con la partnership scientifica di:

AICA

Unità 4

Gestire le informazioni

Livello 3 - approfondimento

Temi trattati all'interno dell'Unità

- Dati e informazione.
- Il valore della conoscenza.
- La Business Intelligence
- Le tecniche di supporto alle decisioni.
- Big Data e le 4V: Volume, Velocità, Varietà, Veridicità.
- Open Data e il valore dei dati pubblici.
- Open Government e Open Service

Sommario

DATI E INFORMAZIONI

Per capire il potere dei dati è utile iniziare a definire cos'è un **dato**, cos'è una **informazione**, cosa significa **presentare** l'informazione e come arriviamo alla **conoscenza** tramite l'informazione.
Andiamo nel dettaglio delle definizioni, distinguendo bene tra dati e informazioni.

Il **dato** è l'unità elementare (grezza) di informazione.

L'**informazione** è l'elaborazione dei dati per rispondere a esigenze specifiche

Esempi di dati possono essere:

- 03/04/2008
- 4/3/2008
- 20080403

E l'informazione conseguente - in diverse forme e rappresentazioni - è che queste stringhe di caratteri potrebbero corrispondere a una data.
Altri esempi di dati grezzi possono essere:

- AB, 5

Elaborati come **informazioni**, questi dati possono fornire significato e corrispondere ad esempio, sempre secondo l'ambito in cui sono definiti e raccolti, a:

- un gruppo sanguigno di tipo AB+,
- un'età anagrafica di 5 anni.

Quando parliamo di dati, parliamo di elementi base, a volte definibili grezzi e non significativi

Quando parliamo di informazione, invece, parliamo di elaborazione dei dati grezzi e/o elementari; in altre parole l'informazione fornisce significato ai dati stessi.
Presentare l'informazione vuol dire valorizzare il senso e il significato di quello che esprime.
La conoscenza è l'utilizzo dell'informazione come elemento di comprensione dei dati raccolti, analizzati e lavorati
L'informazione esiste, quindi, perché è sostanziata dai dati. Senza dati non avremmo informazione e avere **informazione** significa gestire una **conoscenza** tale da permettere attività di analisi, gestione, coordinamento.
I dati, e la conseguente conoscenza, ovvero l'applicazione dell'informazione, sono il fulcro di qualsiasi organizzazione in qualsiasi ambito.

L'informatica nasce come tecnologia di raccolta e trattamento dell'informazione e di elaborazione automatica dei dati e un acronimo la definiva un tempo come EDP, *Electronic Data Processing*.
L'informazione viene trattata e assume valore nei processi e nelle attività di qualsiasi organizzazione e non esiste processo o procedura che non tratti dati e di conseguenza informazioni.

L'analisi dei flussi e dei processi in qualsiasi ambito permette di individuare e registrare le informazioni che assumono di volta in volta valore e rilevanza a ogni passaggio.
L'analisi delle informazioni registrate permette alle organizzazioni di controllare e misurare le attività, prendere decisioni, migliorare i propri processi, il proprio business e la propria missione.

L'informatica ha avuto fin dall'inizio l'esigenza di creare strutture e modelli che facilitassero la raccolta dei dati e la loro ricerca.

Nascono così le "basi dati" e un **Database** è proprio un insieme di informazioni catalogate e organizzate.

Per la progettazione ad alto livello di un database è largamente diffuso l'uso del modello "entità e relazioni" (in inglese **Entity-Relationship**). Questo modello consente infatti di tradurre i risulti dell'analisi dei dati di un contesto applicativo in uno schema concettuale con associabile una rappresentazione grafica comprensibile anche da parte di non specialisti di database.

Il modello si basa sull'identificazione delle classi di elementi (oggetti, persone, ecc.) chiamate appunto **entità**. Dove ogni classe rappresenta l'insieme di elementi che presentano una serie di tipi di dati (detti attributi dell'entità) in comune, dove ogni singolo elemento della classe (chiamato occorrenza) avrà i propri dati.

Per esempio: l'entità lavoratore presenterà come attributi quelli anagrafici, i riferimenti di indirizzo, numeri telefonici, ecc. , la data di assunzione ecc..

