Università di Parma
Dipartimento di Ingegneria e Architettura

Introduzione all'Intelligenza Artificiale

Big Data & Business Intelligence

A.A. 2022/2023

Corso di «Introduzione all'Intelligenza Artificiale»
Corso di «Big Data & Business Intelligence»

Business Intelligence: generalità

Monica Mordonini (monica.mordonini@unipr.it)

SOMMARIO


RAW DATA
EXTRACT
TRANSFORM
LOAD
DATA WAREHOUSE
FORECASTING
Predicting FUTURE customer behaviour
BUSINESS INTELLIGENCE
Useful
DATA MINING
Searching for hidden patterns
Decision-making support
This way
That way
Improve STRATEGY
Spot NEW OPPORTUNITIES
Graphical ANALYSIS

Sommario

- ✓ Business Intelligence
  - Introduzione
  - Il valore della conoscenza
  - Le sfide della Business Intelligence (BI)
- ✓ I dati
  - La materia prima della BI
  - Big Data

- ➢ Business Intelligence vs Data Science

LA BUSINESS INTELLIGENCE:
LEGGERE I DATI PER GUIDARE LE DECISIONI

Il termine è stato utilizzato per la prima volta dall'inventore tedesco Hans Peter Luhn nel 1958, quando lavorava per IBM

Business Intelligence

- ***Fa uso di flussi di dati*** di qualsiasi dimensione per analizzarli e visualizzare **informazioni cruciali**, in relazione all'utilizzo che se ne intende fare.
- Le aziende, a prescindere dalle dimensioni, possono beneficiare di ciò che la BI ha da offrire:
  - informarsi sulle opzioni disponibili aiuta a scoprire la soluzione che può economicamente, efficientemente e abilmente aggiungere valore alla organizzazione su base coerente.

- ***La BI cerca di individuare e rendere percepibili i segnali nascosti***, **contestualizzandoli in un quadro** che permetta di interpretare il significato dei dati attraverso una visione approfondita del passato e del presente;
  - in tal modo è possibile identificare gli schemi che determinano comportamento e prestazioni delle funzioni aziendali.

Business Intelligence

- In sostanza, con la business intelligence è possibile:
  - Raccogliere i dati relativi al rendimento di un'attività;
  - Classificare le informazioni in pattern predefiniti;
  - Analizzare i risultati ottenuti;
  - Realizzare modelli predittivi
  - ➢ Infatti, facendo ricorso alla business intelligence non solo è possibile dare forma a tutte le informazioni in possesso della società, imparando a conoscere e a capire le dinamiche specifiche della stessa, ma anche sfruttarle per predire trend e richieste future dei consumatori, aumentando le vendite e diversificando la propria offerta.

Business Intelligence

La business intelligence si avvale di:

A. piattaforme dedicate, ovvero di programmi informatici che permettono di organizzare tutte le informazioni precedentemente raccolte ed immesse nel sistema in modelli ordinati.
B. modelli matematici, statistici e metodi di analisi approfonditi
   a. Metodo scientifico:
   b. Raccolta dei dati;
   c. Analisi ed elaborazione dei dati inseriti;
C. Verifica dei risultati.
D. Esperti in materia.

***Business intelligence analyst**: figura in grado di seguire l'intero processo dal data mining nel patrimonio aziendale all'analisi approfondita, fino alla restituzione finale dei risultati in informazioni strutturate, da ricavare per le decisioni aziendali.*

Business Intelligence

- Un processo di BI digerisce i flussi di dati in un modo che è più facile da comprendere e indica più chiaramente le azioni necessarie basate su tali dati:
- Si tratta quinid di un percorso di valutazione supportato da strumenti di Data Visualization che consentono di guidare le decisioni verso una maggiore efficienza e maggiori profitti.

- *La BI è dunque la chiave di lettura ottimale per capire cosa è successo e cosa sta succedendo, offrendo soluzioni che aiutano a identificare schemi di comportamento significativi e correlazioni tra le variabili entro un complesso insieme di dati, strutturati e non strutturati, storici, attuali e potenziali.*

Business Intelligence

- Per avere una visione globale di tutte le aree di attività dell'azienda, è necessario organizzare tuti i dati interni ed esterni all'azienda in un unico rapporto.
- Strumenti più utili per il successo di un'azienda prodotti da processi BI sono
  - I report
  - Le "**dahsboard**" (cruscotti) ovvero delle bacheche che al semplice dato aggiungono rappresentazioni grafiche, per rendere più semplice ed efficace il processo di apprendimento.
  - Le dashboard mostrano i **KPI**, Key Performance Indicator, o meglio gli indicatori che svelano tanto i punti forti quanto quelli deboli della azienda, come ad esempio le entrate dalle diverse attività, ma anche le scorte oppure l'engagement sui social network.

Business Intelligence (BI)

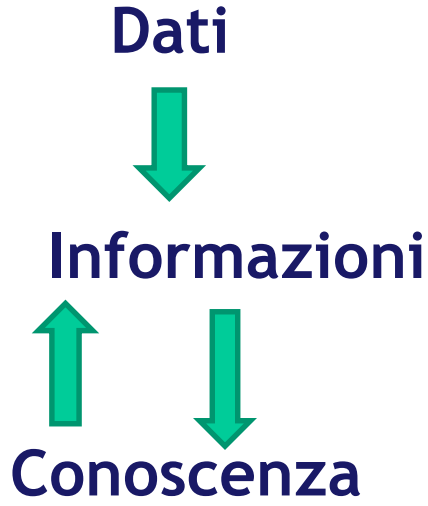
- Rappresenta l'insieme dei processi, delle tecnologie e gli strumenti necessari per trasformare
  1. i dati in informazioni
  2. e informazioni in conoscenza
  3. e conoscenza in piani che promuovono azioni redditizie.

- La BI comprende: data warehousing, analisi aziendale e gestione della conoscenza.
- Anche l'OLAP (OnLine Analytical Processing) è una parte di BI e rappresenta un approccio per rispondere rapidamente alle domande analitiche multidimensionali

Business Intelligence: significa **trasformare** i *dati grezzi* in informazioni **utilizzabili**, distribuire e condividere le informazioni, creando una conoscenza collettiva della propria impresa.

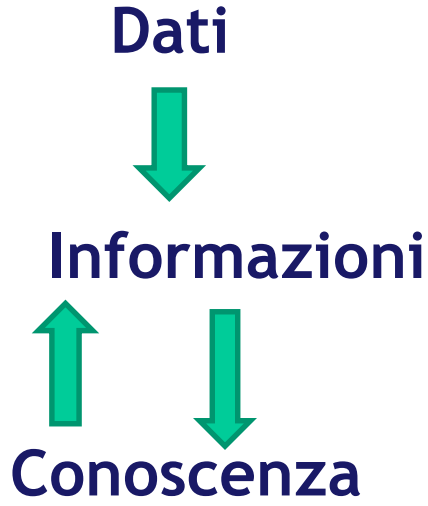
Business Intelligence (BI)

- Dati: sono nozioni grezze e incomplete. Rappresentano un fatto o un oggetto della realtà
  - Mario, 275
- Informazioni: hanno origine dai dati aggiungendo ad essi "valore" attraverso processi di analisi e sintesi, contestualizzazione, calcolo, categorizzazione.
  - Se i due dati vengono forniti in risposta alla domanda «A chi mi devo rivolgere per questo problema , quale è il suo interno? I dati grezzi sono ***interpretati***
- Conoscenza: trasferimento delle informazioni all'interno dell'organizzazione. Identificazione di relazione causa-effetto tra informazioni attraverso esperienza, relazioni sociali etc.
  - L'informazione che Mario il cui interno è 275 sa risolvere quel problema arricchisce la mia conoscenza su quella organizzazione



Business Intelligence (BI)

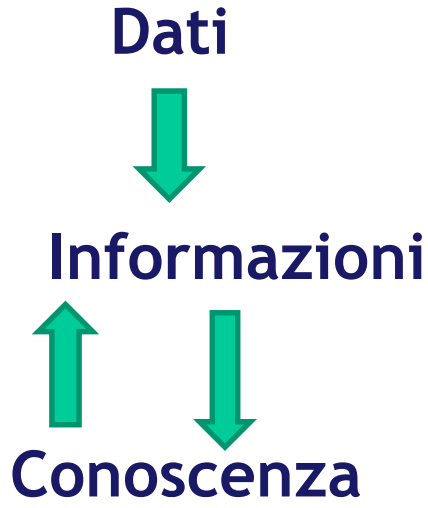
- E’ un processo analitico che trasforma i dati in informazioni a supporto della presa di decisioni ottimizzato da un insieme di tecnologie
- **Con il fine di Migliorare** i **processi decisionali**, di **comunicazione** e ***coordinamento*** delle interdipendenze aziendali, razionalizzare e ottimizzare il processo di **creazione**, **gestione**, **diffusione** e ***condivisione*** *della conoscenza*



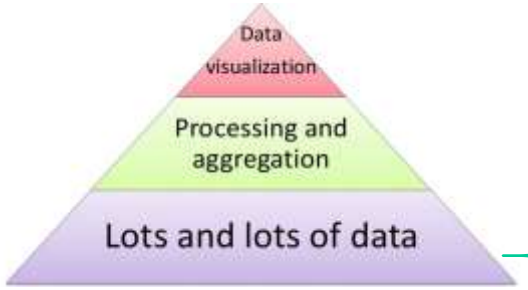
(Grothe e Gentsch, 2000; Davenport e Prusak, 1998)

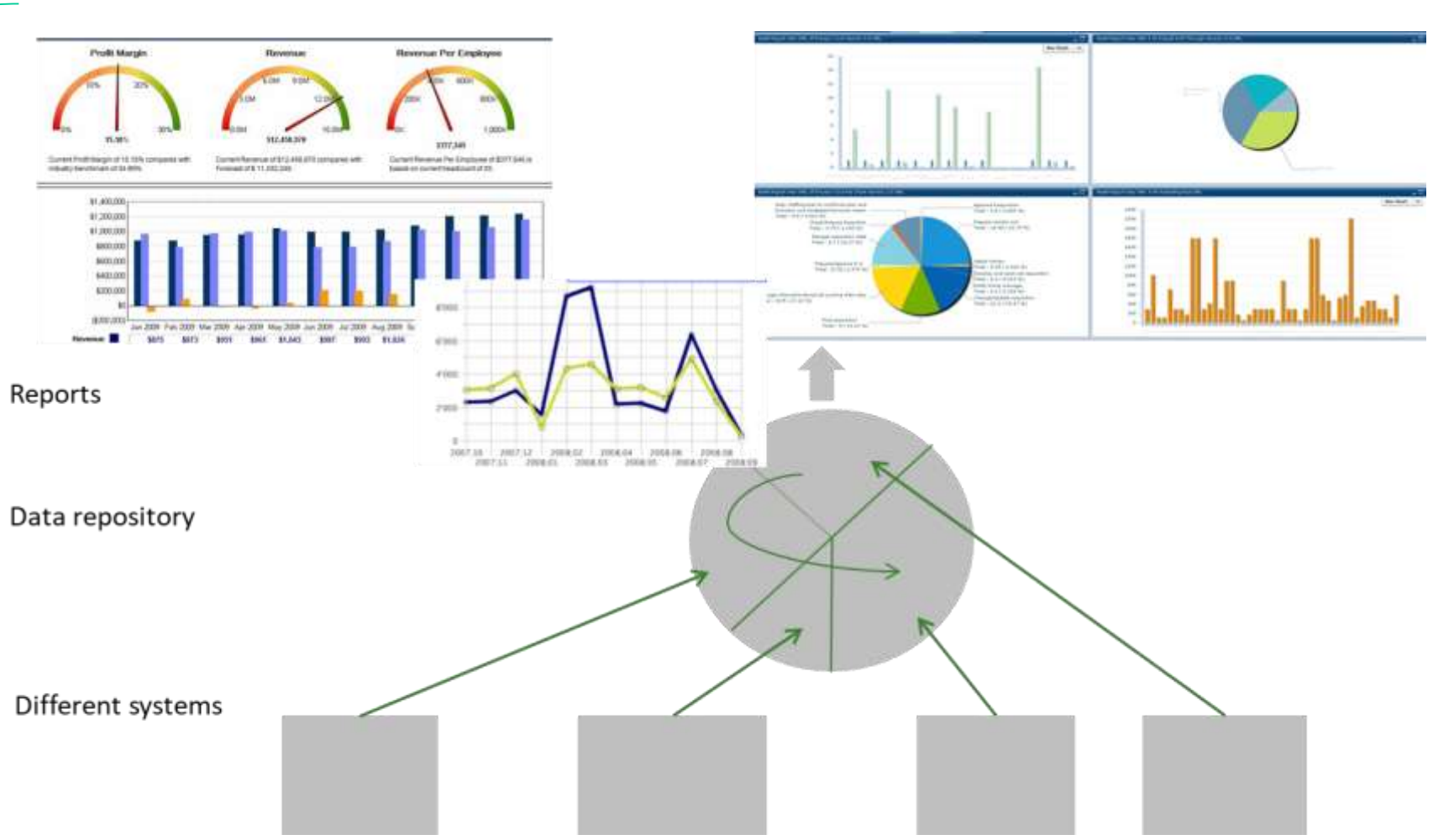
Business Intelligence (BI) & Business management issues

- " Abbiamo montagne di dati in questa azienda, ma non possiamo accedervi".
  - In questa momento: abbiamo montagne di dati ma non riusciamo a fare analisi predittiva
- "Devi rendere facile per gli uomini d'affari ottenere direttamente i dati."
- "Mostrami solo ciò che è importante".
- "Mi fa impazzire il fatto che due persone presentino le stesse metriche di business ad una riunione, ma con numeri diversi. "
- "Vogliamo che le persone utilizzino le informazioni per supportare più processi decisionali basati sui fatti".
- …



Business Intelligence (BI)





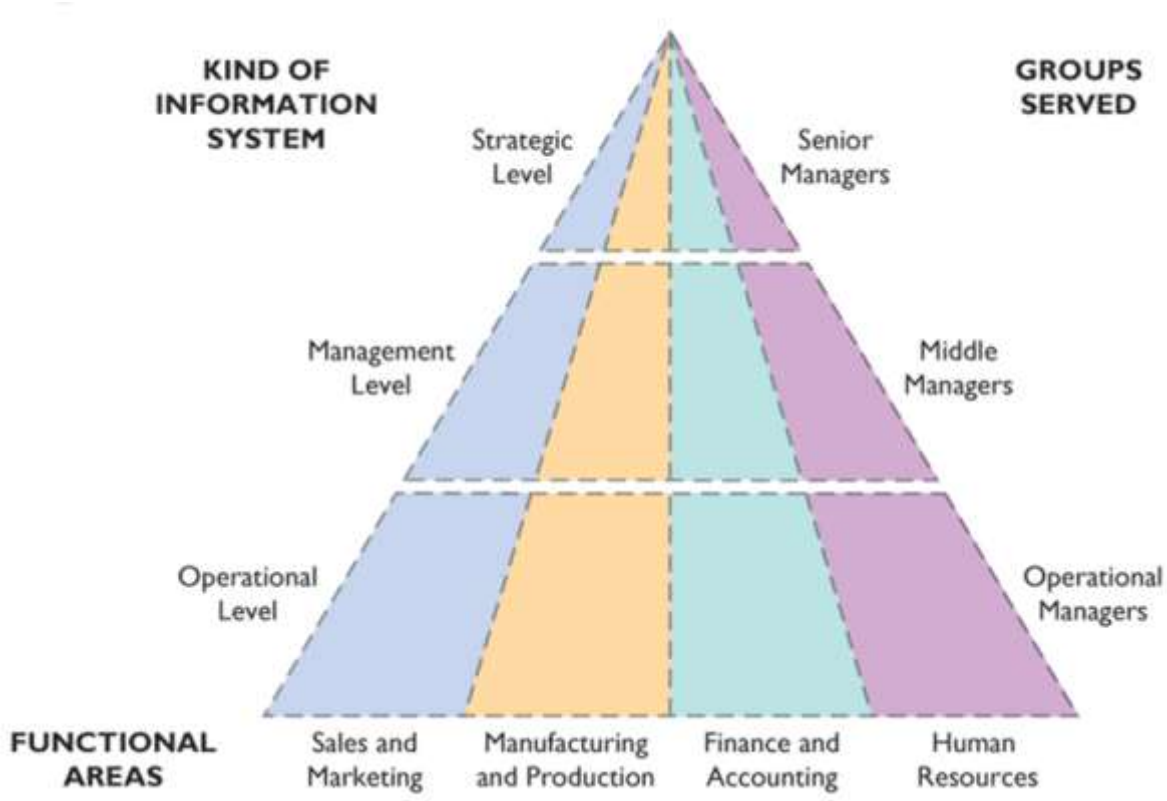
Business Intelligence (BI) e Knowledge management

- ***Knowledge Management*** (*Gestione della conoscenza*) significa:
  - **Creazione**, **raccolta** e **classificazione** di informazioni
  - provenienti da **varie fonti** **di dati** (fonti interne, Web, sistemi ERP)
  - che vengono quindi **distribuite** ai **vari utenti** sulla *base degli specifici interessi*
  - **tramite mezzi** e strumenti diversi.

