# ***CAPITOLO 1: STATO DELL’ARTE***

## ***1.1 COSA SONO I BIG DATA?***

Il termine big data indica una raccolta di dati così estesa in termini di *volume*, *velocità* e *varietà* da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza. Il termine è utilizzato in riferimento alla capacità di analizzare ovvero estrapolare e mettere in relazione un'enorme mole di dati eterogenei, strutturati e non strutturati, allo scopo di scoprire i legami tra fenomeni diversi e prevedere quelli futuri.

Si tratta di una definizione soggettiva in cui non viene esplicitata la dimensione che deve avere un Dataset per poter essere definito Big Data, in quanto, essa aumenterà sicuramente nel tempo grazie i continui avanzamenti tecnologici. Le dimensioni variano nei diversi settori, da dozzine di terabyte a centinaia di petabyte (1000 terabyte), in base anche agli svariati strumenti software a disposizione.

In questa definizione emergono le cosiddette 5V che caratterizzano i Big Data, ovvero il volume, velocità, varietà, veridicità e valore [2].

Il *volume* fa appunto riferimento all’enorme massa di dati generata attraverso numerosi canali.

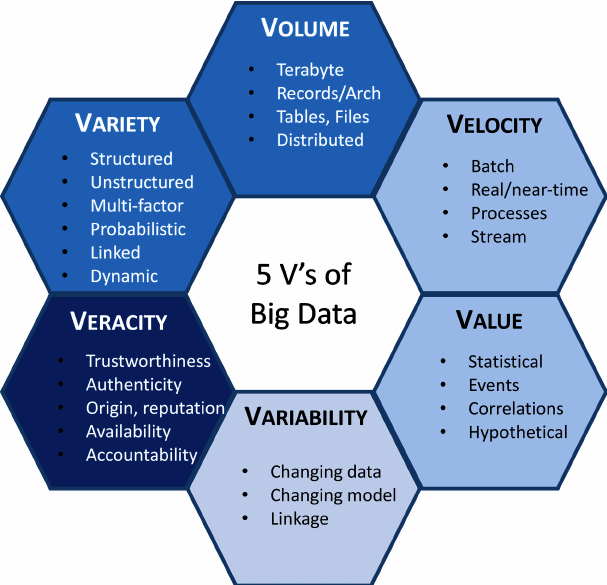
La *velocità* si riferisce alla rapidità con cui i dati vengono acquisiti e utilizzati, in aumento date transazioni sempre più frequenti e veloci: le aziende non solo raccolgono i dati più velocemente, ma cercano di sfruttarli il prima possibile, spesso in real time.

Figura 1: Le 5V dei Big Data

La *varietà* è legata alle differenti tipologie di dati disponibili provenienti da un numero crescente di fonti di dati sia strutturati sia non strutturati; in particolare è possibile identificare cinque categorie di informazioni che costituiscono i Big Data:

* dati generati da smartphone e altri dispositivi mobile relativi a persone, attività e localizzazione, tra cui dati RFID (radio-frequency identification), dispositivi che tracciano il prodotto,e dati da dispositivi di controllo come i contatori per il monitoraggio dell’acqua o del gas;
* dati di vendita e pricing, dati generati dall’attività delle carte fedeltà e degli eventi promozionali;
* computer log Data, come i clickstreams dai siti web;
* informazioni dai social media come Twitter e Facebook;
* social multimediali e altre informazioni da YouTube e siti simili.

La *veridicità* riguarda la questione relativa alla qualità dei dati e al loro livello di sicurezza, la cui garanzia rappresenta una sfida molto importante. Per poter sfruttare i Big Data è necessario saper agire per poter estrarne *valore* e quindi incrementare la produttività e la competitività delle aziende e creare surplus economico per i consumatori.

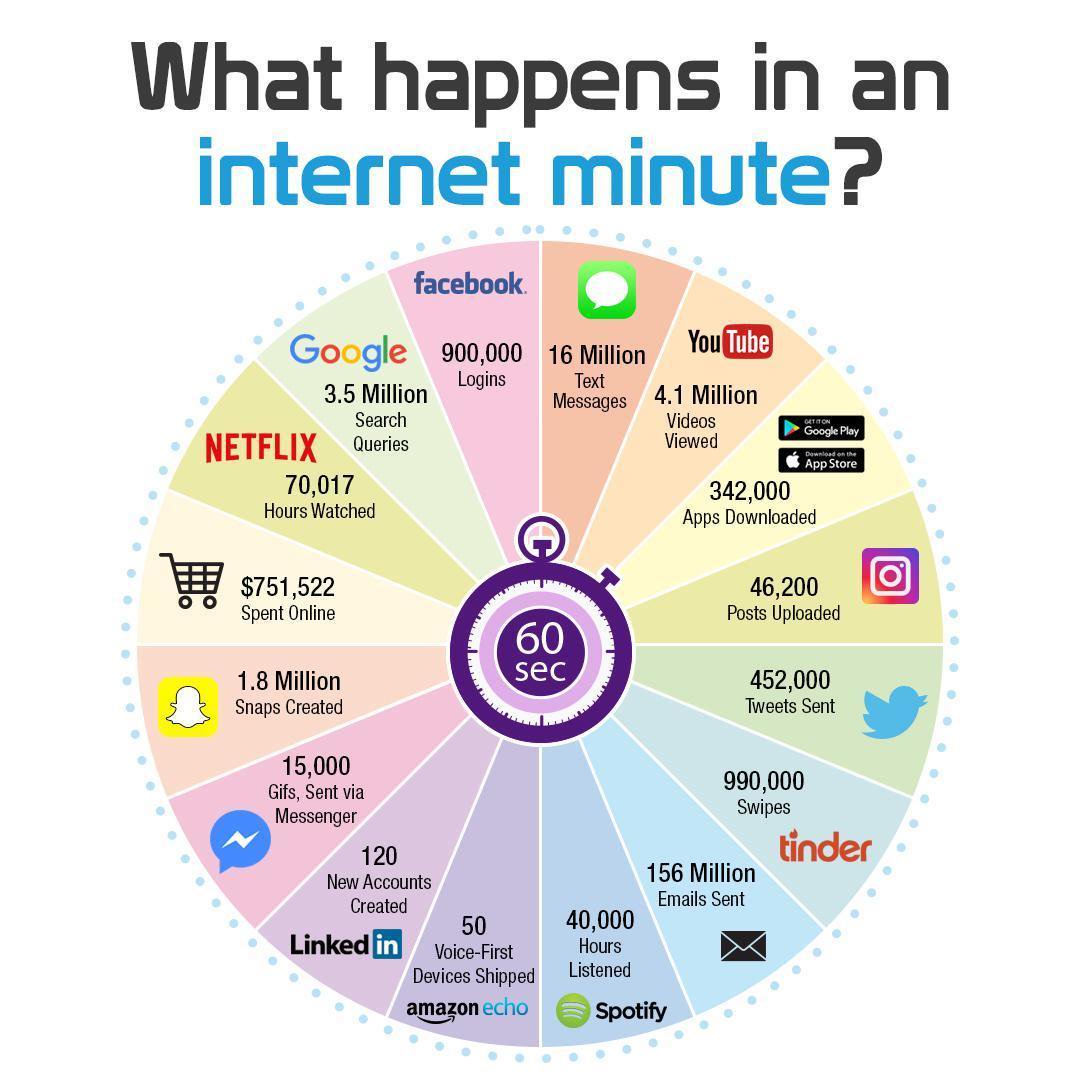
La capacità di memorizzare e aggregare i dati e quindi di utilizzare i risultati per svolgere analisi profonde migliora continuamente grazie alla disponibilità di strumenti software e tecniche sempre più sofisticate combinate a una crescente potenza di calcolo. Stiamo assistendo anche ad un enorme cambiamento della capacità di generare, comunicare, condividere e accedere ai dati dovuto all’aumento del numero di persone, strumenti e sensori ora connessi da reti digitali. Per capire la grandezza del fenomeno, basta osservare la figura nella pagina successiva che mostra quanti dati vengono generati in un minuto.