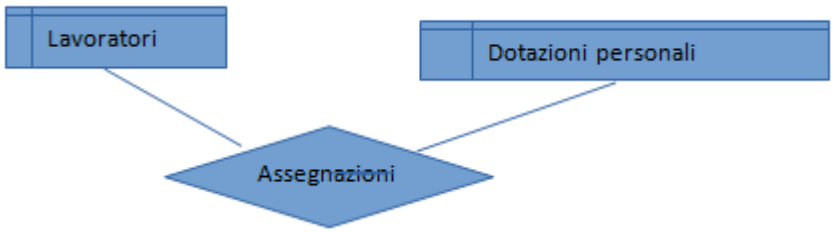
Ogni singola occorrenza, che rappresenta ogni singolo lavoratore conterrà poi i dati che a lui si riferiscono.
In un contesto applicativo le entità sono fra loro in **relazione**, nel senso che ogni occorrenza della prima puo' essere associata ad una o più occorrenze della seconda.

Proseguendo con l'esempio del lavoratore, possiamo considerare le dotazioni che l'azienda consegna al lavoratore, come PC, dispositivo mobile, divisa di lavoro, ecc.

Anche le dotazioni costituiscono un'entità che presenterà come attributi il tipo di dotazione, il suo valore o prezzo di acquisto, la data di acquisto, il fornitore, ecc.

Le due entità "lavoratori" e "dotazioni personali" sono fra loro in relazione poiché un lavoratore può avere o aver avuto in carico nessuna, una o più dotazioni personali, mentre la singola dotazione personale può essere stata assegnata ad uno o più lavoratori in tempi diversi o a nessuno, se è stata appena acquistata.

Il modello entità e relazioni viene usualmente rappresentato in forma grafica, le entità con rettangoli, le relazioni con altra figura geometrica collegata con segmenti alle entità che sono in relazione.



Anche le relazioni possono presentare attributi, per esempio la relazione "Assegnazioni" dell'esempio, può presentare gli attributi: data di consegna e data di restituzione.

A partire da un modello entità e relazioni, gli specialisti di database procedono alla stesura dello schema logico dettagliato, secondo il modello di database adottato e da questi allo schema fisico che precisa i supporti di memoria di massa dove le diverse occorrenze delle entità e delle relazioni devono essere inseriti.

Lo schema logico e quello fisico vengono quindi inseriti nel prodotto che gestirà poi le operazioni sui dati contenuti nel database e che prende il nome di DBMS.

DBMS è infatti l'acronimo di ***Data Base Management System***, e il DBMS è il software che permette di creare e gestire un database.

Insieme ai database nasceva il problema di gestire dei dati coerenti e correlati durante la loro registrazione e il loro trattamento. La **transazione** entrava nel glossario della gestione dei dati.
Quando si cita una **transazione sul database**, si intende un insieme di operazioni che devono essere eseguite in maniera "atomica": ovvero o tutte le operazioni vanno a buon fine e aggiornano i dati o tutte le operazioni vengono rigettate e non modificano o inseriscono dati nel database (regola del "tutto o niente").
Questo permette di rendere sempre congruenti le informazioni correlate relative alla transazione registrata sulla base dati.
Pensate a scritture e registrazioni relative a ordini di acquisto che hanno una loro testata e una serie di articoli che compongono l'ordine, o alla scrittura di una registrazione in contabilità con il conseguente aggiornamento di prima nota, partitari vari, ecc.

L'acronimo **ACID** (**Atomicità, Consistenza, Isolamento, Durabilità**) esprime proprio i concetti associati alla gestione coerente delle transazioni:

- **Atomicità** (***Atomicity***): la transazione è indivisibile nella sua esecuzione, e la sua esecuzione deve essere o totale o nulla, non sono ammesse esecuzioni intermedie. Tecnicamente si parla di COMMIT in caso di transazione totale coerente o ROLLBACK in caso di transazione non coerente, con dati da ripristinare per rispettare il concetto di **Consistenza**.
- **Consistenza** (***Consistency***): quando inizia una transazione il database si trova in uno stato coerente e quando la transazione termina il database deve essere in uno stato coerente, ovvero non deve violare eventuali vincoli di integrità, quindi non devono verificarsi contraddizioni (*inconsistency*) tra i dati archiviati nella base dati, da cui la necessità del concetto di **Atomicità**.
- **Isolamento** (***Isolation***): ogni transazione deve essere eseguita in modo isolato e indipendente dalle altre; l'eventuale fallimento di una transazione non deve interferire con altre transazioni in esecuzione.
- **Durabilità** (***Durability***): persistenza; dopo un *commit* (vedi Atomicità), i cambiamenti apportati non dovranno essere più persi.

