2. Basi di Dati e Data Warehouse

2.1. Le Basi di Dati

Uno dei principali compiti dei sistemi informatici consiste nelle attività di raccolta, organizzazione e conservazione dei dati. Le quotazioni delle azioni nei mercati telematici internazionali, i movimenti dei conti correnti bancari, gli elenchi dei degenti in un ospedale sono esempi di dati indispensabili a gestire alcune attività umane. I sistemi informatici garantiscono che questi dati vengano conservati in modo permanente su dispositivi per la loro memorizzazione, aggiornati in caso di variazione e resi accessibili alle interrogazioni degli utenti. La base di dati può essere definita come una collezione di dati riguardanti uno stesso argomento, o più argomenti correlati tra loro, strutturata in modo tale da consentire che i dati possano venire utilizzati per diverse applicazioni e, normalmente, possano evolvere nel tempo.
Il DBMS o Data Base Management System è un sistema software efficiente ed efficace progettato per assistere al mantenimento e all'utilizzo di grandi collezioni di dati, assicurando condivisione, persistenza e affidabilità. La base di dati è una collezione di dati gestita da un DBMS.
Le caratteristiche generali di un DBMS sono le seguenti[1]:

- permettere agli utenti di creare nuovi database e di specificare i rispettivi schema (struttura logica dei dati), utilizzando un linguaggio specializzato chiamato *data definition language*;
- dare agli utenti l'abilità di eseguire query ai dati e di modificarli, utilizzando un apposito linguaggio detto *query language* o *data-manipulation language*;
- supportare il salvataggio di un ammontare enorme di dati per un lungo periodo di tempo, rendere questi sicuri e garantire un efficiente accesso ai dati;
- controllare l'accesso da più utenti alla volta, evitando accessi simultanei allo stesso dato per garantirne l'integrità.

2.2. I Data Warehouse

Un data warehouse è una collezione di dati di supporto per il processo decisionale, fisicamente separato dai sistemi operazionali che presenta le seguenti caratteristiche:

- è orientata ai soggetti di interesse;
- è integrata e consistente;
- è rappresentativa dell'evoluzione temporale e non volatile.

La costruzione di un sistema di data warehousing non comporta l'inserimento di nuove informazioni bensì la riorganizzazione di quelle esistenti, e implica pertanto l'esistenza di un sistema informativo. Mentre i dati operazionali coprono un arco temporale di solito piuttosto limitato, poiché la maggior parte delle transazioni coinvolge i dati più recenti, il DW permette analisi che spazino sulla prospettiva di alcuni anni. Per questo motivo, il DW è aggiornato ad intervalli regolari ed è in continua crescita. Proprio per il fatto che, in linea di principio, non vengano mai eliminati dati dal DW e che gli aggiornamenti siano tipicamente eseguiti quando il DW è off-line, fa sì che un DW possa essere considerato come un database di sola lettura. Questa caratteristica, insieme all'esigenza degli utenti di contenere i tempi di risposta alle interrogazioni di analisi, comporta varie conseguenze: nei DW perdono importanza le tecniche sofisticate di gestione delle transazioni adottate dai DBMS e la pratica della normalizzazione delle tabelle viene abbandonata a favore di una parziale denormalizzazione mirata al miglioramento delle prestazioni.
Il tipo di elaborazione per cui nascono i DW viene detto *On-Line Analytical Processing* (OLAP), ed è caratterizzato da un'analisi dinamica e multidimensionale che richiede la scansione di un enorme quantità di record (il metro di misura è sull'ordine dei milioni) per calcolare un insieme di dati numerici di sintesi. Le peculiari caratteristiche delle interrogazioni OLAP fanno sì che i dati nel DW siano normalmente rappresentati in forma multidimensionale. L'idea di base è quella di vedere i dati come punti in uno spazio le cui dimensioni corrispondono ad altrettante possibili dimensioni di analisi.

