1 L'ERA DEI BIG DATA

1.1 COSA SONO I BIG DATA?

Prima di iniziare la nostra Ricerca, la domanda che dobbiamo porci è: "Cosa sono esattamente i Big Data?". Essi si riferiscono a un dataset la cui dimensione va al di là della capacità di un database normale di catturare, memorizzare, gestire e analizzare i dati (Manyika J., 2011). 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 dati 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 emerge una delle cosiddette 5V che caratterizzano i Big Data, ovvero il volume, le altre sono velocità, varietà, veridicità e valore (Opresnik D., 2015). Il *volume* fa appunto riferimento all'enorme massa di dati generata attraverso numerosi canali: ogni giorno Google elabora circa 24 petabytes di dati, un motore a reazione può generarne 10 terabyte in 30 minuti così come i contatori intelligenti e i macchinari per l'industria pesante tra cui le raffinerie di petrolio e gli impianti di perforazione. 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. 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 Flickr, 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. Il *valore* è un aspetto fondamentale: 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.

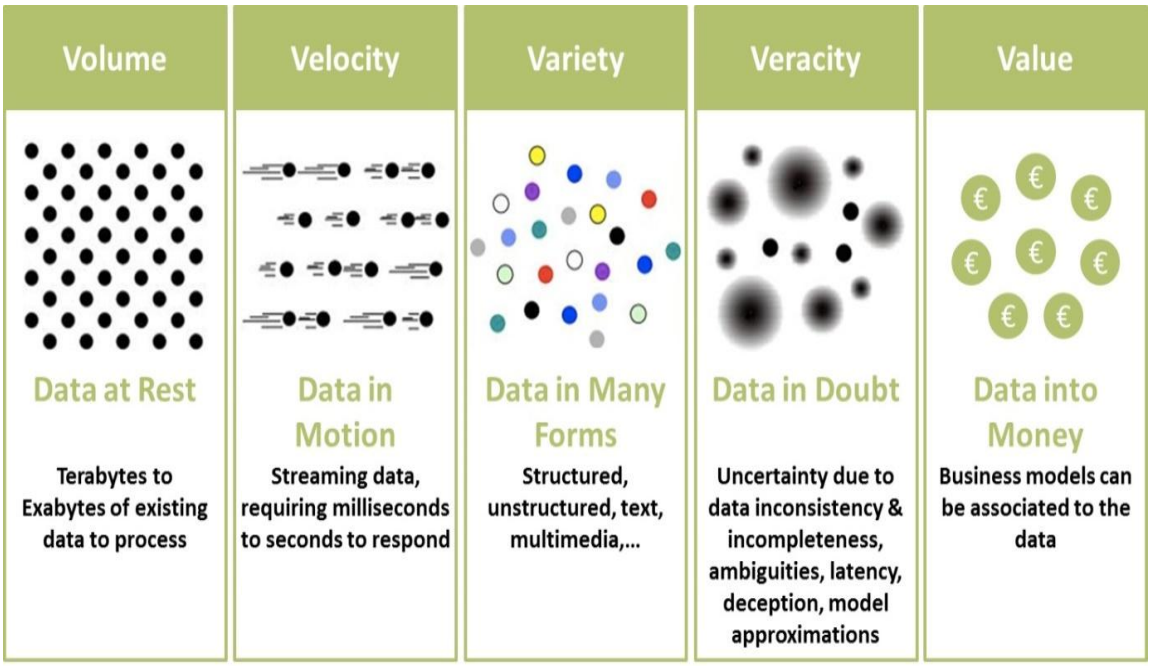

Figura 1: Le 5V dei Big Data (Google, 2014)

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 (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015).

1.2 BENEFICI CONSEGUIBILI DALL'UTILIZZO DEI BIG DATA

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.

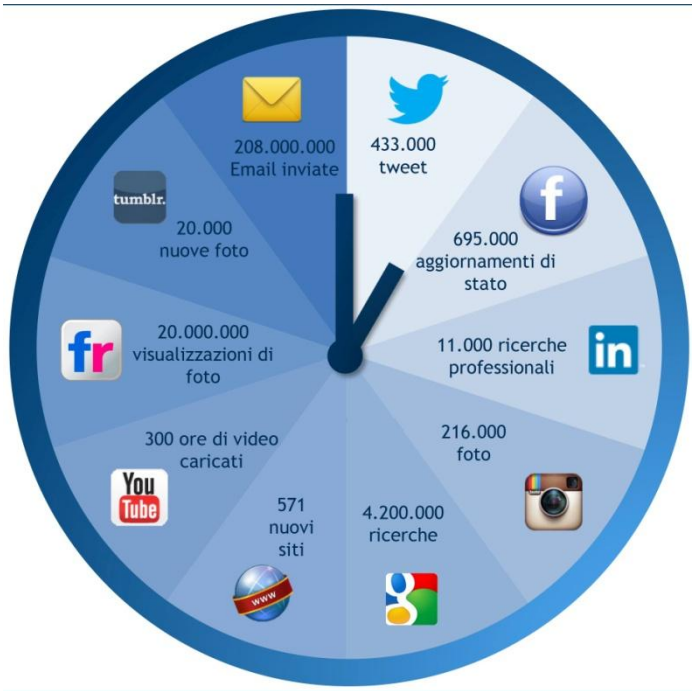

Figura 2: Dati generati in un minuto (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015)

- ✓ **Creare trasparenza.**Un accesso facile e tempestivo ai Big Datarende disponibile una maggiore quantità di informazione e facilita la condivisione dei dati tra le diverse unità organizzative di un'impresa (Bernice P., 2013). Per esempio i dati delle unità di R&S, produzione e ingegneria di un'azienda possono essere integrati al fine di favorire il concurrent engineering, tagliando i tempi e migliorando la qualità (Manyika J., 2011).

- ✓ **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 (Michael K., 2013).

- ✓ **Rivelare le variabilità delle performance e migliorare le prestazioni.**La creazione e la memorizzazione di dati transazionali in forma digitale consente poi 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 (Manyika J., 2011).

Una nota azienda italiana che offre soluzione per la gestione dei dati aziendali, su richiesta di una compagnia assicurativa, desiderosa di migliorare le sue performance, ha analizzato le vendite del suo canale call center, identificato i pattern di successo, determinando quindi la telefonata perfetta.

- ✓ **Personalizzare le azioni.** I Big Data consentono di creare specifici segmenti di clienti e di personalizzare prodotti e servizi sulla base delle loro esigenze. Si tratta di un grande beneficio per vari settori: le aziende di beni di consumo per esempio stanno iniziando ad utilizzare tecniche di Big Data per realizzare promozioni e pubblicità personalizzate per i diversi cluster (Manyika J., 2011).

- ✓ **Migliorare le previsioni.** L'utilizzo dei Big Data e di tecniche adeguate per il loro sfruttamento portano a migliori predizioni e migliori predizioni fruttano migliori decisioni. Per esempio, le principali compagnie aeree statunitensi, venute a conoscenza del disallineamento tra l'orario previsto e quello effettivo di atterraggio, hanno deciso di utilizzare il servizio RightEta offerto da PASSUR Aerospace, un fornitore di tecnologie di supporto decisionale nel campo dell'aviazione, per risolvere questo problema. Più di 155 installazioni raccolgono un ampio range di informazioni su tutti i voli che vedono, generando un costante flusso di dati digitali che vengono analizzati, quindi RightEta si chiede che cosa sia successo in passato quando un dato aereo si è avvicinato ad un dato aeroporto in determinate condizioni e quando effettivamente sia atterrato. In questo modo la compagnia aerea è riuscita ad eliminare il divario tra arrivo previsto e arrivo effettivo dei voli e questo miglioramento delle previsioni ha portato ad un valore di 7 milioni di dollari all'anno in ciascun aeroporto (McAfee A., 2012).
  Diverse Banche italiane stanno mettendo in atto un progetti finalizzati alla raccolta di dati social per arricchire le informazioni sui loro clienti e sfruttarle per predire il tasso di churn tramite adeguati modelli.

