevole valore aggiunto per sistemi operanti in contesti incerti come quello di Trentino Riscossioni. Nel capitolo 6 viene introdotta brevemente la nuova sperimentazione intrapresa.

5.6 I COSTI DELLA SOLUZIONE ADOTTATA

Ai fini di valutare la reale possibilità di'utilizzo in un ambiente di produzione della soluzione sperimentata sino ad ora è importante analizzare anche i costi di un approccio di questo tipo, tenendo in considerazione l'alta eterogeneità delle sorgenti informative da

gestire Per una realizzazione di un progetto di integrazione ICT basata sul prodotto e sulla metodologia utilizzata nella sperimentazione presentata implica diverse tipologie di costo:

- **Costi di infrastruttura**: la realizzazione di un'infrastruttura centrale per la memorizzazione e l'elaborazione dei dati costituisce la base per la realizzazione di un progetto di questo tipo. Nel contesto di Trentino Riscossioni, e più in generale della Provincia Autonoma di Trento, la realizzazione e la gestione di tale infrastruttura è in carico ad Informatica Trentina che già possiede un data center attrezzato. Nel dettaglio si può prevedere una infrastruttura scalabile multinodo, che cresce al crescere del carico. Il costo dell'infrastruttura e della gestione del servizio di data center per un numero limitato di comuni (60 comuni) di dimensione media (3-4000 abitanti) è stimabile tra i 50.000 e i 100.000 euro. Comunque, si tratta di un costo che cresce linearmente in base all'aggiunta di nuovi nodi e quindi prevedibile con una certa facilità;
- **Costi di licenza**: il costo per due licenze del componente cuore del sistema di incrocio dati MIND è stato di 36.000 euro. Il costo per la licenza del prodotto è ininfluente sul totale dei costi del progetto. Application server e database costituiscono altri costi fissi del progetto che aumentano al crescere dei dati;
- **Costo di integrazione nuovo comune**: per l'integrazione del primo comune sono stati stimati 117 giorni di lavoro, il che si traduce in un costo di circa 46.000 euro (stimando in 400 euro al giorno il costo medio di un consulente esterno). A partire da tale dato Gruppo-S stima il lavoro necessario per l'integrazione di 100 comuni in 31 giorni, nel caso la struttura dei dati sia simile, e in 89 giorni in caso di struttura profondamente diversa. Ipotizzando che in metà dei casi la struttura sia simile e nell'altra metà differente, i costi di integrazione di 100 comuni si possono stimare in circa 2 milioni di euro (attribuendo parte del lavoro al consulente esterno e parte al personale interno a Trentino Riscossioni). I tempi di integrazione dal centesimo comune in poi sono stimati in 28 giorni per comuni di medesima struttura e 81 giorni per comuni di diversa struttura;
- **Costi di sviluppo di applicativi**: si tratta dei costi di sviluppo di applicazioni per navigare, analizzare ed elaborare i dati. Il costo di soli strumenti di interrogazione è stimato in almeno 25.000, ai quali dovranno aggiungersi i costi per la realizzazione di strumenti di analisi più complessi, realizzati sulla base delle necessità degli utilizzatori finali. Tuttavia, al momento attuale è difficile dire quale sarà l'ammontare di questa categoria di costo;
- **Costi di gestione**: si tratta di tutti i costi da sostenere per la gestione e la manutenzione dell'applicativo di integrazione e della gestione del sistema informativo su cui è eseguito (server farm presso data center di Informatica Trentina). Per una stima di questi costi sarà necessario analizzare nei dettagli il

progetto del Comune di Firenze, con gli opportuni pesi, per capire quanto tale tipologia di costo incida sul totale del progetto. In tale senso occorre valutare il numero di giornate di personale specialistico necessario per anno considerando una stima di circa 30 giorni per anno ed un costo di circa 15.000-20.000 euro (in base alle attuali tariffe di Informatica Trentina).

Secondo questa prima analisi dei costi l'integrazione della totalità dei comuni trentini richiederebbe quindi un investimento di alcuni milioni di euro. Tuttavia, il modello di business da adottare non è ancora stato definito con precisione, Trentino Riscossioni, Informatica Trentina, Gruppo-S e il Comune di Firenze stanno lavorando al fine di definire un modello di business che soddisfi tutte le parti interessate, attraverso il quale sia possibile sfruttare le norme sul riuso nell'ambito della Pubblica Amministrazione (vedi paragrafo 4.3) senza per questo penalizzare la società fornitrice del prodotto. A titolo di esempio, il progetto intrapreso dal Comune di Firenze, che ha portato oggi all'integrazione di circa 60 basi dati e allo sviluppo di diverse applicazioni per la navigazione e l'analisi del database integrato, ha richiesto finora un investimento di superiore a 1,5 milioni di euro.

5.7 SVILUPPI FUTURI

Conclusa con esito positivo l'attività di sperimentazione Trentino Riscossioni ha individuato i passi necessari per la costituzione di un sistema completo a regime. In particolare l'obiettivo è di coinvolgere man mano un numero crescente di comuni per arrivare a circa un centinaio di comuni integrati. In particolare, Trentino Riscossioni si pone i seguenti obiettivi per il breve/medio periodo:

- Definizione di un modello di business per la messa in produzione della soluzione sperimentale. È necessario trovare un modello di business in grado di soddisfare tutte le parti interessate (Trentino Riscossioni, Informatica Trentina, Provincia Autonoma di Trento, Gruppo-S, Comune di Firenze);
- Affinamento della procedura di accertamento I.C.I.: per poter migliorare l'algoritmo di calcolo automatico del dovuto è necessario affinare la procedura di determinazione della prima casa, magari andando ad integrare nuove fonti informative in grado di migliorare la qualità del dato relativo agli indirizzi di residenza (per questo scopo sarebbe ad esempio molto utile la banca dati dell'Anagrafe Tributaria). Tuttavia, anche con indirizzi corretti al 100% la determinazione della prima casa continuerà a soffrire del problema della mancanza di numeri civici interni alle abitazioni, mancanza che non permette di identificare gli immobili a livello di singolo appartamento in base all'indirizzo;