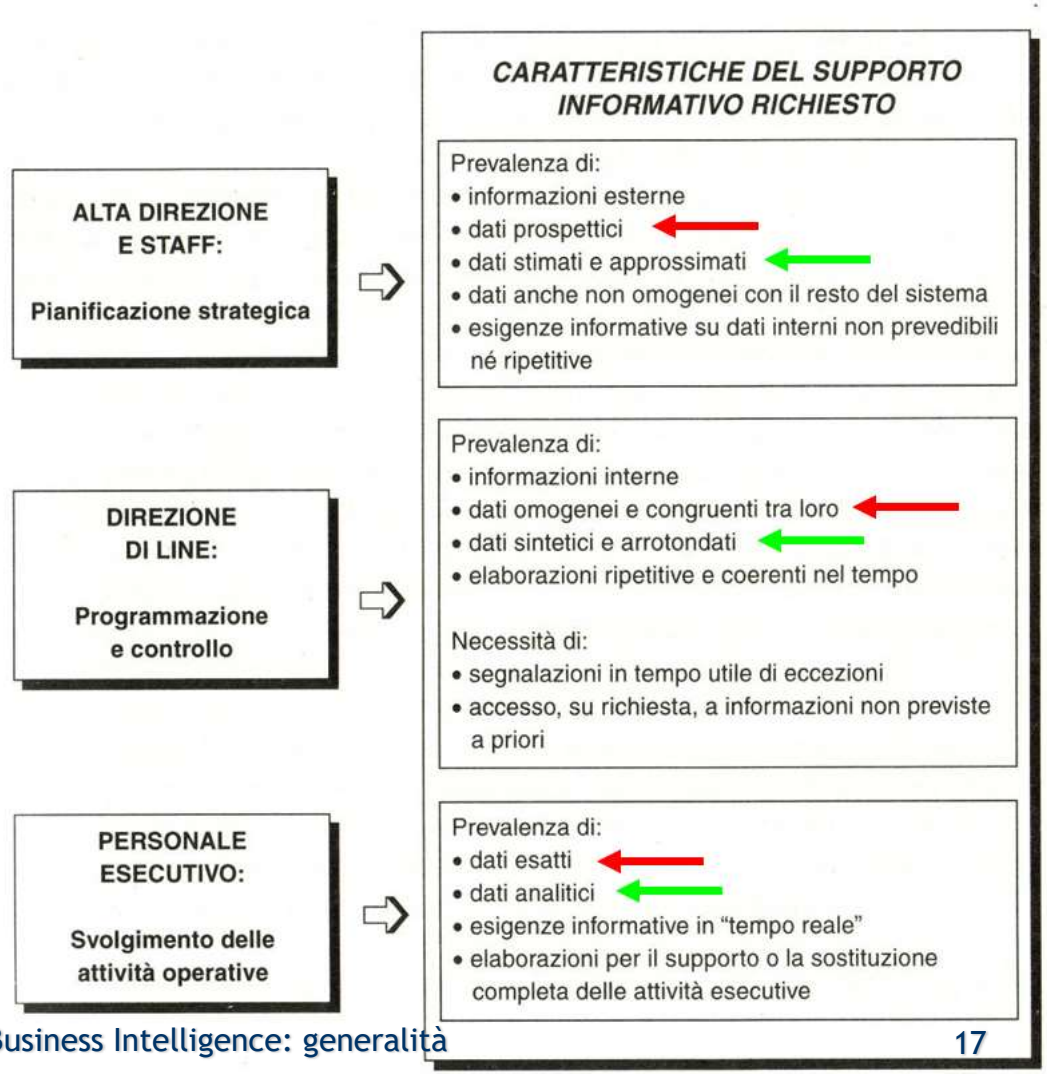
- *Una piattaforma di Knowledge Management, quindi, **raccoglie**, **organizza**, **distribuisce** e rende **facilmente accessibili** le **conoscenze aziendali** a chi ne ha bisogno, nel momento e nel contesto in cui **servono***

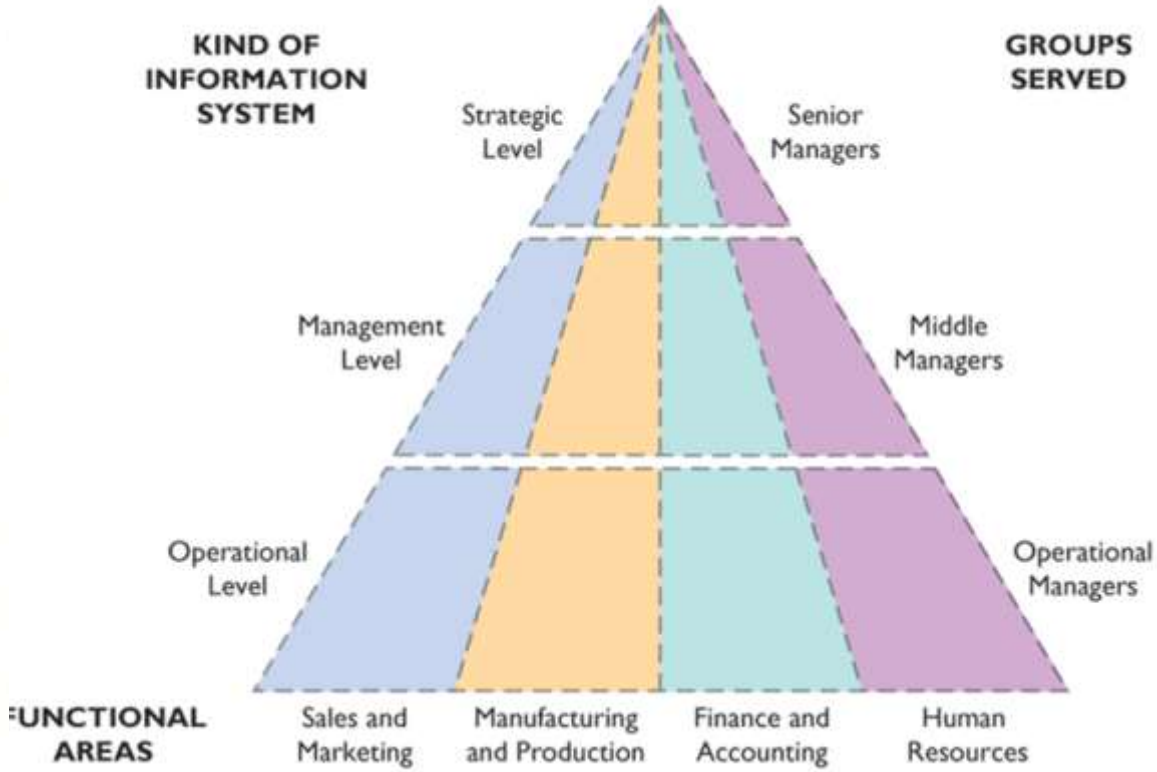
Gli Utenti della Business Intelligence

- Gli utenti di BI possono essere suddivisi in:
  1. Utenti di alto livello con la necessità di avere una visione ampia e capacità di analisi limitate (executives and business decision makers)
  2. Gli utenti specializzati che eseguono analisi dettagliate dei dati e necessitano di strumenti potenti
  3. Lavoratori che necessitano di report di base con possibili funzionalità analitiche (rappresentanti – information workers)
  4. I lavoratori di una azienda che non necessitano di funzioni analitiche e che hanno integrato la BI nei sistemi che utilizzano senza rendersene conto è BI