Figura 2: Dati generati in un minuto nei social

I Big Data rappresentano una grande opportunità per le aziende e per le economie nazionali in quanto consentono di ottenere benefici significativi, che elenchiamo di seguito.

* *Rivelare le variabilità delle performance e migliorare le prestazioni*. La creazione e la memorizzazione di dati transazionali in forma digitale consente alle aziende di avere dati più accurati e dettagliati su svariate performance, dallo stato dei magazzini ai giorni di malattia del personale, in tempo reale o quasi. Inoltre, esse utilizzando i dati per analizzare la variabilità delle prestazioni e per capirne le cause più profonde, possono ottenere risultati migliori.
* *Personalizzare le azioni.* I Big Data consentono di creare specifici segmenti di clienti e di personalizzare prodotti e servizi sulla base delle loro esigenze. Utilizzare tecniche di Big Data per realizzare promozioni e pubblicità personalizzate per i diversi cluster.
* *Migliorare le previsioni e Supportare le persone nel processo di decision making.* Utilizzando Analytics sofisticati su interi Dataset è possibile automatizzare e migliorare i processi decisionali tramite le predizioni dei Key Performance Indicators (KPI), minimizzare i rischi e scoprire preziosi insight; Questi benefici naturalmente non possono essere perseguiti con l’analisi e la gestione di piccoli campioni di dati tramite i fogli di calcolo. I rivenditori per esempio possono utilizzare algoritmi che consentono la messa a punto automatica e l’ottimizzazione degli inventari e dei prezzi a partire dai dati in tempo reale relativi alle vendite nei negozi e a quelle online.
* *Creare trasparenza.* Un accesso facile e tempestivo ai Big Data rende disponibile una maggiore quantità di informazione e facilita la condivisione dei dati tra le diverse unità organizzative di un’impresa.
* *Scoprire i comportamenti nascosti e i bisogni dei consumatori.* La disponibilità quasi in real time di dati da smartphone fornisce caratteristiche dettagliate sui clienti e sul loro complesso processo decisionale quando fanno acquisti: i Big Data permettono infatti di identificare i modelli comportamentali dei consumatori e far luce sulle loro intenzioni.
* *Creare nuovi prodotti e servizi, nuove tipologie di aziende e innovativi modelli di business.* Le società possono sfruttare i Big Data per realizzare nuovi prodotti e nuovi servizi: molte imprese manifatturiere per esempio stanno utilizzando i dati relativi all’utilizzo di prodotti attuali per migliorare lo sviluppo di modelli futuri e per creare servizi post-vendita innovativi; la disponibilità in real time di dati relativi alla location sta comportando lo sviluppo di nuovi servizi che si servono di questi dati, come le assicurazioni danni basate su dove e come le persone guidano le loro automobili.
* *Incrementare la produttività e la profittabilità delle aziende.* Lo sfruttamento dei Big Data può portare ad un aumento dell’efficacia e dell’efficienza delle imprese, le quali potranno realizzare più output utilizzando meno input e migliorare il livello di qualità dell’output stesso.

Questo elenco di benefici mette in evidenza come l’investimento nei Big Data porti alla creazione di valore per le aziende e quindi all’ottenimento di vantaggio competitivo nel lungo termine. Risulta quindi fondamentale per loro sviluppare competenze in questo ambito, pena il declino in un mondo Big Data.

### ***1.1.1 Barriere All’utilizzo Dei Big Data***

Nonostante le opportunità offerte dai Big Data siano enormi, c’è ancora un certo scetticismo all’interno delle aziende sui reali benefici apportati a causa degli scarsi risultati ottenuti in pratica. In uno studio condotto di recente (fine 2014) dal MicKinsey Global Institute (MGI) rivolto agli Analytics leader di alcune importanti aziende americane impegnate nella realizzazione di progetti di Big Data e di advanced Analytics, è emerso come l’utilizzo di queste tecniche abbia portato ad un aumento dei ricavi e ad un abbassamento dei costi inferiore all’1% per i tre quarti degli intervistati [3].

Esistono quindi una serie di barriere all’utilizzo dei Big Data da considerare, che possono essere classificate in 6 categorie: barriere tecniche, barriere legate alle competenze, barriere organizzative/gestionali, barriere culturali, barriere economiche e barriere legate alla privacy.

