Indice

Capitolo 1

Introduzione

La seguente tesi andrà ad illustrare e a documentare il lavoro svolto presso il Gruppo Unicomm nel periodo compreso tra Luglio 2011 e Dicembre 2011; il progetto, guidato dal Dottor Sergio Rizzato, è finalizzato alla creazione di un portale di business intelligence per il marketing. Le principali attività che sono state svolte sono le seguenti:

- analisi della parte del sistema informativo inerente al venduto online e alle tessere punti Fidelity;
- analisi dei report per l'area marketing attualmente presenti;
- analisi delle nuove richieste dell'area marketing;
- individuazione dei KPI (Key Performance Indicators) strategici per le analisi di marketing;
- creazione del modello dati per l'implementazione di un data warehouse contente le informazioni sul venduto online e sulle tessere punti Fidelity.

Oltre a queste attività è stato svolto uno studio teorico per apprendere le conoscenze necessarie allo sviluppo del modello dati del data warehouse, oltre ad un approfondimento sulla business intelligence.

Capitolo 2

Unicomm srl

2.1 L'azienda

Il Gruppo Unicomm è una realtà molto importante nel settore della grande distribuzione alimentare. La sede centrale del gruppo, che riunisce in un'unica grande struttura tutte le funzioni centrali aziendali e il centro distributivo principale, si trova a Dueville (Vi) e si sviluppa su una superficie di 112.400 metri quadrati; il centro distributivo dedicato ai prodotti freschi è situato invece a San Pietro in Gu (Pd) e si estende su una superficie di 20.000 metri quadrati.

I canali di vendita presidiati dal gruppo sono quelli della vendita al dettaglio, il cash and carry, l'ingrosso e il franchising. Per poter rispondere alle diverse esigenze e tipologie di clientela e per poter offrire una copertura capillare sul territorio all'interno del gruppo sono presenti diversi canali di vendita diretta:

- Svelto A&O;
- Super A&O;
- Famila;
- Famila Superstore;
- Emisfero;
- C + C cash and carry;
- Hurrà.

L'area magazzino gestisce una giacenza media di 1.450.000 colli, con 12.900 posti picking e un totale di 45.700 posti pallet. Giornalmente vengono spediti oltre 100.000 colli con un traffico quotidiano di 120 automezzi.

Il Gruppo Unicomm è presente in 7 regioni (Veneto, Friuli Venezia Giulia, Emilia Romagna, Marche, Toscana, Umbria e Lazio). Si è ampliato negli anni e comprende oggi numerose partecipazioni e alleanze locali. Attualmente il gruppo è composto dalle seguenti società:

- Unicomm srl con sede a Dueville (Vi);
- M. Guarnier spa con sede a Belluno;
- Arca spa con sede a Longiano (Fc);
- GMF spa con sede a Ponte San Giovanni (PG);
- Domenico Aliprandi spa con sede a Oderzo (Tv);
- Alter srl con sede a Ponte nelle Alpi (Bl);

Nel 2010 la rete di vendita del gruppo era rappresentata da 204 punti vendita con un fatturato di 1.835.000.000 euro con un trend positivo negli anni.

2.2 Il progetto

Il progetto seguito, riguardante la creazione di un portale di business intelligence per il marketing, è una parte di un lavoro molto più ampio che coinvolge l'intera organizzazione e che si prefigge l'obiettivo di mettere a disposizione di tutti i settori dell'azienda uno strumento comune per poter svolgere analisi di business intelligence. La parte del progetto relativa al marketing è stata avviata nel giugno 2011 in seguito a due ulteriori motivazioni: il cambiamento del sistema di acquisizione dei dati del venduto online e la necessità, da parte del marketing, di avere uno strumento di reportistica più flessibile rispetto al sistema vigente fino a quel momento, e che fosse in grado di integrare i dati sul venduto online con i dati delle tessere punti dei clienti (tessere Fidelity).

Il cambiamento del sistema di acquisizione dei dati del venduto online si è reso necessario in quanto, vista la continua espansione del gruppo e la presenza di realtà molto diverse al suo interno, si era arrivati ad avere una situazione problematica nella quale erano presenti software diversi per la gestione dei dati provenienti dai vari punti vendita. Come si può notare dalla figura sottostante la gestione dei dati era molto problematica perché erano presenti applicazioni distinte per svolgere le stesse funzioni, i dati erano ridondanti e le analisi dei dati venivano eseguite in software distinti.

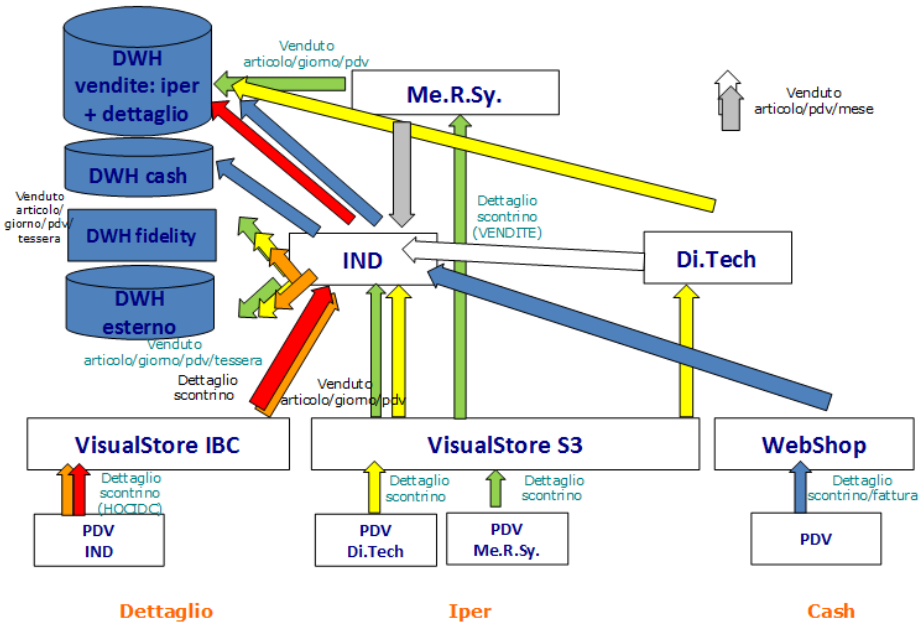

Figura 2.1: Flusso dati del venduto online prima dell'avvio del progetto

In particolare erano presenti diversi software di back-office (server sui punti di vendita): VisualStoreIBC per il canale di vendita dettaglio, VisualStore S3 per il canale di vendita iper e WebShop per il canale di vendita cash. Una ulteriore problematica riguardava la gestione dei dati proveniente da questi software di back-office: i dati confluivano in tre diversi sistemi gestionali per la gestione delle "merci" (acquisti, vendite e magazzino). Anche questi software gestionali, come per il caso precedente erano divisi per canale:

- IND: gestisce i dati relativi alle vendite dei Famila, A&O, Hurrà e Cash and Carry;
- Di.Tech: gestisce le vendite degli Emisfero;
- Me.R.Sy: gestisce i dati di vendita di alcuni Emisfero.

Oltre ai diversi software di gestione dei dati anche le analisi di business intelligence venivano svolte in data warehouse distinti. Questi strumenti di analisi presentavano un front-end di utilizzo diverso e pertanto obbligavano gli utenti finali a conoscere diversi strumenti software per svolgere le analisi.

Questa molteplicità di applicativi, distinti a seconda del canale di vendita, rendeva la gestione del sistema e delle informazioni da esso estratte molto complessa e ormai non più sostenibile; la direzione aziendale ha deciso pertanto di avviare una procedura di unificazione dei sistemi passando ad un unico strumento di acquisizione del venduto online per tutti i punti vendita del gruppo. Questa procedura è iniziata a Luglio 2011 e si concluderà nel 2012.

La configurazione finale del sistema una volta ultimato il cambiamento sarà le seguente:

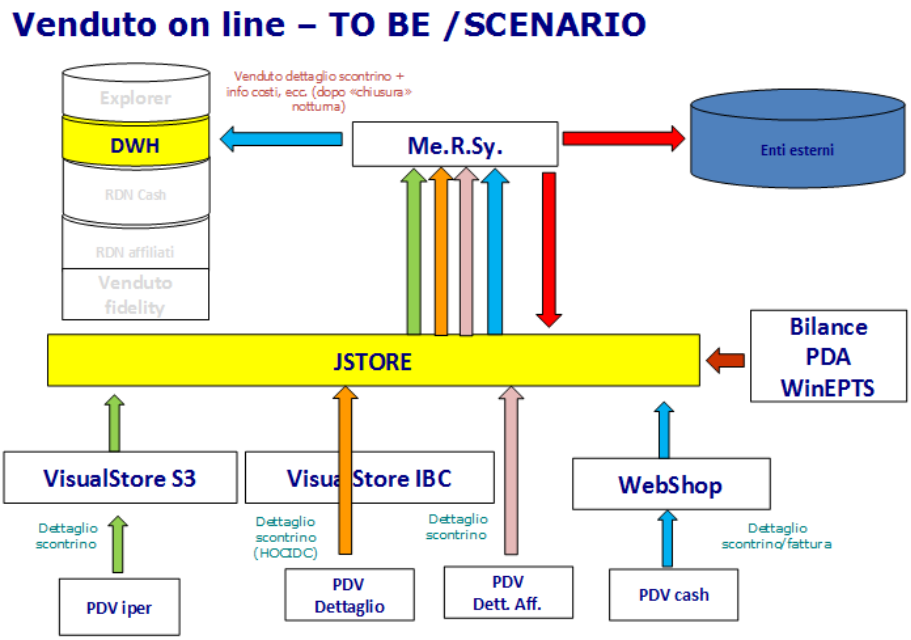

Figura 2.2: Flusso dati finale del venduto online

Nella nuova configurazione del sistema di gestione del venduto si ha un unico applicativo di back-office chiamato JStore il quale invia i dati a Me.R.Sy che è un sistema gestionale integrato in grado di gestire gli acquisti, le vendite e il magazzino delle merci.

Sfruttando questo cambiamento, e viste le continue richieste di modifiche e miglioramenti da parte del marketing, la direzione aziendale ha deciso di avviare un progetto per la realizzazione di un portale di business intelligence che dovrà essere in grado di offrire ai vari operatori report con cadenza periodica e dovrà inoltre permettere la navigazione all'interno di questi report in modo efficace. Questo nuovo portale di BI metterà a disposizione del marketing uno strumento in grado di gestire tutte le informazioni provenienti dai punti vendita e dal nuovo sistema gestionale.

2.3 Il sistema informativo: venduto online e Fidelity Web

Per comprendere il funzionamento del sistema informativo verrà spiegato il ruolo dei vari software, come comunicano tra loro e a che livello si andranno ad estrarre i dati per popolare il data warehouse.

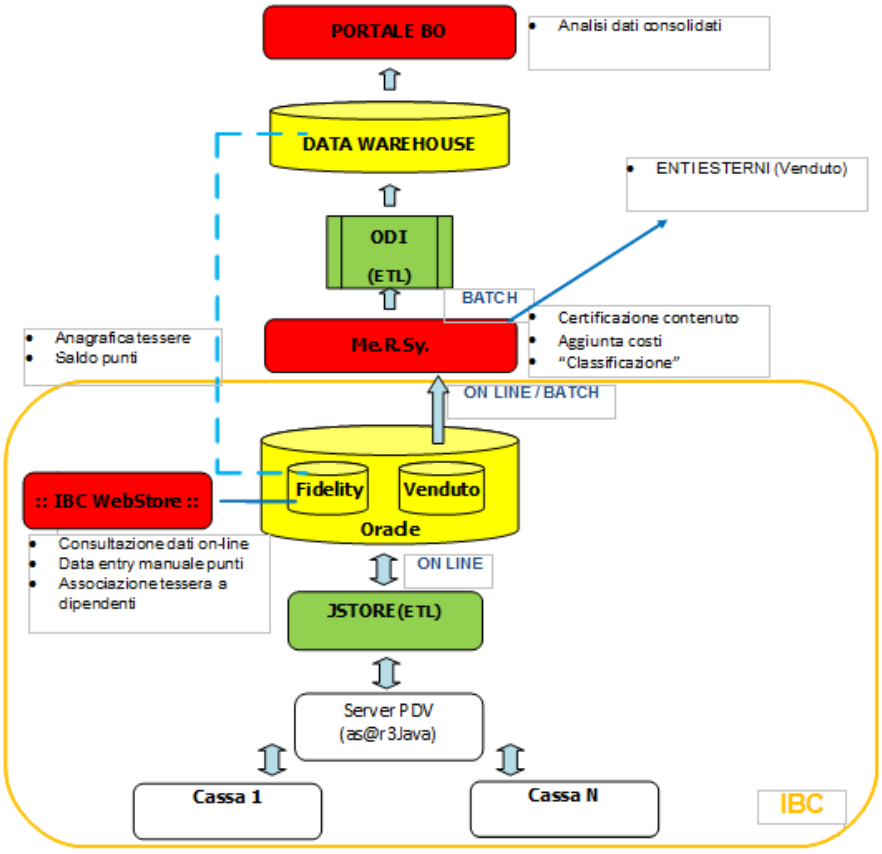

Figura 2.3: Venduto online e Fidelity Web

La parte a livello più basso è gestita da IBC (Information Business Computing) e si occupa di raccogliere i dati provenienti dalle casse dei vari punti vendita. Questi dati vengo trasmessi dai software di gestione delle casse, diversi a seconda del canale di vendita e, in particolare, a seconda che si tratti di iper (Emisfero), cash (C + C cash and carry) o vendita al dettaglio (Svelto A&O, Super A&O, Famila, Famila Superstore, Hurrà). Il primo step di questa raccolta dati prevede l'utilizzo di JStore; successivamente questi dati vengono immagazzinati (online) in due database Oracle: il primo contenente i dati relativi al venduto e il secondo contenente i dati relativi alle tessere Fidelity. A questo punto la parte relativa alle tessere Fidelity viene gestita da un altro software di IBC chiamato IBC Web Store mentre la parte relativa al venduto viene elaborata e certificata da Me.R.Sy (applicativo sviluppato da Wincor Nixdorf). I dati presenti a questo livello sono dati aggregati, che mantengono però un livello di dettaglio sufficiente per consentire di risalire alla singola riga dello scontrino. Il ruolo dei tre applicativi appena descritti è il seguente:

- IBC JStore: è una soluzione rivolta alle esigenze delle aziende del mercato retail che vogliono gestire in maniera centralizzata i punti di vendita. Provvede alla gestione e al supporto in modalità centralizzata sia delle funzionalità dei punti vendita che di quelle di un ambiente di sede che deve gestire reti di punti vendita;

- IBC WebStore: è una soluzione che offre la gestione centralizzata di anagrafiche e altre informazioni (punti accumulati dai clienti possessori di carta Fidelity) che vengono gestite in sede e distribuite ai singoli punti vendita. Consente inoltre la gestione di una struttura merceologica gerarchica con un numero di livelli configurabile in fase di istallazione;

- Me.R.Sy: è una soluzione applicativa che supporta l'impresa e gli utenti nell'ambito della relazione con l'industria, della gestione assortimenti dei punti vendita, del pricing, della concorrenza, della politica promozionale, dell'approvvigionamento merci e dei controlli dei cicli amministrativi. Le fondamenta della suite gestionale Me.R.Sy. si trovano nel suo sistema anagrafico che, governando le informazioni vitali (articoli, fornitori, clienti,...), assicura la corretta gestione di aziende multisocietà e multicanale. L'ambiente applicativo, basandosi su questa solida struttura anagrafica, garantisce una copertura ampia e profonda di tutte le funzionalità inerenti al ciclo passivo (rapporto tra distributore e produttore) e al ciclo attivo (rapporto tra distributore e punto vendita).

I dati che andranno a popolare il data warehouse si otterranno dal databese Fidelity e da Me.R.Sy. L'acquisizione dei dati verrà realizzata con uno strumento ETL (Extract, Trasform, Load) che verrà programmato una volta completato il modello dati. I dati caricati nel data warehouse saranno utilizzabili per le analisi in quanto sono già stati certificati dai sistemi sottostanti.

Capitolo 3

Il marketing e la business intelligence

Dopo aver analizzato e compreso il funzionamento del sistema informativo e aver deciso, insieme alla direzione aziendale, di sviluppare il portale di BI, prendendo i dati dal nuovo sistema unificato di gestione del venduto online e delle tessere Fidelity (quindi da Me.R.Sy e dal database Fidelity), si è passati all'analisi delle esigenze del marketing.

In questa fase del progetto si sono analizzati, attraverso una serie di incontri con il personale del marketing, i report, gli strumenti e le informazioni messe a disposizione dall'attuale sistema gestionale. Successivamente sono state valutate le richieste di miglioramento e le modifiche da apportare ai report attuali. In seguito a questi incontri si è potuto stimare, seppur ancora in modo approssimativo, le dimensioni del progetto e le modalità con cui procedere verso le fasi successive.

3.1 Situazione attuale

Durante i primi incontri sono stati analizzati, con l'aiuto di Margherita Avanzi responsabile CRM e loyalty del settore marketing e comunicazione, i report attualmente presenti su AS400[1]. Per prima cosa sono stati individuati, tra tutti i report messi a disposizione, quelli effettivamente utilizzati. Successivamente è stata svolta la stessa operazione per i report messi a disposizione da Catalina[2]. Dopo questa prima "passata" veloce sui report è iniziata una fase di analisi più dettagliata. Per ogni singolo report è stato valutato il suo funzionamento (parametri d'ingresso, tipologia di filtri e tempi di elaborazione) prendendo nota degli aspetti critici, delle lacune e delle nuove esigenze suggerite dal marketing.

[1] AS400 è il sistema di gestione dei dati relativi al venduto attualmente presente in azienda. Questo software verrà sostituito da Me.R.Sy.

[2] Catalina Connection Builder è un software web-based di business intelligence.

3.1.1 Report su AS400

1. Stampa venduto Fidelity ente/periodo (Codice CLU1).

   Questo report viene utilizzato molto frequentemente in quanto offre una sintesi generale sull'andamento di ogni singolo negozio o di un canale di distribuzione in quanto è possibile visualizzare questo report per tutti i negozi appartenenti a quel canale.

PARAMETRI RICHIESTI:
Ente Negozio.....: XXXXXXXX
Tipo periodo ....: X
Periodo dal .....: XX/XX/XXXX al: XX/XX/XXXX
Tipo di Stampa...: 1 STAMPA PER ENTE

REPORT

Ente XXXX XXXXXXXXXXXXX

Periodo	Clienti	Val.Totale	Vendita Fidelity	%Inc.	Trans.	Num. Vis. Carta Club	%Inc.	Scontr.Medio	Scon.Med.Carta Club	Valore Offerta	% Offerta	Sconto C.Club
01	4738	733.372,420	522.963,040	71,30	30.123	17.340	57,56	24,345	30,159	77.618,170	10,58	0,000
02	4761	730.864,860	527.287,400	72,14	30.495	17.577	57,63	23,966	29,998	76.304,600	10,44	0,000
03	4864	802.835,580	579.118,910	72,13	33.059	19.096	57,76	24,284	30,326	79.250,350	9,87	0,000
04	4974	868.191,890	635.257,560	73,17	33.050	19.205	58,10	26,269	33,077	97.071,800	11,18	0,000
05	4997	793.790,260	580.973,810	73,18	32.465	19.109	58,86	24,450	30,403	84.182,280	10,60	0,000
	24334	3.929.055,010	2.845.600,720	72,42	159.192	92.327	57,99	24,681	30,820	414.427,200	10,54	0,000

Figura 3.1: Stampa venduto Fidelity ente/periodo

Da questo primo report sono emersi alcuni aspetti critici:

- le intestazioni delle colonne non sono immediatamente comprensibili;
- la visualizzazione di questo report per tutti i negozi lo rende difficile da consultare;
- la colonna relativa allo sconto carta club non è valorizzata;
- il totale dei cliente è errato in quanto somma tutti i clienti dei singoli periodi non considerando il fatto che la maggior parte dei clienti è la stessa nei vari periodi.

2. Stampa venduto Fidelity per decili (Codice CLU3).

   Questo report consente di suddividere i clienti in base all'ammontare degli acquisti effettuati, raggruppandoli per decili.