- Sviluppare funzionalità di navigazione visuale dei dati e applicazioni per altri contesti di utilizzo. Innanzitutto, si vorrebbe estendere l'accertamento I.C.I. anche alle aree fabbricabili tramite l'analisi del Catasto Fondiario e del piano regolatore generale. In secondo luogo si potrebbero sviluppare funzioni a supporto dell'accertamento dei tributi TARSU e TIA;
- Aggiunta di nuove basi dati in grado in modo da estendere l'utilizzo del Data Warehouse ad altre tipologie di servizi forniti da Trentino Riscossioni. Ad esempio, si potrebbero integrare le banche dati che riportano informazioni relative ad altri beni in possesso dei contribuenti (automobili, ecc.) in modo da sviluppare applicazioni a servizio delle procedure di riscossione coattiva delle entrate;
- Georeferenziazione delle entità (immobili, strade, civici): la volontà di Trentino Riscossioni è avere in futuro una mappa georeferenziata delle entità rilevanti ai fini fiscali. Tuttavia, un'operazione di questo tipo presuppone il coinvolgimento di altri soggetti, quali la Provincia Autonoma di Trento e il Servizio Catasto, che hanno già avviato attività in questo senso. L'obiettivo di Trentino Riscossioni è dare il via ad una sperimentazione per la georeferenziazione di un particolare comune o di determinate categorie di immobili.

5.8 CONCLUSIONI

La realizzazione di questa prima sperimentazione di integrazione dei dati, con l'obiettivo di supportare le attività di Trentino Riscossioni, ha costituito un passo importante verso la realizzazione del progetto di sistema informativo immaginato dall'azienda. Le attività di sperimentazione hanno permesso alla società di valutare vantaggi e svantaggi di una soluzione di questo tipo, considerando anche i costi che un tale progetto potrebbe avere qualora si coinvolgesse la totalità dei comuni trentini. Nonostante alcuni limiti riscontrati nelle fasi di collezione, scambio ed elaborazione dei dati precedenti alla fase di integrazione e alcune lacune nell'interfaccia e funzionalità offerte dal sistema, i risultati conseguiti con questa sperimentazione sono stati giudicati buoni e hanno contribuito ad estrarre dai dati, in maniera automatica, un rilevante valore aggiunto nel particolare contesto operativo dell'accertamento dell'Imposta Comunale sugli Immobili (ICI). I risultati conseguiti testimoniano inoltre la bontà del motore di integrazione che, nonostante una qualità del dato non sempre opportuna, ha permesso l'integrazione automatica della quasi totalità dei record disponibili, riconducendo il lavoro di bonifica manuale soltanto a qualche centinaio di casistiche. La sperimentazione ha inoltre permesso di identificare le maggiori problematiche sui dati in possesso di Trentino Riscossioni, permettendo affinare le impostazioni delle variabili di integrazione. L'esito positivo della sperimentazione ha determinato il passaggio del progetto ad una fase successiva, ancora in fase di definizione,

che nel corso del 2010 dovrà portare a regime il sistema attraverso l'integrazione dei dati degli enti che hanno affidato i dati a Trentino Riscossioni e lo sviluppo di applicativi di analisi in grado di supportare le attività della società di riscossione.

CAPITOLO 6

APPLICAZIONE DI METODOLOGIE SEMANTICHE

Nel corso dello stage è stata condotta un'attività di ricerca sullo stato dell'arte di soluzioni di Data Integration innovative nell'ottica di valutare le possibili nuove tecnologie da applicare o sperimentare nel contesto di Trentino Riscossioni. In questo capitolo vengono presentati alcuni possibili sviluppi futuri che permetterebbero di avere a disposizione un sistema di integrazione dei dati più efficiente e flessibile rispetto a quanto è possibile ottenere con la soluzione descritta nel capitolo 5. Riprendendo alcune delle tecnologie descritte nel capitolo 3, sono state individuate alcune possibilità di innovazione che potrebbero essere concretamente applicate nello scenario di Trentino Riscossioni, nei termini almeno di una sperimentazione. Di fatto, per una delle possibilità individuate (Entity Name System) è già stata intrapresa un'attività di sperimentazione per valutare l'efficacia e l'utilità di un approccio semantico non solo nel contesto di Trentino Riscossioni, ma anche nell'ottica di realizzare un sistema informativo integrato per i servizi che coinvolgono la Provincia Autonoma di Trento.

6.1 ENTITY NAME SYSTEM (OKKAM)

Nell'ambito dello sviluppo di una prima versione di base dati integrata, Trentino Riscossioni pur ritenendo valido l'approccio classico che ha portato alla realizzazione di un Data Warehouse, ha deciso di non vincolarsi ad una tecnologia specifica ma di sperimentare, per quanto possibile, approcci alternativi che possano portare alla realizzazione di un sistema caratterizzato da una maggiore flessibilità. Tali tecnologie non devono per forza presentarsi come alternative alla metodologia tradizionale ma possono anche essere complementari. L'obiettivo è unire i punti di forza di tecnologie diverse per dare vita ad un sistema il più possibile efficiente, flessibile ed espandibile.

Grazie ai rapporti con la realtà universitaria, sia da parte del laboratorio TasLab che del sottoscritto, è stata identificata una tecnologia innovativa il cui approccio è stato valutato valido e ripercorribile all'interno del progetto di integrazione di Trentino Riscossioni. Si tratta della tecnologia sviluppata all'interno del progetto OKKAM. OKKAM OKKAM è un progetto di integrazione su larga scala fondato dalla Comunità Europea[1]. Al progetto partecipano diverse entità, unite in un consorzio: Università di Trento (coordinatore), L3S

[1] 7th Framework Programme, finanziamento n. 215032.

Research Center (Germania), SAP Research (Germania), Expert System (Italia), Elsevier B.V. (Olanda), Europe Unlimited SA (Belgio), National Microelectronics Application Center (Irlanda), Ecole Polytechnique Fédérale de Lausanne (Svizzera), DERI Galway (Irlanda), Università di Malaga (Spagna), INMARK (Spagna), Agenzia Nazionale Stampa Associata (ANSA, Italia) [66].