Le funzioni di base di uno strumento OLAP sono[2]:

- *Slicing:* è l'operazione di rotazione delle dimensioni di analisi. È un'operazione fondamentale se si desidera analizzare totali ottenuti in base a dimensioni diverse o se si vogliono analizzare aggregazioni trasversali;

- *Dicing:* è l'operazione di 'estrazione' di un subset di informazioni dall'aggregato che si sta analizzando. L'operazione di dicing viene eseguita quando l'analisi viene focalizzata su una 'fetta del cubo ' di particolare interesse per l'analista. In alcuni casi l'operazione di dicing può essere 'fisica' nel senso che non consiste solo nel filtrare le informazioni di interesse ma magari nell'estrarle dall'aggregato generale per distribuirne i contenuti;

- *Drill-down:* è l'operazione di 'esplosione' del dato nelle sue determinanti. L'operazione di drill-down può essere eseguita seguendo due tipologie di *sentiero*: la *gerarchia* costruita sulla dimensione di analisi, oppure la *relazione matematica* che lega un dato calcolato alle sue determinanti. Si può comprendere l'importanza di tale operazione ai fini analitici in termini di comprensione delle determinanti di un dato;

- *Drill-across:* è l'operazione mediante la quale si naviga attraverso uno stesso livello nell'ambito di una gerarchia. Come visto precedentemente il passaggio dalla famiglia di prodotti alla lista dei prodotti è un'operazione di drill-down, il passaggio da una famiglia ad un'altra famiglia è un'operazione di drill-across;

- *Drill-through:* concettualmente simile al drill-down, è l'operazione mediante la quale si passa da un livello aggregato al livello di dettaglio appartenente alla base dati normalizzata.

2.3. Architettura per il data warehousing

Le caratteristiche architetturali principali per un sistema di data warehousing possono essere definite nel seguente modo:

- *Separazione*: l'elaborazione analitica e quella transazionale devono essere mantenute il più possibile separate.
- *Scalabilità*: l'architettura hardware e software deve poter essere facilmente ridimensionata a fronte della crescita nel tempo dei volumi di dati fa gestire ed elaborare e del numero di utenti da soddisfare.
- *Estendibilità*: deve essere possibile accogliere nuove applicazioni e tecnologie senza riprogettare integralmente il sistema.
- *Sicurezza*: il controllo sugli accessi è essenziale a causa della natura strategica dei dati memorizzati.

- *Amministrabilità*: la complessità dell'attività di amministrazione non deve risultare eccessiva.

Esistono diverse architetture per la progettazione di un sistema data warehousing, tra queste si distingue l'*architettura a due livelli*; il nome deriva dalla volontà di evidenziare la separazione tra il livello sorgenti e quello del DW, sebbene in realtà si articola su quattro livelli distinti:

1. *Livello delle sorgenti*. Il DW utilizza fonti di dati eterogenei, essi possono essere estratti dall'ambiente di produzione e quindi originariamente archiviati in database aziendali relazionali, oppure provenire da sistemi informativi esterni all'azienda.
2. *Livello dell'alimentazione*. I dati memorizzati nelle sorgenti devono essere estratti, ripuliti per eliminare le inconsistenze e completare eventuali parti mancanti. I cosiddetti strumenti ETL (*Extraction, Transformation and Loading*) permettono di integrare schemi eterogenei, nonché di estrarre, trasformare, pulire, validare, filtrare e caricare i dati dalle sorgenti nel DW.
3. *Livello del warehouse*. Le informazioni vengono raccolte in un DW. Esso può essere direttamente consultato ma anche usato come sorgente per costruire *data mart*. Con il termine data mart si intende un sottoinsieme o un'aggregazione dei dati presenti nel DW primario.
4. *Livello di analisi*. Permette la consultazione efficiente e flessibile dei dati integrati a fini di stesura di report, di analisi, di simulazione.

2.4. Gli strumenti ETL

Il ruolo degli strumenti di *Extraction, Transformation and Loading* è quello di alimentare una sorgente dati singola, dettagliata esauriente e di alta qualità che possa a sua volta alimentare il DW. Le fasi distinte di questa operazione possono essere suddivise in:

1. *Estrazione*. Durante questa fase i dati rilevanti vengono estratti dalle sorgenti. L'*estrazione statica* viene effettuata quando il DW deve essere popolato per la prima volta. L'*estrazione incrementale* viene usata per l'aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall'ultima estrazione.
2. *Pulitura*. La pulitura è una fase critica nel processo di data warehousing, poiché si incarica di migliorare la qualità dei dati. Esistono diverse inconsistenze che possono rendere *sporchi* i dati: dati duplicati, dati mancanti, valori errati, valori inconsistenti.