PARAMETRI RICHIESTI:
Ente Negozio.....: XXXXXXXX
Tipo periodo ....: X
Periodo dal .....: 0XX/XX/XXXX al: XX/XX/XXXX
Tipo Decile .....: 2 XXXXXXX

REPORT

	Venduto Fidelity	N.ro Clienti	% Venduto su tot.	Valore offerta	%Off.Carta Club	Sconto C. Club	%Inc.di sconto Transizioni	Scontr.Medio
DECILE 1	57.667,840	84	9,90	5.543,350	8,10		1.312	76,121
DECILE 2	57.910,010	126	9,90	5.931,860	8,60		1.524	67,885
DECILE 3	57.835,730	160	9,90	6.735,230	9,80		1.500	20,365
DECILE 4	58.034,700	196	9,90	6.767,440	9,90		1.552	44,855
DECILE 5	58.025,640	238	9,90	5.938,640	8,60		1.674	27,576
DECILE 6	58.006,360	293	9,90	6.505,260	9,50		1.868	88,165
DECILE 7	58.063,330	379	9,90	7.233,070	10,50		1.911	22,213
DECILE 8	58.007,090	518	9,90	7.095,960	10,30		2.120	23,642
DECILE 9	58.066,310	788	9,90	7.870,130	11,50		2.302	28,340
DECILE 10	59.356,800	2214	10,20	8.660,800	12,60		3.346	0,390
	580.973,810	4996	100,00	68.281,740	100,00		19.109	30,403

Figura 3.2: Stampa venduto per decili

Aspetti critici:

- sarebbe interessante, per ogni singolo decile, visualizzare i dettagli dei clienti che lo compongono. Con l'attuale sistema di reportistica non è possibile in quanto i report messi a disposizione da AS400 sono statici.

3. Stampa venduto Fidelity con dettaglio reparto (Codice CLU5).

   Questo report suddivide il venduto dei vari reparti merceologici per ogni periodo in modo da poter analizzare, per ogni singolo reparto, la sua redditività e influenza sul fatturato totale.

PARAMETRI RICHIESTI:
Ente Negozio.....: XXXXXXXX
Tipo periodo ....: X
Periodo dal .....: XX/XX/XXXX al: XX/XX/XXXX

REPORT

TOTALI VENDUTO NEGOZIO XXXXXXXXX

	TOTALE VENDUTO		VAL TOT. CARTA CLUB		OFFERTA CARTA CLUB		SCONTO CARTA CLUB	
REPARTI	791.681		569.106		88.902		0	
SCATOLAME	289.386	36,55%	210.143	36,92%	45.966	51,70%	0	0,00%
CARNE	50.555	6,38%	38.502	6,76%	0	0,00%	0	0,00%
ORTOFRUTTA	117.482	14,83%	86.304	15,16%	0	0,00%	0	0,00%
LATTICINI/GASTRONO	256.246	32,36%	180.565	31,72%	42.852	48,20%	0	0,00%
PESCE	31.643	3,99%	21.500	3,77%	0	0,00%	0	0,00%
NON-FOOD	46.366	5,85%	32.090	5,63%	84	0,09%	0	0,00%

Figura 3.3: Stampa venduto Fidelity con dettaglio reparto

Aspetti critici:

- anche in questo caso sarebbe interessante, partendo dal singolo reparto, poter eseguire un'operazione di drill down[3] per valutare l'andamento dei singoli prodotti o di gruppi di prodotti (es: all'interno dello scatolame valutare l'andamento della vendita dei biscotti).

4. Consultazione venduto Fidelity per tessera (Codice CLU8).

   Questo report permette di consultare per ogni singola tessera il totale dei movimenti sia in termini di transazioni che di venduto. Offre inoltre la possibilità di vedere la percentuale di spesa in offerta acquistata dal singolo cliente.

5. Stampa clienti codificati: dettaglio per tessera (Codice $382).

   La stampa di questo report permette di visualizzare tutte le informazioni relative ad una tessera Fidelity; il suo utilizzo non è frequente ma talvolta necessario.

6. Stampa venduto per articolo dettaglio giorno (Codice CLU13).

   Questo report visualizza per ogni negozio o per un gruppo di negozi tutti i prodotti che sono stati venduti con le loro quantità, il numero di clienti Fidelity e non Fidelity che li hanno acquistati e altre informazioni relative alla vendita.

ENTE	PUNTO VENDITA	DATA DAL	DATA AL	CODICE ARTICOLO	DESCRIZIONE ARTICOLO	CLIENTI TC	TRANS.TC	QTA TC	VAL.TOT.TC	QTA C.CLUB	VAL.C.CLUB	%INC.	VAL. OFFERTA	%INC.	CLIENT NTC	TRANS. NTC	QTA NTC	VAL.TOT.NTC
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	34758	FORM.MAASDAM	305	370	287,755	980,55	0	0	0	0	0	2	46	98,428	287,07
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	37052	MASCARPONE	18	22	22	90,96	0	0	0	0	0	1	4	6	30,28
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	42909	MORTADELLA ORO KG.1CA	8	8	8,413	49,64	0	0	0	0	0	1	4	4,24	25,02
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	73425	CROISSANT SURGELATI	4	4	5	15,72	0	0	0	0	0	1	2	2	6,75
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	73430	STRUDEL DI MELE SURG.	5	5	5	20,8	0	0	0	0	0	0	0	0	0
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	73913	GNOCCHI PATATE SURG.	2	2	2	4,7	0	0	0	0	0	1	1	1	2,35
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	90042	LATTE UHT P.S.	24	35	80	93,6	0	0	0	0	0	1	8	22	25,74
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	96260	PANNA UHT XCAFFE'10X10	36	59	118	73,16	0	0	0	0	0	1	9	12	7,44
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	96285	PANNA CUCINA UHT	10	12	14	15,26	0	0	0	0	0	0	0	0	0
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	121217	POLPA POMODORO 3X400	25	26	27	33,75	0	0	0	0	0	1	6	8	10
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	130105	PISELLI MEDI	22	30	33	19,47	0	0	0	0	0	1	9	16	9,44
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	131675	FAGIOLINI FINISSIMI	7	7	15	11,25	0	0	0	0	0	1	2	3	2,25
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	131732	FAGIOLINI FINI	8	11	15	18,75	0	0	0	0	0	1	1	1	1,25
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	141830	PESCHE META'SCIROP.	3	3	3	4,05	0	0	0	0	0	1	1	1	1,45
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	153048	NETTARE ARANCIA 3X200	49	68	98	81,25	0	0	0	0	0	1	9	10	8
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	160510	FUNGHI TRIFOLATI	120	140	238	311,82	0	0	0	0	0	1	13	29	44,91
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	160970	CARCIOFI SPACCATI	5	5	5	19,9	0	0	0	0	0	0	0	0	0
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	166630	CAPPERI IN ACETO	20	21	23	24,15	0	0	0	0	0	1	6	7	7,35
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	180419	TONNO O.O. 4X80	8	8	9	27,7	0	0	0	0	0	1	6	6	19,4
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	181023	TONNO O.O.	10	18	21	60,48	0	0	0	0	0	1	1	1	2,88
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	181705	SGOMBRI O.O.	4	4	5	6,75	0	0	0	0	0	1	3	7	9,45
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	181790	FILETTI SGOMBRO O.O.	26	30	49	71,05	0	0	0	0	0	1	6	12	17,4
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	182070	FILETTI SGOMBRO O.O.	12	14	18	36,83	0	0	0	0	0	1	3	4	8,09
6101	FAMILA ARZIGNANO BARACCA	1092011	30092011	190102	CARNE LESSATA 3X70	9	14	15	38,85	0	0	0	0	0	1	2	2	5,18

Figura 3.4: Stampa venduto per articolo dettaglio giorno

Aspetti critici:

- l'impaginazione del report non è di immediata comprensione;
- le colonne relative alle offerte non sono valorizzate;
- sarebbe interessante poter selezionare un solo reparto o una sola categoria di prodotti piuttosto che la totalità dei prodotti.

[3] Operazione di "esplosione" del dato nelle sue determinanti.

Oltre alle considerazioni fatte per ogni singolo report ci sono delle osservazioni generali che possono essere applicate a tutti i report. La prima richiesta del marketing riguarda la possibilità di visualizzare il saldo punti dei clienti e i punti emessi dai negozi ai clienti Fidelity. Attualmente non è possibile visualizzare, negli stessi report, sia i dati del venduto che quelli relativi ai punti Fidelity in quanto AS400 non gestisce il saldo punti e IBC Web Store non gestisce i dati sul venduto dei negozi. Se ciò invece fosse possibile, sommando il saldo punti di ogni singolo cliente per ogni negozio, si potrebbe valutare la quantità di premi da erogare e la quantità di punti che non sono ancora stati spesi, informazioni di importanza strategica per l'area acquisiti per la gestione del magazzino dei premi. Questo dato verrebbe inoltre utilizzato per effettuare delle previsioni di spesa: i punti accumulati dai clienti e non ancora spesi sono una voce passiva nel bilancio aziendale e sarebbe molto importante riuscire a valorizzarlo.

Un'altra lacuna riscontrata riguarda le modalità di visualizzazione dei dati: in molti casi le intestazioni dei campi non sono chiare e utilizzano termini che, a seconda dell'utilizzatore, possono assumere significati diversi. Nel nuovo sistema sarebbe auspicabile avere una terminologia per le intestazioni dei campi più chiara e univoca. La visualizzazione dei report attualmente in uso non consente inoltre di effettuare dei confronti immediati tra più punti vendita: gli operatori del marketing sono costretti ad esportare i dati in Excel e ad effettuare delle ritabulazioni delle tabelle.

Come già specificato su alcuni report il marketing vorrebbe avere inoltre la possibilità di navigare all'interno dei singoli campi attraverso delle operazioni di drill down questo consentirebbe di ampliare di molto le possibilità e la flessibilità di analisi in quanto, ad esempio, partendo dai report relativi ai negozi sarebbe possibile andare a vedere l'andamento di vendita dei singoli articoli, dei reparti o, per quanto riguarda i clienti, consentirebbe di effettuare delle considerazioni utili per effettuare delle market basket analysis[4].

3.1.2 Report su Catalina

I report analizzati in questo paragrafo sono i report messi a disposizione da Catalina Connection Builder. Catalina mette a disposizione, tramite un portale web, una serie di report sui quali si possono effettuare analisi di business intelligence. Questo servizio è a pagamento. I dati sul venduto vengono inviati a Catalina ogni due settimane. Dopo alcuni giorni dall'invio dei dati sul portale sono disponibili i report aggiornati. In alcune sessioni è possibile inoltre effettuare delle query specifiche andando a modificare i campi su cui effettuare la ricerca.

Anche in questo caso, come avvenuto per le analisi su AS400, si è visto che non vengono utilizzati tutti i report messi a disposizione.

[4]Letteralmente "analisi del paniere": è un processo di analisi di affinità che analizza le abitudini di acquisto dei clienti nella vendita al dettaglio, trovando associazioni su diversi prodotti comprati; tale processo è utile per l'adozione di strategie di marketing ad hoc.

Report di ricerca

1. Dettaglio scontrini.

   Questo report viene utilizzato principalmente dalla sicurezza patrimoniale in quanto consente di andare a vedere il dettaglio dello scontrino di ogni singolo cliente Fidelity. Anche in questo caso sarebbe molto utile poter vedere il saldo punti del cliente; una ulteriore richiesta per questo report è quella di visualizzare, oltre al codice EAN dei prodotti, anche il codice articolo interno, in quanto gli operatori del marketing hanno più familiarità con questi codici.

2. Performance PV (punto vendita), 2 periodi.

   Report molto utile per avere una sintesi (con dettaglio PV) relativa ai clienti persi-nuovi (e comuni) in 2 periodi liberamente definiti dall'utente.

CMC PV	01	02
N° Clienti/HH (TT), Persi, (Fatt. (TT))	2.674	603
N° Clienti/HH (TT), Nuovi, (Fatt. (TT))	2.598	576
N° Clienti/HH (TT), Nuovi-Persi, (Fatt. (TT))	-76	-27
Variaz. % N° Clienti/HH (TT) (Periodo 1-->2)	-1,0%	-1,0%
Variaz. % N° Clienti/HH (TT) (Periodo 1-->2), Tit. Carta, No Abuser	-1,0%	-1,0%
Variaz. % N° Clienti/HH (TT) (Periodo 1-->2), Altre ID,No Abuser		
Variaz. % N° Clienti/HH (TT) (Periodo 1-->2), Tit. Carta, Abuser		
Variaz. % N° Clienti/HH (TT) (Periodo 1-->2), Altre ID, Abuser		
Variaz. % Fatt. (TT) (Periodo 1->2)	-4,1%	-5,7%
Variaz. % Fatt. (TT) (Periodo 1->2), Tit. Carta, No Abuser	-3,3%	-6,9%
Variaz. % Fatt. (TT) (Periodo 1->2), Altre ID, No Abuser		
Variaz. % Fatt. (TT) (Periodo 1->2), Tit. Carta, Abuser		
Variaz. % Fatt. (TT) (Period 1->2), Altre ID, Abuser		
Variaz. % Fatt. (TT) (Periodo 1->2), Senza Carta	-6,2%	0,8%
Variaz. % Transaz. (TT) (periodo 1 -->2)	-4,9%	-7,5%
Variaz. % Transaz. (TT) Tit. Carta, No Abuser, (periodo 1 -->2)	-3,3%	-8,0%
Variaz. % Transaz. (TT) Altre ID, No Abuser (periodo 1 -->2)		
Variaz. % Transaz. (TT) Tit. Carta, Abuser, (periodo 1 -->2)		
Variaz. % Transaz. (TT) Altre ID, Abuser (periodo 1 -->2)		
Variaz. % Transaz. (TT) Senza carta, (periodo 1 -->2)	-7,0%	-6,3%

Figura 3.5: Performance PV, 2 periodi

3. Profilo sintetico clientela.

   Report che sintetizza tutti gli indicatori di comportamento dei vari clienti.

Mese - anno	2006-12	2006-11
Fatt. (TT), Tit. Carta, No Abuser	€ 455.300,95	€ 436.167,10
Fatt. (TT), Altre ID, No Abuser		
Fatt. (TT), Senza Carta	€ 159.880,15	€ 149.998,19
Fatt. (TT), Abuser		
Fatt. (TT)	€ 615.181,10	€ 586.165,29
Transaz. (TT), Tit. Carta, No Abuser	31.969	33.568
Transaz. (TT), Altre ID, No Abuser		
Transaz. (TT), Senza carta	20.308	21.066
Transaz. (TT), Abuser		
Transaz. (TT)	52.277	54.634
Scontrino medio (TT), Tit. Carta, No Abuser	€ 14,24	€ 12,99
Scontrino medio (TT), Altre ID, No Abuser		
Scontrino medio (TT), Senza carta	€ 7,87	€ 7,12
Scontrino medio (TT), Abuser		
Scontrino medio (TT)	€ 11,77	€ 10,73
% Transaz. (TT), Tit. Carta, No Abuser	61,15%	61,44%
% Transaz. (TT), Altre ID, No Abuser		
% Transaz. (TT), Senza carta	38,85%	38,56%
% Transaz. (TT), Abuser		
% Fatt. (TT), Tit. Carta, No Abuser	74,01%	74,41%
% Fatt. (TT), Altre ID, No Abuser		
% Fatt. (TT), Senza carta	25,99%	25,59%
% Fatt. (TT), Abuser		
N° Clienti (TT), Tit. Carta, No Abuser	8.822	9.004
N° Clienti (TT), Altre ID, No Abuser		
Spesa Media (TT), Tit. Carta, No Abuser	€ 51,61	€ 48,44
Spesa Media (TT), Altre ID, No Abuser		
Transaz. Medie (TT), Tit. Carta, No Abuser	3,62	3,73
Transaz. Medie (TT), Altre ID, No Abuser		

Figura 3.6: Profilo sintetico clientela

4. Ranking fasce orarie.

   Report che permette di analizzare l'andamento delle transazioni suddivise in fasce orarie giornaliere. Queste analisi possono risultare interessanti per valutare i risultati di una campagna promozionale eseguita solo in alcune fasce orarie. Inoltre danno la possibilità all'azienda di capire, per ogni negozio, il numero di casse da aprire durante il giorno.

Estrazione liste clienti

1. Clienti attivi nel periodo.

   Consente di individuare quali sono le ID delle tessere dei clienti attivi in un determinato periodo. Il suo utilizzo viene solitamente sostituito dal report "Clienti in crescita/flessione".

2. Clienti di una categoria di prodotto.

   Report potenzialmente molto interessante in quanto offre la possibilità di vedere quali sono i clienti che hanno acquistato un determinato prodotto. Il marketing vorrebbe avere la possibilità di inserire un paniere di prodotti (attualmente non è possibile) o di poter individuare e, mediante operazioni di drill down, analizzare i clienti che hanno acquistato alcuni prodotti e altri no (per esempio tutti i clienti di un Famila che hanno acquistato biscotti e latte ma non fette biscottate).

3. Clienti in crescita/flessione.

   Permette di confrontare il numero di clienti attivi in due periodi distinti. Si tratta di un report utile per fare considerazioni sulla migrazione dei clienti verso un nuovo negozio della catena o dei concorrenti.

4. Evoluzione spesa cliente, 2 periodi.

   Questo report offre la possibilità di mettere a confronto il fatturato e il numero di transazioni effettuate da un cliente (o da tutti i clienti) in due periodi distinti.

5. Evoluzione spesa cliente, 3 periodi.

ID Cliente	Var. Scontrino medio (TT) (Periodo 1-->2)	Var. % Scontrino medio (TT) (Periodo 1->2)	Var. Scontrino medio (TT) (Periodo 1-->3)	Var. % Scontrino medio (TT) (Periodo 1->3)	Fatt. (TT), Periodo1	Transaz. (TT), Periodo1	Scontrino medio (TT), periodo1	Fatt. (TT), Periodo2	Transaz. (TT), Periodo2	Scontrino medio (TT), periodo2	Fatt. (TT), Periodo3	Transaz. (TT), Periodo3
1	22,03		28,88					22,03	1	22,03	28,88	1
2	-8,69	-100,00%	-8,69	-100,00%	8,69	1	8,69					
3	-21,00	-100,00%	-21,00	-100,00%	21,00	1	21,00					
4	-6,58	-29,61%	-19,34	-87,04%	22,22	1	22,22	31,28	2	15,64	2,88	1

Figura 3.7: Evoluzione spesa clienti, 3 periodi

Report analogo al precedente che permette di mettere a confronto 3 periodi distinti. Viene utilizzato spesso per valutare i risultati delle campagne pubblicitarie o delle offerte in quanto consente di analizzare i dati sul venduto pre-durante-post una campagna promozionale. Sia questo report che il precedente avrebbero una maggior utilità se fosse possibile inserire una lista di prodotti (ad esempio i prodotti in offerta) sui quali effettuare l'analisi.

Gestione prodotti

1. Vendite per prodotto.

   Report utile per monitorare l'andamento di determinati prodotti.

Codice EAN		Senza Carta						Titolari Carta					
		N° Clienti	Fatt. (Prod)	Quantità	Scontrino medio (TcP)	Transaz.	Fatt. (TcP)	N° Clienti	Fatt. (Prod)	Quantità	Scontrino medio (TcP)	Transaz.	Fatt. (TcP)
8002410	AAAAA	0	€ 2,04	1	€ 36,72	1	€ 36,72	5	€ 10,20	5	€ 18,42	5	€ 92,10
8002417	BBBBB	0	€ 2,04	1	€ 18,47	1	€ 18,47	1	€ 2,04	1	€ 26,84	1	€ 26,84
8006790	CCCCC	0	€ 25,20	28	€ 3,33	13	€ 43,35	1	€ 0,90	1	€ 1,68	1	€ 1,68
8009774	DDDDD	0	€ 0,68	2	€ 22,33	2	€ 44,65	3	€ 1,02	3	€ 12,22	3	€ 36,67

Figura 3.8: Vendite per prodotto

Report segmentazione (cliente)

1. Segmenti di scontrino medio VS fasce di transazione.

   Il report segmenta i clienti in base al proprio scontrino medio, incrociandoli con la frequenza d'acquisto (n° di transazioni). Il risultato è una matrice scontrino/frequenza utile per analizzare la correlazione tra i 2 indicatori e per segmentare la clientela. E' un report per certi aspetti molto simile a quello presente su AS400 chiamato "Stampa per decili".

2. Segmenti clienti per numero.

   Offre la possibilità di suddividere i clienti in fasce basate sul numero di transazioni e di analizzare i dati del fatturato di ogni singola fascia.

3. Segmenti clienti per fatturato.

   Report uguale al precedente con l'unica differenza che in questo caso, le fasce vengono identificate in base al numero di transazioni e non al fatturato.

Report segmentazione (transazioni)

1. Analisi transazione per fasce di scontrino.

   Il report indica il numero di transazioni per fasce di scontrino. Permette di valutare la concentrazione degli scontrini in fasce di importi e può essere utilizzato per la pianificazione di attività promozionali.

Report migrazione clienti

1. Confronto fra PV, Periodo1 vs Periodo2.

   Il report richiede la selezione di 2 PV e di 2 periodi di cui visualizzare gli indicatori principali; può essere utile in relazione a casi specifici (ad esempio se il PV1 è in ristrutturazione nel periodo2 e si vuol capire se i clienti si sono spostati in un altro negozio (PV2) della stessa insegna).

