# ***CAPITOLO 2: LA CREAZIONE DELLE DATA MART DI FURLA***

## ***2.1 FURLA DWH***

Furla è un’azienda fondata a Bologna nel 1927 da Aldo e Margherita Furlanetto che produce accessori uomo/donna sulla quale ha attuato un processo di ritorno in Italia dalla Cina. Dagli anni ’90 l’azienda ha attuato un processo di internazionalizzazione nella produzione degli accessori in pelle fino a delocalizzare in Cina parte della produzione mantenendo in Italia tecnici specializzati per il controllo qualità al fine di non diminuire il valore del marchio. La scelta non ha però portato a delle riduzioni dei costi o a dei miglioramenti effettivi, con gli anni si sono resi evidenti alcuni problemi che stavano frenando la competitività dell’azienda. Il controllo qualità di una produzione geograficamente distante era diventato troppo costoso, per i viaggi dei tecnici o per la necessità di trasportare i capi in Italia, alcune operazioni andavano obbligatoriamente svolte in Italia per la presenza di operatori specializzati e con il necessario know-how, infine il time-to-market dell’azienda era diventato un problema, i prodotti ritardavano rispetto ai concorrenti con una produzione più vicina ai mercati di sbocco o con una produzione più concentrata geograficamente (il 23% del fatturato aziendale è realizzato in Italia). Dal 2010 parte della produzione torna in Italia, tramite l’acquisizione di uno stabilimento in Toscana, l’azienda così riesce ad ottenere ulteriori risorse con conoscenze specifiche nel settore e nella fase di progettazione e campionatura delle collezioni, tiene interamente in Italia la produzione dei capi con un impatto delle skill e delle abilità manuali più elevate, aumenta il fascino e il richiamo del brand che nel frattempo riesce anche ad avere un time-to-market minore e un controllo maggiore sulla filiera.

Il ritorno in Italia ha permesso di sfruttare non solo il Made In Italy di per sé ma anche di enfatizzare le abilità di progettazione e realizzazione che secondo Poletto il Paese ha nella produzione di abiti e accessori di alta gamma. Furla, nonostante l’aumento della quota prodotta in Italia resta convintamente un brand internazionale e che riesce a trovare nel proporre uno stile italiano uno dei valori differenziali che danno redditività e crescita all’azione economica, inoltre tra le novità del brand, permesse dall’efficientamento della supply chain, c’è la realizzazione di capi e accessori su misura, proposta appositamente per il flagship di Shangai, con un prodotto interamente Made In Italy consegnato in soli tre mesi dal momento della richiesta e la nuova linea di Sneakers presentata nel febbraio 2019 alla fashion week di Milano.

### ***~~2.1.1 Datawarehouse Del DB Oracle di Furla~~***

## ***2.2 EXTRACTION, TRANSFORMATION AND LOADING (ETL)***

Il ruolo degli strumenti di ETL è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il Data Warehouse. Le operazioni da essi svolte vengono spesso indicate con il termine *riconciliazione* che, durante il processo di alimentazione del Data Warehouse (DW) avviene in due occasioni: quando il DW viene popolato per la prima volta e periodicamente quando viene aggiornato. La riconciliazione consiste di quattro distinti processi detti rispettivamente:

1. Estrazione (extraction o capture);

2. Pulitura (cleaning o scrubbing);

3. Trasformazione (transformation);

4. Caricamento (loading).

In linea generale il confine tra pulitura e trasformazione è abbastanza nebuloso quindi per semplicità si assume che l’operazione di pulitura sia essenzialmente mirata alla correzione dei valori dei dati, mentre la trasformazione si occupa più propriamente del loro formato.

### ***2.2.1 Extraction***

La Data Integration è composta da due sottofasi chiamate estrazione e pulitura.