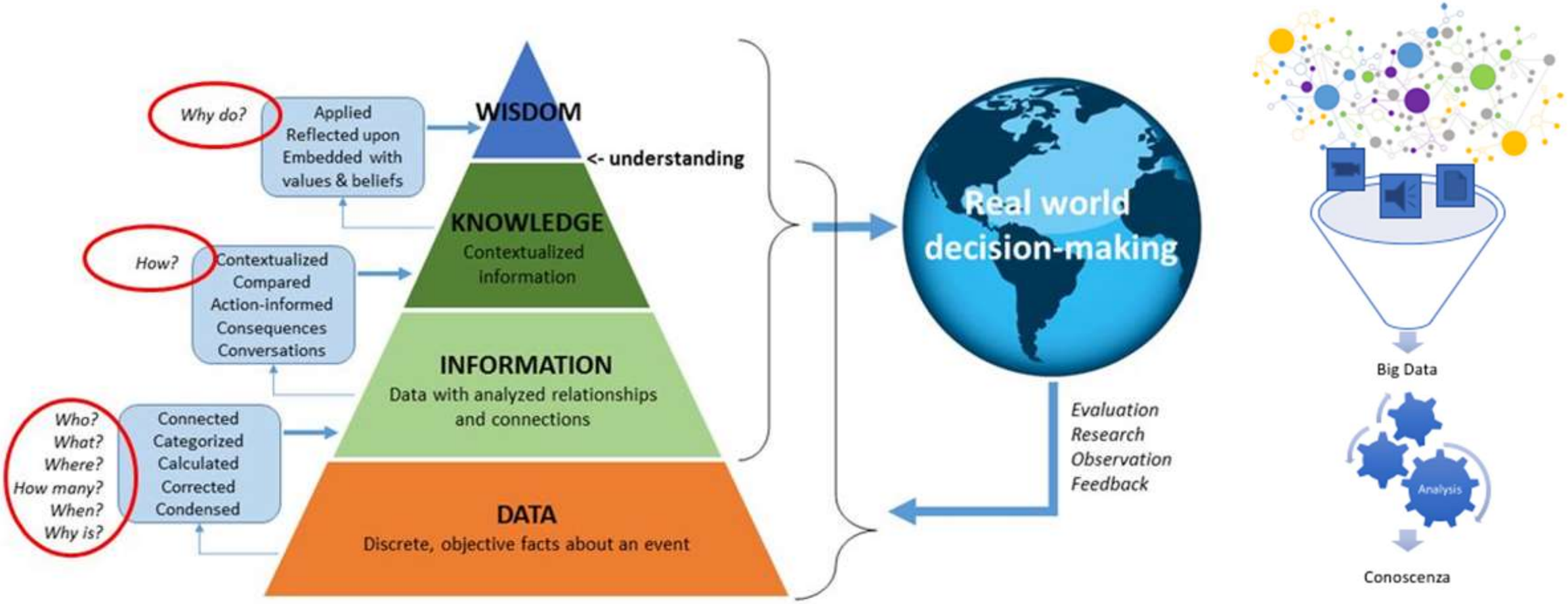
Gli Utenti della Business Intelligence





IL VALORE DELLA CONOSCENZA

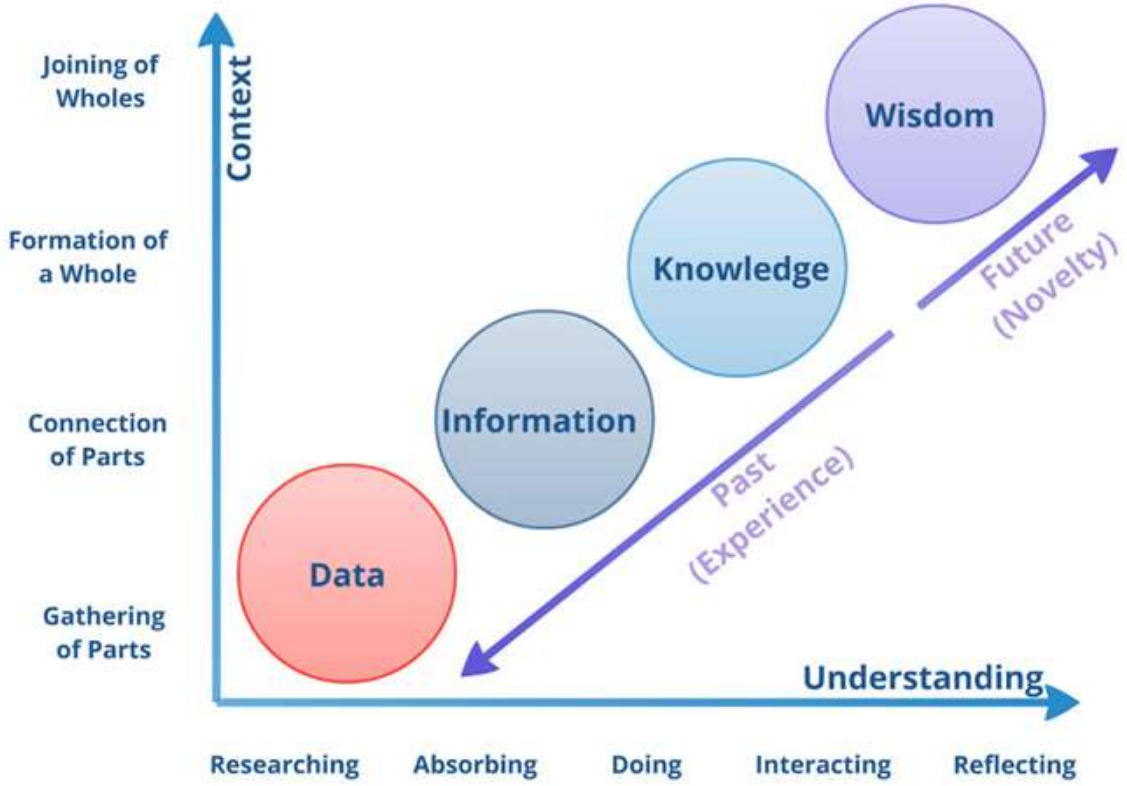
La piramide di KIWD



Il valore della Conoscenza

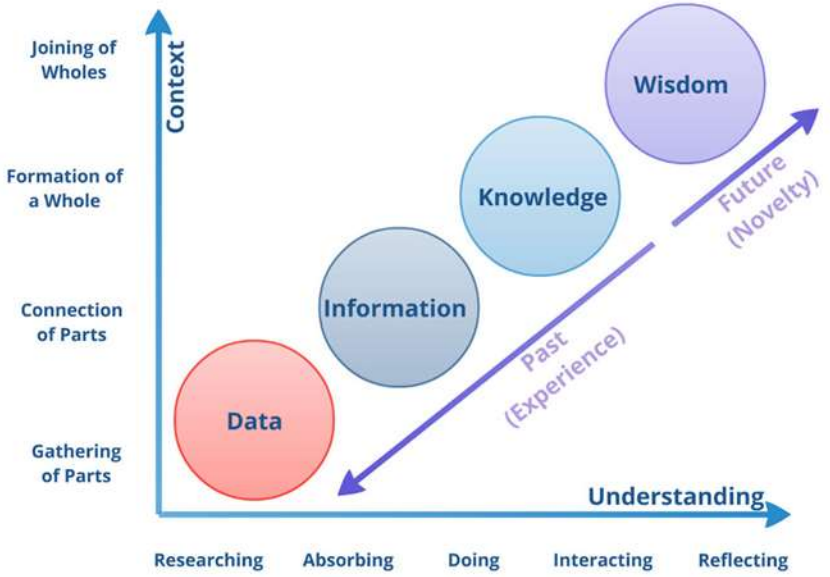
- ***Dati*** sono nozioni grezze, spesso risultanti da una osservazione. In informatica, un dato è un elemento informativo costituito da simboli.
  - Es rossi,mario,180 …
- ***Informazioni:*** rappresentazione dei fatti (dati) organizzati in modo da essere comprensibili e significativi per l'utente destinatario. Sono il risultato dell'elaborazione di più dati a cui viene aggiunto un "valore" attraverso processi di analisi e sintesi, contestualizzazione, calcolo, categorizzazione.

  - Es. se in un db rossi è un cognome, 180 un'altezza in cm.
  - Si può trovare l'informazione Rossi è alto 180cm



Il valore della Conoscenza

- ***Dati*** sono nozioni grezze, spesso risultanti da una osservazione. In informatica, un dato è un elemento informativo costituito da simboli.
  - Es rossi,mario,180 …
- ***Informazioni:*** rappresentazione dei fatti (dati) organizzati in modo da essere comprensibili e significativi per l'utente destinatario . Sono il risultato dell'elaborazione di più dati a cui viene aggiunto un "valore" attraverso processi di analisi e sintesi, contestualizzazione, calcolo, categorizzazione.
  - Es. se in un db rossi è un cognome, 180 un'altezza in cm.
  - Si può trovare l'informazione Rossi è alto 180cm



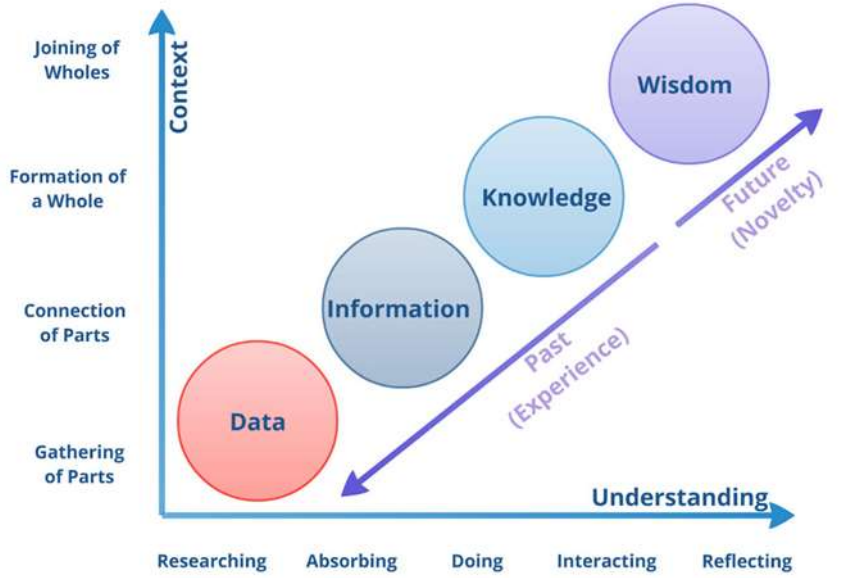
Conoscenza: trasferimento delle informazioni all'interno dell'organizzazione, cioè una conoscenza è il collegamento fra più informazioni quali p.e., l'identificazione di relazione causa-effetto tra informazioni attraverso esperienza, relazioni sociali etc.

- Rappresenta la consapevolezza e la comprensione di fatti, verità o informazioni ottenuti attraverso l'esperienza o l'apprendimento
- La conoscenza esiste solo in quanto esiste un'intelligenza che possa utilizzarla

Il valore della Conoscenza

Conoscenza: trasferimento delle informazioni all'interno dell'organizzazione, cioè una conoscenza è il collegamento fra più informazioni quali p.e., l'identificazione di relazione causa-effetto tra informazioni attraverso esperienza, relazioni sociali etc.

- Rappresenta la consapevolezza e la comprensione di fatti, verità o informazioni ottenuti attraverso l'esperienza o l'apprendimento
- La conoscenza esiste solo in quanto esiste un'intelligenza che possa utilizzarla



La ***saggezza*** è la capacità di aumentare l'efficacia. La saggezza aggiunge valore, il che richiede la funzione mentale che chiamiamo giudizio. I valori etici ed estetici che questo implica sono intrinseci all'attore e sono unici e personali. (*Russell Ackoff*)

L'Analisi. Ciò che genera conoscenza

1. ***Descriptive Analytics***, l'insieme di strumenti orientati a descrivere la situazione attuale e passata dei processi aziendali e/o aree funzionali. Tali strumenti permettono di accedere ai dati secondo viste logiche flessibili e di visualizzare in modo sintetico e grafico i principali indicatori di prestazione;
2. ***Predictive Analytics***, strumenti avanzati che effettuano l'analisi dei dati per rispondere a domande relative a cosa potrebbe accadere nel futuro; sono caratterizzati da tecniche matematiche quali regressione, forecasting, modelli predittivi, ecc;
3. ***Prescriptive Analytics***, applicazioni big data avanzate che, insieme all'analisi dei dati, sono capaci di proporre al decision maker soluzioni operative/strategiche sulla base delle analisi svolte;
4. ***Automated Analytics***, capaci di implementare autonomamente l'azione proposta secondo il risultato delle analisi svolte.

https://www.data-ware.it/competenze/bigdata/

LE SFIDE DELLA BI
DATI DI QUALITÀ PER ESSERE AFFIDABILI

Le sfide della Business Intelligence (BI)

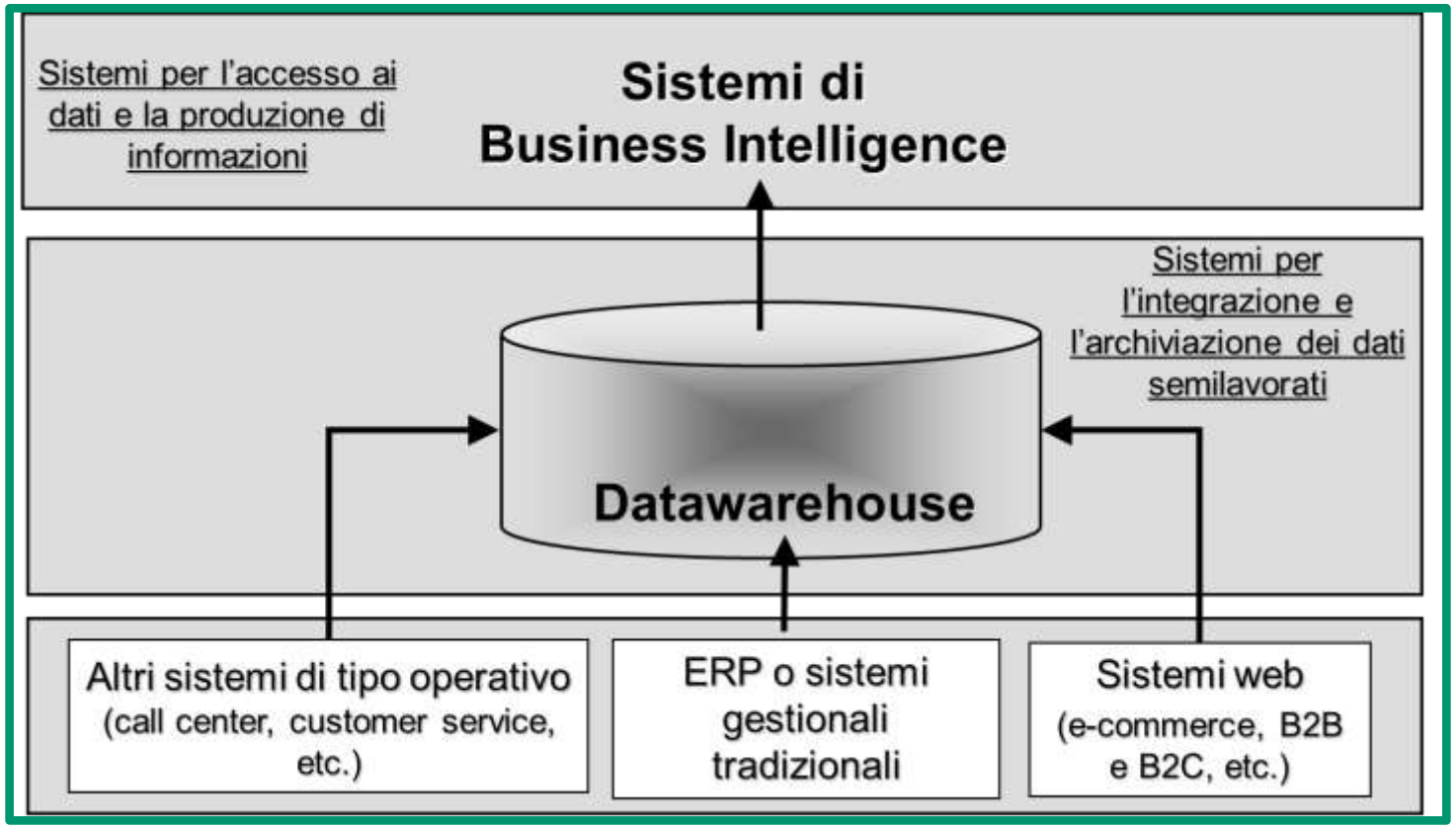
- Esistono diversi problemi inerenti a qualsiasi progetto di BI:
  - I dati stanno in più posti
  - I dati non sono formattati per supportare analisi complesse
  - Diversi tipi di lavoratori hanno esigenze di operare su dati diversi
  - Quali dati dovrebbero essere esaminati e in quali dettagli
  - In che modo gli utenti interagiranno con tali dati

Il primo passo: il consolidamento dei dati

- Il processo di consolidamento dei dati significa
  - spostarli,
  - renderli coerenti
  - e ripulirli il più possibile
- I dati vengono spesso archiviati
  - in diversi formati
  - sono spesso incoerenti tra le fonti
  - potrebbero essere sporchi cioè con valori internamente incoerenti o mancanti

Disparate Data

- Dati possono stare in luoghi diversi e formati diversi:
- Database relazionali
- File XML
- Fogli di calcolo (Excel)
- …
- I dati potrebbero anche trovarsi in database su diversi sistemi operativi e piattaforme hardware



Inconsistent Data

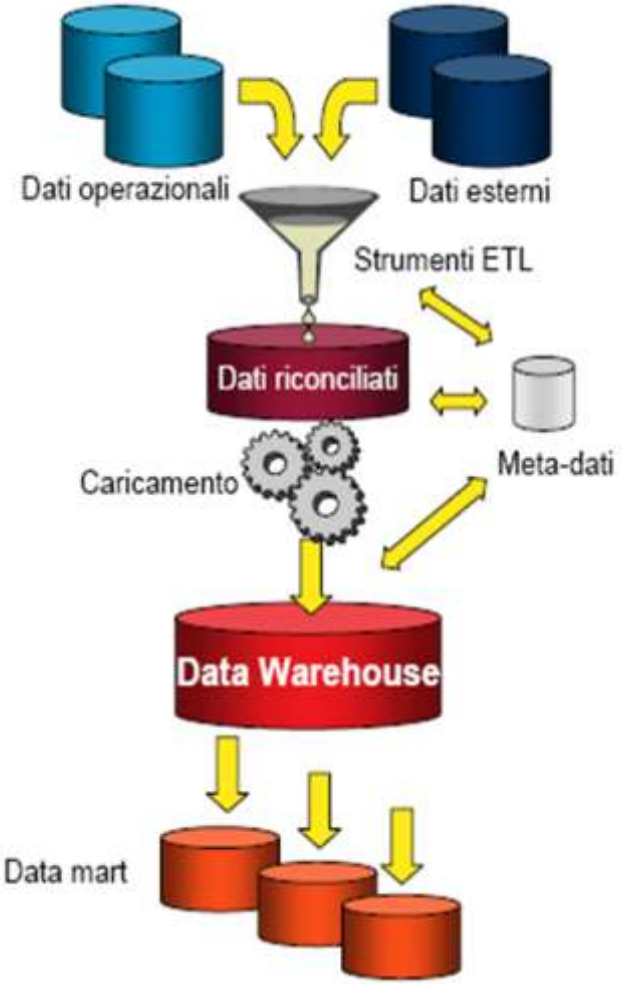
- I dati potrebbero essere incoerenti
- Due mappe potrebbero avere nomi diversi per indicare la stessa parte fisica
- Per rappresentare True e False, un sistema può usare 1 e 0, mentre un altro sistema può usare T e F
- I dati memorizzati in paesi diversi probabilmente memorizzeranno le vendite nella loro valuta locale
  - Queste vendite devono essere convertite in una valuta comune

Creare Dati di Qualità

- I dati puliti facilitano analisi più accurate
- Virtualmente tutte i sistemi possono avere dati sporchi:
  - I dati sporchi sono valori errati che entrano in un sistema. Può essere un semplice refuso quando si immette un numero, ma spesso è un valore errato inserito in un campo di testo in formato libero.
    - Ad esempio possiamo scrivere male il nome di una città
- La pulizia dei dati non validi può richiedere procedure estese che richiedono l'aggiornamento ogni volta che viene rilevato un nuovo valore errato.
- Le organizzazioni possono anche utilizzare algoritmi di data mining per aiutare a ripulire i dati;
  - ad esempio, può essere utilizzata una ricerca fuzzy per aiutare a confrontare i valori di testo che sono simili.

