***Creation Of A Data Mart For Evaluate The Closing Reasons And The Best Geo-Locations For A Fashion Retail Store***

Con il termine **Business Intelligence (BI**) ci si riferisce ad una serie di processi aziendali che ruotano attorno ai dati, con operazioni di raccolta, elaborazione, analisi, cui scopo è quello di produrre informazioni al servizio del management strategico e tattico, che trova supporto analitico, storico e previsionale della Data Driven Strategy. La BI è stata collocata altresì nel sottoinsieme operativo, poiché sta assumendo un ruolo sempre più importante anche nelle normali attività giornaliere delle aziende.

I **Data Warehouse (DWH)** sono il principale strumento a supporto della Business Intelligence. Essi permettono di collezionare dati integrati, consistenti e certificati, afferenti a tutti i processi di business dell’azienda e provenienti dalle fonti operazionali. Questi dati vengono in seguito opportunatamente trasformati attraverso procedure ETL e controllati attraverso il sistema di data quality.

un **Data Mart** è un database analitico progettato per incontrarsi con le esigenze specifiche di un’impresa. Essendo sottoinsieme logico o fisico di un data warehouse di dimensioni maggiori, segue le stesse regole di progettazione con dati aggregati a vari livelli di dettaglio, anche se, talvolta può essere costituito anche in assenza di un sistema di dati integrato. La loro suddivisione è basata sul tipo di business che ogni cliente vuole affrontare.

L’implementazione di una Data Mart viene spesso divisa in vari livelli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SCHEMA** | **L0** | **L1** | **L2** | **L2\_STAR**  **SCHEMA** |
| **Definizione** | Estraggo i dati da vari tipi di file senza trasformazioni. | Principali trasforma-zioni e le operazioni di data quality. | Molto veloce e tabelle snelle. uso di surrogate key per collegare i vari attributi. | Poche tabelle ma corpose, per avere tutti i dati neccessari per i report. |
| **Area** | Staging/ Extraction Area. | Trasformation Area. | ETL  Area. | Visualization Area. |
| **Primary key** | NO. | YES. | YES. | YES. |
| **Surrogate key** | NO. | YES. | YES. | NO. |
| **Azione sulla Tabella** | Truncate table, but maintain schema | Nothing. | Nothing. | Nothing. |
| **Azione sui dati** | Insert. | Update/  Insert. | Update/  Insert. | Update/  Insert. |

Il modello multidimensionale o **DFM** (Dimensional Fact Model) è un modello concettuale dove è possibile rappresentare i dati all’interno di un ipecubo, i cui spigoli rappresentano le dimensioni di analisi, che successivamente verrà suddiviso in tanti “cubetti”, ciascuno dei quali è identificato da una terna di coordinate. Ogni cubetto contiene idealmente i valori assunti dalle misure per quella data terna e viene comunemente denominato “fatto” in quanto rappresenta l’accadimento di un evento di interesse per il dominio di business.

Un modello multidimensionale, quindi, si basa principalmente su 4 concetti chiave:

* *Fatto:* concetto rilevante per il processo di Decision-Making. Tipicamente modella una specifica area di business (Vendite, Ordini, Produzione, etc.), ed è caratterizzato da una a più misure;
* *Misura:* rappresenta l’aspetto quantitativo del fatto che risulta di elevata importanza per l’analisi. Proprio dalle *Misure* vengono estratti dei *KPI (Key Performance Indicator)* che guideranno le imprese nelle proprie strategie di business. Alcuni esempi possono essere la Quantità prodotta, il Profitto, e il Prezzo;
* *Dimensione:* rappresenta le coordinate di analisi del *Fatto*. Tra queste possiamo trovare Data, Prodotto, Negozio;
* *Attributo Dimensionale:* è un raggruppamento logico di alcuni elementi di una stessa dimensione. Classi di elementi che consentono all'utente di selezionare i dati per specifiche caratteristiche.

Una volta costruito il Data Fact Model, viene implementato lo schema logico. Esso viene rappresentato secondo uno **Star Schema**, il cui centro è costituito da una tabella dei fatti; le punte della stella rappresentano invece le tabelle delle dimensioni che si diramano dal centro. Le caratteristiche principali di uno Star Schema sono le seguenti:

* Struttura semplice di facile comprensione;
* Query molto performanti, perché riducono i JOIN da effettuare tra tabelle;
* Tempo di caricamento dei dati relativamente lungo, perché la ridondanza dei dati dovuta alla de-normalizzazione, provoca l’aumento delle dimensioni della tabella;
* Ampiamente supportato da un gran numero di strumenti di business intelligence;
* Le tabelle dei fatti in uno Star Schema sono in terza forma normale, mentre le tabelle dimensionali sono de-normalizzate.

Una base dati è in 3NF (*terza forma normale*) se tutti gli attributi non-chiave dipendono dalla chiave soltanto, ossia non esistono attributi non-chiave che dipendono da altri attributi non-chiave. Tale normalizzazione elimina la dipendenza transitiva degli attributi dalla chiave e prende il nome di schema **SnowFlake**.

* Normalized tables;
* Slower query run than Star Schema;
* Creation of Integer surrogate keys (SK).
* Minimal code changes.

Per svolgere una analisi completa e ottenere un risultato attendibile partendo dai dati ISTAT, evidenziati e trasformati tramite il processo di data quality, sono state necessarie principalmente tre query, tutte estremamente collegate tra loro, con l’obiettivo finale di trovare quali, tra le migliaia di città sono le più appetibili economicamente per aprire un nuovo negozio (**best geo-location)**.

