# Creare valore dai Big Data

## I dati

In una azienda vi sono i **manager** che hanno il compito di prendere decisioni e che necessitano prevalentemente di dati di sintesi, il **personale operativo** che rappresenta il motore delle varie funzioni aziendali e necessita di un maggior dettaglio nei dati, poi vi è il **data scientist**, che ha il compito di preparare i dati, applicare tecniche di elaborazione evolute e creare una piattaforma analitica adatta sia al management sia agli operativi.

In genere i dati presenti in azienda sono provenienti da **fonti operazionali**, ovvero che fanno riferimento all’attività operativa giornaliera dell’azienda (produzione, acquisti, vendite, contabilità, personale, clienti).

**Data warehouse** o data mart, contengono i dati integrati, coerenti e certificati dei processi di business dell’azienda e costituisce il punto di partenza per le attività analitiche del sistema di Business Intelligence

La **Business Intelligence** è un sistema di modelli, metodi, processi, persone e strumenti che rendono possibile la raccolta regolare ed organizzata del patrimonio di dati generato da un’azienda. Permette la trasformazione dei dati in informazioni, la loro conservazione e presentazione in una forma semplice, flessibile ed efficace, tale da costituire un supporto alle decisionistrategiche.

Dati di fonti esterne possono essere usati come integrazione delle anagrafiche o come oggetto di specifiche analisi (esempio: il sentiment relativo all’azienda).

I dati possono essere: strutturati, non strutturati (il testo di un post o i bit di una immagine) o semi-strutturati (ovvero hanno una parte strutturata ed una no).

Per lavorare con i dati non strutturati occorre effettuare trasformazioni in grado di fornire un risultato che possegga una struttura.

I **manager** hanno bisogno dei dati di due categorie di applicativi:

**Knowledge Management System**(KMS): informazioni riguardanti le attività dell’azienda.

**Sistemi di Business Intelligence** (BI), il cui compito è quello di fornire un supporto alle decisioni, tramite la trasformazione dei dati aziendali in informazioni.

I manager richiedono informazioni analitiche di sintesi e analisi predittive, presentate con l’ausilio di un supporto grafico tale da offrire, attraverso un semplice sguardo, la percezione dell’andamento dell’azienda o dei singoli settori.

Il **personale esecutivo** si occupa dell’operatività corrente necessita di dati di dettaglio, forniti con tempestività che sono contenuti nei seguenti applicativi:

EnterpriseResourcePlanning(ERP), strumenti gestionali per le varie aree. Supply Chain Management (SCM), per la gestione delle catene di fornitura. CustomerRelationshipManagement (CRM), che supportano il contatto con i clienti. OperationalBusinessIntelligence, con analisi di dettaglio in tempo quasi reale.

Il **data scientist**, utilizza tecniche avanzate, informatiche, statistiche o di machine learning col compito di svolgere l’analisi dei dati.

Il profilo ideale comprende numerose competenze che è difficile trovare in un’unica persona: Competenze informatiche (data base, piattaforme per big data, linguaggi di programmazione orientati alla analisi), competenze statistiche, predictive analytics e machine learning, gestione della qualità del dato, business e organizzazione aziendale, problematiche e sfide nel settore dell’azienda, colloquiare comunicare i risultati e supportare nelle decisioni tutti i destinatari.

Il migliore approccio è di creare in azienda un team di data scientist ognuno con competenze diverse.

## Big Data

Le nuove tecnologie e le piattaforme cloud hanno reso più semplice e meno costoso l’attivazione di processi volti a sfruttare le opportunità offerte dai Big Data. I sistemi quali Hadoop e Spark permettono la raccolta sistematica e l’elaborazione di grandi volumi di dati strutturati e non. Il valore aggiunto di questi dati è dato principalmente dall’analisi predittiva e prescrittiva in grado di creare vantaggi competitivi per l’azienda.

I BD sono definiti come quei dati che presentano 1 o + di queste caratteristiche:

Volume, Velocità (rapidità con cui i dati sono prodotti), Varietà (in formato fonte e struttura).