2. Segmenti clienti per fatturato, periodo1 VS periodo2.

   Il report richiede la selezione di 2 periodi, in riferimento ai quali il software segmenta i clienti in decili (per fatturato), restituendo l'incrocio delle 2 segmentazioni (metrica = n° clienti), mettendo in evidenza i clienti persi e i clienti nuovi.

Anche per questi report, come già detto per quelli presenti su AS400, si possono riscontrare problemi legati all'intestazione delle colonne e alla possibilità di navigare all'interno dei report. Per quanto riguarda la tabulazione e la stampa dei report Catalina offre la possibilità di esportare i report sia in formato PDF che in formato Excel, consentendo pertanto, nei casi in cui la tabulazione proposta non sia funzionale, di modificarla in modo semplice e veloce.

La reportistica messa a disposizione da Catalina presenta però altri aspetti critici: i dati su cui effettuare le analisi sono disponibili con un ritardo di circa due/tre settimane; in alcuni casi questo non è un problema ma in altri, come per valutare il successo di una campagna promozionale, è un ritardo non accettabile. Un'altra criticità è legata al fatto che, quando gli operatori vanno a modificare una query, per ottenere le risposte sono costretti a tempi di attesa molto lunghi. Questo è dovuto principalmente a due motivi:

1. La struttura dati utilizzata per fare analisi è standard e non è specifica per le richieste del marketing. Quando vengono modificate le query pertanto, soprattutto nei casi in cui si vuole ottenere un livello elevato di dettaglio, è necessario attendere una rielaborazione complessa dei dati a partire da tabelle con milioni di righe. Questo problema è presente anche su AS400 però è meno marcato in quanto le possibilità di modificare le query sono molto più limitate.

2. Il ritardo nella visualizzazione dei dati è legato anche al fatto che i server, nei quali vengono eseguite le elaborazioni, sono dislocati negli USA e la quantità di dati da trasferire in risposta è ingente. Inoltre, come suggerito dagli operatori del marketing, i tempi di risposta variano anche in base al carico di richieste che pervengono a questi server. Conviene quindi utilizzare il portale di Catalina durante la mattina quando negli Stati Uniti le aziende sono ancora chiuse.

Confrontando i report dei due software emerge inoltre che alcuni di essi sono disponibili in entrambi i sistemi, questo è ovviamente uno spreco inutile di risorse.

3.2 Esigenze del marketing

A seguito degli incontri nei quali sono stati analizzati nel dettaglio i report utilizzati, si è svolto un ulteriore incontro al quale oltre a Margherita Avanzi, ha partecipato il Dottor Alessandro Camattari direttore del settore marketing e comunicazione. Durante questo incontro sono emerse e sono state documentate le esigenze a cui il sistema di business intelligence che si andrà a progettare dovrà rispondere:

1. Il sistema dovrà essere in grado di integrare i dati del venduto con i dati delle tessere Fidelity.

2. L'intestazione dei campi dovrà essere univoca e chiara per tutti i report.

3. Il nuovo sistema dovrà essere in grado di gestire sia report statici che dinamici. I report statici dovranno poter essere richiesti saltuariamente, programmati e inviati ai diretti interessati via mail. Questo consentirà ad esempio all'inizio di ogni mese di inviare ai responsabili delle varie aree dei report sull'andamento dei negozi. Per quanto riguarda i report dinamici invece dovrà essere possibile navigare al loro interno tramite

operazioni di drill down e drill across[5] in modo efficiente con tempi di attesa più brevi possibili. A partire da tutti i report dovrà essere possibile raggiungere il livello di dettaglio più elevato, cioè quello della singola riga dello scontrino.

4. Il nuovo sistema dovrà offrire la possibilità di analizzare i volantini delle offerte, funzionalità che non è presente nei due strumenti di analisi attuali ma che è di notevole importanza per l'azienda. Attualmente l'unico modo possibile per analizzare i risultati di una campagna promozionale è quello di andare ad inserire la lista dei prodotti presenti nella campagna e analizzare i dati di vendita su questi prodotti; così facendo il lavoro è molto complesso perché bisogna recuperare tutti i codici di quella campagna, non si possono effettuare inoltre analisi per tutte quelle tipologie di campagne che non prevedono un taglio prezzo del prodotto. Queste offerte possono riguardare uno sconto percentuale sul venduto totale, l'aggiunta di punti Fidelity bonus nel caso di acquisto di determinati prodotti, le offerte tre per due e sconti se si supera una soglia di spesa in determinati reparti. Le analisi sulle campagne pubblicitarie e sulle offerte sono fondamentali per un'azienda che opera sul mercato retail e non avere un sistema in grado di analizzarle è una lacuna che deve essere assolutamente colmata dal nuovo applicativo.

3.3 Nuove analisi proposte

Oltre alle richieste attualmente espresse dal marketing è stato svolto un lavoro di ricerca per individuare nuove tipologie di analisi che potrebbero risultare interessanti una volta che il nuovo sistema entrerà in funzione. Questa ricerca permette di avere una visione sugli sviluppi che potrebbe avere in futuro il sistema. Capire quali altre analisi si possono effettuare con i dati a disposizione può risultare di vitale importanza nella fase di progettazione del modello dati in quanto, se non si prevedono le esigenze e le richieste che potrebbero emergere in futuro, ci si potrebbe trovare tra qualche anno a dover riprogettare il modello dati perché non più in grado di rispondere in modo efficiente alle nuove esigenze aziendali.

Pertanto, dopo alcune analisi teoriche dei dati disponibili e dei possibili report che si possono ottenere partendo da queste informazioni, sono emersi i seguenti nuovi report di analisi:

- Analisi del venduto in base ai dati anagrafici presenti nelle schede dei clienti Fidelity.

  Esempi di utilizzo: si sceglie una categoria di prodotti (esempio alcolici), si individua il venduto medio, si raggruppano i vari clienti in base alla percentuale di scostamento dalla media (o altri criteri) e si analizzano le caratteristiche anagrafiche dei vari raggruppamenti.

[5] Operazione mediante la quale si naviga attraverso uno stesso livello nell'ambito di una gerarchia.

Obiettivi: individuare bisogni diversi e capire qual è il modo per raggiungere i vari segmenti di clienti; individuare inoltre i segmenti di clienti più attraenti per determinati prodotti.

- Analisi degli spostamenti dei clienti.

  Esempi di utilizzo: si sceglie un gruppo di negozi, si trovano le tessere passate in più negozi, si individuano i prodotti passati solo in un negozio e non negli altri e si va a vedere se questi prodotti sono presenti in entrambi i punti vendita o no.

  Obiettivi: capire le motivazioni per cui i clienti si recano in due negozi distinti, capire eventuali vantaggi derivati dal posizionamento strategico dei negozi e, partendo dall'anagrafica delle tessere Fidelity, individuare i clienti per valutare gli acquisti effettuati nei vari negozi e canali distributivi del gruppo. Quest'ultima analisi è più complessa delle altre poiché i clienti, a seconda del canale distributivo, hanno tessere Fidelity diverse e l'unico modo per confrontare i dati è quello di utilizzare le schede anagrafiche dei clienti.

- Analisi delle vendite dei prodotti prima, durante e dopo una campagna promozionale.

  Obiettivi: valutare in che modo la campagna promozionale ha influito sulle vendite dei prodotti.

- Analisi dei reparti in cui i clienti (o un raggruppamento di clienti) acquistano meno.

  Esempi di utilizzo: se si nota che un reparto è poco produttivo in un solo Famila allora si può supporre che il problema sia legato a qualche aspetto di quel negozio e non ai prodotti della catena.

  Obiettivi: individuare i reparti meno produttivi e capirne i motivi (esempio: motivi legati al personale che serve i clienti).

- Analisi sui clienti persi.

  Obiettivi: individuare i clienti persi e realizzare azioni mirate nei loro confronti per tentare di recuperarli.

3.4 Analisi delle esigenze e avvio del progetto

Dopo aver raccolto tutte le esigenze del marketing e aver ipotizzato alcuni sviluppi futuri del sistema, si è passati alla fase di elaborazione delle esigenze e avvio del progetto. Il primo passo è stato quello di illustrare ai consulenti della ditta Miriade[6] quanto emerso dagli incontri con il marketing. Successivamente si è passati, con l'aiuto di Vittorio Favero (consulente di Miriade), alla fase

[6]Ditta di consulenza che collabora con Unicomm nello sviluppo del portale aziendale di business intelligence.

pre-progettuale dove si sono analizzati gli aspetti organizzativi, le tempistiche del progetto e le modalità con cui i vari membri dovranno operare all'interno del progetto.

La prima decisione ha riguardato la scelta della struttura dati su cui sviluppare il progetto: è stato scelto di utilizzare un data warehouse basato su Oracle. Questo database è molto potente ed è in grado di gestire l'ingente quantità di dati che dovrà essere elaborata per le analisi di business intelligence. Si è stabilito poi di utilizzare il portale di business intelligence aziendale, già in uso in altri settori aziendali, come strumento in cui programmare, gestire e reperire i report. L'accesso ai dati presenti in questo portale sarà consentito solo agli utenti in possesso delle credenziali; così facendo sarà possibile gestire e mettere a disposizione i dati aziendali solo al personale autorizzato e in base alle necessità.

Un altro aspetto emerso durante le riunioni con il marketing riguarda la tempistica con cui i dati per le analisi di BI dovranno essere disponibili all'interno del data warehouse: le analisi non verranno svolte sui dati in tempo reale ma a partire dai dati del giorno precedente. Questo è un aspetto fondamentale per la progettazione del data warehouse perché consente di programmare gli strumenti ETL per l'acquisizione dei dati in modo tale da popolare il data warehouse durante la notte, quando il carico di lavoro dei server è ridotto. Se fosse stato necessario svolgere analisi sui dati in tempo reale sarebbero emersi molti problemi implementativi. C'è solo un caso eccezionale per il quale tale tempistica può non essere rispettata: quello in cui una tessera Fidelity, durante la giornata, superi un certo numero prestabilito di transazioni; questo dato serve alla sicurezza patrimoniale per scoprire eventuali abusi della tessera Fidelity. Questa analisi è già svolta da IBC Web Store; si è visto che è possibile accedere a questo dato in tempo reale accedendo al portale di IBC. Nella progettazione del DWH (data warehouse) possiamo pertanto tralasciare questa richiesta.

La decisione più importante presa durante la fase pre-progettuale riguarda la tempistica e le modalità con cui procedere nello sviluppo del progetto. E' stato deciso di suddividere il progetto in due fasi: la prima fase consiste nel creare il modello logico del data warehouse e la documentazione relativa ai campi di analisi e alla tipologia di report richiesti; la seconda fase, che inizierà appena sarà concluso il passaggio a Me.R.Sy, e cioè quando ci saranno i dati del venduto di tutti i negozi disponibili sul nuovo gestionale, consiste nella programmazione del software ETL e nell'implementazione del data warehouse. Tale fase inizierà comunque subito dopo il completamento della prima in quanto ci sono già alcuni punti vendita che utilizzano Me.R.Sy per testare il suo funzionamento e sarà affidata a Miriade in quanto necessita di un elevato livello di conoscenza. Inizialmente verranno caricati i dati relativi ai vari negozi e progressivamente, a mano a mano che verranno portati su Me.R.Sy anche gli altri punti vendita, saranno disponibili tutti i dati. Si prevede che il progetto si concluderà entro la fine del 2012.

Capitolo 4

La business intelligence

4.1 Che cos'è e quando è nata

Per business intelligence (BI) si intende un insieme di tecniche basate sull'utilizzo di strumenti informatici per identificare, estrarre ed analizzare i dati di business. Il suo utilizzo si sta diffondendo molto rapidamente in quanto consente di ottenere informazioni storiche, valutazioni sullo stato attuale, e offre la possibilità di effettuare previsioni future sull'andamento dell'azienda. Tali informazioni sono molto utili per i dirigenti aziendali, soprattutto nelle aziende di medie e grandi dimensioni, dove la vision dell'imprenditore non è in grado di coprire tutti gli aspetti dell'azienda; è una tecnica che consente di prendere decisioni strategiche basandosi su dati reali e previsioni statistiche. Un sistema di BI può essere definito come un sistema di supporto decisionale.

Il termine business intelligence è stato usato per la prima volta nel 1958 da un ricercatore dell'IMB di nome Hans Peter Luhn, che definì la business intelligence come "la capacità di cogliere le interrelazioni dei fatti presentati in modo tale da orientare l'azione verso un obiettivo desiderato". Successivamente, all'inizio degli anni '90, Howard Dresner, un analista del Gartner Group, identificò con questo termine un insieme di modelli, metodi e strumenti rivolti alla raccolta sistematica del patrimonio di informazioni generate da un'azienda, alla loro aggregazione e analisi e infine alla loro presentazione in forma semplice. Questa conoscenza è utilizzabile in processi decisionali e di analisi da parte dei knowledge workers. La Business Intelligence si è di fatto sviluppata in quelle realtà aziendali dove sono presenti enormi quantità di dati, tipicamente nel mondo della telefonia e della finanza, dove è necessario colmare la lacuna esistente tra l'accumulo puro di dati e la loro comprensione. In un'economia guidata dall'informazione, il vantaggio competitivo di un'azienda è legato alla capacità di acquisire, analizzare e utilizzare le informazioni necessarie al processo decisionale in modo migliore e più rapido rispetto agli altri: le indicazioni strategiche sono estrapolate principalmente dalla mole dei dati operazionali contenuti nelle basi di dati aziendali. I software di Business Intelligence possono essere considerati uno strumento sempre più indispensabile

all'interno del processo decisionale dell'azienda.

Ciascuna delle fasi del processo decisionale di tipo strategico può trarre notevoli benefici dalla disponibilità di adeguati supporti informativi di tipo automatizzato. Nella fase di ricognizione, che precede la formulazione dei piani di sviluppo strategici, i sistemi informativi devono essere pronti a fornire una visione sintetica sia della realtà aziendale (reporting interno) sia delle variabili di mercato che più direttamente hanno relazione con lo sviluppo dell'impresa (monitoraggio dell'ambiente esterno). Per quanto riguarda il reporting di interesse strategico l'informatica deve assolvere a due compiti distinti: in primis deve migliorare il reporting sulle prestazioni interne dell'azienda, aumentandone l'accuratezza e la tempestività, in modo da cogliere puntualmente e senza ritardo i segnali allarmanti che dovessero manifestarsi all'interno dell'azienda. Contemporaneamente il sistema deve essere in grado di rispondere alla richieste di informazioni più analitiche e puntuali che possono contribuire ad una conoscenza più precisa dei fenomeni che hanno attinenza con il processo di pianificazione.

Importante, ai fini del miglioramento del processo decisionale di tipo strategico, è anche la disponibilità di modelli che raffigurano lo sviluppo delle attività future dell'impresa, sia al variare di vincoli ambientali sia in funzione di decisioni manageriali. La possibilità di utilizzare modelli basati sulla simulazione è di grandissimo aiuto nelle fasi del processo decisionale che riguardano la ricerca di possibili soluzioni e la valutazione e scelta dell'alternativa migliore per un determinato problema.

Il sistema informativo è di grande aiuto infine nella fase di controllo, che segue l'attuazione di una decisione strategica e precede ulteriori interventi correttivi, se gli esiti delle decisioni prese non dovessero corrispondere alle aspettative (controllo strategico).

4.2 Le necessità delle aziende

Nel corso del tempo, le organizzazioni hanno sempre manifestato il bisogno di gestire flussi di informazioni provenienti tanto dall'interno quanto dall'esterno dei confini aziendali. Le soluzioni che si sono succedute, e che ancora oggi in alcuni casi si alternano, hanno carattere diverso: si va dalla semplice aggregazione dei dati, a un'analisi degli stessi attraverso fogli di calcolo che restituiscono informazioni, a sistemi più complessi che consentono un monitoraggio più ampio della performance aziendale e dei trend ambientali. Viene spontaneo chiedersi perché le soluzioni adottate dalle aziende siano mutate a tal punto nel corso del tempo. La risposta risiede per lo più nell'evoluzione del contesto: le aziende non agiscono in modo isolato, al contrario esse sono collocate all'interno di un ambiente, inteso come insieme di tutti gli elementi posti al di fuori dei confini dell'organizzazione ma capaci di influenzare l'organizzazione stessa o una sua parte. In senso lato, tutto quanto accade nell'ambiente generale può incidere indirettamente sulle performance dell'azienda. Ciò ci induce a riflettere sulla reale complessità che ogni organizzazione deve affrontare e sulla conseguen-

te necessità di monitorare l'ambiente esterno, al fine di rispondere in modo adeguato ai suoi mutamenti.

Le grandi imprese, con un passato quasi secolare alle spalle, rimpiangono probabilmente l'epoca in cui, conquistato il successo, si poteva permanere in tale condizione di privilegio senza molti sforzi per parecchi decenni. E' lecito quindi domandarsi quali minacce, nell'attuale contesto socio-economico, insidino un successo stabile e duraturo. Le insidie alla stabilità sono rappresentate in sintesi da tre diversi fattori:

- la rapidità con cui si manifesta il cambiamento tecnologico;
- la variabilità che caratterizza la domanda del mercato;
- la creatività e la capacità di innovazione derivante dall'esasperata concorrenza.

La rapidità con cui avviene il cambiamento tecnologico è dovuta, in una certa misura, alla velocità con cui circolano le idee nella comunità tecnico-scientifica e alla capacità di assorbimento dell'innovazione da parte sia delle imprese sia dei consumatori. Le aziende produttrici appaiono instancabili nella loro corsa verso prodotti tecnologicamente più avanzati e dalle prestazioni migliori. Anche la domanda di mercato non segue più cicli lenti, ma è soggetta a rapidi cambiamenti; l'aumentare della competizione nei mercati stimola infine la creatività dei produttori che desiderano differenziarsi dalla concorrenza.

In sostanza, se il monitoraggio dell'ambiente è sempre stato essenziale al fine di una discreta performance aziendale, oggi lo è ancor di più a causa delle particolari condizioni ambientali in cui ci troviamo.

Quello che le aziende osservano con preoccupazione è che il livello di complessità del contesto socio-economico in cui esse operano sta crescendo inesorabilmente. Le aziende devono adattarsi a vivere nel contesto corrente, sviluppando le caratteristiche idonee a fronteggiare le nuove condizioni ambientali. Tali caratteristiche sono indicate come competenze, ovvero capacità e abilità nel gestire le problematiche che si presentano.

L'aumento della complessità ha accresciuto il fabbisogno di informazioni puntuali e tempestive per la gestione dell'impresa, in special modo con riferimento alle scelte strategiche. L'elevata complessità impone all'azienda di mantenersi sempre vigile e attenta; il suo sistema informativo deve essere pronto a cogliere e segnalare le varie minacce e opportunità.

Conoscere il proprio ambiente significa in primo luogo essere in grado di raccogliere, aggregare ed analizzare le informazioni provenienti dall'esterno. In secondo luogo, significa saper dare una risposta adeguata ai mutamenti in corso, sia a quelli più ampi sopra citati, sia a quelli più specifici della propria base clienti. All'interno dell'azienda il processo decisionale ha inizio da questo flusso di informazioni, il quale però diventa sempre più articolato e coinvolge sempre più interlocutori: a questo punto i decision maker necessitano di un supporto concreto, capace di fornire in tempi immediati una fotografia della situazione

attuale e una previsione dei trend successivi. Tale supporto viene tipicamente fornito dai sistemi di Business Intelligence.

Un aspetto molto importante per i sistemi di business intelligence riguarda il grado di dettaglio con cui le informazioni vengono messe a disposizione ai vari decision maker: i responsabili delle varie aree aziendali sono generalmente interessati ai dati con elevati livelli di dettaglio mentre per i direttori potrebbero essere più importanti i dati relativi all'andamento generale del loro settore o dell'azienda. I sistemi di business intelligence devono pertanto essere in grado di fornire i dati con la granularità richiesta dai vari operatori.