* *Barriere tecniche*: Difficoltà di integrazione dei dati, Basso grado di influenza del business, Scarsa qualità dei dati
* *Barriere legate alle competenze*: Difficoltà di comprensione degli strumenti analitici e di quantificazione dei benefici, Carenza di talenti, Difficoltà nella scelta del tool adatto.
* *Barriere organizzative/gestionali*: Mancanza di commitment dei I top manager che non sono coinvolti nelle iniziative di Big Data, verso le quali mostrano poco interesse e quindi non forniscono l’aiuto necessario
* *Barriere culturali*: Inerzia. La maggior parte delle aziende non è ancora pronta e del tutto aperta alle innovazioni che i Big Data potrebbero portare, in quanto il loro sfruttamento richiederebbe significativi cambiamenti culturali e organizzativi.
* *Barriere economiche*: Le iniziative Big Data richiedono ingenti spese in termini di tecnologie implementate e di nuove figure professionali da assumere
* *Barriere legate alle privacy:* I consumatori non vogliono che le loro informazioni personali, come i personal location Data e i dati elettronici generati dal loro uso di Internet, vengano utilizzate dalle aziende, soprattutto perchè non sanno dove e come queste verranno sfruttate dalle organizzazioni, le quali devono considerare anche le leggi relative alle privacy dei diversi Paesi. Tools che consentono di tracciare ogni movimento dei dipendenti e di misurare continuamente le loro performance fanno gli interessi delle organizzazioni e non dei singoli individui, che vedono minacciata la loro privacy. Le aziende devono quindi preservare la privacy individuale e questo le limita nello sfruttamento dei Big Data.

### ***1.1.2 Tecniche Per L’analisi Dei Big Data***

Di seguito riportiamo le principali tecniche e le tecnologie utilizzate per aggregare, manipolare, gestire e analizzare i Big Data.

* *A/B testing*: tecnica in cui un gruppo di controllo viene confrontato con gruppi di test al fine di determinare quali modifiche e azioni miglioreranno una data variabile obiettivo, come il tasso di risposta a una campagna di Marketing.
* *Crowdsourcing:* tecnica utilizzata per raccogliere dati, sottoposta a un grande gruppo di persone o a una comunità, attraverso per esempio i network media come il Web.
* *Data fusion e Data integration*: insieme di tecniche che integrano e analizzano dati provenienti da diverse fonti al fine di sviluppare insight più efficienti e accurati rispetto a quelli ottenuti esaminando una singola fonte.
* *Modelli predittivi*: tecniche in cui viene creato o scelto un modello matematico per prevedere la probabilità di un risultato.
* *Data mining*: insieme di tecniche di classificazione, cluster analysis, regole associative e regressione, che permette di estrarre modelli da grandi dataset combinando metodi statistici e di machine learning con la gestione dei database.
* *Machine Learning*: parte della computer science riguardante la progettazione e lo sviluppo di algoritmi che consentono ai computer di identificare i comportamenti basandosi su dati empirici e in particolare di riconoscere schemi complessi e prendere decisioni intelligenti.
* Natural language processing (NLP): insieme di tecniche di computer science e linguistica che si ricorrono ai computer per analizzare il linguaggio umano.
* *Classificazione*: insieme di tecniche che permettono di identificare a quali categorie appartengono nuovi dati, basandosi su un training set i cui i dati sono già stati categorizzati.
* *Cluster analysis*: metodo statistico per classificare gli oggetti, che divide un grande gruppo in piccoli gruppi caratterizzati internamente da omogeneità non nota in anticipo.
* *Network analysis*: insieme di tecniche utilizzate per caratterizzare le relazioni tra nodi in un grafo o in una rete. Nella social network analysis vengono analizzate le relazioni tra individui di una comunità o di un’organizzazione, per esempio come viaggiano le informazioni.
* *Regole associative*: tecniche volte a scoprire relazioni interessanti tra variabili all’interno di un grande dataset.
* *Regressione*: set di tecniche che permettono di determinare come il valore di una variabile dipendente cambia quando una o più variabili indipendenti vengono modificate.
* *Ottimizzazione:* insieme di tecniche numeriche utilizzate per riprogettare sistemi complessi e processi al fine di migliorare le performance relativamente a uno o più aspetti, tra cui costi, velocità e affidabilità.
* *Sentiment Analysis*: applicazione del processing natural language e di altre tecniche analitiche per identificare ed estrarre informazioni soggettive dai testi, per esempio la “polarità” (positiva, negativa o neutra) delle caratteristiche o dei prodotti su cui le persone hanno espresso un’opinione e il grado e la forza dell’opinione stessa.
* *Statistica*: scienza della raccolta, organizzazione e interpretazione dei dati, utilizzata per esprimere giudizi sulle relazioni tra variabili che potrebbero essersi verificate per caso (ipotesi nulla) e su quelle causali (statisticamente significative).
* *Data Visualization*: tecniche di creazione di immagini, diagrammi o animazioni che consentono di comunicare, capire e migliorare i risultati dell’analisi dei Big Data.

## ***1.2 PROGETTI DI BIG DATA NEL MARKETING***

Lo sfruttamento dei Big Data in ambito Marketing rappresenta un enorme potenziale, tanto che le aziende si stanno dedicando e hanno un grande interesse verso progetti che prevedono il loro utilizzo in quest’area. Oltre ai Social Analytics che affronteremo nel capitolo successivo, possiamo classificarne altri 6: il Direct e il Digital Marketing, la Customer Micro-Segmentation, il Location-based Marketing, Price Optimization, l’In-store Analysis e il Cross-Selling/Up-Selling.

### ***1.2.1 Direct E Digital Marketing***

Il Direct Marketing comprende tutte le tecniche di Marketing che consentono alle aziende di comunicare in modo mirato e personalizzato direttamente con il cliente o l’utente finale. La continua e significativa crescita di internet e della sua importanza ha comportato il rapido sviluppo del Digital Marketing, che assume la forma di display advertising, contenuti su Facebook, video clip su Youtube, e-mail personalizzate e molto altro. Le aziende per fare Digital Marketing oggi possono contare sull’enorme ammontare di informazioni degli utenti, che trascorrono ore e ore al giorno su Internet, relative ai loro interessi, ai contenuti delle loro comunicazioni, agli acquisti che fanno e molto altro [4].

Il Direct Marketing si serve di molte tecniche di Big Data, oltre che per identificare i clienti più profittevoli e quelli che risponderanno con maggiore probabilità, soprattutto per profilare i clienti, in modo da prevedere anche il comportamento di quelli sconosciuti. Vengono utilizzate sia tecniche di apprendimento supervisionato, come i modelli di ottimizzazione, le reti neurali bayesiane e gli alberi decisionali sia quelle non supervisionate, tra cui il clustering. Per ottenere risultati migliori l’ideale è combinare diverse tecniche.