Un’altra definizione li descrive come: dati che non è possibile analizzare con tecnologie tradizionali oppure per i quali l’adozione di tecnologie tradizionali risulti molto onerosa economicamente.

Il ciclo di vita dei Big Data prevede: acquisizione, immagazzinamento e organizzazione, trasformazione e analisi.

**L’immagazzinamento** ha due problemi: la mole e l’assenza di struttura.

Il problema della mole è risolto da piattaforme tipo **Hadoop** (Open Source) che è un sistema affidabile e scalabile che sfrutta le risorse di più computer autonomi e collegati in rete (sistemi di calcolo distribuito). Con questi sistemi le varie risorse si sommano e permettono di trattare problemi non risolvibili attraverso un singolo elaboratore. **Spark** è un motore di calcolo distribuito molto efficiente può essere installato sul cluster Hadoop ma lavora anche su altri sistemi.

Una componente di Hadoop è **HDFS** (Hadoop Distributed File System) che permette l’immagazzinamento dei dati distribuito ma con una logica di ridondanza che permette l’integrità anche in caso di guasto ad uno dei nodi. La componente **MapReduce** invece si occupa di gestire la distribuzione dei carichi elaborativi e lavora secondo il principio *divide et impera*.

HDFS non è un DB e non replica tutte le funzionalità (come fa ad esempio HBase) però gestisce anche dati non strutturati. Per gestire dati non strutturati ci sono anche altri sistemi che fanno parte dei cosiddetti database NoSQL (Cassandra, Berkeley DB, MongoDB, Neo4J) che non utilizzano il concetto di tabella.

La **trasformazione** e **l’analisi** nell’ecosistema Hadoop sfruttano la componente MapReduce, esistono alcuni strumenti più ad alto livello che permettono di interfacciarsi con MapReduce, ad esempio Pig (simile al SQL), Hive (consente l’aggregazione di dati e l’interrogazione), **Mahout** (una piattaforma di machine learning dedicata al clustering e alla classificazione). Anche Spark ha numerose funzionalità di calcolo non è Hadoop ma è completamente integrata.

Tra le alternative ad Hadoop citiamo i database definiti GPU based. Le **GPU** (Graphic Processing Unit), sono processori nati per la gestione della grafica, ma che si sono rivelati molto utili anche in altri campi, dove la potenza di calcolo è un requisito importante. La struttura delle GPU è tale da permettere che singole operazioni siano eseguite su blocchi di dati molto più grandi rispetto a quelli elaborati dalle normali CPU, garantendo così un maggiore throughput.

Casi particolari:

Il settore **banking**: l’utilizzo di infrastrutture Hadoop-Spark hanno permesso di realizzare analisi classiche su dati di dettaglio (su singole transazioni invece di saldi mensili o con profondità temporali superiori) il db delle transazioni di grandi banche conta diversi milioni di record al giorno la cui sintesi mensile comporta una perdita di informazione. Le infrastrutture Big Data permettono di realizzare strumenti di analisi come clustering, churn, e modelli predittivi in tempi ragionevoli. Inoltre permettono di mantenere un’adeguata storicizzazione delle transazioni sia per scopi analitici e per imposizioni normative.

**Industry 4.0**: È una parola di moda che non ha una vera e propria definizione; la possiamo descrivere come un processo che ha il suo punto di arrivo nella produzione industriale (quasi) completamente automatizzata e interconnessa. Alcuni fattori sono la disponibilità di nuove generazioni di macchinari (stampa 3D, innovazioni nella robotica, comunicazioni machine-to-machine, additive manufacturing) e nuove tecnologie di elaborazione dati: (IOT, cloud computing, advanced analytics col machine learning). Un esempio già implementato è l’analisi, in tempo quasi reale, della situazione degli impianti: gli strumenti di stream analytics sono in grado di acquisire e monitorare i dati dei sensori con un tempo di bassissima latenza in modo da verificare il corretto funzionamento dei macchinari e di avere una visione in tempo reale dei parametri di produzione; inoltre con modelli predittivi si è in grado di identificare eventuali anomalie o anticipare possibili guasti potendo così programmare fermi macchina o manutenzioni evitando disagi imprevisti (proactive maintenance).