Durante la prima sottofase i dati rilevanti vengono estratti dalle sorgenti e questa operazione può essere di tipo:

* *Statico:* viene effettuata quando il DW deve essere popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali;
* *Incrementale:* viene usata per l’aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall’ultima estrazione. L’idea alla base è quella di utilizzare i cambiamenti registrati a livello dei dati per aggiornare il DW. I benefici derivabili sono: volume molto piccolo dei dati coinvolti di volta in volta nell’operazione rispetto all’estrazione statica; la maggior parte dei dati nel DW restano invariati e vengono analizzati solo i datai che hanno subito modifiche. Vengono usate tecniche CDC (Change Data Capture) che permettono di monitorare le sorgenti dati con l’obiettivo di individuare i cambiamenti avvenuti a livello dei dati. Queste tecniche sono particolarmente importanti per la Data Warehouse maintenance grazie alla propagazione dei cambiamenti rilevati a livello della sorgente

Invece, la Pulitura è la sottofase che si occupa di migliorare la qualità dei dati andando ad eliminare dati “sporchi” dovuti a duplicazioni, inconsistenze, dati mancanti, valori errati etc.

Le principali funzionalità di pulitura dei dati riscontrabili negli strumenti ETL sono la correzione e l’omogeneizzazione, che utilizzano dizionari appositi per correggere errori di scrittura e riconoscere sinonimi, e la pulitura basata su regole, che applica regole proprie del dominio per stabilire le corrette corrispondenze tra i valori.

### ***2.2.2 Trasformation***

È la fase centrale del processo di riconciliazione e ha l’obiettivo di convertire i dati dal formato operazionale sorgente a quello del DW. Tra le funzionalità di questo livello per l’alimentazione del livello dei dati riconciliati si hanno:

* *conversione e normalizzazione*, che operano sia a livello di formato di memorizzazione sia a livello di unità di misura al fine di uniformare i dati;
* *matching*, che stabilisce corrispondenze tra campi equivalenti in sorgenti diverse;
* *selezione*, che riduce, se necessario, il numero di campi e record rispetto alle sorgenti.

Nella fase di alimentazione del DW si hanno invece due sostanziali differenze:

* la normalizzazione viene sostituita dalla denormalizzazione;
* si introduce l’aggregazione che realizza le opportune sintesi dei dati.

### ***2.2.3 Loading***

In questa fase avviene il caricamento dei dati sul DW attraverso due modalità alternative:

* *refresh:* i dati vengono riscritti integralmente sostituendo completamente quelli precedenti. In generale questa tecnica viene utilizzata solo durante la fase iniziale di popolamento del DW;
* *update:* vengono aggiunti al DW solo i cambiamenti avvenuti sui dati senza sovrascrivere ad ogni iterazione tutti i dati. Questa tecnica viene utilizzata, in abbinamento all’estrazione incrementale per l’aggiornamento periodico del DW.

Un modo per ridurre il tempo di caricamento è quello di parallelizzare il processo ETL. Questo può

si verificano in due modi: più passaggi eseguiti in parallelo e un singolo passaggio in esecuzione

in parallelo.

* *Passi di carico multipli*. Il flusso di lavoro ETL è diviso in più indipendenti lavori presentati insieme. È necessario riflettere attentamente su ciò che accade ogni lavoro; l'obiettivo principale è creare posti di lavoro indipendenti.
* *Pipeline*. Il database stesso può anche identificare determinati compiti che può

eseguire in parallelo. Ad esempio, la creazione di un indice può essere in genere parallela

attraverso tutti i processori disponibili sulla macchina.

### ***2.2.4 Possibili Problemi Nell’ ETL E Come Risolverli***

Dopo che il sistema ETL è in produzione, i guasti possono verificarsi per innumerevoli motivi oltre il controllo del processo ETL. Cause comuni di guasti alla produzione di ETL includere:

* Errore di rete
* Errore del database
* Errore del disco
* Errore di memoria
* Errore nella qualità dei dati
* Aggiornamento di sistema senza preavviso

Per proteggersi da questi guasti, è necessario un solido sistema di backup e un sistema compagno di ripristino e riavvio. Devi pianificare per errori irreversibili durante il caricamento perché accadrà. Il sistema dovrebbe anticipare questo e fornire funzionalità di recupero, arresto e riavvio di arresto anomalo.

