PROGETTAZIONE SISTEMA OLAP

Amedeo Racanati
N° matricola 663528

Abstract

Scopo di questo documento è quello di esporre la metodologia applicata per la progettazione e implementazione di un sistema OLAP, al fine di effettuare un'analisi dei dati di un sistema OLTP, utilizzato da un'azienda che effettua vendite online tramite un e-commerce.
L'obiettivo di questo caso di studio è stato quello di analizzare i dati del sistema OLTP al fine di produrre conoscenza utile ai managers dell'azienda. Benché la qualità dei dati sia stata non del tutto ottimale, si è potuto creare un data warehouse (DW) che ne permettesse un'analisi piuttosto accurata, al fine di comprendere ad esempio lo stato dei ricavi prodotti dall'azienda nei diversi anni di vita, o anche la ripartizione territoriale delle vendite stesse. Il tutto è stato gestito tramite un programma ETL creato ad hoc per questo caso di studio, un DBMS contenente il DW e un programma di analisi dei dati chiamato Jaspersoft*.

Introduzione

Nel seguente caso di studio ci si è posto come obiettivo quello di analizzare, comprendere ed esporre i dati storici prodotti da un sistema OLTP utilizzato da un'azienda che effettua vendite online tramite un e-commerce.
L'azienda in questione possiede un archivio di dati operante dal 2010. Le diverse informazioni associate alle vendite sono state prese in esame per la realizzazione di un sistema OLAP, al fine di produrre conoscenza utile per i managers dell'azienda.
Dopo aver compreso il contesto di business dell'azienda, si è passati all' analisi dei dati da loro forniti.

* Sito web di Jaspersoft: https://www.jaspersoft.com/it

Essi sono presenti su un database relazionale, ma al fine di poter avere una procedura standard per il processo di data warehousing, si è stabilito che i dati debbano venir esportati in formato Excel, i quali poi verranno letti, trasformati e caricati su un DW grazie ad un programma ETL creato ad hoc per questo caso di studio.
Dai dati esportati sul file Excel, si è evinto che la qualità degli stessi è piuttosto mediocre, in quanto alcune informazioni sono spesso mancanti, o scorrette.
La filosofia seguita nel processo di data warehousing è stata quella di includere il maggior numero possibile di record, al fine di poter fornire analisi dei dati il più possibile complete, fermo restando l'importanza della qualità dei dati stessi.
Ad esempio, se un record dovesse possedere delle informazioni mancanti, si preferisce associare a queste ultime un valore fittizio nel DW ("Non specificato"); in questo modo il record in questione può essere utilizzato per quelle analisi dei dati per le quali il record è significativo, mentre per le altre verrà escluso.
Come suggerisce lo stesso Hess:
"When all other means for obtaining a meaningful value have failed, the warehouse loading process should use it's own dummy value. The most simplistic version of this involves converting all missing attribute values to a special value like the word UNKNOWN or UNAVAILABLE." (Hess, 1998).

Processo di data warehousing

Definizione del protocollo

Al fine di definire con precisione la struttura del DW, si è dovuto prima di tutto stabilire un protocollo per l'esportazione dei dati sul file Excel.
Il file Excel deve contenere come prima riga l'intestazione delle diverse colonne, e le successive righe devono contenere i dati delle registrazioni di vendita effettuate dall'e-commerce.
Ogni riga, infatti, registra una determinata vendita di un prodotto, il quale possiede diverse caratteristiche, e la vendita è associata ad un determinato

ordine effettuato da un determinato cliente.
Di seguito vengono elencate le colonne dell'intestazione stabilite dal protocollo:

