BIG DATA

BIG DATA E LE 5 V

Da una recente indagine si presume che ad oggi vi siano 3 miliardi di dispositivi connessi alla rete in grado di generare una quantità di dati in ascesa esponenziale.

Giusto per esempio, ogni minuto sono spedite circa 200.000 email, si inviano circa 300.000 tweet, vengono caricate centinaia di ore di video su YouTube. Solo su Facebook vengono caricati contenuti per un totale di oltre 105 terabyte (TB) ogni 30 minuti. In aggiunta, le app dei dispositivi mobili tracciano ed analizzano in ogni momento ogni aspetto della nostra attività quotidiana. Infine, una asta gamma di sensori, preseti nei più disparati oggetti, sono in grado di generare ogni giorno enormi quantità di informazioni utili per differenti contesti applicativi (es. monitoraggio, sorveglianza, controllo prestazioni, ecc.).

Viene utilizzato il termine Big Data per descrivere collezioni di dati, strutturati (come quelli presenti in basi di dati) e non (e-mail, dati multimediali, dati prodotti da Social Network e generati da sensori, log di sistemi, ecc.) per cui i metodi convenzionali di gestione ed analisi non risultano essere più adeguati.

Il modello di base utile a descrivere le caratteristiche dei Big Data è sicuramente quello delle cosiddette 3 V introdotto dall'analista Doug Laney nel 2001. Esso prevede che una collezione di dati, per essere considerata "big", deve possedere le seguenti proprietà:

• Volume, si riferisce al fatto che la dimensione effettiva del dataset deve essere molto grande;

•Velocità, fa riferimento all'elevata velocità di generazione dei dati ed al fatto che l'analisi dei dati necessita di essere effettuata a "ritmo sostenuto", se non addirittura in alcuni casi, in tempo reale;

• Varietà, riferita alla diversa tipologia di dati, provenienti da fonti eterogenee, ed alla necessità di esplorare anche dati non strutturati, oltre all'insieme delle informazioni tradizionali.

In aggiunta alle 3 V, sono state introdotte altre due dimensioni, Veridicità (cioè la qualità e l'accuratezza dei dati, legata al livello di "rumore" statistico che contengono e ne determina l'affidabilità) e Valore (è il potenziale beneficio a cui mira l'analisi dei dati), per il monitoraggio della qualità dei dati, enfatizzando il valore aggiunto al business derivante dall'analisi dei dati stessi.

L'idea alla base dei Big Data va ritrovata dunque nella necessità di analizzare, contemporaneamente, collezioni molto grandi di dati correlati per ricavare informazioni aggiuntive rispetto all'analisi separata di piccoli insiemi di dati. Questo permette, ad esempio, l'analisi degli "umori" dei mercati, analisi strategiche delle aziende ed altre tipologie di analisi che coinvolgono ingenti quantitativi di dati.

Le sorgenti più note sono le seguenti:

• Enterprise Data: Insieme dei dati condivisi all'interno di un'organizzazione (es. email);

• Public Data: Dati aperti al pubblico senza alcuna misura di protezione;

• Sensor Data: Insieme dei dati catturati dai sensori (es. satelliti, GPS);

• Social Media: Informazioni condivise on-line sotto forma di immagini, testo, audio e video;

• Transactions: Insieme di tutte le informazioni derivanti da transazioni.

L'aumentare di questa tipologia di dati con un trend di crescita ben oltre le aspettative e la necessità di avere sistemi in grado di gestire enormi moli di dati veloci, eterogenei e sempre meno strutturati, il tutto mantenendo ottime prestazioni, sono fattori che hanno messo in evidenza come i classici sistemi di basi di

dati relazionali (RDBMS, Relational Database Management System) presentino delle lacune nel trattare tali dati.

I Big Data richiedono la progettazione di sistemi di basi di dati in grado mantenere ottime prestazioni anche con un elevato incremento del volume dei dati. E per questo motivo che sono state proposte nuove tecnologie (sistemi NoSQL) per la soluzione dei problemi sopra descritti.

