UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"BIG DATA: EVOLUZIONE DEL SISTEMA IMPRENDITORIALE"

RELATORE:

PROF.SSA LUISA BISAGLIA

LAUREANDO: BONFIO DAVIDE

MATRICOLA N. 1112989

ANNO ACCADEMICO 2018-2019

“Il candidato, sottoponendo il presente lavoro, dichiara, sotto la propria personale responsabilità, che il lavoro è originale è che non stato già sottoposto, in tutto in parte, dal candidato o da altri soggetti, in altre Università italiane o straniere ai fini del conseguimento di un titolo accademico. Il candidato dichiara altresì che tutti i materiali utilizzati ai fini della predisposizione dell’elaborato sono stati opportunamente citati nel testo e riportati nella sezione finale ‘Riferimenti bibliografici’ e che le eventuali citazioni testuali sono individuabili attraverso l’esplicito richiamo al documento originale.”

INDICE

INTRODUZIONE

Sebbene il fenomeno dei *Big Data* si sia manifestato agli inizi del nuovo millennio, non è tutt'ora stata formulata una loro precisa definizione; tra le varie proposte, quella che sembra più oggettivamente descrittiva è quella proposta da McKinsey (2011): "*un sistema di Big Data si riferisce a un dataset il cui volume è talmente grande che eccede la capacità dei sistemi di database relazionali di catturare, immagazzinare, gestire ed analizzare*".

Da questa definizione si può comprendere la difficoltà riscontratasi globalmente in un primo momento della comprensione di questo fenomeno e successivamente della sua analisi ed al suo impiego empirico. Queste nuove entità sono il motore trainante di quella che viene denominata "Quarta Rivoluzione Industriale" e, come le rivoluzioni precedenti, non limita la propria influenza al solo campo economico: con l'avvento dell'era dei *Big Data* si costituisce un nuovo approccio al raccoglimento, all'analisi ed all'interpretazione dei dati che, venendo generati anche dagli umani stessi, portano con sé informazioni sensibili, che chiaramente necessitano di un trattamento riservato a cui sia garantito un elevato livello di sicurezza. Con questi presupposti ci si avvia ad un cambiamento radicale: un nuovo empirismo sociale (Arbia, 2018) al quale ogni individuo, nel pubblico e nel privato, dovrà adattarsi acquisendo un approccio empirico-deduttivo, in contrasto con l'attuale logico-deduttivo e gli enti preposti, siano essi pubblici o privati, al raccoglimento e all'analisi di questi dati dovranno dotarsi di strutture e mezzi idonei ad ottemperare a questo compito, assicurando riservatezza e sicurezza.

Il compito centrale delle discipline statistico-informatiche continua ad essere quello di creare nuove strumentazioni e modelli per l'interpretazione dei *Big Data*, adottando un approccio cerebrale per la risoluzione di questo problema, in contrasto con una soluzione di tipo muscolare che consisterebbe nel continuare a potenziare gli strumenti utilizzati dalla statistica classica. Ad oggi persiste ancora una mancanza di trattazioni accademiche a riguardo, ma la comunità scientifica sta riservando una discreta importanza a questa tematica, il che preannuncia che in un prossimo futuro la comprensione di queste entità sarà ancora più completa.

In questa trattazione verrà dapprima affrontato il fenomeno dei *Big Data*, fornendo le principali caratteristiche e i metodi più diffusi per il loro rilevamento ed analisi, con un breve inciso riguardante la sicurezza dei dati *user-generated*; in un secondo momento verranno discusse le loro applicazioni a livello industriale ed infine verrà analizzato il metodo con cui Amazon, mediante l'utilizzo di dati e algoritmi, sia divenuta l'impresa leader mondiale nel proprio settore.

CAPITOLO 1 BIG DATA: CARATTERISTICHE ED ANALISI

Al giorno d'oggi, esiste una enorme quantità di dati generati quotidianamente nella produzione e fornitura di beni e servizi, nelle ricerche scientifiche e nelle vite private delle persone. L'aumento del volume dei dati nel mondo digitale sembra crescere ad un tasso superiore rispetto alle infrastrutture computazionali; processi adatti al raccoglimento e all'interpretazione dei dati sono in grado di rivelare nuove conoscenze in merito ai mercati finanziari, società e ambiente e, se correttamente interpretati, possono delineare perfettamente la personalità di un soggetto o individuare dei Blue Ocean in campo economico. Le convenzionali tecnologie di processamento dati, come i *database* e le *data warehouse*, sono diventate inadeguate nella gestione di questi volumi di dati che vengono processati. Questa nuova sfida è conosciuta anche come *Big Data* e a causa della loro natura ubiquitaria, a questa tematica è stata riservata un'enorme attenzione negli ultimi anni.

1.1 CARATTERISTICHE DEI BIG DATA

Non esiste una definizione universalmente accettata dei *Big Data*, anche se si è soliti credere che i Big Data dovrebbero includere data set con ampiezza maggiore a quella normalmente utilizzata dagli strumenti software per catturare, gestire e processare dati in un accettabile lasso di tempo.
Sulla base di questo concetto, la comunità scientifica (George, Haas, & Pentland, 2014) è solita riassumere tre importanti aspetti dei *Big Data*, conosciuti anche come "le 3 V" e sono:

- Volume
- Velocità
- Varietà.

La problematica rappresentata dal volume dei dati è la più facilmente comprensibile. Nelle scienze, quali biologia, meteorologia, astronomia, i ricercatori incontrano costantemente limitazioni computazionali a causa del continuo aumento del volume dei dati. Nel web, applicazioni quali Google e Facebook stanno operando con numeri di utenti che non sono mai stati presi in considerazioni dalle applicazioni locali. La dimensione dei data set utilizzati dalle odierne applicazioni Internet può essere di così grande ampiezza da causare problemi computazionali; questa tipologia di problemi può essere riscontrata anche nelle aree di finanza, comunicazione e business informatici data la grande applicazione delle tecnologie dell'informazione e all'aumento dell'intensità delle transazioni online. Nonostante i problemi

legati al volume, non esistono tutt'ora opinioni comuni sulla quantificazione dei *Big Data* in quanto questa quantificazione dipende da innumerevoli fattori. In primo luogo, la complessità della struttura dei dati: un *dataset* relazionale di diversi Petabyte può non essere considerato un *Big Data* in quanto può essere prontamente gestito dai più aggiornati software di gestione *database* (DBMS). Al contrario, un *dataset* grafico di svariati Terabyte è comunemente considerato un *Big Data* in quanto il processamento dei grafici è molto laborioso e complesso per le attuali tecnologie. In secondo luogo, si dovrebbero tenere in considerazione anche requisiti richiesti dalle loro applicazioni: nelle ricerche biologiche un tempo di attesa di diverse ore è considerato inaccettabile, sistemi di *trading* automatizzati richiedono frazioni di secondo per ottenere una risposta, senza contemplare la grandezza dei dati. In realtà, i *Big Data* di grandi dimensioni sono un oggetto di studio in costante mutamento, con un range di grandezza che varia da pochi Terabyte a molti Zettabyte per singolo *dataset*, a seconda del contesto in cui questi dati sono utilizzati e generati.

Nome	Simbolo	Grandezza*
Chilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}

* espressa in multipli di byte

Tabella 1 Unità di misura per la memorizzazione delle informazioni

La problematica della velocità consiste nel dover gestire e migliorare la rapidità con il quale i nuovi dati sono creati o sono elaborati dati già esistenti; questa viene spesso riscontrata nei dati generati dai macchinari, come quelli provenienti da *device* elettronici personali. In queste applicazioni, grandi volumi di dati in *upload* generati senza sosta viaggiano tra sistemi e ciò fa nascere la necessità di applicazioni che diano un senso ai dati appena dopo la loro creazione. La velocità di generazione dei dati porta a delle complicazioni in ogni aspetto delle piattaforme di gestione dati: dai livelli di archiviazione ai livelli di elaborazione che necessitano di *query* veloci e scalabili[1]. La tecnologia dei data streaming ha impiegato anni a

[1] In informatica, la caratteristica di un sistema software o hardware facilmente modificabile nel caso di variazioni notevoli della mole o della tipologia dei dati trattati.

raggiungere prestazioni soddisfacenti riguardanti la velocità. La capacità degli attuali sistemi di streaming è comunque limitata, specialmente quando si deve affrontare il crescente numero di dati in arrivo negli odierni *network* di sensori, sistemi di telecomunicazioni, ecc.

Nelle applicazioni empiriche su dati reali, questi spesso non provengono da una singola fonte. L'implementazione dei *Big Data* richiede una gestione incrociata degli stessi da varie fonti in cui i dati possono presentarsi in diversi formati. Questo aspetto rappresenta la terza sfida dei *Big Data*: la varietà dei dati che porta molte informazioni per la risoluzione di problemi e migliora le qualità dei servizi. La questione è come riuscire a catturare le informazioni provenienti da diversi tipi di dati e come correlarli tra loro. Usualmente i dati possono essere classificati in 3 grandi categorie:

- **Dati strutturati**: sono quelli caratterizzati da uno schema, quindi di fatto quelli gestiti dai DBMS classici;
- **Dati semi-strutturati**: qui s'incontrano alcune delle caratteristiche dei dati strutturati e alcune delle caratteristiche dei non strutturati. Nonostante non vi siano limiti strutturali all'inserimento dei dati, le informazioni vengono, comunque, organizzate secondo logiche strutturate e interoperabili. XML, un linguaggio di testo per lo scambio di dati nel Web, è un classico esempio di dati semi strutturati. I documenti di XML contengono tag di dati definiti dall'utente che li rendono leggibili dalle macchine;
- **Dati non strutturati**: si tratta di dati completamente privi di schema. Possono essere identificate due categorie di dati non strutturati: dati grezzi, ad esempio immagini, e dati senza schema, ad esempio porzioni di testo.

Esistono tecnologie molto sofisticate per gestire ognuna di queste tipologie di dati, come i database relazionali e tecniche per il recupero dei dati. Ciò nonostante un'integrazione senza soluzione di continuità di queste tecnologie rimane ancora una sfida.

Nel corso del tempo sono state proposte altre 3V per cercare di rendere la descrizione di questi oggetti di studio più precisa e completa. Queste sono:

- **Veridicità:** IBM ha coniato questo termine come quarta V che rappresenta l'inaffidabilità inerente ad alcune fonti di dati. Per esempio, le emozioni dei consumatori nei *social media* hanno natura incerta, dal momento che si compongono di giudizi umani quali valutazioni di prodotti o previsioni dell'andamento dei mercati finanziari; inoltre, contengono informazioni discrezionali e perciò non oggettive. Sorge quindi il bisogno di trattare con dati imprecisi e incerti.
- **Variabilità:** sono stati introdotti i concetti di Variabilità e Complessità come due dimensioni aggiuntive di Big Data. La Variabilità si riferisce alla variazione dei tassi di

flussi dei dati. Spesso, la velocità di generazione dei dati non è constante nel tempo, presentando periodici picchi. La Complessità si riferisce al fatto che i Big Data sono generati attraverso una miriade di risorse e fonti. Ciò impone un compito imprescindibile: il bisogno di pulizia e trasformazione dei dati ricevuti da differenti risorse.

- **Valore:** la Oracle ha introdotto il termine Valore come attributo definitivo di Big Data. Basandosi sulla definizione data da Oracle, i Big Data sono spesso caratterizzati dalla loro relativa "bassa densità di valore." Significa che i dati ricevuti nella forma originale spesso hanno un basso valore di contenuto relativamente al loro volume. Tuttavia, un alto valore può essere ottenuto analizzando un ampio volume di questi dati.

1.2 OPERAZIONI SUI BIG DATA

Nel grafico sottostante (Chen, et al., 2013) sono sintetizzati i principali processi per ottenere informazioni utili partendo da dati grezzi.

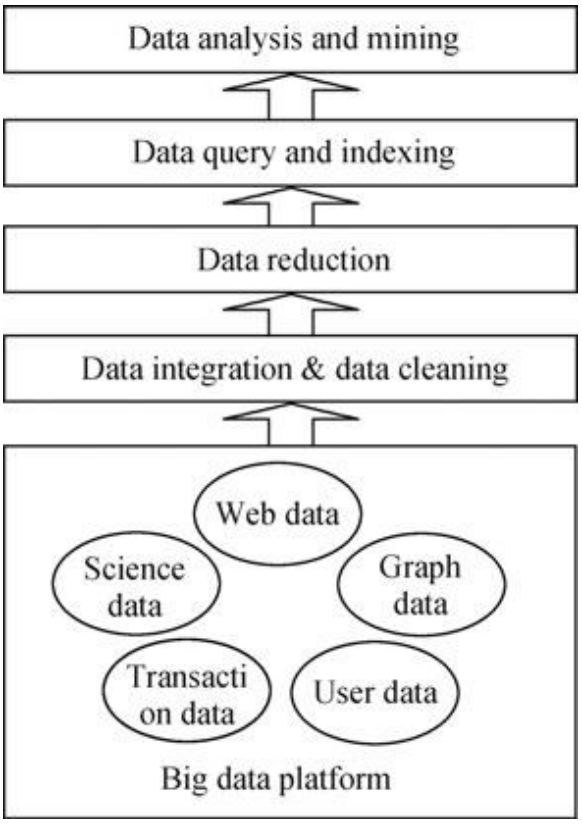

Grafico 1 CHEN J, CHEN Y, DU, LU, ZHAO, ZHOU (2013)

1.2.1 Le fonti dei Big Data

I *Big Data* sono generati da una crescente pluralità di fonti, inclusi gli "*Internet clicks*", le transazioni compiute da *smartphone*, i dati creati dall'utente e i dati generati dai *Social Media* come i contenuti appositamente generati attraverso reti di sensori o transazioni commerciali come ad esempio le domande di vendite online; questi dati richiedono, per essere analizzati, una forte capacità computazionale e tecniche specifiche per individuare *trend* e modelli ricorrenti all'interno di una moltitudine ed eterogeneità di dati. Nuove intuizioni ottenute dall'estrazione dei dati possono coadiuvare ed essere complementari alla statistica ufficiale, ai sondaggi ed agli archivi che rimangono in larga misura statici aggiungendo intuizioni ottenute dall'esperienza collettiva, il tutto compiuto in tempo reale, migliorando sia l'informazione in sé e riducendo il tempo necessario ad ottenere questi risultati.

