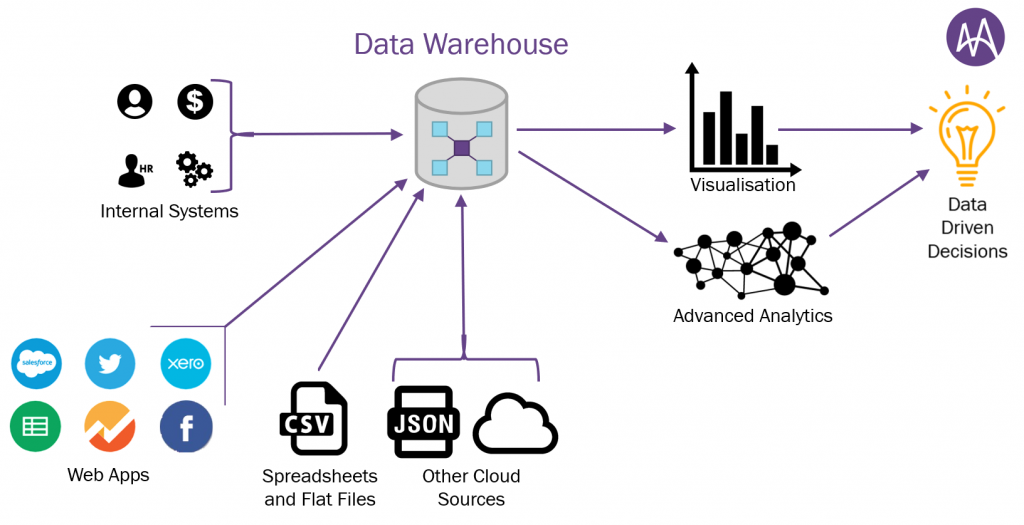
architettura concettuale per l’analisi dei Big Data



*Figura 3: Architettura concettuale per l’analisi dei Big Data (Raghupathi W., 2014)*

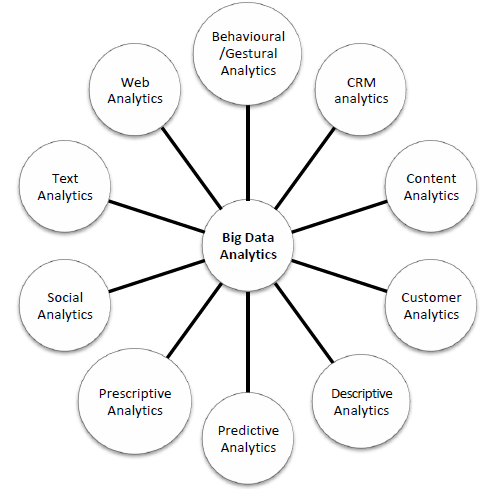
Innanzitutto i Big Data provengono da divere fonti, sia interne che esterne, spesso sono in formati differenti e risiedono in posizioni multiple in numerosi sistemi legacy e altre applicazioni. I dati possono essere sia strutturati (dati conservati in Database relazionali, organizzati secondo schemi e tabelle rigide), sia non strutturati (dati conservati senza alcuno schema come forme libere di testo tra cui articoli e parti di e-mail, audio senza tag, immagini e video) sia semi-strutturati (dati che presentano caratteristiche sia di quelli strutturati che di quelli non strutturati; un esempio è rappresentato dai file compilati con sintassi XML per i quali non ci sono ci sono limiti strutturali all’inserimento dei dati, ma le informazioni vengono organizzate secondo logiche strutturate e interoperabili). Dopo che i dati sono stati uniti, questi hanno bisogno di essere processati o trasformati, essendo in uno stato grezzo. Ci sono diverse opzioni a disposizione:

Service-oriented architecture combinata con web services (middleware): i dati rimangono grezzi e i services sono utilizzati per chiamare, recuperare e processare i dati;

Data warehousing: dati provenienti da svariate fonti vengono aggregati e preparati per essere processati, anche se non sono disponibili in real-time;

Extract, transform and load (ETL): dati che derivano da diversi fonti vengono puliti e preparati per lo step successivo.

Il passo successivo consiste nella scelta della piattaforma e della tecnologia da utilizzare, tra quelle elencate nella tabella. L’ultima fase invece è relativa alle applicazioni di Big DataAnalytics che includono queries, reports, OLAP e Data mining e alla visualizzazione, compresa in tutte queste applicazioni (Raghupathi W., 2014).

Un ruolo centrale in quest’ambito viene svolto dai Big Data Analytics, tecnologie di Business Intelligence &Analytics basate sulle tecniche descritte in tabella. Riportiamo nel seguente grafico quelli più importanti, che ritroveremo più volte in questa ricerca.  **

* *Behavioural/Gestural Analytics*: analisi automatizzata delle attività umane catturate da video che tracciano i movimenti e i gesti per individuare e comprendere comportamenti e intenzioni;
* *Content Analytics*: insieme di tecnologie che processano i contenuti digitali e i comportamenti degli utenti nelle conversazioni con altre persone, nelle discussioni sui social network o relativamente al livello di consumo ed engagement di documenti e nuovi siti, per rispondere a determinate domande;
* *CRM Analytics*: soluzioni che raccolgono, organizzano e sintetizzano i dati dei clienti per aiutare le organizzazioni a risolvere i problemi di business riguardanti i consumatori
* attraverso tool, dashboard, portali e altri metodi negli ambiti di Marketing, Sales e Customer Service;
* *Customer Analytics*: tecnologie che sfruttano i dati per capire la composizione, i bisogni e la soddisfazione dei consumatori, per poi segmentarli in gruppi sulla base dei comportamenti adottati, implementare azioni di Marketing personalizzate e determinare trend generali;
* *Descriptive Analytics*: analisi di dati e contenuti pe rispondere alla domanda “Cosa è successo?” o “Cosa sta succedendo?” attraverso strumenti tradizionali di BI e visualizzazione;
* *Predictive Analytics*: Analytics avanzati che implementano tecniche quali la regressione, i modelli predittivi e la statistica per analizzare i dati e i contenuti e rispondere alle domande “Cosa succederà” o “Cosa accadrà molto probabilmente?”;
* Prescriptive Analytics: altra forma di Analytics avanzati che esamina i dati e i contenuti per rispondere alle domande “cosa dovrebbe essere fatto?” o “cosa dobbiamo fare per far sì che succeda una determinata cosa?” e per far questo utilizza tecniche quali l’analisi di grafici, la simulazione, le reti neurali e la machine lerarning;
* Social Analytics: tools che estraggono, analizzano e sintetizzano automaticamente i contenuti generati dagli utenti online. Questa tecnologia verrà descritto in modo approfondito nel successivo capitolo;
* *Text Analytics*: processo di estrazione delle informazioni dai testi, utilizzato per diversi scopi, tra cui il *riepilogo*, ovvero il tentativo di trovare i contenuti chiave in un grande insieme di informazioni, la *sentiment analysis*, già spiegate o per determinare cosa ha guidato un determinato commento di una persona e quindi per un fine esplicativo;
* *Web Analytics*: applicazioni analitiche utilizzate per capire e migliorare l’esperienza online del consumatore, l’acquisizione di utenti e l’ottimizzazione del digital Marketinge delle campagne pubblicitarie. Questi offrono reporting, segmentazione, gestione delle campagne e integrazione con altre fonti dati e processi.

**L’IMPATTO DEI BIG DATA NEL SETTORE RETAIL**

I Big Data offrono enormi opportunità anche al mondo Retail che ha disposizione non solo dati come le transazioni e le operazioni dei clienti, ma anche dati dagli RFID e informazioni sul comportamento online e sul sentiment dei consumatori. Le aziende che utilizzano tecniche e tecnologie in grado di sfruttare questi dati, riescono a migliorare l’efficacia delle loro azioni di Marketing e di merchandising, a ridurre i costi delle operazioni e della supply chain e quindi a migliorare la loro profittabilità e a ottenere un vantaggio competitivo rispetto agli atri concorrenti (Manyika J., 2011).

I Retailer in particolare possono utilizzare due diverse tipologie di Analytics (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015):

*Performance Management & Basic Analytics*: strumenti di Descriptive Analytics che consentono di accedere ai dati secondo viste logiche flessibili e dinamiche e di visualizzare sinteticamente e graficamente i principali indicatori di prestazione;

*Advanced Analytics*: strumenti avanzati che permettono di svolgere un’analisi attiva dei dati sfruttando metodologie di prescriptive&predictive analysis, determinando trend e prevedendo il valore futuro di variabili numeriche e categoriche.

In quest’analisi ci concentriamo sia sui processi di back-end, ovvero i processi di interazione Retail-fornitore o processi interni del Retailer sia su quelli di front-end, ovvero quelli di interazione Retailer-consumatore.

Processi di back-end

*Logistica*. Oggi i fornitori logistici gestiscono un enorme flusso di beni creando nel contempo un set rilevante di dati (origine e destinazione dei viaggi, dimensione, peso, contenuto del trasporto, posizione)considerati i milioni di viaggi intrapresi ogni giorno.

Le tecniche di Big Data possono essere innanzitutto utilizzate per ottimizzare i costi del viaggio per la consegna dei prodotti. Focalizziamoci su due diverse modalità.

