Business Intelligence (BI)

La Business Intelligence (BI) è per definizione una raccolta di processi, strumenti e tecnologie utili per ottenere maggiori profitti con la capacità di utilizzare tramite indicatori di prestazioni chiave per misurare le prestazioni.

**L’ Evoluzione della BI**

L’ evoluzione della BI è iniziata decenni fa con i primi report mainframe, chiamati output di sistema. Essi venivano principalmente stampati su carta, per poi essere periodicamente distribuiti ai manager.

Ci si aspettava che i manager si facessero strada attraverso di loro per identificare le informazioni pertinenti che potevano usare nella loro decisione tattica e strategica. Questo però portava anche delle complicazioni a livello logistico ed organizzativo, in quanto avendo molti fogli, la probabilità di perderne qualcuno era molto elevata.

Le prime query hanno velocizzato il processo e hanno consentito ai manager tecnicamente esperti di creare report personalizzati ad hoc, ma pochi manager avevano il tempo e le competenze per farlo. L'emergere del data warehouse ha dato un grande impulso alla BI aggregando tutti i dati in un'unica posizione, dove potrebbe essere interrogato in modo interattivo senza impatto sulle applicazioni di produzione.

Gli strumenti di query e rapporti online con interfacce grafiche sempre più facili da utilizzare hanno reso la BI accessibile a più gestori e ha consentito a tali manager di ottenere informazioni e risposte critiche in modo efficiente e rapidamente. Le data warehouse sono stati seguiti da data store specializzati in data mart che hanno ulteriormente accelerato il processo di acquisizione di informazioni ai responsabili ai fini di un processo decisionale informato. Poi c'erano gli strumenti di analisi analitica online (OLAP) e altri strumenti analitici multidimensionali, che permettevano ai manager di tagliare a dadi e giuntare i dati in una varietà di modi e di estrapolarli per informazioni nascoste altrimenti. A quel punto, la BI ha iniziato a fondersi con Business Analytics (BA).

Oggi BI e BA vengono forniti come applicazioni che si basano su un'infrastruttura di database, sistemi di gestione dei dati, funzionalità di estrazione, trasformazione e caricamento (ETL) e altro ancora. L'infrastruttura BI può includere dashboard esecutivi, scorecard e altri strumenti che rendono più facile per i manager trovare e comprendere le informazioni e utilizzarle in modo proattivo nel processo decisionale. Questa evoluzione decennale non ha avuto un prezzo. Durante questo periodo le aziende hanno acquistato e distribuito un'ampia varietà di prodotti correlati alla BI. Hanno utilizzato diversi modelli di implementazione, diverse interfacce utente e una gestione diversa

interfacce e avere diversi requisiti di integrazione. A questo punto le aziende possono disporre di otto, dieci o più prodotti di BI diversi. Il costo per mantenere questa proliferazione di strumenti e tecnologie è già alto e aumenterà. Business Intelligence (BI) è il portafoglio più completo di tecnologia e applicazioni.

**Strumenti per implementare un Business Intelligente di successo**

Alcuni degli strumenti necessari per implementare un sistema intelligente di business moderno con successo includono:

**Data Warehousing**

Un data warehouse è un archivio unico, completo e coerente di dati ottenuti da varie fonti. È un database che contiene dati che sono stati puliti e trasformati in un formato informativo**.** I data warehouse sono strumenti per sistemi di informazione organizzati. Di solito sono database relazionali, cioè database dove si possono confrontare, abbinare e mettere in relazione attributi di dati simili.

Altra definizione di un data warehouse è un repository di informazioni integrate, disponibile per la realizzazione di interrogazioni tramite query e analisi principalmente di business.

Dati e informazioni sono estratti da fonti eterogenee man mano che vengono generati. Ciò rende molto più facile ed efficiente eseguire query su dati originariamente provenienti da fonti diverse.

Alcune delle caratteristiche di base per l'implementazione di un data warehouse di successo includono:

* Una serie di programmi che estraggono i dati da un ambiente operativo.
* Un database che conserva i dati del data warehouse.

* Una caratteristica distintiva di un data warehouse è la separazione delle funzionalità di supporto.