Ad esempio, Per un processo di caricamento dovrebbe impegnare serie relativamente piccole di record alla volta e tenere traccia di ciò che è stato commesso. La dimensione del set dovrebbe essere regolabile perché le dimensioni della transazione hanno implicazioni di prestazioni su diversi DBMS.

Il sistema di ripristino e riavvio viene utilizzato, ovviamente, per riprendere un lavoro che è entrato in errore e si è fermato o per far riportate l'intero lavoro indietro tramite backup e riavviarlo. Questo sistema è significativo dipende dalle capacità del sistema di backup. Quando si verifica un errore,la reazione iniziale istintiva è tentare di salvare qualsiasi cosa sia stata elaborata e riavviare il processo da quel punto. Ciò richiede uno strumento ETL solido e affidabile funzionalità di checkpoint, in modo che possa determinare perfettamente cosa ha elaborato e cosa non deve riavviare il lavoro esattamente nel punto giusto. In molti casi, potrebbe essere meglio uscire da tutte le righe che sono state caricate come parte del processo e riavviare dall'inizio.

Per questo motivo è consigliato di progettare tabelle dei fatti con un surrogato primario a singola colonna chiave. Questa chiave surrogata è un numero intero semplice che viene assegnato in sequenza come le righe vengono create per essere aggiunte alla tabella dei fatti. Con la chiave surrogata della tabella dei fatti, puoi facilmente riprendere un carico che viene fermato o estrarre tutte le righe nel carico limitando un intervallo di chiavi surrogate.

Quanto più un processo ETL è lungo, tanto più devi essere consapevole delle vulnerabilità a causa di un errore. La progettazione di un sistema ETL modulare composto da processi efficienti, resistenti agli arresti anomali e alle interruzioni impreviste, può ridurre il rischio di un guasto con conseguente notevole recupero. Un'attenta considerazione di quando mettere fisicamente i dati scrivendoli su disco, insieme a punti di recupero accuratamente predisposti e caricamento di data / ora o di tabelle sequenziali dei fatti consente di specificare la logica di riavvio appropriata.

## ***2.3 ETL DAL DB ORACLE DI FURLA***

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

Tuttavia, le più recenti evoluzioni del mercato hanno portato ad un aumento della domanda di prodotti completi di Data Integration, portando al 60% la percentuale delle imprese che utilizza uno dei pacchetti di prodotti di integrazione offerti sul mercato, con lo scopo di effettuare attività di Business Intelligence [8].

La recente crisi economica ha portato inoltre ad una diminuzione dei budget assegnati allo sviluppo dell’Information Technology nelle aziende, determinando un incremento dell’adozione di soluzioni di integrazione open source.

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

* **Software personalizzati**: con l’emergere delle prime necessità di integrazione dei dati molte imprese svilupparono internamente prodotti ad hoc in grado di rispondere alle esigenze specifiche del proprio ambito di business. Con la maturazione del mercato dei prodotti di Data Integration questo tipo di approccio è divenuto sempre meno conveniente. Inoltre, l’emergere di architetture SOA e applicazioni SaaS sta decretando la fine dei prodotti sviluppati in casa. Oggi le suite di Data Integration presenti sul mercato offrono sicuramente funzionalità e affidabilità migliori;
* **Software proprietari**: lo sviluppo di applicativi di Data Integration ha contribuito ad aumentare la produttività delle attività collegate all’integrazione dei dati. I prodotti di integrazione dei dati sono maturati costantemente negli anni garantendoun ventaglio di funzionalità sempre più ricco e variegato, rendendo tali applicativi idonei a supportare la grande maggioranza degli scenari di business che richiedono l’utilizzo di dati integrati. Il numero di applicativi sul mercato è oggi elevato, si va dalle suite di prodotti in grado di coprire la quasi totalità delle necessità aziendali a prodotti specializzati in particolari contesti di business o specifiche problematiche;
* **Software open source**: il limite dei maggiori prodotti proprietari presenti sul mercato sono i costi necessari per la loro implementazione. Per venire in contro alle necessità delle aziende più piccole e con risorse limitate si sono da poco affacciate sul mercato i primi prodotti open source, prodotti in grado di supportare una discreta quantità di funzioni ma con un costo decisamente minore rispetto ai prodotti proprietari (costi di licenza nulli, costi di infrastruttura ridotti, servizi pagati in base all’utilizzo).