Il progetto ha come obiettivo quello di dare vita ad un Web fatto di entità, dove collezioni di dati e informazioni su entità diverse (persone, luoghi, organizzazioni, eventi, ecc.) sono pubblicati sul Web in un formato che ne permetta l'integrazione automatica in una singola base di conoscenza virtuale e decentralizzata. In particolare, OKKAM punta a realizzare un servizio aperto e decentralizzato, Entity Name System (ENS), a supporto della raccolta e dell'integrazione di dati e informazioni sulle entità presenti sul Web. La realizzazione di un sistema di questo tipo prevede l'assegnazione di un identificativo unico e globale per ogni entità nominata sul Web. Il sistema Entity Name System (ENS) potrebbe rappresentare per il Semantic Web lo stesso ruolo che il Domain Name Service (DNS) gioca per il Web tradizionale [67]. L'idea di base è che il sistema ENS sia disponibile in ogni applicazione per creare dei contenuti, in modo da rendere possibile l'interazione con il server ENS che attribuisce gli identificativi alle entità.

Per realizzare questa visione, OKKAM deve seguire il seguente percorso [66]:

1. Rendere disponibile un'infrastruttura scalabile e sostenibile (ENS) per assegnare e gestire gli identificativi unici per le entità;
2. Sostenere una crescita rapida del cosiddetto Web of Entities, promuovendo l'okkamizzazione dei contenuti e l'utilizzo di applicazioni compatibili con il servizio;
3. Dimostrare i benefici di un Web fatto di entità univocamente identificate e più in generale i vantaggi, dal punto di vista della gestione della conoscenza, di un approccio orientato alle entità, anche attraverso la realizzazione di infrastrutture, applicazioni e servizi che abilitino la ricerca semantica di informazioni, rafforzando le tecniche di gestione della conoscenza.

La realizzazione di un sistema di questo tipo permetterà di risolvere i tipici problemi di interconnessione tra i dati, dando vita ad un'interoperabilità semantica che permetterà di integrare con grande facilità informazioni relative ad una determinata entità in tempi molto brevi su scala globale (Web).

Il problema dell'identificazione univoca delle entità si presenta, seppur in un contesto più limitato, anche nel progetto di integrazione dei dati di Trentino Riscossioni. Per questo la società ha deciso di intraprendere una sperimentazione, parallela a quella descritta nel capitolo 5, per testare l'efficacia dell'applicazione dell'approccio di OKKAM in un contesto chiuso ma comunque caratterizzato da una scarsa interoperabilità informativa. La

sperimentazione è partita ufficialmente nel mese di novembre 2009 e coinvolge i seguenti attori:

- Provincia Autonoma di Trento
- Trentino Riscossioni S.p.A.
- Informatica Trentina S.p.A.
- Università di Trento
- Università di Galway

Ogni attore partecipa alla sperimentazione con i propri obiettivi. In particolare, il consorzio OKKAM si è dimostrato molto felice e determinato nel dare vita a questa sperimentazione in quanto costituisce un banco di prova importante per l'infrastruttura creata. Un eventuale esito positivo dell'applicazione del servizio in un caso d'uso reale costituirà infatti un elemento utile per dimostrare la validità del progetto. Il consorzio, tramite questa sperimentazione, si pone quindi l'obiettivo di testare la propria tecnologia in un contesto applicativo reale che possa dimostrare la validità della tecnologia sviluppata e che permetta al consorzio di avere un futuro una volta concluso il progetto europeo (giugno 2010). Un futuro che dovrebbe portare alla fondazione di un ente no-profit che gestisca l'infrastruttura del servizio ENS, affiancato da una o più continuazioni profit in grado di sviluppare servizi commerciali basati sull'infrastruttura disponibile.

Gli obiettivi che invece si pongono la Provincia e Trentino Riscossioni sono differenti. La PAT si è dimostrata interessata a tale sperimentazione in quanto, dal punto di vista del sistema informativo provinciale, potrebbe dare vita ad un sistema che permetta di integrare con maggiore facilità i diversi database provinciali (Parix, Catasto, dati geografici, ecc.) e i database degli Enti (anagrafe, tributi, stradari, ecc.). Lo sviluppo di una tecnologia di questo tipo potrebbe fornire inoltre la possibilità di sviluppare servizi Web per una moltitudine di soggetti diversi (cittadini, istituzioni, forze dell'ordine, ecc.), mantenendo la tecnologia OKKAM legata al territorio trentino. Per Trentino Riscossioni l'obiettivo rimane quello di dare un valore aggiunto ai dati in modo da:

- Acquisire una visione globale e integrata della base dati impositiva, creando un profilo tributario del contribuente il più possibile esteso e flessibile per l'aggiunta di ulteriori dati in futuro;
- Ottenere una base dati integrata con un alto livello di correttezza e qualificazione del dato, sulla quale costruire i processi di gestione delle entrate (accertamento, programmazione, estratto conto del cittadino, gestione del territorio);
- Acquisire un sistema di integrazione più flessibile e performante rispetto alla soluzione di Data Warehousing sperimentata;

Per valutare l'applicazione dell'approccio semantico di OKKAM nel contesto di Trentino Riscossioni sono stati definiti dei casi d'uso attraverso i quali misurare l'efficacia (sia in

termini quantitativi che qualitativi) della tecnologia utilizzata. In particolare, il caso d'uso base consiste nella realizzazione del profilo delle entità (persone, organizzazioni, eventi, luoghi, immobili) rappresentate nelle basi dati a disposizione (paragrafo 5.1.1). Nel caso del contribuente sarà quindi un profilo contenente i dati anagrafici e i dati utili ai fini tributari (immobili posseduti, tributi pagati, ecc.). L'interfaccia grafica del sistema di interrogazione è basata sulla tecnologia Sig.ma (http://sig.ma). Il motore di ricerca Sig.ma permette di integrare in una vista unica e personalizzabile tutte le informazioni relative ad una determinata entità, identificabile come tale in seguito all'attribuzione di un identificativo univoco tramite il servizio ENS. Il sistema permette inoltre la navigazione attraverso le entità collegate all'entità ricercata, permettendo una ricerca reticolare di informazioni che dovrebbe semplificare la ricerca dei dati utili ai fini tributari. Il passo successivo sarà quello di realizzare un'interfaccia ad hoc per l'utilizzo del profilo nel campo dell'accertamento. Un'interfaccia che permetta di filtrare le informazioni disponibili sulla base di alcuni filtri o regole schematizzate a priori. In questo modo l'accertatore avrà la possibilità di isolare determinate casistiche o tipologie di contribuenti per i quali vuole procedere all'accertamento della situazione tributaria.