Con il termine NoSOL (Not only SQL) si indica una famiglia di tecnologie per la gestione dei dati che non segue unicamente il classico approccio relazionale per l'organizzazione, gestione e memorizzazione dei dati. In altri termini, non si adotta più un modello logico in cui i dati sono rappresentati solo sotto forma di prefissate relazioni con vincoli di integrità, ma si cerca invece di rendere più flessibile la gestione dei dati per una maggiore velocità di esecuzione e una più libera rappresentazione delle informazioni.

La definizione NoSQL fu usata per la prima volta nel 1998 per una base di dati relazionale open-source che non usava un'interfaccia SQL. Il termine fu poi reintrodotto nel 2009 da Eric Evans in occasione di un evento per discutere di basi di dati distribuite open-source: il nome fu un tentativo per descrivere l'emergere di un numero consistente di sistemi di basi di dati distribuite non relazionali, non preservanti tutte le classiche proprietà ACID per le transazioni.

Le grandi aziende informatiche hanno investito negli ultimi anni molte risorse nel progettare le proprie tecnologie NoSQL, sperimentando diverse implementazioni alla ricerca della migliore soluzione. Google, Amazon e Facebook sono solo alcune tra queste, ma negli anni recenti anche aziende di medie/piccole dimensioni si sono avvicinate a questa tipologia di database, portando alla creazione di un grande numero di soluzioni. Alcuni esempi sono Cassandra originariamente sviluppato all'interno di Facebook e oggi usato anche da Twitter e Digg, Volde-mort sviluppato e usato da Linkedin, il database NoSQL di Amazon SimpleDB, il servizio di memorizzazione e sincronizzazione dei dati Ubuntu One basato sul database non relazionale CouchDB.

I database di tipo NoSQL hanno un'ottima capacità nel gestire grandi quantità di dati e una particolare idoneità nell'essere implementati su infrastrutture di Cloud Computing, ciò li sta rendendo sempre più popolari negli ultimi anni.

In sintesi, la ricerca di alternative ai database relazionali può essere spiegata da due principali esigenze: la continua crescita del volume di dati da memorizzare e la necessità di elaborare grandi quantità di dati in poco tempo: i nuovi sistemi di basi di dati si sono quindi evoluti per essere più flessibili, utilizzando modelli di dati meno complessi, cercando di aggirare il rigido modello relazionale e di aumentare il livello delle prestazioni nella gestione e interrogazione dei dati.

Caratteristiche dei sistemi NOSQL

I database relazionali sono fondati su relazioni univoche fra i dati: tutti i dati da esaminare sono organizzati in tabel-le, e memorizzati come singole righe di ciascuna tabella. Tale struttura impone la frammentazione delle informazioni fra tabelle differenti, anche quando i dati descrivono un medesimo oggetto (per individuare le informazioni riferite a uno stesso oggetto, occorre effettuare operazioni logiche come il JOIN). L'utilizzazione di tabelle di dimensioni troppo grandi, di conseguenza, risulta di difficile gestione, anche con sistemi molto efficienti.

Nei database non relazionali, seguendo i dettami del modello object-oriented, i dati sono incapsulati in appositi tipi di dati astratti, quindi le informazioni non sono più distribuite su diverse strutture logiche, ma aggregati in oggetti. Ogni oggetto ingloba tutti i dati associati ad un*occorrenza di entità e può essere implementato utilizzando differenti tipologie di strutture dati. In questo modo qualsiasi applicazione può studiare l'entità come oggetto, e valutare, in un solo sguardo d'insieme, tutti i dati informativi ad esso collegati.

Le proprietà comuni alla maggioranza dei sistemi di tipo NoSQL:• scalabilità orizzontale per "operazioni semplici":

• possibilità di replicare e distribuire partizioni di dati su più server;

• un'interfaccia o protocollo di chiamata semplificato;

• un modello di concorrenza più debole rispetto a quello garantito dai RDBMS (Relational database management system) che supportano di contro tutte le proprietà ACID;

• uso efficiente della memoria e di indici distribuiti;

• capacità di aggiungere dinamicamente nuovi attributi.