- ✓ **Supportare le persone nel processo di decision making.** Utilizzando Analytics sofisticati su interi Dataset è possibile automatizzare e migliorare i processi decisionali, minimizzare i rischi e scoprire preziosi insight, benefici che 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 (Manyika J., 2011).

- ✓ **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. Nasceranno inoltre aziende che si occuperanno di aggregare ed analizzare i dati aziendali relativi a prodotti, servizi, fornitori, consumatori e loro preferenze e si assisterà addirittura a nuovi modelli di business (Manyika J., 2011).

- ✓ **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 vantaggio interesserà svariati settori: in quello manifatturiero è stata prevista una riduzione di più del 50% dei costi di sviluppo del prodotto e di quelli di assemblaggio e una diminuzione di più del 7% del capitale circolante, nel Retail in America è stato stimato un incremento del margine operativo netto del 60% e un aumento della crescita di produttività annua pari allo 0,5-1%, mentre nel settore sanitario americano e nel settore della PA europeo è stato previsto che la produttività aumenterà rispettivamente dello 0,7% e dello 0,5% annuo (Manyika J., 2011).

  Uno studio condotto al MIT Center for Digital Business ha dimostrato che le aziende Data-driven perseguono migliori performance finanziarie e operative. 330 dirigenti di aziende pubbliche del Nord America sono state intervistate su aspetti quali la gestione organizzativa e tecnologica e sono stati raccolti i risultati dai loro report annuali e da fonti indipendenti: le tre migliori aziende di ciascun settore nell'utilizzo dei Big Data, in media, sono più produttive dei loro competitor per il 5% e più profittevoli per il 6% (McAfee A., 2012).

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.

1.3 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 (Court D., 2015).

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.

Tipologie barriere		**Descrizione**
Barriere tecniche	Difficoltà di integrazione dei dati	Integrare dati disomogenei, provenienti da svariate fonti strutturate e non, è un compito molo complesso, così come integrare dati che arrivano da diverse aree aziendali. Di conseguenza estrarre *insight* e trasformare questi in azioni non è così semplice.
	Basso grado di influenza del business	Molto spesso gli Analyticsnon portano a risultati significativi per le aziende, frenando ulteriormente la loro adozione da parte delle organizzazioni.Il dirigente di una famosa casa automobilistica ha investito recentemente in un'iniziativa per capire come i social media possano essere utilizzati per migliorare la

		pianificazione della produzione e le previsioni e ha osservato come l'analisi abbia fatto emergere dettagli interessanti sulle preferenze dei consumatori, ma non abbia fornito una guida su come migliorare effettivamente l'approccio alle previsioni (Court D., 2015).
	Scarsa qualità dei dati	Non sempre le aziende hanno a disposizione dati aggiornati e affidabili e questo è un altro ostacolo da superare.
Barriere legate alle competenze	Difficoltà di comprensione degli strumenti analitici e di quantificazione dei benefici	I manager di linea e i vari utilizzatori non capiscono gli strumenti analitici utilizzati o i consigli che suggeriscono, soprattutto quando questi sono difficili da utilizzare o quando non sono incorporati con i flussi di lavoro e i processi. Risulta perciò complesso per loro anche stimare i benefici derivanti dall'utilizzo dei tool e quindi il ritorno sull'investimento.
	Carenza di talenti	La scarsità di talenti, ovvero di persone con competenze di statistica e di machine learning e manager in grado di sfruttare gli insight dai Big Data nei loro business, rappresenta una delle barriere più importanti. All'interno delle aziende mancano quindi figure organizzative specializzate, quali il Data Scientist e lo Chief Data Officer, che

		approfondiremo più avanti. Negli Stati Uniti nel 2018 si stima una domanda di talenti pari 440000/490000, un'offerta pari a 300000 e quindi un gap di 140000/190000. Persone con questo tipo di abilità sono difficili da trovare in quanto lo sviluppo di tali competenze richiede anni di formazione e inoltre non si può pensare di colmare questo gap cambiando i requisiti di laurea o aspettando persone che si laureeranno in questo ambito o importando talenti, ma risulta necessario riqualificare individui che sono già a disposizione. (Manyika J., 2011).
	Difficoltà nella scelta del tool adatto	La grande varietà di tool, che cambiano molto velocemente e fanno cose molto diverse, rende il processo di scelta molto difficile per le aziende, dato il livello non ancora adeguato di competenze nell'ambito Big Data.
Barriere organizzative/gestionali	Mancanza di commitment da parte del top management	I top manager 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. Per esempio,

		per un'azienda sarebbe importante avere a disposizione dati in tempo reale e meccanismi automatici di pricing, ma se i processi di management prevedono di stabilire i prezzi su base settimanale, l'organizzazione non sarà in grado di sfruttare le opportunità offerte dalla tecnologia (Court D., 2015). Il problema è che nelle organizzazioni non c'è consapevolezza sull'impatto che i Big Data avranno sulla gestione aziendale. Questo comporta lunghi tempi di implementazione.
Barriere economiche	Investimento elevato	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.

Tabella 1: Barriere all'utilizzo dei Big Data

1.4 TECNICHE E TECNOLOGIE PER L'ANALISI DEI BIG DATA

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

Tecniche	**Tecnologie**
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.	*Cassandra*: sistema open source di gestione dei Database, progettato per trattare grandi quantità di dati su un sistema distribuito. Questo sistema è stato sviluppato originariamente da Facebook e ora è gestito come progetto dalla fondazione Apache Sotfware.
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.	*Database NewSQL*: classe di moderni sistemi di Database relazionali che cercano di fornire le stesse prestazioni scalabili dei sistemi NoSQL per l'elaborazione delle transazioni online in lettura e scrittura
Cluster analysis: metodo statistico per classificare gli oggetti, che divide un grande	*Database relazionali*: Database costituito da un insieme di tabelle, in cui dati sono

gruppo in piccoli gruppi caratterizzati internamente da omogeneità non nota in anticipo.	memorizzati in righe e colonne. I sistemi di gestione dei Database relazionali memorizzano solo dati strutturati e utilizzano per lo più il linguaggio SQL (linguaggio progettato per la gestione dei dati di questi Database. Questa tecnica permette di inserire, interrogare, aggiornare e cancellare i dati, di controllarne l'accesso e di gestire la struttura del Database).
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.	*Database non relazionali (NoSQL)*: Database che non memorizza i dati in tabelle (righe e colonne).
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.	*Data warehouse*:Database specializzato per il reporting e spesso utilizzato per memorizzare grandi quantità di dati strutturati. I dati sono caricati attraverso strumenti di ETL (extract, transform e load) da Database operazionali e i report sono generati tramite strumenti di BI.
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.	*Google File System***:** sistema proprietario di file distribuito, sviluppato da Google. È stata l'ispirazione per Hadoop.
Machine Learning: parte della computer science riguardante la progettazione e lo sviluppo di algoritmi che consentono ai	*Hadoop***:** software framework open source che processa grandi Data set relativamente a un certo tipo di problema su un sistema