*1)Ottimizzazione del percorso in real time*. Quando il veicolo viene caricato per partire, il percorso di consegna ottimale viene pianificato utilizzando i dati di spedizione rilevati dai sensori sui prodotti in viaggio. Durante il trasporto sistemi dinamici suggeriscono cambiamenti del percorso a seconda delle condizioni del traffico, dei fattori geografici e dello stato del ricevente. Questo approccio, che si basa sull’utilizzo di dati reali, permette di tagliare i costi e di ridurre l’emissione di CO2 diminuendo per esempio le distanze da percorrere. 2)*Pick-up e consegna basate sulle persone*. Pendolari, taxisti o studenti potrebbero essere pagati per occuparsi della consegna nell’ultima parte del percorso se questa coincide con il viaggio che devono effettuare. Questo approccio, che porta ad una notevole riduzione dei costi, richiede l’utilizzo di tecniche di Big Data: flussi di dati real-time vengono tracciati al fine di assegnare la spedizione alle persone disponibili, basandosi sulle loro posizioni e destinazioni. Attraverso dispositivi mobile, possibili trasportatori pubblicano la loro posizione e accettano l’assegnazione della consegna.

Analitycs avanzati possono essere utilizzati anche per prevedere la domanda al fine ottimizzare la capacità di trasporto e la quantità di personale necessario in ciascuna zona. Per l’allocazione delle risorse vengono utilizzate informazioni real- time sulle spedizioni (item che sono appena entrati nella rete distributiva, che sono in transito o che sono in magazzino), informazioni dai clienti (apertura di nuove industrie, fallimenti inaspettati) e anche informazioni di eventi locali (epidemie regionali e disastri naturali).

Le soluzioni Big Data permettono inoltre il recupero di informazioni utili per il rilevamento dei rischi di supply chain: dati sugli sviluppi politici locali, sull’economia, sulla salute, sulla natura provenienti da diverse fonti quali siti, social media, blog vengono aggregati e analizzati attraverso sematic Analytics e altre tecniche. Queste soluzioni individuano dei pattern tra le diverse informazioni e quando si verifica una condizione critica per la supply chain del cliente, questo viene avvisato e gli viene inviato un report contenente informazioni quali la probabilità e l’impatto del rischio e contromisure per mitigarlo (Jeske M., 2013).

*Gestione del magazzino.* L’integrazione di strumenti di ottimizzazione del magazzino e dei sistemi ERP consente di fare la migliore analisi possibile dei dati a disposizione quali dati delle vendite, degli acquisti, finanziari, di fornitura e di produzione. Sviluppando un algoritmo, il sistema di ottimizzazione può creare un’interfaccia grafica in grado di illustrare una sintesi di tutti i dati, che permette di identificare i cambiamenti stagionali dei prodotti richiesti, quando si verificheranno gli stock-out, le vendite perse e gli ordini in eccesso, ecc… Questa analisi viene quindi utilizzata da uno tool di ottimizzazione del magazzino per prevedere la domanda in modo accurato e stabilire quindi il livello ottimale del magazzino, che soddisfi le richieste dei clienti ed eviti sia lo stock-out che la sovrabbondanza di item. Questo strumento dà poi dei suggerimenti per la successiva pianificazione delle scorte e definisce la soglia del riordino. Data l’integrazione tra questo tool e i sistemi ERP, tutti questi dati vengono comunicati sia all’interno che agli stakeholder all’esterno dell’organizzazione (Sage, 2013)Alcune aziende utilizzano anche sistemi bar-code collegati ai processi di rifornimento automatico per ridurre l’incidenza dello stock-out (Manyika J., 2011).

*Pianificazione dei turni*. È possibile utilizzare un algoritmo predittivo che prenda in considerazione un ampio range di parametri individuali e locali: dati dei ricavi storici, orari di apertura dei negozi, orari di arrivo dei prodotti dai centri distributivi ma anche i giorni del mercato, i giorni di vacanza delle località vicine e i dati delle previsioni meteorologiche che influenzano il comportamento dei clienti. L’algoritmo dà come soluzione le vendite giornaliere previste, a partire dalle quali vengono pianificati i turni in modo ottimale, evitando surplus e carenze del personale, che impattano negativamente sulle performance finanziarie del negozio e sulla soddisfazione dei clienti e dei dipendenti (Jeske M., 2013).

Processi di front-end

Le nuove abitudini dei consumatori, che comportano la creazione di uno tsunami di dati, giustificano l’utilizzo delle tecnologie di Big Data da parte dei Retailer (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015):

2 consumatori su 3 si informano online prima di acquistare un prodotto, comprano in un negozio, ma hanno preso la decisione prima su canali digitali (Fonte: Net Retail –Netcomm, Campione: 3.055 individui);

il 54% dei consumatori preferisce percorsi di acquisto che contemplino almeno un’interazione con i canali digitali online e mobile (Fonte: Total Retail PwCCampione:1.000 consumatori che hanno acquistato almeno 1 volta online);

degli oltre 30 milioni di Internet user, il 31% utilizza per navigare un solo device, il 69% utilizza due o più device per navigare tra cui PC, Smartphone e Tablet (Fonte: SurveyCAPI 2013 –Doxa, Campione: totale Internet userdaily);

l’89% degli utenti Smartphone utilizza il device all’interno del negozio. Di questi circa il 40% dichiara di farlo sempre o spesso (Fonte: SurveyCAWI 2014 –Doxa, Campione: 1.503 utenti smartphone);

tra gli utenti che usano lo Smartphone in negozio, il 42% degli utenti confronta i prezzi e il 30% degli utenti invia foto dei prodotti da acquistare ad amici (Fonte: SurveyCAWI 2014 –Doxa, Campione: 1.341 utenti che usano lo smartphonein punto vendita);

molti consumatori utilizzano applicazioni quali RedLaser che permette loro di scannerizzare il bar code su un item in un negozio con il loro smartphone e ottenere immediatamente il prezzo e il prodotto dei concorrenti (Manyika J., 2011).

I Retailer hanno a disposizione una serie di leve per sfruttare i Big Data e trarre quindi vantaggio nei processi di interazione con il cliente. In questo paragrafo ci limitiamo ad elencarle, in quanto saranno oggetto di approfondimento del secondo capitolo. Nei processi di Marketing le tra le leve principali citiamo il Cross-Selling, il location based Marketing, l’in-store behavior analysis, la customer micro-segmentation, la sentiment analysis e la multichannel consumer experience; mentre nelle attività di merchandising l’ottimizzazione dell’assortimento, l’ottimizzazione del prezzo e l’ottimizzazione del posizionamento e del design (Manyika J., 2011). L’utilizzo di queste tecniche impatta positivamente sulla customer experience del consumatore, dalla fase di prevendita a quelle di acquisto e di pagamento fino alla fase di post-vendita: il cliente riesce a

trovare il prodotto soddisfa al meglio i suoi bisogni e spende meno per trovare i prodotti al prezzo più vantaggioso. Anche su questo argomento ci focalizzeremo nel capitolo successivo.

Il DataWarehouse

Generalità

La tecnologia delle basi di dati è finalizzata alla gestione efficiente e affidabile di dati “in linea” (On Line Transaction Processing, OLTP); questi dati però potrebbero rivelarsi utili non solo per la gestione, ma anche per la pianificazione e il supporto alle decisioni. In pratica, i dati del presente e del passato possono consentire un’attività di analisi essenziale nella programmazione delle attività future di un’impresa.

Con l’inizio degli anni ‘90, parallelamente allo sviluppo delle reti e dei prodotti per la distribuzione dei dati, si sono imposte nuove architetture, caratterizzate dalla separazione degli ambienti: a fianco dei sistemi per OLTP si sono sviluppati sistemi dedicati all’elaborazione e analisi dei dati, cioè On Line Analytical Processing (OLAP). Questa sigla mette in evidenza che l’analisi dei dati avviene “in linea”, tramite strumenti interattivi. L’elemento principale dell’architettura OLAP, è il **Data Warehouse** (come il DBMS lo è per OLTP).

Alla base del Data Warehouse (da intendersi in prima istanza come un Data Base di grandi dimensioni) c’è il modello dimensionale che serve per rappresentare l’oggetto d’osservazione.

Se ad esempio si è interessati ad avere informazioni di sintesi (reports) sui visitatori del museo, possiamo rappresentare i visitatori a seconda della sede e del tempo. Per riuscire a visualizzare questa struttura logica si disegna un cubo che rappresenta il Fatto Visitatori del museo, qualificato dalle dimensioni Sede, Biglietti e Tempo: si veda la figura 3. In particolare si osservi l’aggregazione dei dati sulla dimensione tempo tale da avere nel primo cubo una rappresentazione per mese, nel secondo per trimestre e infine nell’ultimo, per anno.

Il cubo può avere più di tre dimensioni (da cui il nome di cubo multidimensionale), ma il limite al numero delle dimensioni utilizzabili è dovuto alla comprensibilità della struttura.