La scalabilità orizzontale è una caratteristica di particolare importanza quando si parla di Big Data. Con questo termine si intende la possibilità di distribuire i dati e le operazioni su macchine fisiche differenti (nodi) al fine di parallelizzare le operazioni e ottenere un minore quantitativo di dati da elaborare per singolo server. Nonostante sia possibile aumentare le prestazioni anche tramite scaling verticale (aumentando il numero di core e processori di una singola macchina), l'approccio distribuito permette di ridurre i costi utilizzando macchine assemblate e permette di superare i limiti hardware dati dal quantitativo massimo di core e memoria installabili su un singolo nodo.

Un ulteriore aspetto particolarmente importante è la replicazione: avere una grande mole di dati distribuiti su una grande quantità di macchine, porta alla necessità di voler replicare i propri dati permettendo alla rete di aver sempre una copia del dato qualora un nodo dovesse avere un malfunzionamento o non essere più raggiungibile dagli altri nodi.

Riguardo gli svantaggi dei sistemi NoSQL, la semplicità di questi database porta alla mancanza dei controlli fondamentali sull'integrità dei dati; tale compito ricade quindi totalmente sull' applicativo che dialoga col database che ovviamente dovrebbe essere testato in modo molto approfondito prima di essere messo in produzione. La mancanza di uno standard universale (come può essere il linguaggio SQL) è un'altra delle pecche di questi database: ogni sistema ha infatti le proprie API (application programming interface) e il suo metodo di accesso ai dati. Pertanto, se lo sviluppo del database sul quale si è basato un dato applicativo venisse interrotto, il passaggio ad un altro database non sarebbe sicuramente una cosa immediata. Infine, si noti che ogni sistema possiede limitate funzionalità di query.

1. Informazione

Quando si parla di dati, ci si riferisce alla misurazione primaria di un fenomeno che si è interessati a osservare; dall'elaborazione dei dati si ottengono delle informazioni, che ci permettono di aumentare lo stato di conoscenza relativo al fenomeno osservato.

La qualità dell'informazione che otteniamo a partire dai dati può condizionata da fattori di varia natura:

• fattori quantitativi: maggiore è la quantità di dati a disposizione, migliore sarà l'informazione e la conoscenza del fenomeno;

• fattori qualitativi: tra tutti i dati che caratterizzano un certo scenario, occorre saper individuare quelli effettivamente necessari ai propri bisogni distinguendoli da quelli ininfluenti o fuorvianti;

• fattori elaborativi: riguardano le tecniche e gli algoritmi usati per attribuire un significato ai dati;

• fattori temporali: l'informazione ha spesso una validità limitata nel tempo e diventa vitale ottenerla il più velocemente possibile.

Attualmente molte aziende private e pubbliche producono e acquisiscono grandi quantità di dati, che vengono raccolti per essere analizzati e capitalizzati. Le informazioni estratte sono necessarie per rimanere competitivi sul mercato (Figura 1).

Figura 1 Piramide DIKW (Data, Information, Knowledge, Winsdom).

Con il termine Big Data si identificano grandi collezioni di dati (che vanno oltre le caratteristiche dei classici strumenti di database) e le tecnologie usate per estrarre da queste le informazioni desiderate.

Questo fenomeno si è evoluto in funzione della rapida crescita del numero e della varietà di dati resi disponibili da Internet:

• Web and Social media data: dati clickstream (cioè la sequenza di click di un utente su una pagina internet) e interazione con social media, per esempio Facebook, Twitter, Instagram, LinkedIn ecc.

• Machine to machine data: letture di sensori, misuratori e altri dispositivi facenti parte del sistema Internet of Things (IoT), in cui oggetti «intelligenti» interconnessi si scambiano dati in automatico.

• Big transaction data: dati economici-finanziari, di tipo sanitario o di telecomunicazioni.

• Biometric data: impronte digitali e facciali, scansioni della retina, dati genetici, calligrafia, ecc.

• Human-generated data: grandi quantità di dati (spesso non strutturati) come email, registrazioni vocali, note di call center, documenti digitalizzati, sondaggi, cartelle cliniche ecc.

Nelle attività quotidiane su tecnologie ICT, lasciamo, più o meno consapevolmente, tracce digitali dei nostri interessi, affari, relazioni, acquisti, comunicazioni, movimenti. Le nostre ricerche su Internet, l'uso del GPS, i messaggi vocali, le e-mail, i tweet o i post su qualsiasi social network generano dati che seminiamo dietro di noi (Figura 2) .