3. *Trasformazione*. Durante questa operazione i dati vengono convertiti dal formato operazionale sorgente a quello del DW. La corrispondenza con il livello sorgente è in genere complicata dalla presenza di più fonti distinte eterogenee, che richiede durante la progettazione una complessa fase di integrazione.
4. *Caricamento*. L'ultima fase da eseguire è il caricamento dei dati nel DW, che può avvenire secondo due modalità:

- *Refresh*. I dati del DW vengono riscritti integralmente, sostituendo quelli precedenti.
- *Update*. I soli cambiamenti occorsi nei dati sorgente vengono aggiunti nel DW, tipicamente senza distruggere o alterare i dati esistenti.

2.5. Il modello multidimensionale

Il concetto di multidimensionalità è di interesse centrale nel tema dei data warehouse. E' il concetto di dimensione che ha origine alla metafora del cubo per la rappresentazione dei dati multidimensionali. Secondo questa metafora, gli eventi corrispondono a celle di un cubo i cui spigoli rappresentano le dimensioni di analisi. Ogni cella del cubo contiene un valore per ciascuna misura.

Dunque, un cubo multidimensionale è incentrato su un fatto di interesse per il processo decisionale. Esso rappresenta un insieme di eventi, descritti quantitativamente da misure numeriche. Ogni asse del cubo rappresenta una possibile dimensione di analisi; ciascuna dimensione può essere vista a più livelli di dettaglio individuati da attributi strutturati in gerarchie.

2.6. Fasi della progettazione del DW

Le fasi principali per la progettazione di un data mart possono essere riassunte nel seguente modo[3]:

1. *Analisi e riconciliazione dei dati*. La prima fase di progettazione prevede la documentazione dello schema dei dati operazionali da dove verrà alimentato il data mart. Quindi occorre analizzare e comprendere gli schemi delle sorgenti disponibili, determinare eventuali correlazioni tra le sorgenti e infine valutare la qualità dei dati.

2. *Analisi dei requisiti*. In questa fase il progettista raccoglie, filtra e analizza i requisiti degli utenti finali, con il fine di delineare quali informazioni sono di interesse strategico.
3. *Progettazione concettuale*. Questa fase prevede l'uso dei requisiti utente ottenuti durante la fase precedente per disegnare uno schema concettuale per il data mart.
4. *Progettazione logica*. Essa include l'insieme di passi che, a partire dallo schema concettuale, permettono di determinare lo schema logico del data mart.
5. *Progettazione dell'alimentazione*. Nell'ultima fase vengono prese tutte le decisioni che riguardano il progetto di alimentazione del data mart.

2.6.1. Analisi e riconciliazione delle fonti dati

Come già accennato in precedenza, i dati contenuti nel DW sono ricavati da un insieme di sorgenti che si possono differenziare sia per la tecnologia che le gestisce (ad esempio i DBMS), sia per il modello attraverso il quale rappresentano la realtà aziendale. Le varie sorgenti operazionali possono essere fortemente o completamente indipendenti, dunque è fondamentale per il progettista acquisire una conoscenza quanto più approfondita delle sorgenti dati.
Uno dei principi fondamentali del data warehousing è il concetto di dato integrato che permette di derivare informazioni consistenti e prive di errori. Il raggiungimento di questo risultato necessita di un *processo di riconciliazione* che comporta integrazione, pulizia e trasformazione dei dati.

2.6.2. Analisi dei requisiti

La fase di analisi dei requisiti ha l'obiettivo di raccogliere le esigenze di utilizzo del data mart espresse dai suoi utenti finali. Essa ha un'importanza strategica all'interno della progettazione del data mart, poiché ha un ruolo sostanziale nel determinare lo schema concettuale dei dati, il progetto dell'alimentazione, l'architettura del sistema e l'evoluzione di esso. Esistono diverse tecniche per l'analisi dei requisiti utente, tra queste si distingue l'approccio basato sull'utilizzo del formalismo di Tropos.
Tropos è una metodologia di sviluppo software orientata agli agenti e presenta due caratteristiche originali. Primo, la metodologia è basata sui concetti di agente e obiettivo, che

sono di supporto a tutte le fasi di sviluppo. Secondo, viene assegnato un ruolo cruciale all'analisi dei requisiti preliminari.