I Big Data consentono perciò di guadagnare insight per profilare clienti e prospect efficacemente, permettendo quindi di rivolgersi a target molto più stretti e precisi e di conseguenza di ridurre la dispersione dei contenuti pubblicitari, con ovvi vantaggi anche in termini economici [5].

I vantaggi apportati dai Big Data al Direct Marketing sono, oltre alla personalizzazione del messaggio, la visione a 360° del cliente, l’identificazione dei contenuti, del timing e del canale più appropriato per inviare il messaggio e la possibilità di fare questo in real time.

Da ciò deriva un incremento del tasso di conversione, ovvero del numero di visitatori che decidono di cliccare su un certo contenuto casuale o di visitare un sito web come risultato di un’azione guidata, e quindi la massimizzazione del Digital ROI, l’acquisizione di nuovi clienti e la fidelizzazione di quelli che già si rivolgono all’azienda.

### ***1.2.2 Customer Micro-Segmentation***

L’era dei Big Data ha creato la segmentazione 2.0: la molteplicità di nuove tipologie di dati e lo sviluppo di Analytics avanzati permette di avere dettagli granulari e un numero maggiore di informazioni sui consumatori e quindi di generare micro-segmenti molto precisi, costituti da un piccolo numero di persone. Molti Retailer affermano addirittura di essere impegnati nella personalizzazione e non più nella semplice segmentazione [1]. I tradizionali segmenti B2C e B2B basati rispettivamente su dati demografici, psicografici e comportamentali e sulle dimensioni delle aziende o sui criteri di acquisto adottati sono ormai superati.

Sfruttando quindi:

• *Activity-Based Data*: clickstream Data dal web, le storie degli acquisti, i dati dei call center, i dati mobile;

• *Profili dei social network*: la storia lavorativa e l’appartenenza a gruppi;

• *Sentiment Data*: associazioni a prodotti e aziende (like o follows) e commenti online;

• *Dati tradizionali*: dati delle ricerche di mercato e quelli transazionali;

è possibile costruire segmenti molto più stretti. Gli uomini di Marketing possono quindi creare offerte, prodotti e servizi personalizzati e su misura per ciascun cluster, con ovvi benefici sui ritorni. Questi dati possono inoltre essere aggiornati in real time, riuscendo quindi ad identificare i cambiamenti dei clienti e delle loro preferenze.

### ***1.2.3 Price Optimization***

Le aziende possono sfruttare la crescente granularità dei dati sul pricing e sulle vendite e i potenti Analytics per ottimizzare i prezzi. L’ammontare di informazioni a loro disposizione è enorme, dalle serie storiche della domanda, ai dati relativi alle scorte, a quelli riguardanti i competitor, fino al livello delle vendite attuali. Questa base di dati è in continuo aumento considerata l’esplosione di nuovi canali di vendita online dove i consumatori possono confrontare i prezzi, pratica che ha incrementato la competizione e la frequenza nelle variazioni dei prezzi stessi e considerate le tecnologie digitali e i social media attraverso i quali è possibile tracciare il comportamento di milioni di utenti [6].

Da queste ingenti quantità di dati, attraverso opportuni tools, i pricing manager sono in grado di estrarre insight per definire il prezzo ottimale che un consumatore è disposto a pagare per ciascun prodotto quasi in real- time, basandosi sulle sue caratteristiche.

La price optimization può considerare per esempio l’elasticità della domanda al prezzo, con specifici modelli che analizzano i dati delle vendite storiche per ricavare insight sul pricing di ciascuna unità, che possono poi essere utilizzati per fare promozioni o per ridurre i prezzi, valutando i costi conseguenti. I benefici che le aziende riescono a conseguire in questo modo sono un aumento dei ricavi, dei margini e della quota di mercato.

Tuttavia, per riuscire a sfruttare i Big Data in quest’area è necessario costruire una fiducia verso il cliente, Identificare le opportunità più promettenti, che comprendono determinare quanto il consumatore vuole pagare esattamente per un dato prodotto attraverso la customer segmentation e le promozioni personalizzate, e Ascoltare i dati su cui le organizzazioni devono saper far leva e utilizzare adeguati Analytics per identificare elementi che spesso vengono trascuratie determinare quindi i fattori guida dei prezzi per ciascun cliente e prodotto [7].

### ***1.2.4 Location-Based Marketing***

Il Location-based Marketing si basa sull’adozione crescente di smartphone e di altri device mobile che generano i personal location Data, i quali permettono di conoscere dove si trovano le persone in real-time e offrono alle aziende la possibilità di tracciare geograficamente tramite GPS o i WI-FI il comportamento del cliente, favorendo lo sviluppo di una strategia di Marketing che considera le abitudini lavorative e di divertimento e non solo le preferenze dei consumatori. Altre fonti sono i segnali delle torri di triangolazione cellulare e i pagamenti tramite carte di credito e di debito, le quali, attraverso il terminale del punto di vendita, rendono disponibili i dati di identificazione personale.

Quello che le aziende fanno solitamente è il Geo-targeted advertising, ovvero effettuare azioni di advertising in tempo reale in base alla localizzazione dei propri clienti. Infatti, per ottenere enormi vantaggi, le aziende ricorrono alle push notifications, cioè delle offerte customizzate e aggiornate per un determinato cliente mentre, per esempio, cammina con in mano lo smartphone all’interno del negozio.

Pertanto, lo sfruttamento dei dati di geolocalizzazione a disposizione delle aziende per implementare azioni di Marketing mirate può portare ad un aumento delle vendite, quindi ad un incremento dei profitti e ad un miglioramento della customer experience e perciò alla fidelizzazione della clientela.

Tuttavia, relativamente a questo progetto le aziende si trovano a dover affrontare due sfide: La privacy ed un trade-off cioè se gli utenti desiderano ricevere offerte mobile quando si trovano in prossimità dello store stesso.

### ***1.2.5 In-Store Analysis***

L’in-store analysis prevede l’analisi dei dati relativi al comportamento dei consumatori negli store in real time, è un altro importante progetto Big Data in ambito Marketing.