La caratteristica di "*BIG*" nei *Big Data* è data dall'ampiezza dei dataset. Nella comunità scientifica è sorta la disputa riguardante il fatto che *BIG* non sia più un parametro di definizione bensì quanto "intelligenti" questi dati siano.

I *Big Data* sono prodotti incessantemente da una crescente base di soggetti, siano essi d'origine umana o generati da macchine o da processi; questa pluralità di fonti può essere categorizzata in cinque macroaree come segue:

Dati pubblici: sono generati dai singoli individui e solitamente archiviati e gestiti dal governo e da organizzazioni governative o da comunità locali. Esempi di questi dati sono quelli concernenti l'utilizzo di servizi pubblici, l'uso di energia, dati anagrafici e dati sulla sanità a cui si ha accesso, anche se con dei limiti dati dalle leggi sulla privacy.

Dati privati: sono dati generati da imprese private, organizzazioni *non profit* e da individui, che riflettono informazioni riservate che non possono essere direttamente imputate alle fonti pubbliche. Per esempio, i dati privati includono le transazioni degli acquirenti, tag dei prodotti via onde radio utilizzate dalle imprese nelle *supply chain*, nelle movimentazioni di risorse, nel *browsing* e dati prodotti dalla navigazione in Internet.

Dati residui: sono riferiti a dati che sono raccolti passivamente, dati non importanti con valore nullo o semi-nullo per i soggetti che li raccolgono. Questi dati sono raccolti per diversi scopi e possono creare valore aggiunto solamente se ricombinati ed associati ad altri dati. Quanto degli individui svolgono attività, siano esse particolari o quotidiane, utilizzando *device* elettronici come gli *smartphone*, generano dei dati senza alcun valore in sé.

Community data: sono un distillato di dati non strutturati, specialmente testi, creati in *network* dinamici che rappresentano i *trend* nei social. Esempi comuni di questi tipi di dati sono: le recensioni di prodotti date dagli utilizzatori, le valutazioni di applicazioni, fino ad arrivare ai

post su Twitter. Questi *community data* possono essere elaborati con metodo inferenziale per cercare di costruire modelli che rispecchino e spieghino determinati comportamenti sociali.

Dati qualificanti: sono tipi di dati prodotti da individui attraverso azioni particolari che rispecchiano i propri comportamenti e atteggiamenti. Nel campo della psicologia, gli individui dichiarano le loro preferenze di ciò che vorrebbero fare in contrapposizione alle "preferenze rivelate" le quali sono ottenute solamente tramite da un'analisi del comportamento. I dati qualificanti aiutano le interconnessioni tra psicologia e i comportamenti umani; i ricercatori di scienze sociali provenienti da differenti aree come psicologia, *marketing* possono beneficiare, utilizzandole nelle proprie ricerche, dei dati riguardanti le preferenze implicite e dichiarate, riuscendo a costruire dei modelli di comportamento che rispecchino le reali preferenze degli individui.

In molti casi, *i Big Data* non hanno delle chiare strutture relazionali tra loro; molto spesso contengono informazioni di diverso tipo come immagini, testi, *metadata*, ecc. Questo significa che le informazioni utili non possono essere semplicemente estratte utilizzando un singolo modello di analisi. Per diversi tipi di dati possono essere applicate diverse strategie. Per grandi dati relazionali, le performance dei più comuni *DBMS* che si basano su di un singolo nodo[2], calano significativamente quando il volume di dati supera il centinaio di GB.
La forte richiesta delle cosiddette performance ACID[3] rendono meno efficaci le performance della tradizionale RDBM. Parallelamente, soluzioni di gestione dei dati come le OceanBase (piattaforme di storage online che permettono altresì l'elaborazione dei dati) hanno provato a risolvere le problematiche della scalabilità per processare online grandi volumi di dati relazionali. *Database in memory* come VoltDB e HANA sono state recentemente commercializzate in risposta alle sfide di performance per OLTP e OLAp.

Per grandi dati di tipo grafico molte soluzioni riguardanti il processamento dei dati grafici non possono essere contenuti nelle memorie proposte recentemente. Pregel è una di queste, in più numerosi algoritmi sono stati proposti per l'estrazione di queste tipologie di dati e la gestione delle applicazioni, data la natura computazionale molto complessa degli algoritmi grafici.

I dati non strutturati non possono essere effettivamente compresi e processati efficientemente quando si presentano allo stato grezzo. Le tecniche di estrazione delle informazioni, che vengono applicate per estrarre dati importanti e facilmente gestibili

[2] In informatica i computer su cui sono memorizzate ed elaborate le informazioni dei Database vengono denominati nodi.
[3]. Nell'ambito dei database, ACID deriva dall'acronimo inglese Atomicity, Consistency, Isolation, e Durability (Atomicità, Coerenza, Isolamento e Durabilità) ed indica le proprietà logiche che devono avere le transazioni per risultare corrette.

proveniente da dati senza struttura, sono in costante miglioramento per cercare di ridurre la perdita di informazione, causata dal processamento di dati allo stato grezzo, in quanto le informazioni estratte risultano essere sommarie; in molte applicazioni, la dimensione dei dati è talmente grande che un piccolo gruppo di essi è sufficientemente accurato da poter supportare il bisogno di analizzare e processare i *Big Data*.

1.2.2 Integrazione e pulizia dei dati

Per quanto concerne la fase di pulizia ed integrazione dei dati è stata condotta una ricerca (Chen, et al., 2013) per valutare il cambiamento delle parole chiave nelle trattazioni accademiche riguardanti *Big Data* in fase di integrazione dei dati. La ricerca evidenzia come le problematiche della fase di integrazione siano suddivise in vari *step*, con l'intento di ridurre la complessità di questa macro-operazione, suddividendole in operazioni di minor entità e quindi più maneggevoli come riportato di seguito.

• **Utilizzare dati integrati da comunità di utenti o crowdsourcing:** Dati i limiti delle intelligenze artificiali, sono sempre presenti errori in fase di mappatura degli schemi dei dati generati automaticamente dai processi. Solitamente questi errori vengono sistemati dagli esperti di dominio e risulta evidente come questo approccio non possa funzionare nell'era dei *Big Data*; molti ricercatori suggeriscono di adottare utenti e/o comunità di utenti per migliorare la qualità dei dati integrati. Gli autori dell'esperimento propongono un metodo per confermare i *feedback* degli utenti valutando l'utilità di campioni. Diversi algoritmi sono stati proposti per incorporare automaticamente i *feedback* e migliorare i già presenti programmi di integrazione dati. Si noti che le comunità possono non essere formate dagli stessi utenti del sistema di integrazione dati ma il *crowdsourcing* solitamente significa utilizzare volontariamente la conoscenza di molte persone.

• **Incertezza sulla provenienza:** L'incertezza è divenuta una proprietà intrinseca in molte applicazioni per l'integrazione dati. Si necessita di condurre l'integrazione e la pulizia basandosi su dati imprecisi; a questo scopo solitamente viene costruito un modello probabilistico per rappresentare l'incertezza dei dati, così da poter essere comunque prese decisioni; tutto ciò con la necessità di mantenere traccia della loro provenienza e le interconnessioni tra loro.

• **Pay-as-you-go:** è impossibile costruire una perfetta integrazione per i *Big Data*, a causa del loro volume e dell'alta velocità di accumulazione. Quindi una ragionevole via di operare è quella di costruire un sistema imperfetto che può provvedere a fornire le informazioni

necessarie e incrementalmente migliorare questi sistemi qualora fossero disponibili nuove tecnologie e risorse. Questa è esattamente l'idea che il *pay-as-you-go* sottende: i sistemi di questo tipo solitamente si occupano anche dell'incertezza dei dati e della loro provenienza, tendendo a utilizzare l'intelligenza umana per migliorare la qualità dell'*output***.**

• **Entity matching and resolution:** L'obiettivo della risoluzione dell'entità è di identificare quali *record* si riferiscono alla stessa entità nel mondo reale, compito fondamentale nell'integrazione dati. Recentemente, questo si è dimostrato l'argomento che ha catturato maggiormente l'attenzione in quanto è crescente l'interesse nell'estrazione di tabelle da pagine web. Svariati approcci sono stati proposti per migliorare la qualità della risoluzione dell'entità come la ricombinazione di differenti modelli, approcci interattivi, usare dipendenze funzionali, ecc.

1.2.3 Sintesi dei dati

La sintesi dei dati consiste nella loro riduzione, partendo da una moltitudine di essi fino ad estrarre solamente le loro parti più importanti. In altre parole, la sintesi dei dati è la trasformazione numerica o alfabetica di informazioni digitali ottenute empiricamente o sperimentalmente in forme di dati corretti, ordinati e semplificati. Gli strumenti di sintesi dei *Big Data* permettono di operare con i dati analitici, evitando così di dover operare con complessi e grandi volumi di dati grezzi. Sono due le metodologie più utilizzate per la sintesi dei dati:

- **Machine learning**: questa tecnica consiste in differenti meccanismi che permettono, nel tempo, ad una macchina intelligente di migliorare le proprie capacità e prestazioni; la macchina, quindi, sarà in grado di imparare a svolgere determinati compiti migliorando, tramite l'esperienza, le proprie capacità, le proprie risposte e funzioni. Alla base dell'apprendimento automatico ci sono una serie di differenti algoritmi che, partendo da nozioni primitive, renderanno la macchina autonoma nel prendere una specifica decisione piuttosto che un'altra o effettuare azioni apprese nel tempo. I *Big Data* richiedono tecnologie così sofisticate per processare efficientemente un grande ammontare di dati in un lasso di tempo tollerabile, visto che le tradizionali tecniche di sintesi dati potrebbero non considerare tutte le modalità ottenute. È possibile che alcune tecniche di *machine learning* possano aiutare a comprendere i *trend* dei dati, classificandoli in categorie, sottolineando possibili fattori comuni, predicendo il futuro basandosi sul passato. Utilizzando una soluzione completamente *machine learning-based* sui *Big Data* si potrebbero individuare frodi, portare prodotti sul mercato più rapidamente ed essere generalmente più competitivi.

- **Massively parallel processing:** consiste nell'elaborazione coordinata di un programma da parte di più processori che agiscono su diverse parti del programma stesso, con ciascun processore che utilizza la propria memoria e sistema operativo. È considerata la seconda possibilità per la sintesi efficace dei *Big Data*; alcune tecnologie di processamento parallelo quali il *cloud computing database*, griglie di estrazione dati e sistemi di distribuzione file sono stati proposti dalla comunità scientifica. Tra questi, il *cloud computing* è uno di quelli più interessanti. Il *cloud computing* è un metodo che permette accessi ai *network on demand* a delle risorse informatiche condivise che possono essere rapidamente approvvigionate e rilasciate con un minimo sforzo gestionale con un numero minimo di interazioni col provider. Migliorando le tecniche di *cloud computing*, queste possono essere utilizzate per sintetizzare i dati con l'aiuto di processi paralleli.

1.2.4 Indicizzazione e interrogazione dei big data

Oggigiorno, varie forme di dati in tutti i campi hanno invaso ogni aspetto della nostra vita quotidiana. Quando si considerano le interrogazioni e l'indicizzazione dei *Big Data*, inevitabilmente alcune problematiche sorgono: la prima consiste nella pesantezza di un'informazione digitale che risulta essere troppo grande per molti dei software e per le persone che gestiscono i processi, inoltre, una singola macchina non può elaborare tutti i tipi di dati che dovrebbero essere immagazzinati nei sistemi di archivio. L'indicizzazione dei *Big Data*, distinguibile da quella tradizionale, dovrebbe essere basata sui sistemi di distribuzione ed ulteriori teorie sull'interrogazione di queste entità dovrebbero essere proposte, per consentire un'efficace operazione su queste entità, sanando quello che ancora oggi persiste ad essere una mancanza in campo accademico. In secondo luogo, i *Big Data* non si riferiscono solamente a *dataset* di grande volume ma oltretutto di complessa natura e composto da dati eterogenei. Infine, le strutture ad albero che hanno raggiunto un'ottima popolarità nell'indicizzazione tradizionale, nel campo dei *Big Data* non sono per niente ottimali in quanto è molto difficile evitare tempi lunghi di attesa a causa dei colli di bottiglia che si vengono a creare quando i dati cominciano ad essere molto numerosi. D'altra parte, la tolleranza d'errore è un fattore che non può essere sottovalutato in fase di indicizzazione.

I ricercatori, per quanto riguarda le problematiche sopracitate, hanno confrontato differenti metodi sperimentali, tra questi, il *B-tree* è uno dei più accreditati. L'obiettivo di tale metodo è di operare consistenti update simultanei e nel mentre permettere la lettura di numerosi dati contemporaneamente. Inoltre, può fornire la migrazione dei nodi di alberi online ed addizioni e sottrazioni dinamiche dei server.

Un'altra metodologia per assolvere questi compiti è l'uso del BATON per supportare il processo di interrogazione. Il BATON (*Balanced Tree Overlay Network)* è una struttura ad albero bilanciata che permettere la connessione dei network peer to peer. In un *network* ad N, il BATON può garantire che sia le *query* che le *range query* siano disposte in O(log N) *step*, risparmiando risorse per O(log N). Un *framework* di indicizzazione basato sul CGI[4] può ridurre l'ammontare dei dati trasferiti all'interno del *cloud* e facilitare la distribuzione dei *database* di tipo *back-end*. Quando si riassumono i *Big Data* in parallelo si può, inoltre, presentare un *trade-off* tra accuratezza ed efficienza senza per questo trascurare la scalabilità.