- IdOrdine, numerico: identifica un determinato ordine associato alla vendita;
- DataOrdine, data (GG/MM/AAAA): data dell'ordine;
- CodStatoFattura, stringa: codice dello stato di fatturazione dell'ordine;
- CodProvincia, stringa: codice della provincia di fatturazione dell'ordine;
- Sesso, numerico: indica il sesso del cliente che ha effettuato l'ordine (3 = maschio, 4 = femmina);
- Quantità, numerico: numero di prodotti venduti (valori negativi = vendite, valori positivi = resi di vendita);
- Prezzo, numerico: prezzo di acquisto di un singolo prodotto (pertanto il ricavo totale si ottiene moltiplicando la quantità per il prezzo);
- NomeDes, stringa: nome del design del prodotto venduto;
- LinguaCollezione, stringa: collezione di appartenenza del prodotto venduto;
- LinguaColore, stringa: colore del prodotto venduto;
- NomeSes, stringa: genere di appartenenza del prodotto venduto (es. Kids Boys, Kids Girls, Uomo, Donna);
- PagamentoOrdine, stringa: metodologia di pagamento dell'ordine;
- ValoreTagliaEffettivo, stringa: numero di taglia del prodotto venduto;
- NomeCat, stringa: nome della categoria di riferimento del prodotto venduto;
- NomeMac, stringa: nome della macro-categoria di riferimento del prodotto venduto.

I campi obbligatori per ciascuna riga, al fine di poter essere inserita nel DW, sono: IDOrdine, DataOrdine, Quantità, Prezzo, CodStatoFattura.
Senza questi campi non sarebbe possibile né stabilire informazioni importanti, quali la temporalità e il luogo della vendita, né determinare le misure oggetto dell'analisi dei dati, ossia Quantità e Prezzo.

Altri vincoli imposti dal protocollo sono:

- Durante la fase di importazione si controlla che l'ordine delle colonne sia lo stesso di quello elencato dal protocollo, al fine di assicurare una corretta importazione dei dati.
- Si controlla per ciascun ordine (il quale è determinato tramite il campo IDOrdine) che le righe ad esso associate possiedano sempre la stessa DataOrdine, lo stesso CodStatoFattura e la stessa CodProvincia. In caso negativo, non sarebbe possibile stabilire quale informazione sia corretta e quale mendace, e pertanto vengono escluse le righe dell'ordine in questione;

Per gli altri campi, invece, qualora il dato non sia stato specificato o sia scorretto, si procederà nell'assegnare di default una dimensione fittizia, come specificato nell'introduzione.

Progettazione del data warehouse

Dopo la definizione del protocollo, si è proceduto con la definizione dei fatti, delle misure e delle dimensioni.
Il **fatto**, come si può già dedurre, riguarda la vendita di un determinato numero di prodotti di egual genere associato ad un ordine ben specifico.
Le **misure** specificate per il fatto sono Quantità venduta e Prezzo totale di vendita. La quantità venduta indica il numero di prodotti venduti; mentre il prezzo totale di vendita altro non è che il ricavo rilevato dalla vendita, calcolato come prodotto tra il Prezzo unitario e la quantità.
Si è deciso di adottare una grana fine, in modo tale da poter effettuare analisi dei dati sotto diverse prospettive di interesse.
Le prospettive definite, ossia le **dimensioni**, sono le seguenti:

- Data, che possiede a sua volta varie informazioni associate, tra le quali: nome del giorno, numero giorno nel mese, numero giorno nell'anno, tipologia di giorno (feriale/weekend), nome del mese, numero del mese nell'anno, numero di settimana nell'anno, trimestre;
- Stato, formato dal codice ISO e dal nome;