LA BUSINESS INTELLIGENCE

La *business intelligence* raggruppa un insieme di pratiche capaci di estrarre poche e significative informazioni da un grande insieme di dati, e di presentarle in maniera immediatamente comprensibile per chi deve prendere decisioni.

Esistono strumenti molto sofisticati per l'analisi dei dati, ma la cosa più importante sono i ragionamenti per definire e capire quali sono gli indicatori di cui abbiamo bisogno per prendere decisioni per comprendere quali possono essere i dati di ingresso da cui si estraggono le informazioni.

Con ***business intelligence*** (**BI**) ci si riferisce quindi:

1. ai processi aziendali per la raccolta e l'analisi delle informazioni importanti e strategiche,
2. alla tecnologia utilizzata per realizzare questi processi,
3. alle informazioni ottenute come risultato di questi processi.

Le organizzazioni raccolgono informazioni per effettuare valutazioni e stime sul proprio contesto aziendale e di mercato (ricerche di mercato e analisi della concorrenza).
I dati raccolti sono elaborati e utilizzati per supportare concretamente le decisioni di chi occupa ruoli direzionali (esempio: capire l'andamento delle *performance* dell'azienda, generare stime previsionali, ipotizzare scenari futuri e future strategie di risposta).
In secondo luogo le informazioni possono essere analizzate a differenti livelli di dettaglio e gerarchici per le necessità di ogni altra funzione aziendale: *marketing*, commerciale, finanza, personale o altre.
In letteratura la *business intelligence* è citata come il processo di trasformazione di dati e informazioni in conoscenza.
Il software utilizzato nella business intelligence ha l'obiettivo di permettere alle persone di prendere decisioni strategiche fornendo informazioni precise, aggiornate e significative nel contesto di riferimento.
I sistemi di business intelligence sono chiamati anche sistemi per il supporto alle decisioni (***Decision support systems*** o **DSS**).
Le tecniche di business intelligence possono presentare informazioni di sintesi utilizzando il cosiddetto cruscotto (*dashboard*), uno strumento indispensabile per guidare l'analisi, in grado di separare indicatori indispensabili dai dati consultabili per pianificare le prossime azioni.

I cruscotti sono interfacce che accelerano i processi decisionali presentando in modo sintetico le informazioni più importanti in termini di business intelligence. Le informazioni sono sintetizzate con indicatori di performance (KPI) (quando parleremo del *business plan* vedremo bene cosa sono i KPI) associati a criteri operativi, tattici e strategici di una organizzazione.
I dati generati dai vari sistemi (ad esempio: contabilità, produzione, ricerca & sviluppo, ecc.) vengono archiviati in ***data warehouse***, che ne conservano le qualità informative.
Il ***Data warehouse*** è il luogo di consolidamento dei dati aziendali.
Le tecniche e i sistemi dedicati alla business intelligence vanno ad interrogare questo archivio complessivo dei dati per estrarre informazioni utili a prendere decisioni.
Il ***data mart*** (deposito di dati) indica un sotto-insieme contenente i dati per un particolare settore (dipartimento, direzione, servizio, gamma prodotto, ecc.). Si parla quindi ad esempio di *data mart marketing*, *data mart* commerciale
Il ***data mining*** (estrazione dei dati) permette l'estrazione di un sapere o di una conoscenza a partire da grandi quantità di dati e l'utilizzazione operativa di questo sapere.
Il *data mining* ha una duplice funzione:

- Estrazione, con tecniche analitiche d'avanguardia, di informazione implicita, nascosta, da dati già strutturati, per renderla disponibile e direttamente utilizzabile.
- Esplorazione e analisi, su grandi quantità di dati, allo scopo di scoprire ***pattern*** (schemi) significativi.

Questa attività è cruciale in molti ambiti della ricerca scientifica, ma anche in altri settori.
È utilizzata per risolvere problemi diversi tra loro, dalla gestione delle relazioni con i clienti (CRM), all'individuazione di comportamenti fraudolenti, all'ottimizzazione di siti web.