Extraction, Transformation, and Loading (ETL)

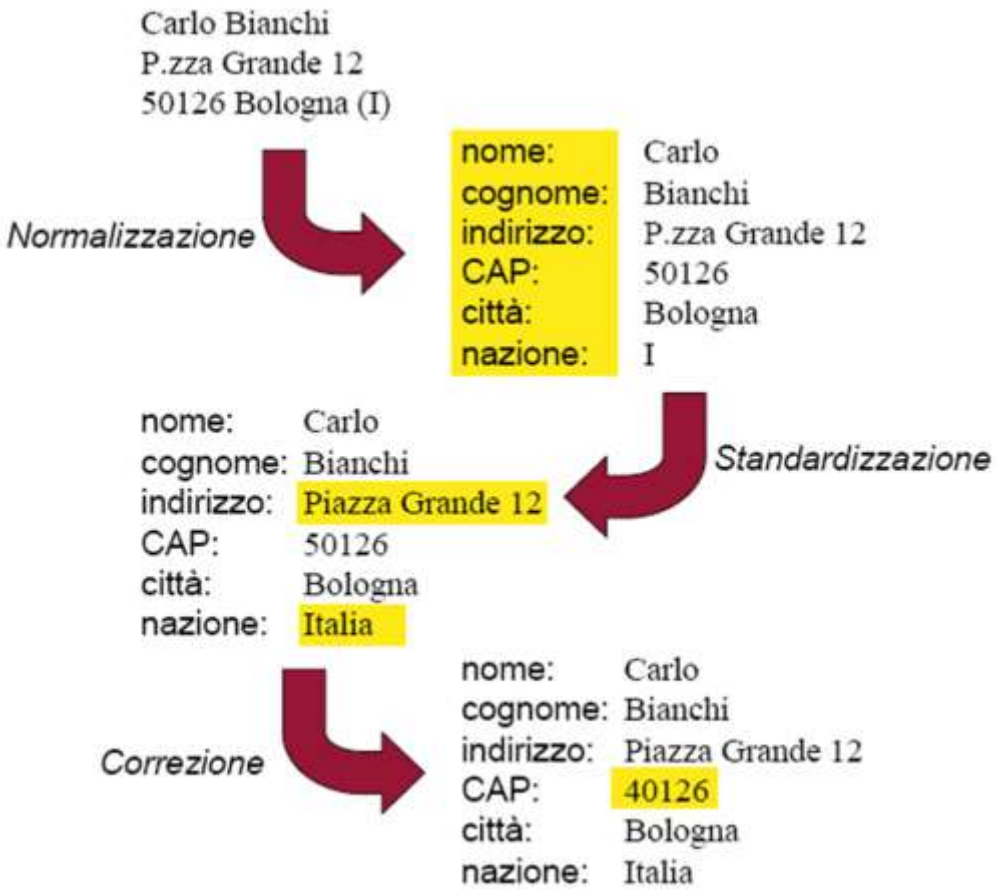
- Il processo di spostamento dei dati dai suoi sistemi di origine, il consolidamento in una posizione centrale e la correzione delle incoerenze dei dati si chiama Estrazione, Trasformazione e Caricamento dei Dati o ETL.
  - La fase di estrazione estrae i dati dalle varie fonti.
  - I dati vengono quindi trasformati o resi coerenti (i valori "Veri" sono tutti impostati sullo stesso valore, le valute vengono convertite e così via).
    - In questo modo aumentandone la loro qualità
  - Infine, viene caricato in un repository di dati (spesso chiamato factory di dati o **data warehouse**).



- ***Lo sviluppo del processo ETL spesso consuma l'80% dei tempi di sviluppo totale***

Extraction, Transformation, and Loading (ETL)

- Esempio di funzionamento



DataWareHouse (DWH)

- **Anello di collegamento** tra i dati, le applicazioni e i sistemi informativi di tipo operativo e i sistemi informativi manageriali di supporto alle attività di controllo e di decisione
- Costruito per supportare i processi decisionali
- Può essere definito come:
  - **una raccolta di dati integrata:** *da fonti transazionali o esterne diverse*
  - **subject oriented:** *organizzati per argomento (non per applicazione)*
  - **time variant:** *organizzati per riferimento temporale (fotografia)*
  - **non-volatile:** *non modificabili:* ***sola lettura***

DataWareHouse (DWH): *caratteristiche*

- **Integrazione**: requisito fondamentale di un DWH è l'integrazione della raccolta dati.
  - Nel DWH confluiscono dati provenienti da più sistemi transazionali e da fonti esterne.
  - L'obiettivo dell'integrazione può essere raggiunto percorrendo differenti strade: mediante l'utilizzo di metodi di codifica uniformi, mediante il perseguimento di una omogeneità semantica di tutte le variabili, mediante l'utilizzo delle stesse unità di misura

- **Subject oriented**: perché il DWH è orientato a temi specifici dell'azienda piuttosto che alle applicazioni o alle funzioni.
  - In un DWH i dati vengono archiviati in modo che possano essere facilmente letti o elaborati dagli utenti.
  - L'obiettivo, quindi, NON è più quello di minimizzare la ridondanza mediante la normalizzazione MA quello di fornire dati che abbiano una struttura in grado di favorire la produzione di informazioni. Si passa dalla progettazione per funzioni alla modellazione dei dati al fine di consentire una visione multidimensionale degli stessi.

DataWareHouse (DWH): *caratteristiche*

- **Time variant:** i dati archiviati all'interno di un DWH hanno un orizzonte temporale molto più esteso rispetto a quelli archiviati in un sistema operazionale.
  - Ciò, tuttavia, comporta che i dati contenuti in un DWH sono aggiornati fino ad una certa data, che nella maggior parte dei casi, è antecedente a quella in cui l'utente interroga il sistema.
  - Situazione del tutto differente, al contrario, si manifesta in un transazionale in cui i dati corrispondono sempre ad una situazione costantemente aggiornata che tuttavia non fornisce un quadro storico del fenomeno analizzato
- **Non-volatile:** tale caratteristica indica la non modificabilità dei dati contenuti nel DWH che consente accessi in sola lettura.
  - Comporta, inoltre, una maggiore semplicità di progettazione del database rispetto a quella di un database relazionale che supporta una applicazione transazionale.
  - In tale contesto non si fronteggiano le possibili anomalie dovute agli aggiornamenti e tanto meno si ricorre a strumenti complessi per gestire l'integrità referenziale o per bloccare record a cui possono accedere altri utenti in fase di aggiornamento

Asking a BI Question

Per recuperare i dati da un «data warehouse», si deve conoscere i principali componenti di una domanda:

- Gli umani tendono a pensare in modo multidimensionale, anche se non se ne rendono conto
  - Spesso vogliamo vedere un valore particolare in un determinato contesto
- Generalmente gli utenti chiedono di vedere "qualcosa" (vendite, spese, numero di unità, ecc.) segmentati "da" alcune cose (tempo, luogo, venditore, ecc.).
  - Mostrami le vendite per mese per prodotto per il Nord America
  - "Cosa" vuoi vedere (le vendite in questo caso) è una misura
  - Come vuoi vederlo (mese, prodotto e Nord America) è una dimensione

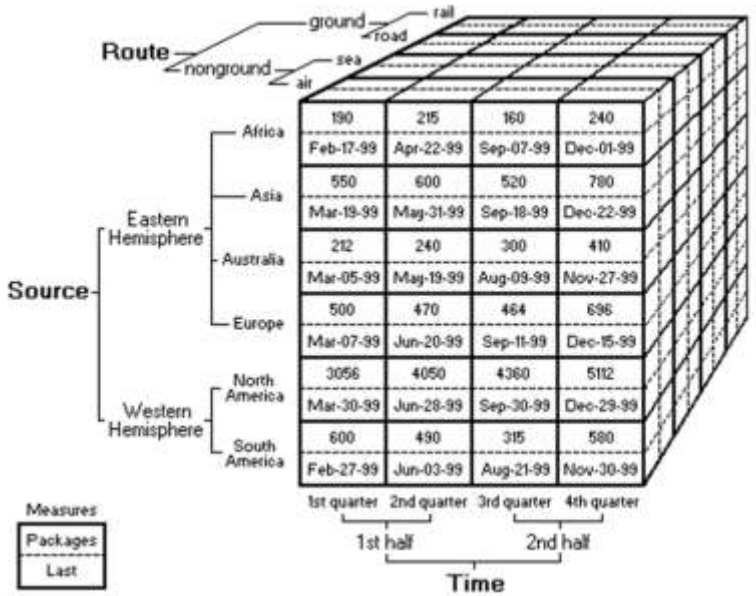
❑ Queste misure e dimensioni possono essere memorizzate in cubi.

Asking a BI Question

- Un **cubo** è il componente base di un data warehouse.
- Un warehouse può contenere uno o più cubi.
- Un cubo è una struttura multidimensionale che contiene dati basati su dimensioni in cui gli utenti accedono ai dati «navigando» attraverso le varie dimensioni.

Esempio: un data warehouse per un'azienda di spedizioni come FedEx o UPS.

- Vediamo tre dimensioni: Tempo, Sorgente e Rotta.
- Ogni intersezione di Tempo, Sorgente e Rotta è una cella.
- All'interno di quella cella ci sono due misure: il numero di pacchetti e la data di spedizione.



- Questo è molto diverso da una configurazione relazionale: i database relazionali sono bidimensionali (righe e colonne) e ogni cella può avere un solo valore.

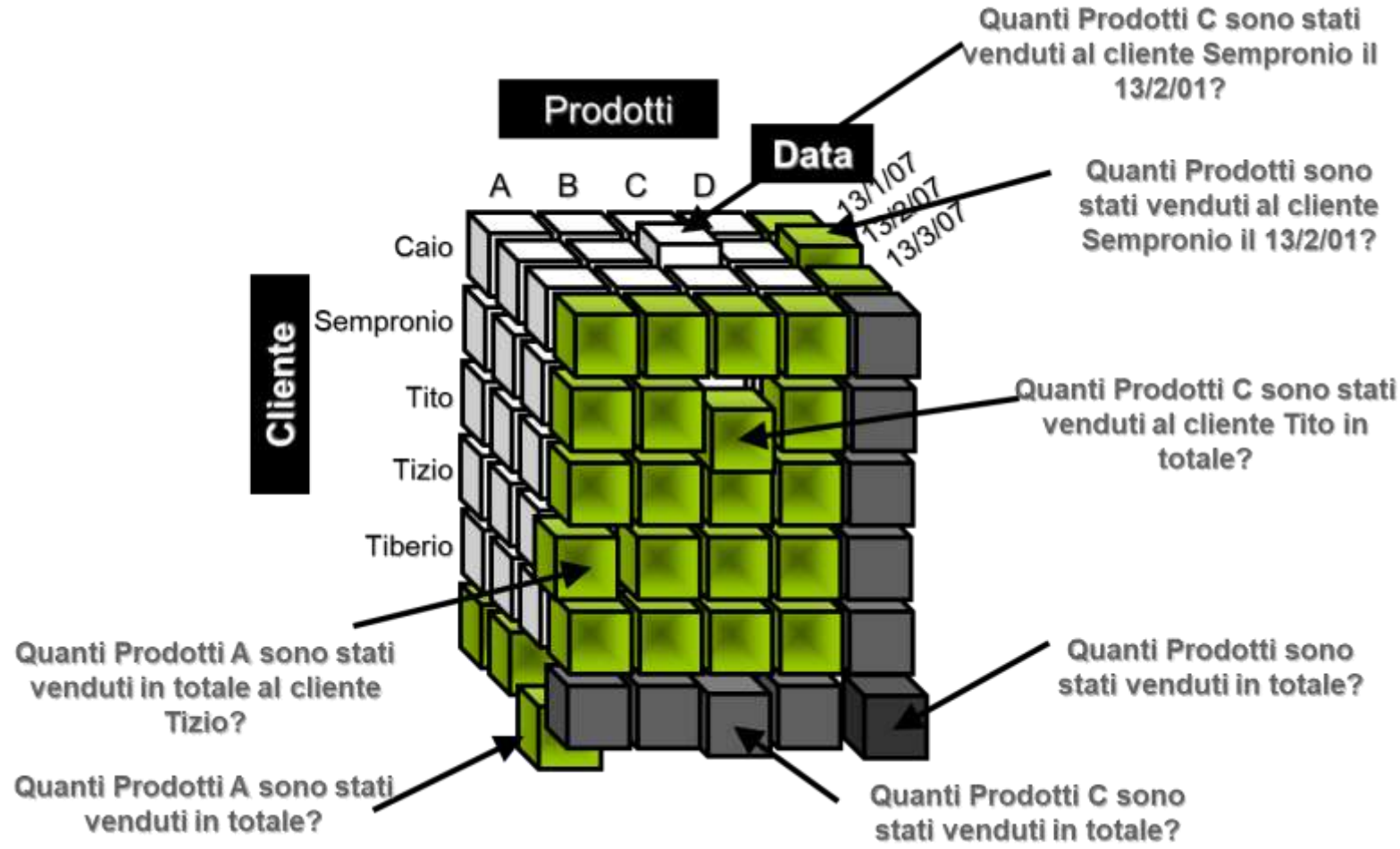
Asking a BI Question

- Le **misure** sono "cosa" la gente vuole vedere. Sono quasi sempre numerici.
  - Vendite in dollari, profitti, spese, data dell'ultima spedizione, inventario etc …
- Le **dimensioni** sono come si vuole vedere i dati : per tempo, geografia, prodotto, account, dipendente, ...
- Le dimensioni sono costituite da **attributi** che rappresentano diversi modi di guardare qualcosa in una dimensione.
  - Ad esempio, in una dimensione di prodotto, un utente potrebbe voler confrontare le vendite di un prodotto per colore; il prodotto rosso vende meglio del blu? Dipende da quale area del paese viene esaminata? Molte delle colonne in una tabella relazionale possono diventare attributi in un magazzino.
  - Una dimensione temporale può avere un attributo mese e un attributo anno e così via

- È possibile inserire attributi in una struttura gerarchica per assistere l'analisi dell'utente
  - Ad esempio, una dimensione temporale ha spesso un livello Anno che può essere suddiviso in Quarti. I Quarti possono essere suddivisi in Mesi e infine Giorni.

Asking a BI Question: OLAP

- Analisi dei dati su strutture multidimensionali
- Rapida, flessibile ed efficiente
- Utente sceglie interattivamente le informazioni da visualizzare
- Strumento DECISIONALE



Asking a BI Question: OLAP

MOLAP

Multidimensional OLAP

- è la tipologia **più utilizzata**.
- ha uno **specifico motore** per l'analisi multidimensionale
- crea le "dimensioni" con un misto di dettaglio ed aggregazioni
- scelta migliore per **quantità di dati ridotte**
- **veloce** nel calcolare le aggregazioni e restituire risultati
- crea enormi quantità di dati intermedi

ROLAP

Relational OLAP

- **lavora** direttamente con database relazionali
- i **dati** e le tabelle sono memorizzati come tabelle relazionali
- memorizzare le informazioni di aggregazione
- necessita di **minor spazio** disco e uso di RAM
- ma è il più lento nella fase di creazione

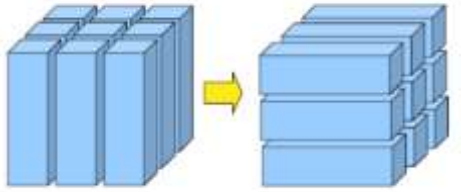
HOLAP

Hybrid OLAP

- **lavora** direttamente con database relazionali
- creato più velocemente dei ROLAP
- è più scalabile di MOLAP.

