BUSINESS INTELLIGENCE E DATA WAREHOUSING: INTRODUZIONE

Slide 2 – Sommario

Benvenuti!

In questa lezione vedremo innanzitutto cosa si intende per Business Intelligence e come i dati possano fornire supporto alle decisioni. Esamineremo, poi, i componenti di un sistema di Business Intelligence e ci focalizzeremo sul Data Warehouse. Affinché tale componente sia un efficace supporto alle decisioni è necessario che sia popolato da dati di buona qualità, quindi, considereremo anche questo aspetto.

Cominciamo…

Slide 3 – Preliminari: dati e informazioni

Iniziamo a introdurre alcuni termini importanti e a chiarirne le differenze:

Con **dati** intendiamo una codifica strutturata di singole entità primarie o transazioni che le riguardano. Un dato è quindi un valore **atomico** (come una stringa, un numero o una data) o **strutturato**, quindi un record composto da una stringa, una data e una sequenza di numeri:

- questi dati possono rappresentare descrizioni e prezzi di prodotti, indirizzi di punti vendita, la data e gli importi sugli scontrini di vendita;
- il contesto interpretativo è quello che permette ai dati di diventare informazioni cioè di acquisire un significato per chi le riceve;
- parliamo di informazioni aggregate per indicare informazioni riassuntive ottenute da elaborazioni e sintesi a partire da dati elementari. Ad esempio, l'importo totale incassato da un punto vendita in una certa data;
- indichiamo, infine, con **conoscenza** le informazioni sulle quali ci possiamo basare per prendere decisioni.

Slide 4 – Valore strategico delle informazioni

Tutte le organizzazioni raccolgono, nella propria operatività quotidiana, grandi volumi di dati che chiameremo **operazionali**. Pensiamo al caso dei supermercati e a tutti i dati relativi alle transazioni di vendita, necessari per verificare l'incasso, garantire l'approvvigionamento delle merci, adempiere agli obblighi fiscali. I dati operazionali sono disponibili in grandissima quantità e sono estremamente dettagliati, ma hanno un bassissimo valore informativo e strategico. La Business Intelligence si basa sul **distillare la conoscenza**, cioè le **informazioni strategiche**, quindi utili ai **fini decisionali**, presenti nei dati operazionali. Mediante filtraggi e aggregazione diminuirà significativamente la quantità e aumenterà il valore.

Slide 5 – Business Intelligence

Con il termine **Business intelligence** si indica quindi un insieme di strumenti e tecniche che danno la possibilità a un'organizzazione di trasformare i propri dati in conoscenza. Ciò consente di prendere decisioni efficaci e tempestive. Tali strumenti permettono infatti di considerare molte alternative e giungere a conclusioni accurate. In riferimento all'esempio delle vendite, guardando l'incasso di un determinato

prodotto in un certo periodo possiamo decidere se stiamo rispettando i nostri target e valutare l'efficacia di campagne promozionali o sconti.

Slide 6 – Decisioni basate sui dati

I sistemi di Business intelligence sono utilizzati dai decisori per ottenere una conoscenza completa del business e dei fattori che lo influenzano, nonché per definire e supportare le proprie strategie.

L'obiettivo è consentire **decisioni basate sui dati** volte a ottenere un vantaggio competitivo, migliorare le prestazioni operative, rispondere più rapidamente ai cambiamenti, aumentare la redditività e, in generale, creare valore aggiunto per l'azienda.

La conoscenza estratta dai dati permette di controllare meglio il rischio insito in ogni decisione.

Slide 7 – Ciclo decisionale

Il ciclo decisionale della Business Intelligence si basa sulla ripetizione di **quattro fasi**:

Analisi: in cui si formula il problema e si individuano le informazioni rilevanti che possiamo ottenere dai dati.

Comprensione: in cui si capisce il problema e come trasformare le informazioni in conoscenza.

Decisione: in cui la conoscenza si traduce in decisioni e azioni.

Misurazione: in cui si valuta l'effetto delle azioni sulle prestazioni dell'organizzazione, ad esempio, calcolando alcuni indicatori.

Slide 8 – Componenti di un sistema di Business Intelligence

Un sistema di Business Intelligence è quindi costituito da diverse componenti. A un estremo abbiamo il punto di partenza, cioè le **sorgenti dati operazionali** che alimentano il processo decisionale. I processi ETL e il Data Warehouse rappresentano le componenti di back-end del sistema, mentre l'utente (il decisore) interagirà con gli strumenti di presentazione o di front-end.