Il piano di lavoro della sperimentazione con OKKAM prevede marzo 2010 come termine della sperimentazione. In caso di esito positivo è prevista la pubblicazione dei risultati in conferenze e pubblicazioni scientifiche, inoltre, verrà realizzato uno studio di fattibilità per l'introduzione di tali tecnologie nel sistema informativo provinciale, valutando i costi della soluzione e la disponibilità di personale con l'esperienza necessaria per la gestire l'evoluzione del sistema OKKAM nella fase successiva alla conclusione del progetto europeo.

6.2 APPLICAZIONE DI ONTOLOGIE

Le ontologie sono state adottate in numerosi settori come sistema per condividere, riutilizzare ed integrare la conoscenza rappresentata in un particolare dominio. Oggi le ontologie sono la componente centrale di molte applicazioni, come portali della conoscenza, sistemi di gestione e integrazione di informazioni, commercio elettronico e servizi di Web semantico in generale. La definizione di un'ontologia per lo scenario di Trentino Riscossioni dovrebbe permettere di realizzare un modello dei dati più ricco e flessibile rispetto a quanto si può ottenere con un modello relazionale, definendo a livello di schema di rappresentazione dei dati vincoli e regole alla base del sistema tributario di riferimento.

6.2.1 ONTOLOGY MATCHING

La sperimentazione con il progetto OKKAM prevede lo sviluppo di un'ontologia da utilizzare come sistema per mappare i dati provenienti da sistemi differenti in un'unica vista integrata. Un prodotto interessante per lo sviluppo e la gestione di un'ontologia in grado di rappresentare il dominio applicativo di Trentino Riscossioni è lo strumento open source Protégé[2]. Protégé è una suite di strumenti per costruire modelli di dominio basati sulle ontologie. Protégé fornisce strumenti a sostegno della creazione, visualizzazione e manipolazione di ontologie in vari formati di rappresentazione. La piattaforma Protégé supporta due modi principali di modellazione di ontologie:

- Il Protégé-editor Frames consente agli utenti di costruire e popolare ontologie che sono frame-based, in conformità con il protocollo Open Knowledge Base Connectivity (OKBC). In questo modello, un'ontologia è costituita da un insieme di classi organizzate in una gerarchia in grado di rappresentare i concetti basilari di un dominio, come le proprietà e le relazioni tra diverse entità;
- Il Protégé OWL editor consente agli utenti di creare ontologie per il Semantic Web, in particolare nel formato Web Ontology Language (OWL), standard approvato dal W3C. Un'ontologia OWL può includere le descrizioni delle classi, le proprietà e le loro istanze. Partendo da un'ontologia di questo tipo è possibile applicare la semantica formale specifica per derivare conseguenze logiche in base ai dati rappresentati.

La realizzazione di un'ontologia in grado di descrivere nei dettagli il contesto di business di Trentino Riscossioni determinerebbe la presenza di una base di conoscenza profonda e flessibile, sulla quale sviluppare poi applicazioni di visualizzazione, analisi ed integrazione di nuovi dati. In particolare, dato il contesto trentino caratterizzato da centinaia di sorgenti di dati eterogenee, si è cercato di ragionare su una possibile soluzione per garantire un'integrazione di nuovi dati almeno semi-automatizzata. Sulla base del progetto Harmony (vedi paragrafo 3.3.2) è stato formalizzato uno scenario per l'applicazione di un approccio di matching dei dati semiautomatico nel contesto di Trentino Riscossioni. Tale idea ha portato alla realizzazione di un breve paper [68], accettato come poster all'International Workshop on Ontology Matching (OM-2009), tenutosi a Washington, DC, USA, il 25 ottobre 2009[3]. Il paper si focalizza su un preciso obiettivo: potenziare l'architettura di integrazione attuale aumentando la flessibilità nella gestione di centinaia di fonti eterogenee riducendo lo sviluppo di software ad hoc per ogni tipologia di sorgente. Come sottolineano altre ricerche, [53] [55], lo sviluppo di una componente di matching automatico in uno scenario di data integration è un campo di applicazione relativamente

[2] http://protege.stanford.edu/

[3] http://om2009.ontologymatching.org/

nuovo. In questo paper viene focalizzato il particolare scenario di Trentino Riscossioni come base per l'eventuale realizzazione di un sistema innovativo in grado di integrare dinamicamente centinaia di fonti differenti. In particolare, quello che si propone è un sistema che permetta l'inserimento, la gestione e la cancellazione di una nuova fonte di dati sfruttando la sintattica e la semantica degli attributi per creare un mapping automatico con lo schema del database integrato. In Figura 44 è rappresentato un sistema di questo tipo, dove le informazioni che provengono da una fonte di dati esterna sono processate dal motore di analisi SSMB (Semantic Schema/Data Matching Box) che calcola automaticamente i collegamenti tra la nuova sorgente e lo schema del database integrato, proponendoli all'utente per la convalidazione o la modifica tramite un'interfaccia grafica. Un sistema di questo tipo si basa su un motore di matching automatico, che potrebbe essere ad esempio l'Harmony integration workbench [52].