Asking a BI Question: OLAP

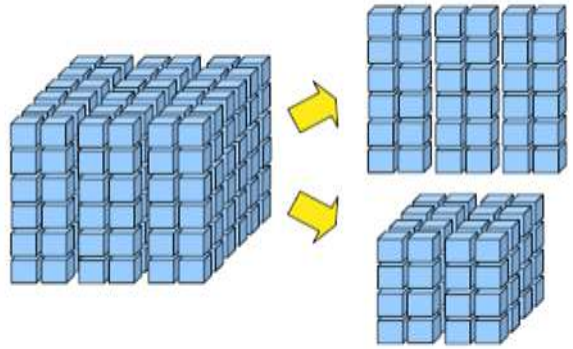
- Principali operazioni:
- **Pivoting**



- è l'operazione di rotazione delle dimensioni di analisi. È un'operazione fondamentale per analizzare totali ottenuti in base a dimensioni diverse o se si vogliono analizzare aggregazioni trasversali;
- La ***tabella pivot*** è la reportistica che risulta da una query OLAP elaborata su dati organizzati all'interno di un ipercubo OLAP.

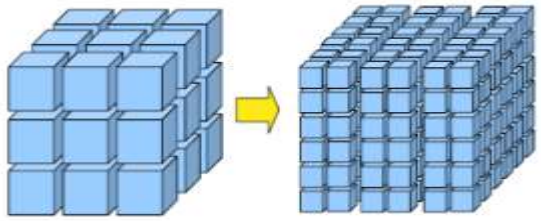
Asking a BI Question: OLAP

- Principali operazioni:
  - estrazione di un subset di informazioni dall'aggregato che si sta analizzando
- **Slicing**:
- Si fissa uno specifico valore per una delle dimensioni del "cubo", estraendo quindi una "fetta" e ottenendo un nuovo cubo con una dimensione in meno rispetto a quello di partenza;
- **Dicing:**
- Si focalizza l'analisi su un sottoinsieme del "cubo" avente particolare interesse per l'analista.



Asking a BI Question: OLAP

- Principali operazioni:
- **Drill-down**

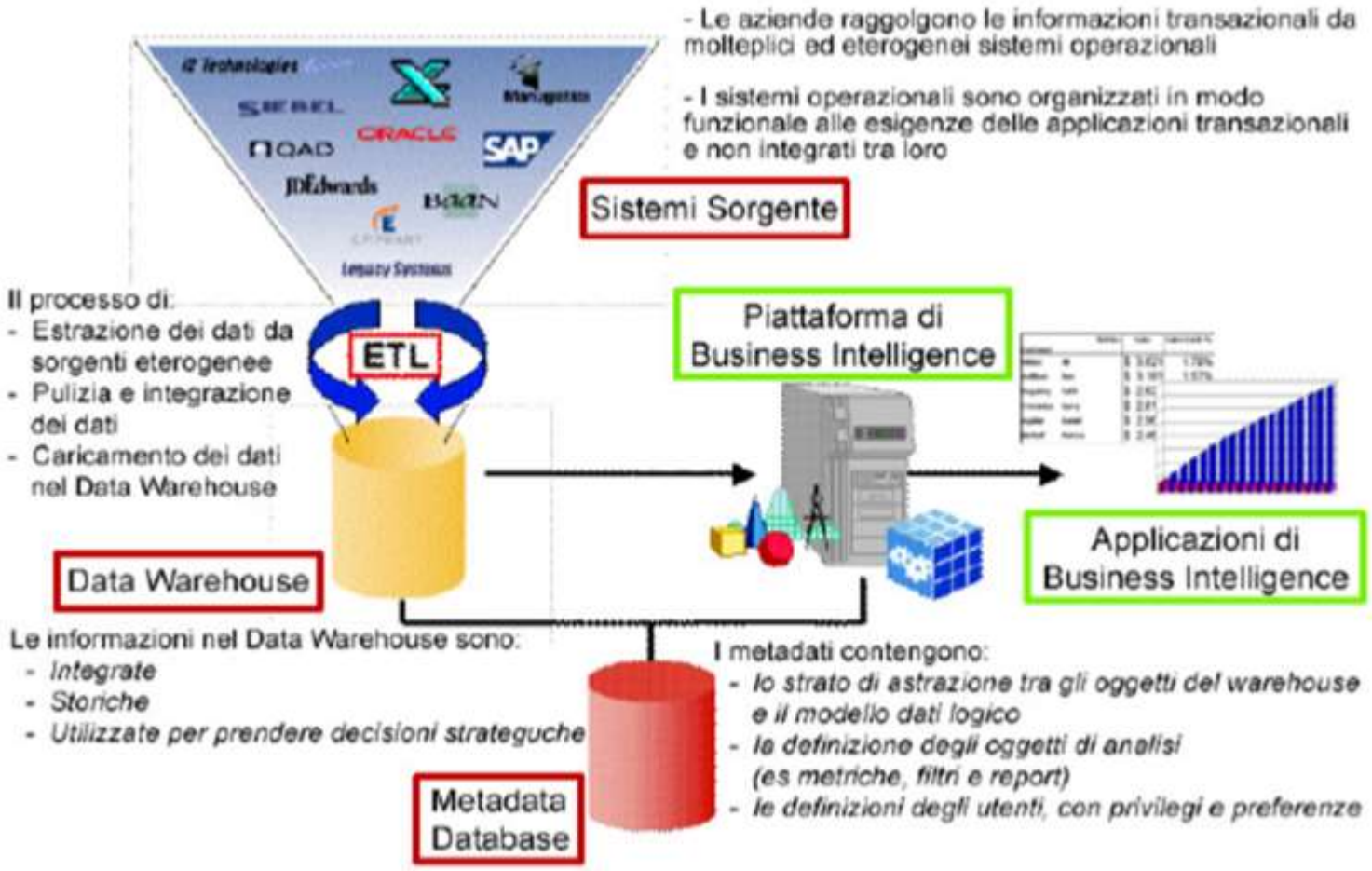


- è l'operazione di "esplosione" del dato nelle sue determinanti. L'operazione di drill-down può essere eseguita seguendo due diversi percorsi: la gerarchia costruita sulla dimensione di analisi (p. es.: passaggio dalla famiglia di prodotti all'insieme dei prodotti che ne fanno parte) oppure la relazione matematica che lega un dato calcolato alle sue determinanti (p. es.: passaggio dal margine al ricavo e costo che lo generano).

Asking a BI Question: Data Mining

- ***Processo di estrazione di conoscenza da banche dati di grandi dimensioni tramite l'applicazione di algoritmi che individuano le associazioni "nascoste" tra le informazioni e le rendono disponibili***
- In altre parole, con data mining si intende l'applicazione di una o più tecniche che consentono l'esplorazione di grandi quantità di dati, con l'obiettivo di individuare le informazioni più significative e di renderle disponibili e direttamente utilizzabili nell'ambito del decision making.
- Serve per:
  - Pianificazione aziendale
  - Ricerche di Mercato
  - Efficacia del marketing
  - Valutazione rischio

Asking a BI Question: Data Mining



Asking a BI Question: Data Mining

- Tipologie di problemi ai quali il DM fornisce una risposta

Problemi	Definizioni
Classificazione	Definizione delle caratteristiche del data set
Clustering	Identificazione delle affinità che definiscono i gruppi in un data set che mostrano comportamenti simili
Sequencing	Identificazione delle correlazioni tra comportamenti all'interno di un periodo definito
Associazione	Identificazione delle correlazioni tra comportamenti che ricorrono nello stesso periodo
Previsione	Identificazione di trend basata su dati storici

Asking a BI Question: Data Mining

- Tipologie di domande alle quali il DM fornisce una risposta

Domande	Tipo di Problema	Tecnica adottabile
Quali sono i tre principali motivi che hanno indotto il mio cliente a passare alla concorrenza?	Classificazione	Reti neurali Alberi decisionali
Quali sono le fasce di clienti a cui posso offrire nuovi prodotti/servizi?	Clustering	Reti neurali Alberi decisionali
Quali sono le probabilità che un cliente che ha aperto un c/c acquisterà anche il prodotto X in breve tempo?	Sequencing	Tecniche statistiche Rule induction
Quali sono le probabilità che un cliente acquisti due prodotti completamente differenti?	Associazione	Tecniche statistiche Rule induction
Quale sarà il prezzo del titolo tra un giorno/mese?	Previsione	Reti neurali Tecniche statistiche

Asking a BI Question: Data Mining e analisi predittiva

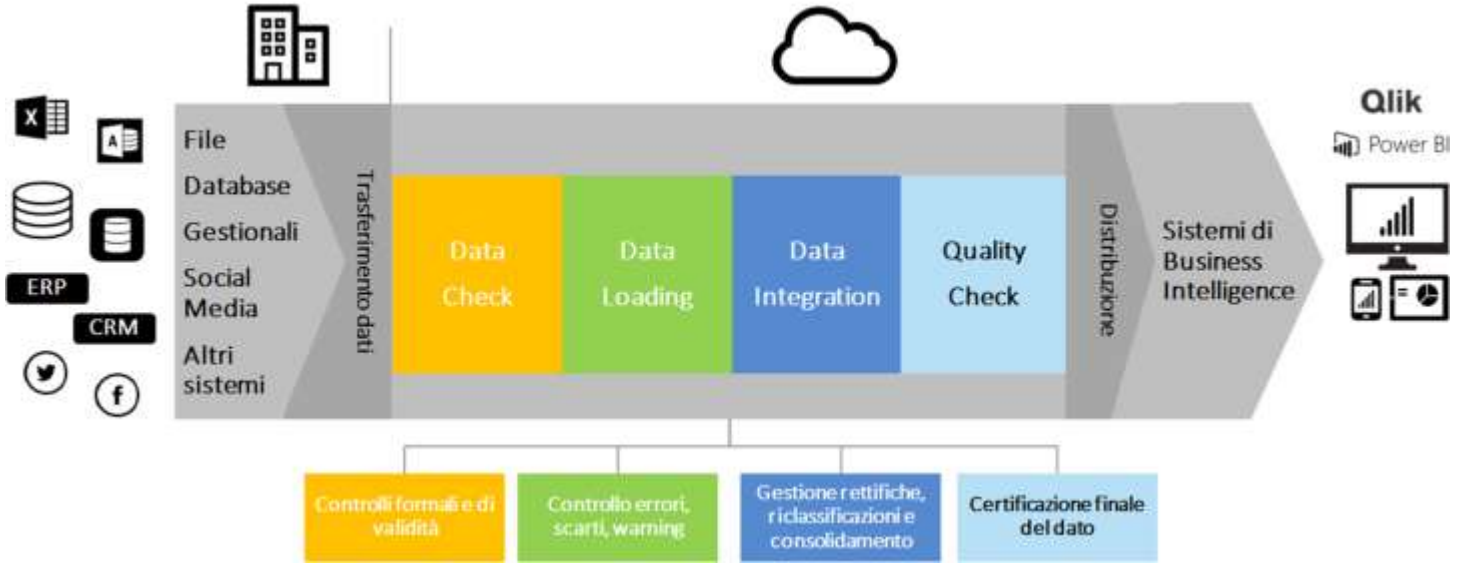
- Cosa c'è di nuovo nel Data Minig?
  - La possibilità di gestire enormi quantità di dati, che rendono obsoleta la definizione classica di grandi campioni
    - (miliardi di record e terabytes di dati non sono inusuali)
  - Le recenti tecniche che provengono dal mondo dell'ingegneria informatica
    - (reti neurali, alberi di decisione, regole di inclusione)
  - Interessi commerciali nel valorizzare le informazioni esistenti al fine di proporre soluzioni "individuali" per una determinata categoria di clienti
  - Text minig
    - Classificazione delle notizie dei giornali
    - Raggruppare e-mail secondo argomenti prestabiliti
    - Archiviare in automatico i documenti in base al loro contenuto

Asking a BI Question: Data Mining e analisi predittiva

- Le tecniche predittive utilizzano i dati del passato per estrarre una visione del futuro, costituita da un modello matematico/statistico.
- ***Le tecniche di predictive analytics non ci dicono cosa accadrà con certezza nel futuro, ma soltanto che cosa potrebbe accadere con un certo grado di probabilità***.
  - Il modello predittivo non è altro che un insieme di parametri, ricavati dai dati del passato, e tali da fornire una rappresentazione della relazione tra le caratteristiche del fenomeno analizzato (le variabili di input, per esempio: dati anagrafici dei clienti, comportamento di acquisto, ecc.) e il valore del fenomeno stesso (variabile di output, o target)

Asking a BI Question: Data Mining e analisi predittiva

- In generale le tecniche predittive, se utilizzate correttamente, portano vantaggi rilevanti in tutti i settori e per qualsiasi problema che debba essere risolto partendo dai dati.
- Ad esempio la predictive analytics è sfruttata con successo in campo medico, per esempio, come supporto all'identificazione di patologie, soprattutto quando occorrerebbe sottoporre il paziente ad esami particolarmente invasivi o estremamente costosi
- Applicazioni di interesse anche nell'industry 4.0 è la **ricerca di anomalie:**
  - *Fraud detection.*
  - *Predictive maintenance*
  - *Intrusion detection*



UNO STANDARD PER LO SVILUPPO DI UN PROCESSO DI BI

I sei passi dello standard CRISP-DM «Cross Industry Standard Process for Data Mining»- primo standard 1996.

- La metodologia fornisce un framework che prevede sei fasi, che possono essere ripetute ciclicamente con l'obiettivo di revisionare e rifinire il modello previsionale:

1. *Business Understanding* - Porre una domanda interessante.
2. *Data Understanding* - Ottenere i dati.
3. *Data Preparation* -Esplorare i dati.
4. *Modeling* - Creare un modello per i dati.
5. *Evaluation* - Creare un modello per i dati.
6. *Deployment* - Comunicare e presentare i risultati.