La posizione e il percorso dei clienti all’interno dello store vengono tracciati attraverso svariate tecnologie: video camere, Wi-Fi e strumenti bluetooth, Wi-fi degli ospiti, sistemi dei punti di vendita, carte di pagamento, trasponder dei carrelli, applicazioni degli smartphone, quali Shopkick e Path Intelligence e tag RFID sulle carte d’acquisto.

Tools e Analytics, quali web dashboard, app mobile, real-time alert, strumenti di Data mining, vengono utilizzati per organizzare, analizzare e visualizzare questo grande ammontare di dati, identificare trend e confrontare le prestazioni dei diversi periodi. Così facendo vengono estratti insight relativi ai comportamenti dei consumatori all’interno dello store, con l’obiettivo ultimo di migliorare la customer experience.

In particolare, gli insight ottenuti sono relativi a:

• chi è il consumatore: conoscere sesso, età, se è la prima volta che entra nel negozio, se ritorna spesso, da dove viene e quali sono i suoi interessi

• quanti clienti entrano nel negozio

• come si comportano gli shopper all’interno dello store.

Le aziende si servono di questi insight per migliorare efficacemente l’organizzazione, ovvero per ottimizzare il layout dello store, le sue caratteristiche, il posizionamento sugli scaffali e il mix di prodotti offerti per trasformare i clienti una tantum in clienti abituali, incrementare la frequenza delle loro visite e delle loro spese migliorando la store experience, aumentare la dimensione media della transazione e attirare un numero sempre maggiore di consumatori. Le organizzazioni mettono in atto aggiustamenti in tempo reale per ottimizzare l’intero processo d’acquisto.

### ***1.2.6 Cross-Selling / Up-Selling***

I Big Data offrono grandi opportunità per aumentare la dimensione media dell’acquisto di un consumatore sia mettendo a disposizione prodotti o servizi collegati con la scelta d’acquisto iniziale sia offrendo qualcosa di maggior valore rispetto a questa, ovvero per migliorare le azioni di Cross-Selling e di Up-Selling. Dati quali le caratteristiche demografiche dei clienti, la posizione real-time, le preferenze, la storia degli acquisti passati vengono utilizzati a tal fine. Gli algoritmi sviluppati si basano su questi dati per prevedere il comportamento dei consumatori in vari scenari di vendita, estrarre insight per capire molto prima cosa vogliono e determinare quindi il miglior approccio [9].

I benefici che le aziende traggono sono un aumento delle vendite e quindi dei profitti e la fidelizzazione dei clienti.

Caso esemplare è quello di Amazon che raccoglie i dati da tutti gli utenti, riconoscendo i trend nelle persone che fanno acquisti simili attraverso tools di Analytics, in modo da cogliere potenziali opportunità e in base a ciascun prodotto o servizio visitato dall’utente sul sito suggerisce “potresti anche volere” ed è proprio in questo modo che riesce ad incrementare significativamente le vendite.

## ***1.3 BUSINESS INTELLIGENCE***

Con il termine Business Intelligence (BI) ci si riferisce ad una serie di processi aziendali che ruotano attorno ai dati, con operazioni di raccolta, elaborazione, analisi, cui scopo è quello di produrre informazioni al servizio del management strategico e tattico, che trova supporto analitico, storico e previsionale alle proprie decisioni. La BI è stata collocata altresì nel sottoinsieme operativo, poiché sta assumendo un ruolo sempre più importante anche nelle normali attività giornaliere delle aziende.

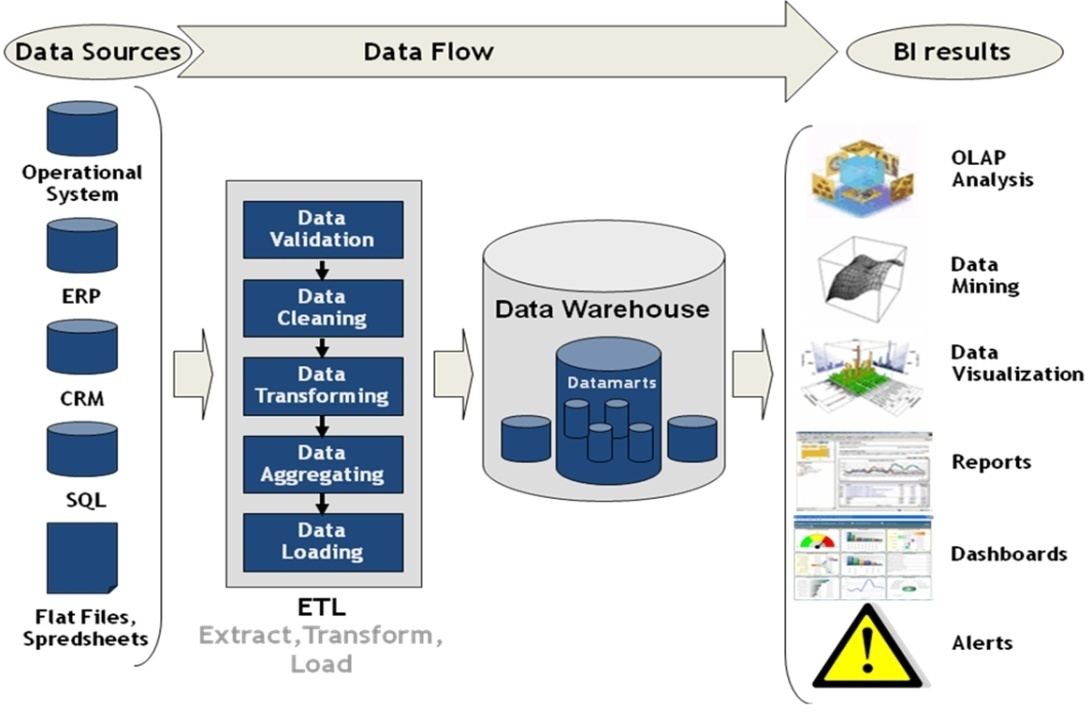


Figura 3: Business Intelligence Framework

Nell’ambito lavorativo moderno, in particolare nel campo della consulenza le aziende, il cui scopo principale è fare business e diventare leader di mercato, si trovano sempre più frequentemente a confrontarsi con realtà differenti dalla propria. Ciò avviene tramite l’analisi del comportamento di aziende facente parte dello stesso settore e lo studio del mercato in cui si trovano.