BIG DATA: UNA ENORME MONTAGNA DI DATI

I dati sono intorno a noi, sono dati che arrivano dalla nostra **localizzazione geografica quando telefoniamo da un cellulare**, dai nostri profili sui **social network**, dagli **indirizzi Internet** che andiamo a visitare, dai pensieri che

esprimiamo su **Twitter**, dai **dati sanitari, economici e finanziari** che affidiamo sempre più spesso e, a volte inconsapevolmente, alle varie nuvole informatiche (***cloud computing***).

Il termine **Big Data** si riferisce alla straordinaria e crescente importanza dei dati digitali, non soltanto per quanto concerne all'aumento quantitativo dei dati in termini di byte accumulati nelle memorie digitali.
C'è infatti il bisogno di analizzare enormi quantitativi di dati in tempo reale (**data analysis, business intelligence**), dati che possono essere distribuiti in data center geograficamente diversificati.

I dati stessi possono essere di varia natura: in genere vengono classificati come dati strutturati e non-strutturati. La differenza è che un dato è strutturato quando risiede in uno specifico campo di una stringa di dati (record, file) e dipende dalla preventiva creazione di un modello dei dati, e cioè dal loro formato, dalle modalità con cui il dato è scritto, elaborato e letto nella memoria. Ne consegue che i dati strutturati possono essere identificati e analizzati con grande velocità.
Il termine **Big Data** si riferisce, ancor più precisamente, alla capacità di eseguire in tempo reale transazioni, interazioni e osservazioni sull'enorme mole di dati che si accumula sempre più vertiginosamente tramite i dispositivi connessi in rete.

Secondo molti analisti, stiamo affrontando un *Big Data* se il dato ha le caratteristiche di:

- Volume.
- Varietà.
- Velocità.

L'azienda informatica IBM, affermando che entro il 2015 il 80% dei dati sarà incerto. ha introdotto un'ulteriore dimensione:

- Veridicità

Vediamo in dettaglio ognuna di queste caratteristiche.

Big Data: Volume

Il **Volume** indica ingenti quantitativi di *dati* non gestibili con i database tradizionali. Stiamo parlando del trattamento di informazioni che partono dai terabytes ai petabytes per entrare nel mondo degli zetabytes, e che i volumi sono in continuo aumento. Per chi non fosse avvezzo a queste dimensioni, lo zetabytes è pari a ben un miliardo di terabytes.

Qualche esempio:
Un motore di un aeromobile genera 10 TB di dati /ogni 30 min e quindi un volo Roma - New York con un quadrimotore genera **640 TB di dati a tratta**.
Dal 2005 al 2012 gli RFID (*Radio Frequency ID Tag*) sono cresciuti da 1,3 a 35 miliardi.
L'RFID (*Radio Frequency IDentification* o Identificazione a radio frequenza) è una tecnologia per l'identificazione automatica di oggetti, animali e persone basata sulla capacità di memorizzare e accedere ai dati usando *transponders* o *tags*.
Il sistema si basa sulla lettura a distanza di informazioni contenute in un tag RFID usando delle specifiche antenne.
I campi di applicazione della tecnologia RFID sono molteplici e spaziano da quelli più diffusi come il controllo delle presenze e accessi, la logistica di magazzino e trasporti, la bigliettazione elettronica o l'identificazione degli animali, a quelli più all'avanguardia come il monitoraggio dei rifiuti, la monetica (il trattamento elettronico e informatico dei pagamenti), i camerini virtuali (una nuova applicazione delle tecnologie digitali diffusa nel settore della moda e nella filiera dell'abbigliamento; permette al consumatore di valutare e provare virtualmente on-line i vestiti e gli accessori-moda) e la rilevazione dei parametri ambientali.

Oggi, in meteorologia ci sono milioni di sensori, telecamere e rilevatori dislocati in tutto il pianeta che producono una quantità elevatissima di dati da misurare in tempo reale.
L'applicazione dell'analisi dei **Big Data** al clima permette di rivoluzionare il mondo delle **previsioni meteorologiche**. L'Arpa (l'Agenzia regionale per la protezione ambientale) dell'Emilia Romagna, ad esempio, ha lanciato una nuova **app** per avere tutti i dati **meteo** in tempo reale. Non solo le previsioni, ma anche temperatura, umidità, vento e tipo di pioggia.

