**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 operazione 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.

*Estrazione*

Durante questa fase 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 Wata Warehouse maintenance grazie alla propagazione dei cambiamenti rilevati a livello della sorgente.

***Pulitura***

È la fase 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.

***Trasformazione***

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.

*Caricamento*

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.