L’adozione della BI da parte delle imprese permette una conoscenza più approfondita non solo di loro stesse ma anche del mercato di riferimento.

Nel periodo attuale, il “cambiamento” è all’ordine del giorno, pertanto saper leggere in anticipo le tendenze dei mercati è un fattore competitivo a cui non si può e non si deve rinunciare.

Data l’elevata mole di dati generata ogni giorno, diventa necessario trovare un metodo che:

Permetta di raccogliere e processare dati ad alta velocità (sempre più spesso si parla di processi Real-Time);

* Fornisca un servizio di pulizia del dato stesso, eliminando dati sporchi, ridondanti o errati tramite i processi di ETL “Extraction, Transformation & Loading” che preleva i dati dai sistemi alimentanti (ERP, fogli Excel etc.) e li porta nel DWH certificandoli attraverso processi di Data Quality. Questo processo sarà spiegato in maniera specifica nel capitolo 2;
* Definisca un sistema consolidato e stabile di memorizzazione per i dati certificati (Data Warehouse);
* Trasformi l’informazione in fonte di conoscenza attraverso analisi di business sui dati stessi, determinando nuovi KPI su cui fare nuove analisi.

### ***1.3.1 Tecnologie Di Business Intelligence & Analytics***

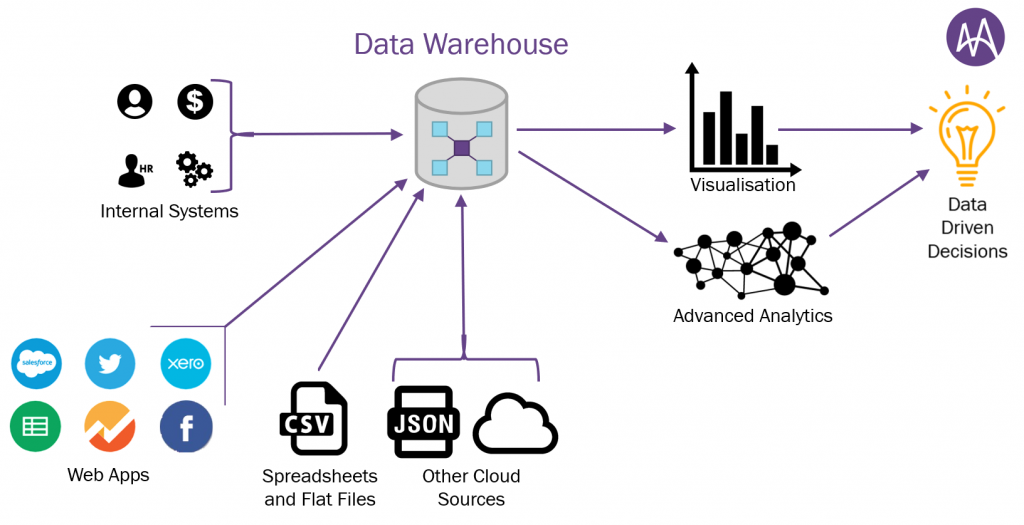


Figura 4: Schema Business Analytics

I Big Data provengono da divere fonti, sia interne che esterne, spesso sono in formati differenti e risiedono in posizioni multiple in numerosi sistemi legacy e altre applicazioni. I dati possono essere sia strutturati (dati conservati in Database relazionali, organizzati secondo schemi e tabelle rigide), sia non strutturati (dati conservati senza alcuno schema come forme libere di testo tra cui articoli e parti di e-mail, audio senza tag, immagini e video) sia semi-strutturati (dati che presentano caratteristiche sia di quelli strutturati che di quelli non strutturati; un esempio è rappresentato dai file compilati con sintassi XML per i quali non ci sono ci sono limiti strutturali all’inserimento dei dati, ma le informazioni vengono organizzate secondo logiche strutturate e interoperabili). Dopo che i dati sono stati uniti, questi hanno bisogno di essere processati o trasformati, essendo in uno stato grezzo. Ci sono diverse opzioni a disposizione:

* Service-oriented architecture combinata con web services (middleware): i dati rimangono grezzi e i services sono utilizzati per chiamare, recuperare e processare i dati;
* Data warehousing: dati provenienti da svariate fonti vengono aggregati e preparati per essere processati, anche se non sono disponibili in real-time;
* Extract, transform and load (ETL): dati che derivano da diversi fonti vengono puliti e preparati per lo step successivo.

Il passo successivo consiste nella scelta della piattaforma e della tecnologia da utilizzare, tra quelle elencate nella tabella. L’ultima fase invece è relativa alle applicazioni di Big Data Analytics che includono queries, reports, OLAP e Data mining e alla visualizzazione, compresa in tutte queste applicazioni [22].

Un ruolo centrale in quest’ambito viene svolto dai Big Data Analytics, tecnologie di Business Intelligence & Analytics basate sulle tecniche descritte in tabella. Riportiamo nel seguente grafico

quelli più importanti, che ritroveremo più volte in questa ricerca.