Un software analizza le informazioni raccolte da più di **250 stazioni**, oltre alle immagini di radar appositi, per prevedere come evolverà il clima nei **tre giorni** successivi. I dati vengono poi passati e lavorati da un team di esperti che corregge le previsioni.
Per il 2020, IDC (*International Data Corporation* - società mondiale specializzata in ricerche di mercato, servizi di consulenza e organizzazione di eventi nei settori IT, TLC e dell'innovazione digitale) afferma che avremo online una quantità di byte pari ad almeno 40 volte la quantità di granelli di sabbia di tutte le spiagge della Terra.

Big Data: Varietà

La Varietà indica elementi di diversa natura e non strutturati come testi, audio, video, flussi di *click*, segnali provenienti da RFID, cellulari, sensori, transazioni commerciali di vario genere. L'era dei Big Data è caratterizzata dalla necessità e dal desiderio di esplorare anche dati non strutturati oltre e insieme alle informazioni tradizionali. Se pensiamo ad un post su un social media, un tweet o un blog, essi possono essere in un formato strutturato, ma il vero valore si trova nella parte dei dati non strutturati.

La **Varietà** è quindi riferita alle varie tipologie di dati provenienti da fonti diverse (strutturate e non); quindi parliamo di:

- **Dati strutturati in tabelle (relazionali)**
  Sono i dati sui quali si basa la tradizionale Business intelligence. I volumi sempre crescenti di dati memorizzabili e le architetture sempre più performanti rendono ancora oggi le tabelle relazionali la principale fonte dati per la *Big Data Analytics*. Tutti i sistemi gestionali esistenti producono dati strutturati o strutturabili in tabelle relazionali.
- **Dati semistrutturati (XML e standard simili)**
  È il tipo di dati che sta sfidando l'egemonia dei dati strutturati. Applicazioni transazionali e non forniscono nativamente output di dati in formato XML o in formati tipici di specifici settori. Si tratta perlopiù di dati *business-to-business* organizzabili gerarchicamente.
- **Dati di eventi e macchinari (messaggi, batch o real time, sensori, RFID e periferiche)**
  Sono i tipici dati definibili Big Data, che sino a pochi anni fa venivano memorizzati solo con profondità temporali molto brevi (massimo un mese) per problemi di *storage*.
- **Dati non strutturati (linguaggio umano, audio, video)**
  Sono enormi quantità di metadati, per lo più memorizzati sul web, dai quali è possibile estrarre informazioni strutturate attraverso tecniche avanzate di analisi semantica. Il metadato - letteralmente (dato) relativo ad un (altro) dato - è un'informazione che descrive un insieme di dati.
- **Dati non strutturati da social media (social network, blog, tweet)**
  Sono l'ultima frontiera delle fonti dati non strutturate. *Crawling* (aggregazione ed analisi di informazioni non strutturate estratte dal web), *Parsing* (o analisi sintattica: è un processo che analizza un flusso continuo di dati in ingresso, letti per esempio da un file o una tastiera, in modo da determinare la sua struttura grazie ad una data grammatica formale; il *parser* è un programma che esegue questo compito), *Entity extraction* (tecnologia per l'estrazione di entità e per la comprensione automatica di parole, frasi e interi documenti) sono tra le tecniche per l'estrazione di dati strutturati e analizzabili. I volumi aumentano esponenzialmente nel tempo. Il loro utilizzo può aprire nuovi paradigmi di analisi prima impensabili.
- **Dati dalla navigazione web (Clickstream)**
  Web Logs, *Tag javascript*, *Packet sniffing* (tutte tecniche di tracciamento e individuazione nella rete) per ottenere la *Web Analytics*. Enormi quantità di dati che portano informazioni sui consumi e le propensioni di milioni di utenti. Anche per questi dati, i volumi aumentano esponenzialmente nel tempo.
- **Dati GIS (Geospatial, GPS)**
  I dati geospaziali sono generati da applicazioni sempre più diffuse. La loro memorizzazione è ormai uno standard e i volumi sono in crescente aumento. I dati geospaziali, analizzati statisticamente e visualizzati cartograficamente, integrano i dati strutturati fornendo, ad esempio, informazioni di business, sulla sicurezza o sociali.
  Alcuni servizi di mappe in rete ad esempio forniscono in tempo reale le condizioni del traffico ricavandole dai navigatori delle persone connesse che hanno abilitato la geolocalizzazione.
- **Dati scientifici (astronomici, genetica, fisica)**

Come i dati di eventi, sono per definizione dei Big Data. Per il loro trattamento e analisi si sono sperimentate tutte le più innovative tecniche computazionali nella storia recente dell'Informatica e per questi dati sono stati progettati, nel tempo, tutti i più potenti calcolatori elettronici. I loro volumi sono enormi e in costante aumento.