Uno dei motivi per cui il data warehousing ha richiesto così tanto tempo per svilupparsi è che si tratta in realtà di una tecnologia molto completa. In effetti, il data warehousing può essere rappresentato al meglio come una struttura a livello aziendale per la gestione dei dati informativi all'interno dell'organizzazione (W. Ken et al 2010). Per capire come tutti i componenti coinvolti in una strategia di data warehousing sono correlati, è essenziale avere un'architettura di Data Warehouse.

Un'architettura di Data Warehouse è un modo di rappresentare la struttura generale dei dati, delle comunicazioni, dell'elaborazione e della presentazione esistenti per l'elaborazione degli utenti finali all'interno dell'azienda. (Tecnologia di Data Warehousing a White Paper di Ken Institute, 2010).

L'architettura è composta da un certo numero di parti interconnesse:

1. Database operativo / Livello database esterno

1. Livello di accesso alle informazioni

1. Livello di accesso ai dati

1. Layer Data Directory (Metadata)

1. Livello di gestione dei processi

1. Livello di messaggistica dell'applicazione

1. Livello del data warehouse

1. Livello di gestione dei dati

1. **Database operativo / Livello database esterno:** I sistemi operativi elaborano i dati per supportare esigenze operative critiche. Per fare ciò, sono stati storicamente creati database operativi per fornire una struttura di elaborazione efficiente per un numero relativamente piccolo di transazioni commerciali ben definite. Tuttavia, a causa della scarsa attenzione dei sistemi operativi, i database progettati per supportare i sistemi operativi hanno difficoltà ad accedere ai dati per altri fini gestionali o informativi. Questa difficoltà nell'accedere ai dati operativi è amplificata dal fatto che molti sistemi operativi hanno spesso 10 o 15 anni. L'età di alcuni di questi sistemi significa che la tecnologia di accesso ai dati disponibile per ottenere dati operativi è essa stessa datata.
2. **Livello di accesso alle informazioni:** il livello di accesso alle informazioni dell'architettura del data warehouse è il livello con cui l'utente finale gestisce direttamente. In particolare, rappresenta gli strumenti che l'utente finale utilizza normalmente giorno per giorno, ad esempio Excel, Lotus 1-2-3, Focus, Access, SAS, ecc. Questo livello include anche l'hardware e il software coinvolti nella visualizzazione e stampa rapporti, fogli di calcolo, grafici e grafici per analisi e presentazione.
3. **Livello di accesso ai dati:** il livello di accesso ai dati dell'architettura del data warehouse è coinvolto nel consentire al livello di accesso alle informazioni di comunicare con il livello operativo. Nel mondo della rete oggi, il linguaggio dei dati comune che è emerso è SQL. Originariamente, SQL è stato sviluppato da IBM come linguaggio di query, ma negli ultimi vent'anni è diventato lo standard di fatto per lo scambio di dati.
4. **Layer Data Directory (Metadata):** per fornire un accesso universale ai dati, è assolutamente necessario mantenere una qualche forma di directory di dati o repository di informazioni sui meta-dati. I meta-dati sono i dati relativi ai dati all'interno dell'azienda.Per avere un magazzino completamente funzionale, è necessario disporre di una varietà di metadati disponibili, dati sulle visualizzazioni dell'utente finale di dati e dati sui database operativi. Idealmente, gli utenti finali dovrebbero essere in grado di accedere ai dati dal data warehouse (o dai database operativi) senza dover sapere dove risiedono i dati o il modulo in cui sono archiviati.
5. **Livello di gestione processo:**Il Process Management Layer è responsabile della pianificazione delle varie attività che devono essere eseguite per creare e mantenere il data warehouse e le informazioni sulla directory dei dati. Il Process Management Layer può essere considerato come lo schedulatore o il controllo di alto livello del lavoro per i numerosi processi (procedure) che devono verificarsi per mantenere aggiornato il Data Warehouse.
6. **Applicazione di messaggistica di livello:** Il livello di applicazione del messaggio ha a che fare con il trasporto delle informazioni giro per la rete enterprise computing. L'Application Messaging viene anche chiamata "middleware", ma può coinvolgere più di un semplice protocollo di rete. Application Messaging può anche essere utilizzato per raccogliere transazioni o messaggi e consegnarli a una determinata posizione in un determinato momento.
7. **Livello (fisico) del Data Warehouse:** il Data Warehouse (principale) è il luogo in cui si verificano i dati effettivi utilizzati principalmente per usi informativi. In alcuni casi, si può pensare al Data Warehouse semplicemente come una visualizzazione logica o virtuale dei dati. In un magazzino di dati fisici, le copie dei dati operativi e / o esterni vengono effettivamente archiviati in un formato di facile accesso ed estremamente flessibile. Sempre più spesso, i data warehouse sono archiviati su piattaforme client / server, ma sono spesso archiviati anche su mainframe.
8. **Livello di gestione temporanea dati:** il componente finale dell'architettura Data Warehouse è Data Staging. La gestione temporanea dei dati viene anche definita gestione della copia o della replica, include tutti i processi necessari per selezionare, modificare, riepilogare, combinare e caricare i dati di accesso al data warehouse e alle informazioni da database operativi e / o esterni. Data Staging comporta spesso una programmazione complessa, ma vengono creati sempre più strumenti di data warehousing che aiutano in questo processo. Data Staging può anche coinvolgere programmi di analisi della qualità dei dati e filtri che identificano modelli e dati.