1.2.5 Analisi ed estrazione dei Big Data

I *Big Data*, se non correttamente analizzati, sono delle informazioni senza valore (Gandomi & Haider, 2015). Il loro valore potenziale è realizzato solo quando viene sfruttato per guidare il processo decisionale. Per affidarsi ad un processo decisionale basato sull'evidenza, le aziende hanno bisogno di processi efficienti per trasformare volumi elevati di dati in movimenti rapidi e differenti in intuizioni significative. Tutto il processo di estrazione dai *Big Data* può essere suddiviso in cinque passaggi, illustrati nel seguente grafico.

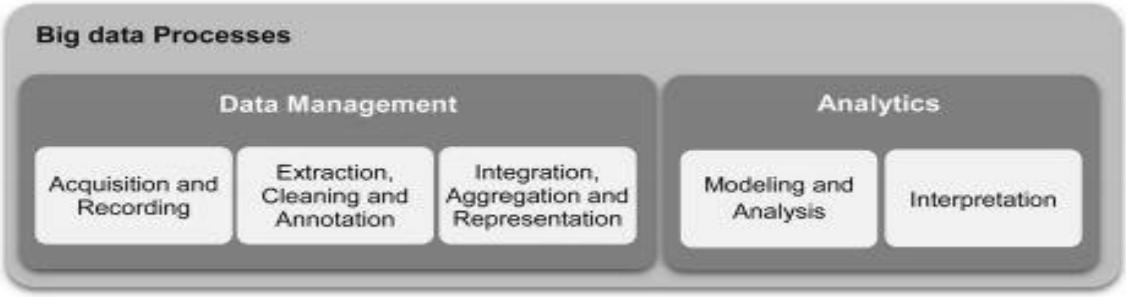

Grafico 2 Fonte: Grandomi, Haider (2014)

Queste cinque fasi formano i due principali sotto processi quali la gestione dei dati e la gestione dell'analisi. La prima coinvolge i processi e le tecnologie di supporto per acquisire e memorizzare i dati, prepararli e recuperarli per l'analisi. *Analytics*, d'altra parte, si riferisce alle tecniche utilizzate per analizzare e acquisire informazioni utili dai *Big Data*. Pertanto, è possibile visualizzare l'analisi dati come sotto processo nel processo complessivo di *'estrazione dell'insight'* da grandi insiemi di dati. L'estrazione delle informazioni, con le relative tecniche applicate, si differenzia a seconda della natura del dato e nella forma questo si presenta. Di seguito sono riportate le varie tipologie d'analisi adottate a seconda dei tipi di dato.

[4] il CG index (*Cloud Global index)* uno schema di indicizzazione scalabile per i sistemi di gestione dati in cloud

Analisi testuale

L'analisi testuale si riferisce alle tecniche che estraggono informazioni da dati di testo. *Feed* di *social network, e-mail, blog, forum online*, risposte al sondaggio, documenti aziendali, notizie e di registri del call center sono esempi di dati testuali detenuti dalle organizzazioni. L'analisi di testi racchiude analisi statistiche, linguistica computazionale e *machine learning*. L'analisi di testo permette di convertire grandi volumi di testi generati dagli utenti in riassunti significativi, che supportano il processo decisionale basato su prove evidenti. Per esempio, l'analisi testo può essere usata per predire l'andamento di mercato basandosi sulle informazioni estratte dai giornali finanziari.

Per estrarre informazioni utili dai testi vengono frequentemente usate queste tecniche: tecniche di *Information extraction (IE)* estraggono dati strutturati da testi non strutturati. Per esempio, algoritmi IE possono estrarre informazioni strutturate come il nome di medicinali, dosaggio e frequenza di assunzione per prescrizioni mediche. Due compiti secondari nella EI sono *Entity Recognition (ER)* e *Relation Extraction (RE)*. ER trova nomi nei testi e li classifica in categorie predefinite come persone, date, luoghi e società. RE trova ed estrae relazioni semantiche tra entità come persone, società, medicine, generi, dai testi. La tecnica di *text summarization* produce automaticamente riassunti succinti di documenti singoli o multipli. Il riassunto che ne risulta trasmette le informazioni chiave nel testo originale. A grandi linee, il riassunto segue due approcci: l'approccio estrattivo e l'approccio astrattivo. Nel sommario estrattivo, un riassunto è creato da unità del testo originale. Il riassunto che ne risulta è un *subset* del documento originale. Basandosi sull'approccio estrattivo, la formulazione di un riassunto implica la determinazione di unità salienti di test e li mette insieme. L'importanza delle unità di testo è valutata dall'analisi della loro posizione e frequenza nel testo. La tecnica di riassunto estrattivo non richiede una "comprensione" del testo. Al contrario, la tecnica di riassunto astrattivo coinvolge l'informazione dell'estrazione semantica del testo; il sommario contiene unità di testo che non sono necessariamente presenti nel testo originale. Al fine di analizzare il testo originale e creare il riassunto, il sommario astrattivo incorpora un'avanzata tecnica di *Natural Language Processing* (NLP). Come risultato, il sistema astrattivo tende a generare riassunti più coerenti rispetto al sistema estrattivo. Comunque, i sistemi estrattivi sono più facili da adottare, specialmente per i *Big Data*.

La tecnica di *Question Answering* (QA) fornisce risposte a domande poste nel linguaggio naturale. Siri di Apple e Watson di IBM sono esempi di sistemi QA commerciali. Questi sistemi sono stati sviluppati per il loro utilizzo nella sanità, nella finanza, nel marketing e nell'istruzione. Similmente al riassunto astrattivo, i sistemi QA fanno affidamento sulle

complesse tecniche di NLP. Tecniche QA sono classificate in 3 categorie: l'approccio basato sul reperimento di informazioni (IR), l'approccio basato sulla conoscenza e l'approccio ibrido.

IR basato sulla QA spesso contiene 3 sottocomponenti. Il primo è il *question processing*, usato per determinare i dettagli, come il tipo di domanda, focus della domanda e tipo di risposta che vengono utilizzati per creare una *query*. Il secondo è il *document processing* usato per recuperare passaggi rilevanti prescritti da set di documenti, usando *query* formulate in elaborazioni di domande. Il terzo componente è l'*answer processing*, usato per estrarre le risposte ottenute dall*'output* del componente precedente, che li classifica e restituisce il punteggio più elevato per la risposta ottimale come output del sistema QA. Sistemi di controllo della qualità basati sulla conoscenza generano descrizioni semantiche del quesito, che viene poi usato per tabellare le risorse strutturate.

Analisi audio

L'analisi audio analizza ed estrae informazione di dati audio non strutturati. Quando applicata al linguaggio parlato umano, l'analisi audio è riferita anche alla cosiddetta analisi del discorso. Da quando questa tecnica è stata applicata all'audio di conversazioni, il termine analisi audio e analisi del discorso sono stati spesso usati in modo intercambiabile. Attualmente, *call center* dei consumatori e la sanità sono le aree primarie di applicazioni delle analisi audio. *I call center* usano l'analisi audio per analisi efficienti di migliaia o milioni di ore di chiamate registrate. Queste tecniche aiutano a migliorare l'esperienza del consumatore, valutare le performance degli agenti, accrescere il tasso di *turnover* delle vendite, monitorare le conformità con le differenti politiche, ottenere informazioni nel comportamento del cliente e identificare i problemi relativi a prodotti o servizi, ecc. I sistemi di audio analisi possono essere progettati per analizzare chiamate live, formulare raccomandazioni di *cross-up selling* basate sulle interazioni passate e presenti del consumatore e fornire *feedback* agli agenti di vendita in tempo reale. Inoltre, i *call center* automatici usano piattaforme di *Interactive Voice Response* (IVR) per identificare e gestire la l'emotività di coloro che effettuano le chiamate.

Rispetto all'assistenza sanitaria, le analisi audio supportano diagnosi e trattamenti per determinare condizioni mediche che affliggono i modelli di comunicazione del paziente (ad es. depressione, schizofrenia e cancro). In aggiunta, l'analisi audio può aiutare ad analizzare i pianti infantili, che contengono informazioni riguardanti lo stato emotivo e la salute del bambino. Il grande ammontare di dati registrati attraverso i sistemi di documentazione clinica basati sul linguaggio sono un altro *driver* per l'adozione dell'analisi audio nell'assistenza sanitaria. L'analisi del discorso segue due tecnologie di approcci comuni: l'approccio basato sulla

trascrizione, conosciuto come *Large-Vocabulary Continuos Speech Recognition* (LVCRS) e l'approccio basato sulla fonetica.

Il sistema LVRCS segue un processo a due fasi: indicizzazione e ricerca. Nella prima fase, il tentativo è di trascrivere il contenuto del discorso contenuto nel file audio. Questo è realizzato grazie all'uso dell'algoritmo *Automatic Speech Recognition* (ARS) che collega i suoni alle parole. Le parole vengono identificate basandosi su un dizionario predefinito, nel caso il sistema non riesca a trovare nel dizionario l'esatta parola, inserirà la parola più simile a quella cercata.

Il sistema basato sulla fonetica funziona attraverso i suoni o fonemi. I fonemi sono unità di suoni percettivamente distinguibili propri di una specifica lingua che caratterizza una parola. Il sistema basato sulla fonetica consiste di due fasi: indicizzazione fonetica e ricerca. Nella prima fase, il sistema traduce gli *input* dal discorso in una sequenza di fonemi. Questo contrasta il sistema LVRCS dove il discorso viene convertito in una serie di parole. Nella seconda fase, il sistema ricerca gli output della prima fare per una rappresentazione fonetica dei termini ricercati.

Video analisi

L'analisi video, conosciuta anche come analisi dei contenuti video (VCA) include una varietà di tecniche di monitoraggio, analisi ed estrazione delle informazioni fondamentali da *video stream*. Sebbene la video analisi sia ancora primordiale rispetto ad altri tipi di estrazione dati, varie tecniche sono già state sviluppate per processare i video in tempo reale allo stesso modo di quelli già registrati. La crescente prevalenza del circuito di telecamere a circuito chiuso (CCTV) e la forte popolarità dei video condivisi in streaming sono i due maggiori contributori alla crescita della video analisi computerizzata. Come sfida chiave, però persiste la grandezza dei dati dei video.

L'applicazione primaria di analisi video, negli ultimi anni, è passata a sistemi automatizzati di sicurezza e sorveglianza. In aggiunta al loro alto costo, i sistemi di sorveglianza basati sul lavoro tendono ad essere meno efficaci dei sistemi automatici. L'analisi video può essere efficace come il rilevamento di violazioni di zone soggette a restrizioni, identificazione di oggetti rimossi o lasciati incustoditi, rilevazioni di vagabondaggio, riconoscimento di attività sospette e rilevazione di manomissione di telecamere, per nominarne alcune. Al rilevamento di una minaccia, il sistema di sorveglianza potrebbe notificare al personale di sicurezza in tempo reale o attivare un'azione automatica.

I dati generati dalle telecamere CCTV nei negozi al dettaglio possono essere estratti per una *business intelligence*. Per esempio, algoritmi intelligenti possono raccogliere informazioni demografiche riguardo i consumatori, come l'età, il sesso e la provenienza etnica. Allo stesso

modo, i negozianti possono contare il numero di consumatori, misurare il tempo di permanenza nello *store*, rilevare i loro modelli di movimento, misurare il tempo di permanenza in diverse aree e monitorare le code in tempo reale. Intuizioni valide possono essere ottenute correlando queste informazioni con i dati demografici del cliente per guidare le decisioni relative a posizionamento del prodotto, prezzo, ottimizzazione dell'assortimento, *design* della promozione, *cross-selling*, *layout* ottimizzazione e personale.

L'indicizzazione automatica dei video e il loro recupero costituiscono un altro dominio per l'applicazione delle analisi video. L'indicizzazione di un video può essere eseguita basandosi su diversi livelli di informazione disponibili inclusi *metadata*, la traccia audio, le trascrizioni ed il contenuto video di un video. Nell'approccio basato sui *metadata*, un sistema di gestione relazionale di *Database* (RDBMS) sono usati per la ricerca dei video. L'analisi di audio e testi può esser applicata per indicizzare un video basandosi sui trascritti e sulla traccia audio ad esso associato.

In termini di architettura di sistema esistono due approcci per l'analisi video chiamati *Server-Based e Edge-Based:*

- **Architettura Server-based.** In questa configurazione, il video registrato attraverso ogni camera viene inviato ad un server centralizzato e dedicato che ne esegue l'analisi. A causa di limiti di banda, il video generato è solitamente compresso al fine di ridurre il numero di *frame* o la risoluzione dell'immagine; ciò comporta una perdita di informazione che può ridurre l'accuratezza dell'analisi.
- **Architettura Edge-based**. In questo approccio, le analisi sono applicate al vertice del sistema. Detto ciò, le analisi video sono eseguite localmente utilizzando dati grezzi registrati dalla camera; l'intero contenuto del video è disponibile in *streaming* per poter essere analizzato, permettendo una sua sintesi dei contenuti più accurata. Sistemi *Edge-Based* sono più costosi da mantenere ed hanno un minor potere di processamento comparati ai sistemi *server-based*.

Analisi dei social media

L'analisi dei *social media* si riferisce all'analisi di dati strutturati e non strutturati provenienti dai social. "*Social Media*" è un termine comune che racchiude al suo interno una varietà di piattaforme online che permettono agli utenti di creare e scambiare contenuti. I *social* media possono essere categorizzate nei seguenti tipi: *social networks* (Facebook e Linkedin), *blog* (WeWordPress), *microblogs* (e.g., Twitter and Tumblr), *social news* (e.g., Digg and Reddit), *social bookmarking* (e.g., Delicious and StumbleUpon), *media sharing* (e.g., Instagram and YouTube), *wikis* (e.g.,Wikipedia andWikihow), *question-and-answer sites* (e.g., Yahoo!