**Internet Of Things** (IOT): riguarda oggetti dotati di sensori e di connettività e che producono dati (principalmente legata ai prodotti consumer). Raccolgono informazioni sulle performance e sulle modalità di utilizzo del prodotto. Ad esempio scatole nere istallate sulle auto dalle assicurazioni RCA, braccialetti con RFID (Radio Frequency Identification) per monitorare gli spostamenti in un parco giochi e sensori vari nei prodotti. I casi d’uso vanno dalla manutenzione proattiva alla proposta di contenuti digitali per i sistemi di entertainment, all’ottimizzazione dei percorsi, etc.

**Smart City:** prevede l’utilizzo intensivo di tecnologie informatiche e di comunicazione per raggiungere elevati livelli di efficienza in vari settori (consumo energetico, l’inquinamento, la mobilità e la sicurezza nelle città) utilizzando dati provenienti da sensori o dai social network. Per esempio secchi che segnalano quando sono pieni per ottimizzare la raccolta, analisi del traffico per calibrare i tempi dei semafori.

Il **data lake** è l’architettura deputata alla raccolta e all’utilizzo dei big data; è costituito da un insieme di componenti in grado di realizzare le fasi del ciclo di vita dei dati: strumenti di data ingestion, strumenti per le trasformazioni e la preparazione dei dati per le analisi, strumenti di analytics per le analisi predittive. Il data lake centralizza i dati provenienti da fonti diverse nel loro formato originale poi vengono estratte, attraverso un processo di trasformazione, le informazioni di interesse delle analisi che vengono registrate in struttare tabellari (dati semi-strutturati).

## Le tecniche di analisi

**Analisi descrittiva**: analizzare i dati del passato, categorizzandoli, filtrandoli, aggregandoli ed applicando funzioni descrittive. I motori OLAP permettono operazioni di slicing (filtro modalità), dicing (filtro modalità non visibili), drill down(+), roll-up (-), drill-trought (accesso al micro), pivoting (cambio dell’ordine di visualizzazione).

L’analisi descrittiva è alla base dei successivi tipi di analisi e permette di identificare problemi, inefficienze e aree di maggior profitto e gli attributi che li determinano (data discovery). Quando la quantità e la complessità aumenta per volume o per numero di attributi l’analisi interattiva non è sufficiente e occorre affidarsi a tecniche automatiche.

**Analisi predittiva** (o data mining):estrae una visione del futuro dai dati del passato usando un modello matematico/statistico o il machine-learning.

**Analisi prescrittiva**: oltre ad una predizione del futuro ne esplicitano le motivazioni attraverso un insieme di regole. Classi: alberi decisionali, Fuzzy Rule-Based System, Logic Learning Machine.

Altri sistemi predittivi come Reti Neurali, Support Vector Machines, o Deep Learning, sono molto efficaci come predizione ma sono delle black-box machine.

[saltate alcune possibili aree di applicazione: CRM – Ricerca di anomalie – Marketing – Altri utilizzi]

# Hadoop

## Hadoop: piattaforma e motore

Hadoop è una piattaforma software open source, affidabile e scalabile finalizzata al calcolo distribuito (prima release del 2008), cioè al calcolo su un sistema di computer autonomi e collegati tra di loro in rete.

Le componenti che costituiscono il nucleo centrale sono:

**Hadoop common**: uno strato di software comune che fornisce funzioni di supporto agli altri moduli

**HDFS**: il file system distribuito che fornisce un’efficace modalità di accesso ai dati e garantisce la ridondanza

**YARN**: un framework che si occupa della gestione delle risorse (memoria/CPU/storage) del cluster e consente di creare applicazioni per il calcolo distribuito

**MapReduce**: sistema di parallel processing che lavora secondo il principio divide et impera

Fino a 4.4