In questo senso, i Big Data diventano un vero e proprio osservatorio sociale, in grado di registrare, misurare e prevedere i bisogni di mobilità, di risorse economiche o energetiche, la diffusione di opinioni, crisi, pandemie ecc.

Figura 2 Stima della quantità di dati generati ogni minuto dall'uso dei servizi internet.

2. Big Data e Cloud Computing

Con il termine **Cloud Computing** o semplicemente Cloud ci si riferisce a un insieme di tecnologie e di servizi offerti da aziende che con i loro data-center favoriscono la fruizione e l'erogazione di applicazioni informatiche, di capacità elaborativa e di stoccaggio dati via web (Figura 3).

La sinergia tra Big Data e Cloud è assodata: dati e contenuti sensibili per svariate necessità fluiscono ininterrottamente nel Cloud che vede così, giorno dopo giorno, aumentare il proprio valore informativo. Tutto ciò non è però esente da critiche. Per il leader del movimento per il software libero, Richard Stallman,

l'idea di utilizzare datacenter ubicati chissà dove, è solo un ulteriore tentativo dei colossi dell'ICT di ingabbiare gli utenti, poiché l'uso delle applicazioni web comporta la perdita del controllo dei propri dati. Egli propone quindi l'uso di network decentralizzati che permettano di navigare anonimamente senza lasciare tracce delle proprie attività. Per esempio, la rete TOR (The Onion Router) utilizza circuiti virtuali crittografati a strati per fare in modo che quando una pagina web arriva a destinazione, abbia un indirizzo IP diverso da quello di partenza.

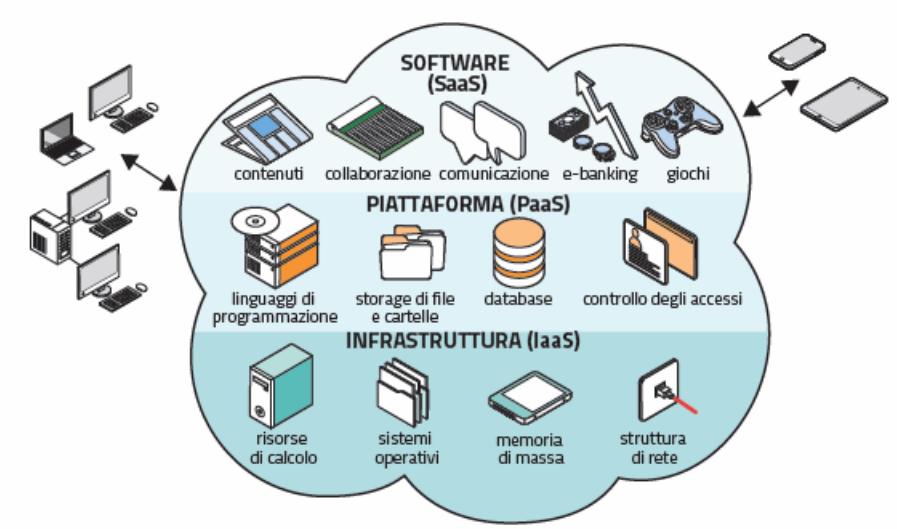

Figura 3 Cloud computing.

3. Big Data Analytics

Per la loro natura di insiemi estremamente grandi di dati, i Big Data non si prestano a essere gestiti con metodi tradizionali perché possono essere anche fortemente inconsistenti. Occorre quindi realizzare collegamenti multipli tra diversi set di dati per individuare correlazioni e gerarchie che ne permettano un adeguato controllo in tempi rapidi. Sviluppare le capacità di gestire e armonizzare tutti questi aspetti è fondamentale per dare un senso ai dati disponibili e sfruttarli al meglio. Per fare ciò si coinvolgono analisti, utenti aziendali e dirigenti in grado di fare ipotesi e porre le domande giuste per individuare tendenze e modelli previsionali.