Answers and Ask.com) and *review sites* (e.g., Yelp, TripAdvisor). Inoltre, molte *app* per *smartphones* come Findmyfriend, forniscono una piattaforma per piattaforme sociali, quindi, fungono da canali di social media.

La caratteristica principale delle attuali analisi sui *social* è la loro natura data-centrica. La ricerca sui *social* media spazia attraverso diverse discipline includendo: psicologia, sociologia, antropologia, informatica, matematica, fisica ed economia. Un'applicazione primaria dei *social media* è quella del *marketing*, questo può essere attribuito ad una forte espansione e ad una crescente adozione dei social media dai consumatori di tutto il mondo. I dati generati dal consumatore e le relazioni ed interazioni tra entità di *network* sono le due fonti di informazione dei *social media*. Basandosi su questa categorizzazione, l'analisi sui *social* possono essere classificare in due gruppi:

- **Analisi basate sui contenuti**. Queste analisi si focalizzano sui dati "postati" dagli utenti di piattaforme *social*, quali *feedback* dei consumatori, recensioni di prodotti, immagini e video. Questi contenuti sono soventemente voluminosi, non strutturati e dinamici. Analisi sui testi, audio e video possono essere applicati per trarre informazioni da tali dati; inoltre, tecnologie *Big Data* possono essere adottate per individuare il canale con il quale processare i dati.
- **Analisi basate sulla struttura**. La struttura di un *social* è modellata attraverso una seria di nodi e collegamenti. Si analizzano due tipi di grafici *network* denominati Grafici Social e Grafici Activity. Nei Grafici Social, un collegamento tra un paio di nodi significa l'esistenza di una relazione tra le entità. Negli *activity networks,* i collegamenti rappresentano interazioni reali avvenute tra una coppia di nodi. Queste interazioni includono scambi di informazioni come like e commenti. I grafici sull'attività sono preferibili ai grafici social in quanto una relazione attività è più rilevante in fase di analisi rispetto ad una mera connessione.

Il rilevamento della comunità è riferita alla scoperta ed estrazione di implicite comunità all'interno dei *network*. Per i *social online*, le comunità si riferiscono a dei *sub-network* composti da utenti che interagiscono tra loro in maniera più intensiva. Molto spesso, queste comunità contengono milioni di nodi e collegamenti, quindi i *social online* tendono ad avere dimensioni colossali. La rilevazione di comunità aiuta a sintetizzare enormi *network* che possono facilitare la scoperta di modelli di comportamento e predire proprietà emergenti al loro interno. Rispetto a ciò, il rilevamento di comunità è simile al *clustering*, una tecnica di estrazione dati usata per sezionare gli stessi in gruppi disgiunti sulla base delle caratteristiche similari dei dati. Il rilevamento di comunità ha trovato diverse aree di applicazione nel

marketing e nel *World Wide Web*. Per esempio, ciò permette alle imprese di sviluppare prodotti più efficienti.

Analisi delle influenze dei social si riferiscono alle tecniche che riguardano la modellazione e valutazione dell'influenza degli attori e delle connessioni dei social. Naturalmente, il comportamento di un attore in un *social* è influenzato dall'atteggiamento degli altri utenti. Infatti, è desiderabile quantificare l'influenza dei partecipanti, la forza delle loro connessioni per scoprire schemi di influenza diffusi all'interno del *network*. Queste tecniche di analisi possono essere influenzate nei *marketing* virale per creare un'efficiente *brand awareness.*

Un saliente aspetto dell'analisi delle influenze dei *social* è di quantificare l'importanza dei nodi componenti i *network*. Varie misure sono state sviluppare a questo scopo includendo centralità di grado, centralità di mezzo, centralità di vicinanza e centralità di auto vettore. Altre misure valutano la forza di connessione rappresentata da un *Edge*, relazioni o modelli di influenza nei social. *The Liner Threshold Model* (LTM) e *Independet Cascade Model* (CICM) sono due esempi di *framework* molto conosciuti.

La *link prediction* (Gandomi & Haider, 2015) individua il problema della previsione di futuri collegamenti tra nodi esistenti nel *network* studiato. Le strutture dei *social* non sono statiche ma crescono in modo continuativo attraverso la creazione di nodi e di *Edge*. Le tecniche di link prediction rimirano all'accuratezza dell'interazione, della collaborazione o dell'influenza attraverso le entità di *network* in intervalli specifici di tempo. Questa tecnica restituisce un'accuratezza probabilistica del 40-50%, suggerendo che l'attuale struttura di *network* analizzata possa contenere informazioni latenti riguardo i futuri collegamenti. Nella sicurezza, la *Link Prediction* aiuta a scovare potenziali collaborazioni di terrorismo o di società criminali. Nel contesto dei *social*, la principale applicazione della *link prediction* sta nello sviluppo dei suggerimenti dei sistemi, nelle amicizie consigliate su Facebook e nei video correlati di YouTube.

Predicative analitica

L'analisi previsionale comprende una varietà di tecniche che predicono i futuri guadagni basandosi su dati storici e attuali. In pratica, questo tipo di analisi può essere applicato a tutte le discipline, dalla previsione di eventuali guasti alle componenti di un aereo basate nello stream di dati derivanti da migliaia di sensori, alla predizione del movimento successivo del consumatore riguardo a ciò che comprerà, quando lo comprerà e cosa comprerà basandosi sui suoi *social*.

Al suo centro, l'analisi previsionale cerca di scoprire modelli e relazioni attraverso i dati. La regressione lineare ha lo scopo di catturare l'interdipendenza tra le variabili esplicative

e le variabili dipendenti, utilizzandole al fine di creare delle previsioni. Basandosi sulla metodologia studiata, le tecniche possono essere categorizzate in due gruppi: tecniche di regressione e tecniche di apprendimento delle macchine *(neural network*). Un'altra classificazione è basata sul tipo di variabili di *outcome*: tecniche come la regressione lineare individuano continue variabili di outcome, mentre altre tecniche come la *random forests*[5] sono applicate con variabili di *outcome* discreti.

Le tecniche di analisi previsionale sono principalmente basate su metodi statistici: molti fattori vengono utilizzati al fine di sviluppare nuovi metodi statistici per i *Big Data*. In primo luogo, metodi statistici convenzionali traggono validità dalla significatività statistica: un piccolo campione è estratto da una più grande popolazione e, dopo aver elaborato un modello statistico atto a spiegare il fenomeno studiato, la significatività delle variabili è confrontata con valori critici standard per verificarne la significatività. La conclusione è poi generalizzata all'intera popolazione. In contrasto, i campioni di *Big Data* sono molto voluminosi e rappresentano la maggior parte della popolazione quindi la nozione di significatività statistica non risulterà più rilevante. In secondo luogo, in termini di efficienza di calcolo, molto metodi convenzionali usati per i campioni di piccole dimensioni non possono essere utilizzati per i *Big Data*. Il terzo fattore corrisponde alle caratteristiche distintive inerenti ai *Big Data*: eterogeneità, accumulazione dell'errore, correlazioni spurie ed endogeneità accidentale.

- **Eterogeneità**. I *Big Data* sono spesso ottenuti da diverse fonti e rappresentano informazioni da diverse sottopopolazioni. Come risultato, i dati raccolti sono altamente eterogenei. I dati della sottopopolazione, in piccoli campioni, sarebbero considerati *outliers* dagli strumenti statistici classici, a causa della loro bassa frequenza. Tuttavia, il grande volume dei *dataset* dei *Big Data* crea un'opportunità unica di modellare l'eterogeneità attraverso l'aumento dei dati della sottopopolazione.
- **Accumulazione dell'errore**. Modelli di previsione per i *Big Data* spesso includono stime simultanee di diversi parametri. Gli errori di stima accumulati da differenti parametri possono dominare la moltitudine di variabili che hanno effetti corretti all'interno del modello. In altre parole, alcune variabili con potenza esplicativa possono essere confuse a causa di errori di stima.
- **Correlazione spurie**. Queste correlazioni si riferiscono a variabili che sono falsamente considerate come correlate data l'enorme grandezza del *dataset*.

[5] La Random Forest è un metodo di apprendimento di insieme per la classificazione, la regressione e altre attività che operano costruendo una moltitudine di alberi decisionali al momento dell'addestramento e hanno come output la modalità delle classi (classificazione) o la previsione media (regressione) dei singoli alberi. La Random Forest corregge l'overfitting degli alberi decisionali.

- **Endogeneità accidentale**. Un'assunzione comune nelle analisi di regressione è che le variabili esplicative siano indipendenti dal termine d'errore. La validità di molteplici metodi statistici usati nell'analisi di regressione dipendono da questo assunto. In altre parole, l'esistenza dell'endogeneità accidentale (la dipendenza con il termine d'errore di alcune variabili esplicative) mette a rischio la validità del metodo statistico usata per l'analisi. Sebbene l'assunzione di esogeneità sia usualmente incontrata in piccoli campioni, l'endogeneità accidentale è comunemente presente nei *Big Data*. È degno di nota menzionare che, in contrasto con le correlazioni spurie l'endogeneità accidentale si riferisca ad effettive correlazioni tra variabili e termine d'errore.

  L'irrilevanza statistica, la sfida di capacità computazionale e le caratteriste uniche dei Big Data succitate sottolineano la necessità di sviluppare nuove tecniche statistiche per ottenere informazioni da modelli previsionali.

1.3 SICUREZZA E PRIVACY DEI BIG DATA

Nel momento in cui si comprende la pervasività di generazione e analisi di questi dati, si può facilmente intuire come queste entità portino ad una problematica riguardante la privacy degli individui.

1.3.1 Scandalo Cambridge Analytica

"Addio Cambridge Analytica. La controversa società di analisi e consulenza politica, finita nella bufera per violazione della privacy e abuso di dati personali raccolti attraverso il social network Facebook, ha chiuso definitivamente i battenti." (Valsania, 2018)

Con queste parole inizia l'articolo de "Il Sole 24 Ore" che annuncia la chiusura della società di analisi dati inglese, dovuta allo scandalo denominato "*Datagate*". Cambridge Analytica aveva ottenuto contratti per 15 milioni di dollari durante la campagna elettorale del 2016 negli Stati Uniti, che si è rivelata allo stesso tempo l'apice del suo business e del suo successo e l'inizio della sua precipitosa caduta (Cadwalladr & Graham-Harrison, 2018).

Questo scandalo riguarda la profilazione di oltre 50 milioni di individui, attraverso l'estrazione e analisi di dati provenienti da Facebook, senza alcuna autorizzazione. La raccolta dati è stata effettuata attraverso un'applicazione chiamata Thisisyourdigitallife, creata a scopi accademici. La compagnia del creatore di questa *app* (Aleksandr Kogan) era affiliata a Cambridge Analytica che, tramite la app, è entrata in possesso delle informazioni Facebook non solo degli utenti pagati per rispondere ai questionari della app stessa, ma anche dei loro amici, creando una base dati per la profilazione di milioni di utenti. Questo al fine di fornire, allo staff

di Donald Trump, all'inizio del 2014, una descrizione psicologica della base votante dei cittadini USA, così da poter creare una propaganda mirata per ogni persona; utilizzando la stessa metodologia e per lo stesso fine, Cambridge Analytica ha collaborato altresì con il partito di Nigel Farage, per indirizzare una propaganda mirata in favore alla Brexit.

1.3.2 GDPR e nuove direttive sulla privacy

Dal 25 maggio 2018, si trova piena applicazione il Regolamento generale sulla protezione dei dati personali: si tratta del Regolamento europeo 2016/679, noto anche come GDPR – *General Data Protection Regulation*, direttamente applicabile negli Stati membri UE.

L'Italia ha recepito questa normativa con la pubblicazione nella Gazzetta ufficiale del Decreto 101 del 10 agosto 2018. Il suo obiettivo è di dare all'Europa, ai propri Stati membri e ai suoi cittadini una normativa comune sul trattamento dei dati personali dei cittadini stessi, anche alla luce dell'innovazione tecnologica ed economica degli ultimi anni.

Il Decreto sulla privacy:

- definisce in modo chiaro cosa si intenda per comunicazione e diffusione dei dati personali;
- individua nel Garante della privacy l'autorità incaricata del controllo e della promozione delle regole deontologiche in materia;
- stabilisce che il consenso al trattamento dei dati personali potrà essere espresso solo al compimento dei 14 anni di età. Chi ha un'età inferiore necessita del consenso di chi esercita la sua responsabilità genitoriale. Il consenso poi, deve essere richiesto dal titolare del trattamento in modo chiaro e semplice, facilmente comprensibile dal minore (Capo II art. 2 del Decreto);
- tutti gli organi giudiziari avranno l'obbligo di nominare il DPO[6] e si precisano le limitazioni ai diritti degli interessati in relazione a ragioni di giustizia. Si rafforza il divieto di pubblicazione dei dati dei minori, e si prevede una relativa sanzione penale a riguardo;

[6] il Data Protection Officer è un supervisore indipendente, il quale sarà designato obbligatoriamente, da soggetti apicali di tutte le pubbliche amministrazioni e nello specifico è previsto l'obbligo nel caso in cui "il trattamento è effettuato da un'autorità pubblica o da un organismo pubblico, eccettuate le autorità giurisdizionali quando esercitano le loro funzioni giurisdizionali".

