BIG DATA oltre l'hype e le mode: dove stiamo andando e dove indirizzare gli investimenti

La prima metà del 2018 ha evidenziato sempre più come la definizione originaria di Big Data non riesca più adeguatamente a descrivere il fenomeno, gli ambiti di intervento e le tecnologie. L'accelerazione esponenziale nel 2017 di IoT, Machine Learning e Intelligenza Artificiale, ha reso necessario un superamento delle 3 V (in realtà 4) e del resto della classica definizione di Gartner. Ecco la guida per aiutarvi a non restare soffocati dalla valanga dei dati (*di Cosimo D'Amicis, Data Scientist*).

In questo articolo ti guiderò attraverso le trasformazioni continue che subisce l'universo Big Data, così che al termine della lettura avrai un'idea più precisa.
L'adozione dei Big Data, necessita di un cambiamento profondo dei processi che riguardano la gestione dei dati di qualsiasi natura, in qualsiasi azienda. Facciamo quindi il punto per riprendere in mano il filo di Arianna e non perderci tra i diversi trend in ascesa, per capire quale sia l'approccio migliore per ogni attività.

Partiremo dalle origini per meglio capire cosa siano i Big Data e comprendere perché vanno definendosi differenti direzioni e ingegnerie, in base allo scenario di impiego, alle risorse a disposizione e a quelle che dovranno essere approntate.

LA STORIA E IL PASSATO RECENTE

Chiunque abbia provato ad affacciarsi al mondo dei big data, ha impattato con il modello delle 3 V (Volume, Velocità e Varietà dei dati), opera di Doug Laney di Gartner e ormai risalente al 2001: un'era fa.[1]
La locuzione Big Data, risale al 1998 e appare per la prima volta in una presentazione dell'accademico John Mashey[2], al tempo Chief Scientist di SGI.

Nella definizione classica, che comprende **tre** parti, il termine Big Data descrive:

- dati caratterizzati da grandi Volumi, Velocità e Varietà (prima parte);
- le tecnologie cost-effective che ne consentono la gestione (seconda parte);
- il fine ultimo: la capacità di trarre valore e vantaggi competitivi dai dati, allo scopo di migliorare la visione di insieme per supportare le intuizioni e le decisioni che rendono un business performante (terza parte).

---

[1] "3D Data Management: Controlling Data Volume, Velocity, and Variety" di Doug Laney per Meta Group (ora Gartner), https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2] "Big Data ... and the Next Wave of InfraStress", di John Mashey per SGI, http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Con *Volumi* è facile immaginare che ci riferiamo alla quantità dei dati; *Velocità* fa riferimento non solo alla rapidità con cui i dati vengono acquisiti e ingeriti ma anche alle diverse logiche temporali, che sottintendono alle diverse fonti dei dati, da conciliare e a volte sincronizzare; *Varietà* rappresenta la co-esistenza di dati di natura eterogenea (provenienti da differenti fonti, human o machine), più o meno strutturati e organizzati in maniera e in **formati** differenti.

La quarta V può corrispondere invece a *Veridicity*, secondo le diverse interpretazioni, e coincide con le esigenze di qualità del dato.
In ogni progetto che riguarda le decisioni assistite dai dati, la qualità è *cruciale* per impedire che i risultati siano privi di significato: se i dati non rappresentano adeguatamente la realtà, vanifichiamo l'investimento e restituiamo indicazioni errate al business.

Il principale vantaggio delle tecnologie Big Data consiste nella disponibilità immediata del dato per l'analisi, per cui il concetto di Veridicity diventa peculiare e va adeguato al singolo contesto: "si" agli arrotondamenti necessari per conciliare velocemente diverse fonti, entro i limiti consentiti dallo scopo dell'analisi, assolutamente "no" a dati corrotti che necessitano di ulteriori elaborazioni.

In molti si concentrano sulle prime due parti della definizione classica di Big Data, che seppur essenziali ne possono far perdere di vista lo scopo: dotare il business di una grande ma agile capacità di adattarsi a ogni genere di cambiamento con rapidità, l'unica costante del mondo odierno.

Per dare un quadro più completo che non solo descriva i Big Data ma che ne illustri le implicazioni, è stato suggerito un modello a tre tier, dove le 4 V originarie sono alla base: Mark Beyer di Gartner (il padre del concetto di "logical data warehouse") ha definito un modello a 12 dimensioni che contempla tutte le necessità dell'*Extreme Information Management*[3].
Nel suo modello i Big Data si trovano alla base e ogni dimensione corrisponde a un aspetto legato ai dati che il business deve gestire e presidiare al meglio per trarre il massimo valore dai Big Data.