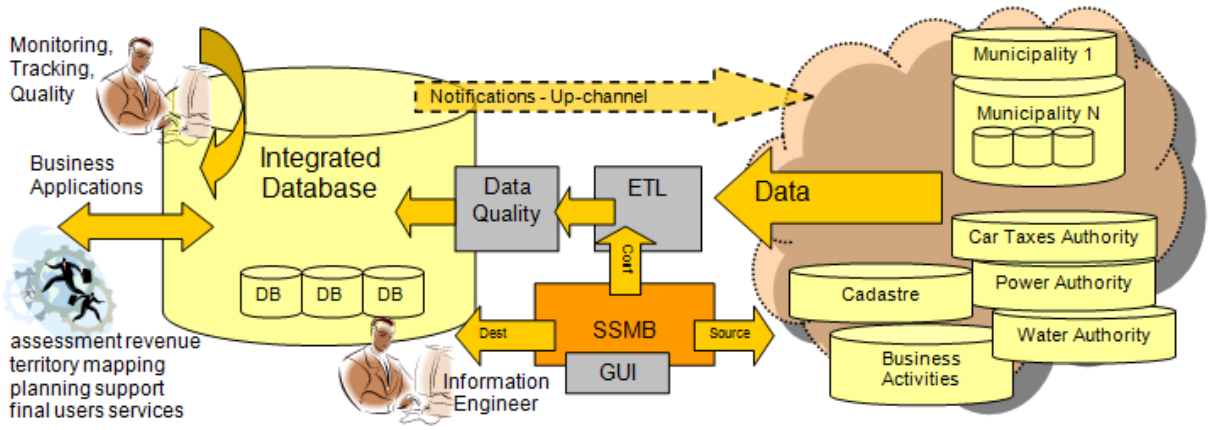

Figura 44 – Una possibile architettura di matching e mapping semiautomatico dei dati (Fonte: [68]).

La realizzazione di un sistema di integrazione di nuove fonti di questo tipo dovrebbe portare ad una maggiore interoperabilità tra i diversi sistemi e schemi di dati ed aumentare la quantità di informazioni e conoscenza condivisa all'interno del sistema di integrazione.

Nel paper realizzato si è quindi proposto un possibile scenario di applicazione di innovative tecnologie di matching dei dati, per cui, in ottica futura, ci si potrebbe muovere nelle seguenti direzioni:

1. Formalizzare nei dettagli lo scenario di applicazione;
2. Valutazione e test delle funzionalità dell'Harmony integration workbench;
3. Sviluppo di uno specifico prototipo per lo scenario di Trentino Riscossioni.

La realizzazione di un sistema di integrazione delle nuove fonti di dati più flessibile e gestibile rispetto al sistema attualmente sperimentato sarebbe un notevole valore aggiunto visto il contesto altamente eterogeneo in cui la società di riscossione si trova ad operare.

6.3 CONCLUSIONI

La sperimentazione di un prodotto di Data Integration tradizionale ha permesso di individuare limiti e criticità di tale soluzione, determinando la volontà di sperimentare metodologie alternative o complementari. Per questo, Trentino Riscossioni, grazie ai rapporti con l'Università degli Studi di Trento, ha deciso di intraprendere una nuova sperimentazione individuando nel progetto OKKAM un sistema in grado di rispondere non solo alle proprie necessità, ma di rivestire un ruolo centrale anche a livello provinciale, nell'ottica di costituire un sistema informativo provinciale in grado di favorire l'integrazione di dati provenienti da fonti differenti. La collaborazione tra Università, Provincia e Trentino Riscossioni garantisce le competenze e le risorse necessarie per esplorare ed eventualmente intraprendere strade innovative, nell'ottica della realizzazione di un sistema informatico flessibile, espandibile e in grado di abilitare l'interoperabilità tra i diversi sistemi informativi (provinciali, comunali, ecc.). La ricerca di soluzioni innovative e di valore è quindi un'attività molto importante non solo per Trentino Riscossioni ma anche per gli altri Enti trentini, di conseguenza la ricerca di nuove soluzioni da sperimentare rivestirà un ruolo di primo piano nel progetto di integrazione dell'azienda al fine di realizzare un sistema in grado di collocare il Trentino in una situazione di vantaggio nel panorama nazionale che va verso l'adozione del federalismo fiscale.

CONCLUSIONI

Il concetto di Data Integration nasce in seguito allo sviluppo dei primi sistemi di gestione dei dati, in particolare negli anni in cui prendono piede i primi sistemi gestionali per specifici contesti di business. Con l'avvento, all'interno delle realtà aziendali, di più sistemi informativi che producono, memorizzano e gestiscono dati si evidenziano i primi problemi determinati dalla presenza di dati duplicati, ridondanti, disallineanti e discordanti. Questo rende fin da subito chiara ed evidente la necessità di tecnologie in grado di dare una visione unica, consolidata e storicizzata dei dati in possesso dell'azienda. Se agli albori della disciplina del Data Mangement tale necessità non era così stringente, lo sviluppo e la crescita delle tecnologie informatiche, sia a livello hardware che software, ha portato all'adozione su larga scala dei sistemi per la gestione dei dati, con un conseguente aumento della quantità di dati e informazioni disponibili, rendendo sempre più evidente la necessità di strumenti di integrazione.

Nel corso degli anni '80-'90, in un contesto caratterizzato da una forte espansione dell'ICT, nascono quindi le prime metodologie di integrazione (data warehouse e database federati) con l'obiettivo di risolvere i sempre più evidenti problemi di gestione dei dati attraverso sistemi, schemi e tecnologie differenti. In particolare, è la disciplina del Data Warehousing che trova inizialmente maggior successo sul mercato e nel corso degli anni '90 vengono definiti diversi approcci per la realizzazione, la gestione e l'utilizzo di un Data Warehouse. A partire dagli anni 2000 cominciano ad affermarsi sul mercato soluzioni di integrazione più complete che uniscono in un ambiente unico funzionalità di gestione dei dati differenti (data warehousing, data federation, data quality, data cleaning, data mining, ecc.). Tali prodotti cercano di fornire un ambiente più semplice ed usabile ma ricco di funzionalità, in modo da rendere meno complesse le operazioni di integrazione dei dati, rendendo disponibili una moltitudine di funzionalità aggiuntive che permettono di costruire una base dati integrata solida e di qualità per esigenze e finalità differenti.