**Partizionamento del database: nozioni di base**

Storicamente, coloro che hanno trovato il massimo beneficio dal partizionamento del database hanno avuto una o più delle seguenti caratteristiche presenti nel loro ambiente:

* Sistemi di database che supportano applicazioni mission-critical.
* Sistemi caratterizzati da carichi di lavoro OLTP elevati.
* Sistemi di database che supportano attività OLAP, DSS, data warehousing o data mining.

Alcuni sostenitori del partizionamento del database vedono un'applicazione ancora più ampia per il suo utilizzo. -In ambienti aziendali di grandi dimensioni, non appena i singoli oggetti diventano più grandi di 1 a 10 GB, il partizionamento può avere senso, ‖ afferma Hermann Baer, ​​un product manager principale per il data warehousing su Oracle. Note di Baer, ​​occasioni in cui i clienti sono stati in grado di utilizzare il partizionamento per suddividere i dati in una tabella importante e isolare quindi il rischio che la tabella generale non sia più disponibile. Un tale approccio, afferma, consente anche un recupero più rapido di partizioni isolate da vari tipi di interruzioni . LaYuhanna di Forrester ha visto la partizione utilizzata quasi sempre in ambienti di data warehousing, dove la motivazione per l'uso è piuttosto semplice.

-Se hai un database di grandi dimensioni, devi gestirlo in blocchi per ottenere una migliore gestione dei dati. L'obiettivo principale del partizionamento dei database nell'ultimo decennio è stato quello dei data warehouse e dei data mart, in cui si suddividono le query in unità più piccole in modo che possano essere eseguite in parallelo su partizioni diverse , " afferma. -È dividere e conquistare. Dividi i dati e conquista. Se si dispone di grandi quantità di dati in esecuzione su un singolo processore, tale query può essere scomposta in unità più piccole che vengono eseguite contemporaneamente a più processori, in cui ogni processore va dopo una singola partizione di dati‖.Yuhanna afferma che i fornitori di database hanno diversi modi per implementare il partizionamento del database, con alcuni modi per suddividere e tagliare i dati rispetto ad altri. Il metodo di partizionamento del database più comune consiste nel suddividere una tabella più ampia basata sul tempo, come una tabella contenente dati di vendita

- in singole partizioni per settimana, mese, trimestre o anno. Questo è noto come partizionamento dell'intervallo perché consente di rappresentare un intervallo di date o numeri nella partizione, con ovvi punti di inizio e fine per l'intervallo. La Figura 2.1 mostra una tipica tabella delle vendite suddivisa per mese.