Questo elenco, in ogni caso non esaustivo, indica quale sia la potenziale varietà di dati da trattare in un'applicazione sviluppata per trasformare i dati in informazioni di business.

Big Data: Velocità

Per Velocità si intende la quantità dati che affluisce e necessita di essere processata a ritmi sostenuti o in tempo reale.
Contrariamente a quanto si potrebbe pensare la velocità non si riferisce alla crescita, ovvero al volume, ma alla necessità di comprimere i tempi di gestione e analisi: in brevissimo tempo il dato può diventare obsoleto. È dunque strategico presidiare e gestire il ciclo di vita dei Big Data.

Informazioni **non aggiornate** hanno un **bassissimo valore intrinseco** e potrebbero risultare quasi inutili o addirittura **dannose**, ad esempio la geolocalizzazione che attiviamo sui nostri smartphone: se non fossero informazioni aggiornate continuamente, i percorsi ottimali sulla mappa non sarebbero reali e utili.

Il **tempestivo** allineamento delle basi di dati, l'elaborazione delle interrogazioni in **tempo reale** e la restituzione dei risultati necessitano di **tecnologie, architetture e applicazioni ottimizzate e dedicate.**

Anche quello della velocità non è comunque un concetto nuovo per le applicazioni analitiche. Da sempre ci si pone il problema di come rendere le interrogazioni più performanti, di come ottenere in tempo reale le informazioni di cui abbiamo bisogno e più in generale di come si riesce, nel più breve tempo possibile, a trasformare i dati in informazioni e le informazioni in decisioni di business.

Con lo scenario e la complessità visti in precedenza, però, le cose si complicano ulteriormente. Si ha bisogno di velocità sia per catturare rapidamente i dati sia per memorizzarli immediatamente in forma strutturata. Le informazioni estratte devono essere coerenti, confrontabili e aggiornate. Informazioni non aggiornate, anche se basate su Big Data, impoveriscono il loro valore fino a renderle inutili se non addirittura dannose. L'allineamento delle basi dati, l'elaborazione delle interrogazioni e la restituzione dei risultati necessitano di tecnologie, architetture e applicazioni ottimizzate e dedicate.

Big Data: Veridicità
Per Veridicità si intende la qualità dei dati intesa come il valore informativo che se ne può estrarre.
La Veridicità è un valore. Tutti i dati raccolti costituiscono un *valore* per un'azienda. È dall'analisi che si colgono le opportunità e supporto ai processi decisionali.

Tuttavia un grande volume di dati non garantisce da solo la qualità dei dati. Bisogna essere sicuri della loro affidabilità. La *veridicità* dei dati diventa quindi il quarto requisito fondamentale affinchè i dati possano essere considerati un valore e possano generare nuove idee.

Ciò che è immediatamente ed intuitivamente chiaro dallo studio del fenomeno dei Big Data è il fatto che si tratta di miniere di informazioni da cui si possono estrarre strutture di conoscenza e di sapere, profili di trend in atto, previsioni per l'immediato futuro spaventosamente potenti.

Ciò **che non si percepisce**, invece, è il fatto che connettendo queste singole miniere si ottiene un insieme che è molto di più della somma dei singoli *data set*. Un insieme reticolare che ci può fornire non solo **le risposte a vecchie domande, ma che può far emergere domande nuove di particolare importanza strategica.**

Le realtà del mondo IT sia private che pubbliche sono già in corsa frenetica per mettere a punto efficaci strumenti intelligenti (es. **strumenti semantici**) che permettano di analizzare e gestire queste masse di dati che non si possono affrontare con gli strumenti standard usati per catturare, gestire e processare i normali *data set* in tempi accettabili.

Forse non sappiamo immaginare che cosa si può fare con centinaia di miliardi di miliardi di byte (ovvero con le **decine/centinaia di exabyte** dei *big data* attuali, peraltro in veloce crescita) ma possiamo intuire che da questi big data, attraversabili e interpretabili con strumenti di tipo semantico, emergeranno continuamente nuovi *pattern (schemi)*:

- sia quelli che appaiono in risposta a **nostre specifiche domande**, richieste, esigenze,
- sia quelli completamente imprevisti, che emergono per pura ***serendipity*** (fare scoperte piacevoli per puro caso)

Attraverso i nuovi strumenti matematico/linguistici si riesce ad estrarre **sapere da grossi *set* di dati**, con risultati concreti non indifferenti.