computer di identificare i comportamenti basandosi su dati empirici e in particolare di riconoscere schemi complessi e prendere decisioni intelligenti.	distribuito. Il suo sviluppo è stato ispirato da Google File System e da MapReduce di Google. Hadoop è stato sviluppato originariamente da Yahoo! E ora è gestito come progetto di Apache Software Foundation.
Modelli predittivi: tecniche in cui viene creato o scelto un modello matematico per prevedere la probabilità di un risultato.	*HBase*: Database open source, distribuito e non relazionale, modellato sul BigTable di Google. È stato sviluppato originarimante da Powerset e ora è gestito come progetto di Apache Software Foundation, come parte di Hadoop.
Natural language processing (NLP): insieme di tecniche di computer science e linguistica che si ricorrono ai computer per analizzare il linguaggio umano.	*Hive*: software sviluppato da Facebook e da poco open source che gestisce e interroga dati memorizzati in un Hadoop cluster utilizzando un linguaggio simile al linguaggio SQL. È più familiare rispetto ad Hadoop per gli utilizzatori di strumenti di BI.
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.	*In-memory Database:* Database Management System che gestisce i dati nella memoria centrale, molto più veloce dei DBMS su memorie di massa, ma le moli di dati sono molto inferiori.
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.	*Lucene*: progetto utilizzato per ricerche e text Analytics, incorporato in diversi software open source.

Regole associative:tecniche volte a scoprire relazioni interessanti tra variabili all'interno di un grande dataset.	*Mahout*: progetto Apache il cui obiettivo è sviluppare applicazioni gratis per algoritmi di machine learning distribuiti e scalabili, al fine di supportare i Big DataAnalytics sulla piattaforma Hadoop.
Regressione: set di tecniche che permettono di determinare come il valore di una variabile dipendente cambia quando una o più variabili indipendenti vengono modificate.	*MapReduce*: software framework introdotto da Google e implementato anche in Hadoop per processare grandi Data set relativamente a certi tipi di problemi su un sistema distribuito.
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.	*MPP Database*: Database che forniscono un'interfaccia SQL e lavorano segmentando su più nodi i dati i quali vengono elaborati in parallelo
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).	*Oozie*: progetto open source che ottimizza il flusso di lavoro e il coordinamento tra i task.
Visualizzazione: tecniche di creazione di immagini, diagrammi o animazioni che consentono di comunicare, capire e migliorare i risultati dell'analisi dei Big Data.	*PIG*:altro software sviluppato da Yahoo e ora open source, che cerca di portare Hadoop più vicino alla realtà di sviluppatori e utenti aziendali, analogamente a Hive. A differenza di quest'ultimo però utilizza un linguaggio "Perl-like" per eseguire query sui dati memorizzati

	in un cluster Hadoop.
	Sistemi distribuiti: insieme di computer che comunicano attraverso una rete, utilizzati per risolvere un problema computazionale, il quale viene diviso in diversi task ognuno dei quali viene risolto da uno o più computer che lavorano in parallelo. Questi sistemi presentano tre vantaggi: costi bassi, alta affidabilità e una maggiore scalabilità.
	WibiData:combinazione di web Analytics e Hadoop, costruito su HBase. Permette ai siti web di eseguire esplorazioni migliori e di lavorare con i dati dei propri utenti, consentendo risposte in real-time relativamente al comportamento degli utenti e fornendo quindi contenuti personalizzati e decisioni.
	Zookeeper: infrastruttura centralizzata con vari servizi, in grado di sincronizzare un cluster di servers. Le applicazioni di Big DataAnalytics si servono di questi servizi per coordinare processi paralleli tra grandi cluster.

Tabella 2: Tecniche e tecnologie per l'analisi dei Big Data

Dopo aver descritto le principali tecniche e tecnologie, vediamo ora un'architettura concettuale per l'analisi dei Big Data, rappresentata nella figura riportata nella pagina seguente.

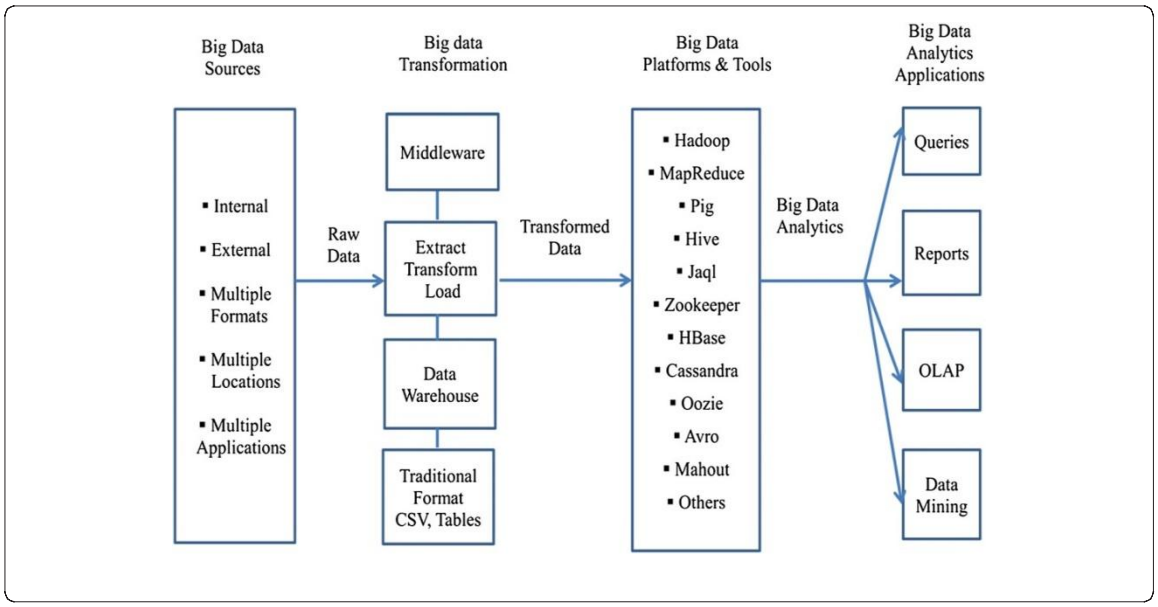

Figura 3: Architettura concettuale per l'analisi dei Big Data (Raghupathi W., 2014)

Innanzitutto 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 DataAnalytics che includono queries, reports, OLAP e Data mining e alla visualizzazione, compresa in tutte queste applicazioni (Raghupathi W., 2014).

Un ruolo centrale in quest'ambito viene svolto dai Big DataAnalytics, 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.