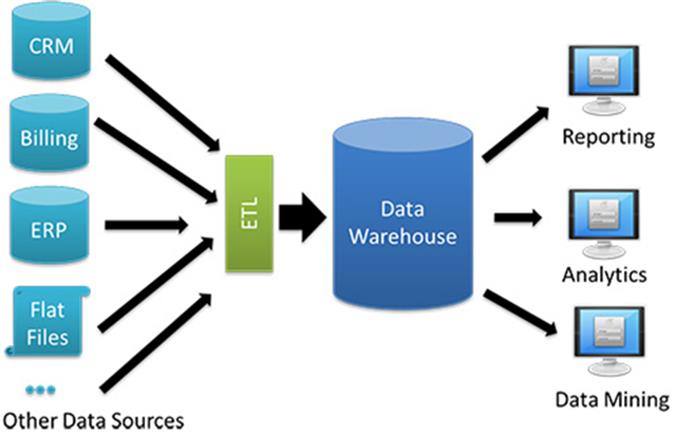
Una volta che le partizioni sono state definite in una tabella, le query che vengono eseguite rispetto alla tabella vengono in genere ottimizzate per eseguire automaticamente la "prune" delle partizioni non necessarie fuori dal processo di query prima di eseguire la query. Ad esempio, se si desidera analizzare tutte le vendite effettuate nelle ultime due settimane di aprile, la query verrà inviata alla tabella delle vendite ed elaborerà automaticamente i dati solo dalla partizione -April‖ della tabella, rispetto ai dati di tutte le 12 partizioni .Gli utenti don' t bisogno di conoscere la struttura di partizionamento, neanche. Le query utilizzano la stessa lingua, indipendentemente dal fatto che la tabella sia partizionata o meno. Ciò che è diverso è il partizionamento di efficienza sottostante che ora si inserisce nel processo. Questo processo può comportare significative riduzioni dei tempi di elaborazione . Baer lo spiega in questo modo: -Nell'esempio più semplice di una strategia di partizionamento basata sul tempo partizionata per mese, potresti solo voler guardare all'ultimo mese (o un dodicesimo) dei dati in una tabella. Ciò significa che le risorse per soddisfare le richieste di business sono anche un dodicesimo come mostrato nella figura 2.1. Rimanendo in questo esempio semplicistico, si potrebbe teoricamente eseguire 12 volte il numero di utenti sullo stesso hardware o eseguire query 12 volte più velocemente.‖

Sviluppi e ricerca in tecnologia di Business Intelligence

Dai primi sistemi di Business Intelligence (BI) che sono iniziati con i primi report mainframe, uno degli obiettivi principali dei ricercatori èstato quello di ridurre i costi dei sistemi intelligenti aziendali. Questo può essere osservato dall'evoluzione del sistema intelligente di business descritto sopra. Pochi possono discutere più con i benefici della conoscenza e della gestione guidata dalle informazioni. A lungo annunciati da luminari come Peter Drucker, l'informazione e la conoscenza sono emerse come la chiave del successo aziendale nel 21 ° secolo. Sebbene la BI svolga un ruolo centrale nel consentire alle organizzazioni di sfruttare il potere delle informazioni e dei dati, le aziende si sono a lungo dibattute con la natura frammentata della BI come si è evoluta. La natura frammentata della BI aumenta il costo delle iniziative di BI, limita la sua applicazione e riduce la sua efficacia nel guidare il miglioramento del business. Un altro problema significativo con un approccio frammentato alla BI è la possibilità di fornire informazioni incoerenti e inaffidabili. Con questo e gli altri difetti di un approccio frammentato, ciò che è richiesto è un approccio unificato per BI in cui i vari pezzi vengono indirizzati in un unico, integrato solutio n

Concetti di modellazione ETL

La struttura generale per i processi ETL è mostrata in Fig. 1. I dati vengono estratti da diverse fonti di dati, e poi propinati al DSA dove viene trasformato e pulito prima di essere caricato nel data warehouse. Fonte, area di sosta, e di destinazione ambienti possono avere molti formati di struttura dati diversi come fl a fi le, insiemi di dati XML, tabelle relazionali, fonti non relazionali, fonti web log, tems legacy sistem e fogli di calcolo.