- Provincia, formata dal codice, dal nome della regione e dal nome della nazione (si è evitato di inserire chiavi referenziali verso lo stato o le regioni, in modo tale da poter migliorare le prestazioni durante la fase di analisi dei dati);
- Sesso del cliente, formato dalla denominazione dello stesso (Maschio/Femmina);
- Nome del design, formato dal nome dello stesso;
- Lingua collezione, formata dal nome della stessa. Per questa dimensione è presente anche l'ordinamento delle collezioni, in quanto si riferiscono ad un intervallo temporale ben determinato, e quindi ordinabile (utile per le analisi dei dati ordinati temporalmente);
- Lingua colore, formata dal nome della stessa;
- Sesso dell'abbigliamento, formato dal nome dello stesso;
- Valore della taglia, formato dal nome della stessa;
- Tipologia di pagamento, formata dal nome della stessa;
- Nome macro-categoria, formato dal nome della stessa;
- Nome categoria, formato dal nome della stessa e dal nome della macro-categoria di riferimento (anche qui per migliorare le prestazioni durante l'analisi dei dati).

Ogni tabella delle dimensioni possiede inoltre un ID auto incrementante utilizzato come chiave primaria.
La tabella del fatto contiene i riferimenti a tutte e 12 le dimensioni sopra citate, e in aggiunta contiene anche l'ID dell'ordine. Tutte queste informazioni - tranne lo stato, la provincia e la data - costituiscono la chiave primaria del fatto.
Oltre ai campi su citati, il fatto possiede altre due colonne, ossia le due misure QuantitàVenduta e PrezzoTotale.
Per le tabelle delle dimensioni è stata impostata una chiave "Unique" con indice per tutte quelle colonne che vengono utilizzate per ricercare, e quindi determinare, la dimensione di ciascun fatto.
Ad esempio per la tabella della dimensione Categoria, è stata impostata una chiave "Unique" sulla colonna Nome, in modo tale che in fase di determinazione della dimensione per ciascuna riga, la ricerca sia più veloce,

con conseguente riduzione dei tempi di importazione dei dati. In merito alla riduzione dei tempi. è stata creata una stored procedure che velocizza la ricerca della dimensione data, in quanto questo dato è sempre presente per ciascuna riga, e la quantità di righe per questa dimensione è elevata.

Infine è stata creata anche una stored procedure per l'inserimento dei fatti, in quanto si è notato che così facendo le tempistiche di esecuzione della procedura di importazione si sono notevolmente ridotte.

Il DW, avente le tabelle su citate, è stato creato su un DBMS dal nome SqlServer.

Ecco lo schema del database così definito.