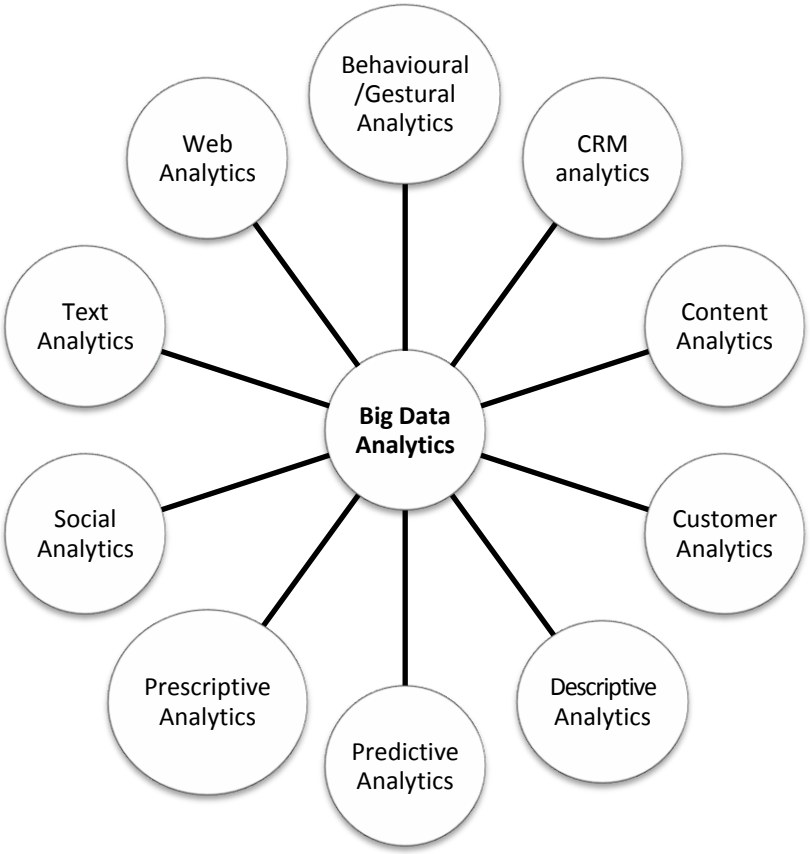

Grafico 2: I Big Data Analytics

- *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 Marketinge delle campagne pubblicitarie. Questi offrono reporting, segmentazione, gestione delle campagne e integrazione con altre fonti dati e processi.

1.5 TENDENZE IN ATTO

La presenza di nuovi strumenti e i miglioramenti degli approcci di analisi dei dati offrono nuove opportunità per trarre ulteriori vantaggi. In particolare possiamo individuare tre tendenze in atto (Court D., 2015).

La prima è rappresentata dallo sviluppo di soluzioni analitiche specifiche, focalizzate su un'area determinata, come la logistica, la gestione del rischio, il pricing e la gestione del personale. Si

tratta di soluzioni che possono essere sviluppate molto rapidamente e che permettono di ottenere notevoli benefici nell'area specifica, ma che richiedono un cambiamento nella cultura organizzativa e la creazione di enfasi sulla loro adozione.

Un second trend è la democratizzazione degli "Analytics", su cui stanno investendo molte aziende quali American Express, Procter & Gamble e Walmart (Court D., 2015). Lo sviluppo di nuovi strumenti self-service sta aumentando la fiducia negli Analytics degli utilizzatori frontline, i quali, senza conoscere la singola riga di codice, possono collegare i dati da molteplici fonti e fare previsioni. Inoltre gli strumenti di visualizzazione rendono più facili le operazioni di slice e dice, permettono di individuare i dati da esplorare per affrontare le problematiche di business e di supportare il processo di decision making. Un'azienda di hardware, per esempio, ha sviluppato un set di soluzioni analitiche self-service e di strumenti di visualizzazione che aiutano l'azienda a condurre le analisi dei clienti e a identificare le opportunità di vendita e di rinnovo al fine di migliorare le decisioni della forza di vendita. L'implementazione di questa piattaforma ha portato ad un incremento dei ricavi pari a 100 milioni di dollari.

Infine oggi sta diventando sempre più semplice automatizzare i processi e prendere decisioni: i miglioramenti tecnologici permettono di catturare un numero molto più grande di dati in tempo reale, facilitando quindi i processi di elaborazione di un'enorme base di dati e di analisi in real time. Questi avanzamenti tecnologici stanno aprendo nuovi percorsi per l'automazione e l'apprendimento automatico per tutte le aziende e non solo quelle leader nella tecnologia. Pe esempio un'importante società di assicurazione ha raggiunto notevoli progressi implementando uno strumento che permette di prevedere la gravità dei reclami, attraverso un confronto istantaneo di milioni di dati registrati, riducendo quindi il bisogno dell'intervento umano.

1.6 L'IMPATTO DEI BIG DATA IN QUATTRO SETTORI

Affrontiamo ora il modo in cui i Big Data possono creare valore in quattro diversi domini: il *settore sanitario*, il *settore della Pubblica Amministrazione*,il *Manufacturing* e il *Retail*.

1.6.1 SANITÀ

Oggi il settore sanitario sta affrontando uno tsunami di dati relativi alla salute e alla Sanità, creati e accumulati continuamente. Si tratta di dati clinici generati da sistemi di supporto alle decisioni quali note e prescrizioni dei medici, immagini mediche comprese quelle 3D più recenti, dati dai laboratori, dalle farmacie, dalle assicurazioni e altri amministrativi; electronic

health records (EHR, archivi di dati digitali sanitari); dati generati dalle macchine e sensor data, come quelli provenienti dal monitoraggio dei segnali vitali; genomic data, tra cui il genotipo e l'espressione genica; post dai social media; blog (Raghupathi W., 2014).In particolare sono gliEHR che consentono una profonda conoscenza clinica e dei quadri patologici dei pazienti: questi archivi di dati possono essere utilizzati per cercare associazioni nelle diagnosi mediche e considerare le relazioni temporali tra eventi al fine di scoprire la progressione delle malattie (Chen H., 2012).Ogni persona porta 4 terabyte di dati e ciascun payer-provider (i payors sono compagnie di assicurazione, organizzazioni sanitarie, imprenditori e gestori di richieste di rimborso nell'ambito dei programmi di assistenza medica statali o federali mentre i providers sono ospedali, personale sanitario e cliniche) potrebbe costruire una matrice con centinaia di migliaia di pazienti con diverse informazioni e parametri (demografia, cura e risultati) raccolti per un lungo periodo di tempo (Miller K., 2011-2012).

I dati sono accumulati velocemente in tempo reale. Future applicazioni real-time, come il rilevamento di infezioni il prima possibile, permetteranno di identificarle rapidamente e di mettere in pratica le giuste terapie, prevenendo infezioni e riducendo il tasso di mortalità.

Anche la varietà è una caratteristica dei dati del settore. Ci sono infatti dati strutturati che possono essere facilmente memorizzati, interrogati, analizzati e manipolati, i quali, insieme a quelli semistrutturati, includono la lettura degli strumenti e la conversione dei documenti cartacei in EHR. Oltre a questi ci sono dati non strutturati che comprendono note di medici e infermieri scritte a mano, prescrizioni cartacee, radiografie e immagine mediche.

La veridicità dei dati, soggetti a errori(soprattutto quelli non strutturati caratterizzati da grande variabilità), rappresenta un obiettivo, in quanto la qualità dei dati nella Sanità è fondamentale: decisioni mediche che impattano sulla vita o la morte delle persone dipendono infatti dall'accuratezza delle informazioni (Raghupathi W., 2014).
Le principali applicazioni di Big DataAnalytics utilizzate nel settore sanitario sono le queries, i report, l'OLAP e il Data mining. Tecniche e tecnologie, di cui abbiamo già discusso in precedenza, vengono utilizzate per aggregare, manipolare, analizzare e visualizzare tutti i dati.
Per riuscire ad ottenere vantaggio competitivo dagli insight forniti dai Big Data, gli strumenti tradizionali non sono sufficienti in quanto essi si focalizzano solo sulla riduzione dei costi e non sul miglioramento dei risultati delle terapie e quindi delle condizioni e del livello di soddisfazione dei pazienti. È quindi necessario sviluppare strumenti incentrati sul paziente che

prendano in considerazione entrambi gli aspetti; per questo gli stakeholder devono focalizzarsi sui cinque seguenti aspetti:

- right living: i pazienti devono avere un ruolo attivo nel miglioramento della loro salute, facendo scelte adeguate relative alla dieta, all'esercizio e alla prevenzione;
- right care: i medici e tutto il personale sanitario devono avere accesso alle medesime informazioni per favorire il coordinamento e lavorare per lo stesso obiettivo, al fine di evitare la duplicazione dello sforzo e strategie subottimali;
- right provider: tutti i provider devono accedere agli archivi di dati ed essere in grado di raggiungere i risultati migliori;
- right value: incrementare contemporaneamente il valore e la qualità della cura;
- right innovation: sviluppare nuovi approcci per migliorare i servizi sanitari.