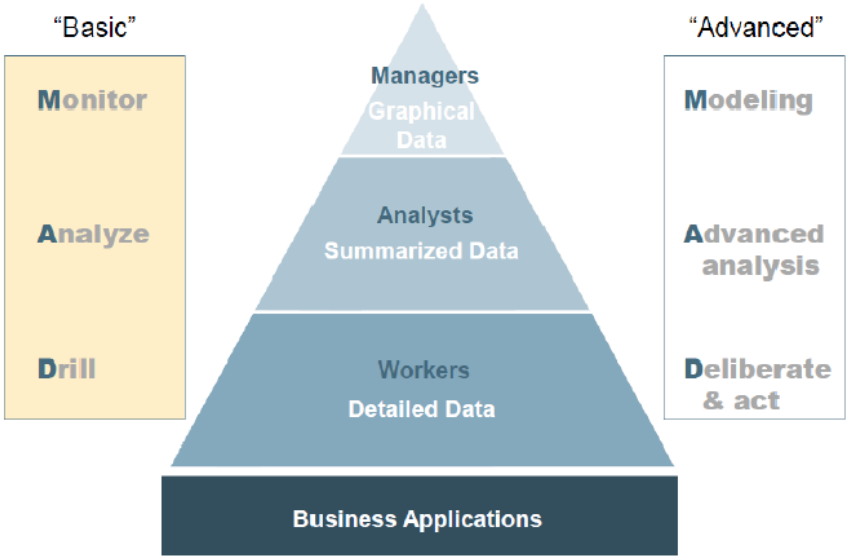

Figura 4.1: BI: accuratezza delle informazioni richiesta

Questo aspetto è molto importante in quanto consente di mettere a disposizione ai vari operatori le informazioni più interessanti per i loro ambiti di analisi. Capita molto spesso che le informazioni messe a disposizione dal sistema informativo non vengano prese in considerazione perché non sono nel formato desiderato o, ad una prima occhiata, risultano inutili.

4.3 La Business Intelligence in Unicomm

All'interno del gruppo Unicomm già da diversi anni si adottano strumenti di business intelligence a supporto alle decisioni. Tali strumenti di analisi si sono sviluppati tra loro però in modo indipendente e con tempistiche diverse nei vari settori aziendali. Questo sviluppo disomogeneo ha portato ad una situazione che presenta più soluzioni di DWH, implementate nel tempo in diverse aree

aziendali, con basi dati separate, flussi di alimentazione indipendenti e front-end eterogenei.

Questa frammentazione delle soluzioni porta a delle criticità nella gestione e nell'utilizzo di questi sistemi in quanto:

- la conoscenza delle diverse soluzioni è frammentata tra gli utenti finali, i fornitori che le hanno realizzate e le diverse figure professionali all'interno dei sistema informativo (sistemista, referente business, operatore helpdesk);
- i dati all'interno dei vari sistemi di data warehouse sono duplicati e in molti casi inconsistenti (ad esempio la tabella degli articoli varia a seconda sia quella relativa alle vendite o agli acquisti);
- gli utilizzatori finali sono costretti a conoscere e utilizzare applicazioni diverse per reperire i report; questo provoca confusione nell'utilizzo degli strumenti software.

Visti i molteplici problemi dovuti alla frammentazione degli attuali strumenti di BI e cogliendo l'occasione del cambio del sistema informativo commerciale, la direzione aziendale, su suggerimento dei sistemi informativi, ha deciso di far evolvere la business intelligence aziendale verso una soluzione di tipo "enterprise": unica infrastruttura tecnologica, basi dati integrate e condivise, strumenti di analisi uniformi e supporto centralizzato (figure professionali adeguatamente formate).

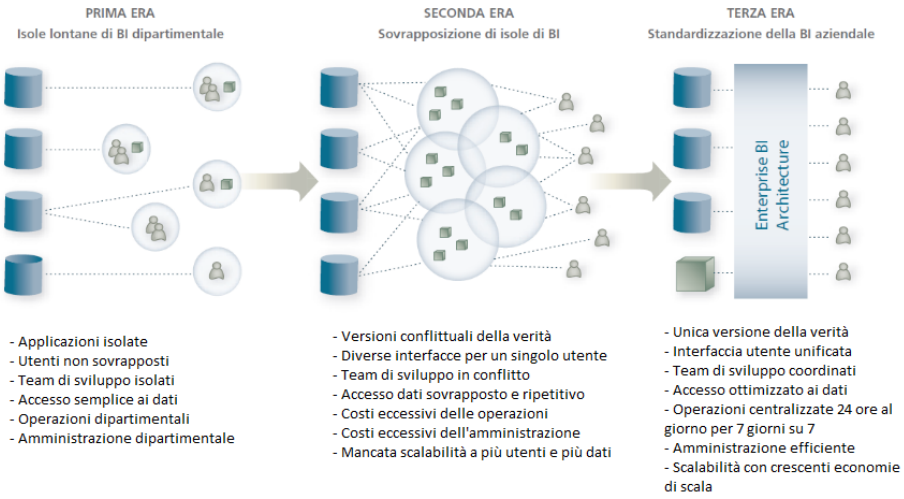

Figura 4.2: Evoluzione della business intelligence aziendale.

Il processo di cambiamento è agevolato in quanto il nuovo sistema informativo commerciale (Me.R.Sy.) prevede il supporto di soluzione DWH/BI per

la generazione di report "complessi", l'analisi dei dati e la pubblicazione di indicatori di performance (KPI).

Questo processo di cambiamento è in fase di svolgimento e si prevede finirà nei primi mesi del 2013; una volta concluso i report e gli strumenti di BI per tutti i settori aziendali saranno disponibili nel portale aziendale di business intelligence consentendo l'accesso ai dati anche agli agenti operanti sul territorio.

Capitolo 5

Data warehouse

Tra i sistemi di supporto alle decisioni, i sistemi di data warehousing sono probabilmente quelli su cui negli ultimi anni si è maggiormente focalizzata l'attenzione sia nel mondo accademico sia in quello industriale. E' possibile definire il data warehousing come una collezione di metodi, tecnologie e strumenti di ausilio al cosiddetto "lavoratore della conoscenza" per condurre analisi dei dati finalizzate all'attuazione di processi decisionali ed al miglioramento del patrimonio informativo. Le esigenze che ne hanno decretato la nascita furono quelle di ottenere le informazioni necessarie nella forma desiderata trovando il modo di accedere in maniera efficace alla montagna di dati posseduta.

In un contesto aziendale medio-grande sono tipicamente presenti più basi di dati, ciascuna relativa ad una diversa area del business, spesso memorizzate su piattaforme logico-fisiche differenti e non integrate dal punto di vista concettuale. I risultati prodotti all'interno delle diverse aree saranno allora, molto probabilmente, inconsistenti tra loro. Le parole chiave che diventano fattori distintivi e requisiti indispensabili del processo di data warehousing, ossia del complesso di attività che consentono di trasformare i dati operazionali in conoscenza a supporto delle decisioni, sono:

- accessibilità a utenti con conoscenze limitate di informatica e strutture dati;
- integrazione dei dati sulla base di un modello standard dell'impresa;
- flessibilità di integrazione per trarre il massimo vantaggio dal patrimonio informativo esistente;
- sintesi per permettere analisi mirate ed efficaci;
- rappresentazione multidimensionale per offrire all'utente una visione intuitiva ed efficacemente manipolabile delle informazioni;
- correttezza e completezza dei dati integrati.

Al centro del processo vi è il data warehouse, un contenitore di dati che diventa garante dei requisiti esposti.

Inmon ne diede una definizione nel 1996: un data warehouse (DWH) è una collezione di dati di supporto per il processo decisionale che presenta le seguenti caratteristiche:

- è orientata ai soggetti di interesse;
- è integrata e consistente;
- è rappresentativa dell'evoluzione temporale e non volatile.

Il DWH è orientato ai soggetti in quanto si incentra sui concetti di interesse dell'azienda, quali clienti, prodotti, vendite e ordini; i database operazionali sono organizzati invece intorno alle differenti applicazioni del dominio aziendale. La condizione di integrità e consistenza è molto importante in quanto il DWH si impegna a restituire una visione unificata dei dati provenienti da più fonti eterogenee: dati estratti dall'ambiente di produzione, e quindi originariamente archiviati in basi di dati aziendali, o provenienti da sistemi informativi esterni all'azienda. La costruzione di un sistema di data warehousing non comporta l'inserimento di nuove informazioni bensì la riorganizzazione e la gestione di quelle esistenti. Un data warehouse può essere sia consultato direttamente che usato come sorgente per costruirne delle parziali repliche orientate verso specifiche aree dell'impresa. Tali repliche vengono dette data mart, ovvero un sottoinsieme o un aggregazione dei dati presenti nel data warehouse primario, contenente l'insieme delle informazioni rilevanti per una particolare area del business, una particolare divisione dell'azienda o una particolare categoria di soggetti.

5.1 Componenti di un data warehouse

Il data warehouse è composto da quattro elementi base, ognuno dei quali ha le proprie funzionalità e il proprio ruolo all'interno del sistema.

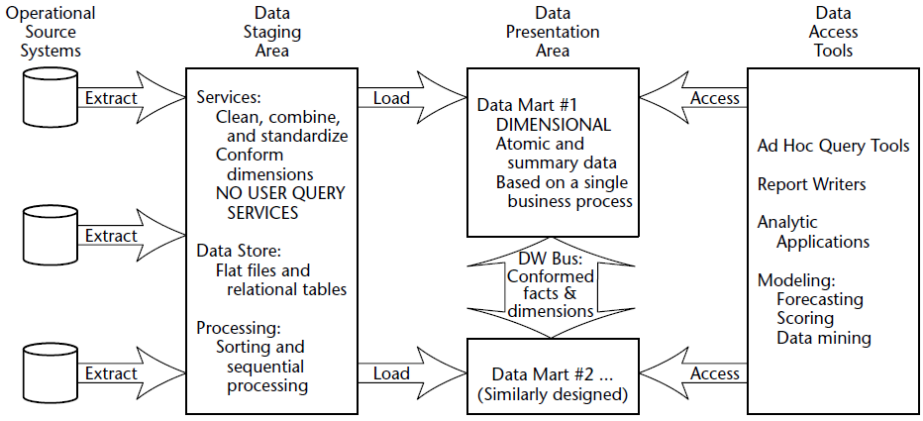

Figura 5.1: Elementi base di un data warehouse

- Sistemi sorgente: sono costituiti dai sistemi gestionali e amministrativo-contabili di tipo tradizionale o ERP, dai sistemi che interfacciano il mercato (sistemi di CRM), dai sistemi Web e da tutti gli altri sistemi informativi di tipo operativo e/o transazionale. Devono essere visti come parti esterne rispetto al sistema di data warehousing poiché probabilmente si avrà poco o nessun controllo sul contenuto e la forma dei dati che essi contengono. Questi sistemi solitamente mantengono pochi dati storici. Avere a disposizione un buon data warehouse solleva la gran parte della responsabilità di rappresentare il passato ai sistemi sorgente.

- Staging area: è composta da due parti: un'area di memorizzazione dei dati e un insieme di procedure comunemente dette extraction-transformation-loading (ETL). Si colloca tra i sistemi sorgente e l'area di presentazione. Kimball[1] paragona la staging area alla cucina di un ristorante, nella quale gli ingredienti vengono trasformati per un buon pasto. I dati operazionali vengono infatti trasformati ed elaborati per produrre informazioni utili all'azienda. La staging area è accessibile solamente a professionisti qualificati e pertanto risulterà essere off-limits per gli utenti business; non sarà predisposta inoltre per servizi di interrogazione e di presentazione.

- Area di presentazione: è la parte dove i dati sono organizzati, conservati e resi disponibili per l'interrogazione diretta da parte di utenti, autori di report e altre applicazioni analitiche. Per gli utilizzatori finali l'area di presentazione coincide con il data warehouse in quanto è tutto quello che possono vedere e toccare mediante gli appositi strumenti in loro possesso.

- Strumenti di accesso ai dati: è l'insieme degli strumenti di front-end che gli utenti business hanno a loro disposizione per consultare l'area di pre-

[1]Ralph Kimball è universalmente riconosciuto come uno dei "guru" del data warehouse; ha scritto "The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling".

sentazione. Possono essere semplici strumenti per eseguire query ad hoc oppure strumenti che eseguono analisi più complesse. Nell'80-90 per cento dei casi tuttavia gli utenti utilizzano applicazioni che forniscono automaticamente e ad intervalli di tempo prestabiliti informazioni strutturate in modo pressoché invariabile e che quindi non implicano la costruzione diretta di query.

5.2 Architetture per il data warehousing

Come accennato nel paragrafo precedente la presenza e le modalità di utilizzo della staging area definiscono l'architettura del sistema di data warehousing. La scelta dell'architettura da utilizzare dipende dalle esigenze e dal tipo di organizzazione entro la quale il progetto dovrà essere realizzato; esistono tuttavia delle caratteristiche indispensabili per un sistema di data warehousing:

- SEPARAZIONE: l'elaborazione analitica e quella transazionale devono essere mantenute il più possibile separate;
- SCALABILITÀ: l'architettura hardware e quella software devono poter essere facilmente ridimensionate a fronte della crescita nel tempo dei volumi di dati da gestire ed elaborare e del numero di utenti da soddisfare;
- ESTENDIBILITÀ: deve essere possibile accogliere nuove applicazioni e tecnologie senza riprogettare integralmente il sistema;
- SICUREZZA: il controllo sugli accessi è essenziale a causa della natura strategica dei dati memorizzati;
- AMMINISTRABILITÀ: la complessità dell'attività di amministrazione non deve risultare eccessiva.

5.2.1 Architettura a un livello

E' un'architettura scarsamente utilizzata nella pratica. Ha come obiettivo quello di minimizzare la quantità di dati memorizzati eliminando le ridondanze.

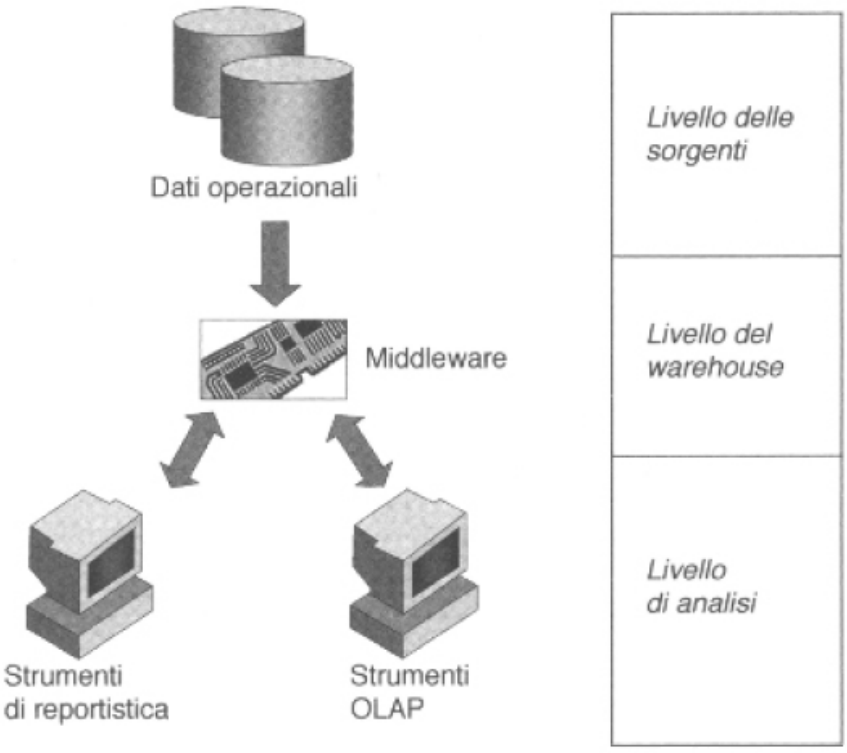

Figura 5.2: Architettura a un livello

Come mostrato nella figura il data warehouse in questo caso è virtuale, poiché viene implementato come una vista multidimensionale dei dati operazionali da un apposito middleware. Una tale architettura presenta i seguenti punti deboli:

- non rispetta il requisito di separazione dell'elaborazione analitica da quella transazionale: le interrogazioni di analisi vengono infatti ridirette sui dati operazionali dopo essere state reinterpretate dal middleware, interferendo così con il normale carico di lavoro transazionale;
- i requisiti di integrazione e correttezza dei dati possono essere soddisfatti, ma con un'elevata complessità;
- è impossibile avere un livello di storicizzazione superiore a quello delle sorgenti.

5.2.2 Architettura a due livelli

Con questa architettura si riesce a soddisfare il requisito di separazione, come si può notare nella figura sottostante.

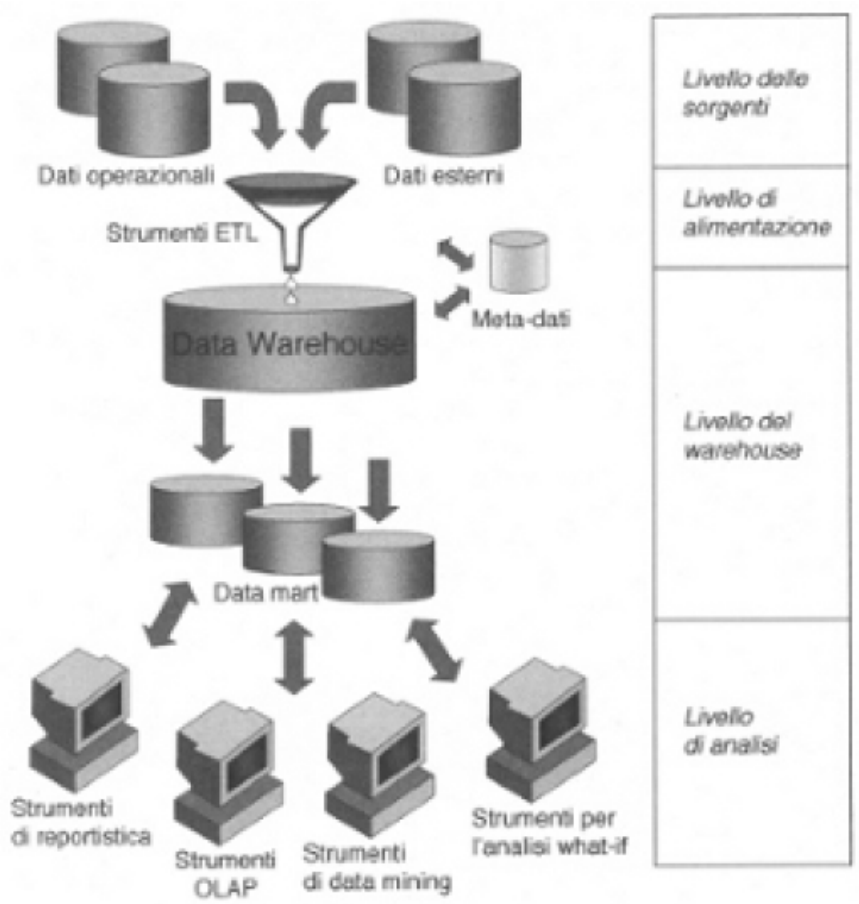

Figura 5.3: Architettura a due livelli

Nonostante si articoli su quattro livelli distinti viene chiamata architettura a due livelli per evidenziare la separazione tra le sorgenti e il data warehouse. I dati che il data warehouse utilizzerà sono contenuti in database aziendali relazionali o legacy, oppure provenienti da sistemi informativi esterni all'azienda (livello delle sorgenti). Tali dati saranno estratti, ripuliti per eliminare le inconsistenze e completare eventuali parti mancanti, integrati per fondere sorgenti eterogenee secondo uno schema comune, mediante gli strumenti ETL (livello di alimentazione). Le informazioni vengono raccolte nel data warehouse che potrà essere direttamente consultato o usato come sorgente per costruire data mart (livello del warehouse). Accanto al DWH, il contenitore dei metadati mantiene informazioni sulle sorgenti, sui meccanismi di accesso, sulle procedure di pulitura e alimentazione, sugli utenti, sugli schemi dei data mart ecc. Infine si potranno consultare in modo efficiente e flessibile i dati integrati a fini di stesura di report, di analisi e di simulazione (livello di analisi).

5.2.3 Architettura a tre livelli

Al livello delle sorgenti e quello del data warehouse, viene aggiunto un terzo livello che viene chiamato livello dei dati riconciliati. Questo livello materializza i dati operazionali ottenuti a valle del processo di integrazione e ripulitura dei dati sorgente. Il data warehouse non viene più alimentato direttamente dalle sorgenti, ma dai dati riconciliati.