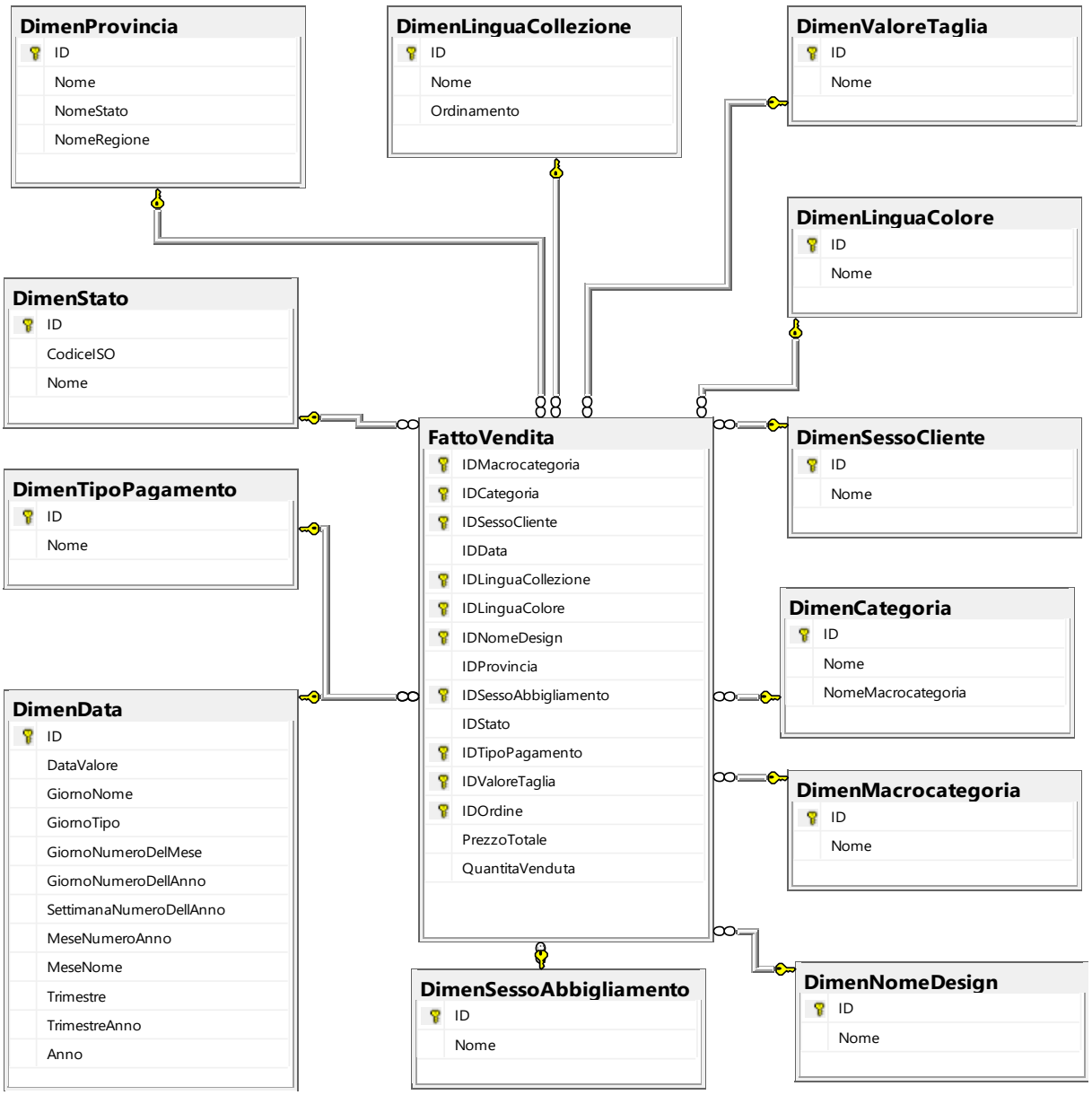

Fig. 1: Diagramma relazionale del DW progettato

PROGRAMMA ETL

Dopo la progettazione del DW, si è passati alla progettazione e implementazione del programma ETL.
Il programma è stato scritto in C#, ed è una semplice applicazione Windows Form, avente un'interfaccia grafica essenziale.
Il programma ETL è stato pensato per essere eseguito ogni mese da un operatore umano, il quale dovrà specificare al programma il file Excel, dal quale verranno importati i dati secondo il protocollo definito in precedenza.
La procedura di importazione mensile richiede non più di 10 minuti per essere portata a termine.
Il programma ETL preleva i dati dal file Excel, li processa e li salva nel DW su citato.
Oltre ad eseguire i controlli specificati dal protocollo, il programma effettua altre operazioni essenziali per l'importazione dei dati:

- Controllo dei resi: si fa in modo che i resi contribuiscano al calcolo del totale della quantità e del prezzo delle rispettive vendite a cui si riferiscono. Mentre le vendite vengono sommate ai totali, i resi vengono sottratti;
- Eliminazione delle misure aventi come valore 0: esse rappresentano delle informazioni prive di utilità per i managers, in quanto l'aggregazione dei dati è sempre sommativa;
- Ordinamento delle dimensioni delle collezioni in base alla loro temporalità;
- Controllo dei dati già importati: se le righe sono state già importate in precedenza nel DW, e se si riesegue nuovamente una importazione degli stessi dati, il programma bloccherà l'importazione, in quanto sarebbe errato inserire nuovamente le stesse informazioni.

Il programma ETL fornisce inoltre un link per l'accesso al server utilizzato per l'analisi dei dati, il quale verrà presentato nella prossima sezione.