Già alcuni leader nel settore sanitario hanno iniziato a focalizzarsi su queste soluzioni o comunque a porre le basi per il futuro. Per esempio Kaiser Permanente, consorzio sanitario in Oakland, California, United States, ha implementato HealthConnect, un nuovo sistema che permette lo scambio dei dati tra le varie strutture mediche e promuove l'utilizzo degli EHR; questa soluzione ha migliorato le prestazioni nell'ambito delle malattie cardiovascolari e comportato un risparmio pari a 1 miliardo di dollari dovuto alla riduzione delle visite e dei test di laboratorio (Kayyali B., 2013).

1.6.1.1 Benefici conseguibili dall'utilizzo dei Big Data

Il ricorso ai Big Data in questo settore permette di conseguire una serie di benefici, connessi a quelli dei Big Data in generale. In tabella riportiamo i 6 principali (Manyika J., 2011), (Raghupathi W., 2014), (Sun J., 2013).

Benefici	**Descrizione**
Riduzione dei costi	L'aumento dell'efficienza è uno dei più importanti vantaggi ed è reso possibile da una serie di pratiche. Innanzitutto la comparative effectiveness research (CER), ovvero l'analisi di grandi Dataset in cui sono contenuti dati quali le caratteristiche dei pazienti e i costi e i risultati delle terapie, permette di identificare trattamenti economicamente

	vantaggiosi. In secondo luogo la creazione di mappe dei processi e di dashboard a partire dai Dataset dei provider consente di identificare le fonti di variabilità e gli sprechi e quindi di ottimizzare i processi clinici. Infine l'analisi dei quadri patologici e dei trend per stimare la domanda futura dei farmaci e il ricorso a modelli predittivi sviluppati dalle aziende farmaceutiche tramite l'aggregazione dei dati di ricerca, favoriscono un'allocazione più vantaggiosa delle risorse R&D.
Incremento dei risultati grazie a decisioni migliori (maggiore qualità delle terapie ed alta soddisfazione dei clienti)	La CER e la trasparenza dei dati, garantita dallo sviluppo di mappe di processo e di dashboard, permettono inoltre di ridurre l'incidenza delle terapie dannose e di quelle che dovrebbero essere prescritte ma che non vengono messe in pratica, nonché di migliorare la qualità delle cure. Anche i sistemi di supporto per le decisioni cliniche, in cui i medici inseriscono le loro terapie che vengono poi confrontate con le linee guida, mettono in guardia da eventuali errori come la prescrizione di farmaci che avranno effetti negativi sui pazienti. Questi ultimi possono essere individuati anche attraverso l'analisi dei test clinici e quindi si eviterà di immetterli sul mercato salvaguardando l'immagine aziendale. Gli stessi modelli predittivi rendono le attività di R&D sui farmaci più veloci e più specifiche, favorendo un'immissione molto più rapida dei medicinali sul mercato e la produzione di composti specifici con alto tasso di successo. Infine l'utilizzo di un Database con i dati di tutti i pazienti e di tutte le terapie a livello nazionale assicura una rilevazione rapida delle malattie infettive e il controllo di eventuali epidemie globali grazie ad un apposito programma, garantendo quindi il monitoraggio della salute pubblica ed un miglioramento della qualità della vita stessa.

Personalizzazione della cura	Enormi Dataset vengono analizzati per esaminare le relazioni tra le variazioni genetiche, la predisposizione a malattie specifiche e le risposte a determinati farmaci per poter sviluppare farmaci personalizzati, garantendo terapie più efficaci e diagnosi precoci.
Prevenzione	L'utilizzo di Analytics avanzati quali modelli predittivi e di segmentazione dei pazienti per identificare gli individui che possono trarre beneficio da cure proattive o da cambiamenti dello stile di vita, favorisce la prevenzione di eventuali problemi di salute.
Aumento del livello di soddisfazione di tutti gli attori del sistema	Le piattaforme e le comunità online, come i già citati Sermo.com e PatientsLikeMe.com, migliorano la comunicazione tra gli individui e consentono la condivisione di esperienze e quindi un aumento del sostegno reciproco e del loro coinvolgimento.
Sviluppo di nuovi modelli di business	I dati clinici dei pazienti e dei Dataset delle richieste, aggregati e sintetizzati, possono essere venduti a terze parti. Questi robusti Dataset clinici rendono possibile la nascita di nuovi business, quali l'analisi dei risultati clinici per i payors che possono così prendere decisioni migliori o l'analisi dei Database per scoprire i biomarcatori che selezionano le terapie.

Tabella 3: Benefici conseguibili dall'utilizzo dei Big Data nel settore sanitario

È quindi chiaro che l'estrazione di conoscenza dai Big Data nel settore sanitario sia fonte di rilevanti vantaggi, tuttavia essa richiede di affrontare sfide quali la garanzia della privacy in modo da mantenere la fiducia delle persone e la conduzione di ricerche etiche. Il rischio è infatti quello di essere puniti per il mancato rispetto dei requisiti stabiliti dall'HIPAA (Health Insurance Portability and Accountability Act) e dall'IRB (Institutional Review Board).

1.6.2 PUBBLICA AMMINISTRAZIONE

I Governi di tutto il mondo stanno riconoscendo l'importanza dei Big Data come leva per migliorare i servizi ai cittadini e ottenere un vantaggio competitivo, soprattutto in un periodo come quello attuale, in cui i Governi locali stanno affrontando un calo del budget, un deterioramento delle infrastrutture e il bisogno di attirare e trattenere business per la crescita economica. Per la Pubblica Amministrazione è fondamentale capire come i Big Data e gli Analytics possano incrementare l'efficienza e la produttività, migliorare i processi decisionali, aumentare la trasparenza verso i cittadini, aiutare a controllare sprechi, frodi e abusi e gestire budget e costi (Fiorenza P., 2014). Cisco nel 2014 ha stimato che lo sfruttamento dei Big Data rappresenterà un'opportunità pari a 4600 miliardi di dollari nei prossimi 10 anni (Bikar P., 2015)

Gli enti organizzativi hanno accesso a vari flussi di dati, per lo più testuali e numerici e per il 90% in formato digitale (Manyika J., 2011): documenti, video, foto, dati dai social media e dal mobile. Si tratta quindi di un mix di dati strutturati e non, che necessitano, per l'estrazione di insight, di un Database NoSQL, più agile e flessibile nel raccogliere, processare e analizzare un'ampia varietà di fonti di dati. MarkLogic è la piattaforma NoSQL collauData per le applicazioni Big Data governative, realizzate per condurre analisi real-time e trasformare tutti i dati in informazioni di valore e azioni da implementare (Fiorenza P., 2014).

Comunque, nonostante ci siano programmi di e-government avanzati, le agenzie non condividono i dati con i cittadini e con le organizzazioni: molto spesso succede che l'impiegato di un'agenzia riceva una copia di dati via fax o via mail da un'altra, quando questi potrebbero essere memorizzati elettronicamente. Esistono addirittura delle restrizioni legali e delle policy che impediscono la condivisione.