Definizione di data warehouse

Non è facile definire univocamente un DW, dato il proliferare di interpretazioni su cosa realmente s’intenda: un processo, una tecnica, oppure un sistema. In questo lavoro si farà riferimento principalmente alla definizione data da William Inmon: “Un insieme di dati orientati al soggetto, integrato, variante rispetto al tempo, non volatile costruito per supportare il processo decisionale”. Questa definizione focalizza l’attenzione su alcuni aspetti peculiari del DW:

Un DW è orientato al soggetto (gli utenti (fruitori) del museo, i servizi offerti, …), cioè esclude tutti i dati che non saranno utilizzati nel processo di pianificazione delle strategie, raggruppandoli per aree o temi d’interesse e finalizzandoli a chi li utilizza più che a chi li genera;

Un DW è integrato, ossia consistente rispetto ad un sistema concettuale globale dei dati e rispetto alle unità di misura e alle strutture di codifica. In pratica occorre che il significato di ciascun dato sia chiaro, concordato a livello aziendale e definito con precisione sia in termini applicativi che informatici;

I dati di un DW hanno un orizzonte temporale molto ampio (5, 10 anni o più), per implementare il quale è necessario disporre di informazioni che contengano la dimensione temporale;

Un DW è non volatile: il dato viene caricato e letto off line; una volta memorizzato correttamente può essere rilevato, ma non modificato dall’utente.

Una definizione più operativa di Data Warehouse è quella data da Ralph Kimball:

“A data warehouse is a copy of transaction data specifically structured for querying and reporting[1](#_bookmark0)”.

Applicazioni

La tecnologia del Data Warehousing, nata per analizzare grandi quantità di dati e trasformarli in informazioni, può essere applicata anche ai musei: infatti, il personale direttivo deve continuamente prendere decisioni su come utilizzare le risorse disponibili per raggiungere gli obiettivi prefissati e come valutarne i risultati.

Queste operazioni sono quelle che consentono ad un museo di coniugare un’efficace impostazione strategica con un uso efficiente delle risorse a disposizione. La pianificazione di strategie può essere facilitata proprio con l’introduzione di un sistema di Data Warehouse (nel seguito denominata DW).

In altre parole, poiché l’obiettivo di un progetto DW è quello di consentire agli utenti la gestione più efficace ed efficiente l’attività dell’impresa, il primo passo nella sua costruzione è di individuare il tipo di dati che dovranno entrare a farvi parte e successivamente le tecniche di analisi più appropriate ai fini del supporto decisionale

Una volta inseriti i dati, il DW permette di ottenere attraverso un’analisi incrociata l’ottimizzazione del loro utilizzo a vantaggio dell’impresa e di ridurre al minimo i margini di errore, tenendo sempre presente che il vantaggio principale ottenibile è sempre legato dall’abilità dell’utente di interrogare i dati in modo innovativo.

La trasformazione dei vari dati in informazioni congruenti può potenziare e/o ottimizzare le attività del museo, per es. attraverso:

il rapporto con gli sponsor (del passato, del presente e possibili)

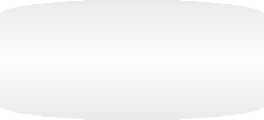
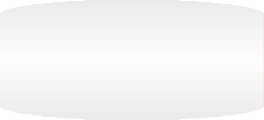
l’offerta delle esposizioni

la migliore gestione contabile - collegata ad a), b) e d)

la miglior gestione del personale - collegata ad a), b) e c).

Caratteristiche di base

I dati contenuti nel DW dovranno poter essere sintetizzati in modo da fornire una visione globale caratterizzata da informazioni aggregate, le quali diventano dettagliate, al punto da risultare atomiche ai livelli più bassi, in modo da spiegare aspetti particolarmente critici durante l’analisi (si veda figura 4).



Data Mart

**Storico**

DataBase Operazionale

**Storico**

DW

Informazionale

**Attuale**

**Dati Dettagliati**

**Sintesi spinta**

**Sintesi lieve**

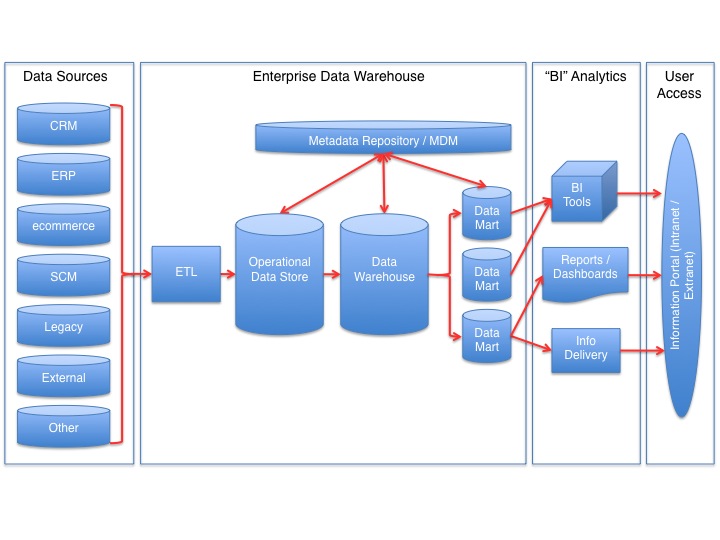
**Figura 4**

In un’ottica tecnica la disponibilità di dati dettagliati offre il vantaggio di rendere necessaria una sola estrazione dei dati dall’ambiente operazionale, che quindi non vedrebbe penalizzate le proprie prestazioni, poiché la successiva attività di sintesi sui database del **Warehouse** (e l’eventuale distribuzione verso i Data Mart dipartimentali) impegnerebbe solo i sistemi informazionali. D’altra parte è evidente che alcuni dati di sintesi (specie quelli di più largo uso) se preaggregati consentono di elevare le prestazioni del DW.

La storicizzazione dei dati è un altro dei concetti cardine del DW. E’ proprio la necessità di storicizzare una rilevante mole di dati che spinge ad operare una sintesi di questi su più livelli nel sistema informazionale, senza perdere completamente traccia del dato in dettaglio, pur rinunciando alla sua attualità. Ciò conduce ad un’architettura dei dati organizzata su più piani.

Nello strato più alto troviamo informazioni ad un livello di sintesi molto elevato (tanto da giustificare la loro allocazione su DW di tipo tematico, i **Data Mart**), che vengono dettagliate man mano che si scende nella struttura fino ad essere storicizzati al livello più basso.

L’utente dovrebbe essere in grado di accedere ed operare sui dati desiderati anche senza sapere dove essi risiedono, in quale forma siano stati memorizzati e quali tool provvedano alla loro consegna.

****

Evidentemente, a causa delle dimensioni in gioco, non tutte le informazioni possono essere tenute on line e tipicamente, i dati storici dettagliati saranno archiviati off line: quindi le risposte alle interrogazioni che riguardano i livelli di sintesi più alti saranno senz’altro interattive, mentre analisi che si spingono fino al reperimento di dati storici archiviati determineranno risposte differite; ai livelli intermedi la scelta sarà legata alle caratteristiche dei sistemi e dei DBMS che ospitano e gestiscono il DW.

In ogni caso nel rispetto della filosofia stessa del Data Warehousing, la demarcazione tra dati on line e off line non deve ostacolare la possibilità dell’utente di navigare tra i percorsi indicati: in altri termini il sistema di Data Warehousing deve assicurare attraverso le funzioni di accesso ai dati la possibilità, a partire da un percorso intrapreso, di approfondire il grado di indagine sulla dimensione, in maniera diretta o differita secondo la disponibilità del dato dettagliato, o di regredire ad un più elevato livello di sintesi.

Per completare il quadro progettuale dell’architettura dati che caratterizza la soluzione di Data Warehousing, occorre individuare una tecnica che consenta di definire i contenuti del sistema, permettendo di includere all’interno del DW l’intero orizzonte dell’attività che si sta considerando.

La necessità di accedere ai dati storici è uno degli elementi caratteristici di un progetto di Data Warehousing poiché questo tipo di dati gioca un ruolo primario nell’alimentare tutti quei sistemi usati per analisi di tendenza che sono determinanti nel supporto decisionale.

Architettura dei Data Warehouse

Vi si individuano tre componenti principali:

il back-end, che gestisce il processo di estrazione, filtraggio, integrazione e trasformazione dei dati prelevati da vari ambienti operativi; esso attiva inoltre il processo di caricamento e di indicizzazione dell’informazione all’interno della base dati.

Il sistema di gestione dei dati (DBMS) su cui risiedono i dati e i metadati del DW. Per metadati si intende Informazioni che descrivono il significato e la struttura dei dati aziendali, come sono creati, come vi si accede e come possono essere usati. In questa definizione troviamo diverse tipologie di metadati: informazioni sui dati di origine, su come accedervi e trasformarli, sulla struttura del DW, su come utilizzare i dati da parte dell’utente.

Il front-end, l’interfaccia grafica che permette all’utente finale di interagire con i dati.

Fermi restando i concetti finora espressi, la realizzazione di un sistema di Data Warehousing comporta diverse scelte progettuali. Qui verrà presentata l’architettura DW più completa ed esaminati i relativi vantaggi e svantaggi.

Presentiamo qui di seguito un’estensione della architettura generale di un DW, orientata al raggiungimento di un alto grado di flessibilità a costo di una ridondanza dei dati memorizzati nei vari livelli.

Il DW è composto da database OLTP, datamart, OLAP-Server e Client.

I dati contenuti nei database subiscono delle operazioni di estrazione trasposto e trasformazione per essere immessi nei datamart e nel DW. In ogni architettura vi possono essere diversi datamart ma un solo DW. I data mart condividono con il DW la stessa struttura, ma sono limitati nel loro scopo a fornire informazioni utili a una singola divisione aziendale (nel caso del museo i data mart sono finalizzati ai diversi reparti museali).

L’OLAP-Server permette di ottimizzare l’utilizzo delle risorse hardware in due modi distinti: da un lato trattando le operazioni a più alta richiesta di potenza elaborativa; dall’altro riduce i dati che passano dal server ai client, i computer utilizzati dal personale amministrativo.