Il **valore** che ogni adozione tecnologica può generare deve essere sempre al centro: può sembrare fin troppo ovvio ad alcuni, ma è l'obiettivo reale e fondamentale. Esempi concreti di come le conoscenze raccolte dalle analisi Big Data sono per esempio la validazione di un'idea per una nuova linea di prodotti, un'opportunità di cross-selling, un'intuizione per la riduzione dei costi e i risultati specifici per un certo settore (pensiamo al farmaceutico e alla scoperta di un effetto causale che può tradursi in una cura di una malattia, pensiamo ai modelli in grado di predire i fermi macchina per rotture).

[3] Una semplice critica di questo concetto può trovarsi qui https://forwardthinking.pcmag.com/show-reports/289401-beyond-big-data-gartner-on-extreme-information-management

Ogni grande progetto che riguarda i dati deve quindi generare un **valore** per l'azienda che si impegna nell'eseguire le analisi. Altrimenti, il rischio è quello di eseguire un qualche esperimento tecnologico per il gusto di farlo.

Dove sono i Big Data e perché ne ho bisogno?

In tutte le aziende che hanno provato a interrogarsi sui Big Data ci si pone una domanda: dove sono i miei Big Data? I dati in mio possesso sono già tali? Passare ai Big Data significa stravolgere l'attuale sistema di gestione dei miei dati a cancellare il mio DWH? Ho già BI, CRM e altri sistemi di supporto alle decisioni che tutte le funzioni aziendali utilizzano, quale contributo possono dare i Big Data? Ha senso parlare di Big Data nella mia organizzazione?

Uno dei nodi fondamentali da sciogliere riguarda la democratizzazione del dato a livello aziendale. Possiamo avere diversi applicativi e database che lavorano sui dati, ma è facile che questi siano disponibili solo secondo logiche verticali, all'interno di silos analitici sigillati non comunicanti tra loro.

Passare a un approccio Big Data vuol dire riuscire a prendere quel dato e a renderlo più liquido, orizzontale: il risultato è un insieme di sistemi dove gli algoritmi di *Advanced Analytics* possano attingere ai dati provenienti da siti web, dal marketing, dai sensori installati in produzione, dalle vendite, dalla logistica e così via.

In altri casi, la necessità è puramente tecnologica: se voglio fare sentiment analysis o sfruttare le potenzialità della NLP (Natural Language Processing, analisi del linguaggio naturale) dovrò dotarmi di strumenti in grado di gestire questo tipo di dato molto poco strutturato, in un senso classico e "relazionale" del termine. Pensiamo a tutta la conoscenza che passa attraverso le e-mail che corrono attraverso ogni azienda. Big Data vuol dire mettersi in condizione di disporre di questi *dark data/dati oscuri* per estrarre conoscenza definita come *actionable insights,* in italiano traducibile come *intuizioni fattibili*.

Spesso questi dati (pensiamo proprio per esempio ai flussi di e-mail) più che non strutturati, presentano una struttura variabile che i sistemi predisposti dall'IT devono essere in grado di gestire.

La risposta alla domanda posta nel titolo del paragrafo è: tutte le organizzazioni sono in possesso di dati strutturati, semistrutturati e non strutturati, da cui ancora non traiamo il massimo della conoscenza e del valore possibile.

Con il concetto classico di Data Warehouse e Data Mart, siamo obbligati a definire lo schema prima di poter caricare i dati e lavorarci. Lavorare con i Big Data invece prevede il cosiddetto *Data Lake*, che permette di superare il concetto di *schema* e di *single version of the truth,* che permette di memorizzare, gestire e rendere subito disponibile l'informazione, *qualsiasi* sia la sua struttura.

Il Data Lake non cancella quindi il Data Warehouse ma lo *complementa,* per andare incontro alle necessità imposte dalle ormai famose 3 V di cui abbiamo parlato e per consentire agli utenti intermedi e finali di sperimentare con il dato e di analizzarlo non appena acquisito, per poterlo poi strutturare in uno schema ed eventualmente essere caricato nel DWH.

Data Lake è oggi quasi-sinonimo di **Hadoop** giunto alla terza versione nel dicembre 2017 e che sostanzialmente nasce come risposta alle necessità sollevate dai Big Data: non ci soffermeremo qui sulle caratteristiche e le peculiarità di Hadoop.
Per giustificare l'affermazione ci basti ricordare che Hadoop è stato concepito per essere *nativamente cost-effective* (venne inizialmente concepito per sfruttare vecchio hardware ammassato nei data center), e che memorizza l'informazione su diverse macchine e che ha dei meccanismi di consistenza del dato che rendono disponibile lo stesso, anche quando una parte dell'infrastruttura smette di funzionare.