Inizialmente, le query saranno relative agli indicatori regionali presi dal sito dell’ISTAT (spesa media mensile familiare per beni e servizi non alimentari, reddito medio, rete ferroviaria in esercizio, tasso di disoccupazione, PIL Pro Capite). Sarà creato un fattore di somma dei vari rank raggruppati per regione e ordinati per rank decrescente. In questo modo, si otterrà come risultato un ordine di potenzialità di sviluppo delle regioni Italiane. Ad esso, si sommeranno i dati aziendali relativi allo storico del numero di negozi chiusi e ancora aperti per ogni regione e, il valore finale su cui sarà basata la nostra classifica, prenderà il nome di Rank. Le tre regioni migliori su cui investire un capitale economico sono Trentino Alto Adige, Valle D’Aosta e Friuli Venezia Giulia.

L’ultimo step da svolgere per trovare la posizione ideale per aprire un nuovo negozio, è quella di selezionare i comuni presenti nelle tre regioni precedentemente scelte come le più appetibili e andare in profondità analizzando i dati relativi al numero di turisti riferiti all’anno scorso e il numero di residenti per ogni città. Il risultato ottenuto evidenzia come Andalo, Corvara In Badia e Gressoney-La-Trinitè sono le migliori città dove poter aprire un negozio nel 2019.

**L’algoritmo di classificazione CART** è una procedura non parametrica che costruisce un albero decisionale con l’obiettivo di etichettare un attributo; Infatti, con il termine classificazione si intende il processo che data una collezione di record, denominata *Training Set,* cerca di costruire un modello in grado di attribuire una caratteristica, denominata *attributo Classe,* basandosi sulla combinazione delle altre proprietà che caratterizzano il singolo individuo della popolazione. Una volta ottenuto il modello, esso può essere usato per predire la classe di nuove istanze di record per cui la classe è sconosciuta.

Gli step importanti da seguire quando si costruisce un albero decisionale con la procedura CART sono principalmente due: adottare un criterio di bontà della tecnica con i cui i nodi vengono suddivisi da parent nodes a child nodes (split criterion) e stabilire una regola di arresto di crescita dell’albero (stopping rule).

Nel progetto implementato, utilizzeremo il processo CART per definire e predire quali negozi continueranno ad esercitare e quali negozi chiuderanno nel 2019, avendo come training set i negozi che hanno chiuso nel 2018, aventi i dati dal 2017 al primo semestre 2019.

Questi dati saranno ulteriormente divisi in training and test set grazie ad una partizione randomica 80/20 sul 100% degli elementi analizzati, dove verrà creato il modello partendo dai Training set di dati per testarlo sui dati di test successivamente.

Una volta eseguita l’operazione di test, sarà verificata la sua diseguaglianza distribuzione tramite il Gini index, dove 0 rappresenta la perfetta uguaglianza, mentre un indice di 100 implica una perfetta disuguaglianza e Inoltre, sarà studiata l’accuratezza tramite la media.

Ottimizzato al massimo il modello, sono stati introdotti i dati del database da analizzare e predire nel modello creato, ottenendo la previsione. Il modello è stata ulteriormente valutata tramite l’indice di Gini [0.1705464] e l’accuratezza [0.8701299], e infine, confrontato con i risultati iniziali, per capirne la vera efficienza. Il risultato ha portato ad una previsione di 4 negozi a rischio chiusura su un totale complessivo di 29.

La **data visualization** è un termine generico che descrive qualsiasi tentativo di aiutare le persone a comprendere il significato dei dati analizzati posizionandoli in un contesto visivo. Modelli, tendenze e correlazioni che potrebbero non essere rilevati nei dati basati su testo possono essere esposti e riconosciuti più facilmente tramite report utilizzando dei software di visualizzazione dei dati come per esempio Microsoft Power BI.

Il documento prodotto viene chiamato *report.* Nel primo esempio presentato, sono mostrati i risultati ottenuti grazie ad una analisi dei negozi chiusi in Italia dal 2017 ad oggi dell’azienda cliente del progetto analizzato, identificando la causa della chiusura di essi, classificata come una chiusura per un cattivo margine di profitto, o una chiusura per un mercato coperto negli anni o la nascita di un nuovo negozio nei dintorni. Il risultato è visibile attraverso la mappa sottostante che racchiude due esempi per ciascun tipo di chiusura. I negozi di Serravalle mostrano una chiusura di tipo new shop, mentre il negozi di padova e modena di un margine non adeguato.

Le immagini seguenti mostrano un esempio in Power BI di cosa si intende per una dashboard. La sottostante, comprende una analisi generale dei canali di vendita con associati i relativi guadagni, dei prodotti, e del made in. Sarà la visualizzazione di default per il cliente. Il primo grafico (BarChart) comprende la visualizzazione del fatturato totale del 2019 riferente al canale del negozio, suddiviso in Property Full Price e Property Otlet. Il secondo grafico è la rappresentazione 80/20 di Pareto, dove è evidende che gran parte delle vendite deriva dalle borse. L’ultimo grafico, invece, mostra la percentuale sul totale della provenineza e fattura dei prodotti.

Tramite l’utilizzo di filtri o semplice navigazione è possibile passare da una analisi molto generale ad una analisi molto più dettaglia in ogni singolo particolare. L’immagine iniziale è quella seguente, spiegano le variazioni di prodotti venduti e made in grazie ad una selezione del tipo di negozio che si vuole prendere in considerazione. L’ultima, invece, usa la stessa idea, ma la selezione avviene per il prodotto borse.

È molto importante osservare come i grafici interagiscono tra loro portando ad una analisi rapida ed efficacie, scelta in base all’esigenza del cliente.