- considera ovviamente rilevante l'interesse pubblico, che può portare ad utilizzare i dati personali di determinati soggetti;
- dovranno essere adottate misure adeguate di sicurezza, come tecniche di cifratura e di pseudonomizzazione a tutela del dato personale, misure di minimizzazione e le specifiche modalità per l'accesso selettivo ai dati;
- le misure di garanzia che riguardano i dati genetici e il trattamento dei dati relativi alla salute per finalità di prevenzione, diagnosi e cura sono adottate sentito il Ministro della salute che, a tal fine, acquisisce il parere del Consiglio superiore di sanità;
- è ammesso l'utilizzo dei dati biometrici con riguardo alle procedure di accesso fisico e logico ai dati da parte dei soggetti autorizzati, nel rispetto delle misure di garanzia e protezione;
- al Garante viene assegnato il compito di scrivere le misure di garanzia per il trattamento di dati genetici, biometrici, sanitari;
- viene introdotto il concetto di diritto all'eredità del dato in caso di decesso, con l'introduzione di una norma che consente di disporre *post mortem* dei propri dati caricati nei servizi informativi delle società;
- viene data la possibilità (su autorizzazione dell'interessato) di comunicare i dati personali degli studenti universitari, per favorirne l'inserimento nel mondo del lavoro, la formazione e l'orientamento professionale;
- come forma di tutela, viene introdotto il reclamo, alternativo al ricorso in tribunale.

CAPITOLO 2 BIG DATA APPLICATI ALLA GESTIONE DELLE AZIENDE

Il concetto di Big Data ed il loro efficiente impiego nell'ambito aziendale hanno modificato profondamente i concetti di prodotti e servizi, della filiera produttiva necessaria alla loro produzione e somministrazione. L'uso dei *Big Data* nel *management* si può denotare in quattro ambiti costituenti un'impresa: macchinari, prodotti, logistica e cultura manageriale.

2.1 I BIG DATA APPLICATI AI MACCHINARI

Nell'attuale ambiente competitivo, le società dimostrano sempre più interesse alle problematiche riguardanti la gestione dei dati con lo scopo di aumentare la rapidità del *decision-making* per migliorare la produttività (Lee, Kao, & Yang, 2014). La Germania è *leader* nella trasformazione delle proprie imprese verso la rivoluzione industriale di quarta generazione, basata sull'innovazione della produzione e dei servizi, in particolare i servizi abilitati al sistema ciberfisico. Le "*Smart Industries*" si concentrano maggiormente sull'ottimizzazione del controllo centrale e dell'*intelligence*, raggiungendo così una fluente interazione con i sistemi circostanti trasformando, mediante l'utilizzo di tecniche di *machine learning*, i normali macchinari in sistemi autoregolativi migliorano tutte le prestazioni e la gestione della manutenzione. La completa implementazione dei sistemi di *machine learning* richiede ulteriori progressi nella scienza per affrontare diverse questioni fondamentali. I problemi possono dividersi in cinque categorie distinte:

- **Interazioni con dirigenti e operatori:** gli operatori controllano le macchine, che eseguono i compiti assegnati mentre i dirigenti designano la logistica. Sebbene questi compiti siano spesso ottimizzati dall'esperienza degli operatori e dirigenti, una parte significativa non viene considerata: la condizione fisica delle componenti delle macchine.
- **Insiemi di macchine:** è frequente che macchine simili o identiche vengano usate per svolgere diverse mansioni in diverse condizioni di lavoro. In contrapposizione, i metodi predittivi sono progettati per supportare un singolo o limitato numero di macchinari e condizioni di lavoro. Questi metodi e quelli di gestione della condizione dei macchinari disponibili non traggono vantaggio dal considerare queste macchine identiche come un insieme, raccogliendo conoscenze utili da casi diversi.
- **Qualità del processo e del prodotto:** la qualità del prodotto può fornire un'ulteriore informazione delle condizioni delle macchine attraverso algoritmi

di *backward induction*, inoltre può fornire *feedback* per sistemi di *management* che possono migliorare la gestione della produzione.

- **Big data e Cloud:** la gestione dei dati e la distribuzione nell'ambiente dei *Big data* è fondamentale per il raggiungimento di una perfetta implementazione del *machine learning*. L'importanza di riuscire a sfruttare la flessibilità e le capacità aggiuntive fornite dal *cloud* è chiara, ma l'adattamento di algoritmi predittivi e di gestione della condizione dei macchinari richiede ulteriori sviluppi al fine di ottenere un'implementazione efficiente delle attuali tecnologie di gestione dei dati.
- **Rete di sensori per il controllo**: i sensori sono i *gateway* delle macchine per percepire l'ambiente fisico circostante. Ad ogni modo, i guasti dei sensori e il loro degrado potrebbero far circolare informazioni errate o inaccurate per gli algoritmi di *decision-making.*

All'interno del concetto di industria 4.0 risulta una sorprendente crescita nell'avanzamento e adozione della tecnologia dell'informazione e della rete dei social media, divenuta sempre più influente riguardo la percezione dei consumatori rispetto all'innovazione del prodotto, della qualità e della varietà e velocità di spedizione. Questo richiede che la fabbrica sia dotata di meccanismi di *machine learning*; parallelamente a questa nuova tecnologia, due tipi di sviluppo innovativo stanno ricevendo maggior attenzione da parte del mondo accademico e delle industrie: innovazione dei servizi e *Big Data* industriali.

2.1.1 Sistemi di machine learning e di auto-manutenzione per ambienti industriali

Il recente sviluppo della struttura dell'*Internet of Things* e l'emergere della tecnologia di rilevamento hanno creato una griglia di informazioni unificate che connette sistemi umani e tecnologici. Con analisi specifiche, l'avvento del *cloud computing* e della struttura del sistema ciberfisico, le industrie future saranno capaci di raggiungere un sistema informativo che aiuti gli insiemi di macchine ad essere consapevoli e manutentori di sé stessi e attivamente pronti ad affrontare potenziali problemi di produzione.

Un sistema auto-manutentore è definito (Lee, Kao, & Yang, 2014) come un sistema che possa auto-valutare la propria condizione e deperimento, e oltretutto usare informazioni provenienti da altre macchine simili per prendere decisioni intelligenti riguardo alla manutenzione, al fine di evitare problemi potenziali. Un'analitica intelligente per ottenere un determinato livello di emergenza verrà usato a livello di macchine singole o di insiemi di macchine.

Per sistemi meccanici, auto-consapevolezza significa essere in grado di stimare e valutare l'attuale o la passata condizione delle macchine e reagire di conseguenza al risultato della valutazione. Questa stima può essere generata usando un algoritmo *data-driven* per analizzare dati ed informazioni raccolte da macchine apposite alla generazione di essi o dal proprio ambiente. Le condizioni dei macchinari possono essere recensite in tempo reale e spedite al controllore specializzato per un controllo adattivo ed ai responsabili della macchina per la manutenzione repentina. Inoltre, per la maggior parte delle applicazioni industriali, specialmente per le flotte di macchinari, la autoconsapevolezza delle macchine è ancora molto lontana dalla realizzazione. Recenti algoritmi di diagnosi o pronostici sono di norma utilizzati per determinate macchine o impieghi ma non sono molto flessibili per gestire informazioni più complesse.

La ragione per la quale l'auto-consapevolezza non sia stata ancora del tutto realizzata è da riassumere come segue:

- **Mancanza di un'interazione strettamente connessa tra uomo-macchina:** il fattore che influenza maggiormente le condizioni delle macchine e la loro produzione è la gestione e il funzionamento dell'uomo. La produttività e la qualità della produzione possono essere influenzate dalla progettazione e dalla pianificazione delle attività delle attrezzature; quest'ultime possono solo eseguire passivamente i comandi dell'operatore, anche quando il compito assegnato non è ottimale per le sue condizioni correnti. Un sistema intelligente di macchinari, d'altra parte, dovrebbe essere in grado di suggerire attivamente attività ed aggiustare i parametri operativi al fine di massimizzare la produttività e la qualità dei prodotti.
- **Mancanza di apprendimento adattivo e intero utilizzo delle informazioni disponibili:** i sistemi PHM[7] non possono essere ampiamente implementati nelle industrie a causa del loro scarso livello di adattamento, che porta ad una mancanza di robustezza negli algoritmi di monitoraggio della salute dei macchinari. Il problema dello sviluppo e l'implementazione sono solitamente separati nel sistema PHM: l'algoritmo PHM è sviluppato da esperimenti di raccolta dati ed esso non cambia durante la realizzazione a meno che non venga aggiornato da esperti. In molteplici casi, l'algoritmo gestisce solo i dati di monitoraggio delle condizioni dal reale utilizzo dei macchinari con procedure predefinite senza lo stimolo di imparare da esse. Questa situazione è molto distante dall'idea ottimale, poiché i dati raccolti empiricamente dalle macchine,

[7] Prognostic and Health Managment: è un framework che offre soluzioni individuali e personalizzati per la gestione della salute dei macchinari.

provengono da molteplici unità e si riferiscono ad un maggior lasso di tempo, il che significa che contengono molte più informazioni rispetto ai dati generati in laboratorio.

Gli algoritmi capaci di apprendere da questi dati saranno in grado di raggiungere un'ottima flessibilità e robustezza al fine di gestire diverse situazioni.
Risolvendo i problemi sopracitati, un sistema di macchinari CPS[8] unificato per consapevolezza e l'auto-mantenimento è stato sviluppato per poter estrarre efficientemente informazioni rilevanti dai *Big Data*, e inoltre coadiuvare il processo di *decision-making*.

2.1.2 Analisi dello status delle macchine con base di auto-apprendimento

A differenza della maggior parte dei CPS esistenti, che sono orientati al controllo o alla simulazione, il CPS proposto usa una conoscenza di base e un algoritmo relativo per rappresentare l'usura delle macchine e il comportamento di produzione materiale. Utilizzando l'apprendimento adattivo e gli algoritmi di *data mining*, una conoscenza di base rappresentata dalla performance della macchina e il suo stato di usura possono essere generati automaticamente. La conoscenza di base sarà capace di crescere proporzionalmente coi dati e con l'eventuale sviluppo della capacità di rappresentare fedelmente le condizioni di lavoro delle macchine reali. Con esempi di dati e le informazioni associate e raccolte dalle macchine, saranno confrontate sia orizzontalmente (macchina a macchina) sia verticalmente (di volta in volta) utilizzando appositi algoritmi sviluppati per l'estrazione delle conoscenze. A causa della completezza delle conoscenze di base, gli algoritmi PHM possono essere molto più flessibili nella gestione di eventi non previsti e forniranno risultati PHM più accurati.

I macchinari che svolgono compiti, in tempi simili possono avere produzioni e condizioni di salute analoghe. Basandosi su queste similitudini, i gruppi di macchine possono essere costruite come una base di conoscenza che rappresenti le diverse prestazioni della macchina e le condizioni di lavoro.
Algoritmi di apprendimento senza supervisione come il *Self-Organizing Map* ed il modello *Gaussian Mixture* possono essere utilizzati per creare automaticamente cluster per diversi regimi di lavoro e condizioni della macchina. Una ricerca di simili *cluster* può terminare con due risultati:

[8] Un sistema ciberfisico (Cyber-Physical System) è un sistema informatico in grado di interagire in modo continuo con il sistema fisico con il quale opera. Il sistema è composto da elementi fisici dotati ciascuno di capacità computazionale e riunisce strettamente le cosiddette "tre C": capacità computazionale, comunicazione e capacità di controllo.

- **Si trovano cluster simili:** la macchina dalla quale è stato raccolto il campione verrà etichettata come dotata della condizione di salute definita dal gruppo identificato. Nel frattempo, a seconda della deviazione tra il *cluster* esistente e l'ultimo campione, l'algoritmo aggiornerà il *cluster* esistente utilizzando nuove informazioni prese dall'ultimo campione.
- **Non si trovano cluster simili:** l'algoritmo manterrà la sua operazione di campionamento corrente fino a quando non otterrà un numero sufficiente di campioni fuori dal *cluster*. Quando il numero di campioni supera una certa quantità, significa che esiste un nuovo comportamento della macchina che non è ancora stato modellato e l'algoritmo crea automaticamente un nuovo *cluster* per rappresentare tale comportamento. In questo caso, l'algoritmo di *clustering* può essere molto adattivo alle nuove condizioni. Inoltre, il *cluster* sarà utilizzato come conoscenza di base per la valutazione della salute nello spazio cibernetico proposto. Con tale meccanismo, è possibile accumulare diversi comportamenti di prestazione della macchina nella conoscenza di base e utilizzarli per una futura valutazione della salute.

2.1.3 Vantaggi della struttura CPS

L'innovazione chiave di queste strutture del CPS riguarda la realizzazione di un sistema di autoconsapevolezza e manutenzione, integrando sia i dati dei sensori che le informazioni dell'insieme di macchine; in questo modo il volume dei dati può essere ridotto e simili modelli possono essere mappati. Il vantaggio chiave di questa struttura può essere riassunto nei seguenti punti:

- **Struttura unificata di CPS dal modello di salute macchina-macchina:** il CPS non è specifico per una macchina, ma per un insieme di macchinari e operatori umani; permette alle macchine di raccogliere informazioni dai propri pari, da operatori umani e da altri ambienti cosicché le macchine possano raggiungere l'auto-consapevolezza della loro condizione di salute attraverso la comparazione e l'apprendimento dal passato dei macchinari simili.
- **Abilitare gli algoritmi intelligenti di consapevolezza e auto-manutenzione usando l'algoritmo di auto-apprendimento PHM:** Algoritmi adattivi abilitano anche il sistema di apprendimento dai dati che sono stati accumulati mediante l'utilizzo; rigidità e inabilità di utilizzo rispetto ad eventi mai accaduti sono i maggiori ostacoli che prevengono gli attuali algoritmi di PHM dall'essere ampiamente implementati nell'industria.