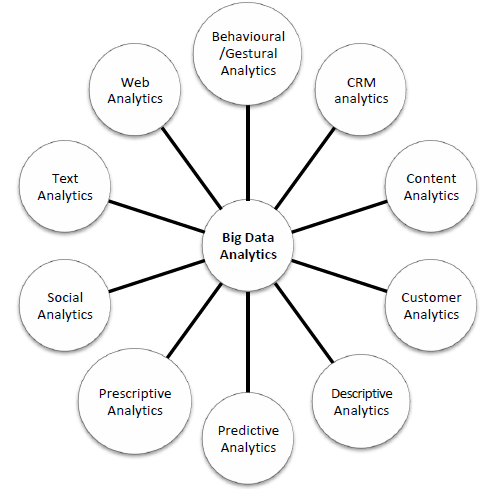
**

Figura 5: Tecniche di Big Data Analytics

* *Behavioural/Gestural Analytics*: analisi automatizzata delle attività umane catturate da video che tracciano i movimenti e i gesti per individuare e comprendere comportamenti e intenzioni;
* *Content Analytics*: insieme di tecnologie che processano i contenuti digitali e i comportamenti degli utenti nelle conversazioni con altre persone, nelle discussioni sui social network o relativamente al livello di consumo ed engagement di documenti e nuovi siti, per rispondere a determinate domande;
* *CRM Analytics*: soluzioni che raccolgono, organizzano e sintetizzano i dati dei clienti per aiutare le organizzazioni a risolvere i problemi di business riguardanti i consumatori
* attraverso tool, dashboard, portali e altri metodi negli ambiti di Marketing, Sales e Customer Service;
* *Customer Analytics*: tecnologie che sfruttano i dati per capire la composizione, i bisogni e la soddisfazione dei consumatori, per poi segmentarli in gruppi sulla base dei comportamenti adottati, implementare azioni di Marketing personalizzate e determinare trend generali;
* *Descriptive Analytics*: analisi di dati e contenuti pe rispondere alla domanda “Cosa è successo?” o “Cosa sta succedendo?” attraverso strumenti tradizionali di BI e visualizzazione;
* *Predictive Analytics*: Analytics avanzati che implementano tecniche quali la regressione, i modelli predittivi e la statistica per analizzare i dati e i contenuti e rispondere alle domande “Cosa succederà” o “Cosa accadrà molto probabilmente?”;
* Prescriptive Analytics: altra forma di Analytics avanzati che esamina i dati e i contenuti per rispondere alle domande “cosa dovrebbe essere fatto?” o “cosa dobbiamo fare per far sì che succeda una determinata cosa?” e per far questo utilizza tecniche quali l’analisi di grafici, la simulazione, le reti neurali e la machine lerarning;
* Social Analytics: tools che estraggono, analizzano e sintetizzano automaticamente i contenuti generati dagli utenti online. Questa tecnologia verrà descritto in modo approfondito nel successivo capitolo;
* *Text Analytics*: processo di estrazione delle informazioni dai testi, utilizzato per diversi scopi, tra cui il *riepilogo*, ovvero il tentativo di trovare i contenuti chiave in un grande insieme di informazioni, la *sentiment analysis*, già spiegate o per determinare cosa ha guidato un determinato commento di una persona e quindi per un fine esplicativo;
* *Web Analytics*: applicazioni analitiche utilizzate per capire e migliorare l’esperienza online del consumatore, l’acquisizione di utenti e l’ottimizzazione del digital Marketing e delle campagne pubblicitarie. Questi offrono reporting, segmentazione, gestione delle campagne e integrazione con altre fonti dati e processi.

## ***1.4 LE DATAWAREHOUSE***

I Data Warehouse (DWH) sono il principale strumento a supporto della Business Intelligence. Essi permettono di collezionare dati integrati, consistenti e certificati, afferenti a tutti i processi di business dell’azienda e provenienti dalle fonti operazionali. Questi dati vengono in seguito opportunatamente trasformati attraverso procedure ETL e controllati attraverso il sistema di Data Quality.

La Qualità dei dati è un requisito fondamentale per l’intero sistema informativo. In quanto, se i dati risultano sporchi, possono oltre che causare un peggioramento delle performance aziendali, portare a prendere decisioni inopportune, comportando costi aggiuntivi e perdita di opportunità.

L’obiettivo di un DWH è pertanto quello di supportare il “*knowledge Worker*” (dirigente, amministratore, gestore, analista) per aiutarlo a condurre analisi finalizzate all’attuazione di processi decisionali e al miglioramento del patrimonio informativo, e fornire un unico punto di accesso per tutti i dati dell’azienda resi consistenti e affidabili attraverso i processi di ETL. Il DWH garantisce inoltre una profondità storica completa dei dati, poiché in esso viene persistito anche lo stato passato delle informazioni permettendo così un’analisi temporale.



Figura 6: Datawarehouse Key Word

Dovranno quindi essere attentamente progettati per gestire in maniera efficiente ed efficace le caratteristiche dei Big Data.

I DWH sono realizzati come principale base per i Decision Support System (DSS), cioè un *sistema* di supporto alle decisioni è un sistema in grado di fornire chiare informazioni agli utenti, in modo che essi possano analizzare dettagliatamente una situazione e prendere le opportune decisioni sulle azioni da intraprendere in modo facile e veloce [12]. Il DSS si appoggia su dati di uno o più database, spesso organizzati in strutture diverse con dati non omogenei.

In altre parole, un sistema di questo tipo deve supportare le attività di analisi e controllo manageriale di routine, le attività di ricerca delle cause di un problema (*focused search*) e le attività di gestione manageriale complessa (*decision making*), permettendo inoltre un facile utilizzo ad un’utenza con un tempo disponibile ridotto e riluttante verso nuove tecnologie (soprattutto nei casi in cui non riesce a percepire in breve tempo i benefici).

Andiamo a descriverne in dettaglio le caratteristiche:

* *Orientato al soggetto*: nel data warehouse i dati sono organizzati per soggetti rilevanti – prodotti, clienti, fornitori, periodi di tempo – al fine di offrire tutte le informazioni inerenti una specifica area.
* *Integrato*: il data warehouse deve essere in grado di integrarsi perfettamente con la moltitudine di standard utilizzati nelle diverse applicazioni. I dati devono essere ricodificati, per risultare omogenei dal punto di vista semantico, e devono utilizzare le stesse unità di misura
* *Variabile nel tempo*: a differenza dei dati operazionali, quelli di un data warehouse hanno un orizzonte temporale molto ampio (anche 5-10 anni), risultando riutilizzabili in diversi istanti temporali.
* *Non volatile*: i dati operazionali sono aggiornati in modo continuo; nel data warehouse i dati sono caricati inizialmente con processi integrali e successivamente aggiornati con caricamenti parziali; i dati, una volta caricati, non vengono modificati e mantengono la loro integrità nel tempo.

È possibile che un DWH sia suddiviso in diversi Data Mart, ognuno dei quali specifico per un solo processo di business fra quelli presenti all’interno dell’azienda (ordini, vendite, clienti, marketing, etc.)

### ***1.4.1 Architettura Di Un Data Warehouse***

In fase di progettazione risulta fondamentale stabilire quali tipologie di architettura adottare. Chiaramente, da quando sono stati idealizzati, i modelli (descritti successivamente) si sono evoluti e, di conseguenza, un DWH deve essere costruito secondo i principi moderni [10].

I pattern descritti in questo paragrafo rimangono comunque delle basi da cui partire.