1.6.2.1 Benefici conseguibili dall'utilizzo dei Big Data

Analizziamo ora più in dettaglio i benefici apportati dai Big Data alla Pubblica Amministrazione.

La virtualizzazione degli eterogenei Dataset pubblici permette di guidare il processo decisionale e di migliorare i risultati delle politiche. Ci sono però degli ostacoli da superare per perseguire questo beneficio:

- ci sono troppi dati da utilizzare efficacemente;
- i dati si sviluppano ampiamente e in modo nascosto;
- i formati non sono facilmente condivisibili;

- le politiche di sicurezza variano dal dominio pubblico a quello privato e segreto.

La soluzione è rappresentata dal Data Virtualization (DV) che fornisce un'unica soluzione consentendo l'accesso a tutti i dati, senza che ci siano duplicazioni. Per esempio il Governo Tedesco ha implementato un programma decennale che prevede la digitalizzazione di regole e regolamenti riguardanti la pianificazione, le zone edificabili e i permessi ambientali. Il programma si pone come obiettivi decisioni migliori e più veloci, un flusso di lavoro completamente digitale, una visione unica delle informazioni per cittadini, governi e aziende e una condivisione dei dati accurata e affidabile (Bikar P., 2015).

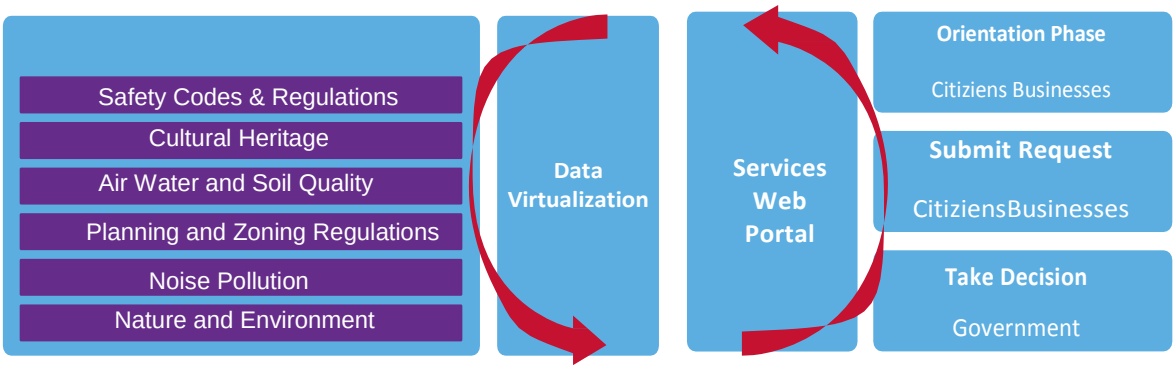

Figura 4: Programma di digitalizzazione di regole e regolamenti implementato dal Governo Tedesco (Bikar P., 2015)

I Big Data danno vita a nuove politiche basate su informazioni più smart relative ai cittadini e alle aziende, rendendo i servizi pubblici più efficaci in diversi modi: garantendo alle famiglie la distribuzione di benefit e di altri supporti di loro diritto, rispondendo alle richieste pubbliche più velocemente e accuratamente, rilevando in anticipo problemi e priorità in modo da aggiustare le politiche. Inoltre prendendo decisioni migliori su come gli enti devono essere organizzati e quali lavori hanno la priorità, i costi delle operazioni governative si riducono.

Rendere le informazioni del Settore Pubblico disponibili a tutti i cittadini e a tutti i dipendenti della PA genera trasparenza, un aumento della soddisfazione e della fiducia dei cittadini stessi nei confronti della PA, facilita l'interoperabilità all'interno dell'amministrazione e la creazione di nuovi servizi.

Analizzando i grandi volumi di dati in real-time e trasformandoli in informazioni di valore tramite i modelli predittivi, è possibile scoprire anomalie passate e attuali, rilevare trend che evidenziano rischi potenziali e trovare quindi dei modi per ridurre errori, abusi e frodi e ottimizzare i processi (Bikar P., 2015).

I Big Data permettono anche di scoprire la variabilità delle performance delle diverse parti di un ente governativo che svolgono funzioni simili, attraverso per esempio l'utilizzo di dashboard che visualizzano tutti i dati operativi e finanziari, alimentando la competizione tra queste; competizione che porta quindi ad un miglioramento delle performance e ad un aumento della produttività.

L'utilizzo dei Big DataAnalytics dà anche la possibilità segmentare e personalizzare i servizi per gli individui e la popolazione. I vantaggi sono un incremento dell'efficacia e dell'efficienza delle agenzie, un miglioramento della relazione tra manager e clienti degli enti e un aumento della soddisfazione dei clienti grazie a servizi che rispondono ai loro bisogni. Un'agenzia di lavoro tedesca per esempio ha analizzato i dati storici dei clienti, le azioni implementate e i risultati al fine di effettuare interventi su misura per i disoccupati: l'iniziativa ha portato a una riduzione annua della spesa di 10 miliardi nei tre anni successivi, ad una diminuzione del tempo necessario per trovare lavoro da parte disoccupati e a una loro maggiore soddisfazione (Manyika J., 2011).

1.6.3 MANUFACTURING

Anche il Manufacturing sta cercando di applicare Analytics a grandi pool di dati al fine di dedurre informazioni utili per il business e azioni da perseguire. I benefici che ne derivano sono un incremento dell'efficienza nelle attività di progettazione e di produzione, un aumento della qualità e del grado di innovazione dei prodotti, una maggiore soddisfazione dei bisogni dei consumatori, una previsione della domanda più accurata, una produttività più alta e un miglioramento della gestione della complessa e globale catena del valore.

I dati provengono da innumerevoli fonti: macchinari per la produzione, sistemi per la gestione della supply chain, sistemi che monitorano le performance dei prodotti che sono già stati venduti, RFID (radio-frequency identification) ovvero dispositivi che tracciano il prodotto e per i quali è previsto un aumento dai 12 milioni del 2011 ai 209 miliardi del 2021, commenti sui social media e diversi sistemi tra cui i computer-aided design, i computer-aided engineering e i computer-aided manufacturing (Manyika J., 2011).

Segnali quali le vibrazioni e la pressione estratti da sensori presenti nelle macchine e dati storici relativi a questi elementi possono essere utilizzati da Analytics avanzati che con l'avvento del cloud computing e del framework Cyber-Physical Systems, comporteranno la creazione di un sistema informativo della flotta macchine che consentirà a queste ultime di essere self-aware e di prevedere eventuali problemi di prestazione. Un sistema di macchine self-aware e self-

manteined è un sistema in grado di autovalutare il suo stato di salute e di peggioramento ed utilizza ulteriori informazioni provenienti dalle altre macchine per prendere decisioni relative alla manutenzione intelligente al fine di evitare potenziali problemi. In particolare questi smart Analytics verranno utilizzati sia a livello di singola macchina che a livello di flotta (Lee J., 2014).