- **Il sistema di supporto decisionale intelligente per una proattiva pianificazione della manutenzione:** attraverso la connessione dei macchinari e la loro auto-consapevolezza delle condizioni, i piani di manutenzione saranno organizzati ed ottimizzati a partire dal livello di sistema. La produzione e le *performance* dei macchinari possono essere ottimizzate mediante il bilanciamento e la compensazione del carico di lavoro basandosi sulla condizione di salute e dello stress di ogni macchina.

Il sistema di monitoraggio predittivo è una tendenza della produzione intelligente *data-based*. Con l'avvento della quarta rivoluzione industriale emergono quattro aree di forte impatto:

- La predizione di salute delle macchine riduce il tempo di inattività della macchina, e le informazioni pronosticate supporteranno il sistema ERP9 per ottimizzare la gestione della produzione, la programmazione della manutenzione e garantire la sicurezza dei macchinari;
- Il flusso di informazioni attraverso la linea di produzione, il livello dirigenziale delle aziende, la gestione della catena di distribuzione crea una gestione aziendale più trasparente e organizzata;
- Il nuovo trend industriale ridurrà il costo del lavoro e fornirà un ambiente di lavoro migliore;
- Ridurrà il costo del risparmio energetico, la programmazione della manutenzione ottimizzata e la gestione della catena di distribuzione.

2.2 BIG DATA APPLICATI AI PRODOTTI

Per comprendere le modalità in cui i *Big Data* siano presenti, migliorando i processi produttivi e il prodotto stesso, all'interno del ciclo di vita di un prodotto è necessario dividere quest'ultimo in tre fasi differenti (Lee, Thao, Ying, & Zhao, 2015): iniziale, fase dell'utilizzo e fase finale.[10]

[9] Enterprise resource planning: è un software di gestione che integra tutti i processi di business rilevanti di un'azienda

[10] Ci si riferirà a queste 3 fasi con le abbreviazioni: BOL (Beginning of life), MOL (Middle of Life) e EOL (End of Life)

2.2.1 Fase BOL

Durante questa fase le operazioni più salienti sono il design e la produzione; per fornire una base di dati agli esperti di design si necessita di un'approfondita analisi del mercato di riferimento, al fine di individuarne le caratteristiche e cercare di soddisfarle nella fase di *concept* del prodotto.

Analisi di mercato

L'analisi di mercato può essere divisa in due parti: la prima, trovare quali sono i consumatori che compongono il target obiettivo e successivamente individuarne i bisogni primari, data l'eterogeneità della domanda e dei desideri personali. Nella prima fase di analisi perviene il concetto di *Big Data*, che rende possibile l'individuazione di clienti con bisogni simili, partendo da una vastità di potenziali consumatori; esistono 3 tipi di dati che possono contribuire a determinare un target di consumatori: dati storici riguardanti acquisti simili effettuati dal consumatore e le sue preferenze su come investire le proprie finanze, reperibili nella fase di vendita; dati su ricerche di mercato provenienti da questionari somministrati ai consumatori; la cronologia dei siti usualmente frequentati e utilizzati da una persona.

Product design

Il design del prodotto è un processo iterativo e complesso: inizia con l'identificazione di particolari bisogni dei clienti che vengono successivamente processati lungo una serie di attività per cercare una soluzione ottimale al problema; questo processo è diviso in 3 fasi: specificazione del prodotto, design preliminare e design dettagliato.

Le decisioni riguardanti il design hanno un forte impatto sui costi, sulle performance, sull'affidabilità, sulla sicurezza e sull'impatto ambientale del prodotto. I requisiti e le caratteristiche del design durante i primi stadi del ciclo di vita sono spesso imprecise ed approssimate. Con lo scopo di presentare specifiche soluzioni che incontrino le necessità dei consumatori, sono stati creati vari strumenti che utilizzano le tecniche di interpretazione dei dati web. La condivisione e il riutilizzo di conoscenze di design e informazioni di prodotti mediante piattaforme specialistiche, permette di creare una base di conoscenze su cui operare uno sviluppo di prodotto, migliorando la conoscenza manageriale di design per ricercare una crescente modularità. Implementando questi strumenti via Web per migliorare il supporto in fase di design, permette agli sviluppatori di creare un ambiente in cui possano tradurre efficacemente le loro idee su come soddisfare una molteplicità di desideri di mercato in prodotti.

Il design dettagliato è l'ultima attività della fase di progettazione prima che la vera produzione inizi; in relazione al codice sorgente il design dettagliato è ancora astratto ma

dovrebbe essere abbastanza specifico da permettere una traduzione pratica oggettiva e non un'interpretazione di essa.

Produzione

Nella fase di approvvigionamento, il compito primario è quello di scegliere fornitori qualificati, basandosi sulla qualità dei prodotti, sulla loro reputazione e sul prezzo dei fattori produttivi. Nell'economia globale molti dati riguardanti i fornitori e le loro *performance* sono reperibili: una combinazione del TOPSIS[11] con un processamento gerarchico sono utilizzati per scegliere senza perfetta informazione.

I sistemi di produzione futuri necessiteranno di un ammontare di dati e di complessi processi per fornire ai lavoratori un adeguato livello di performance dei macchinari a causa di una crescente domanda di integrazione verticale con più alti livelli dei sistemi. Il *testing* del prodotto è una fase indispensabile della produzione, specialmente per prodotti complessi che vengono assemblati da vari componenti e un singolo difetto può risultare fatale per l'intero funzionamento del prodotto finale.

Le tecniche dei *Big Data* sono state gradualmente utilizzate per la stima dello stato di usura, grazie alla loro grande capacità di acquisire, integrare, trasportare, processare ed analizzare in tempo reale dati dinamici provenienti da un numero sempre in aumento di sensori. L'intero ciclo di produzione non include solamente il *flow* dei materiali lungo il processo ma anche l'energia utilizzata in esso: esistono due metodi principali per ridurre il consumo dell'energia dei macchinari; primo, migliorare le configurazioni dei macchinari e secondo, predisporre metodi per la riduzione dei consumi. Mediante l'utilizzo di tecniche per la gestione dati, l'ottimizzazione delle strumentazioni e delle procedure possano essere più facilmente raggiunte: ottenendo dati in tempo reale sui consumi si possono prendere decisioni riguardanti il risparmio energetico basandosi su maggiori informazioni.

2.2.2 Fase MOL

Nella fase di utilizzo di un prodotto può apparire che la fabbrica produttrice non svolga alcun ruolo attivo per il cliente, ma non è così: con la rilevazione e l'interpretazione dei dati prodotti in tempo reale del prodotto può essere implementata una manutenzione più efficace; nei prodotti non elettronici l'uso dei *Big Data* aiuta le imprese a ottimizzare la gestione del magazzino e l'organizzazione delle spedizioni.

[11] The Technique for Order of Preference by Similarity to Ideal Solution è un metodo aggregazione che considera i pesi relativi dei criteri di scelta tra un set di alternative; dopo aver effettuato la normalizzazione dei punteggi assegnati ai vari criteri, la scelta ricadrà sull'alternativa che avrà la minor distanza complessiva dei valori dei propri criteri rispetto all'alternativa ideale.

Ottimizzazione del magazzino e processamento degli ordini

La gestione dell'archiviazione dei dati è una parte vitale della logistica dal momento che le operazioni di gestione possono garantire un'efficiente circolazione dei beni dalle fabbriche ai venditori. Di pari passo con l'espansione delle aree di *trading*, gli archivi dati sono diventati complesse strutture che possono individuare relazioni tra fattori a livello globale. Per ottenere risposte alle sfide portate dal commercio globale, la gestione dell'inventario dovrebbe essere migliorata portandola ad un livello di maggior intelligenza: bilanciando il bisogno di disponibilità del prodotto e la necessità di minimizzare i costi di magazzino, dovrebbe essere costruito un inventario intelligente per identificare richieste, fornire più accuratamente i rifornimenti e tracciare lo status di un prodotto.

Per le imprese i cui modelli operativi sono *pull-production*, il processamento degli ordini può essere effettuato prima della effettiva produzione; in questi casi, le aspettative dei consumatori hanno subito una crescita e il volume di informazioni ad esse relative ed il loro processamento è un problema che può essere risolto solamente tramite l'utilizzo di tecniche di gestione di *Big Data*.

Training e manutenzione

L'assistenza è una forma di servizio che viene venduta indipendentemente o parallelamente al prodotto in sé, le tecniche di gestione dati dovrebbero essere introdotte nel processo di assistenza clienti come un metodo efficace per aumentare la soddisfazione del cliente, quando il prodotto viene utilizzato. È possibile che i prodotti talvolta non rendano alla perfezione a causa di comportamenti non corretti dei consumatori: è necessaria un'educazione del cliente all'uso del prodotto quando questo è sofisticato o fragile. Il problema dell'educazione del cliente è divenuto di tipo organizzativo, come nelle scuole: edifici per l'insegnamento, attrezzatura e programmazione dei corsi producono un grande volume di dati che dev'essere analizzato per prendere le decisioni più correttamente.

Con lo sviluppo del RFID[12] si è resa possibile l'instaurazione di una cooperazione tra casa produttrice e il cliente: i prodotti sono dotati di sensori che forniscono al produttore informazioni sullo status del prodotto; questo permette la fornitura di assistenza attraverso un monitoraggio attivo dello stato dell'attrezzatura ispezionando regolarmente il prodotto attraverso segnali remoti.

[12] Radio-frequency identification: si intende una tecnologia per l'identificazione automatica di informazioni inerenti ad oggetti, animali o persone basata sulla capacità di memorizzazione di dati da parte di particolari etichette elettroniche, chiamate e sulla capacità di queste di rispondere all'interrogazione a distanza da parte di appositi apparati fissi o portatili, chiamati reader.

Le tecnologie RFID e l*'Internet of Things* applicato agli impianti produttivi hanno reso possibile tracciare i prodotti dalla nascita fino allo smaltimento. Per questo, l'implementazione dei *Big Data* può fornire opportunità per una nuova generazione di manutenzione come quelle di tipo preventivo e previsionale che, a differenza di quella di tipo correttivo, intraprendono azioni prima dell'effettivo manifestarsi di un problema. La manutenzione si è gradualmente sviluppata in questa direzione e i confini tra utilità e manutenzione non sono più chiaramente delineati, in quanto un certo grado di assistenza è fornita durante l'uso del prodotto, tuttavia come ideare e progettare l'assistenza di tipo previsionale è ancora un compito gravoso, dato l'ammontare di dati da processare.

2.2.3 Fase EOL

Con *l'upgrade* dell'alta velocità di produzione e consegna dei prodotti, specialmente nell'elettronica e nella *e-commerce*, un *focus* maggiore è stato posto nello spreco da loro generato, per ridurre l'impatto negativo di questi verso l'ambiente nella fase finale del ciclo di vita. Dopo aver ricavato il tempo residuo delle varie componenti alla fine del ciclo di vita, gli addetti allo smaltimento possono ottimizzare il recupero del prodotto con l'obiettivo di massimizzarne il valore. In questa fase, le tecniche di gestione *Big Data* possono essere applicate alle varie opzioni quali riciclo del prodotto, rigenerazione e il loro smantellamento.

Uno degli scopi più importanti del riciclaggio è di ridurre gli scarti di produzione dispersi nell'ambiente: è necessario assicurare che il processo di riciclaggio sia esso stesso a basso consumo di energia. Per raggiungere questo scopo tecniche di gestione dei *Big Data* dovrebbero essere applicate per creare sistemi intelligenti di supporto alle decisioni per garantire attività di riciclaggio che minimizzino l'impatto ambientale e il consumo di risorse durante la fase di riciclo; come soluzione sono state proposte tecniche di ricerca senza *backtracking*, *query* efficienti e flessibili per l'interrogazione dei *Database* applicate a dei DSS intelligenti per garantire la sincronizzazione nelle operazioni di riciclaggio.

2.3 BIG DATA APPLICATI ALLA LOGISTICA

Innovazione e tempistiche sono i maggiori vantaggi competitivi per aziende il cui *core business* è rappresentato dalla logistica (Witkowski, 2017). Una conseguenza del boom del mercato globalizzato ha portato al cambiamento dell'organizzazione delle imprese: l'elemento più importante appare essere l'accorciamento del ciclo di vita del prodotto. La prima generazione di Volskwagen Golf fu prodotta dal 1974 al 1983; l'ultima generazione è stata prodotta dal 2008 al 2012. In queste competizioni per le acquisizioni di più clienti, le aziende

sono viste come offerenti di nuovi prodotti, ma con un livello minore di qualità rispetto alle versioni precedenti, un fenomeno che si può notare soprattutto nel mercato dei beni di consumo.

L'innovazione nella logistica non è solamente associata al coinvolgimento delle moderne soluzioni informatiche ma un segno di modernità può essere trovato nel modo di pensare. Soluzioni innovative nella logistica possono manifestarsi in:

- Continuo miglioramento del team che realizza l'innovazione e la continua verifica del lavoro ed impegno;
- Vigilanza costante sulle qualità delle attività;
- *Focus* costante nel lavoro del team che lavora sulle pratiche da implementare e nei valori condivisi;
- Attività che includono la costante ricerca di nuovi e migliori modi di perfezionamento dei compiti della logistica;
- Soddisfazione del consumatore con il lavoro e l'onestà, eliminazione dell'inerzia decisionale, dei comportamenti e delle barriere associate al cambiamento delle attività nell'area della logistica.