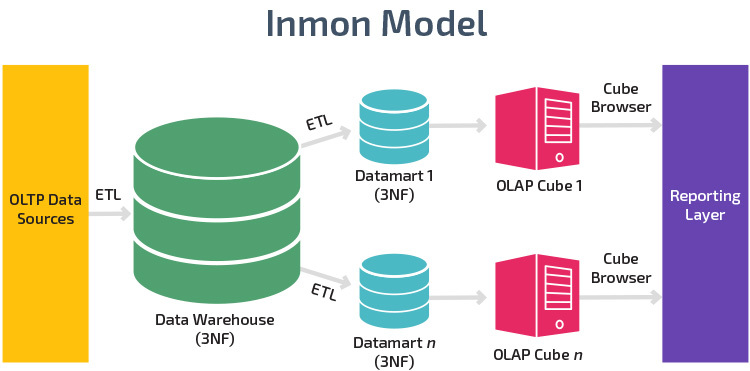
***Modello di Inmon - Corporate Information Factory****:* I DWH si costruiscono nella loro totalità fin dal principio come un unico blocco monolitico; non è possibile vederli come la composizione dei DM. Viene adottata una visione Top-Down.

Figura 7: Modello di Inmon

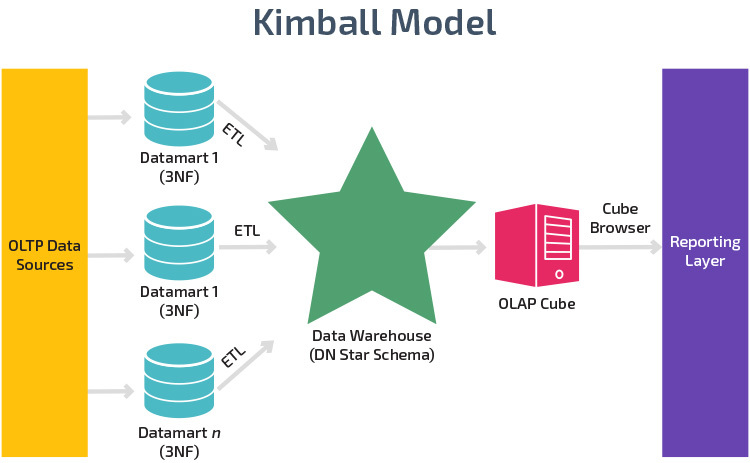
***Modello di Kimball - Dimensional Model****:* adotta un approccio Bottom-up in cui il DWH nasce dall’unione dei vari Data Mart che riferiscono ognuno ad una specifica area di business.

Figura 8: Modello di Kimball

È stato dimostrato che gli approcci di Inmon e Kimball funzionano per consegnare con successo i data warehouse. Esistono persino organizzazioni in cui è stata implementata una combinazione di entrambi ("modello ibrido"). In un modello ibrido, il data warehouse viene creato utilizzando il modello Inmon e, oltre al data warehouse integrato, i data mart orientati ai processi aziendali vengono creati utilizzando lo schema a stella per la creazione di report. Non possiamo generalizzare e affermare che un approccio è migliore dell'altro; entrambi hanno i loro vantaggi e svantaggi, ed entrambi funzionano bene in diversi scenari. L'architetto deve selezionare un approccio per il data warehouse in base ai diversi fattori; alcuni di quelli chiave sono stati identificati in questo documento. Infine, affinché qualsiasi approccio abbia successo, deve essere attentamente studiato, discusso in dettaglio e progettato per soddisfare le esigenze di reporting della BI dell'organizzazione e dovrebbe anche integrarsi con la cultura dell'organizzazione.

## ***1.4.2 OLTP vs OLAP***

**On-Line Transaction Processing (OLTP)**

A livello di database, gli On-Line Transaction Processing si basano su query multi-access veloci ed efficaci. Le principali operazioni svolte sono INSERT, DELETE e UPDATE in quanto modificano direttamente i dati. Questi ultimi vengono quindi costantemente aggiornati e, di conseguenza, richiedono un efficiente supporto alle operazioni di riscrittura. Una caratteristica fondamentale di questi sistemi è la normalizzazione, la quale fornisce un modo rapido ed efficace per effettuare scrittura nel database.

**On-Line Analytical Processing (OLAP)**

L’On-Line Analytical Processing è un insieme di tecniche software per l'analisi accelerata e interattiva di grandi moli di dati, con la possibilità di farlo da punti di vista differenti. Questi sistemi si riveleranno molto utili per l’ottenimento di informazioni di sintesi. Queste ultime avranno lo scopo di supportare e migliorare i processi decisionali aziendale. Esempi di strumenti OLAP sono i Data Warehouse, i Cubi Multidimensionali.

Le maggiori differenze fra i due sistemi sono riportati in tabella [10]:

Tabella 1: OLTP vs OLAP

|  |  |  |
| --- | --- | --- |
|  | **OLTP** | **OLAP** |
| **Finalità** | Supporto all’operatività | Supporto al processo decisionale |
| **Modalità di utilizzo** | Guidata, per processi e stati successivi | Interrogazione ad hoc |
| **Quantità di dati per operazione elementare** | Bassa: centinaia di record per ogni transazione | Alta: milioni di record per ogni query |
| **Qualità** | In termini di integrità | In termini di consistenza |
| **Orientamento** | Per processo/applicazione | Per Soggetto |
| **Frequenza di aggiornamento** | Continua, tramite azioni | Sporadica, tramite funzioni esplicite |
| **Copertura temporale** | Dati correnti | Storica |
| **Ottimizzazione** | Per accessi in lettura e scrittura su una porzione di dati | Per accessi in sola lettura su tutta la base di dati |

In base alla memorizzazione dei dati, si avranno diverse architetture OLAP, ognuna delle quali con i propri pro e contro [10]:

* + - *Relational OLAP (ROLAP):*i dati vengono memorizzati in un database relazionale come supporto al motore OLAP. Le analisi multidimensionali vengono tradotte in query, resti- tuendo risultati in forma multidimensionale.
    - *Multidimensional OLAP (MOLAP):*si ha sia il database che il motore multidimensionale. Ci saranno difficoltà per operazioni di Drill-Down.
    - *Hybryd OLAP (HOLAP):*unisce i vantaggi dei due sistemi precedenti. In particolare, pre-aggrega i dati in sistemi multidimensionali per un’analisi efficiente e veloce, mentre vengono ricercate in un data base relazionale in caso di Drill-Down.
    - *Desktop OLAP (DOLAP):*i dati vengono caricati in un sistema client e vengono calcolati dal motore in locale.