Le tecnologie e le metodologie di Big Data Analytics hanno tre scopi di fondo:

• descrittivo: ottenere le informazioni necessarie a trasformare la conoscenza derivante da esse in vantaggio competitivo;

• predittivo: stimare i risultati da conseguire. Per esempio, predire la tendenza all'acquisto di un prodotto per particolari tipi di clienti, o rilevare elementi comuni in reati come le frodi per poterli prevenire;

• prescrittivo: valutare l'effetto di future decisioni per consigliare possibili scelte in merito. Si mira non solo a prevedere che cosa accadrà, ma anche a spiegarne il perché e, quindi, a fornire consigli sulle azioni da adottare.

Esistono varie **tecniche di analisi dei Big Data**. Le principali si rifanno al Business Intelligence (BI) e al Machine Learning (ML).

• Business Intelligence: raccoglie dati grezzi su cui vengono poi usati strumenti ETL (Extract, Transform and Load) per manipolare, trasformare e classificare i dati in database strutturati detti data warehouse. Con

strategie di data mining si esplorano i dati usando analisi OLAP (On Line Analytical Process) e KDD (Knowledge Discovery in Database), quindi con dashboard semplificate di visualizzazione si rendono le informazioni accessibili a chi deve analizzare e comprendere le prestazioni passate per pianificare le future strategie di miglioramento dei KPI (Key Business Indicators) aziendali.

• Machine learning: la prima importante differenza rispetto alla BI è che il ML usa l'intelligenza artificiale per rilevare i modelli in milioni di dati. A questa differenza, si possono aggiungere tre aspetti:

− invece dei dati aggregati, il ML utilizza dati individuali, a livello di singola istanza, così che migliaia di variabili possono essere usate per definire modelli dei fenomeni esaminati;

− il ML offre analisi *predittive*, e non descrittive;

− gli ETL sono sostituiti da applicazioni basate su algoritmi che, in automatico, apprendono costantemente dai dati rilevati per integrarne i modelli e fornire capacità predittive.

Supponiamo, per esempio, che un'azienda di e-commerce effettui un'analisi del comportamento dei propri clienti per sapere quanti ne potrebbe perdere nel breve periodo. Un approccio basato sulla BI utilizzerebbe dati di mesi o anni precedenti, insieme ad altre variabili come le tendenze del mercato o il numero di clienti attuale rispetto al passato. Verrebbero così creati dashboard sulla variazione percentuale di clienti prevista, e da questa si potrebbero realizzare campagne di marketing mirate per acquisire nuovi clienti. In un approccio ML, invece, si userebbe l'intero database di clienti persi con i loro profili, per cercare modelli di comportamento e determinare quali di loro mostravano segnali di disaffezione e perché. I dati da usare sarebbero i dettagli degli acquisti storici di tutti i clienti, i loro dati anagrafici, i dati dei prodotti in catalogo (codici identificativi, categorizzazioni, prezzi), i dati delle promozioni o delle campagne di marketing. Mentre la BI fornisce un'analisi delle tendenze con la percentuale di clienti che probabilmente verranno persi, il ML effettua previsioni cliente per cliente così da intraprendere azioni personalizzate per riconquistarne l'interesse.

4. Big Data e Smart City

Secondo la definizione dell'Unione Europea «una smart city è un luogo in cui le reti e i servizi tradizionali sono resi più efficienti con l'uso di soluzioni digitali a beneficio dei suoi abitanti e delle imprese. Una città intelligente va oltre l'uso delle tecnologie digitali per un migliore utilizzo delle risorse e minori emissioni. Significa reti di trasporto urbano più intelligenti, impianti di approvvigionamento idrico e di smaltimento dei rifiuti migliorati e modi più efficienti per illuminare e riscaldare gli edifici. Significa anche un'amministrazione cittadina più interattiva e reattiva, spazi pubblici più sicuri e un migliore soddisfacimento delle esigenze di una popolazione che invecchia» (Figura 4).

In una smart city, Big Data e Data Analysis consentono di rilevare e interpretare quantità sempre più importanti di dati strutturati (presenti nei database) e non strutturati (flussi di messaggi, documenti di testo, foto, video, segnali GPS, file audio e social media ecc.) in maniera funzionale per la gestione della città.

Figura 4 Trama informativa di una smart city.