Per la mia Tesi abbiamo deciso di utilizzare un Software Open Source che offre tutte le funzionalità necessarie per lo svolgimento, minimizzando al massimo i costi: Talend.

### ***2.3.1 Talend***

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

* *Talend Open Studio*: suite gratuita scaricabile gratuitamente con licenza open source (GPL). Talend Open Studio si presenta come prodotto di Data Integration completo e contraddistinto da un ampia gamma di funzionalità, sufficienti per la maggior parte delle necessità;
* *Talend Integration Suite:* è una versione potenziata del prodotto gratuito che aggiunge funzionalità avanzate come lo sviluppo collaborativo e monitoraggio avanzato del progetto.

Per chi non possiede l’hardware necessario per supportare il sistema c’è una terza opzione costituita da Talend On Demand, ovvero un offerta di tipo Software ad a Service (SaaS).

I prodotti di Talend offrono ad oggi le seguenti funzionalità:

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

Vediamo quindi quali sono i punti di forza dell’approccio di open source di Talend [7]:

* *Nessuna barriera all’adozione*: la disponibilità gratuita del prodotto di base rende praticamente immediata l’installazione del software. Talend supporta il cliente attraverso tutorial sull’utilizzo di base, inoltre è possibile fare affidamento ad una vasta comunità di utilizzatori;
* *Curva di apprendimento veloce*: il prodotto si presenta graficamente amichevole e facile da utilizzare. L’interfaccia grafica è intuitiva e l’utilizzo delle funzionalità di base non richiede particolari addestramenti;
* *Modello di prezzi stabile e prevedibile*: i prodotti proprietari prevedono spesso costi elevati man mano che si espandono le funzionalità e le capacità del prodotto, con costi di licenza che aumentano all’aumentare delle macchine installate. Questo rende spesso difficile una corretta previsione dei costi nelle fasi iniziali del progetto, soltanto a lavoro ultimato è possibile rendersi conto del costo effettivo della soluzione adottata. Talend prevede un modello di costo basato sul numero di sviluppatori e sull’utilizzo del servizio, indipendente da licenze, hardware e quantità di dati da integrare;
* *L’importanza di una comunità a supporto*: la comunità online di esperti ed utilizzatori del prodotto è già oggi molto vasta ed è un fattore di grande importanza per facilitare l’implementazione e il mantenimento delle soluzioni offerte da Talend. Forum, wiki, guide e contributi gratuiti degli utenti rappresentano un valore aggiunto che solo un prodotto di questo tipo può offrire;
* *Ampio supporto a tipologie di dati differenti*: con oltre 400 connessioni preimpostate la soluzione di Talend garantisce la compatibilità con un grande numero di sistemi, database, pacchetti di software, applicazioni gestionali, servizi
* web, ecc. Nessun’altra soluzione sul mercato vanta un numero di possibili connessioni così elevato;
* *Flessibilità, versatilità e riuso del prodotto*: Talend non si limita ad un supporto alle tecniche standard di ETL ma permette l’implementazione di diverse strategie di integrazione. La possibilità di riuso di progetti già perfezionati costituisce inoltre un altro punto di forza dell’approccio open source;
* *Funzionalità e performance*: il livello di funzionalità offerto è paragonabile a quello dei prodotti proprietari. Tuttavia, si registrano alcune lacune nel campo della modellazione dei dati, data quality e data mining. Un team di ricerca e sviluppo dedicato permette al prodotto di essere sempre aggiornato alle ultime esigenze del mercato e di proporre funzionalità innovative;
* *Costi e tempi ottimizzati*: le soluzioni offerte da Talend risultano da un 50% a un 80% più economiche rispetto ai prodotti tradizionali, essendo meno costose da acquisire e mantenere e permettono uno sviluppo più rapido del sistema di integrazione.