BIG DATA: ANALISI PREDITTIVA E BIG DATA ANALYTICS

Le **analisi di business** sono un sistema evoluto di *data processing* che permette di acquisire una visione accurata della situazione attuale, di individuare scenari futuri e di favorire e suggerire decisioni efficaci e risultati tangibili.
Le analisi possono essere di tipo descrittivo, diagnostico, predittivo, prescrittivo e preventivo.

- **L'analisi descrittiva** rappresenta quello che accade e presenta informazioni servendosi di strumenti di Business Intelligence. In particolare, si occupa di analizzare gli eventi passati, per riassumere e chiarire le dinamiche e le *performance* delle metriche di interesse e ricavarne indicazioni su come approcciarsi alle prossime attività. Categorizzare i clienti sulla base di caratteristiche note è un esempio di questo tipo di analisi, che guarda alle informazioni disponibili e le utilizza per ottenere una visione d'insieme o di dettaglio.
- **L'analisi diagnostica** individua le cause che hanno portato alla situazione attuale. Conoscere lo stato attuale del business non è sufficiente per comprendere la situazione corrente e adottare strategie mirate: analisi diagnostiche permettono di individuare i motivi di determinate tendenze o avvenimenti, per valorizzare e ripetere le azioni più efficaci e ottimizzare le attività che non hanno portato i risultati previsti.
- **L'analisi predittiva** è la previsione di ciò che accadrà nel futuro. Permette di migliorare la comprensione del business, contribuendo a prevedere il comportamento degli utenti e le *performance* dell'organizzazione. È ampiamente utilizzata nelle strategie di *marketing digitale*, per far emergere nuove opportunità di business e per ottimizzare le campagne. Questo è possibile grazie alla raccolta, all'analisi e alla modellazione dei dati sul comportamento dei clienti e sullo storico delle performance delle campagne effettuate.
- **L'analitica prescrittiva** si spinge oltre la previsione dei risultati futuri e fornisce raccomandazioni in maniera automatica sulle azioni da intraprendere. Questa analisi è possibile grazie alla sintesi dei dati e all'utilizzo congiunto di scienze matematiche, regole di business e tecnologie denominate ***Machine Learning***.
- Per ultima, **l'analisi preventiva** indaga le ulteriori azioni da intraprendere per evitare risultati negativi, come per esempio la perdita di fidelizzazione da parte dei clienti. Questa analisi quindi si occupa della correzione e dell'ottimizzazione delle strategie e dei processi, per anticipare i possibili problemi e garantire migliori performance.

I Big Data si accumulano in mille settori: sanità, sicurezza, borsa, meteo, traffico, relazioni sociali, stili di consumo, inclinazioni sessuali, politiche e cicli economici, universo finanziario.
Generano un'economia del loro trattamento e, allo stesso tempo, strutturano e modificano l'economia secondo i criteri che producono il loro trattamento. Compongono tecnologie del sapere per favorire processi decisionali. I Big Data sono un asse nuovo, materiale e digitale, dalle conseguenze potenti e ancora tutte da esplorare sia nell'economia che nella politica.

OPEN DATA

Gli ***Open Data*** sono dati che possono essere liberamente **utilizzati, riutilizzati e ridistribuiti**, con la sola limitazione, al massimo, della richiesta di attribuzione dell'autore e della ridistribuzione senza che vengano effettuate modifiche.