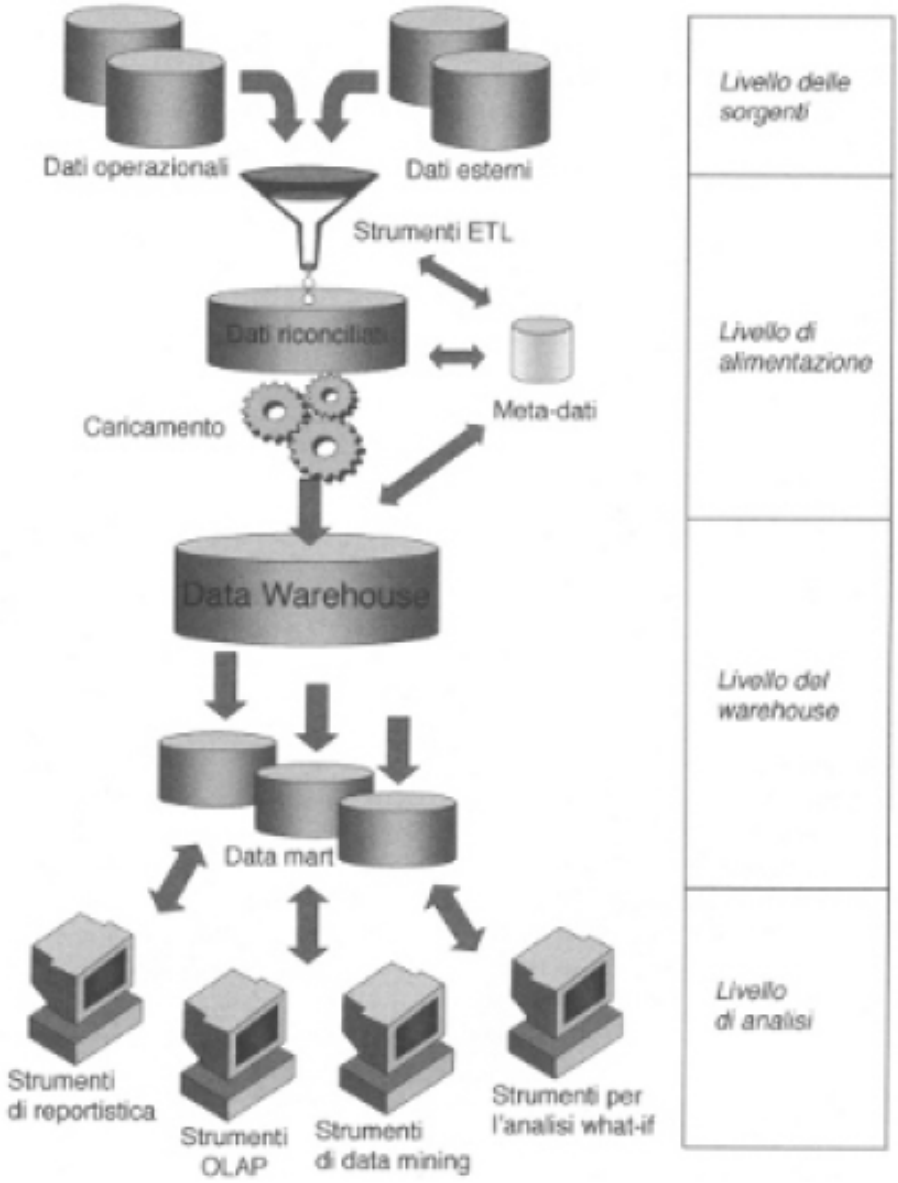

Figura 5.4: Architettura a tre livelli

Un vantaggio di questa architettura è che il livello dei dati riconciliati crea un modello di dati comune e di riferimento per l'intera azienda, introducendo al contempo una separazione netta tra le problematiche legate all'estrazione e integrazione dei dati dalle sorgenti e quelle inerenti l'alimentazione del DWH. Presenta tuttavia lo svantaggio di introdurre un'ulteriore ridondanza rispetto

ai dati operazionali sorgente.

5.3 Gli strumenti ETL

Il ruolo degli strumenti ETL (Extraction-Trasformation-Loading) è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il data warehouse. Le procedure di popolamento del data warehouse possono raggiungere elevati livelli di complessità, in relazione alle discrepanze esistenti tra le sorgenti, al loro livello di correttezza e al livello di precisione (temporale) rappresentativa nel tempo che si desidera mantenere nel sistema informativo. Sono caratterizzate da una sequenza di fasi che dipendono dalle politiche di aggiornamento che si è deciso di adottare e che prevedono azioni più o meno articolate. La complessità di queste procedure è tale che sul mercato sono presenti diversi prodotti software orientati specificatamente al supporto delle fasi di estrazione, pulizia, trasformazione e caricamento dei dati nel processo di alimentazione del data warehouse. Occorre precisare che, nella letteratura, i confini tra pulitura e trasformazione sono spesso sfumati dal punto di vista terminologico, per cui è spesso poco chiara l'attribuzione di una specifica operazione all'uno o all'altro processo.

5.3.1 Estrazione

Le operazioni di estrazione sono eseguite all'atto dell'inizializzazione del livello riconciliato per essere poi ripetute periodicamente, in base all'intervallo di aggiornamento stabilito dal progettista, al fine di acquisire informazioni relative agli eventi verificatisi durante la vita del sistema. I dati che andranno a popolare il data warehouse sono solo quelli essenziali all'analisi e non tutti i dati ospitati sui sistemi di origine. Esistono due tipologie di approcci all'estrazione:

- estrazione statica: vengono trattati tutti i dati presenti nelle sorgenti operazionali. E' l'unica soluzione possibile all'atto dell'inizializzazione, ma può essere impiegata ogni qual volta la quantità ridotta dei dati lo permetta;
- estrazione incrementale: con questo approccio vengono presi in considerazione i soli dati prodotti o modificati dalle sorgenti nell'intervallo di tempo intercorso dall'ultimo aggiornamento del data warehouse. Può essere suddiviso ulteriormente in immediato e ritardato, in base al momento in cui viene registrata una modifica ai dati.

L'estrazione incrementale si basa generalmente sui log mantenuti dal DBMS transazionale, l'estrazione può essere guidata inoltre dalle sorgenti nei casi in cui è possibile ricevere, in modo asincrono, le notifiche delle modifiche dalle applicazioni operazionali.

5.3.2 Pulizia

Spesso i dati provenienti dalle sorgenti non sono di qualità adeguata agli standard richiesti per il sistema informativo, per questo devono essere applicate delle analisi in grado di rilevare, e possibilmente correggere, le situazioni che potrebbero essere critiche o che potrebbero condurre ad errori. Tra gli errori e le inconsistenze tipiche che rendono "sporchi" i dati si segnalano:

- dati duplicati (per esempio un cliente che compare più volte nell'anagrafica clienti);
- inconsistenza tra valori logicamente associati (per esempio tra i dati della persona ed il suo codice fiscale);
- dati mancanti (per esempio la professione di un cliente);
- uso non previsto di un campo (per esempio il campo destinato al codice fiscale usato per memorizzare il numero di telefono dell'ufficio);
- valori impossibili o errati (per esempio 30/02/2011);
- valori inconsistenti per la stessa entità dovuti a differenti convenzioni (per esempio la nazione indicata mediante sigla piuttosto che con il nome completo) e abbreviazioni (per esempio "Piazza Garibaldi" e "P.za Garibaldi");
- valori inconsistenti per la stessa entità dovuti a errori di battitura (per esempio "Piazza Garibaldi" e "Piaza Garibaldi").

Per correggere errori di scrittura e riconoscere i sinonimi vengono utilizzati degli appositi dizionari, mentre per stabilire le corrette corrispondenze tra valori vengono applicate regole proprie del dominio applicativo.

5.3.3 Trasformazione

Durante la fase di trasformazione vengono eseguite le trasformazioni necessarie a conformare i dati delle sorgenti alla struttura del data warehouse; in caso di architettura a tre livelli l'output di questa fase è il livello dei dati riconciliati. La presenza di più fonti eterogenee complica notevolmente questa fase, in quanto viene richiesta una complessa fase di integrazione. Nella fase di trasformazione possono essere effettuate molte operazioni, tra cui:

- *conversione e normalizzazione*: operano a livello di formato di memorizzazione e di unità di misura per uniformare i dati;
- *matching*: stabilisce le corrispondenze tra campi equivalenti in sorgenti diverse;
- *selezione*: riduce il numero di campi e di record rispetto a quello nelle sorgenti.

Negli strumenti ETL le attività di pulitura e trasformazione sono spesso allacciate e sovrapposte.

5.3.4 Caricamento

E' la fase in cui i dati vengono caricati nel data warehouse. Questa procedura può avvenire in due modalità:

- *refresh*: i dati vengono completamente riscritti all'interno del DWH. Viene solitamente utilizzata insieme all'estrazione statica durante la fase di inizializzazione;
- *update*: vengono aggiunti al DWH i soli cambiamenti verificatisi nelle sorgenti operazionali. Questa tecnica viene solitamente utilizzata in abbinamento all'estrazione incrementale al fine di ottenere un aggiornamento periodico del DWH.

5.4 Il modello multidimensionale

La progettazione di data warehouse e data mart si basa sul paradigma di rappresentazione multidimensionale dei dati in grado di offrire un duplice vantaggio: sotto il profilo funzionale risulta efficace per garantire tempi di risposta rapidi a fronte di interrogazioni complesse; sul piano logico le dimensioni corrispondono in modo naturale ai criteri di analisi utilizzati dai knowledge worker.

Il modello multidimensionale si basa sul fatto che gli oggetti che influenzano il processo decisionale sono fatti del mondo aziendale, come ad esempio le vendite o le spedizioni. Le occorrenze di un fatto vengono dette eventi: ogni singola vendita o spedizione effettuata è un evento. Per ciascun fatto interessano in particolare i valori di un insieme di misure che descrivono quantitativamente gli eventi. La quantità degli eventi all'interno di una azienda è troppo elevata per poter analizzare ogni singolo evento singolarmente; per questo motivo, per poterli agevolmente selezionare e raggruppare, si immagina di collocarli in uno spazio n-dimensionale in cui gli assi vengono chiamati appunto dimensioni di analisi (per esempio, nel caso in cui il fatto in questione siano le vendite, le dimensioni di analisi potrebbero essere: i prodotti, i negozi e le date).

Il concetto di dimensione genera la metafora del cubo. Un CUBO MULTIDIMENSIONALE è incentrato su un *fatto* di interesse per il processo decisionale. Esso rappresenta un insieme di *eventi*, descritti quantitativamente da misure numeriche. Ogni asse del cubo rappresenta una possibile *dimensione* di analisi.

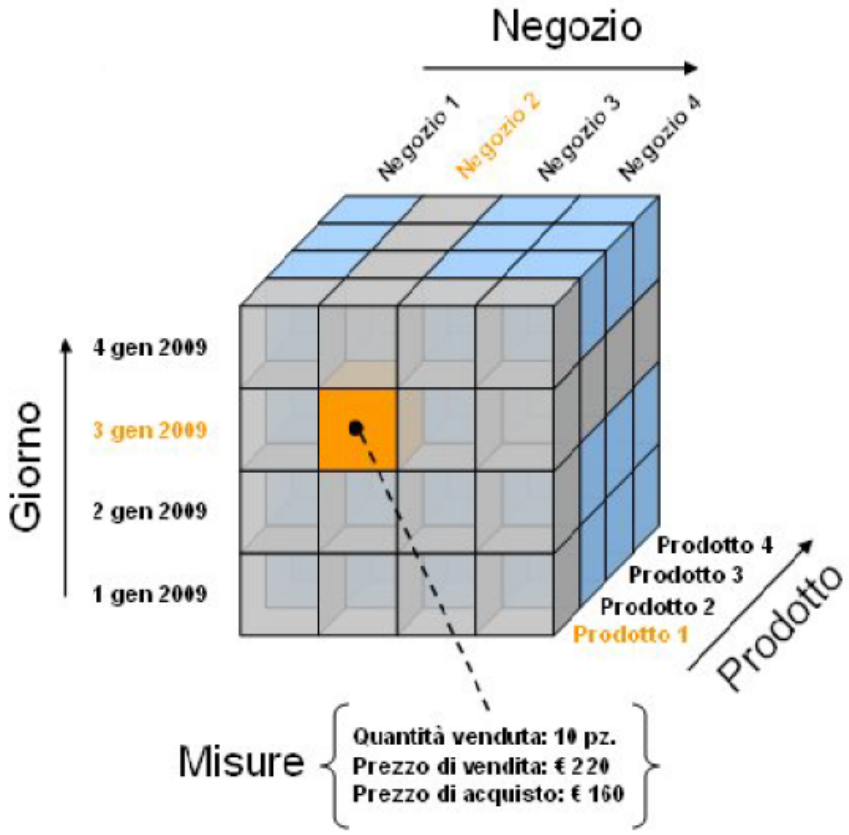

Figura 5.5: Cubo multidimensionale che modella le vendite di una catena di negozi

Se le dimensioni sono più di tre, si definirà un ipercubo.

Ogni dimensione è generalmente associata ad una gerarchia di livelli di aggregazione che ne raggruppa i valori in diversi modi; i livelli che compaiono nella gerarchia vengono detti attributi dimensionali.

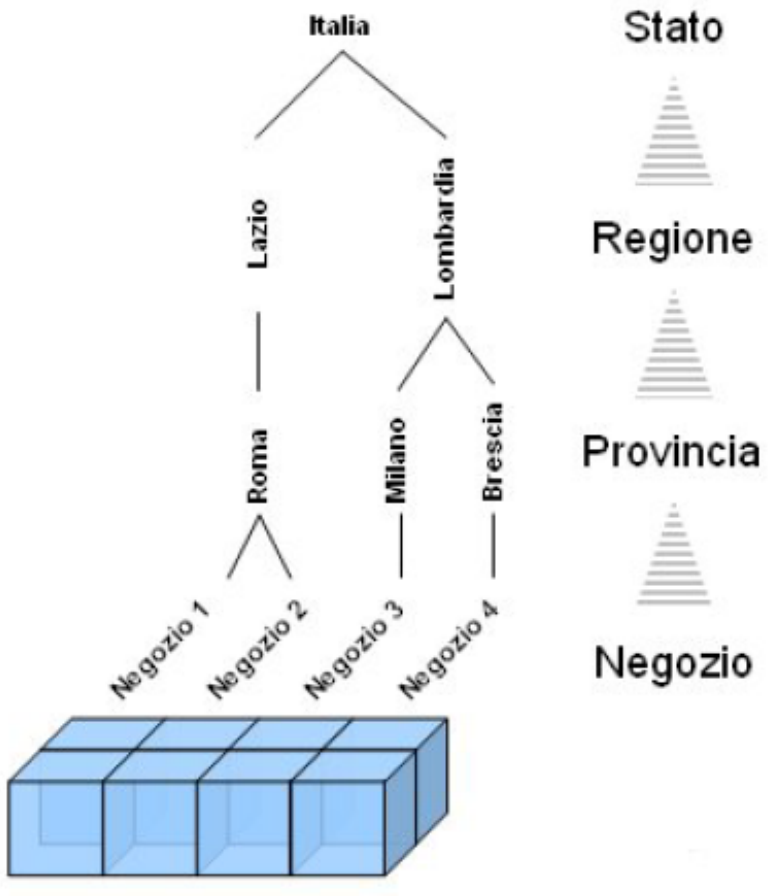

Figura 5.6: Una possibile gerarchia per la dimensione negozi

5.4.1 Modellazione concettuale: il Dimensional Fact Model

Un modello concettuale deve per definizione fornire una serie di strutture, dette costrutti, atte a descrivere la realtà di interesse in una maniera facile da comprendere e che prescinde dai criteri di organizzazione dei dati nei calcolatori. Il modello Entità/Relazione è un modello concettuale molto diffuso nelle imprese per la progettazione e documentazione di basi di dati relazionali. E' invece ormai universalmente riconosciuto che un data mart si appoggia su una visione multidimensionale dei dati. Il modello Entità/Relazione non risulta essere adatto a tale scopo in quanto non è in grado di mettere correttamente in luce gli aspetti peculiari del modello multidimensionale, senza contare che risulterebbe poco economico dal punto di vista grafico-notazionale.

Il Dimensional Fact Model (DFM), proposto da Golfarelli nel 1998, è un modello concettuale appositamente concepito per fungere da supporto alla progettazione di data mart; è essenzialmente di tipo grafico, e può essere considerato come una specializzazione del modello multidimensionale per applicazioni di data warehousing. La rappresentazione concettuale generata dal DFM consiste in un insieme di schemi di fatto.

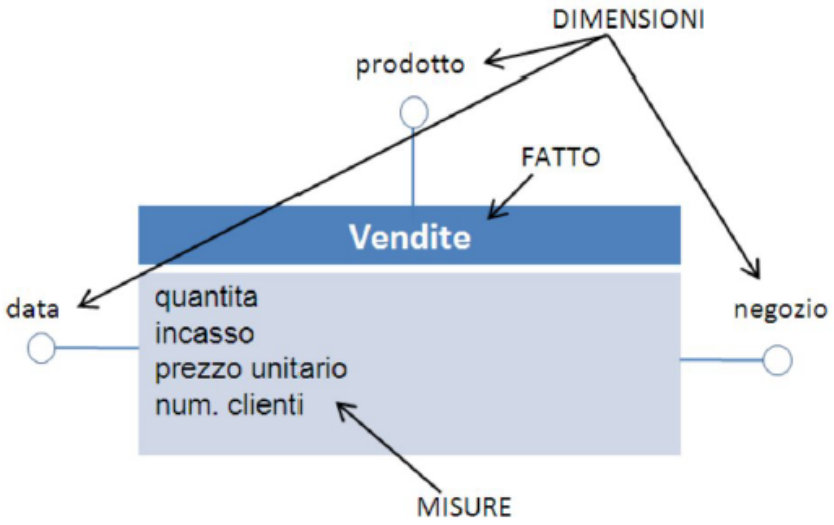

Figura 5.7: Esempio di schema di fatto delle vendite

Gli elementi di base modellati dagli schemi di fatto sono:

- *fatto*: è un concetto di interesse per il processo decisionale che modella un insieme di eventi che si verificano in una realtà aziendale (per esempio per un'azienda che lavora nel campo del commercio i fatti possono essere: le vendite, le spedizioni, gli acquisti, i reclami);
- *misura*: è una proprietà numerica di un fatto e ne descrive un aspetto quantitativo di interesse per l'analisi. Ad esempio una vendita può essere misurata dal numero di unità vendute, dal prezzo unitario e dallo sconto applicato. Un fatto può anche essere privo di misure, in questo caso si registra solo il verificarsi di un evento;
- *dimensione*: è una proprietà con dominio finito di un fatto e ne descrive una coordinata di analisi. E' caratterizzata da numerosi attributi generalmente di tipo categorico. Un fatto ha in genere più dimensioni che ne determinano la granularità minima di rappresentazione. Ad esempio le dimensioni tipiche per il fatto di vendita sono il prodotto, il negozio e la

data. In questo esempio l'informazione elementare che può essere rappresentata riguarda le vendite di un prodotto effettuate in un negozio in un determinato giorno. Non è possibile distinguere tra vendite effettuate da impiegati diversi o in differenti ore del giorno;

- *gerarchia*: è un albero direzionato i cui nodi sono attributi dimensionali e i cui archi modellano associazioni molti-a-uno tra coppie di attributi dimensionali;
- *attributi dimensionali*: sono i campi non chiave di una tabella dimensione e memorizzano gli attributi dei membri.

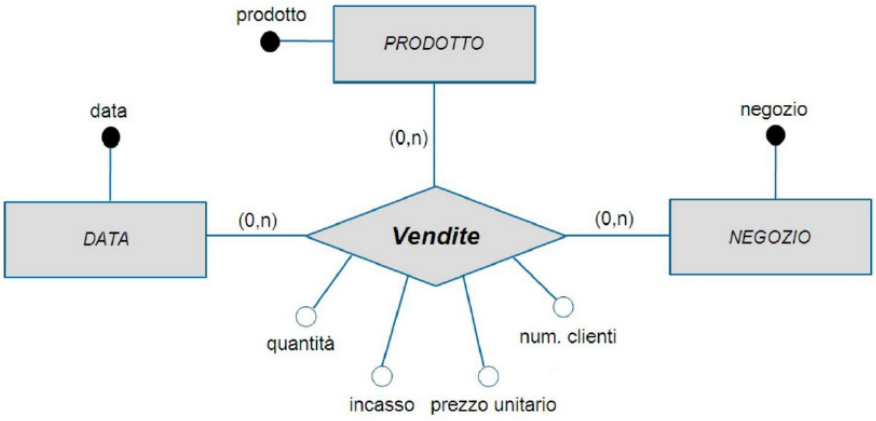

Figura 5.8: Schema Entità/Relazione relativo alle vendite

Come si può osservare nella figura che illustra lo schema di fatto relativo alle vendite, un fatto è raffigurato da un rettangolo che ne riporta il nome insieme ai nomi delle eventuali misure; le dimensioni sono rappresentati da piccoli cerchi collegati al fatto tramite linee. E' importante evidenziare come un fatto esprime un'associazione molti-a-molti tra le dimensioni. Per tale motivo lo schema Entità/Relazione corrispondente ad uno schema di fatto consiste in un'associazione n-aria tra entità che modellano le dimensioni. Le gerarchie vengono rappresentate da alberi direzionati i cui nodi sono attributi dimensionali e i cui archi modellano associazioni molti-a-uno tra coppie di attribuiti dimensionali.