Tropos gestisce quattro fasi di sviluppo software:

1. Analisi dei requisiti preliminari
2. Analisi dei requisiti progettuali
3. Progetto architetturali
4. Progetto esecutivo

La metodologia Tropos è stata applicata con successo in differenti aree applicative, la notazione di questa metodologia può essere riassunta di seguito:

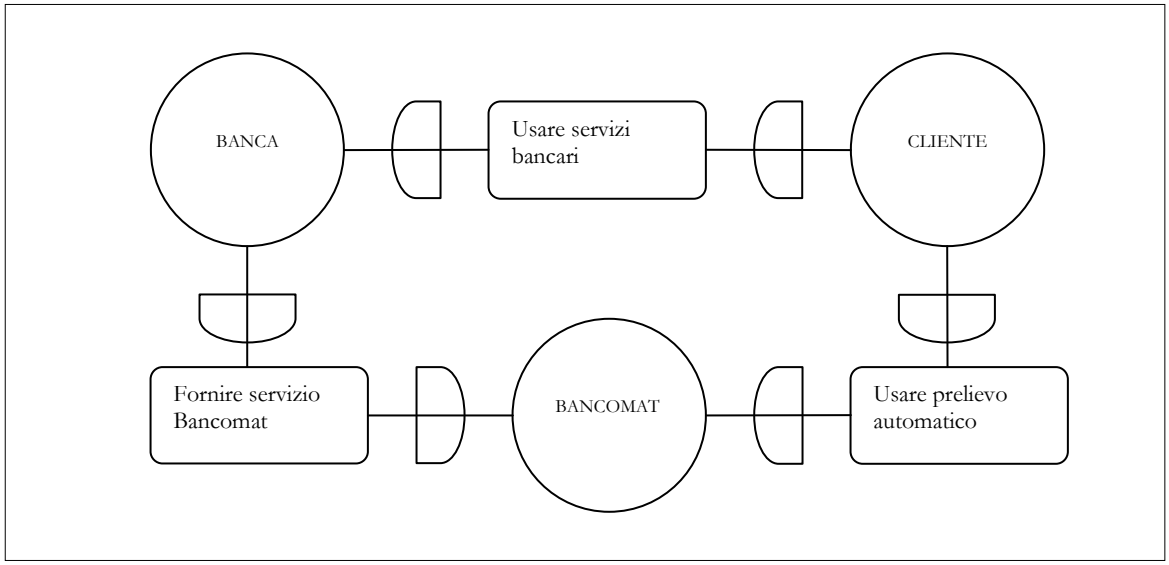
- *Attori*. Un attore rappresenta uno *stakeholder* aziendale. Un attore può modellare un agente (fisico o software), un ruolo (caratterizzazione astratta di un comportamento assunto da uno o più agenti in uno specifico contesto) o una posizione (insiemi di ruoli giocati generalmente da un singolo agente).
- *Dipendenze strategiche*. Una dipendenza strategica rappresenta un accordo tra due attori, uno dei quali dipende dall'altro per il rispetto dell'accordo, esso può consistere in un obiettivo da raggiungere.
- *Diagramma degli attori*. Il diagramma degli attori è un grafo di attori legati da dipendenze strategiche.
- *Diagramma di ragionamento*. Viene utilizzato per rappresentare i fondamenti logici e razionali che regolano le relazioni di ciascun attori con gli altri.

Nel contesto specifico dei data warehouse, è necessario introdurre alcuni nuovi concetti:

- *Fatti*. Un fatto modella un insieme di eventi che si verificano quando un obiettivo viene raggiunto.
- *Attributi*. Sono dei campi la cui valorizzazione si accompagna alla registrazione di un fatto ad opera di un obiettivo.
- *Dimensioni*. Una dimensione è una proprietà di un fatto che ne descrive una possibile coordinata di analisi.
- *Misure*. Una misura è una proprietà numerica di un fatto che ne descrive un aspetto quantitativo.

La metodologia Tropos propone un tipo di analisi centrata sugli obiettivi dei decisori, ossia degli attori protagonisti del processo decisionale. Dapprima vengono identificati tutti i decisori; poi, per ciascuno di essi, vengono effettuati quattro passi d'analisi:

1. *Analisi degli obiettivi*. L'analisi degli obiettivi inizia con uno studio del diagramma degli attori per i decisori. Gli obiettivi abbinati a ciascun decisore sono quindi scomposti e analizzati nel dettaglio, al fine di generare un insieme di diagrammi di ragionamento.
2. *Analisi dei fatti*. I diagrammi di ragionamento vengono estesi identificando fatti e associandoli agli obiettivi dei decisori.
3. *Analisi dimensionale*. In questa fase ciascun fatto viene associato alle dimensioni che i decisori ritengono necessarie per soddisfare i singoli obiettivi decisionali
4. *Analisi delle misure*. Infine l'analista associa un insieme di misure a ciascun fatto precedentemente identificato.