In ogni processo scientifico lo sviluppo di un modello deve prevedere step di valutazione

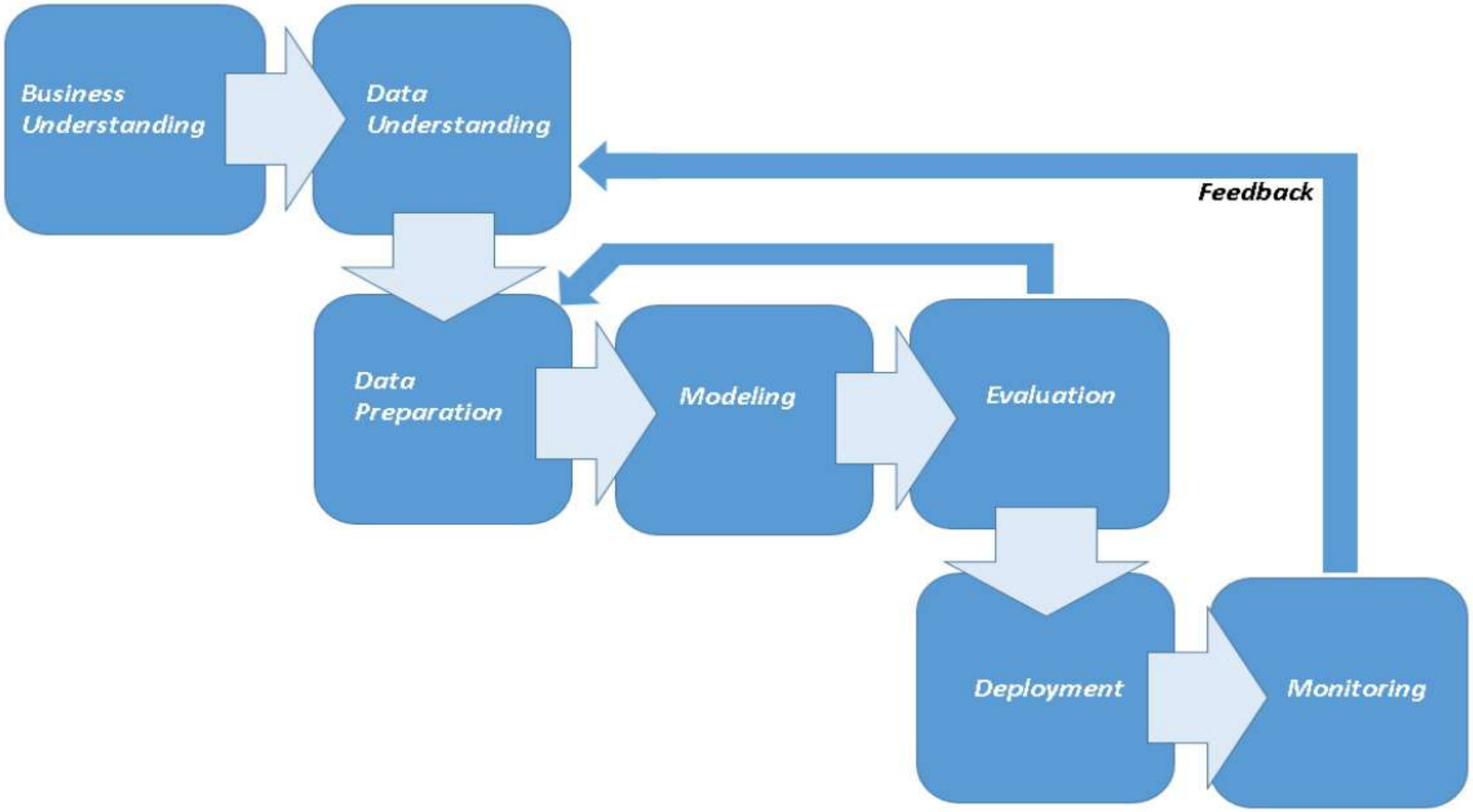
In una azienda i dati sono «comunicati» anche attraverso l'utilizzo del modello

I sei passi dello standard CRISP-DM «Cross Industry Standard Process for Data Mining»- primo standard 1996.

- La metodologia fornisce un framework che prevede sei fasi, che possono essere ripetute ciclicamente con l'obiettivo di revisionare e rifinire il modello previsionale:

1. *Business Understanding* – Si deve conoscere il business per implementare un sistema predittivo ed è necessaria una chiara definizione degli obiettivi da raggiungere
2. *Data Understanding* - Raccogliere correttamente, capire e valutare la bontà dei dati
3. *Data Preparation* -I dati presenti nei database aziendali nella loro forma originaria, sono praticamente inutilizzabili dagli algoritmi predittivi …
4. *Modeling* - creazione del modello predittivo tramite la scelta dell'algoritmo e la definizione dei parametri
5. *Evaluation* - effettuare un'attività di testing e valutazione della sua capacità predittiva
6. *Deployment* - Il modello può essere finalmente utilizzato dagli utenti di business

La metodologia CRISP-DM «Cross Industry Standard Process for Data Mining» in un grafo



Skills in development a BI project

- **Statistica e Matematica** – Si sa usare la regressione lineare? Si sa creare un modello per il pocesso?
- **Software Development** – R, python, or MATLAB? Database? Operazioni SQL? Attendibilià e controllo delle fonti dei dati?
- ***Business Experience*** – Come si sviluppa un processo aziendale? Che tempi ha?
- ***Adaptability*** – Si ha esperienza nel trovare una soluzione a un problema completamente nuovo senza una guida sostanziale?

I DATI
LA MATERIA PRIMA PER LA BUSINESS INTELLIGENCE

I Dati

Essere nell' «era dei dati»

- l'era dell'informazione ha fatto letteralmente esplodere la produzione di dati elettronici.
- Secondo le stime, nel 2011 si sono creati circa 1800 miliardi di GB di dati che nel 2012 sono saliti a 2800 e nel 2020 a 40.000…
- Non solo si creano ma si consuma dati anche a un tasso sempre più accelerato: nel solo 2013, un utente medio di uno smartphone usava circa 1 GB di dati al mese, oggi, tale cifra si stima si sia più che raddoppiata.

➢ Problema non risolto: riuscirne a dargli un senso

I Dati:le fonti

- *Interne*
  - Fonti operazionali, cioè quelle che fanno riferimento all'attività operativa giornaliera dell'azienda e che, per questo motivo, variano a seconda della tipologia di business e settore economico.
- *Esterne*
  - Es l'analisi del sentiment, volta a verificare quale sia l'opinione delle persone che scrivono sui social rispetto ad una certa tematica, un certo prodotto o una certa azienda.
    - Per realizzare questo tipo di attività occorrono dati provenienti dai social network (Facebook, Twitter,…), dai blog o da forum e dunque esterni all'azienda.
  - Il reperimento e l'utilizzo di dati esterni pongono alcuni problemi. Uno di essi, forse il principale, consiste nella loro qualità, che potrebbe presentare difetti di accuratezza, completezza e coerenza.
  - Occorre precisare che problematica relativa alla qualità riguarda anche i dati interni; tuttavia, sui dati esterni l'azienda non ha alcuna possibilità di manovra.

I Dati:Aspetto

- *Dati quantitativi*:
  - dati che possono essere descritti tramite numeri; su di essi è possibile eseguire semplici operazioni matematiche, compresa la somma.
  - Possono essere discreti o continui
  - Si potrebbero porre domande tipo:
    - Qual è il valore medio?
    - Questa quantità cresce o decresce con il trascorrere del tempo (se il tempo è un fattore)?
    - Esiste una soglia di attenzione?
- *Dati qualitativi*:
  - dati che non possono essere descritti tramite numeri e semplici operazioni matematiche. Questi dati, in genere, vengono descritti usando delle categorie e un linguaggio "naturale".
  - Si potrebbero porre domande tipo:
    - Quale valore è più/meno frequente?,
    - Quanti valori univoci esistono?

I Dati

Esempio: caratteristiche delle caffetterie

- *Dati quantitativi o qualitativi?*:
  - Nome della caffetteria
  - Fatturato (in migliaia)
  - Codice Zip (CAP)
  - Numero di clienti mensili (in media)
  - Origine del caffè

I Dati: struttura

- *Dati organizzati/ strutturati:*
  - si tratta di dati ordinati in una struttura a righe e colonne, dove ogni riga rappresenta un'unica osservazione e le colonne rappresentano le caratteristiche di tale osservazione.
  - oppure tramite un formato quali l'XML o il JSON, che assieme ai dati contengono i metadati che definiscono i nomi dei campi e la loro struttura.
- *Dati non organizzati/ strutturati*
  - questo è il tipo di dati in formato libero, normalmente testo o audio grezzo o segnali che devono essere analizzati meglio per poter essere organizzati
- *Dati semi organizzati/ strutturati*
  - I dati semi-strutturati presentano una parte dotata di struttura e una parte non strutturata. Per esempio un documento Word, o PDF, possiede una serie di metadati che sono molto ben strutturati (titolo, autore e molto altro), mentre il corpo del documento è costituito da testo.

I Dati: struttura

- *Dati non organizzati/ strutturati*
  - ***Rappresentano dall'80 al 90 % dei dati del mondo***
  - ➢ Il 90% informazioni disponibili al mondo si trova intrappolata in un formato difficilmente utilizzabile:
  - Si deve prevedere delle tecniche di pre-elaborazione (preprocessing), per dare una struttura ad almeno una parte dei dati, da passare alla successiva analisi
- Es su un testo «libero come un tweet:

This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies.

Posso eseguire:

- conteggio di parole/frasi;
- l'esistenza di determinati caratteri speciali;
- la lunghezza relativa del testo;
- l'individuazione degli argomenti.

	this	wednesday	morn	are	this wednesday	?	Lunghezza relativa	Argomento
Conteggio parole	1	1	1	1	1	1	4.03	astronomia

I Dati: I quattro livelli dei dati

- ***Il livello nominale*:**
  - è costituito dai dati descritti unicamente per nome o categoria.
    - ES:il genere, la nazionalità, la specie o il tipo di luppolo in una birra…
    - Non sono descritti da numeri e pertanto sono qualitativi.
  - Non possiamo svolgere operazioni matematiche sul livello nominale dei dati, tranne le operazioni di uguaglianza e appartenenza a insiemi:
    - Es: Una figura descritta come un quadrato rientra nella descrizione di un rettangolo, ma non vale il viceversa
  - Per trovare il **centro dei dati nominali, in genere si considera la moda** (l'elemento più frequente) del dataset.
- Poiché in genere si possono usare solo parole per descrivere i dati, queste si possono perdere in una traduzione o essere scritte in modo errato.
- Che conoscenze che se ne possono trarre? Avendo a disposizione solo la moda come misurazione del centro, siamo incapaci di trarre conclusioni su un'osservazione media, né possiamo ordinarli.

I Dati: I quattro livelli dei dati

- ***Il livello ordinale***:
  - Esiste un ordine di valutazione e gli strumenti necessari per collocare un'osservazione prima di un'altra e per confrontarle, tuttavia non si hanno le differenze relative fra le osservazioni, cioè non possiamo sommarle o sottrarle
    - Es: L'indice di gradimento è una delle più comuni fra le scale del livello ordinale
    - In questo caso si sa che un punteggio di 7 è meglio di 3, ma nulla ci dice che la differenza fra 5 e 6 è la stessa che fra 1 e 2.
  - **la mediana** (ovvero il valore centrale fra i dati numerici) è, normalmente, un modo appropriato per definire il **centro dei dati.** Si può anche usare la moda ma non la media perché a questo livello la somma o la sottrazione non hanno senso e neanche la divisione.
  - Es: Risultati di un sondaggio:

5, 4, 3, 4, 5, 3, 2, 5, 3, 2, 1, 4, 5, 3, 4, 4, 5, 4, 2, 1, 4, 5, 4, 3, 2, 4, 4, 5, 4, 3, 2, 1

Ordinati: *1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5*

  - ***Mediana: 4.0***
  - Media: 3.4375 meno adatta a rappresentare il risultato

I Dati: I quattro livelli dei dati

- ***Nominale o Ordinale?***
  - L'origine dei semi nella vostra tazzina di caffè.
  - La posizione ottenuta dai partecipanti a una gara a piedi.
  - Il metallo usato per la medaglia ricevuta dopo aver partecipato alla suddetta gara.
  - Il numero telefonico di un cliente.
  - Quante tazzine di caffè bevete in una giornata.

I quattro livelli dei dati

- ***Il livello degli intervalli:***
  - I dati di questo tipo consentono di eseguire sottrazioni fra i punti dei dati. Si possono ordinarli e confrontarli ma anche sottrarli e sommarli.
    - Es: La temperatura.
      - Se a Roma c'è una temperatura di 37° e a Mosca c'è una temperatura di 27°, si può inferire che a Roma ci sono 10° in più che a Mosca.
  - La più accurata descrizione del **centro dei dati è in questo caso la media** aritmetica (ma è possibile usare anche moda e mediana).

I quattro livelli dei dati

- *Il livello degli intervalli*:
  - ***E' possibile misurare attraverso un numero la "dispersione" dei dati***.
  - Insieme a una misurazione del centro, una misurazione della variabilità può descrivere quasi interamente un dataset con due soli numeri.
  - La deviazione standard (o scarto quadratico medio) è la misurazione della variabilità dei dati più comune e rappresenta la "distanza media di un punto dei dati rispetto alla media".
  - Utile se si è interessati alle fluttuazioni nei dati (es. resa percentuale delle azioni).

la "differenza al quadrato" fra i punti e la media invece della "semplice differenza" permette di enfatizzare i valori erratici, quei punti dei dati che sono eccezionalmente lontani.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

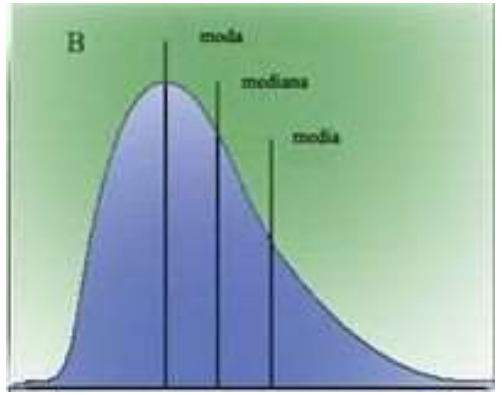
I Dati

Principali indici statistici

- **INDICE**
  - **di posizione**
    - **MODA**
    - **MEDIANA**
    - **MEDIA**
  - **di dispersione**
    - **SCARTO QUADRATICO MEDIO**
    - **VARIANZA**
    - **RANGE**
  - **di forma**
    - **ASIMMETRIA (SKEWNESS)**
    - **CURTOSI (KURTOSIS)**

	DEFINIZIONE	VANTAGGI	SVANTAGGI
MEDIA	somma dei dati / numero dei dati	adatta a manipolazioni matematiche	molto influenzata dai valori estremi
MEDIANA	livello di misura al di sotto del quale cade la metà dei dati	non influenzata dai valori estremi	non adatta a manipolazioni matematiche
MODA	valore che ricorre con maggiore frequenza	di significato facilmente intuibile	possibili distribuzioni bi-, tri-modali ecc.

www.quadernodiepidemiologia.it



I Dati:I quattro livelli dei dati

- ***Il livello dei rapporti*** :
  - I dati al livello degli intervalli non hanno un "punto iniziale naturale o uno zero naturale"
    - Es: i 0°C non indicano affatto una "assenza di temperatura", i 0° Kelvin sì e non si tratta di uno zero collocato in modo arbitrario.
    - Es: il denaro depositato in banca si colloca a questo livello: si può definire "niente denaro sul conto" .
  - ***Su questi dati possiamo fare anche moltiplicazioni e divisioni***
  - La media aritmetica rimane il metodo più usato ricavare il centro dei dati, ma esiste anche la media geometrica (radice quadrata del prodotto di tutti i valori).
- I dati al livello dei rapporti sono normalmente non-negativi (avendo un punto iniziale..) e si può avere qualche problema nel trattarli.
- Es: se nel conto bancario si consente la presenza di debiti ed un saldo di 50 000 euro, il seguente rapporto non avrebbe affatto senso:
- 50 000euro/-50 000 euro = -1

I Dati:I quattro livelli dei dati

- ***Il livello nel quale collocare i dati è un'importante decisione all'inizio di ogni analisi.***
- Es: su dei dati che in genere vengono considerati al livello ordinale e si può applicare loro strumenti come la media aritmetica e la deviazione standard che prima o poi possono dare dei problemi imponendo ai dati una struttura che essi non hanno.