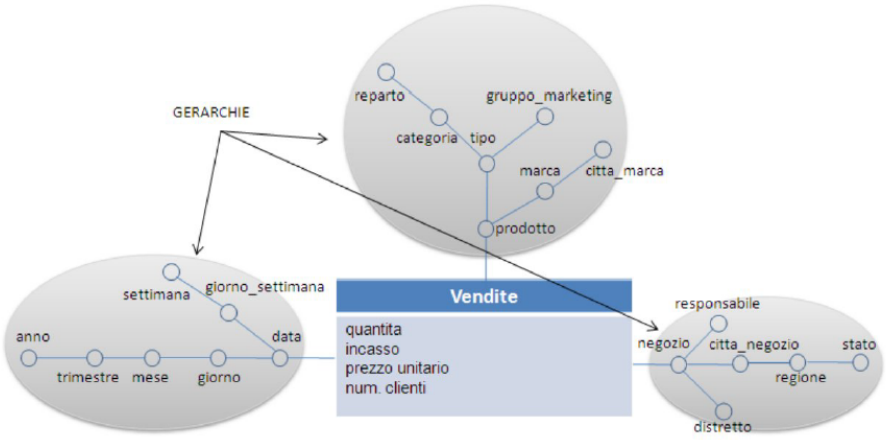

Figura 5.9: Schema di fatto delle vendite arricchito inserendo le gerarchie

Se si volesse tradurre questo schema di fatto nel corrispondente schema E/R si avrebbe un'esplosione in termini grafico-notazionali, come già si era accennato in precedenza. Tutti gli attributi dimensionali all'interno di uno schema di fatto devono avere nomi diversi tra loro. Nomi uguali possono essere differenziati qualificandoli con il nome di un attributo dimensionale che li precede nella gerarchia (per esempio, città negozio e città marca).

5.4.2 Modellazione logica

Mentre per la fase di modellazione concettuale non ci si deve preoccupare delle scelte che si dovranno fare durante la fase di modellazione logica, per quest'ultima non si può dire la stessa cosa. Sarà infatti in questa fase che si dovrà scegliere il DBMS da utilizzare durante la progettazione fisica. I dati soggetti ad analisi possono essere rappresentati secondo due modelli logici: quello relazionale, che da luogo ai cosiddetti sistemi ROLAP (Relational OLAP[2]), e quello multidimensionale, per il quale i sistemi utilizzati vengono detti MOLAP (Multidimensional OLAP). Esiste anche una terza soluzione, intermedia alle due appena menzionate, il cosiddetto HOLAP (Hybrid OLAP).

5.4.2.1 I sistemi ROLAP

Adottare una soluzione di questo genere implica il dover modellare i concetti multidimensionali osservati finora in elementi bidimensionali, ovvero le tabel-

[2]On-Line Analytical Processing: insieme di tecniche software per l'analisi interattiva e veloce di grandi quantità di dati. Gli strumenti OLAP si differenziano dagli OLTP (On-Line Transaction Processing) per il fatto che i primi hanno come obiettivo la performance nella ricerca e il raggiungimento di un'ampiezza di interrogazione quanto più grande possibile; i secondi invece hanno come obiettivo la garanzia di integrità e sicurezza delle transazioni.

le del modello relazionale. Una tale operazione viene effettuata mediante il cosiddetto star schema.

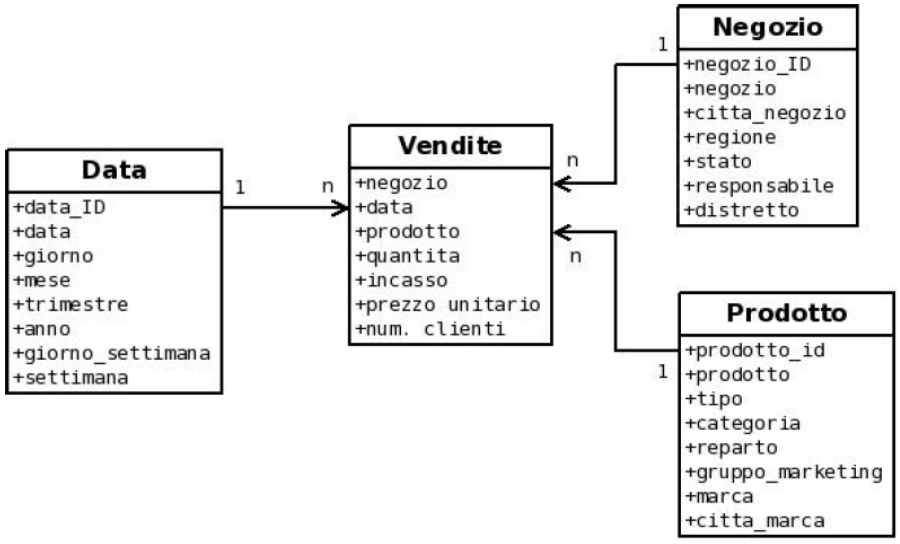

Figura 5.10: Star schema delle vendite

Uno schema a stella è composto da:

- un insieme di tabelle chiamate tabelle delle dimensioni (DIMENSION TABLE). Ciascuna di queste tabelle è caratterizzata da una chiave primaria e da un insieme di attributi che descrivono le dimensioni di analisi a diversi livelli di aggregazione;
- una tabella chiamata tabella dei fatti (FACT TABLE) in cui sono presenti le chiavi di tutte le tabelle delle dimensioni. La chiave primaria di questa tabella sarà data dall'insieme delle chiavi esterne delle dimension table. La tabella dei fatti contiene inoltre un attributo per ogni misura.

La visione multidimensionale si ottiene eseguendo un join tra la fact table e le dimension table.

Si noti come le dimension table violino la terza forma normale, ovvero contengano attributi che dipendono transitivamente da una chiave. Una tale situazione introduce una ridondanza e per tanto richiede più spazio per la memorizzazione dei dati, ma allo stesso tempo richiede un minor numero di join per reperire le informazioni. Si potrebbe però essere interessati ad avere uno schema logico più vicino agli enunciati della teoria relazionale; lo snowflake schema lo permette in quanto caratterizzato da una parziale normalizzazione delle dimension table.

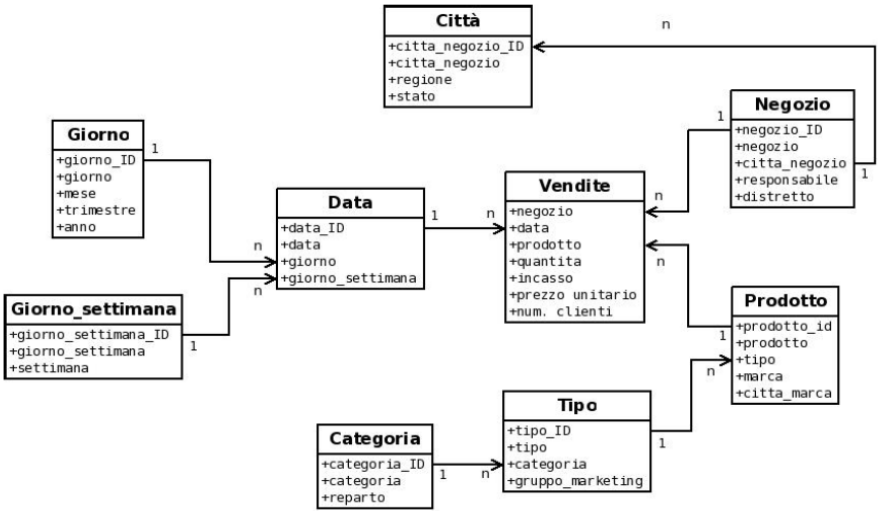

Figura 5.11: Snowflake schema ottenuto mediante una parziale normalizzazione dello star schema relativo alle vendite

Uno schema snowflake è ottenibile da uno schema a stella scomponendo una o più dimension table in più tabelle, in modo tale da eliminare alcune delle dipendenze funzionali transitive in esse presenti. Le tabelle delle dimensioni le cui chiavi sono importate nella fact table vengono dette primarie, mentre chiameremo secondarie le rimanenti. In questo modo è possibile trovare il giusto compromesso tra spazio in memoria utilizzato e numero di join da effettuare per ricavare l'informazione desiderata. Si noti come ad ogni passo di normalizzazione corrisponda un arco nello schema di fatto e una sotto-gerarchia che invece verrà memorizzata in una tabella a parte. Affinché lo snowflaking sia efficace tutti gli attributi del sottoalbero dell'attributo da cui ha origine la normalizzazione devono essere spostati nella nuova relazione. La scelta di mappare elementi del mondo multidimensionale nel modello relazionale potrebbe apparire una forzatura; una tale scelta tuttavia è giustificata da un insieme di motivazioni di varia natura, prima fra tutte la constatazione che il modello relazionale è di fatto lo standard nel settore dei database. Inoltre l'evoluzione subita dai DBMS relazionali nell'arco degli anni ne fa degli strumenti estremamente raffinati ed ottimizzati.

5.4.2.2 I sistemi MOLAP

Nell'approccio MOLAP il data warehouse memorizza i dati usando strutture intrinsecamente multidimensionali: i dati vengono fisicamente memorizzati in vettori e l'accesso è di tipo posizionale. Il sistema alloca una cella per ogni possibile combinazione dei valori delle dimensioni e l'accesso ad un fatto avviene in modo diretto, sulla base delle coordinate fornite. L'utilizzo di una tale

soluzione rappresenta la soluzione naturale per un sistema di data warehousing e può fornire prestazioni ottimali, in quanto le operazioni di query multidimensionale non devono essere simulate mediante complesse istruzioni. Il principale problema a cui però è soggetta la soluzione MOLAP è la sparsità dei dati, come rappresentato in figura.

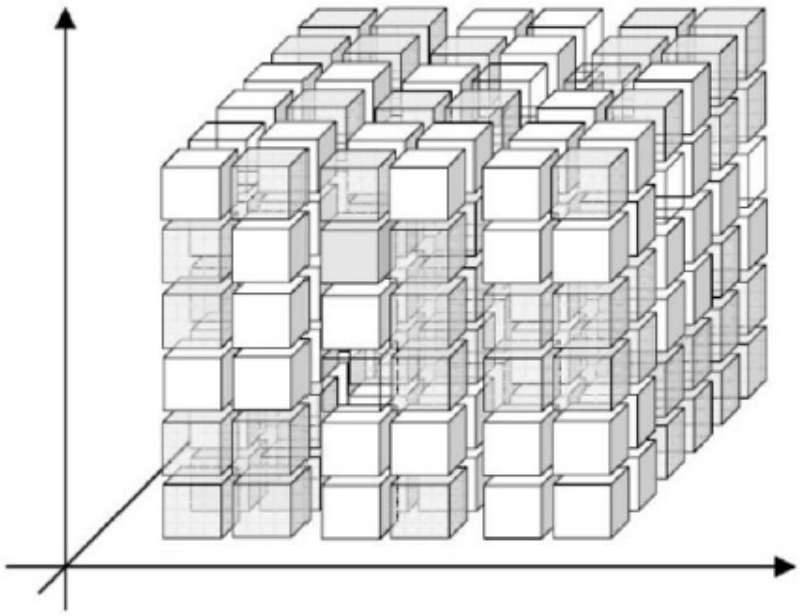

Figura 5.12: Rappresentazione del fenomeno di sparsità dei dati: in bianco le celle relative ad eventi effettivamente accaduti.

Mediamente in un cubo multidimensionale meno del 20% delle celle contiene effettivamente delle informazioni, mentre le restanti celle risultano essere vuote poiché corrispondono ad eventi non accaduti. La memorizzazione di celle non informative provoca uno spreco di spazio su disco. Il fenomeno della sparsità dei dati viene affrontato partizionando il cubo n-dimensionale in questione in più sottocubi n-dimensionali che vengono detti chunk. Si parla di chunk densi se la maggior parte delle celle contengono dati, chunk sparsi altrimenti. Un tale approccio permette di operare su blocchi di dati di dimensione inferiore e che quindi potranno essere caricati agevolmente in memoria. Si osserva però che la memorizzazione diretta di chunk sparsi comporta un notevole spreco di spazio dovuto alla rappresentazione delle celle che non contengono informazioni. Per questo motivo i chunk sparsi vengono utilizzati mediante un indice che riporta l'offset delle sole celle che contengono informazioni.

5.5 Strumenti utilizzati in Unicomm

Nel Gruppo Unicomm per la gestione del portale di business intelligence e del sistema di data warehouse vengono utilizzati i seguenti sistemi hardware e software:

- Il software di gestione delle merci Me.R.Sy. gira sulla seguente macchina:
  - Sistema operativo: IBM i V6R1M1;
  - CPU: power 6;
  - RAM: 128GB;
  - spazio disco: 11,3 TB;
  - versione DB: vedi S.O.
- Server Oracle (cluster) nel quale risiede il data warehouse:
  - Sistema operativo: Red Hat Enterprise Linux Server release 5.4;
  - CPU: Intel Xeon X5560 (2 x);
  - RAM: 48GB;
  - spazio disco: 2,2TB;
  - versione DB: Oracle 11gR2.
- Lo strumento ELT (Extraction-Loading-Transformation) è Oracle Data Integrator (ODI) versione 10.1.3.2.0 (2 = alta affidabilità). L'architettura di ODI consiste principalmente in:
  - 4 moduli grafici (Designer, Topology Manager, Operator, Security Manager);
  - 2 repository su database relazionale (master repository e work repository);
  - un agente schedulatore;
  - una applicazione web (metadata navigator).

  Server virtuale con Sistema Operativo: Windows Server 2003 SP2.
- Il portale di business intelligence è basato su SAP BI platform. Di questa suite attualmente vengono utilizzati i seguenti moduli:
  - BusinessObject BI package (1 produzione + 1 sviluppo/test):
    - BI Portal;
    - BO Web Intelligence (analisi interattiva dati);
    - BO Voyager (analisi MOLAP, anche su cubi diversi e db "esterni").

Capitolo 6

Modellazione del data warehouse

Concluse le fasi di analisi delle esigenze e dello studio teorico degli strumenti necessari alla realizzazione del data warehouse, si è passati alla fase di progettazione vera e propria.

6.1 Analisi pre-progettuale

Si sono svolti alcuni incontri con il consulente di Miriade, Vittorio Favero, per riassumere quanto emerso dalle analisi delle esigenze del marketing e per programmare e decidere come procedere con lo sviluppo del progetto.

Durante questi incontri è emerso che le analisi svolte sono più che sufficienti per la prima parte di sviluppo del progetto in quanto offrono una visione globale del problema e permettono di passare alla fase di modellazione concettuale. Successivamente, durante la fase implementativa del progetto, saranno necessari ulteriori incontri per valutare i dettagli relativi all'intestazione e alla tabulazione dei report e altri aspetti sull'utilizzo dell'applicativo e sull'interfaccia utente che avrà il portale di BI.

E' stata individuata inoltre la tipologia di architettura del sistema di data warehouse: per quanto riguarda il progetto di business intelligence relativo alle analisi del venduto online e delle tessere Fidelity si ha sia un'architettura a tre livelli che una a due livelli. In particolare, per quanto riguarda i dati relativi alle tessere Fidelity, viene utilizzata un'architettura a due livelli perché si vanno a prelevare i dati direttamente nel database operativo Fidelity, così facendo non si passa attraverso il livello dei dati riconciliati. Per quel che riguarda i dati del venduto online invece viene utilizzata un'architettura a tre livelli in quanto non si vanno a prelevare i dati direttamente dal database operativo del venduto ma si utilizzano i dati messi a disposizione da Me.R.Sy; questi dati sono dati riconciliati perché sono distinti da quelli utilizzati nella gestione delle vendite.

Per il progetto in esame si è deciso di adottare una strategia che consenta di ottimizzare e velocizzare ulteriormente i processi di estrazione ed elaborazione dei dati: al posto di utilizzare software ETL per l'estrazione, la trasformazione e il caricamento dei dati dai database operativi al data warehouse viene utilizzato

Oracle Data Integrator (ODI), un software di tipo ELT (Extraction-Loading-Transformation). Queste due tipologie di software svolgono lo stesso compito, la differenza sta nel modo e nell'ordine in cui lo svolgono. Gli strumenti ETL estraggono i dati, li trasformano per adattarli alle necessità del data warehouse e poi li caricano all'interno del DWH, questo processo richiede una grande disponibilità di risorse in quanto è necessario svolgere questa operazione per tutte le fonti dati che alimentano il data warehouse. La trasformazione dei dati deve avvenire in una piattaforma intermedia che deve essere adatta a gestire ed elaborare quantità di dati potenzialmente molto elevate. Per questo motivo è molto probabile che lo strumento ETL diventi il collo di bottiglia del sistema di gestione delle informazioni. Al contrario gli strumenti di ELT mettono in comunicazione, solitamente tramite protocollo Ftp, i singoli database con il data warehouse: così facendo ogni singolo DB comunica con il DWH evitando la formazione di colli di bottiglia. In questo caso la trasformazione dei dati avviene solamente alla fine nel data warehouse. L'utilizzo di questa tecnica consente inoltre di migliorare la scalabilità del sistema: nel caso sia necessario aggiungere nuove sorgenti dati per aumentare il data warehouse basta inserire all'interno dell' ELT le nuove fonti senza dover, come accade per gli ETL, andare a riprogrammare tutto il sistema. Un altro vantaggio di questa tecnica riguarda i costi iniziali: nel caso di uno strumento ETL è necessario predisporre una struttura hardware in grado di gestire grandi quantità di dati durante la fase di trasformazione; se si utilizzano gli ELT invece questo non è necessario in quanto la trasformazione avviene all'interno del data warehouse.

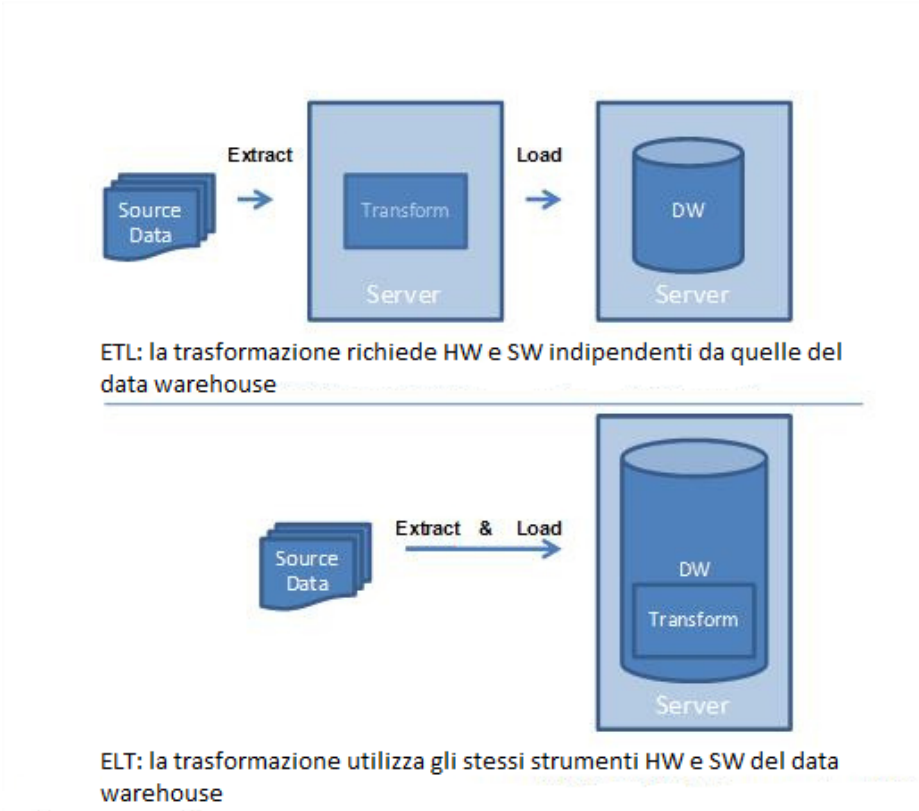

Figura 6.1: ETL vs ELT

Le fasi da seguire per lo svolgimento del progetto sono:

1. individuazione delle fonti da cui estrarre i dati per popolare il data warehouse;
2. realizzazione del modello Entità/Relazione;
3. realizzazione del Dimensional Fact Model;
4. realizzazione dello star schema;
5. individuazione e realizzazione dei Data Mart.

Una volta completate queste fasi si potrà passare alla fase di programmazione degli strumenti ELT e alla fase di progettazione del data warehouse. Queste due fasi, visto l'elevato livello di specializzazione richiesto, saranno svolte dalla ditta di consulenza esterna Miriade.

6.2 Individuazione delle fonti dei dati per alimentare il DWH

Prima dello sviluppo del modello relazionale sono state individuate le fonti dei dati che, attraverso gli strumenti ELT, saranno utilizzate per popolare e aggiornare quotidianamente il data warehouse.