Slide 9 – Sorgenti dati

Le sorgenti dati contengono dati operazionali, o di dettaglio, e possono essere basi di dati dell'organizzazione o sorgenti dati a diverso livello di strutturazione (ad es. fogli di calcolo o file di testo) sia interne che esterne all'organizzazione.

Slide 10 – Processi ETL

I processi di **estrazione, trasformazione e loading** (ETL) si occupano di selezionare, integrare, riconciliare i dati, trasformandoli nel formato per l'analisi e migliorandone anche la qualità. Per effettuare le trasformazioni viene utilizzata un'area di memorizzazione temporanea detta **Data Staging area**.

Slide 11 – Data Warehouse

I dati così trasformati sono memorizzati nel Data Warehouse dell'organizzazione, un repository di informazioni che raccoglie e integra dati provenienti da fonti diverse ed eterogenee rendendoli disponibili per analisi finalizzate alla pianificazione e al processo decisionale.

Il Data Warehouse può essere organizzato in Data Mart, cioè componenti che contengono un sottoinsieme o aggregazione dei dati archiviati in un Data Warehouse primario. Include una serie di informazioni rilevanti per una specifica area o reparto aziendale, categoria di utenti o linea di business.

Slide 12 – Strumenti di presentazione

La componente più vicina all'utente, o componente di front-end, consiste in un insieme di strumenti di presentazione che possono essere strumenti OLAP per eseguire sessioni interattive di interrogazione di tipo analitico, strumenti di reportistica, strumenti statistici o di Data Mining.

Slide 13 – Business Intelligence & Data Warehousing

Abbiamo individuato le componenti del sistema di Data Warehousing che può essere definito come un sistema che estrae, pulisce, rende conformi i dati sorgente e li consegna a un archivio dati dimensionale, abilitando e implementando le interrogazioni e le analisi finalizzate al prendere decisioni. I task di estrazione, pulizia e riconciliazione sono effettuati dalla componente ETL e sono una delle parti più costose e difficili. Il Data Warehouse è l'archivio dati dimensionale e gli strumenti di presentazione sono la parte più visibile che permette di effettuare definizioni e analisi.

Slide 14 – Data Warehouse – Caratteristiche

Al centro del sistema abbiamo quindi il Data Warehouse: una raccolta dati che supporta i processi decisionali e la cui più importante caratteristica è essere orientato ai **soggetti** (anziché alle applicazioni come le basi dati operazionali). Le basi di dati operazionali, infatti, sono a supporto di particolari applicazioni e servizi offerti come la gestione del magazzino, la gestione dei dipendenti, la contabilità, la vendita online. Nel caso di un Data Warehouse il servizio abilitato è sempre l'analisi e quindi l'interesse è focalizzato sui soggetti coinvolti: i clienti, i dipendenti, ecc.

Il Data Warehouse deve inoltre essere **integrato e consistente** (riconcilia tutti i dati di interesse per l'analisi). Mostra non solo l'istantanea corrente dei dati, ma la loro evoluzione nel tempo. Inoltre non è volatile perché mantiene tutta la storia dei dati.

Questa caratteristica permette infatti di individuare tendenze e di validare l'esito di azioni intraprese in passato.

Slide 15 – OLAP vs OLTP

Ma perché c'è bisogno di un altro repository? Le esigenze di elaborazione operazionali, dette anche OLTP, Online Transactional Processing, sono molto diverse da quelle analitiche, dette OLAP, Online Analytical Processing. In riferimento all'esempio della catena di supermercati, esempi di elaborazioni operazionali possono essere:

- la transazione di vendita;

- la verifica della disponibilità in un punto vendita o del prezzo di un prodotto.

Un esempio di interrogazione OLAP invece può riguardare:

- il confrontare gli incassi di diversi punti vendita in un certo periodo;
- la verifica dell'andamento stagionale delle vendite di latticini.

Gestire nello stesso repository il carico OLTP e OLAP non soddisfa adeguatamente le esigenze di nessuno dei due.

Slide 16 – Differenze tra un Data Warehouse e un database

Vediamo meglio queste differenze. Come discusso prima, lo scopo per cui viene realizzata una base di dati dipende dall'applicazione, mentre nel caso di un Data Warehouse lo scopo è sempre il supporto alle decisioni. Il focus è quindi nel primo caso sull'**applicazione**, mentre nel secondo è sui **soggetti**. La qualità dei dati è definita in una base dati operazioni in termini di **integrità**: vogliamo cioè assicurarci — tramite opportuni vincoli — che i dati riflettano accuratamente situazioni possibili nel mondo reale.