I Dati: I quattro livelli dei dati

- Con un nuovo dataset, ci si deve chiedere:
  - ***I dati sono organizzati o non organizzati?*** (es, i nostri dati sono disponibili in una struttura tabellare?)
  - ***Ogni colonna è quantitativa o qualitativa?*** (es. i valori sono numeri, stringhe o rappresentano quantità?)
  - ***In quale livello dei dati si situa ogni colonna?*** (es, i valori sono nel livello nominale, ordinale, degli intervalli o dei rapporti?)
- Le risposte stabiliscono i tipi di grafici che si potranno usare e l'interpretazione dei dati nei successivi passi.
  - Talvolta si dovrà con attenzione convertire i dati da un livello a un altro per allargare la visione del problema.

BIG DATA
Business Analytics

Big Data e Business Intelligence

- ***I Big Data sono considerati il nuovo petrolio***

- Nel 2018, il mercato Big Data Analytics ha fatto registrare una forte accelerazione, con un tasso di crescita del 26%, raggiungendo un valore complessivo di 1,393 miliardi di euro.
  - Tradotto in termini più concreti ciò significa incrementare il fatturato, ampliare la base clienti, creare nuovi servizi e prodotti e, in ultima analisi, accrescere il profitto.
- Questo dato ha trovato anche riscontro nell'indagine relativa alle priorità d'investimento di CIO e Innovation Manager italiani, con i Big Data Analytics sempre più al centro delle attenzioni.

https://blog.osservatori.net/it_it/big-data-aspetti-positivi

Big Data e Business Intelligence

- *I Big Data sono considerati il nuovo petrolio*

- Quella dei big data è un'industria estrattiva:
  - così come si ricava il petrolio dalle profondità del suolo o il carbone dalle miniere,
  - i nostri dati personali vengono
    1. estratti in forma grezza da internet
    2. e poi raffinati (o meglio, aggregati)

- per creare **conoscenza e valore** per chi li analizza e li sfrutta a fini commerciali.

Big Data e Business Intelligence

- *I Big Data sono considerati il nuovo petrolio*
- Non si tratta solo di Facebook e Google,
- *Ma*:
  - tutte le informazioni sulla nostra attività fisica raccolte dagli smartwatch,
  - gli spostamenti memorizzati dagli smartphone,
  - la musica che ascoltiamo su Spotify,
  - i film che vediamo su Netflix,
  - la cronologia web
  - ….

Informazioni che possono essere utilizzate per creare un ***profilo consumatore quanto più possibile accurato*** *e inviare pubblicità personalizzata, offerte su misura o sconti sulle assicurazioni.*

Big Data e Business Intelligence

- ***I Big Data sono considerati il nuovo petrolio***

- E ancora l'esplosione di **Internet of Thing..**
- **Oggetti intelligenti** connessi alla rete in grado di estrarre qualunque tipo d'informazione:
  - ✓ frigoriferi intelligenti =>cosa mangiamo,
  - ✓ Televisori intelligenti => le nostre trasmissioni,
  - ✓ Visori per la realtà aumentata => ristoranti in cui ci siamo recati, monumenti che abbiamo visitato
  - ***Assistenti digitali (eg, Alexa…)***
    - *Che analizzeranno con sempre maggiore precisione le nostre conversazioni per dedurne ulteriori informazioni*

Big Data e Business Intelligence

- ***I Big Data sono considerati il nuovo petrolio***

- Il problema è che spesso gli utenti non si rendono conto di quanti sono i dati che forniscono spontaneamente alle aziende.
- Si immagini uno smartphone connesso ad internet con il GPS attivo, casomai mentre fate jogging con una app che registra i vostri passi e la frequenza cardiaca.

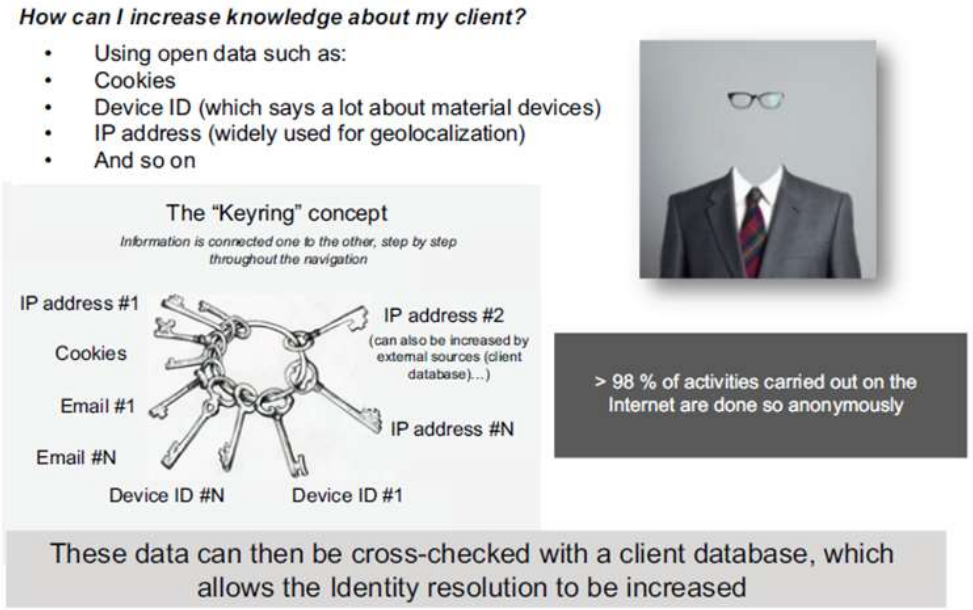


Figure 1. *Identity resolution*

- La quantità di dati inviati volontariamente alle aziende è enorme
  - perché scaricando e utilizzando le app avete accettato il loro uso
- ma la gente non si preoccupa
  - almeno fin quando l'assicurazione non disdetta la polizza perché si accorge che guidate troppo velocemente

https://www.valigiablu.it/big-data-petrolio-digitale/

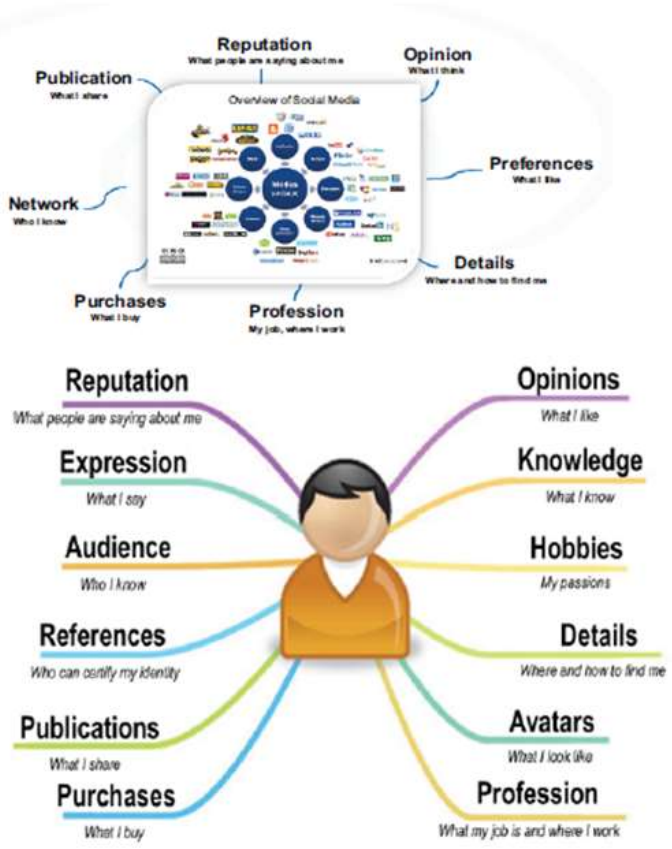


Figure I.2. *The traces we leave on the Internet (whether voluntarily or not) form our Digital Identity*

Identità digitale

- ✓ **identità dichiarativa**: il dato che noi inseriamo volontariamente (sui social network, blog, ecc.);
- ✓ **identità comportamentale** (download, navigazione, cookie …)

- ➢ Ogni nuova connessione, navigazione o altra attività sul Internet arricchisce questo patrimonio informativo su di noi, e di cui non siamo i custodi.
- ➢ E qui sta il problema: abbiamo infatti organizzato le nostre vite attorno a questa Identità Digitale
- ➢ E "delegato" la gestione del nostro identità a terzi (come i motori di ricerca).

Questa identità alla fine diventa la nostra **e-reputation**

Big Data e Business Intelligence

- *Per un mondo connesso «senza latenza»…*
- **Il tempo** è la parola chiave
  - tutto ruota intorno alla recitazione più veloce e migliore dei concorrenti nell'ambiente digitale, dove le informazioni viaggiano attraverso Internet alla velocità della luce.
- Il tempo rappresenta quindi un "patrimonio immateriale" con alto valore aggiunto
  - Dato che gran parte delle nostre decisioni e azioni successive (personali o professionali) dipendono dal mondo digitale (che mescola informazioni e algoritmi per elaborare queste informazioni a velocità mai viste).

- ***Questo "nuovo" mondo è strutturato su Internet e richiede alle aziende di prendere decisioni e agire in un ambiente altamente competitivo, gestendo dati complessi in pochi millisecondi (o meno).***
- **Un mondo in cui "l'esperienza del cliente" è fondamentale:** ***un'azienda non possiede un cliente ma solo il tempo che sceglie di dedicare***

Big Data e Business Intelligence

- Le soluzioni attualmente in atto (*recommendation systems*) non sono molto interattive con il loro ambiente:
  - Si basano su modelli predefiniti basati su un numero limitato di variabili descrittive per la situazione
  - Non mostrano molto autoapprendimento (cioè aggiornamento dei modelli dopo l'analisi dei dati reali)
  - E il risultato è che le stesse cause (identificate da poche variabili) innescano gli stessi effetti.

- Inoltre, spesso non prendono in considerazione le variazioni del contesto in tempo reale
  - come un utente è arrivato su una pagina web, quali contenuti hanno visto prima, qual è la natura della loro ricerca, le azioni risultanti, ecc.
- O prendono in considerazione solo i risultati di decisioni e azioni passate.

Big Data e Business Intelligence

- L'Intelligenza Artificiale legata ai Big Data si dice che sarà la chiave di volta **dell'apprendimento digitale.**

- L'Intelligenza Artificiale verrà utilizzata in aggiunta ai Big Data per
  - estrarre significato,
  - determinare risultati migliori basati sull'apprendimento continuo
  - e consentire il processo decisionale in tempo reale.

F. Iafrate (2018). Artificial Intelligence and Big Data: The Birth of a New Intelligence. Wiley Online Library

Big Data e Business Intelligence

Vantaggi di AI rispetto alla Business Intelligence tradizionale basata su metodi statistici:

1. capacità di analizzare e prendere decisioni in pochi millisecondi all'interno di un contesto di analisi molto complesse sulla base di dati grezzi quali sono i Big Data
2. capacità di imparare dall'esperienza (analisi, decisioni, azioni) fornita dagli esperti dei settori e ricordarla in un qualche modo
   - «non c'è scelta buona o cattiva, esistono solo esperienze»

- ***Questa memoria digitale, che si arricchisce quando si verificano e si sviluppano esperienze diverse sarà la chiave di volta dei processi decisionali e nel tempo costituirà la memoria dell'azienda.***

La nostra società di informazione e comunicazione (media audiovisivi, Internet) potrebbe suggerire che l'apprendimento equivale ad essere informato.

Questo approccio è incompleto:

- ***l'informazione è certamente una parte importante del ciclo di apprendimento, ma informarsi non è la stessa cosa che «allenarsi» (training).***

Perché ci sia un vero apprendimento, lo studente (l'algoritmo, nel caso dell'Intelligenza Artificiale) deve essere in grado di scegliere tra diverse soluzioni e imparare dalle sue scelte alla luce dell'obiettivo da raggiungere.

- ***Questo alla fine si tradurrà in esperienze (migliori o peggiori in termini di obiettivi) che formeranno gradualmente un livello iniziale di apprendimento.***
  - È questo livello di apprendimento che troviamo nelle soluzioni di Intelligenza Artificiale, chiamato "**autoapprendimento**".

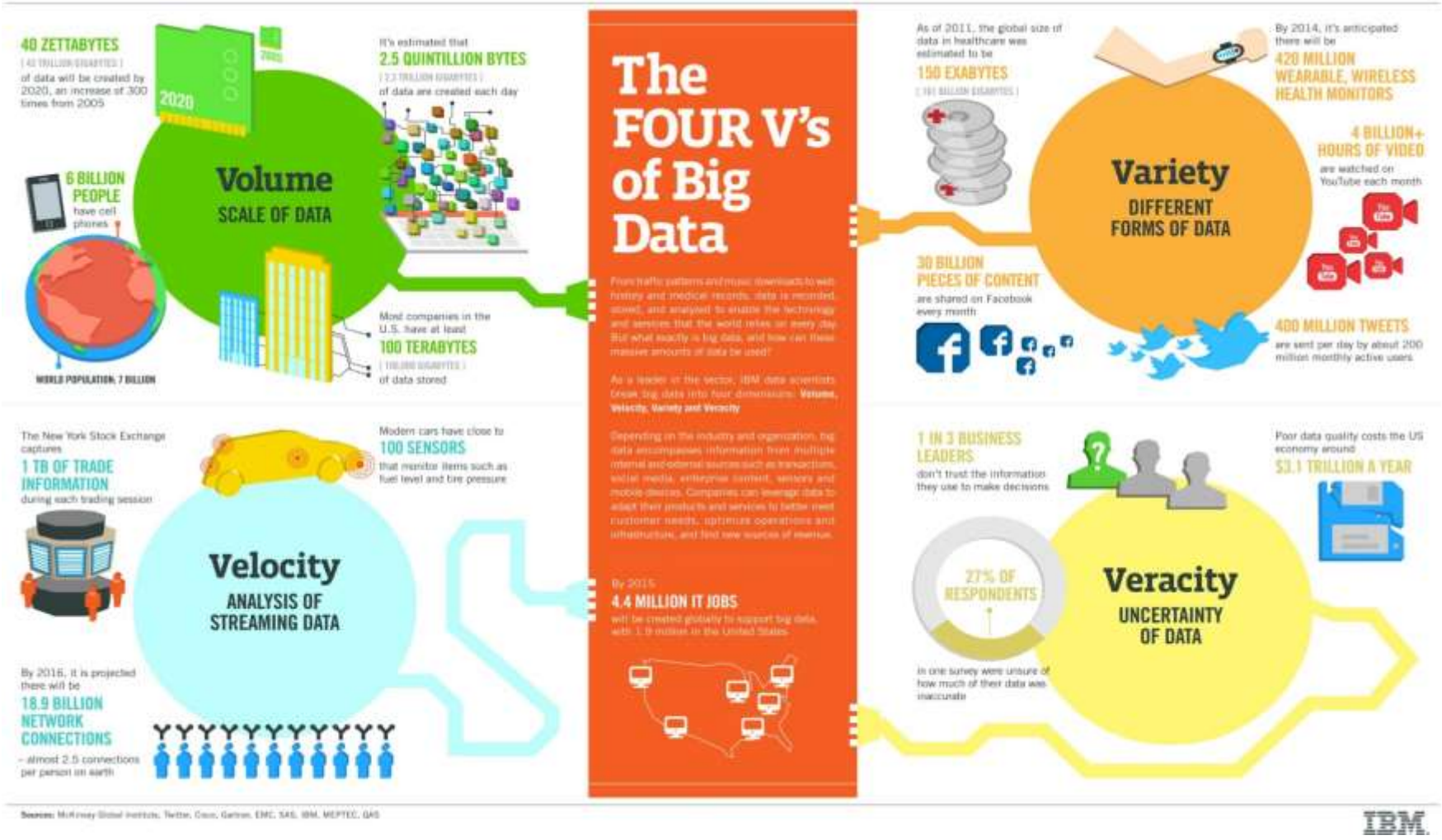
Big Data

- I big data sono generalmente **definiti** come quei dati che presentano una o più delle seguenti quattro caratteristiche, chiamate le quattro V:
  1. **Volume** :le quantità elevate di dati. I dati generati automaticamente da macchine (sensori, DCS - Distributed Control System, strumenti scientifici) e quelli relativi a transazioni bancarie e movimenti sui mercati finanziari possono assumere volumi imponenti, soprattutto se considerati al loro massimo livello di granularità.
  2. **Velocità:** la rapidità con cui i dati sono prodotti. L'Internet Of Things e i sensori, sono in grado di generare dati con una velocità elevatissima.
  3. **Varietà**
  4. **Veracità**