Progettazione del Data Warehouse

Ci sono due aspetti principali nella costruzione di un DW: la progettazione dell’interfaccia a partire dai sistemi operazionali e il progetto del DW in sé. In un certo senso, parlare di progettazione non descrive in modo preciso quello che avviene nel processo di realizzazione che procede in maniera euristica.

Infatti dapprima viene popolata una porzione del DW per essere utilizzata e verificata dall’analista. In seguito, sulla base delle indicazioni di ritorno da parte dell’utilizzatore finale, i dati sono modificati ed altri vengono aggiunti.

Il ciclo di retroazione continua lungo tutta la vita del DW. Quindi il DW non può essere costruito con il consueto modello di sviluppo guidato dai requisiti, tuttavia non bisogna commettere l’errore di pensare che non serva del tutto definire dei requisiti di partenza.

Strategie di progettazione

La scelta della strategia dovrà tener conto di una serie di fattori legati indissolubilmente alla realtà aziendale, in questo caso al tipo di museo.

*La strategia Top-Down* è stata storicamente la più utilizzata nell’implementazione di grandi DW: il “dipartimento Sistemi Informativi” è responsabile della catalogazione di tutte le informazioni disponibili e di quelle richieste dai vari dipartimenti.

Viene progettato un Modello (di business del Museo) della struttura organizzativa del Museo, al quale il DW corrisponderà e al quale verrà fatto riferimento per l’interpretazione delle informazioni disponibili. Questo fa in modo che qualunque decisione venga presa partendo da una base di conoscenza comune, condivisa e più estesa e significativa di quella normalmente in possesso dei singoli “dipartimenti”.

I vantaggi principali della strategia top-down risiedono nel fatto che i dati vengono estratti, trasformati e caricati una sola volta, indipendentemente dal tipo di utilizzo e dal numero di dipartimenti che ne fanno uso, e nel fatto che l’architettura multi - tier permette di utilizzare il DW come fonti di dati comuni per più data mart dipartimentali.

Tuttavia gli svantaggi non sono da sottovalutare: il progetto e la costruzione di una simile struttura hanno, infatti, bisogno di notevoli sforzi in termini di modellazione, implementazione, istruzione e assistenza agli utilizzatori finali.

La strategia bottom up *si articola in due momenti: l’implementazione di una serie di data mart indipendenti, e successivamente la fusione dei singoli data mart in una visione globale, che corrisponde al DW.* L’idea alla base di questa strategia è di avvicinare la costruzione del DW ai suoi fruitori finali.

Gli aspetti positivi di questa strategia, rispetto alla top down, si possono riassumere come segue:

relativa facilità di implementazione

tempo ridotti per ottenere i primi risultati parziali

ritorno economico maggiormente controllabile

maggiore semplicità per il fruitore finale che ha a disposizione tutte le informazioni di cui ha bisogno;

minori requisiti iniziali in termini di hardware.

Naturalmente anche questa metodologia porta con sé alcuni punti deboli:

Aumenta la ridondanza dei dati, in quanto più data mart indipendenti conservano più copie degli stessi dati.

Aumenta il carico per il sistema OLTP (i data base tradizionali) per il moltiplicarsi di estrazioni degli stessi dati.

Aumenta il pericolo di una proliferazione incontrollata di data mart che può portare a problemi di integrazione nel futuro DW.

Schema del DW

Anche nel caso di un Museo, il successo di un sistema di DW sta nel determinare un disegno globale dell’impresa, salvo poi, per ragioni tecniche, economiche e anche temporali, implementare dei sottoinsiemi, i Data Mart (DM), i quali risulteranno integrabili in qualunque successivo momento.

I dati di un DM sono organizzati secondo una semplice struttura, detta schema multidimensionale o più semplicemente schema a stella.

Il primo termine mette in luce la presenza di molteplici dimensioni di analisi; il secondo la struttura “stellare” dello schema, una volta che esso venga interpretato con il modello Entità-Relazioni classico.

La principale caratteristica dello schema a stella è la sua struttura regolare e indipendente dal problema considerato.

In ambiente informazionale è indispensabile ristrutturare i dati e configurare correttamente il sistema in modo da soddisfare le necessità degli utenti finali ai quali ci si rivolge. In tal modo si consente al DBMS di elaborare le query (le interrogazioni) che si rivolgono al sistema, tenendo bene a mente che il segreto del successo verso tale comunità è molto spesso la semplicità della struttura dati con la quale questa deve interfacciarsi. Il modello dimensionale va proprio incontro a questo obiettivo di semplicità.

La figura 8 mostra un semplice esempio di schema a stella, al cui centro è stata posta la tabella dei fatti (fact table), mentre le entità satelliti più piccole sono dette tabelle delle dimensioni (dimension table). Al contrario degli schemi Entità-Relazione tradizionali, lo schema a stella è evidentemente asimmetrico: c’è una tabelladominante di grandi dimensioni al centro del diagramma, ed è l’unica a possedere collegamenti multipli, che si concretizzano attraverso identificatori univoci (campi chiavi) in essa contenuti, con le altre tabelle; queste ultime hanno tutte un solo collegamento (join) alla tabella centrale, con lo scopo di minimizzare il numero di join richiesti per ciascuna query.

A volte durante la fase di progettazione del DW, non sempre è chiaro se un campo estratto dal Data Base operazionale sia un attributo piuttosto che un fatto; spesso il dubbio può essere risolto osservando se il campo ha la caratteristica di variare il proprio valore con continuità (o meglio, ogni volta che viene rilevato), proponendosi come una potenziale misura dell’attività da inserire nella fact table; oppure se è più ragionevole considerarlo come la descrizione di qualcosa che è più o meno costante e quindi inserirlo come attributo dimensionale. Un ruolo fondamentale degli attributi delle tabelle dimensionali è quello di essere usati all’interno dei vincoli nelle interrogazioni sul DW o come intestazioni delle colonne negli output per gli utenti .

Nello schema a stella di figura 8 la fact table contiene i fatti: Ammontare delle vendite (dei biglietti d’ingresso) e Quantità venduta (dei biglietti), mentre le dimensional table ospitano ognuna la descrizione funzionale della dimensione del business presa in considerazione nella fact table, dettagliandone i parametri rilevanti ai fini dell’analisi. Per estendere l’esempio fatto precedentemente si potrebbero aggregare le vendite dei biglietti rispetto alle mostre per cercare di capire se le varie mostre hanno successo oppure no: esse potrebbero diventare un importante strumento di presentazione del museo al pubblico. Oppure si potrebbe, ancora una volta, aggregare i visitatori per tipologia di biglietto per verificare quali siano le classi di utenti (giovani, adulti, militari, comitive, ecc,.....) che maggiormente frequentano il museo per poi prendere, di conseguenza, decisioni appropriate. Ma naturalmente, le possibilità non si fermano qui.

STAR SCHEMA / FACT TABLE

1.4. Tipologia dei dati in un Data Warehouse

I dati coinvolti nella progettazione e nell’utilizzo di un DW si possono classificare in due gruppi:

dati di business;

metadati.

1.4.1. Dati di business

Sono i dati necessari allo svolgimento e alla gestione del business stesso. Essi rappresentano sia le attività svolte dall’organizzazione sia gli oggetti del mondo reale con i quali essa si trova a trattare. E’ possibile classificare i dati di business nelle seguenti quattro categorie:

**Dati Real-Time**. Rappresentano lo stato corrente del business e ne consentono lo svolgimento. Sono i dati al più alto livello di dettaglio nell’organizzazione e sono creati, manipolati e acceduti esclusivamente dai sistemi operazionali .

**Dati Storici**. Rappresentano lo stato dell’azienda nei tempi passati. Questi dati sono memorizzati su supporti poco costosi e più lenti appunto perché non acceduti di frequente.

**Dati Derivati**. Sono derivati attraverso un processo di trasformazione a partire dai dati real-time. Possono esistere sia ad un elevato livello di dettaglio sia in forma aggregata e possono essere acceduti esclusivamente in lettura da applicazioni DSS.

**Dati Riconciliati**. Sono generati attraverso un processo di razionalizzazione e di integrazione dei dati real-time provenienti da più sorgenti. Tale processo di generazione comporta la creazione e il mantenimento di un insieme di dati storici. Nei sistemi DSS raramente i dati riconciliati sono esplicitamente definiti e dove esistono raramente sono fisicamente realizzati: il più delle volte sono soltanto il risultato di operazioni eseguite sui dati real-time. La costruzione di un livello di riconciliazione dei dati è il passo fondamentale per la creazione di un DW poiché la mancanza di tale passo potrebbe portare alla proliferazione di inconsistenze tra dati operativi.

1.4.1.1. Temporalità dei dati di business.

Dal momento che il business cambia nel tempo, i dati che lo rappresentano devono essere in grado di rappresentare tale cambiamento.

Il concetto di timestamp è la modalità classica di introduzione di aspetti temporali nei databases: un timestamp è un campo in formato data-ora che definisce le proprietà temporali ad esso associate (data e ora di creazione di modifica o di cancellazione). La rappresentazione del tempo attraverso l’uso di timestamps può essere espressa a tre livelli di dettaglio:

Si può tenere traccia del tempo a *livello di campo*: viene associato un timestamp ad ogni campo del record. Raramente però, la rappresentazione a tale livello di dettaglio è utilizzata anche perché è una tecnica molto dispendiosa che aggiunge al database un volume pari a quello dei dati memorizzati.