Dall'analisi del sistema informativo e del suo processo di cambiamento è emerso che le fonti ottimali da utilizzare per popolare il DWH sono Me.R.Sy per quanto riguarda i dati del venduto e il database Fidelity per quanto riguarda la gestione delle tessere Fidelity. Si è deciso di utilizzare il database Fidelity per i dati sulle tessere e sui punti in quanto i dati relativi a questi due aspetti non richiedono particolari certificazioni e, inoltre, sono già archiviati in un DB Oracle: questo facilita il lavoro degli strumenti ELT. Per i dati relativi al venduto invece si è deciso di utilizzare le informazioni messe a disposizione dal nuovo software gestionale piuttosto che prelevare i dati direttamente dal database del venduto. E' stata fatta questa scelta in quanto i dati del venduto, prima di poter essere messi a disposizione degli utenti per effettuare analisi, devono essere sottoposti a verifiche e certificazioni in quanto i dati presenti sul database provengono direttamente dai software di gestione delle casse e pertanto vengono archiviati nel DB del venduto senza subire alcun tipo di controllo. La mancanza di certificazioni è un problema perché si potrebbero avere dei dati incongruenti (ad esempio il fatturato totale diverso dalla somma dei singoli scontrini perché non sono stati presi in considerazioni dei buoni spesa). Le certificazioni potrebbero essere svolte anche sul data warehouse, questo porterebbe però un ulteriore carico di lavoro e, inoltre, sarebbe uno spreco di risorse in quanto queste operazioni sono già svolte dal gestionale Me.R.Sy. I dati messi a disposizione hanno un elevato livello di dettaglio perché tengono traccia delle singole righe dello scontrino e sono inoltre presenti anche dei dati aggregati ed elaborati. Ad esempio, oltre alle singole righe dello scontrino, sono già presenti i dati con il totale di ogni singolo scontrino e le informazioni relative ai totali giornalieri di ogni singola cassa. La disponibilità di queste informazioni è molto utile perché consente di diminuire il carico di lavoro del DWH e di sviluppare il modello dati partendo dai dati aggregati e limitando pertanto il numero di tabelle necessarie alla rappresentazione dell'universo di interesse. Nel caso in cui questi dati aggregati non fossero disponibili sarebbe stato necessario calcolarli all'interno del data warehouse; questo passaggio porterebbe inoltre ad un notevole aumento delle dimensioni e della complessità del modello Entità/Relazione.

6.3 Individuazione dei report principali

I report più significativi sono i report principali per il marketing. A partire da questi report si possono modificare i filtri per visualizzare altre informazioni e si possono effettuare operazioni di drill down per navigare all'interno dei report.

Successivamente, per i report più utilizzati, verranno creati dei data mart per velocizzare il processo di reperimento delle informazioni.

I principali report di analisi individuati sono i seguenti:

- venduto Fidelity;
- venduto per decili;
- performance punto vendita;
- ranking fasce orarie;
- venduto per articolo;
- venduto per tessera;
- venduto Fidelity dettaglio reparto;
- dettaglio scontrino;
- clienti codificati;
- venduto campagne promozionali;
- clienti persi;
- migrazione di clienti.

A partire da questi report è possibile soddisfare tutte le esigenze emerse durante gli incontri con i responsabili dell'area marketing ed è possibile effettuare le nuove analisi che sono state proposte.

6.4 Modellazione concettuale

La modellazione concettuale si divide in due fasi: la realizzazione del modello relazionale e quella del dimensional fact model. Successivamente si passa alla modellazione logica anch'essa suddivisa in due fasi: realizzazione dello star schema e dei data mart.

6.4.1 Modello relazionale (ER)

Il primo passo, e forse il più complesso per la realizzazione del modello relazionale, è quello di capire come poter modellare i dati in modo da ottenere un unico schema ER, e successivamente un unico DFM, in grado di contenere tutte le informazioni necessarie per realizzare i report richiesti dal marketing. Per arrivare ad ottenere il modello finale sono stati realizzati diversi altri modelli i quali però non riuscivano a rappresentare nel modo corretto la realtà: alcuni presentavano un livello di dettaglio troppo basso mentre altri non riuscivano a schematizzare il reale funzionamento del sistema. E' emerso quindi che per

ottenere tutte le informazioni necessarie alla realizzazione dei report è necessario mantenere un elevato livello di dettaglio; si è deciso pertanto di utilizzare il dettaglio dello scontrino andando a rappresentare ogni sua singola riga. Così facendo, all'interno del DWH, sarà possibile avere tutte le informazioni relative ai singoli acquisti di ogni cliente.

Nella figura sottostante è riportato lo schema Entità/Relazione relativo al progetto. Per ogni entità sono riportate solamente le chiavi (primarie e secondarie) per facilitare la lettura dello schema.

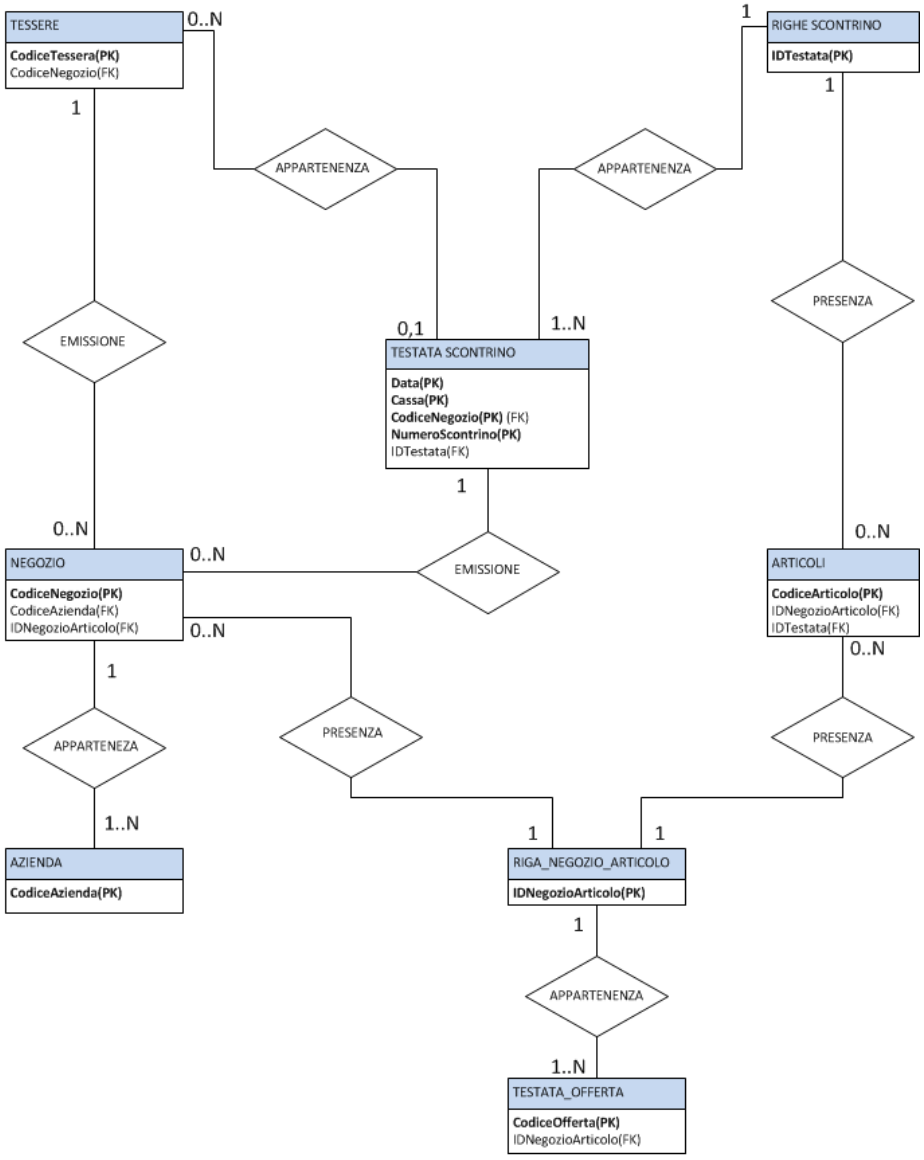

Figura 6.2: Schema Entità/Relazione

Come si può notare nella figura, l'entità principale dello schema è quella della testata dello scontrino. Partendo da questo dato si può risalire ai prodotti acquistati, al negozio che ha emesso lo scontrino e al cliente che ha effettuato l'acquisto.

Un altro punto di interesse dello schema è quello della gestione delle offerte.

All'interno del Gruppo Unicomm sono presenti diversi tipi di offerta:

- 3 x 2: prendi tre prodotti e ne paghi due;
- sconto: sconto sul totale dello scontrino;
- taglio prezzo: prodotti venduti ad un prezzo inferiore rispetto a quello di listino;
- punti jolly: l'acquisto di alcuni prodotti prevede l'aggiunta di punti nella tessera Fidelity del cliente;
- sconti vendite abbinate (basket): sconti previsti nel caso di acquisto di più prodotti contemporaneamente;
- sconti reparto: sconti nel caso di acquisto di prodotti per un importo superiore ad una determinata soglia in un determinato reparto;
- buoni industria: buoni di pagamento per determinati articoli (ad esempio 30 centesimi in meno al prossimo acquisto di yogurt).

In ogni negozio e per ogni prodotto possono essere presenti più tipologie di campagne promozionali e di offerte. Nel modello relazionale tutto questo è stato rappresentato nel seguente modo: ogni singola offerta può essere presente in uno o più negozi appartenenti anche a canali distinti; ogni offerta può contenere uno o più prodotti (nel caso un'offerta non contenga prodotti questa offerta si riferisce a sconti percentuali) e ogni prodotto può essere caratterizzato da una o più offerte. Nel caso delle offerte che non contengono prodotti si è deciso, per riuscire a modellare correttamente la realtà, di introdurre un articolo fittizio; così facendo in ogni offerta è presente la coppia articolo-negozio. Per individuare in modo univoco le varie offerte è stata utilizzata come chiave primaria un numero progressivo (IDNegozioArticolo) in quanto se fosse stata utilizzata la combinazione articolo-negozio, nel caso di più offerte che non prevedono articoli nello stesso negozio, si avrebbero avuto più chiavi primarie con lo stesso identificativo (sarebbe stato violato il vincolo di integrità referenziale).

6.4.2 Dimensional Fact Model (DFM)

In seguito alla realizzazione del modello relazionale si è passati allo studio e allo sviluppo del dimensional fact model. Questo passaggio è necessario in quanto consente di ottenere una visione multidimensionale del problema, necessaria per la progettazione dei data mart. Come già detto nel Paragrafo 5.4.1 gli elementi fondamentali di questo modello sono i fatti, le misure e le dimensioni.

In questo caso il fatto è lo scontrino, le dimensioni di analisi sono invece le seguenti:

- la data;
- il cliente;

- gli articoli;
- il negozio;
- le offerte.

Ognuna di queste dimensioni a sua volta è rappresentata da una gerarchia che rappresenta associazioni molti-a-uno tra coppie di attributi dimensionali.

Le misure non vengono rappresentate nella figura sottostante per facilitarne la comprensione; si è evitato di rappresentare anche lo schema Entità/Relazione relativo al fatto dello scontrino perché si sarebbe ottenuto un'esplosione in termini grafico-notazionali.

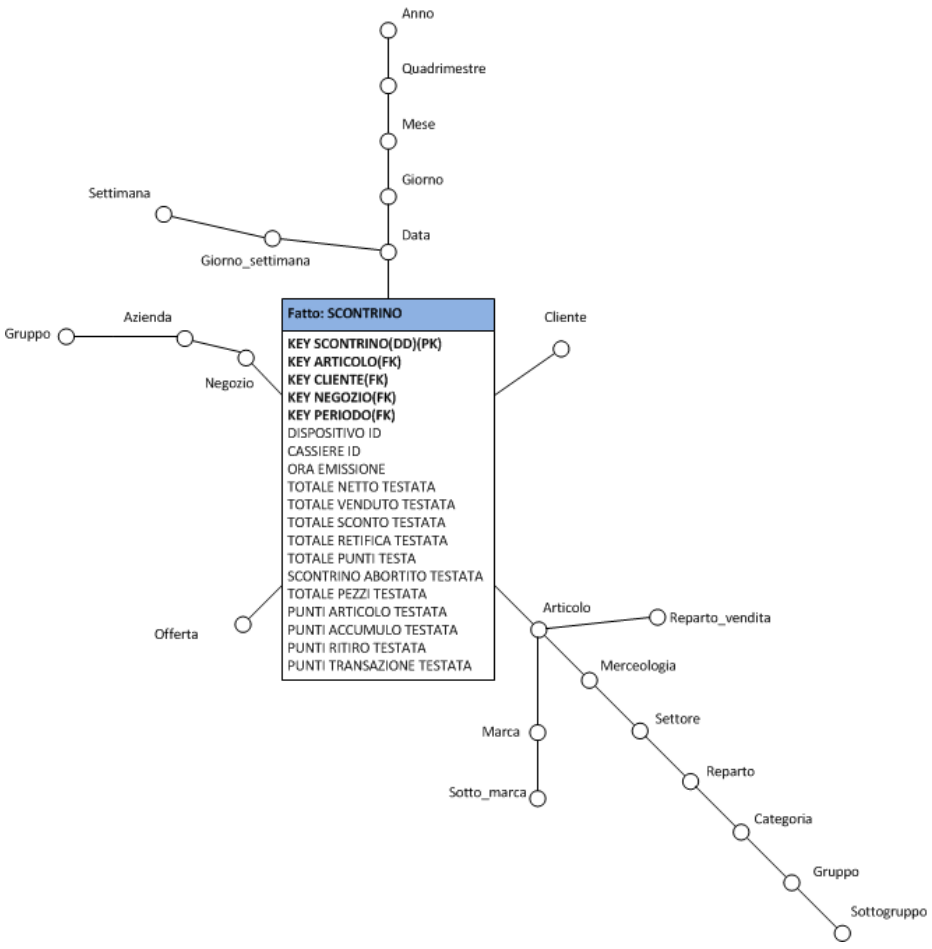

Figura 6.3: Dimensional Fact Model

6.5 Modellazione logica

Durante questa fase verranno fatte delle scelte che, a differenza di quelle fatte nella modellazione concettuale, andranno ad influire sull'implementazione del data warehouse.

In primo luogo è emerso che si dovranno utilizzare due sistemi logici: il sistema ROLAP e il sistema MOLAP. Questo è necessario in quanto è emerso che sono necessari sia report predefiniti, che si ottengono con un sistema di tipo ROLAP, che report dinamici, che permettano di navigare all'interno delle dimensioni, per i quali è necessario l'utilizzo di un sistema di tipo MOLAP.

6.5.1 Il sistema ROLAP

Per l'utilizzo del sistema ROLAP è necessario modellare i concetti multidimensionali (DFM) in elementi bidimensionali. Questo passaggio avviene tramite la realizzazione dello star schema.

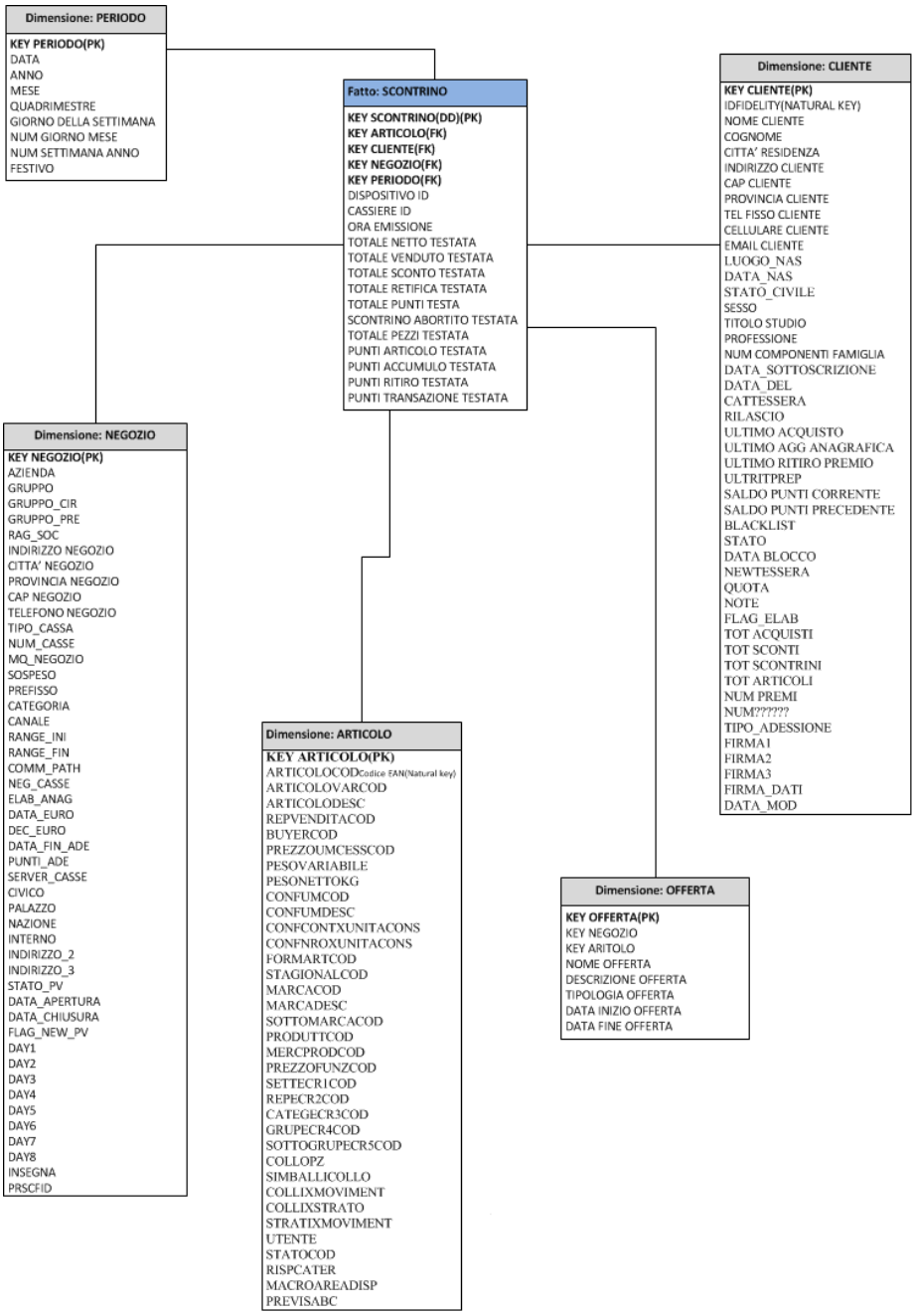

Figura 6.4: Star schema

La visione multidimensionale di questo schema si ottiene effettuando una

join tra le tabelle dei fatti e quelle delle dimensioni. Ad esempio, se sono interessato al venduto dei clienti Fidelity in un negozio nel mese di giugno, dovrò unire le informazioni relative al periodo, al negozio, quelle presenti sulla fact table scontrino e quelle sulla dimensional table dei clienti.

All'interno della fact table è stata inserita una dimensione degenere che, anche se non è strettamente necessaria, consente di identificare in modo univoco lo scontrino, e tutti i dati in esso presenti, utilizzando una sola chiave. L'utilizzo di questo accorgimento semplifica la scrittura delle varie join e velocizza le loro esecuzioni.

Un'altra scelta progettuale molto importante, che consente di ottimizzare il numero di join necessarie per visualizzare i report, è quella di limitare il più possibile l'espansione a fiocco di neve dello star schema. Questa scelta consiste nell'inserire all'interno delle dimensioni principali anche le informazioni relative alle gerarchie del DFM. Nella figura sottostante è riportato lo schema a fiocco di neve relativo alla dimensione articoli.