Big Data

- I big data sono generalmente **definiti** come quei dati che presentano una o più delle seguenti quattro caratteristiche, chiamate le quattro V:
  1. **Volume**
  2. **Velocità**
  3. **Varietà:** la diversità dei formati, delle fonti e delle strutture..
  4. **Veracità:** bias, rumore e anormalità nei dati. I dati che vengono archiviati e estratti sono significativi per il problema analizzato?

Big Data


40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005
2020
6 BILLION PEOPLE have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The FOUR V's of Big Data
By 2015
4.4 MILLION IT JOBS
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
IBM

Big Data Analytics

Caso	Caratteristiche	Esempi di utilizzo
Sensori e DCS	Velocità e volume	Analisi dei guasti e manutenzione predittiva
Radio Frequency Identification (RFID)*	Velocità e volume	Analisi del percorso d'acquisto in un negozio della grande distribuzione. Analisi e tracking delle merci, in combinazione con altri dati (ambientali, geografici, ecc.)
Quotazioni e transazioni su mercati finanziari.	Velocità e volume	Sistemi automatici di high frequency trading; analisi previsionale
Dati da strumenti scientifici	Velocità e volume	Riconoscimento di pattern. Simulazioni.
Dati astronomici	Volume e varietà	Analisi del quantitativo enorme di dati raccolti da osservatori e radiotelescopi.
Dati metereologici	Volume	Previsioni meteo. Monitoraggio di eventi atmosferici estremi.

Esempi di Big Data

Big Data Analytics

Caso	Caratteristiche	Esempi di utilizzo
Informazioni sanitarie	Volume e varietà (diversità di formati)	Ministero della salute, enti di ricerca: identificazione e monitoraggio della diffusione di malattie
Dati fiscali, bancari e patrimoniali	Volume	Il Ministero delle Finanze, Agenzia delle Entrate e Guardia di Finanza possono utilizzare le enormi banche dati a loro disposizione per l'identificazione di comportamenti anomali che indicherebbero casi di evasione fiscale.
Social Network	Varietà: diversità di formati, dati semi-strutturati	Sentiment Analysis (come si sta parlando della nostra azienda? Come è stato accolto il nuovo prodotto?) CRM (Customer Relationship Management) Utilizzo da parte di servizi di intelligence.
Blog, Forum	Varietà: Dati semi-strutturati	Sentiment analysis Utilizzo da parte di servizi di intelligence.
Web server log	Volume	Analisi del traffico sui web server, identificazione dei comportamenti di navigazione degli utenti

Esempi di Big Data

Big Data Analytics

Caso	Caratteristiche	Esempi di utilizzo
Log del traffico di un router	Volume e Velocità	Utilizzo da parte di provider.
Dati provenienti da sistemi di sorveglianza	Volume, Velocità e diversità di formati	Utilizzo da parte di polizia, enti di vigilanza, servizi di intelligence.
Documenti	Volume e assenza di struttura	Fraud detection: per esempio, l'analisi di richieste di risarcimento effettuate alle assicurazioni possono essere analizzate e associate a casi fraudolenti ed esaminate con più attenzione.

Esempi di Big Data

Big Data Analytics

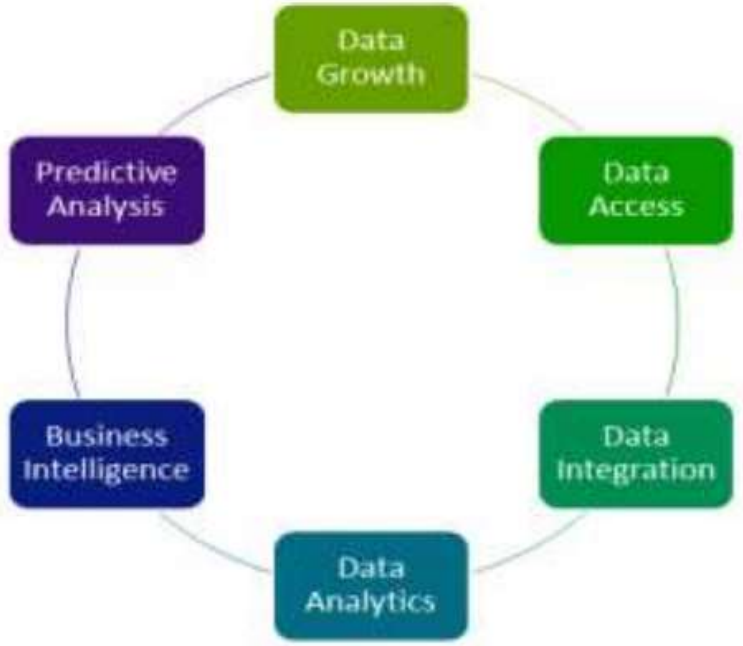
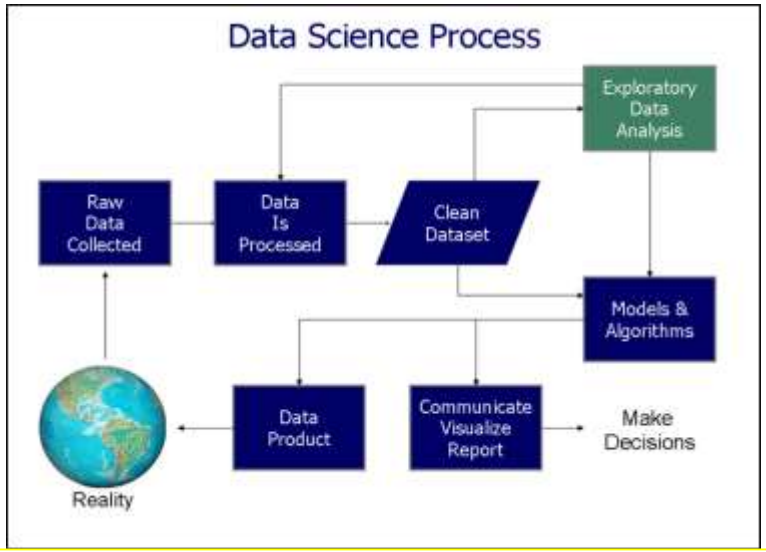


Figure 6: Life Cycle of Big Data

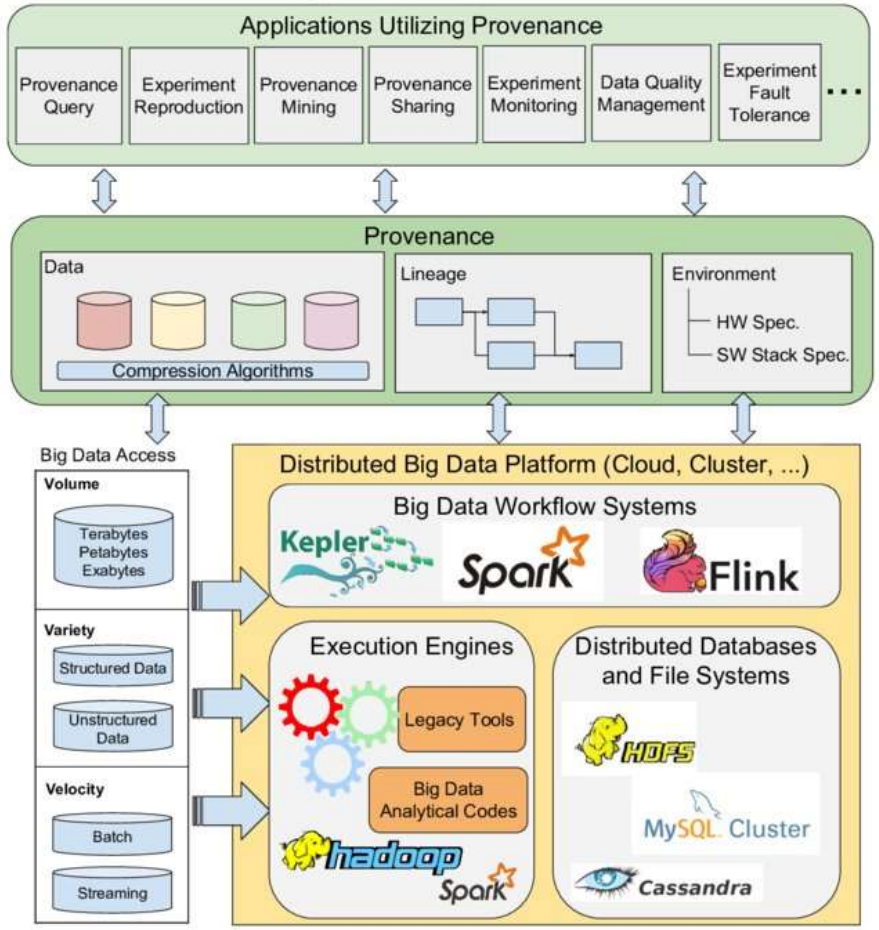
Real time Predictive Analysis Using Big Data

Big Data Analytics

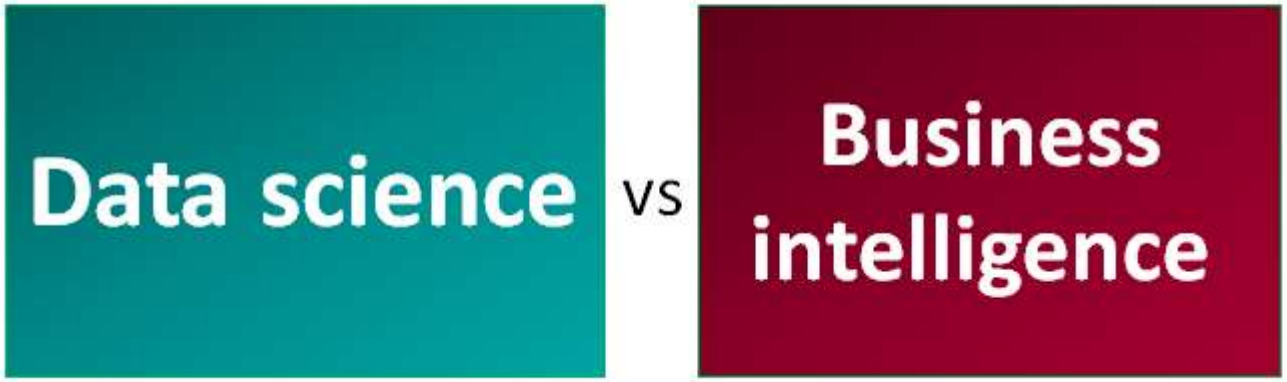
Per un system di Data Analysis funzionante si deve progettare e mantenere tutto il framework per il Big Data Engine



"A scientist can discover a new star, but he cannot make one. He would have to ask an engineer to do it for him."
-Gordon Lindsay Glegg, The Design of Design (1969)



Proposed Big Data provenance reference architecture by Jianwu Wang.



I DATI

LA MATERIA PRIMA PER LA BUSINESS INTELLIGENCE MA ANCHE PER LA DATA SCIENCE

Business Intelligence

- E’ fondamentalmente un insieme di tecnologie, applicazioni e processi utilizzati dalle aziende per l'analisi dei dati aziendali.
- Esegue la conversione di dati grezzi in informazioni significative che vengono quindi utilizzate per il processo decisionale aziendale e azioni redditizie.
- Si tratta dell'analisi di dati strutturati e talvolta non strutturati che aprono la strada a nuove e redditizie opportunità di business.
- Supporta il processo decisionale basato sui fatti piuttosto che il processo decisionale basato su ipotesi.
- Quindi ha un impatto diretto sulle decisioni aziendali di un'impresa.
- Gli strumenti di business intelligence aumentano le possibilità di un'impresa di entrare in un nuovo mercato e aiutano a studiare l'impatto degli sforzi di marketing.

Data Science

- E’ fondamentalmente un campo in cui le informazioni e la conoscenza vengono estratte dai dati utilizzando vari metodi e algoritmi.
- Può quindi essere definito come una combinazione di vari strumenti matematici, algoritmi, statistiche e tecniche di apprendimento automatico che vengono utilizzati per trovare i modelli nascosti nei dati che aiutano nel processo decisionale.
- Si occupa sia di dati strutturati che non strutturati.
- È correlato sia al data mining che ai big data.
- Implica lo studio delle tendenze storiche e quindi l'utilizzo delle sue conclusioni per ridefinire le tendenze attuali e anche prevedere le tendenze future.

Analisi con Business Intelligence

- BI lavora su una formula preesistente.
- Un'azienda si avvicinerà a un esperto di BI con un'idea di quali dati vogliono essere estratti e analizzati e della formula che vogliono utilizzare.
- La risposta che l'azienda spera di ottenere sarà già definita, almeno in termini di ambito, anche se i fatti e le cifre esatti che comporranno la risposta sono ancora da definire.

Analisi con la Data Science

- La scienza dei dati inizia con una domanda.
- Piuttosto che attenersi a un metodo collaudato per misurare il successo e il fallimento all'interno dell'azienda, la scienza dei dati cerca di anticipare tendenze, opportunità e problemi futuri.

- Ma l'attuale relazione tra business e dati e l'avvento dei Big Data e e Machine Learning suggerisce che ci stiamo muovendo verso un futuro in cui BI e data science vanno necessariamente di pari passo.
- La BI fornirà le basi su cui i data scientist possono formulare le proprie previsioni.

Business Intelligence vs Data Science