Si può tenere traccia del tempo a *livello di record*: viene associato un timestamp ad ogni record di una tabella e tale timestamp verrà aggiornato ogni qualvolta viene modificato un campo del record ad esso associato.

Si può tenere traccia del tempo a *livello di tabella*: viene associato un timestamp all’intera tabella e tale timestamp è modificato ogni qualvolta viene modificato un campo di record nella tabella. E’ spesso il livello di controllo più richiesto dall’utente finale nonostante non sia sempre il metodo più appropriato di rappresentazione temporale.

Aspetti interessanti riguardano la struttura dei dati e come nuovi eventi modificano quelli esistenti. Da questo punto di vista, le informazioni possono essere classificate come segue:

**Dati transitori**: sono tipicamente i dati esistenti nei sistemi operazionali. La loro caratteristica fondamentale consiste nel fatto che gli aggiornamenti e le cancellazioni effettuate sui record distruggono l’informazione precedentemente memorizzata. Questo preclude ovviamente la possibilità di mantenere la storia dei singoli record; infatti le informazioni relative ad un record saranno disponibili per un tempo non determinabile, che terminerà al verificarsi di una transizione che sovrascriverà o cancellerà il record.

**Dati periodici o semi-periodici**. Un dato è definito periodico quando è disponibile l’insieme dei valori che esso ha assunto nel tempo. Sostanzialmente un dato periodico non è mai cancellato né sovrascritto; se si deve effettuare un aggiornamento di alcuni campi viene inserito un nuovo record con un timestamp aggiornato rispetto al precedente: tale timestamp dovrà quindi necessariamente far parte della chiave. Dati di tipo periodico sono tipicamente quelli riconciliati, in quanto essi vengono usati per analisi temporali comparative. Raramente invece sono periodici i dati real-time. Più spesso i dati real-time sono di tipo semi periodico; in questo tipo di dati le informazioni storiche sono mantenute soltanto parzialmente, ad esempio mantenendo soltanto gli ultimi dieci stati per ogni record.

1.4.2. I metadati

Al crescere della quantità e della varietà dei dati usati e memorizzati in un’impresa è fondamentale formalizzarne la descrizione e l’impiego affinché vengano utilizzati in maniera completa e consistente. Nasce quindi l’esigenza di conoscere quali dati sono disponibili, dove sono collocati, quali procedure di trasformazioni sono state effettuate ecc…E’ per soddisfare queste necessità che prendono vita i metadati.Esistono varie tipologie di metadati.

**Metadati build-time**.Tali metadati sono creati ed utilizzati durante la progettazione iniziale e la costruzione dei database e dei processi sia dei sistemi operazionali che dei sistemi informazionali . Le sorgenti dei metadati, quando esistono, sono in svariate forme come ad esempio tools di modellizzazione dei dati o documentazione cartacea.Un tipico esempio di metadati buid-time sono gli script per la creazione delle tabelle delle basi di dati o i modelli (ad esempio E/R) degli stessi.

**Metadati di controllo**.Tali metadati garantiscono il corretto funzionamento dei processi gestiti dai sistemi operazionali e del DW stesso.Forniscono informazioni temporali sui dati, ad esempio quando una tabella è stata modificata l’ultima volta,quando è stata lanciata un’applicazione ecc..Inoltre vengono impiegati per gestire le autorizzazioni per l’accesso sicuro ai dati e per indagare sulla frequenza di utilizzo sia dei dati che delle funzioni disponibili in modo da stabilire la loro effettiva utilità.Questa tipologia a di metadati è di interesse fondamentale per l’amministratore del DW.

**Metadati di utilizzo**.Sono i più importanti per gli utenti finali.Rappresentano essenzialmente una guida che consente all’utente di recuperare le informazioni di interesse attraverso l’enorme mole di dati disponibili. Contengono (o dovrebbero contenere) le seguenti indicazioni:

*Un sommario* che contenga le aree di business interessate dal DW,le dimensioni disponibili, i livelli di aggregazione presenti e il livello di dettaglio raggiungibile. Dovrebbero ancora essere disponibili i differenti aliases usati dall’organizzazione per denotare lo stesso concetto;

*L’origine dei dati*: l’utente ha la necessità di conoscere da quale sistema provengano i dati;

*Le sequenze di trasformazione*: l’utente deve essere in grado di comprendere attraverso quali trasformazioni gli giungono i dati, in modo da capire meglio come utilizzarli;

*Livelli di accesso*: sono informazioni riguardanti i permessi di accesso ai dati;

*Informazioni temporali* che forniscono la data e l’ora dell’ultimo aggiornamento effettuato sul sistema;

*Stima dei tempi di accesso*: in modo che l’utente sappia quanto debba aspettare per il risultato di una query; senza questa informazione l’utente potrebbe pensare ad un blocco del sistema o ad un suo errore.

Questo tipo di metadati è generalmente trascurato nella fase di progettazione, ma assume una certa importanza quando la crescita del volume dei dati comincia a creare problemi di “orientamento”. La loro realizzazione in un secondo momento favorisce, in genere, la locazione unica e controllata, ma porta al ritardo nella creazione di strumenti software per la loro gestione.

I Data Marts

Il termine Data Mart è usato per denotare una classica architettura a due livelli che si rivolga ad uno specifico problema soddisfacendo specifiche esigenze di un’area limitata ed omogenea di utenti. Le principali differenze tra un Data Mart e un Data Warehouse sono riassunte nella Tabella 1-2.

|  |  |  |
| --- | --- | --- |
|  | **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 |

Tabella 1-2 **Data Warehouse e Data Mart**

Essendo un Data Mart fortemente orientato a soddisfare le esigenze informative di un gruppo omogeneo di utenti,vengono meno alcuni vincoli fondamentali di un Data Warehouse :

mantenere i dati ad un forte livello di dettaglio per poter soddisfare richieste non previste;

garantire che il significato di ogni dato sia univoco e valido per l’intera organizzazione.

Proprio per la sua specificità un Data Mart non è in genere in grado di soddisfare richieste non previste o di supportare con modifiche minime l’aggiunta di ulteriori data sources. I Data Marts prodotti per un’azienda possono essere scalati in un DW a patto che tra essi non vi siano incoerenze che escludono la possibilità di creare un modello di dati comune. Si è soliti utilizzare il termine Data Marts anche nei DW quando si parla di una parte limitata di dati dedicata ad una singola area dell’azienda; in questo caso, però, i dati provengono da un modello comune a tutta l’impresa, il che evita i problemi d’incoerenza a patto di una piccola duplicazione di dati.

2.3. Progettazione dei Business Information Warehouses

I Business Information Warehouses costituiscono nel loro insieme il livello dei dati derivati. I passi che portano alla realizzazione di un BIW sono concettualmente gli stessi incontrati nella realizzazione del BDW, e la diversità sta nelle esigenze che i due livelli devono soddisfare: se da una parte il BDW, e quindi il livello riconciliato, deve costituire il modello-dati comune a tutta l’impresa e in cui confluiscono tutti i dati (operazionali) di interesse dell’organizzazione, dall’altra i singoli BIW sono focalizzati su particolari aspetti di business, coinvolgendo quindi soltanto una piccola porzione dei dati (non operazionali) contenuti nel BDW. In altre parole, i dati dei BIWs sono *derivati* dal BDW e trasformati per essere efficacemente utilizzati.

Quindi le caratteristiche dei BIWs sono:

Limitazione dei dati contenuti ad una ristretta area del business di interesse;

Strutture dati finalizzate all’efficienza delle interrogazioni;

Utilizzo completo da parte di applicazioni di analisi OLAP, Data Mining , reportistica.

2.3.1. Il modello multidimensionale

I BIWs devono sostanzialmente essere interrogati dagli utenti e devono essere progettati per questo scopo. Il modello entità-relazione si è rivelato inadeguato in ambienti di supporto alle decisioni sia per le basse prestazioni offerte sia per la difficoltà di essere facilmente comprensibili agli utenti finali. Un modello ER è infatti altamente simmetrico, ogni entità sembra essere uguale alle altre e risulta difficile sia identificare facilmente le entità fondamentali del business, sia identificare quali entità contengono misure numeriche di interesse e quali invece soltanto descrizioni statiche di oggetti. Inoltre mentre il modello ER è disegnato per offrire un efficiente accesso ai dati ad un alto numero di transazioni concorrenti operanti ognuna su pochi records, in un BIW tale caratteristica passa in secondo piano rispetto all’esigenza che tale modello sia facilmente comprensibile agli utenti finali e che fornisca il supporto per sofisticate analisi su un grande volume di dati. A tale scopo si è andato affermando un nuovo modello, detto **modello multidimensionale,** che struttura i dati intorno ai concetti di maggiore interesse per il business. Il tradizionale modello ER descrive “entità” e “relazioni”: ogni entità rappresenta una classe di oggetti nel mondo reale che hanno proprietà omogenee, mentre le relazioni corrispondono a classi di fatti del mondo reale significativi nel contesto in cui si opera. In un modello dimensionale invece, la struttura dati è organizzata per descrivere :

**Fatti**: un *fatto* rappresenta il processo di business da modellare (vendite, acquisti, etc) ;

**Dimensioni**: una *dimensione* determina la granularità adottata per rappresentare i fatti (prodotti, tempo, negozi) ;

**Gerarchie**: una *gerarchia,* definita su una dimensione,determina come le istanze dei fatti possono essere selezionate e aggregate ;

**Misure**: una *misura* rappresenta un attributo numerico di un fatto.

Un fatto esprime una relazione molti-a-molti tra le dimensioni: ogni combinazione di valori delle dimensioni definisce una *istanza del fatto* in corrispondenza della quale si individuano precisi valori delle misure. Un’istanza di fatto è detta primaria se è caratterizzata da uno ed un solo valore per ogni misura, secondaria se esiste un’aggregazione sulle dimensioni. Le istanze dei fatti sono le informazioni elementari rappresentate in un Business Information Warehouse.