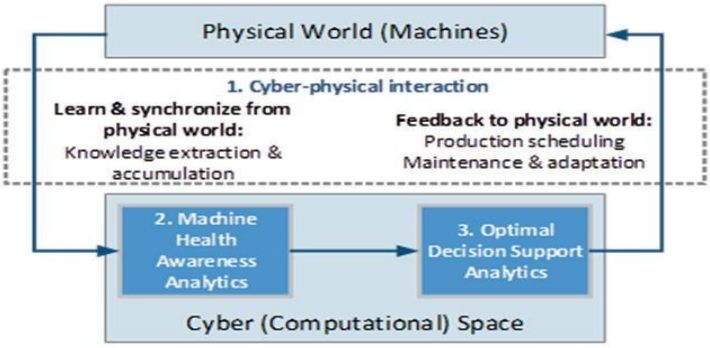

Figura 5: Framework Cyber-Physical Systems per le macchine self-aware e self-maintenance(Lee J., 2014)

Per fare in modo che i team di Big Data e di Analytics riescano ad estrarre la business intelligence dalle diverse fonti dati e ad ottimizzare i processi manifatturieri, devono valere le tre seguenti condizioni (Kerschberg B., 2014):

- *Trasparenza dei dati* che permette l'integrazione dei dati provenienti dalle diverse funzioni manifatturiere;
- *Visibilità del processo*che consente ai manager di vedere come stanno avvenendo i processi, in modo da implementare delle correzioni;
- *Visualizzazione dei dati* risultanti dall'applicazione degli Analytics ai Big Data in modo che gli utilizzatori finali come i responsabili dell'impianto possano vedere i dati nascosti e il loro valore. Questo è fondamentale per il dinamismo real-time dei dati stessi.

Vediamo di seguito i benefici derivanti dallo sfruttamento dei Big Data nelle diverse parti dellacatena del valore manifatturia.

- Sfruttando i dati di input e gli insight sui clienti, realizzare un prodotto con caratteristiche adeguate alle loro esigenze e ridurre i costi di sviluppo del prodotto;
- creare più valore attraverso la piattaforma di product lifecycle management (PLM) che integra i Dataset di sistemi multipli per rendere efficace la collaborazione;
- prendere le decisioni migliori, che comportano una riduzione dei costi: produttori e designer condividono i dati e creano in modo rapido ed economico simulazioni per testare diversi progetti, diversi fornitori e i costi di produzione associati (importante perché le decisioni nella parte di progettazione comportano l'80% dei costi);
- ridurre il tempo di sviluppo di nuovi prodotti ed eliminare i difetti prima della costruzione del prototipo;
- migliorare i prodotti esistenti e sviluppare nuovi modelli e varianti di prodotti già esistenti;
- aumentare il grado di innovazione: i produttori invitano gli stakeholder esterni a esprimere idee innovative o a partecipare alla sviluppo di nuovi prodotti attraverso piattaforme web-based.

- Prevedere la domanda e migliorare la pianificazione della supply chain, in modo da utilizzare i soldi nel modo più efficiente possibile e migliorare il livello di servizio. Per far questo è necessario integrare i dati dei Retailer come quelli relativi alle promozioni(item, prezzi,..) e al magazzino (livello di scorte nel magazzino, vendite per negozio);
- ridurre il tempo di risposta, il livello delle scorte e il tempo per lanciare un nuovo prodotto (da articolo introduttivo);
- estrarre nuove idee e comprendere meglio i propri prodotti, clienti e mercati (da articolo introduttivo).

- Aumentare l'efficienza del processo produttivo attraverso tecniche di simulazione applicate ai grandi volumi di dati generati dai prodotti;
- ridurre il numero di cambiamenti del progetto, i costi dei tool di design e costruzione, le ore di assemblaggio e migliorare l'affidabilità della consegna. Per ottenere tali vantaggi è necessario utilizzare i dati di sviluppo del prodotto e i dati storici di produzione per progettare e simulare il sistema produttivo, dal layout alla sequenza di fasi da seguire per un dato prodotto;
- controllare e ottimizzare i processi produttivi e di supply al fine di ridurre gli scarti e

massimizzare la resa utilizzando i dati real-time dei sensori in tali processi.
• Riprogettare il prodotto e sviluppare nuovi prodotti, sfruttando i dati appena citati; • migliorare la previsione della domanda; • migliorare i servizi post-vendita offerti (esempio: i produttori di aerei o di ascensori utilizzano i dati dai sensori presenti sui prodotti per programmare pacchetti proattivi di servizi di manutenzione intelligente).

Tabella 4: Benefici derivanti dallo sfruttamento dei Big Data nelle diverse parti della catena del valore manifatturiera

1.6.4 RETAIL

I Big Data offrono enormi opportunità anche al mondo Retail che ha disposizione non solo dati come le transazioni e le operazioni dei clienti, ma anche dati dagli RFID e informazioni sul comportamento online e sul sentiment dei consumatori. Le aziende che utilizzano tecniche e tecnologie in grado di sfruttare questi dati, riescono a migliorare l'efficacia delle loro azioni di Marketing e di merchandising, a ridurre i costi delle operazioni e della supply chain e quindi a migliorare la loro profittabilità e a ottenere un vantaggio competitivo rispetto agli atri concorrenti (Manyika J., 2011).

I Retailer in particolare possono utilizzare due diverse tipologie di Analytics (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015):

- *Performance Management & Basic Analytics*: strumenti di Descriptive Analytics che consentono di accedere ai dati secondo viste logiche flessibili e dinamiche e di visualizzare sinteticamente e graficamente i principali indicatori di prestazione;
- *Advanced Analytics*: strumenti avanzati che permettono di svolgere un'analisi attiva dei dati sfruttando metodologie di prescriptive&predictive analysis, determinando trend e prevedendo il valore futuro di variabili numeriche e categoriche.

In quest'analisi ci concentriamo sia sui processi di back-end, ovvero i processi di interazione Retail-fornitore o processi interni del Retailer sia su quelli di front-end, ovvero quelli di interazione Retailer-consumatore.

- Processi di back-end

- *Logistica*. Oggi i fornitori logistici gestiscono un enorme flusso di beni creando nel contempo un set rilevante di dati (origine e destinazione dei viaggi, dimensione, peso, contenuto del trasporto, posizione)considerati i milioni di viaggi intrapresi ogni giorno.

Le tecniche di Big Data possono essere innanzitutto utilizzate per ottimizzare i costi del viaggio per la consegna dei prodotti. Focalizziamoci su due diverse modalità.

1)Ottimizzazione del percorso in real time. Quando il veicolo viene caricato per partire, il percorso di consegna ottimale viene pianificato utilizzando i dati di spedizione rilevati dai sensori sui prodotti in viaggio. Durante il trasporto sistemi dinamici suggeriscono cambiamenti del percorso a seconda delle condizioni del traffico, dei fattori geografici e dello stato del ricevente. Questo approccio, che si basa sull'utilizzo di dati reali, permette di tagliare i costi e di ridurre l'emissione di CO_2 diminuendo per esempio le distanze da percorrere.

2)*Pick-up e consegna basate sulle persone*. Pendolari, taxisti o studenti potrebbero essere pagati per occuparsi della consegna nell'ultima parte del percorso se questa coincide con il viaggio che devono effettuare. Questo approccio, che porta ad una notevole riduzione dei costi, richiede l'utilizzo di tecniche di Big Data: flussi di dati real-time vengono tracciati al fine di assegnare la spedizione alle persone disponibili, basandosi sulle loro posizioni e destinazioni. Attraverso dispositivi mobile, possibili trasportatori pubblicano la loro posizione e accettano l'assegnazione della consegna.

Analitycs avanzati possono essere utilizzati anche per prevedere la domanda al fine ottimizzare la capacità di trasporto e la quantità di personale necessario in ciascuna zona. Per l'allocazione delle risorse vengono utilizzate informazioni real-time sulle spedizioni (item che sono appena entrati nella rete distributiva, che sono in transito o che sono in magazzino), informazioni dai clienti (apertura di nuove industrie, fallimenti inaspettati) e anche informazioni di eventi locali (epidemie regionali e disastri naturali).