ANALISI OLAP

Una volta presentato il programma ETL, non manca che mostrare il programma utilizzato per l'analisi dei dati.
A tal proposito è stato scelto il programma Jaspersoft, il quale possiede una versione gratuita ed è provvisto di una community online affiatata, grazie alla quale è stato possibile risolvere qualche problematica riscontrata durante l'utilizzo del software. Inoltre la presenza di tutorial semplici ed efficaci ne ha permesso un facile apprendimento.
Dopo aver installato il software in locale, si è configurato il sistema affinché si possa collegare al DW presente su SqlServer.
In seguito, si è deciso di estrapolare dal DW alcune delle informazioni presenti, al fine di effettuare analisi dei dati mirate.
Si è deciso di analizzare le vendite sotto tre prospettive:

- Ricavi annuali di vendita, ossia i ricavi raggruppati per anno, al fine di comprendere l'andamento delle vendite nel tempo e lo stato dell'azienda;
- Ricavi effettuati in Italia raggruppati per regioni e province, utile per comprendere in quali luoghi vengono generati più introiti;
- Ricavi e quantità vendute raggruppate per macro-categorie, al fine di comprendere quali tipi di prodotti costituiscano la fonte primaria delle vendite.

Una volta deciso ciò, si è proceduto nel creare su Jaspersoft i relativi domini e le viste ad hoc. Essi sono facilmente configurabili e permettono un'analisi dei dati personalizzabile in ogni minimo dettaglio.
In seguito, si sono potuti creare i report relativi alle tre prospettive sopra indicate, che vengono mostrati qui di seguito.