Per fissare meglio tali nuovi concetti, facciamo riferimento al seguente esempio:

*Esempio 2-1*

In una catena di negozi il *fatto* è tipicamente il processo di vendita; le *dimensioni* scelte (che determinano la granularità di ogni istanza del fatto) sono il prodotto venduto, il negozio in cui è avvenuta la vendita e il giorno di vendita; la scelta delle *misure* ricade sulla quantità di prodotto venduta e il guadagno ricavato ;

le *gerarchie* definibili sulle dimensioni saranno :

**Dimensione Prodotto :** Prodotto🡪Marca🡪Categoria ;

**Dimensione Negozio :** Negozio🡪Città🡪Regione🡪Nazione ;

**Dimensione Tempo :** Giorno🡪Settimana🡪Mese🡪Anno.

Risulta chiaro allora che una gerarchia definisce un insieme di dipendenze funzionali all’interno di una dimensione. Si può anche osservare che scegliendo come dimensione temporale il mese di vendita invece che il giorno o la marca venduta invece che il singolo prodotto, si andrà a cambiare la granularità nel rappresentare i fatti. Una combinazione ‘prodotto venduto’, ‘giorno di vendita’, ‘negozio in cui è stata effettuata la vendita’, determina in maniera univoca le due misure ‘quantità venduta’ e ‘prezzo unitario’. Inoltre la presenza di gerarchie sulle dimensioni permette differenti livelli di aggregazione sulle istanze dei fatti. E’ possibile ad esempio aggregare opportunamente le misure di vendita per determinare la quantità di prodotti venduta in una data regione in un particolare anno, o la quantità di prodotti venduta per una data marca in una città in un dato giorno dell’anno e così via. L’utilizzo di un cubo o di un iper-cubo fornisce un modo conveniente ed intuitivo per visualizzare un modello di dati multidimensionale: ogni spigolo del cubo rappresenterà una dimensione,mentre ogni punto all’interno del cubo, determinato dall’intersezione delle coordinate definite dalle istanze di ogni dimensione (istanza del fatto) determinerà univocamente le misure di interesse.

Il modello multidimensionale è dettagliatamente trattato nel capitolo successivo in cui saranno trattati tutti gli aspetti riguardanti OLAP.

La modellizzazione di un fatto non sempre ha delle misure numeriche associate: in questo caso ogni istanza registrerà l’occorrenza di un evento. Si consideri ad esempio il fatto PRESENZA che tiene traccia delle presenze degli studenti in un college, con le dimensioni *data, studente, corso* ed *insegnante*: una istanza del fatto rappresenta la presenza di uno studente in un corso tenuto da un insegnante in una certa data.

2.3.1.1. Il Modello Dimensional Fact

Mentre è da tutti riconosciuta la validità del modello dimensionale per l’implementazione dei BIWs, soltanto di recente è stato proposto un modello dati concettuale (**Dimensional Fact Model)** che adotti per i concetti espressi in precedenza una rappresentazione grafica indipendente dalla rappresentazione logica o fisica che si adotterà in fase di realizzazione. Il modello rappresentante un fatto è strutturato come un albero la cui radice è il fatto stesso. Un fatto è rappresentato da un rettangolo che contiene il nome del fatto e, tipicamente una o più misure. In Figura 2-4 è rappresentato lo schema concettuale dell’esempio 2-1 opportunamente ampliato.



Figura 2-4, **Esempio di schema del fatto ‘Vendite’.**

Ogni nodo (rappresentato da un cerchio) direttamente unito al fatto è una dimensione (prodotto, negozio, giorno) mentre ogni sotto-albero che ha come radice una dimensione rappresenta una gerarchia (prodotto, marca, categoria).I nodi di ogni sotto-albero rappresentano attributi che possono assumere un insieme di valori discreto; un arco che unisce due nodi rappresenta una relazione molti-a-uno tra coppie di attributi. Alcune fogli dei sotto-alberi possono essere rappresentate da linee invece che da cerchi (indirizzo): questi nodi vengono detti *non-dimensional attributes* e contengono informazioni addizionali su un attributo della gerarchia senza poter essere usati per eseguire aggregazioni (ad esempio non ha senso aggregare le vendite per indirizzo del negozio).

Un interessante aspetto da prendere in esame è il *concetto di additività tra fatti*.

Il tipico processo di interrogazione di un Data Warehouse è mirato all’estrazione di dati aggregati secondo uno specifico operatore. Nella quasi totalità dei casi l’operatore di aggregazione sarà quello di somma: come conseguenza le istanze dei fatti e quindi le misure ad esse associate dovrebbero essere additive lungo le dimensioni definite. Un esempio di misura additiva è la quantità di prodotto venduta nell’esempio 2-1. Non sempre però tutte le misure risultano essere additive su ogni dimensione. Avremo allora che una misura è detta *semi-additiva* se non è additiva su una o più dimensioni, e *non-additiva* se tale misura non è additiva su nessuna dimensione. Un esempio di misura semi-additiva è rappresentato dalle rimanenze di scorte in magazzino.

Figura 2-5 ,**Esempio di misura semi-additività relativa al fatto Inventario**

Infatti tale misura, pur non essendo additiva nel tempo, è additiva lungo le dimensioni prodotto e negozio. Una misura non-additiva è la temperatura: sommare due temperature infatti difficilmente ha senso. Nel modello DF le misure sono additive su tutte le dimensioni per definizione. La semi-additività è rappresentata esplicitamente collegando attraverso una linea tratteggiata le misure semi- o non-additive alle dimensioni lungo le quali tali misure non possono essere sommate. Se un operatore di aggregazione può essere usato per un attributo non additivo su una dimensione va esplicitamente indicato.

Lo schema di fatto può non essere un “vero” albero: infatti, può accadere che due o più percorsi distinti connettano due attributi dimensionali all’interno di una gerarchia, purché ogni percorso orientato continui a rappresentare una relazione molti-a-uno (vedi Figura 2-5).

2 “Extraction, Transformation & Loading” è il processo che preleva i dati dai si- stemi alimentanti (ERP, fogli Excel etc.) e li porta nel DWH certificandoli attraverso processi di Data Quality

### *Data Warehouse*

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 op- portunatamente trasformati attraverso procedure ETL e controllati at- traverso il sistema di Data Quality.

La Qualità dei dati è un requisito fondamentale per l’intero sistema informativo. In quanto, se i dati risultano sporchi, possono oltre che causare un peggioramento delle performance aziendali, portare a pren- dere decisioni inopportune, comportando costi aggiuntivi e perdita di opportunità.

L’obiettivo di un DWH è pertanto quello di supportare il “*knowledge Worker*” (dirigente, amministratore, gestore, analista) per aiutarlo a con- durre analisi finalizzate all’attuazione di processi decisionali e al miglio- ramento del patrimonio informativo, e fornire un unico punto di accesso per tutti i dati dell’azienda – resi consistenti e affidabili attraverso i processi di ETL. Il DWH garantisce inoltre una profondità storica com- pleta dei dati, poiché in esso viene persistito anche lo stato passato delle informazioni permettendo così un’analisi temporale.

Dovranno quindi essere attentamente progettati per gestire in ma- niera efficiente ed efficace le caratteristiche dei Big Data3 [1].

I DWH sono realizzati come principale base per i Decision Support System (DSS) 4.

*“Un sistema di supporto alle decisioni è un sistema in grado di for- nire chiare informazioni agli utenti, in modo che essi possano analizzare dettagliatamente una situazione e prendere le opportune decisioni sulle azioni da intraprendere in modo facile e veloce”* [8].

3 Con il termine Big Data si intende una collezione di dataset così grandi e complessi che sono difficilmente processabili con normali database relazionali.

4 Il DSS si appoggia su dati di uno o più database, spesso organizzati in strutture diverse con dati non omogenei.

In altre parole, un sistema di questo tipo deve supportare le attività di analisi e controllo manageriale di routine, le attività di ricerca delle cause di un problema (*focused search*) e le attività di gestione manage- riale complessa (*decision making*), permettendo inoltre un facile utilizzo ad un’utenza con un tempo disponibile ridotto e riluttante verso nuove tecnologie (soprattutto nei casi in cui non riesce a percepire in breve tempo i benefici).

È possibile che un DWH sia suddiviso in diversi Data Mart, ognuno dei quali specifico per un solo processo di business fra quelli presenti all’interno dell’azienda (ordini, vendite, clienti, marketing, etc.)

### *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 [7].

L’implementazione può essere di due tipi:

* + - *Top-Down:* costruzione del DWH, e conseguente aggregazione ed esportazione nei vari Data Mart.
    - *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.