Tuttavia, tali tipologie di prodotti si basano su schemi e tecnologie di gestione dei dati poco flessibili e scarsamente espandibili; con l'avvento della rete internet e la conseguente esplosione della quantità di dati i prodotti di integrazione presenti sul mercato cominciano a manifestare dei limiti. La comparsa di nuovi formati di dati, la necessità di esporre dati sul Web, la crescente complessità dei contesti di business ha determinato una situazione globale caratterizzata da alta eterogeneità e variabilità di dati e informazioni. Per questo motivo è necessaria un'ulteriore evoluzione dei prodotti di integrazione, un'evoluzione verso modelli di dati più ricchi e flessibili, in grado di espandersi senza problemi e che possano essere facilmente integrati tra loro. L'evoluzione dei sistemi di integrazione dei

dati punta quindi verso l'adozione di tecnologie e schemi semantici per la rappresentazione dei dati e verso la realizzazione di contenitori di metadati comuni attraverso i quali favorire l'integrazione e il riuso dei modelli di rappresentazione dei dati utilizzati. Inoltre, i prodotti di integrazione tradizionali si basano su un modello di prezzi obsoleto che rende difficile l'accesso a tali tecnologie alle aziende più piccole e con risorse limitate. L'evoluzione del contesto della Data Integration punta quindi anche sull'affermazione di nuovi modelli di prezzi, modelli basati sull'effettivo utilizzo dell'infrastruttura di integrazione anziché su il vecchio concetto di licenza per macchina. I nuovi prodotti di integrazione, semantici e open source, comparsi sul mercato negli ultimi anni, puntano infatti sul nuovo modello di prezzi, basato sull'utilizzo del servizio, rendendo di fatto accessibili ad ogni tipologia di azienda tecnologie di integrazione fino a poco tempo disponibili solo per i più grandi contesti aziendali.

Nonostante i limiti dettati dalle più recenti evoluzioni del contesto informatico, i tradizionali prodotti di integrazione garantiscono la realizzazione di database integrati e storicizzati che permettono di effettuare analisi ed elaborazioni a supporto di diverse attività altrimenti molto dispendiose e difficili da realizzare. In particolare, nel caso di reale applicazione di tecnologie di integrazione, nell'ambito del progetto Trentino Riscossioni, l'utilizzo di tecniche tradizionali (data warehouse) ha garantito la realizzazione di una base dati integrata di qualità superiore rispetto alle basi dati di partenza, storicizzata (per quanto possibile) e consolidata, dove ogni entità è identificata univocamente. L'esperienza avuta nel progetto di Trentino Riscossioni da un lato ha confermato l'utilità di un database integrato e dall'altro ha posto una serie di problemi e interrogativi sui limiti della tecnologia utilizzata, in particolare in previsione di una futura espansione del sistema per l'intera realtà trentina (223 comuni più diversi enti di tipologia differente).

Per quanto riguarda gli effetti positivi e benefici di un database integrato, l'esperienza della sperimentazione realizzata da Trentino Riscossioni ha permesso di constatare come in un particolare ambito di applicazione (accertamento tributario) l'utilizzo di dati integrati possa portare ad una netta diminuzione del tempo di analisi manuale, dando all'operatore la possibilità di concentrarsi sui casi sospettati di maggiore evasione, evitando una dispendiosa analisi manuale di tutte le posizioni contributive che risultano corrette. L'integrazione di più basi dati in un database unico ha altresì permesso di unificare in un'anagrafica unica tutte le informazioni relative ad una determinata entità. Questo permette di ottenere un'anagrafica delle entità il più possibile completa e corretta, alleviando nuovamente l'operatore dall'esecuzione di laboriose e dispendiose operazioni manuali di ricostruzione dell'anagrafica pescando i dati da diversi database, collocati in schemi e ambienti differenti. L'aggregazione delle informazioni sulla medesima entità in un record unico ha quindi permesso di compiere un'operazione di pulizia e consolidazione dei dati disponibili, garantendo la realizzazione di una base dati integrata di buona qualità.

Inoltre, laddove la fonte di origine lo permettesse il data warehouse è impostato per mantenere una visione storica dei dati, cosa di fondamentale importanza nel contesto tributario che prevede la possibilità di operare fino a 5 anni indietro nel tempo; visione storica che non sempre le basi dati di origine permettono di ottenere.

Tuttavia, per arrivare alla creazione del database integrato sono state necessarie una serie di tappe, talvolta lunghe e complesse, che vanno dal recupero dei dati presso i fornitori (Comune, PAT, o società private), alla loro analisi, passando per la descrizione e la mappatura degli attributi. Inoltre, lo sviluppo di funzioni verticali ha comportato una fase di forte collaborazione con il personale addetto allo specifico contesto di utilizzo dei dati, in modo da raccogliere i requisiti che l'applicazione deve soddisfare per garantire un utilizzo utile ed efficace del dato integrato. Sono state attività abbastanza onerose in termini di tempo e sicuramente l'espansione del sistema all'intera realtà trentina richiederà la creazione di un team di lavoro che si occupi a tempo pieno delle fasi di preparazione, valutazione e analisi dei dati, e in generale di tutte le attività preliminari e di contorno alla fase di integrazione vera e propria che verrà gestita dal fornitore del prodotto. Ci si scontra inoltre con un contesto di riferimento altamente variabile e piuttosto complesso da rappresentare attraverso procedure automatizzate, quale è il contesto tributario. In aggiunta, il contesto altamente frammentato tipico della realtà trentina non facilita sicuramente l'integrazione dei dati in un ambiente unico. Ci si scontra infatti con un contesto caratterizzato dalla presenza di 223 comuni e diversi enti, ognuno con le proprie basi dati, costruite secondo logiche differenti, tramite l'utilizzo di applicativi diversi e, talvolta, nemmeno informatizzate ma gestite tramite archivi cartacei. Nel complesso ci si trova di fronte ad un contesto operativo sicuramente complesso che richiederà investimenti consistenti per poter realizzare a pieno il progetto di integrazione promosso da Trentino Riscossioni ma verso il quale anche la PAT nutre forte interesse.

Per quanto riguarda la soluzione tecnologica sperimentata si tratta sostanzialmente di una soluzione di Data Warehousing che tramite l'utilizzo di tecniche di inferenza statistica permette di riconoscere in maniera automatica la stessa entità attraverso basi dati differenti e, di conseguenza, permette di integrare tutti i dati che la riguardano in un record unico. Il fornitore del prodotto si è dimostrato valido e molto disponibile così come il prodotto utilizzato ha permesso di realizzare una base dati integrata di buona qualità, in grado di dare un effettivo valore aggiunto nel caso d'uso dell'accertamento, definito per testare l'efficacia della base dati integrata. Pur avendo determinato il conseguimento di un buon risultato la soluzione utilizzata presenta dei limiti. In particolare, le operazioni di preparazione dei dati per l'integrazione sono eccessivamente macchinose e il caricamento dei dati richiede personale con una discreta esperienza dell'ambiente. Inoltre, la soluzione adottata non si dimostra molto flessibile e scalabile, il che, in un ambiente altamente variabile come quello tributario, potrebbe rappresentare un potenziale problema. Infine, il

contesto trentino altamente frammentato richiederà onerose fasi di preparazione e mappatura dei dati date dall'aggiunta di un nuovo comune alla base dati integrata.