I più importanti *driver* di innovazione (Witkowski, 2017), che in qualche modo a creano valore aggiunto nella logistica, sono le risorse umane e la cultura organizzativa.
Oltre a citare le due innovazioni rivoluzionarie nella logistica, vale a dire il container, il quale ha totalmente rivoluzionato il flusso dei materiali e la tecnologia RFID, che ha contribuito alla trasparenza della catena di distribuzione, si menzionano altri fattori chiave di successo della logistica:

- la struttura di una rete regionale, flessibilità, gestione dei rischi e rotazione dei mezzi;
- richieste maggiori da parte dei consumatori in termini di gestione del tempo della spedizione, della loro disponibilità ed affidabilità;
- servizio predisposto in accordo con i bisogni dei consumatori, perciò una risposta rapida ai loro bisogni;
- segmentazione della catena di distribuzione concentrata sulla domanda e sugli specifici bisogni dei consumatori, ciò aiuta a ridurre il volume dello *stock* e di ottimizzare i costi;
- requisiti di sicurezza e potenziali pericoli nella catena di approvvigionamento;
- gestione dei rischi nella catena di distribuzione;
- strategie per uno sviluppo sostenibile delle imprese riguardo l'aspetto ambientale.

Tutti i fattori e tendenze sopracitati dovrebbero riflettersi in soluzioni innovative della logistica. Grazie alle economie di scala, saranno capaci di condurre a soluzioni che

permetteranno agli operatori logistici di incontrare le necessità dei consumatori nel 21esimo secolo; tuttavia, c'è un'entità che non possono essere trascurate e sulla quale si può compiere analisi e, ovvero i *Big Data*, che permettono una gestione veloce ed efficiente ed un uso costante dei *Database* (grazie alla capacità di raccogliere informazioni da risorse differenti).

Un esempio dell'uso della tecnologia dei *Big Data* nell'area logistica è fornito dalla DHL, con il loro utilizzo nel miglioramento del "Resilience360", uno strumento progettato per gestire i rischi nella catena di distribuzione. L'azienda può fornire ai consumatori informazioni sulle potenziali interferenze delle rispettive catene di approvvigionamento. Questo avviene grazie alla raccolta e valutazione dei dati che rendono possibile, non solo la protezione dell'azienda stessa (qualora un pacco non arrivasse in tempo l'azienda subirebbe una perdita in termini di immagine), ma anche il miglioramento dell'efficienza nella catena di distribuzione; non esiste un'interruzione nelle operazioni ed è possibile raggiungere in modo costante la soddisfazione dei clienti.

La DHL dimostra che l'uso dell'analisi *Big Data* aumenta l'efficienza operativa, mentre fornisce l'opportunità di esplorare nuovi modelli aziendali. "DHLresilience360" contiene due elementi che sono associati all'analisi della valutazione del rischio, tanto quanto strumenti di monitoraggio della catena di distribuzione che lavorano in tempo reale. La forza della catena e le correlate perdite di ricavi dipendono solamente da eventuali *break* nella produzione. La DHL è nella fase pilota del modello "The forecast number of packages DHL", che è considerato come l'implementazione effettiva con l'analisi dei *Big Data*. Questo modello semplifica il volume pianificato del trasporto dei pacchi, prendendo in considerazione fattori di dati correlati. Un altro modello "DHL Geovista" permettere analisi dettagliate e valutazioni di dati geografici complessi da ottenere, che facilitano notevolmente il compito ai fornitori dei servizi logistici di anticipare la molteplicità delle vendite generate dalle piccole e medie imprese.

I *Big Data* permettono ai fornitori di servizi di ottimizzare i processi di logistica, migliorare il servizio ai clienti e presentare un inizio promettente per lo sviluppo di nuovi modelli di business. Tra le risorse del sistema di approvvigionamento, vi sono informazioni provenienti da rivenditori, trasporti, fatture, dati da profili di clienti, profili di *social network*, ordini, previsioni di mercato e schemi che non possono essere tralasciate se si vuole perseguire un continuo perfezionamento del servizio. Utilizzando i dati dei consumatori per analizzare le informazioni del sistema di consegna, i rivenditori possono soddisfare le aspettative dei clienti anticipando i loro comportamenti, passando in questo modo da una concezione di logistica passiva ad una di logistica proattiva.

2.4 BIG DATA COME BASE PER UNA NUOVA CULTURA MANAGERIALE

"I *Big Data* non sono solo un altro modo per dire dati analizzati?"
Con questa frase si esprime il dubbio di un'intera classe dirigenziale riguardo all'utilizzo dei dati all'interno delle aziende come base effettiva su cui operare decisioni gestionali. Uno studio effettuato dall'Harvard Business School in collaborazione con il MIT e col McKinsey Institute (Mcafee & Brynjolfsson, 2012) evidenzia come più le aziende si definiscano come *data-driven*, più pervengano a misure oggettive di risultati finanziari ed operativi. In particolare, si evidenzia come le aziende facenti parte del quartile più redditizio nel loro settore, con l'utilizzo di decisioni basate sui dati, abbiano avuto, in media, il 5% in più di produttività e il 6% in più di profitti rispetto ai loro concorrenti. Questa differenza di rendimento è rimasta significativa anche dopo la contabilizzazione dei contributi di manodopera, capitale, servizi acquistati e investimenti IT tradizionali. Due impieghi empirici per la risoluzione di questi dubbi manageriali sono pervenute dalla Silicon Valley: promozioni più veloci e personalizzate e una nuova cultura del *decision making*.

2.4.1 Promozioni più veloci e più personalizzate

Un paio d'anni fa, Sears Holdings[13] arrivò alla conclusione che c'era bisogno di creare valore dall'enorme quantità di dati relativi ai consumatori, dai prodotti e dalle promozioni ricavate dai brandi di Sears, Craftsman e Lands'End. Sarebbe stato prezioso, ma allo stesso tempo difficile, combinare e usare tutti questi dati per creare promozioni e offerte su misura per il cliente e personalizzare le offerte al fine di avere vantaggi rispetto alle condizioni locali: Sears utilizzava 8 settimane circa per creare promozioni personalizzate, al punto che molti di queste non erano più ottimali per l'azienda. Era impiegato così tanto tempo perché i dati richiesti per le analisi di larga scala erano altamente frammentati e voluminosi ed erano ospitati in molti *database* e *data warehouse* gestiti dai vari marchi.

Al fine di trovare una via più veloce ed economica per fare il suo lavoro analitico la Sears Holdings aveva adattato le tecnologie e le pratiche al sistema dei *Big Data*. Come uno dei suoi primi passi, la Sears aveva creato un cluster Hadoop, un semplice gruppo di server di merce poco costosa le cui attività erano coordinate da una struttura di software. La Sears iniziò ad usare il software per memorizzare i dati che ricevevano dai suoi brand e a tenere i dati in magazzini di dati esistenti. Ciò ha condotto analisi direttamente nei cluster, evitando la complessità dispendiosa in termini di tempo che avrebbero incontrato nell'estrazione di dati da

[13] In questo paragrafo si riferirà a Sears Holding nel passato, in quanto questa società ha dichiarato bancarotta nel 2018 ed è stata acquisita da ESL Investments; le metodologie proposte restano comunque valide.

varie fonti e la loro combinazione in modo che potessero essere analizzati. Questo cambiamento permise all'azienda di essere molto più veloce e precisa nelle promozioni. Secondo il CTO della società, Shelley, il tempo di cui si aveva bisogno per generate un set di promozioni comprensive calò dalle 8 settimane ad una settimana per ogni promozione ottenendo promozioni di qualità maggiore e maggiormente personalizzate. I cluster della Sears Hadoop archiviavano e processavano molti Petabytes di dati ad una frazione minore di costo comparata con le *data warehouse* standard.

2.4.2 Una nuova cultura di decision making

Uno degli aspetti maggiormente criticati dell'uso dei *Big Data* è il loro impatto nell'aspetto decisionale: quando i dati sono scarsi, costosi da ottenere o non disponibili in versione digitale, una decisione diviene sensata se operata da persone qualificate, le quali svolgono il loro compito in base ad esperienze; "intuizione" è l'etichetta che si è data allo stile di deduzione e *decision-making*.

Per decisioni particolarmente importanti, queste persone sono di alto livello nell'organizzazione oppure sono costosi consulenti introdotti a causa della loro esperienza e dei loro *track record*, in molti ripongono la loro fiducia nelle aziende che prendono la maggior parte delle loro decisioni facendo affidamento sull'HIPPO[14](Mcafee & Brynjolfsson, 2012).

I dirigenti interessati nella gestione delle transizioni dei *Big Data* possono iniziare con due semplici metodi. In primo luogo, possono fissare l'obiettivo chiedendosi cosa significhino i dati quando devono prendere una decisione importante e poi continuare con domande più specifiche come "Da dove arrivano i dati? Che tipo di analisi sono stata fatte?". Gli obiettivi prefissati possono essere respinti dai dati e poche cose sono più rilevanti per cambiare la cultura del *decision-making* rispetto a vedere un dirigente modificare le proprie decisioni sull'oggettiva analisi dei dati. Quando si ha bisogno di conoscere quali problemi affrontare l'esperienza del dominio rimane critica.

Le società non raccolgono tutti i benefici della transizione dall'uso dei *Big Data* se non sono capaci di gestire in modo efficiente il cambiamento. Le 5 aree rilevanti in questo processo di cambiamento culturale sono:

- *LEADERSHIP*. Le aziende primeggiano nell'era dei Big Data non perché abbiano maggiori quantità di dati ma perché hanno una squadra di leader che pone degli obiettivi chiari, definisce come sarà il successo e richiede delle proposte corrette. La potenza dei

[14] Acronimo di Highest Paid Person's Opinion ed è usato per esprimere la tendenza degli impiegati a rinviare ai superiori (aventi stipendi più alti) il compito di prendere decisioni.

Big Data non ha bisogno di intuizioni o visioni da parte dell'uomo, al contrario, si necessitano di leader societari che trovino le giuste opportunità aziendali e capiscano come si sviluppi il mercato.

- *TALENT MANAGEMENT*. I dati diventano più economici, la corrispondenza dati acquisisce maggior valore. Alcuni dei più importanti di questi sono gli esperti di dati e altri professionisti che lavorano con grandi quantità di informazioni. Le statistiche sono importanti, ma molte tecniche chiave nell'uso dei *Big Data* sono raramente insegnate nei corsi di statistica tradizionali. Forse, sono ancora più importanti le abilità nella pulizia e organizzazione di grandi *dataset*; i nuovi tipi di dati raramente arrivano già in formati strutturati. Gli strumenti di visualizzazione e le tecniche allegate sono motivo di aumento di valore.
- *TECNOLOGIA*. Gli strumenti disponibili per gestire il volume, la velocità, la quantità di *Big Data* sono migliorati notevolmente negli ultimi anni. In generale, queste tecnologie non sono troppo costose e molti software sono *open source*. Hadoop, il *framework* più usato, combina un hardware di base con un software *open source*. Esso prende stringe di dati in entrata e li distribuisce in economici dischi; inoltre fornisce strumenti per l'analisi dei dati.
- *DECISION MAKING*. Un'organizzazione efficace mette sullo stesso piano le informazioni e le corrette decisioni rilevanti. Nell'era dei *Big Data*, le informazioni sono create e trasferite e le conoscenze possono essere utilizzate da altre organizzazioni oltre a quella che ha diffuso per prima tali innovazioni. Il buon leader sa creare un'organizzazione flessibile per minimizzare la sindrome del "non inventato qui" e massimizzarne la cooperazione plurifunzionale.
- *COMPANY CULTURE*. La prima domanda che un'organizzazione *data-driven* si pone preliminarmente ad una decisione non è "Cosa pensiamo?" ma "Cosa conosciamo?". Questo richiede uno spostamento dall'agire in base alle sensazioni ed istinto. Richiede inoltre la rottura delle cattive abitudini che non sono mai state notate nelle società, pretendendo si essere molto più basate sui dati di quanto realmente lo fossero.

CAPITOLO 3 AMAZON

Amazon, la 28-esima azienda al mondo per fatturato e tra le prime al mondo come valore di capitalizzazione, trova le sue radici nella vendita online di libri e supporti elettronici per la loro lettura in formato digitale. Ad oggi, l'azienda vanta un valore di circa 916.1 miliardi[15], raggiunti grazie all'incontro di intuizioni geniali dei propri ricercatori ed alla volontà del proprio fondatore di osare ed innovare, Jeff Bezos.

Analizzando il bilancio di fine 2018, si intuisce come la maggior parte degli introiti non provengano dalla vendita diretta di prodotti ma da servizi che l'azienda può permettersi di offrire, come l'Amazon Web Service, la vendita di prodotti per conto terzi e l'Advertising.

3.1 AMAZON WEB SERVICE

Amazon Web Services (AWS) è una società interamente controllata da Amazon che gestisce la piattaforma *cloud* più completa e utilizzata al mondo.
La sua caratteristica distintiva è l'ampiezza dei servizi offerti, più di 165: una lista approssimativa delle funzionalità di base di questo strumento è offerta dalla stessa Amazon nel proprio sito web e include:

- Calcolo;
- *Storage*;
- *Database*;
- Migrazione e trasferimento di informazioni;
- Reti di distribuzioni per contenuti;
- Strumenti per sviluppatori di videogiochi;
- Gestione di governance per aziende;
- Servizi multimediali;
- *Machine learning* e *IoT* industriali;
- *Blockchain*.

3.1.1 Chi utilizza AWS?

Questo servizio, integrato di funzionalità riguardanti tutto ciò che può essere eseguito con un computer a scopo industriale e gestionale, è adatto per qualsiasi tipologia di impresa, sia essa UniCredit per gestire le informazioni dei propri clienti e per creare promozioni customizzate o una *start-up* innovativa che utilizza questa infrastruttura per reperire risorse finanziarie.

[15] Dato Forbes risalente al 17/5/19

Le funzionalità non si riducono all'applicazione dei propri algoritmi ma assicurano un alto livello di sicurezza dei dati che vengono immessi e processati. Basti pensare che la CIA, nel 2013, stipulò con Amazon un contratto da 600 milioni per l'utilizzo del *Cloud Computing*. AWS supporta 85 standard di sicurezza e certificati di conformità e tutti i 116 servizi AWS che archiviano i dati dei clienti offrono la possibilità di crittografare tali dati.