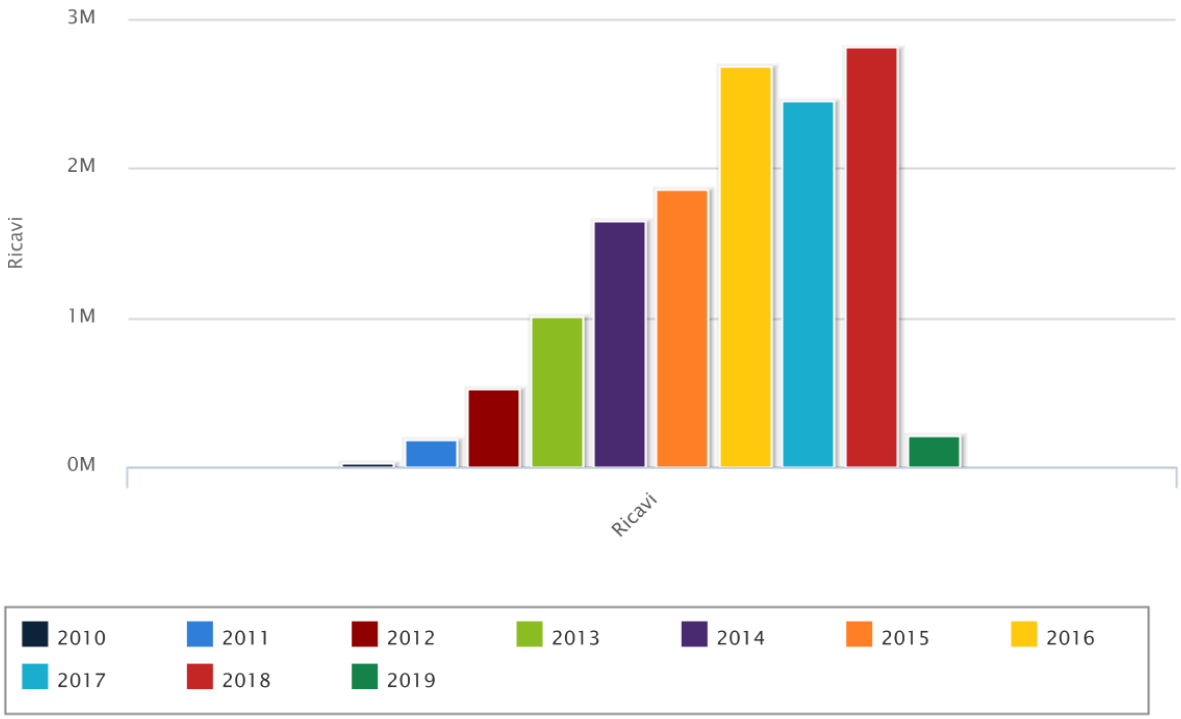

Fig. 2: Analisi dei **ricavi annuali**. Si può notare come l'azienda nel tempo abbia sempre più ingrandito il proprio business. Si può notare un costante incremento delle vendite, ad esclusione dell'anno 2017. L'anno 2019, essendo lo stesso in cui è stato redatto questo caso di studio, non possiede ancora tutti i dati di sua competenza.

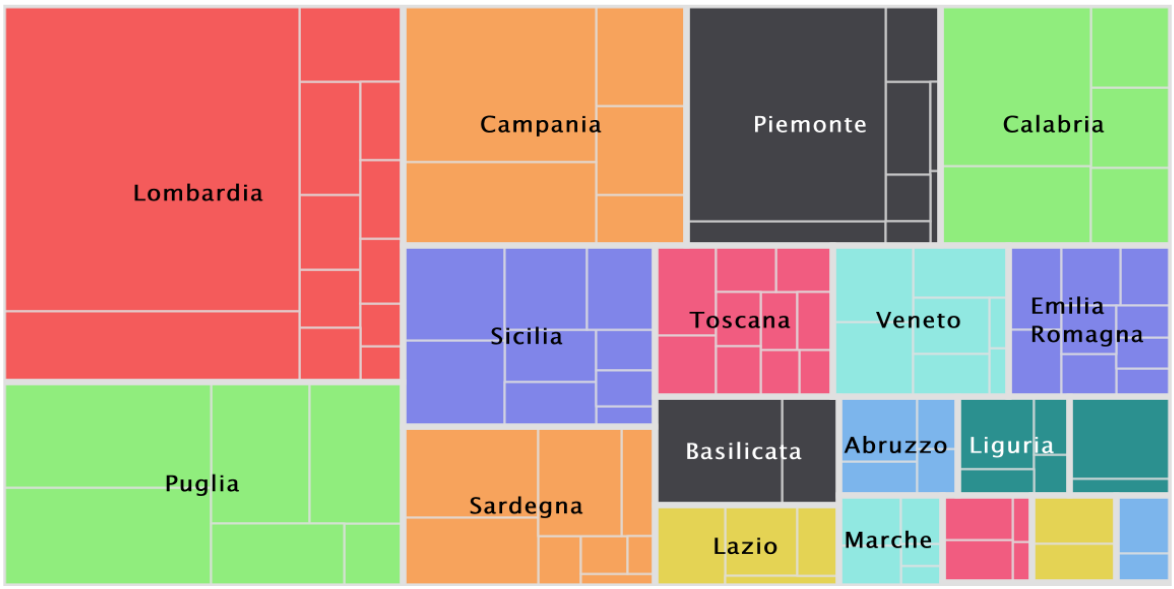

Fig. 3: Analisi delle **vendite territoriali**. Si può notare come il business sia diffuso in tutta l'Italia, ma con una prevalenza di vendite in Lombardia e Puglia. C'è una leggera tendenza di concentrazione delle vendite nel sud Italia e nelle isole.