Per questi motivi Trentino Riscossioni si sta muovendo alla ricerca di tecnologie alternative o complementari a quella sperimentata, tecnologie in grado di garantire un approccio più flessibile e scalabile soprattutto sia per quanto riguarda le fasi di analisi, mappatura e integrazione di nuove basi dati sia per garantire una struttura del database integrato flessibile, espandibile e facilmente modificabile in seguito ad eventuali evoluzioni del contesto di riferimento. In tutto questo la ricerca di un modello di business valido e vantaggioso per tutte le parti coinvolte costituisce sicuramente il prossimo passo sul quale lavorare intensamente per la messa in produzione del sistema di integrazione dei dati, seguito dalla definizione di nuovi casi d'uso e di nuove applicazioni verticali per sfruttare le potenzialità del database integrato.

BIBLIOGRAFIA

[1] Frank Hayes, *The Story So Far*, 2002 [Online]. http://www.computerworld.com/s/article/70102/The_Story_So_Far

[2] Cark W. Olofson, *Worlwide Database Management Systems 2009-2013 Forecast and 2008 Vendor Shares*, IDC Market Analysis, 2009.

[3] William Harvey Inmon, *The Evolution of Integration*, Inmon Consulting Services, 2008.

[4] Vincent McBurney, *Data Integration*, 2009 [Online]. http://it.toolbox.com/wiki/index.php/Data_Integration

[5] Ted Friedman, Mark A. Beyer, e Andreas Bitterer, *Magic Quadrant for Data Integration Tools*, Gartner Research, 2008.

[6] Vincent McBurney, *Data Integration Techniques*, 2009 [Online]. http://it.toolbox.com/wiki/index.php/Data_integration_techniques

[7] Mark R. Madsen, *The Role of Open Source in Data Integration*, 2009, Third Nature Technology Report.

[8] William Harvey Inmon, "What is a Data Warehouse?," in *Prism*, vol. 1, no. 1, 1995.

[9] Ralph Kimball e Margy Ross, *The Data Warehouse Toolkit*, 2nd ed. Wiley Computer Publishing, 2002.

[10] Michael Read, *A Definition of Data Warehousing*, 2009 [Online]. http://www.intranetjournal.com/features/datawarehousing.html

[11] William Harvey Inmon, *Building the Data Warehouse*, 3rd ed. Wiley Computer Publishing, 2002.

[12] Fon Silvers, *Building and Maintaining Data Warehouse*. Auerbach Publications, 2008.

[13] Ralph Kimball, *The Data Warehouse Lifecycle Toolkit*. Wiley, 1998.

[14] William Harvey Inmon, "Data Mart Does Not Equal Data Warehouse," in *DM Review*, 1998.

[15] Dennis Heimbigner e Dennis McLeod, "A federated architecture for database systems," in *AFIPS National Computer Conference 1980*, Anaheim (California), 1980, pp. 283-289.

[16] Dennis Heimbigner e Dennis McLeod, "A Federated Architecture for Information Management," in *ACM Transactions on Office Information Systems*, vol. 3, no. 3, pp. 253-278, 1985.

[17] Alon Halevy et al., "Enterprise Information Integration: Successes, Challenges and Controversies," in *ACM SIGMOD Conference*, Baltimore, Maryland, USA, 2005.

[18] Laura Haas, Eileen Lin, e Mary Roth, "Data Integration through database federation," in *IBM Systems Journal*, vol. 41, no. 4, 2002.

[19] Michael Stonebraker, *To ETL or federate? That is the question*, 2007 [Online]. http://databasecolumn.vertica.com/2007/12/to-etl-or-federate.html

[20] Jim Harris, *All I Really Need To Know About Data Quality I Learned In Kindergarten*, 2009 [Online]. http://www.ocdqblog.com/home/all-i-really-need-to-know-about-data-quality-i-learned-in-ki.html

[21] Colin White, *Developing a universal approach to cleansing customer and product data*, 2008, SAP BusinessObject White Paper.

[22] Thomas C. Redman, *Data Driven*. Boston, Massachusetts Harvard Business Press, 2008.

[23] Leo L. Pipino, Yang W. Lee, e Richard Y. Wang, *Data Quality Assessment*, Communications of the ACM, 2002.

[24] Jonathan Geiger, "Data Quality Management: The Most Critical Initiative You Can Implement," in *SUGI 29 Proceedings*, Montréal, Canada, 2004, pp. Paper 098-29.

[25] Ted Friedman, *Data Integration Technology and Architecture: Building Your Data Circulatory System*, Gartner Summit Events, 2007.

[26] Ted Friedman, *The Benefits (and Challenges) of Deploying Data Integration Tools*, 10 agosto 2007.

[27] SAP Press Conferences, *SAP to Acquire Business Objects in Friendly Takeover*, 2007 [Online]. http://www.sap.com/about/press/businessobjects/20071007_005046.epx

[28] Ted Friedman, Mark A. Beyer, e Andreas Bitterer, *Magic Quadrant for Data Integration Tools*, Gartner Research, 2007.