**Le fasi ETL**

Durante il processo ETL, i dati vengono estratti da un database OLTP, trasformati per corrispondere allo schema del data warehouse e caricati nel database (Berson and Smith, 1997; Moss, 2005). Molti data warehouse incorporano anche dati provenienti da sistemi non-OLTP, ad esempio testo fi, sistemi legacy, e fogli di calcolo. ETL è spesso una complessa combinazione di processi e tecnologie che consuma una porzione significativo degli sforzi di sviluppo di data warehouse e richiede le competenze di analisti di business, i progettisti di database e sviluppatori di applicazioni. Il processo ETL non è un evento una tantum. Poiché le fonti di dati cambiano, il data warehouse verrà periodicamente aggiornato. Inoltre, poiché il business cambia il sistema DW deve cambiare - al fine di mantenere il suo valore come uno strumento per i decisori, come conseguenza l'ETL cambia e si evolve. I processi ETL devono essere progettati per facilità di modificazione. Un sistema ETL solido, ben progettato e documentato è necessario per il successo di un progetto di data warehouse.

Un sistema ETL consiste di tre fasi funzionali consecutive: estrazione, trasformazione e caricamento:

**Estrazione**

Il passo primo in qualsiasi scenario ETL è l'estrazione dei dati. La fase di estrazione dell'ETL è responsabile dell'estrazione dei dati dai sistemi di origine. Ogni fonte di dati ha la sua serie distinta di caratteristiche che devono essere gestite al fine di estrarre in modo efficace i dati per il processo ETL. Il processo deve integrare in modo efficace sistemi che hanno piattaforme diverse, ad esempio diverse

Durante l'estrazione dei dati da diverse fonti di dati, il team di ETL deve essere a conoscenza di:

(a) utilizzando i driver ODBC JDBC connect a fonti di database

(b) comprendere la struttura dei dati di fonti

(c) sa come gestire le fonti con natura diversa come i mainframe.

Il processo di estrazione consiste in due fasi, estrazione iniziale e modifica dell'estrazione dei dati. Per l'estrazione iniziale (Kimball et al., 1998), è la prima fi tempo di ottenere i dati dalle diverse fonti operative da caricare nel data warehouse. Questo processo viene eseguito una sola volta dopo aver creato il file DW per popolarlo con un'enorme quantità di dati dai sistemi di origine. L'estrazione incrementale è chiamato cambiato la cattura dei dati (CDC), dove i processi ETL rinfrescano il DW con la modi fi cato e ha aggiunto dati nei sistemi di origine dopo l'ultima estrazione. Questo processo è periodico in base al ciclo di aggiornamento e alle esigenze aziendali.Cattura anche solo chan - dati GED dall'ultima estrazione utilizzando molte tecniche come colonne di audit, di registro del database, data di sistema, o la tecnica delta.

**Trasformazione**

Il secondo passo in ogni scenario ETL è la trasformazione dei dati. La fase di trasformazione tende a rendere la pulizia e la formazione dei dati in entrata per ottenere dati accurati corretti, completi, coerenti e non ambigui. Questo processo include la pulizia, la trasformazione e l'integrazione dei dati. Si de- definisce la granularità delle tabelle dei fatti, le tabelle delle dimensioni, dello schema DW (sguardo o neve fl Ake), fatti derivati, lentamente cambiando dimensioni, tabelle dei fatti privi di dati utente. Tutte le regole di trasformazione e gli schemi risultanti sono descritti nel repository dei metadati.

**Caricamento in corso**

Caricamento dei dati alla struttura multidimensionale obiettivo è la fi - passo ETL finale. In questo passaggio, i dati estratti e trasformati vengono scritti nelle strutture dimensionali effettivamente accessibili dagli utenti finali e dai sistemi applicativi. La fase di caricamento include sia il caricamento delle tabelle delle dimensioni che il caricamento delle tabelle dei fatti.

**Modelli di processi ETL**

Questa sezione navigare attraverso gli sforzi fatti per concep - tualize i processi ETL. Sebbene i processi ETL siano fondamentali per la creazione e la manutenzione dei sistemi DW, è evidente la mancanza di un modello standard che possa essere utilizzato per rappresentare gli scenari ETL. Dopo aver creato il nostro modello, faremo un confronto tra questo modello e i modelli discussi in questa sezione. La ricerca nel campo dei processi di modellazione ETL possono essere classificati in tre approcci principali:

1.               Modellazione basata su mappature di espressioni e linee guida.

2.               Modellazione basata su costrutti concettuali