La gestione ed il processamento dei dati è considerabile come la nuova frontiera per le imprese, siano esse piccole o multinazionali, ed Amazon, operando con anni di anticipo, sviluppando algoritmi efficienti e creandosi enormi spazi di *storage*, è in grado di sopperire alle mancanze delle singole imprese con un prodotto unico, complesso e multiservizio.

Tra le aziende che utilizzano questa infrastruttura vanno citate la MLB, Netflix ed ENI.

3.1.2 Punti di forza dell'AWS

Il primo vantaggio di questo prodotto è che non esistono prodotti sostitutivi nel mercato: una soluzione di *cloud computing* che fornisca algoritmi così precisi e spazi di *storage* così ampi non è disponibile sul mercato, se non l'AWS.

Nel 2003 Benjamin Black, un esperto informatico, viene assunto in Amazon per cercare di risolvere un problema di coordinazione tra ingegneri: da qui prese vita un progetto per fornire una maggiore scalabilità. Nel 2006 venne rilasciato il primo servizio di AWS: EC2 (Elastic Compute Cloud). Quando Bezos comprese le potenzialità dello sviluppo di questi prodotti, cercò un modo per renderli alla portata di tutti, basandosi sul continuo miglioramento delle prestazioni e sul costante aumento dei servizi offerti alle imprese. Da qui deriva il secondo punto di forza: la velocità di innovazione. La società ha un modello organizzativo basato su team di prodotti molto snelli e autonomi, così da velocizzare i rilasci di funzionalità e servizi, sviluppati sempre partendo dalle esigenze dei clienti, in una modalità "*working backward*".

Il terzo vantaggio risiede nelle enormi risorse finanziarie disponibili per la ricerca e lo sviluppo di tale piattaforma e l'esperienza accumulata nella comprensione e nel soddisfacimento delle esigenze dei clienti. Nel 2018, AWS è diventato il primo provider di servizi *cloud* a offrire 100 Gbps di larghezza di banda di rete con i tipi di istanza EC2 C5n. AWS vanta esperienza, maturità, affidabilità, sicurezza e prestazioni che gli hanno permesso di distribuire servizi cloud per 12 anni a milioni di clienti in tutto il mondo per l'esecuzione di una vasta gamma di casi d'uso.

3.1.3 Risultati nel fatturato di AWS

Parallelamente all'implementazione a livello globale nei sistemi economici di strutture per l'analisi dei dati, essendo AWS il servizio di *Cloud Computing* più diffuso al mondo, sta conoscendo un periodo di crescita che ha portato, nel primo trimestre del 2019, a una crescita del 41% rispetto all'anno precedente con utili operativi pari a 2.2 miliardi, il che fa supporre agli analisti che, alla chiusura del 2019, i ricavi di AWS supereranno i ricavi del 2018 che si attestano a 25 miliardi di fatturato, con un utile del 30%.

3.2 VENDITA PER CONTO TERZI

I prodotti reperibili su Amazon non sono tutti acquistati e gestiti dall'impresa stessa: vengono offerte ai venditori le possibilità di usufruire della rete logistica internazionale, dello stoccaggio, dell'invio dei prodotti e della gestione dei resi.

3.2.1 Come funziona la vendita per conto terzi?

Una volta che il venditore decida di usufruire della struttura di Amazon, questo deve sottoscrivere un abbonamento mensile di €39 se si vendono più di 40 prodotti al mese, altrimenti l'abbonamento mensile è gratuito, ma viene applicata una franchigia di € 0.99 a prodotto venduto (account da venditore base o premium); in entrambi i casi, alla vendita di un prodotto, verranno applicate delle ritenute in base a dei criteri prestabiliti[16]. Un'ulteriore franchigia per le commissioni di rimborso è trattenuta qualora il prodotto venga rimborsato: a carico del venditore viene addebitato l'importo minore tra €5.00 ed il 20% delle commissioni sostenute da Amazon per il rimborso del cliente.

[16] Tabella disponibile all'indirizzo: <https://sellercentral.amazon.it/gp/help/external>

Per quanto concerne i costi di magazzino, calcolati in base all'unità di spazio volumetrico occupato dai prodotti, vengono riassunti nella seguente tabella:

	Categoria	Periodo	Prezzo
Dimensioni standard	Abbigliamento e scarpe	Da Gennaio a Settembre	€15.60
		Da Ottobre a Novembre	€21.60
	Tutte le altre categorie	Da Gennaio a Settembre	€26.00
		Da Ottobre a Novembre	€36.00
Fuori misura	Tutte le categorie	Da Gennaio a Settembre	€18.00
		Da Ottobre a Novembre	€25.00

Tabella 2: Tariffe applicate da Amazon per i costi di magazzino. Fonte: <https://services.amazon.it/servizi/logistica-di-amazon/tariffe.html >

3.2.2 Vantaggi offerti da Amazon

Affidando la propria logistica ad Amazon significa affidare i propri prodotti all'impresa di *e-commerce* migliore al mondo: negli anni, grazie alle superbe prestazioni nella logistica e agli accordi stipulati con le maggiori società di trasporto a livello internazionale, Amazon ha creato una *brand perception* nei consumatori positiva; essere loro partner commerciali è percepito come segno di professionalità e flessibilità data la velocità di spedizione e la loro politica su resi e rimborsi. Inoltre, un venditore si libera della fase di distribuzione affidandola a dei professionisti ed a una rete logistica globale; i prodotti acquisiranno ulteriore *appeal* per gli utenti Prime, in quanto saranno prontamente spediti e senza costi aggiuntivi.

Amazon offre, oltre che alla condivisione del valore del marchio, anche facilitazioni procedurali: in Italia, nel 2019, è entrato in vigore l'obbligo di fatturazione elettronica e questo ha creato non pochi problemi ai rivenditori; Amazon offre assistenza durante la fase di fatturazione elettronica includendo l'IVA ed offrendo opzioni di scorporazione dell'imposta per una facile operatività sul prezzo reale del prodotto.

3.2.3 Risultati nel fatturato della vendita per conto di terzi

Il dato relativo al fatturato della vendita per conto terzi va stimato, in quanto non è possibile scorporare i costi totali della logistica che Amazon sopporta per la vendita dei propri prodotti da quelli che sostiene per la vendita per conto di terzi. Si può, tuttavia, approssimando per difetto, rilevare una media di ritenuta del 6% nel prezzo di vendita globale per i terzi; sommando i valori degli abbonamenti dei vari *retailer*, si ottiene un valore approssimativo ma significativamente veritiero di 2 miliardi di fatturato. Questo valore è in forte crescita dato l'aumento continuo dei *retailer* che scelgono Amazon come partner commerciale in quanto, superata una certa soglia di vendita, i vantaggi per il singolo venditore sono economicamente sensibili.

3.3 AMAZON ADVERTISING

L'advertising di Amazon è la possibilità, offerta ai commercianti che utilizzano Amazon per la vendita dei propri prodotti, di comprare degli spazi pubblicitari al fine di sponsorizzare i propri prodotti o l'intero *brand*. La comparsa delle inserzioni non è casuale ma apparirà a video agli utenti che stanno acquistando o hanno storicamente acquistato prodotti complementari o sostitutivi al proprio; la suddetta sponsorizzazione avviene attraverso quella che è chiamata DPS (*Amazon Demand-Side Platform*) e permette agli algoritmi di Amazon di riproporre programmaticamente una certa inserzione a tutti gli utenti che potrebbero essere interessati.

3.3.1 Come funziona l'Advertising?

Il soggetto promotore sceglie quali prodotti vuole che siano promossi, inserendo parole chiave a cui questo è associato e ogni volta un utente farà una ricerca inerente a queste parole chiave la suddetta inserzione apparirà come "Prodotto sponsorizzato", e al click di un utente questo verrà reindirizzato alla scheda tecnica del prodotto.

Qualora si sia scelto di promuovere l'intero *brand*, al fine di creare *brand awareness*, l'utente verrà reindirizzato ad una *landing page* dedicata al *brand*, contenente informazioni del *brand* stesso e ai suoi prodotti sponsorizzati per un minimo di tre fino ad un massimo di cento. I prodotti sponsorizzati appariranno per primi come risultati delle ricerche di utenti e verranno riproposti nella sezione correlati al momento dell'acquisto.

La sponsorizzazione viene pagata seguendo il principio del *pay-per-click*: all'inizio della campagna pubblicitaria si scelgono i prodotti da pubblicizzare e si imposta un *budget* che determinerà per quanti *click*, da parte degli utenti, l'inserzione rimarrà attiva.

3.3.2 Limiti imposti ai venditori

La possibilità di pubblicizzare i prodotti su Amazon è stata recentemente tolta ai prodotti cosiddetti CRAP, acronimo di "*Can't Realize A Profit*": in sostanza Amazon sta modificando la propria politica aziendale, decidendo di concedere la possibilità di pubblicizzare i prodotti secondo criteri oggettivi, tra cui la redditività che la vendita di un prodotto può rappresentare per l'azienda stessa; non è più sufficiente pagare la quota della campagna pubblicitaria ma è necessario che un prodotto risulti redditizio[17] per Amazon. Questo viene comunicato attraverso una mail ai commercianti direttamente dall'azienda in cui vengono indicati quali prodotti non sono adatti alla sponsorizzazione; questa politica è stata dettata dalla volontà dirigenziale di aumentare il margine dei profitti e di tagliare i costi per le attività non produttive, come la chiusura dei *pop-up* store in favore dei negozi automatizzati Amazon Go.

3.3.3 Risultati nel fatturato dell'Advertising

Il fatturato generato dall'advertising è una voce in continua crescita nel bilancio: a fine 2018, Amazon ha fatturato solamente mediante sponsorizzazioni 10 miliardi di dollari e nelle stime future questa attività è destinata a superare il fatturato di AWS, avendo come obiettivo 16 miliardi di utile operativo entro la fine del 2021.

[17] se un prodotto è venduto a $15 e ad Amazon lo stoccaggio e la logistica costano $15 questo non può essere sponsorizzato

4 CONCLUSIONI

La scelta dell'analisi di Amazon non è stata casuale: il successo di una società leader a livello globale nel proprio settore risiede nell'utilizzo e l'implementazione di tecniche *data-based*, dimostrando come questa sia la naturale evoluzione delle imprese in ogni settore.

Il vantaggio competitivo raggiunto negli anni da Amazon, rispetto alle aziende concorrenti, risale al momento in cui, con anni di anticipo rispetto a tutti i *player* internazionali, si è compresa la potenzialità dell'analisi dei dati e dello sviluppo di algoritmi per il *computing*. In ogni ramo aziendale è presente un'applicazione pervasiva dell'analisi dei dati e l'utilizzo di algoritmi computazionali a fini economici: siano essi utilizzati per lo *storage* e l'analisi di dati generati internamente nella fase logistica oppure offerti come strumento ad imprese esterne per coadiuvare la loro operatività; questo elemento fa di Amazon la società che più di tutte, assieme ad Alphabet, ha modificato la geografia socio-economica mondiale accorciando idealmente le distanze tra mercati che fino al decennio scorso risultavano essere tra loro impermeabili, data la loro distanza non percorribile in tempi oggettivamente brevi.

Raggiunto questo livello di performance, Amazon continua a reinvestire in numerosi progetti, per migliorare prodotti o servizi già esistenti o per crearne di nuovi per rispondere ai bisogni latenti dei consumatori, al fine di conservare la propria posizione di leader nel mercato mondiale. Tra questi progetti si annoverano la costruzione di negozi Amazon Go: negozi senza commessi in cui i prodotti comprati verranno addebitati direttamente nel conto dell'acquirente al momento dell'uscita dal punto vendita, oppure l'implementazione di varie funzioni di Alexa, l'intelligenza artificiale sviluppata da Amazon come assistente personale. I progetti di Amazon non sono solamente legati al commercio: con la creazione della società Blue Origin si è interessata altresì di viaggi spaziali, ridefinendo i confini tra impresa economica e umanitaria; gli sviluppi e le nuove possibilità per l'uomo che porteranno queste innovazioni non possono che essere estremamente interessanti.[i]

[i] Numero parole:14997

RIFERIMENTI BIBLIOGRAFICI

ARBIA, G. (2018). *Statistica, nuovo empirismo e società nell'era dei Big Data.* Roma: Edizioni Nuova Cultura.

CADWALLADR, C., & GRAHAM-HARRISON, E. (2018, Marzo 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, p. [online].

CHEN, J., CHEN , Y., DU, X., LI, C., LU, J., ZHAO, S., & ZHOU, X. (2013). Big Data challenge: a data managment perspective. *frontiers computer science, 7 (2)*, 157-164.

GANDOMI, A., & HAIDER, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Managment, 35*, 137-144.

GEORGE, G., HAAS, R., & PENTLAND, A. (2014). Big Data and managment. *Academy of Managment Journal, 57*, 321-326.

LEE, J., KAO, H., & YANG, S. (2014). Service innovation and smart analytics for Industry 4.0 and big data enviroment. *Procedia CIRP, 16*, 3-8.

LI, J., TAO, F., YING, C., & ZHAO, L. (2015). Big Data in product lifecycle managment. *The International Journal of Advanced Manufacturing Technology, 81*, 667-684.

MCAFEE, A., & BRYNJOLFSSON, E. (2012). Big Data: The Managment Revolution. *Harvard Business Review, 5*, 4-9.

VALSANIA, M. (2018, Maggio 2). Cambridge Analytica travolta dal Datagate: bancarotta e chiusura immediata. *Il Sole 24 Ore*, p. [online].

WITKOWSKI, K. (2017). Internet of Things, Big Data; Industry 4.0- Innovative Solutions in Logistics and Supply Chain Managment. *Procedia engineering, 182*, 763-769.