### ***~~2.3.2 Talend Data Integration~~***

### ***~~2.3.3 Talend Trasformation & Loading~~***

## ***~~2.4 ETL IN REAL TIME DA TWITTER~~***

### ***~~2.4.1 Python Code For Twitter API~~***

### ***~~2.4.2 Kafka-Confluent~~***

### ***~~2.4.3 Sink Connectors~~***

### ***~~2.4.3 K-SQL~~***

## ***2.5 FURLA DATA MART***

Generalmente un Data Mart viene estratto da un DWH, ma talvolta può essere costituito anche in assenza di un sistema di dati integrato. Nello specifico, 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 [11].

L’implementazione può essere di due tipi:

* + 1. *Top-Down:* costruzione del DWH, e conseguente aggregazione ed esportazione nei vari Data Mart.
    2. *Bottom-Up:* concentrandosi su aree specifiche del business si costruiranno i vari Data Mart per poi giungere alla costruzione del DWH. In questo modo si avrà un approccio scalabile.

Tabella 2: DWH vs DATA MART

|  |  |  |
| --- | --- | --- |
|  | **Data Warehouses** | **Data Marts** |
| Finalità | Application-neutral  Centralizzati e condivisi  Intera impresa | Applicazioni specifiche  Dipartimenti o aree |
| Dati | Bassa denormalizzazione | Alta denormalizzazione |
| Soggetti utilizzatori | Soggetti di molte aree | Soggetti di una singola area |
| Sorgenti dei dati | Molte  Dati esterni operazionali | Poche  Dati esterni operazionali |
| Caratteristiche | Flessibile, estensibile  Lunga vita  Data-oriented | Ristretto, non estensibile  Vita breve  Project-orientation |
| Tempo d’implementazione | 9-18 mesi per il primo stadio | 4-12 mesi |

### ***2.5.1 Modello Multidimensionale - DFM***

Il modello E-R (Modello Entità – Relazione), diffuso per progettare sistemi informativi relazionali, non è adatto per esprimere e analizzare in modo dettagliato grandi moli di dati. Sarà necessaria l’adozione di un nuovo modello concettuale: il modello multidimensionale o DFM (Data Fact Model) [10].

Attraverso questa modellazione è possibile rappresentare i dati all’interno di Hypercube che forniscono istantaneamente le dimensioni di analisi e i fatti di interesse legati al business.

Gli scopi principali del sistema sono:

* Fornire supporto al design concettuale;
* Creare un ambiente dove gli utenti possano fare query in maniera intuitiva e formale;
* Favorire la comunicazione tra designer a utenti al fine di formalizzare i requisiti di progetto;
* Costruire una stabile piattaforma di design logico;
* Fornire una documentazione chiara e efficace.

Un modello multidimensionale 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.

Per navigare all’interno del cubo multidimensionale esistono differenti operazioni che permettono di organizzare i dati al suo interno, attraverso diverse prospettive [10].

La prima è il *Pivoting che* permette di modificare rapidamente la visualizzazione dei dati girando gli assi del cubo. Questo consentirà per l'appunto di cambiare il punto di vista da cui si analizza i dati del cubo. La seconda è invece lo *Slice* & *Dice che* seleziona e proietta i dati del cubo. Nello specifico si estrarranno sotto-cubi filtrando su una (Slice) o più (Dice) dimensioni e, infine, abbiamo il *Roll-Up* & *Drill-Down che* consentono di spostarsi all'interno di una gerarchia, scegliendo il livello di aggregazione secondo il quale l'utente desidera analizzare i dati. Nello specifico si salirà di un livello gerarchico con il roll- up, viceversa si userà il drill-down.

### ***2.5.1 Star Schema***

Una volta costruito il DFM, 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.

Solitamente, le tabelle dei fatti in uno Star Schema sono in terza forma normale, mentre le tabelle dimensionali sono denormalizzate [10]. 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.

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