Qui l'IoT gioca un ruolo cruciale nel fornire flussi di dati in tempo reale che permettono di:

• capire ciò che avviene sul territorio per gestirlo in modo puntuale, rapido, efficiente ed efficace;

• prevedere il verificarsi situazioni critiche per poterle prevenire (Figura 5).

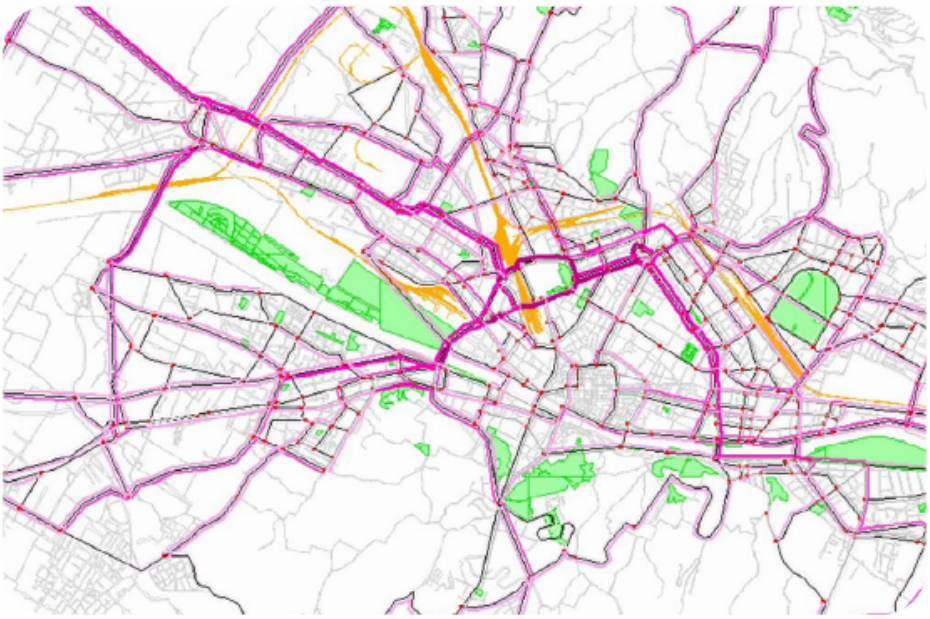

Figura 5 Flussi di traffico stimati su una rete viaria urbana in base alle tracce GPS lasciate dagli smartphone degli automobilisti.

5. Big Data ed etica

I Big Data ci fanno trovare prodotti, auto o mezzi di trasporto per spostarci (magari scegliendo il percorso meno trafficato); tante attività, oggi, senza Big Data non sarebbero così immediate.

Ma il problema dei dati che lasciamo in rete è legato al fatto che qualcuno potrebbe farne un uso illecito o non trasparente. Per evitare che i Big Data possano diventare pericolosi è necessario fissare e rispettare regole certe. Un esempio di problema, di natura politica, è quello di garantire che i sondaggi non siano usati per pilotare onde emotive.

Le più recenti campagne elettorali hanno dimostrato che, con la rivoluzione digitale, le elezioni si vincono anche misurando tutto. Per esempio, negli Stati Uniti, incrociando i dati degli elettori registrati e quelli raccolti in rete o sui social network è possibile condurre campagne elettorali mirate per far arrivare ai singoli individui messaggi sulla propensione di un candidato a prendersi cura proprio dei loro problemi (previdenziali, sanitari, lavorativi ecc.). Per far ciò gli staff elettorali dei vari candidati usano i Big Data per eseguire molte analisi al giorno con diverse decine di migliaia di simulazioni al computer. Questa nuova tecnica, interconnessa e su scala più ampia, si è quasi sempre dimostrata vincente rispetto a quella classica dei sondaggi e della statistica.

I dati personali devono essere trattati in termini di liceità, correttezza, trasparenza. Gli interessi di chi gestisce i Big Data devono sempre bilanciare quelli dei fornitori di dati rispettando la limitazione della finalità, la minimizzazione dei dati e la loro esattezza. In Europa, trattandosi di dati personali, devono essere gestiti in conformità con quanto previsto all'articolo 6 del GDPR – Regolamento Europeo Privacy.