Esempio di diagramma degli attori; gli attori e i loro obiettivi sono rappresentati, rispettivamente, da cerchi e ovali

2.6.3. Progettazione concettuale

Per la progettazione di un data mart esistono diversi approcci, distinti sulla base della rilevanza assegnata alle fasi di analisi del database operazionale e di analisi dei requisiti utente. Un approccio misto prevede un ruolo attivo dei requisiti utente nel limitare la complessità dell'analisi delle sorgenti. Il quadro metodologico prevede quindi l'utilizzo del formalismo Tropos per l'analisi dei requisiti. La procedura si articola in tre fasi:

1. *Mappatura dei requisiti*. Obiettivo di questa fase è stabilire una corrispondenza tra i fatti, le dimensioni e le misure individuate precedentemente e le relazioni e attributi presenti nello schema operazionale.

2. *Costruzione dello schema di fatto*. Ciascuna dimensione e misura mappati con successo da un diagramma di ragionamento esteso allo schema operazionale vengono inclusi nello schema di fatto. Successivamente vengono aggiunti anche gli attributi, eliminando quelli inutilizzati.
3. *Raffinamento*. Lo schema di fatto viene affinato per renderlo più aderente ai bisogni dell'utente

2.6.4. Progettazione logica

La fase di progettazione logica include gli insiemi di passi che, a partite dallo schema concettuale, permettono di determinare lo schema logico del data mart. L'obiettivo primario è la massimizzazione della velocità di reperimento dati, autorizzando il ripetuto utilizzo di dati ridondanti e denormalizzati. Essa può essere riassunta in tre fasi:

- *Traduzione degli schemi di fatto in schemi logici*. Nel caso in cui lo schema logico sia a stella allora la fact table contiene tutte le misure e gli attributi descrittivi direttamente collegati al fatto, e per ogni gerarchia viene creata una dimension table che ne contiene tutti gli attributi.
- *Materializzazione delle viste*. Per materializzazione delle viste si intende il processo di selezione di un insieme di viste che esaltano gli obiettivi di un progetto.
- *Frammentazione delle viste*. Si intende la suddivisione di una tabella in più tabelle dette frammenti al fine di aumentare le prestazioni del sistema

2.6.5. Progettazione dell'alimentazione

Durante la fase di progettazione dell'alimentazione vengono definite le procedure necessarie a caricare all'interno del data mart i dati provenienti dalle sorgenti operazionali. In presenza del livello riconciliato il processo di alimentazione risulta suddiviso in due fasi:

- Dalle sorgenti operazionali al livello riconciliato.
- Dal livello riconciliato al livello dei data mart.

Come vedremo in seguito, nella costruzione di questo progetto è stata decisa una soluzione architetturali a tre livelli poiché la presenza di uno stadio intermedio ha facilitato il compito del gruppo di progettazione.
L'alimentazione dello schema riconciliato avviene nel seguente modo:

- *Estrazione*: indica le operazioni che permettono di acquisire i dati delle sorgenti. Esistono diverse modalità di estrazione: assistita da un'applicazione, basata sui trigger oppure basata sui timestamp.
- *Trasformazione*: indica le operazioni che conformano i dati dalle sorgenti allo schema riconciliato. Alcuni dati possono essere convertiti oppure concatenati tra loro.
- *Caricamento*: indica le operazioni necessarie a inserire i dati trasformati nel database riconciliato aggiornando eventualmente quelli già presenti.

2.7. Gli ambiti applicativi del data warehouse

Nelle banche e in generale nelle istituzioni finanziarie gli ambiti di utilizzo sono molteplici, poiché tutte le aree gestionali di tali organizzazioni sono caratterizzate da volumi considerevoli di dati su cui devono essere prese decisioni strategiche. Poiché il data warehouse può avere un valore strategico, all'interno di tali tipi di organizzazioni è fondamentale per il management definire una strategia per il data warehouse. Essa è essenzialmente un percorso evolutivo che porta l'azienda da applicazioni DW non 'mission-critical' a una situazione in cui il data warehouse è una componente fondamentale del sistema informativo aziendale.

La strategia di data warehousing di un'azienda può essere classificata in base a due dimensioni fondamentali:

- utilizzo del DW esistente: livello di maturità degli utenti e delle funzioni di supporto del DW nell'utilizzo dell'esistente;
- utilizzo prospettico del DW: obiettivi strategici di utilizzo del DW come piattaforma di *decision support*.