Le soluzioni Big Data permettono inoltre il recupero di informazioni utili per il rilevamento dei rischi di supply chain: dati sugli sviluppi politici locali, sull'economia, sulla salute, sulla natura provenienti da diverse fonti quali siti, social media, blog vengono aggregati e analizzati attraverso sematic Analytics e

altre tecniche. Queste soluzioni individuano dei pattern tra le diverse informazioni e quando si verifica una condizione critica per la supply chain del cliente, questo viene avvisato e gli viene inviato un report contenente informazioni quali la probabilità e l'impatto del rischio e contromisure per mitigarlo (Jeske M., 2013).

- *Gestione del magazzino.* L'integrazione di strumenti di ottimizzazione del magazzino e dei sistemi ERP consente di fare la migliore analisi possibile dei dati a disposizione quali dati delle vendite, degli acquisti, finanziari, di fornitura e di produzione. Sviluppando un algoritmo, il sistema di ottimizzazione può creare un'interfaccia grafica in grado di illustrare una sintesi di tutti i dati, che permette di identificare i cambiamenti stagionali dei prodotti richiesti, quando si verificheranno gli stock-out, le vendite perse e gli ordini in eccesso, ecc… Questa analisi viene quindi utilizzata da uno tool di ottimizzazione del magazzino per prevedere la domanda in modo accurato e stabilire quindi il livello ottimale del magazzino, che soddisfi le richieste dei clienti ed eviti sia lo stock-out che la sovrabbondanza di item. Questo strumento dà poi dei suggerimenti per la successiva pianificazione delle scorte e definisce la soglia del riordino. Data l'integrazione tra questo tool e i sistemi ERP, tutti questi dati vengono comunicati sia all'interno che agli stakeholder all'esterno dell'organizzazione (Sage, 2013)Alcune aziende utilizzano anche sistemi bar-code collegati ai processi di rifornimento automatico per ridurre l'incidenza dello stock-out (Manyika J., 2011).

- *Pianificazione dei turni*. È possibile utilizzare un algoritmo predittivo che prenda in considerazione un ampio range di parametri individuali e locali: dati dei ricavi storici, orari di apertura dei negozi, orari di arrivo dei prodotti dai centri distributivi ma anche i giorni del mercato, i giorni di vacanza delle località vicine e i dati delle previsioni meteorologiche che influenzano il comportamento dei clienti. L'algoritmo dà come soluzione le vendite giornaliere previste, a partire dalle quali vengono pianificati i turni in modo ottimale, evitando surplus e carenze del personale, che impattano negativamente sulle performance finanziarie del negozio e sulla soddisfazione dei clienti e dei dipendenti (Jeske M., 2013).

➢ Processi di front-end

Le nuove abitudini dei consumatori, che comportano la creazione di uno tsunami di dati, giustificano l'utilizzo delle tecnologie di Big Data da parte dei Retailer (Osservatorio Big Data Analytics & Business Intelligence, maggio 2015):

- 2 consumatori su 3 si informano online prima di acquistare un prodotto, comprano in un negozio, ma hanno preso la decisione prima su canali digitali (Fonte: Net Retail –Netcomm, Campione: 3.055 individui);
- il 54% dei consumatori preferisce percorsi di acquisto che contemplino almeno un'interazione con i canali digitali online e mobile (Fonte: Total Retail PwCCampione:1.000 consumatori che hanno acquistato almeno 1 volta online);
- degli oltre 30 milioni di Internet user, il 31% utilizza per navigare un solo device, il 69% utilizza due o più device per navigare tra cui PC, Smartphone e Tablet (Fonte: SurveyCAPI 2013 –Doxa, Campione: totale Internet userdaily);
- l'89% degli utenti Smartphone utilizza il device all'interno del negozio. Di questi circa il 40% dichiara di farlo sempre o spesso (Fonte: SurveyCAWI 2014 –Doxa, Campione: 1.503 utenti smartphone);
- tra gli utenti che usano lo Smartphone in negozio, il 42% degli utenti confronta i prezzi e il 30% degli utenti invia foto dei prodotti da acquistare ad amici (Fonte: SurveyCAWI 2014 –Doxa, Campione: 1.341 utenti che usano lo smartphonein punto vendita);
- molti consumatori utilizzano applicazioni quali RedLaser che permette loro di scannerizzare il bar code su un item in un negozio con il loro smartphone e ottenere immediatamente il prezzo e il prodotto dei concorrenti (Manyika J., 2011).

I Retailer hanno a disposizione una serie di leve per sfruttare i Big Data e trarre quindi vantaggio nei processi di interazione con il cliente. In questo paragrafo ci limitiamo ad elencarle, in quanto saranno oggetto di approfondimento del secondo capitolo. Nei processi di Marketing le tra le leve principali citiamo il Cross-Selling, il location based Marketing, l'in-store behavior analysis, la customer micro-segmentation, la sentiment analysis e la multichannel consumer experience; mentre nelle attività di merchandising l'ottimizzazione dell'assortimento, l'ottimizzazione del prezzo e l'ottimizzazione del posizionamento e del design (Manyika J., 2011). L'utilizzo di queste tecniche impatta positivamente sulla customer experience del consumatore, dalla fase di prevendita a quelle di acquisto e di pagamento fino alla fase di post-vendita: il cliente riesce a

trovare il prodotto soddisfa al meglio i suoi bisogni e spende meno per trovare i prodotti al prezzo più vantaggioso. Anche su questo argomento ci focalizzeremo nel capitolo successivo.

1.7 GOVERNANCE DEI SISTEMI BDA&BI

Affinché le aziende riescano a sfruttare i Big Data, è necessario che al loro interno ci siano delle figure che si occupino della governance dei sistemi di Big DataAnalytics& Business Intelligence (BDA&BI). I due ruoli principali sono quelli del *Data Scientist* e dello *Chief Data Officer* (CDO), che descriviamo nel dettaglio di seguito.

1.7.1 DATA SCIENTIST

Il Data Scientist rappresenta il lavoro più affascinante (Ariker M. M. T., 2013)del ventunesimo secolo e, come abbiamo già visto in precedenza, la domanda di questa figura professionale è molto alta.

Ma chi è il Data Scientist? Il Data Scientist è una persona con una solida formazione in computer science, modellazione, statistica, matematica e Analytics, dotata di un forte sesto senso nel business e una grande abilità nel comunicare le sue scoperte ai leader dell'azienda tanto da influenzare l'approccio dell'organizzazione nell'affrontare le sfide di business. Egli esplora i dati da molteplici fonti al fine di ricavare insight nascosti che, trasformati, possono fornire un vantaggio competitivo o risolvere un problema urgente. Non si limita a cercare e riferire i dati, ma li osserva da diverse angolature, determinando il loro significato e scovando dei modi per applicarli. Per questo deve avere una conoscenza, almeno di base, della strategia aziendale (IBM).

In particolare, il Data Scientist deve sviluppare le seguenti 8 competenze (Holtz D., 2014):

- *basic tools*: saper utilizzare un linguaggio di programmazione statistica come R e un linguaggio di querying del Database come SQL;
- *basici statistic*: essere familiari con testi statistici, distribuzioni, stimatori di probabilità e capire quali tecniche rappresentano un approccio valido;
- *machine learning*: conoscere le diverse tecniche di machine learning e capire quando implementare l'una o l'altra;
- *Data munging*: sapere come affrontare le imperfezioni dei dati, quali valori mancanti, formattazione errata di stringhe o dati;