[29] Oracle Press Release, *Oracle Buys Sun*, 20 aprile 2009 [Online]. http://www.oracle.com/us/corporate/press/018363

[30] Oracle Press Release, *U.S. Department of Justice Aprroves Oracle Acquisition of Sun*, 20 agosto 2009 [Online]. http://www.oracle.com/us/corporate/press/029738

[31] Jordan Robertson, *EU objects to Oracle's takeover of Sun*, in Yahoo! Tech News, 10 novembre 2009 [Online]. http://tech.yahoo.com/news/ap/20091110/ap_on_hi_te/us_tec_oracle_sun

[32] IBM Press room, *IBM to Acquire Cognos to Accelerate Information on Demand Business Initiative*, 2007 [Online]. http://www-03.ibm.com/press/us/en/pressrelease/22572.wss

[33] SAP Press Conferences, *Business Objects to Acquire Firstlogic, Inc.*, 2006 [Online]. http://www.sap.com/about/newsroom/businessobjects/20060208_005974.epx

[34] Oracle Press Release, *Oracle to Acquire BEA Systems*, 16 gennaio 2008 [Online]. http://www.oracle.com/corporate/press/2008_jan/bea.html

[35] Zoomix Press Release, *Microsoft Signs Agreement to Purchase Data Quality Start-up Zoomix*, 14 luglio 2008 [Online]. http://www.zoomix.com/pressreleases_article.asp?id=26

[36] Vincent McBurney, *Gartner Versus Open Source Data Integration*, 2009 [Online]. http://it.toolbox.com/blogs/infosphere/blog-fight-gartner-versus-open-source-data-integration-29169

[37] Talend, *The Top 10 Reasons for Choosing Open Source Data Integration*, 2009.

[38] The Yankee Group Report, *Uncovering the Hidden Costs in Data Integration*, 2004.

[39] Philip Howard, *Comparative costs and uses of Data Integration Platforms*, 2008, Research paper by Bloor Research.

[40] David Waddington, *Ten Tips for Selecting a Data Integration Tool. Total Cost of Ownership Really Matters*, 2008, Research from Tyson Consulting.

[41] Neil Raden, *Data Integration: What's Next?*, 1 maggio 2008, Smart (enough) Systems Report.

[42] Patrick Ziegler e Klaus Dittrich, "Three Decades of Data Integration: All Problems Solved?," in *18th IFIP World Computer Congress*, Toulouse, France, 2004.

[43] Neil Raden, *The Emerging Role of Semantic Technology in the Enterprise*, 2006.

[44] Tim Berners-Lee, James Hendler, e Ora Lassila, "The Semantic Web," in *Scientific American Magazine*, 2001, http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

[45] Emanuele Della Valle, Irene Celino, e Dario Cerizza, *Semantic Web - Modellare e condividere per innovare*, 1st ed. Pearson Addison/Wesley, 2008.

[46] Mitchell Ummel et al., "The Rise of the Semantic Enterprise," in *Cutter IT Journal*, vol. 22, no. 9, pp. 1-37, 2009.

[47] Thomas Gruber, "A Translation Approach to Portable Ontology Specifications," in *Knowledge Acquisition*, vol. 5, no. 2, pp. 199-220, 1993.

[48] Rudi Studer, Richard Benjamins, e Dieter Fensel, "Knowledge engineering: principles and methods," in *IEEE Transactions on Data and Knowledge Engineering*, vol. 25, no. 1-2, pp. 161-197, 1998.

[49] Neil Raden, *Semantic Integration: Tapping the Full Potential of Enterprise Data*, 2006.

[50] John Talburt, *Entity Resolution vs. Entity Identification*, 2009 [Online]. http://www.identityresolutiondaily.com/533/entity-resolution-vs-entity-identification/

[51] John Talburt, *The Myth of Matching: Why We Need Entity Resolution*, 2009 [Online]. http://identityresolutiondaily.com/493/the-myth-of-matching-why-we-need-entity-resolution/

[52] Peter Mork, Arnon Rosenthal, Joel Korb, e Chris Wolf, *The Harmony Integration Workbench*, MITRE, 2009.

[53] Peter Mork et al., *The Role of Schema Matching in Large Enterprises*, MITRE, 2009.

[54] H. Wache et al., *Ontology-Based Integration of Information: A Survey of Existing Approaches*, 2001, Intelligent Systems Group, Center for Computing Technologies, University of Bremen, Germany.

[55] Pavel Shvaiko e Jerome Euzenat, *Ten Challenges for Ontology Matching*, Dipartimento di Ingegneria e Scienza dell'Informazione, 2008.

[56] Semyon Axelrod, *MDM is Not Enough - Semantic Enterprise is Needed*, 20 marzo 2008 [Online]. http://www.information-management.com/specialreports/2008_69/10000964-1.html

[57] Oracle, *Oracle Database 11g Semantic Technologies, Semantic Data Integration for the Enterprise*, 2009, An Oracle White Paper.

[58] Expressor Press Release, *expressor 2.0 announcement*, 12 maggio 2009 [Online]. http://www.expressor-software.com/expressor-v2-announcement.htm

[59] Expressor Software, *Expressor™ - redefining data integration* , 2009, Expressor 2.0 product overview.

[60] Philip Howard, *Expressor Software*, 2009, A InDetail Paper by Bloor Research.

[61] Expressor Software, *expressor/Netezza intelligent load & go: low-cost, high-performance data warehousing*, 15 novembre 2009 [Online]. http://www.expressor-software.com/ilag-netezza.htm

[62] Trentino Riscossioni S.p.A., *Il modello di Governance*, 2006.

[63] Claudio De Stasio, *Compendio di diritto tributario*, 1st ed., 2004.

[64] Raffaello Lupi, *Diritto Tributario. Oggetto economico e metodo giuridico nella teoria della tassazione analitico-aziendale*. Giuffrè Editore, 2009, pp. 1-27, 139-213.

[65] Roberto Martini e Luigi Lovecchio, *ICI: guida ragionata all'applicazione dell'imposta*. Sistemi Editoriali, 2003.

[66] Heiko Stoermer, 6 marzo 2009, *More about the OKKAM Project* [Online]. http://www.okkam.org/okkam-more

[67] Paolo Bouquet, Heiko Stoermer, Daniele Cordioli, e Giovanni Tummarello, "An Entity Name System for Linking Semantic Web Data," in *Proceedings of the Linked Data on the Web Workshop CEUR*, 2008.

[68] Stefano Brida, Marco Combetto, Silvano Frasson, e Paolo Giorgini, "Tax&Revenue Service scenario for Ontology Matching," in *The Fourth International Workshop on Ontology Matching*, Washington, DC, USA, 25 ottobre 2009. http://www.dit.unitn.it/~p2p/OM-2009/om2009_poster3.pdf