### *Architettura di un Data Warehouse*

In fase di progettazione risulta fondamentale stabilire quali tipologie di architettura adottare. Chiaramente, da quando sono stati idealizzati, i modelli (descritti successivamente) si sono evoluti e, di conseguenza, un DWH deve essere costruito secondo i principi moderni [1]. I pattern descritti in questo paragrafo rimangono comunque delle basi da cui par- tire.

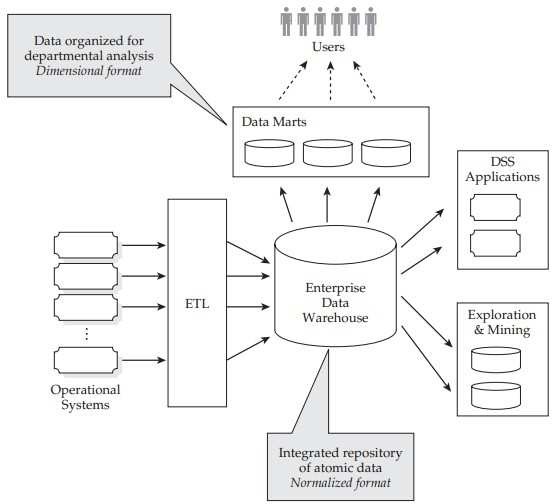
***Modello di Inmon - Corporate Information Factory****:* I DWH si costruiscono nella loro totalità fin dal principio come un unico blocco monolitico; non è possibile vederli come la composizione dei DM. Viene adottata una visione Top-Down (fig. 2).

Figura 2: Modello di Inmon

***Modello di Kimball - Dimensional Model****:* adotta un approccio Bottom-up in cui il DWH nasce dall’unione dei vari Data Mart che rife- riscono ognuno ad una specifica area di business (fig. 3).

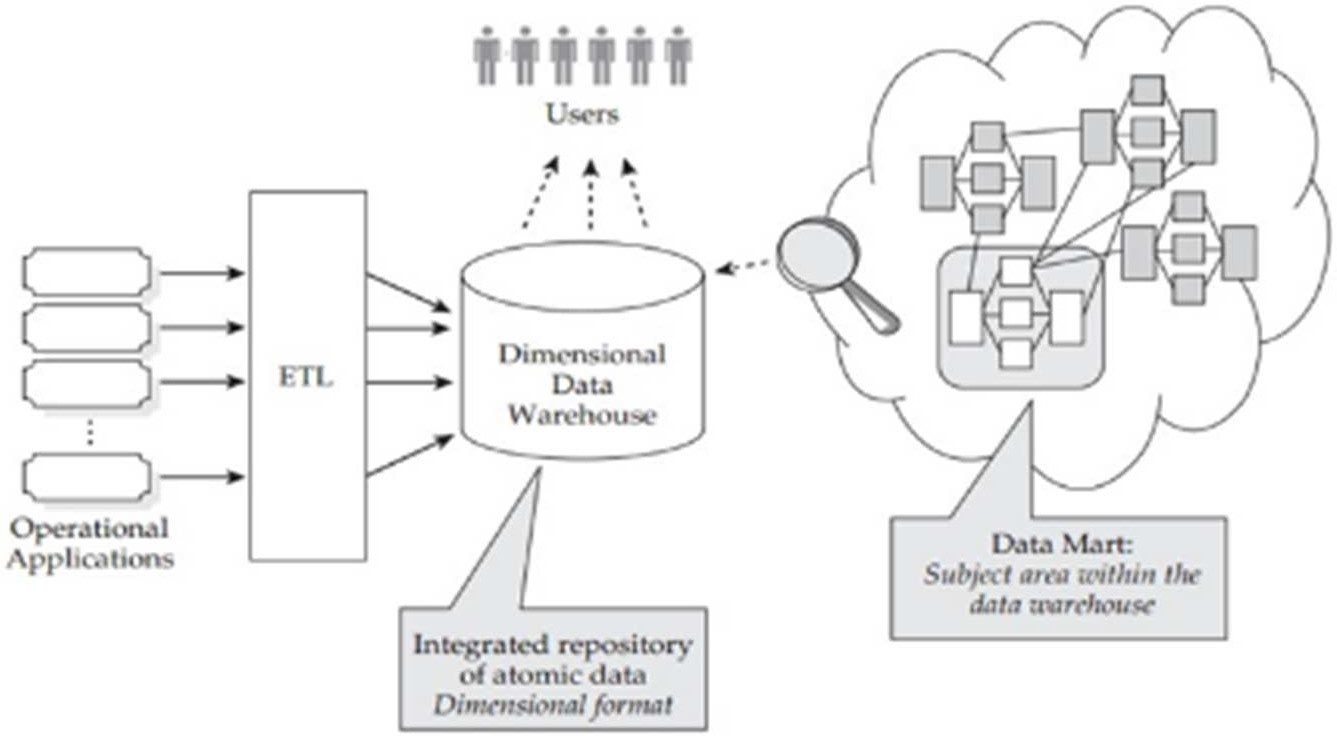


Figura 3: Modello di Kimball

### *OLTP vs OLAP*

**Online Transaction Processing (OLTP)**

A livello di database, gli Online Transaction Processing si basano su query multi-access veloci ed efficaci. Le principali operazioni svolte sono INSERT, DELETE e UPDATE in quanto modificano direttamente i dati. Questi ultimi vengono quindi costantemente aggiornati e, di conse- guenza, richiedono un efficiente supporto alle operazioni di riscrittura. Una caratteristica fondamentale di questi sistemi è la normalizzazione, la quale fornisce un modo rapido ed efficace per effettuare scrittura nel database.

**Online Analytical Processing (OLAP)**

L’Online Analytical Processing è un insieme di tecniche software per l'analisi celere e interattiva di grandi moli di dati, con la possibilità di farlo da punti di vista differenti. Questi sistemi si riveleranno molto utili per l’ottenimento di informazioni di sintesi. Queste ultime avranno lo scopo di supportare e migliorare i processi decisionali aziendale. Esempi di strumenti OLAP sono i Data Warehouse, i Cubi Multidimensionali.

Le maggiori differenze fra i due sistemi sono riportati in tabella [1]:

|  |  |  |
| --- | --- | --- |
| OLTP OLAP | | |
| Finalità | Supporto all’operatività | Supporto al processo deci- sionale |
| Modalità di utilizzo | Guidata, per processi e stati suc- cessivi | Interrogazione ad hoc |
| Quantità di dati per ope- razione elementare | Bassa: centinaia di record per ogni transazione | Alta: milioni di record per ogni query |
| Qualità | In termini di integrità | In termini di consistenza |
| Orientamento | Per processo/applicazione | Per Soggetto |
| Frequenza di aggiorna- mento | Continua, tramite azioni | Sporadica, tramite funzioni esplicite |
| Copertura temporale | Dati correnti | Storica |
| Ottimizzazione | Per accessi in lettura e scrittura su una porzione di dati | Per accessi in sola lettura su tutta la base di dati |

In base alla memorizzazione dei dati, si avranno diverse architetture OLAP, ognuna delle quali con i propri pro e contro [1]:

* + - ***Relational OLAP (ROLAP):*** i dati vengono memorizzati in un database relazionale come supporto al motore OLAP. Le analisi multidimensionali vengono tradotte in query, resti- tuendo risultati in forma multidimensionale.
    - ***Multidimensional OLAP (MOLAP):*** si ha sia il database che il motore multidimensionale. Ci saranno difficoltà per ope- razioni di Drill-Down.
    - ***Hybryd OLAP (HOLAP):*** unisce i vantaggi dei due sistemi precedenti. In particolare, pre-aggrega i dati in sistemi multi- dimensionali per un’analisi efficiente e veloce, mentre vengono ricercate in un data base relazionale in caso di Drill-Down.
    - ***Desktop OLAP (DOLAP):*** i dati vengono caricati in un sistema client e vengono calcolati dal motore in locale.

### *Modello Multidimensionale*

Il modello E-R5, 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**) [1].

Attraverso questa modellazione è possibile rappresentare i dati all’in- terno di ipercubi6 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 ma- niera intuitiva e formale;
    - Favorire la comunicazione tra designer a utenti al fine di for- malizzare i requisiti di progetto;
    - Costruire una stabile piattaforma di design logico;
    - Fornire una documentazione chiara e efficace.

