2.2.3. Popolamento del BDW

Il popolamento del BDW è uno degli aspetti più difficoltosi e dispendiosi nella realizzazione di un DW. Nel processo di popolamento si distinguono tre fasi fondamentali:

1. Cattura dei dati;
2. Trasformazione dei dati catturati;
3. Applicazione delle modifiche al BDW;

2.2.3.1. Cattura dei dati

Il processo di cattura consiste nell’interagire con le data sources per ottenere o una copia dei dati contenuti in essi o la lista delle modifiche avvenute.

Avremo sostanzialmente due differenti tipologie di cattura :

* **Cattura Statica.** La cattura statica fornisce una copia dei dati sorgente in un determinato istante di tempo. Tale copia potrà contenere tutti i dati trovati nei sistemi operazionali o soltanto un sottoinsieme di essi. La cattura statica è comunemente utilizzata la prima volta che un insieme di dati è caricato nel BDW o quando i sistemi operazionali mantengono la storia dei dati in essi contenuti.
* **Cattura Incrementale.** La cattura incrementale a differenza della cattura statica fornisce soltanto l’insieme dei dati che sono stati aggiunti, modificati o cancellati in un certo intervallo di tempo. Se da una parte si ha che il volume dei dati è sensibilmente minore rispetto alla cattura statica, dall’altra si ha una maggiore complessità negli algoritmi di cattura.

Nel caso di una cattura incrementale possono essere utilizzate *tecniche di cattura immediate*, che catturano i cambiamenti nel momento stesso in cui si verificano, e *tecniche di cattura differite* che vengono applicate ad intervalli di tempo predefiniti e che quindi cattureranno le modifiche avvenute in quel particolare intervallo di tempo. E’ ovvio che nel caso di dati transitori o semi-periodici, le uniche tecniche utilizzabili sono quelle immediate.

2.2.3.2. Trasformazione dei dati catturati

Il processo di trasformazione si applica ai dati catturati prima che questi vengano ad essere caricati nel BDW: attraverso un certo numero di funzioni di trasformazione, i record in input vengono opportunamente modificati per prestarsi meglio alle esigenze del Data Warehouse. Le funzioni di trasformazioni possono essere classificate in sei fondamentali categorie: le prime quattro operano a livello di record mentre le ultime due operano a livello di attributi.

1. **Funzioni di selezione**: sono le più semplici funzioni di trasformazione, e consentono di selezionare una parte o tutto il record in input;
2. **Funzioni di separazione/concatenazione**: una funzione di separazione ha il compito di separare le informazioni contenute in un singolo record, mentre una funzione di concatenazione ha il compito di unire in uno stesso record informazioni provenienti da più record distinti. Nel processo di popolamento le funzioni di separazione sono poco usate (in quanto vanno contro le esigenze dei dati riconciliati), mentre invece sono molto usate le funzioni di concatenazione.
3. **Funzioni di normalizzazione/denormalizzazione**: le funzioni di normalizzazione e denormalizzazione hanno pressappoco la stessa utilità di quelle di separazione e concatenazione: la differenza con le precedenti sta nel modo di trattare le chiavi. La normalizzazione attua sempre la divisione di un record, però mantiene il collegamento tra i campi aggiungendo o delle chiavi o dei nuovi records in base al tipo di relazione che si crea tra le parti ottenute. Un esempio di processo di normalizzazione è il seguente:

Matricola

Nome

Esame

Voto

#### **Matricola**

Nome

#### **Codice\_Esame**

Nome

**Matricola**

**Codice\_Esame**

Voto

*Studente*

*Studente*

###### Esame-Superato

*Esame*

1. **Funzioni di aggregazione**: le funzioni di aggregazioni prendono in input dati ad un alto livello di dettaglio e producono in output dati aggregati. Dal momento che i dati coinvolti nel BDW sono ad un alto livello di dettaglio, tali funzioni raramente sono coinvolte nel processo di popolamento.
2. **Funzioni di conversione**: le funzioni di conversione non operano a livello di record ma a livello di singolo attributo. Preso in input un attributo le funzioni di conversione vi applicano una regola di trasformazione ottenendo in output lo stesso attributo ma in un’altra forma (ad esempio conversione tra unità di misura).
3. **Funzioni di arricchimento**: sono funzioni che combinano i dati provenienti da due o più campi in uno o più record per creare un nuovo campo o nuovi campi nel record di output.