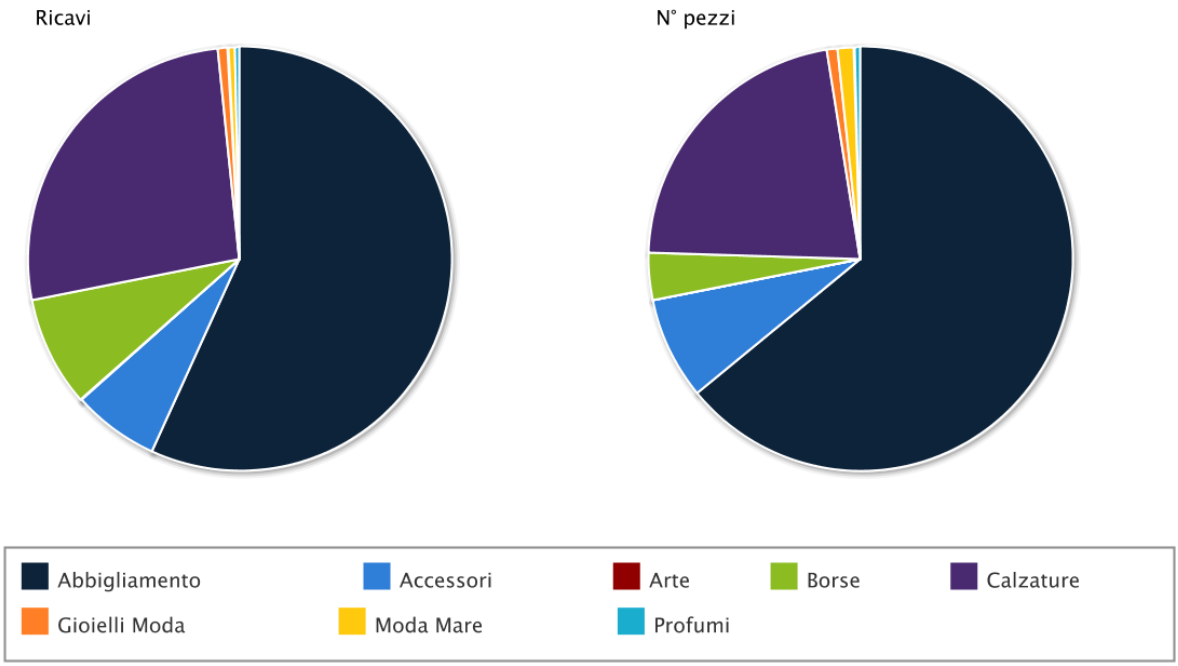

Fig. 4: Analisi delle vendite **raggruppate per macro-categorie**. C'è una netta predominanza del settore abbigliamento, seguito dal settore Calzature. Gli altri settori invece sono quasi del tutto marginali.

ESPERIENZA

Questo caso di studio è stato formativo sotto diverse prospettive.
Innanzitutto ha potuto mostrare un esempio pratico di realizzazione di un data-warehouse, partendo da dati non del tutto ottimali dal punto di vista qualitativo.
Questo ha permesso pertanto di porre l'attenzione su particolari problemi di diversa natura, per i quali si sono trovate le giuste soluzioni.
Infine, tramite l'utilizzo del software Jaspersoft, si è scoperto uno strumento ricco di potenzialità, dato che l'analisi dei dati è un settore in continua espansione e quindi ricco di molte opportunità negli anni a venire.

CONCLUSIONI

Tramite questo caso di studio, si è potuto analizzare con efficacia la situazione dell'azienda in questione, che risulta essere in espansione e in buona salute. Nulla vieta in futuro di poter effettuare ulteriori analisi, mirate a conoscere l'azienda sotto differenti prospettive, dato che la presenza di diverse dimensioni permette un uso dei dati molto variegato.
Infine, tramite la progettazione di questo data-warehouse, si potrebbe spingere l'azienda ad investire maggiormente nella qualità del software OLTP utilizzato per l'e-commerce, al fine di garantire una maggiore qualità dei dati e, pertanto, una conoscenza migliore della situazione aziendale.

Bibliografia

Hess, J. (1998). *Dealing With Missing Values In The Data Warehouse.*