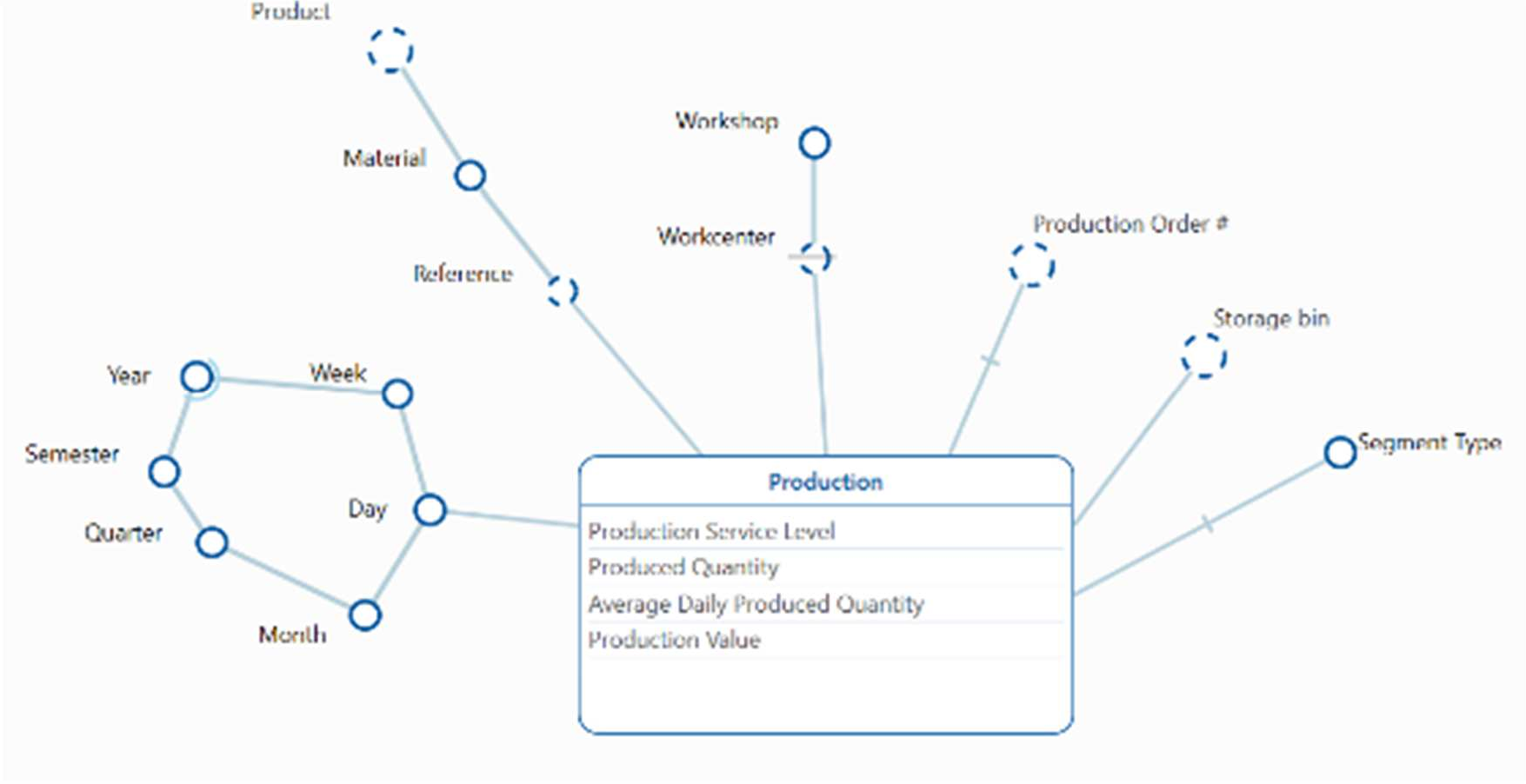


Figura 4: esempio di DFM estratto da Indyco (software sviluppato in Iconsulting)

5 Modello Entità - Relazione

6 Vengono definiti ipercubi proprio perché sono multidimensionali.

* + 1. *Concetti chiave di un DFM*
       - ***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 ri- sulta 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. Al- cuni 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.
    2. *Star Schema*

Una volta costruito il DFM, viene implementato lo schema logico. Esso viene rappresentato secondo uno Star Schema (fig. 5), 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 (3NF), mentre le tabelle dimensionali sono de-normaliz- zate [1].

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

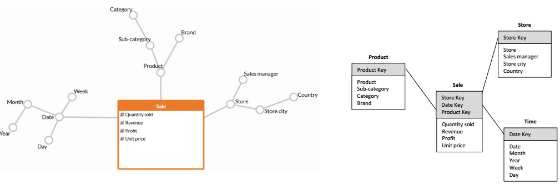


Figura 5*:* Trasformazione da modello Logico DFM a Tabellare

* + 1. *Operazioni nel modello Multidimensionale*

Per navigare all’interno del cubo multidimensionale esistono diffe- renti operazioni che permettono di organizzare i dati al suo interno, at- traverso diverse prospettive [1].

* + - * ***Pivoting****:* permette di modificare rapidamente la visualizza- zione 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.
      * ***Slice* & *Dice****:* seleziona e proietta i dati del cubo. Nello spe- cifico si estrarranno sotto-cubi filtrando su una (Slice) o più (Dice) dimensioni.
      * ***Roll-Up* & *Drill-Down****:* Queste operazioni 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.

### *Real-Time Analitycs*

L’analisi in tempo reale si basa spesso su tecniche di processamento au- tomatico dei dati ed estrazione di conoscenza attraverso modelli stati- stici e derivati dal machine learning, che permettono di fornire informa- zioni utili sfruttate successivamente per prendere decisioni opportune in modo celere. Per alcuni casi d'uso, con l’espressione “in tempo reale” si intende che l’analisi sarà completata entro pochi secondi o minuti dall'arrivo di nuovi dati [9/11].

Esiste una differenza tra *Real-Time On-demand* e *Continuos.* La prima necessita di una query elaborata dal sistema o dagli utenti per fornire un risultato analitico. La seconda, invece, è più proattiva in quanto avvisa gli utenti o innesca risposte man mano che gli eventi ac- cadono.

Nelle realtà aziendali avere una risposta continua è costoso e spesso innecessario.

Inoltre, bisogna tenere conto che esiste una differenza tra *acquisire* i dati in tempo reale e *utilizzarli* in tempo reale.

I dati catturati dai *data streaming*, vengono trasmessi in un luogo dove possono essere immagazzinati, per far sì che non vengano persi.

Essi quindi, non vengono memorizzati, ma i campioni prelevati a inter- valli regolari vengono archiviati per un ulteriore utilizzo.

Questo intervallo può variare da secondi ad anni, oltre che essere guidato da eventi.

La capacità di acquisire ed elaborare dati in tempo reale e usarli insieme a dati che sono stati pre-elaborati non è facile, non tanto a livello tecnico quanto a livello di qualità dei dati stessi.

*“Nel 98% dei casi i dati di cui hanno bisogno le aziende sono storici*” [10]

Essendo molto difficile determinare la qualità e l’affidabilità in tempo reale, l’elaborazione di molti dati in un DWH permette di avere il tempo di capire qual è la qualità, grazie al confronto con altri dati, e prendere la decisione più adatta in base all‘informazione ottenuta.

Il costo di analisi dei dati rispetto al valore che ne deriva, definisce il confine economico che viene raggiunto molto prima del limite tecnolo- gico.

Pertanto, le aziende devono porsi delle domande preliminari:

* Bisogno effettivo di **acquisire** i dati in tempo reale?
* Con che **intervallo** memorizzare i dati? memorizzarli o gene- rare eventi?
* Necessita di **elaborare** i dati in tempo reale?
* Serve anche combinarli con **dati storici**?
* Bisogna **rispondere** a un evento in tempo reale?

(Molto spesso nelle situazioni aziendali, si desidera rispondere a eventi sconosciuti)

Dunque, la quantità di casi d'uso in cui è necessario elaborare o uti- lizzare dati in tempo reale è limitata. La maggior parte dei requisiti sono espressi da un senso di urgenza causato dalla paura di cambiare le con- dizioni di mercato, incomprensioni di architetture di dati presso altre aziende o fastidio al time-to-market dell'IT in progetti correlati ai dati.

Iconsulting, per conto del cliente, ha svolto queste analisi determi- nando che una realtà Real-Time permette di mantenere le attività di analisi operative anche senza avere il consolidamento dei dati che si avrebbe da un DWH. Ciò è permesso anche dalle potenzialità dello stru- mento, già utilizzato dall’azienda, SAP HANA – e, andando più nello specifico HANA LIVE (di cui parlerò più avanti).

Ciò consente all’utente finale l’accesso al prodotto in evoluzione mentre è in fase di sviluppo, ottenendo così feedback in modo incremen- tale per evolvere e modellare in corso d’opera. Allo stesso tempo vengono consolidate le regole per la costruzione dei report finali (Fast Proto- typing7).

È indispensabile però tenere in considerazione che all’aumentare della mole di dati questa soluzione non sarà più sostenibile. Proprio per

7 In simbiosi con la metodologia Agile, l’idea è quella di generare un prototipo il più rapidamente possibile in parallelo con la capacità del business partner di articolare i requisiti. Una volta che vengono perfezionati si passa ad altri requisiti e così via.

questo, parallelamente, viene definito nello specifico come dovrà essere costruita la parte di Data Mart per il consolidamento e storicizzazione dei dati.

### *Data Visualization*

Gli strumenti di reportistica rappresentano la parte visibile della BI. Dal momento in cui esistono svariati modi di rappresentare i KPI, di- venta importante capire quale sia quello più adatto a rappresentare l’in- formazione in base al tipo di dato fornito. I fattori che dominano la progettazione devono essere semplicità e chiarezza, oltre che attendibi- lità e tempestività, per far sì che un report possa essere davvero efficace, permettendogli di svolgere il proprio ruolo quale presentatore della BI[2].

Alcuni esempi di rappresentazione possono essere:

***Report****:* presenta sia tabelle con dati disposti su righe e colonne che grafici con dati su assi cartesiani o su diagrammi a torta. La rappresen- tazione dei KPI avviene in maniera *statica*.

***Dashboard (o cruscotto)****:* è un contenitore di report ed altri ele- menti di analisi (istogrammi, mappe, diagrammi a torta) che permettere una lettura intuitiva, pertinente e di facile comprensione. La rappresen- tazione dei KPI avviene in maniera *dinamica*.