Le regole e le definizioni per gli Open Data sono:

- Disponibilità e accesso: i dati devono essere distribuiti nel loro insieme e in un formato utile e modificabile.
- Riutilizzo e ridistribuzione: i dati devono essere forniti a condizioni tali da permetterne il riutilizzo e la ridistribuzione; ciò comprende la possibilità di combinarli con altri basi dati.
- Partecipazione universale: tutti devono essere in grado di usare, riutilizzare e ridistribuire i dati.
- Completezza: i dati devono comprendere tutte le componenti (metadati) che consentano di esportarli ed utilizzarli online e offline.
- Primarietà: devono essere presenti in maniera granulare.
- Tempestività: devono essere accessibili in modo rapido ed immediato.
- Accessibilità: devono essere disponibili attraverso protocolli stan dard (HTTP) e senza richiesta di alcuna sottoscrizione di contratto.
- Leggibili da computer: devono essere processabili in automatico da computer.
- In formati non proprietari: devono essere disponibili in formati aperti e pubblici.
- Liberi da licenze che ne limitino l'uso, la diffusione o la retribuzione (Es. IODL).
- Ricercabilità: deve essere assicurato agli utenti l'opportunità di ricercare con facilità e immediatezza dati e informazioni di proprio interesse, mediante strumenti di ricerca ad hoc, come database, cataloghi e search engine.
- Permanenti: le caratteristiche elencate devono perdurare per l'intero ciclo di vita del dato.

L'*Open Government* si basa sul principio per il quale tutte le attività dei governi e delle amministrazioni dello Stato devono essere aperte e disponibili per favorire azioni efficaci e garantire il controllo diffuso sulla gestione della cosa pubblica:

- **Trasparenza**: le informazioni devono essere facilmente reperibili.
- **Partecipazione**: coinvolgimento attivo dei cittadini nei processi decisionali.
- **Collaborazione**: le istituzioni non si intendono più come strutture a sé stanti ma inserite in una rete collaborativa.

L'utilizzo degli Open Data in ambito pubblico, oltre che in conformità agli standard tecnici, deve avvenire nel rispetto della normativa vigente.

Un cittadino interessato ad una particolare tipologia di avvisi pubblici deve monitorare ogni giorno il sito del proprio comune e magari quelli di tutti i comuni della provincia.
Quanto spende il mio comune per i servizi sociali? Come è distribuita la spesa? Quale è la spesa media negli altri comuni?
Oggi molte banche dati sono in possesso di privati, spesso accessibili solo a pagamento. Nel mondo anglosassone la *divulgazione* dei dati in possesso della pubblica amministrazione ha portato alla nascita di nuove imprese.
Vedi http://www.opendata500.com/
L'apertura dei dati *abilita* molte iniziative di gestione partecipata dei beni condivisi (bilancio partecipato, leggi di iniziativa popolare, co-progettazione, volontariato sociale, ...) (**Cittadinanza Attiva**).

Open Data in Italia: molti enti e pubbliche amministrazioni italiane forniscono Open Data.
Vedi http://www.dati.gov.it/, dati.gov.it è il portale dei dati aperti della pubblica amministrazione che dal 2011 ospita il catalogo degli open data pubblicati da Ministeri, Regioni ed Enti Locali. I dataset sono organizzati in maniera razionale e semplice, comparando le classificazioni di riferimento usate dalla Comunità Europea e quelle di alcuni tra i migliori portali Open Data mondiali, in modo da favorire una migliore valutazione e lo scambio di informazioni con altri stati. I focus tematici che raggruppano dataset e contenuti editoriali, sono navigabili in maniera semplificata anche dai non addetti ai lavori. L'opportunità più grande degli Open Data è rappresentata dall'interoperabilità.
Il valore dei dati è tanto più alto quanto più è possibile effettuare correlazioni tra più *dataset* indipendenti.
Per migliorare la loro gestione ed erogazione, gli **Open Data vengono arricchiti di informazioni a corredo (metadati)** che ne definiscono le proprietà significative e che saranno **funzionali alla ricerca e al recupero dei dati stessi**.

BIBLIOGRAFIA

TITOLO	AUTORE	EDIZIONI	ANNO
Business intelligence	Rezzani Alessandro	Apogeo Education	2012
Miniere di dati	Ferrari Antonella	Franco Angeli	2002
Data Warehouse. La guida completa	Margy Ross - Ralph Kimball	Hoepli	2003
Big data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà	Mayer-Schönberger Viktor;	Garzanti	2013
Social Media E Sentiment Analysis L'evoluzione dei fenomeni sociali attraverso la rete	Ceron Andrea; Curini Luigi; Iacus Stefano	Hoepli	2014
Il fenomeno open data. Indicazioni e norme per un mondo di dati aperti	Simone Aliprandi	Ledizioni (collana Copyleft Italia)	2014
Reti di indignazione e speranza Movimenti sociali nell'era di Internet	Manuel Castells	Università Bocconi Editore	2012