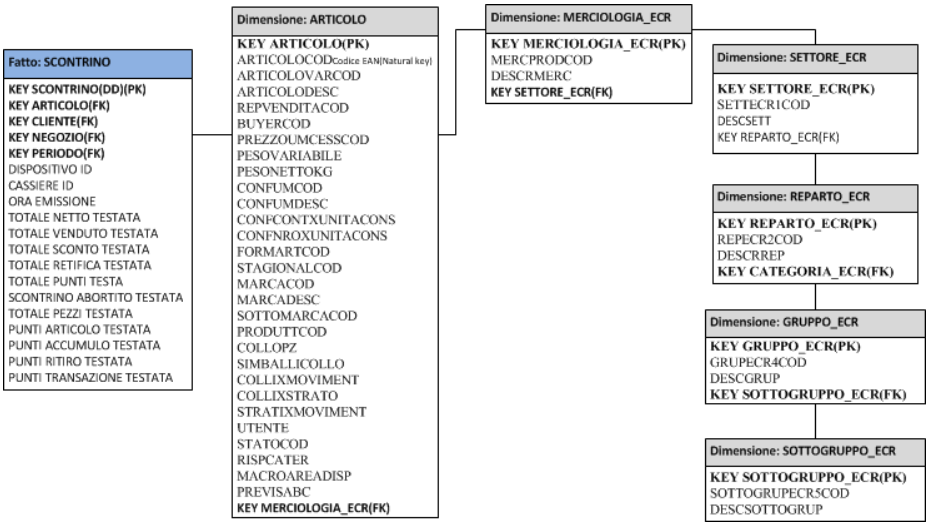

Figura 6.5: Schema a fiocco di neve della dimensione articoli

Come si può notare dalla figura le gerarchie della dimensione articolo presentano al loro interno pochissime informazioni, si è deciso pertanto di inglobare queste informazioni all'interno della dimensione articolo altrimenti, ogni volta che in un report viene richiesto di visualizzare gli articoli, sarebbe necessario effettuare tutte le join con tutte le dimensioni della gerarchia articolo. Questa scelta comporta però un aumento delle informazioni all'interno del DWH dovuto alla ridondanza delle informazioni relative alla classificazione merceologica di ogni articolo.

Questa scelta progettuale è stata applicata, oltre che nella dimensione articolo, anche nella dimensione periodo e nella dimensione negozio.

6.5.2 Il modello MOLAP

Dall'analisi delle esigenze del marketing è emersa la necessità di avere uno strumento che permetta di navigare all'interno dei report e svolgere analisi "dinamiche" andando a variare le dimensioni di analisi. Per poter soddisfare queste esigenze si è reso necessario affiancare il sistema MOLAP a quello ROLAP. Questo sistema prevede la creazione di cubi multidimensionali che consentano di variare le dimensioni di analisi.

Il primo passo di questa fase di modellazione è stato quello di individuare e raggruppare all'interno di data mart i report con dimensioni di analisi simili. Per farlo si è utilizzata una matrice architettura a bus.

REPORT PRINCIPALI	DIMENSIONI DI ANALISI				
	PERIODO	NEGOZIO	CLIENTE	PRODOTTO	OFFERTA
VENDUTO FIDELITY	X	X	X		X
STAMPA PER DECILI	X	X	X		
CLIENTI PERSI	X	X	X		
SPOSTAMENTI DEI CLIENTI	X	X	X		
PERFORMANCE PUNTO VENDITA	X	X			
RANKING FASCE ORARIE	X	X			
VENDUTO PER ARTICOLO	X	X	X	X	X
VENDUTO PER TESSERA	X	X	X	X	X
VENDUTO FIDELITY DETTAGLIO REPARTO	X	X	X	X	X
DETTAGLIO SCONTRINO	X	X	X	X	X
VENDUTO CAMPAGNE PROMOZIONALI	X	X	X	X	X

Tabella 6.1: Matrice architettura a bus

Da questa matrice sono emersi tre data mart principali che richiedono le stesse dimensioni di analisi. Uno di questi data mart prevede tutte le dimensioni di analisi; si potrebbe utilizzarlo per rappresentare anche gli altri ma si è preferito non farlo per alleggerire il carico di lavoro del data warehouse nel momento in cui vengono fatte delle richieste per i data mart con meno dimensioni di analisi.

Come si può notare dalla matrice architettura a bus i datamart principali sono:

1. Venduto generale e gestione clienti: questo datamart prevede come dimensioni di analisi il periodo, il negozio, il cliente e le offerte. I suoi obiettivi sono:

    - rappresentare un resoconto sul venduto dei vari negozi individuando in particolare i totali di vendita relativi alle offerte;

- individuare i clienti che influiscono maggiormente sul fatturato dei vari negozi;
- individuare gli spostamenti dei clienti e i clienti persi.

2. Performance negozio: in questo "cubo" sono presenti solamente due dimensioni di analisi (il periodo e il negozio). L'obiettivo è valutare l'andamento dei punti vendita confrontando i risultati ottenuti in periodi diversi.

3. Dettaglio venduto: questo datamart presenta il livello di dettaglio maggiore e offre la possibilità di effettuare tutti i tipi di analisi richieste dal marketing. Serve in particolare per analizzare i singoli prodotti venduti ai clienti e per vedere nel dettaglio i risultati delle campagne promozionali.

6.6 Report di prova

L'ultima fase del tirocinio ha riguardato la simulazione di alcuni report a partire da un campione di dati aziendali al fine di illustrare al marketing e alla direzione aziendale le potenzialità del nuovo strumento di business intelligence.

Per effettuare tale test sono stati utilizzati dati provenienti dai 5 punti vendita nei quali si sta testando il nuovo sistema gestionale Me.r.Sy.

Il primo passo della fase di test è stato quello, partendo dalle tabelle messe a disposizione dai software di gestione del venduto (JSteore e Me.R.Sy.), di creare l'universo[1] del progetto.

[1]Nome con cui viene chiamata la staging area all'interno del software SAP BusinessObject.

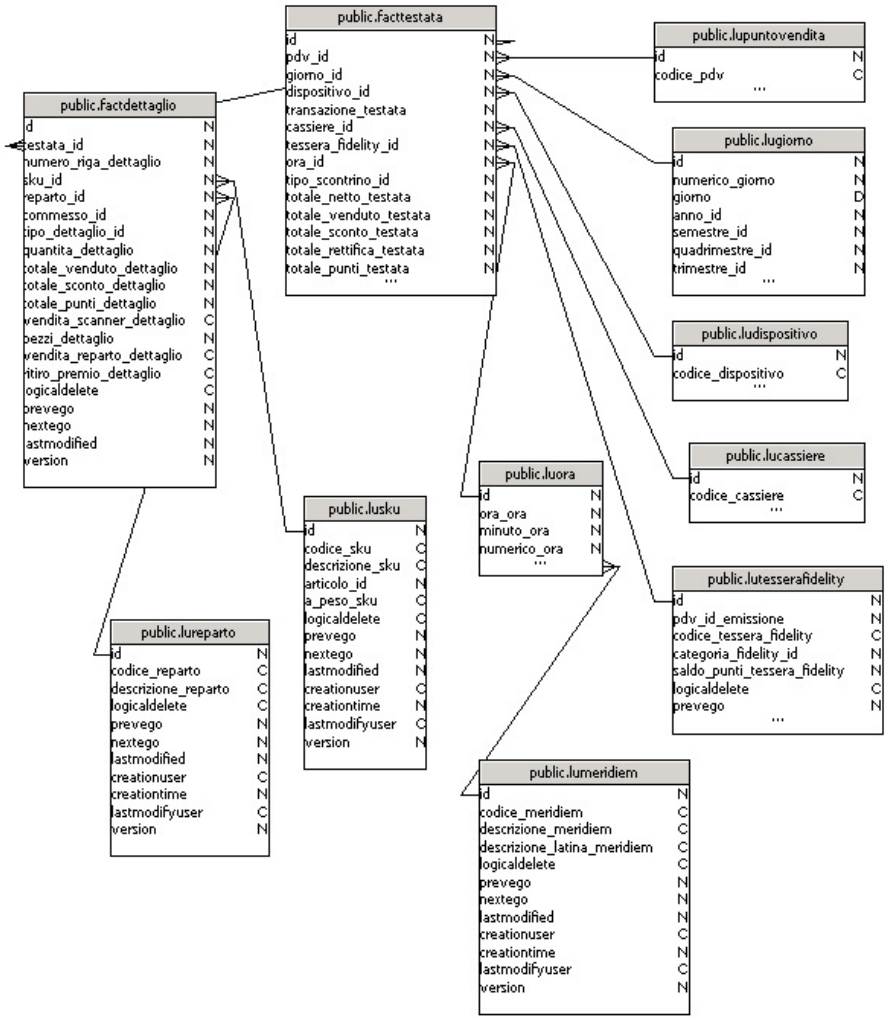

Figura 6.6: Universo

Partendo da questa struttura dati, utilizzando il portale di business intelligence SAP BI platform ed in particolare il modulo SAP BusinessObject, è stato possibile creare alcuni report di prova. SAP BusinessObject, come si può vedere dalla figura sottostante, è un applicativo per la realizzazione e la gestione di report e di analisi di business intelligence molto potente ma allo stesso tempo intuitivo e facile da utilizzare.

Utilizzando questo strumento sono stati realizzati alcuni report che hanno mostrato la bontà del modello dati precedentemente progettato. Questi report sono inoltre stati presentati alla direzione aziendale per illustrare le potenzialità del nuovo portale di business intelligence.

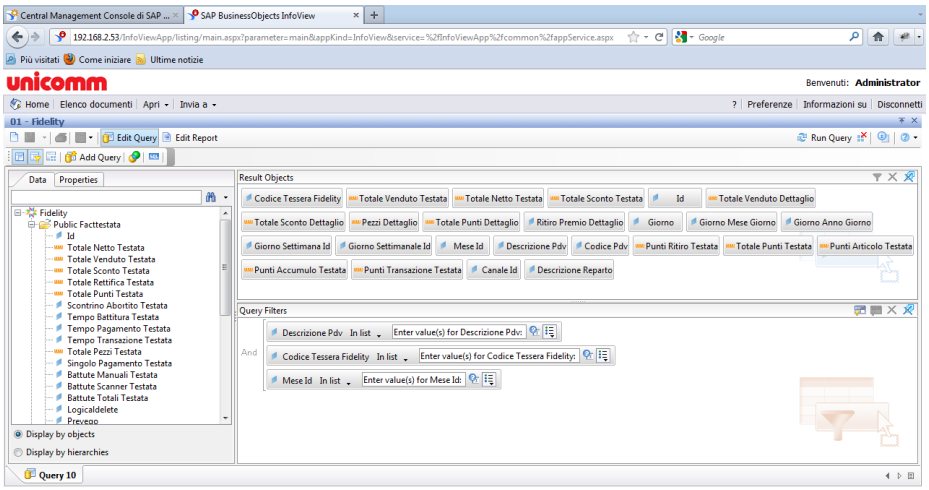

Figura 6.7: SAP BusinessObject

Nelle figure sottostanti sono rappresentati alcuni report realizzati durante il test.

Ora	Clienti Fidelity Fascia Oraria	Vendita Fascia Oraria	Vendita Fidelity Fascia Oraria	Articoli Venduti Fascia Oraria	Scontrini Fascia Oraria	Scontrini Fidelity Fascia Oraria	Orario Rank Nro Scontrini
8	246	12.428,35	8.879,41	6.950	724	443	11
9	866	63.371,37	48.501,73	35.428	2.619	1.728	10
10	1.256	92.101,23	69.288,52	51.314	3.879	2.535	12
11	1.394	107.818,83	77.041,01	59.497	4.292	2.613	19
12	1.151	87.584,38	60.107,31	47.528	3.582	2.034	18
13	606	47.003,50	30.272,19	26.241	1.799	953	17
14	467	36.732,53	25.871,66	19.738	1.244	700	9
15	642	45.782,23	32.171,95	25.048	1.648	942	16
16	849	53.636,31	38.557,00	29.529	2.038	1.193	13
17	1.042	73.509,00	48.294,47	40.265	2.830	1.549	15
18	1.081	82.790,46	51.684,54	45.563	3.148	1.614	14
19	935	79.406,90	47.652,57	43.922	3.219	1.521	8
20	105	9.450,01	4.992,16	5.240	338	142	20
	4.056	**791.615,10**	**543.314,52**	**436.263**	**31.360**	**17.967**	

Figura 6.8: Report Andamento Giornaliero: Ranking Fasce Orarie

Il primo report proposto è simile a quello presente su Catalina e illustra le vendite dei prodotti suddivise per fasce orarie in un determinato periodo. Rispetto al suo predecessore si può notare la chiarezza delle intestazioni delle colonne e l'aggiunta di alcuni campi, come ad esempio la presenza sia dei dati totali che di quelli relativi ai clienti fidelity, che migliorano le potenzialità di analisi.

Un altro report realizzato utilizzando SAP BusinessObject è lo "Stampa venduto Fidelity ente/periodo (Codice CLU1)" su AS400.

Come si può notare in questo report i vari campi hanno un significato molto chiaro, il totale dei clienti Fidelity è corretto e sono presenti i dati relativi ai punti erogati e ai punti premio relativi al negozio durante il periodo selezionato.

Negozio: XXXXXXXXXXXXXX

Mese	Clienti Fidelity	Vendita Totale	Vendita Fidelity	% Inc. Vendita	Scontrini Totali	Scontrini Fidelity	% Inc. Scontrini	Scontrino Medio	Scontrino Medio Fidelity	Punti Erogati	Punti Premio
Giugno	4.056	791.615,10	543.314,52	68,63	31.360	17.967	57,29	25,24	30,24	543.302	526.296
Luglio	3.889	745.222,63	534.856,01	71,77	29.740	17.704	59,53	25,06	30,21	534.843	526.283
	4.951	**1.536.837,73**	**1.078.170,53**	**70,16**	**61.100**	**35.671**	**58,38**	**25,15**	**30,23**	**1.078.145**	**1.052.579**

Negozio: YYYYYYYYYYYYYY

Mese	Clienti Fidelity	Vendita Totale	Vendita Fidelity	% Inc. Vendita	Scontrini Totali	Scontrini Fidelity	% Inc. Scontrini	Scontrino Medio	Scontrino Medio Fidelity	Punti Erogati	Punti Premio
Giugno	3.442	853.037,31	578.711,41	67,84	32.859	17.228	52,43	25,96	33,59	578.698	566.200
Luglio	3.462	863.167,58	592.907,20	68,69	33.668	17.650	52,42	25,64	33,59	592.894	578.856
	4.210	**1.716.204,89**	**1.171.618,61**	**68,27**	**66.527**	**34.878**	**52,43**	**25,8**	**33,59**	**1.171.593**	**1.145.057**

Figura 6.9: Venduto Fidelity Ente - Periodo

L'ultimo report proposto infine riguarda la suddivisione del venduto per reparto.

Reparto	Vendita Totale	Totale % Inc. Totale	Vendita Fidelity	Totale % Inc. Fidelity
Alimentari	244.981,88	14,34%	167.798,09	14,38%
Formaggi	103.275,71	6,05%	73.686,12	6,32%
Non Alimentare	100.170,00	5,86%	67.802,52	5,81%
Frutta	59.543,58	3,49%	40.387,41	3,46%
Salumi	51.741,43	3,03%	35.045,18	3,00%
Verdura	43.950,94	2,57%	29.487,40	2,53%
Restanti	**249.373,77**	**29,23%**	**164.504,69**	**28,43%**
Totale top reparti per Giugno	**603.663,54**	**70,77%**	**414.206,72**	**71,57%**
Totale tutti i reparti per Giugno	**853.037,31**		**578.711,41**	

Alimentari	248.213,58	14,53%	171.514,51	14,70%
Formaggi	104.473,38	6,12%	74.731,25	6,40%
Non Alimentare	92.283,52	5,40%	64.015,86	5,49%
Verdura	47.338,72	2,77%	32.643,44	2,80%
Salumi	44.146,92	2,58%	30.633,36	2,63%
Frutta	41.714,00	2,44%	29.209,09	2,50%
Restanti	**276.809,54**	**32,38%**	**185.333,35**	**31,51%**
Totale top reparti per Luglio	**578.170,12**	**67,62%**	**402.747,51**	**68,49%**
Totale tutti i reparti per Luglio	**854.979,66**		**588.080,86**	

Figura 6.10: Venduto Fidelity per Reparto

Questo report, rispetto a quello presente su AS400, è stato radicalmente modificato; ora si possono vedere i 6 reparti che vendono maggiormente all'interno del negozio selezionato e si possono effettuare delle analisi basandosi sulle differenze tra clienti Fidelity e non Fidelity.

Anche da questi semplici esempi di report emergono gli enormi vantaggi che si possono ottenere dal nuovo sistema di business intelligence, riferiti non solamente sull'aspetto e sulla leggibilità dei report ma soprattutto sulla velocità con cui si ottengono i risultati e sulla facilità con cui è possibile realizzare nuove

interrogazioni. Grazie all'utilizzo dell'interfaccia grafica è possile modificare in modo semplice i vari filtri e ottenere nuove analisi.

Capitolo 7

Conclusioni

Questo lavoro di tesi descrive quanto svolto nei mesi di tirocinio presso Unicomm Srl azienda che opera nella GDO (Grande Distribuzione Organizzata). La prima parte del tirocinio è stata dedicata all'analisi ed alla comprensione del funzionamento del sistema informativo per la gestione del venduto nei vari negozi del gruppo. Successivamente sono iniziati gli incontri con i responsabili dei report per il settore marketing, al fine di capire il funzionamento degli strumenti di reportistica presenti in azienda; in seguito è iniziata l'attività di analisi dettagliata dei singoli report per comprendere le logiche di funzionamento, le modifiche e i cambiamenti da includere nel nuovo strumento di business intelligence che verrà poi implementato. La fase seguente ha riguardato lo studio teorico sulla business intelligence e sui data warehouse. Lo studio sulla BI ha permesso, oltre ad approfondire le conoscenze generali sulla materia, di individuare nuovi report, utilizzando dati già presenti in azienda da proporre al marketing. Lo studio dei data warehouse è stato necessario invece per poter passare alla fase successiva di progettazione del modello dati. Durante la fase di progettazione sono stati sviluppati il modello concettuale e il modello logico del data warehouse. Infine sono stati realizzati alcuni report per testare e validare il modello dati progettato.

Questa esperienza all'interno del Gruppo Unicomm è stata molto interessante e stimolante sia dal punto di vista dell'apprendimento che dal punto di vista umano.

Ho avuto la possibilità di lavorare in un team di progetto all'interno di una delle realtà più grandi a livello nazionale nella grande distribuzione di generi alimentari; questo mi ha permesso di capire le logiche operative presenti all'interno delle grandi aziende, di osservare la collaborazione tra i vari settori aziendali e di sperimentare le modalità di lavoro all'interno di un gruppo di progetto.

Fin dai primi giorni del tirocinio ho avuto la possibilità di partecipare attivamente allo sviluppo del progetto di business intelligence per l'area marketing, a partire dall'analisi delle esigenze degli utenti finali. Ho potuto inoltre apprendere le modalità con cui svolgere un'analisi dei requisiti e di interagire con i

destinatari del progetto sia a livello tecnico (modifiche di report esistenti e analisi di nuove esigenze) che psicologico (aspettative in termini operativi e temporali).

Sia nella mappatura della soluzione esistente, che nella definizione dello scenario futuro, è stato fondamentale confrontarsi con gli utenti sulla base di informazioni reali e delle funzionalità applicative disponibili.

Molto importante per la mia formazione è stata anche la progettazione del modello dati del data warehouse. Dopo aver effettuato lo studio teorico ho messo in pratica quanto appreso: ho sviluppato il modello concettuale e poi il modello logico del data warehouse, in affiancamento con Dottor Rizzato e un consulente di Miriade.

In conclusione questa esperienza presso il Gruppo Unicomm è stata molto importante sia dal punto di vista tecnologico che metodologico e relazionale. Devo ringraziare inoltre il tutor aziendale Dottor Rizzato che mi ha seguito in maniera impeccabile e precisa durante tutto il tirocinio.

Elenco delle figure

Bibliografia

- Pier Franco Camussone. *Informatica organizzazione e strategie.* McGraw-Hill 2000.
- Paolo Atzeni, Stefano Ceri, Stefano Paraboschi e Riccardo Torlone. *Basi di dati. Modelli e linguaggi di interrogazione. Seconda edizione.* McGraw-Hill 2006
- Matteo Golfarelli e Stefano Rizzi. *Data Warehouse - teoria e pratica della progettazione.* McGraw-Hill, second edition, 2006.
- Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit Second Edition The Complete Guide to Dimensional Modeling.* Wiley Computer Publishing 2002
- Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, and Riccardo Torlone. *Basi di dati - modelli e linguaggi di interrogazione.* McGraw-Hill, second edition, 2006.
- Giulio Destri. *Introduzione ai sistemi informativi aziendali.* Monte Universita Parma, 2007.
- Matteo Golfarelli and Stefano Rizzi. *Data Warehouse - teoria e pratica della progettazione.* McGraw-Hill, second edition, 2006.
- Carlo Vercellis. *Business Intelligence - modelli matematici e sistemi per le decisioni.* McGraw-Hill, 2006.
- Luca Cabibbo. *Introduzione al Data Warehousing ed alla Progettazione di Data Warehouse Dimensionali.* http://www.dia.uniroma3.it/~cabibbo/dw/
- Danilo Montesi. *Data Warehouse Architettura e Progettazione.* http://www.cs.unibo.it/~montesi/CBD/07DWH-Arch&Proj.pdf