Alcuni problemi di qualità, riscontrabili nei sistemi operazionali, possono rendere estremamente complesso il processo di trasformazione:

* *Uso di un insieme di chiavi non consistente in data sources distinti.* Nella maggior parte dei casi la scelta della codifica delle chiavi non è univoca per l’intera organizzazione, ma è trattata a livello di singola applicazione. Questo può portare a chiavi differenti per identificare lo stesso prodotto in data sources distinti. Il processo di trasformazione deve portare all’uso di un identificatore univoco per ogni oggetto presente nei dati operazionali.
* *Uso di un alto grado di codifica.* Nei sistemi operazionali molto vecchi era pratica comune ricorrere a codifiche dei dati che permettessero di migliorare le prestazioni del sistema e di ridurre lo spazio su disco occupato; tali codifiche hanno però il grosso svantaggio di essre sconosciute all’utente del sistema e vanno quindi rimosse nel passaggio dei dati nel BDW.
* *Presenza di errori nei dati.* L’esistenza di errori nei dati operazionali è sfortunatamente molto estesa. Il processo di trasformazione dovrà per quanto possibile risolvere tali errori.
* *Uso di campi per scopi diversi da quelli preventivati.* Spesso alcuni campi vengono utilizzati per uno scopo diverso da quello per cui erano statiprogettati .Nel popolare il BDW andrà fatta molta attenzione nel riconoscere il significato di un dato memorizzato in un campo usato per diversi scopi.

2.2.3.3. Applicazione delle modifiche al BDW

I dati in output dal processo di trasformazione vanno trasportati nel Business Data Warehouse in due diverse modi :

* inizialmente come parte del processo di creazione del BDW ;
* successivamente come parte del processo di mantenimento.

Le tecniche di applicazione delle modifiche al BDW sono sostanzialmente quattro:

1. **Load.** L’insieme dei dati da inserire viene caricato nel BDW sostituendo, eventualmente ce ne fossero, i dati preesistenti.
2. **Append.** L’insieme dei dati da inserire viene aggiunto ai dati preesistenti che quindi sono sempre preservati. Alcuni nuovi records potranno essere la duplicazione di records già presenti nel sistema ed eventualmente si può decidere di non permetterne l’inserimento.
3. **Merge distruttivo.** Attraverso questa tecnica di cattura, l’insieme dei dati da inserire viene unito con l’insieme dei dati già esistenti nel BDW. Se la chiave di un nuovo record da inserire coincide con la chiave di un record esistente, quest’ultimo viene aggiornato di conseguenza, altrimenti un nuovo record viene inserito.
4. **Merge costruttivo.** E’ simile alla tecnica precedente, con la differenza che un dato catturato è sempre aggiunto come nuovo record. Se esiste già un record con la stessa chiave, questo non viene sovrascritto ma opportunamente

marcato per indicare che esiste nel sistema un record più aggiornato con la stessa chiave di business.

E’ abbastanza evidente che le diverse tecniche presentate hanno complessità crescente. Dalle caratteristiche di queste funzioni si capisce che non possono essere usate indifferentemente per il processo di popolamento iniziale e per il mantenimento; in particolare, nel processo di popolamento iniziale si utilizzano la load e l’append, mentre nella fase di gestione le rimanenti due.

2.2.4. Archiviazione e recupero dei dati nel BDW

Il BDW deve mantenere dati storici, ovvero deve conservare l’evoluzione del business negli anni; data l’enorme quantità di informazione, che quindi potenzialmente deve conservare, a volte è preferibile memorizzare i dati di uso meno frequente in supporti più economici anche se meno facilmente accessibili ed efficienti. I dispositivi fisici di memoria che si possono utilizzare hanno caratteristiche differenti di costo e di velocità di accesso: si va dai dischi magnetici per i dati più usati , ai dischi ottici o ai nastri per quelli meno usati. I dati archiviati sono a tutti gli effetti dati del warehouse; assieme ai dati, per la comprensione degli stessi, vanno archiviati anche i metadati. La scelta del particolare dispositivo di memorizzazione da utilizzare può avvenire in base a diversi criteri: oltre quello di frequenza d’uso altri possono essere l’età dei dati, le categorie di appartenenza,le distribuzioni geografiche e qualunque classificazione che abbia un senso in termini del business. Ci sono due modi generali per scegliere il momento in cui i dati vanno archiviati :

1. *archiviazione al tempo di cancellazione.* Un dato viene archiviato nel momento in cui viene cancellato dall’insieme dei dati on-line;
2. *archiviazione al tempo di creazione.* Un dato è archiviato contemporaneamente alla sua creazione pur continuando a restare on-line.

Il secondo metodo è preferibile in quanto fornisce i seguenti vantaggi:

* nel caso di perdita accidentale dei dati non servono procedure di data recovery, ma basta utilizzare le procedure di data retrivial;
* l’eliminazione dei dati on-line diventa una semplice cancellazione;
* la sincronizzazione dell’archiviazione dei metadati diventa più semplice;
* già nelle prime fasi della progettazione la scelta del posto fisico dell’archiviazione è più flessibile di quanto non avvenga quando la disponibilità dello spazio su disco si esaurisce.