In un Data Warehouse, al contrario, l'enfasi è sulla consistenza di dati che provengono da sorgenti autonome e eterogenee: vogliamo cioè assicurarci di avere incastrato correttamente i diversi pezzi presi dalle varie sorgenti del puzzle complessivo. I dati di una base di dati operazionale sono elementari e di tipo sia numerico che alfanumerico. In un Data Warehouse abbiamo invece dati aggregati e principalmente numerici.

Il carico di lavoro operazionale è costituito da transazioni predefinite, mentre quello analitico è dato da interrogazioni ad hoc, difficilmente prevedibili e molto variabili.

Gli utenti di una base di dati operazionale sono molti di più di quelli di un Data Warehouse (decisori).

Slide 17 – Differenze tra un Data Warehouse e un database (2)

Una base di dati prevede accesso in lettura e scrittura, a centinaia di record, mentre un Data Warehouse contempla accessi principalmente in lettura a un numero molto maggiore di record. Gli accessi OLTP sono relativi a una piccola frazione dei dati, mentre quelli OLAP riguardano una frazione molto più significativa.

Una base di dati operazionale è aggiornata continuamente (si pensi ad esempio alla disponibilità o al prezzo applicato per un prodotto) mentre a fini analitici l'aggiornamento può essere effettuato periodicamente (ad esempio una volta al giorno).

Infine, una base di dati operazionale ha una copertura temporale limitata (i dati sono sovrascritti quando aggiornati), mentre un Data Warehouse contiene anche dati storici.

Slide 18 – Differenze tra OLAP e OLTP – Conseguenze

Come conseguenza di queste differenze, alcuni dei cardini dell'organizzazione e della gestione dei dati a fini operazionali sono stati quindi rivisti, in particolare i meccanismi transazionali (non più necessari), la normalizzazione degli schemi e le tecniche di indicizzazione (pensate per accessi selettivi).

Anche il modello stesso con cui organizziamo i dati può essere esplicitamente finalizzato all'analisi: sarà il modello dimensionale di cui parleremo nella prossima video lezione.

Slide 19 – Processi ETL

Andiamo ora a esaminare cosa succede ai dati prima di essere consegnati al Data Warehouse.

I processi ETL trasformano i dati dal formato sorgente (operazionale) a quello di destinazione (analitico). Questa trasformazione si effettua quando il repository di destinazione viene popolato per la prima volta e ogni volta che viene aggiornato regolarmente.

Slide 20 – Perché trasformare?

Perché c'è bisogno di una fase di trasformazione?

Per potere estrarre conoscenza utile i dati devono essere **integrati e di buona qualità**.

Avere fonti dati operative diverse ed eterogenee richiede di stabilire una mappatura tra il livello di dati di origine e il livello di analisi di destinazione affrontando due problematiche:

1. Eterogeneità (pulizia e riconciliazione)
2. Passaggio da dettaglio a aggregato, dal modello dei dati operazionale a quello analitico (che, come vedremo, sarà multidimensionale).

Slide 21 – Qualità dei dati

La prima «metà» della trasformazione include anche una fase di pulizia per migliorare la qualità dei dati. I problemi di qualità possono essere presenti anche nelle singole sorgenti, ma sono amplificati dall'integrazione di sorgenti eterogenee. Esempi di problemi includono:

- dati mancanti;
- dati duplicati;
- dati impossibili o errati;
- uso imprevisto dei campi;
- valori incoerenti per una singola entità perché sono state utilizzate diverse rappresentazioni o pratiche diverse;
- valori incoerenti per una singola entità a causa di errori di battitura;
- valori incoerenti che sono logicamente associati.

Slide 22 – Come trasformare?

Questa trasformazione può essere effettuata mediante linguaggi (di scripting, generali, SQL procedurale, ecc.) o utilizzando strumenti di ETL.

Si parla di ELT invece che di ETL quando la trasformazione avviene all'interno del sistema che gestisce il repository analitico.

Gli strumenti di ETL forniscono supporto per la definizione di una pipeline (connessione alla sorgente dati, workflow di trasformazioni, alimentazione della destinazione) e per lo scheduling e il monitoraggio dell'esecuzione automatizzata della pipeline.

Slide 23 – Riepilogo e conclusioni finali

Bene, siamo giunti alla fine di questa video lezione.

Ti ricordo che abbiamo come visto come la Business Intelligence usa i dati, trasformandoli in conoscenza, per aiutare a prendere decisioni.

Abbiamo introdotto:

- le componenti dei sistemi di Business Intelligence, distinguendo componenti di back-end e front-end
- il Data Warehouse come repository integrato per l'analisi
- i processi ETL per popolare il Data Warehouse a partire dalle sorgenti operazionali e migliorare la qualità dei dati.

Grazie per l'attenzione!