Esistono diverse versioni e distribuzioni commerciali di Hadoop, mentre l'alternativa cloud secondo il modello IaaS (Infrastructure as a Service) può essere ospitata su EC2 di Amazon AWS oppure su Microsoft Azure.

L'errore di molte organizzazioni in questo senso, è di vedere il *Data Lake* come un pezzo di storage intermedio, dove poter esclusivamente riporre le informazioni prima di una fase ETL per il caricamento nel DWH.
Limitarsi a questo significa *violare* la terza e la più importante parte della definizione classica di Big Data perché non sfruttiamo al massimo le possibilità analitiche offerte da questi mezzi tecnologici, e i possibili benefici.

Hadoop e "cugini" non rappresentano l'unica tecnologia nè l'unica strategia di memorizzazione dei dati quando si parla di Big Data. Anche le tecnologie basate sul concetto di *blockchain* fanno parte di tutto ciò che possiamo inscrivere nella locuzione. Potremmo considerare le tecnologie blockchain come una sorta di *Peer-to-peer Big Data.*

Hadoop è stato un cavallo vincente ed è determinante tuttora nella maggior parte delle implementazioni Big Data grazie al suo approccio basato su disco e quindi cost-effective, tuttavia oggi **Spark** riveste uguale importanza. Spark nasce come componente dell'ecosistema che gravita attorno a Hadoop e si evolve rapidamente fino a ricoprire uguale importanza. Scopriamo perché.

Spark è basato sull'elaborazione dei Big Data in memoria, un approccio completamente diverso da Hadoop: un altro motivo per cui la "vecchia" definizione di Big Data non basta più per dire di avere familiarità con la tecnologia.

Il superamento della definizione classica di Big Data

Il 2017 ha visto l'aumento esponenziale dell'adozione e dello sviluppo di tutte le tecnologie che fanno riferimento a questi tre ambiti: IoT, Intelligenza Artificiale/Machine Learning e Cyber Security.

Ognuno di questi tre ambiti ha stimolato gli sviluppatori a creare nuovi sistemi e moduli che fossero adatti al particolare campo di applicazione, e alle mutate condizioni di *Volumi, Velocità e Varietà* che ogni caso impone.

Come ulteriore complicazione e in pieno *stile Big Data,* i tre ambiti specificati non sono compartimenti stagni: affrontare e risolvere un problema con tecnologie IoT ci offre possibilità di *advanced analytics* che possono essere sfruttate al meglio con approcci AI/ML e che impongono diversi challenge di Cyber Security; lo stesso accade quando il punto di partenza è l'AI o la Cyber Security.

Le 3 V, nel 2018, assumono quindi significati diversi, sopratutto quando si arriva alla fase di implementazione. Vediamo cosa è accaduto nel 2017 e quali sono le prospettive future.

Il mashup tra IoT, AI/ML e Cyber Security, e la spinta all'adozione necessaria dei Big Data

Come previsto da molte analisi, il mercato legato all'IoT è esploso e abbiamo esempi reali che vanno ben oltre l'accensione di un condizionatore attraverso un'app su uno smartphone.
L'Internet of Things sta diventando parte essenziale di alcuni business, attraendo enormi investimenti anche nei servizi, come nelle assicurazioni.
Poiché si prevede che la spesa totale delle imprese per l'IoT raggiungerà i 6.000 miliardi di dollari entro il 2021, il mercato si concentrerà su dispositivi sempre più reattivi per costruire una rete superlativamente più intelligente.
L'integrazione di IoT con le tecnologie di Intelligenza Artificiale ha permesso nel 2017 a oltre il 60% delle multinazionali di utilizzare i dati per ottimizzare i processi e risparmiare milioni di euro.[4]

Nel 2018 c'è una maggiore attenzione sul rafforzamento della sicurezza, mentre lo sviluppo delle sinergie IoT/AI continuerà a svilupparsi.

[4] Dati provenienti da diversi report e analisi Gartner, IDC, Machina. Una raccolta esaustiva di questi report può trovarsi agli indirizzi https://www.forbes.com/sites/louiscolumbus/2017/12/10/2017-roundup-of-internet-of-things-forecasts/#6fb5b93f1480 e https://www.postscapes.com/internet-of-things-market-size/

Le aziende si sono rese conto della necessità imperativa di un sistema di sicurezza informatica impeccabile, e le applicazioni abilitate dall'intelligenza artificiale servono proprio a questo: sistemi di elaborazione di dati storici degli attacchi informatici e per la previsione delle strategie di difesa saranno sempre più richiesti.
Anche la sicurezza di Hadoop è stata migliorata da pacchetti come Apache Atlas e Ranger, per raggiungere i livelli richiesti dalle organizzazioni enterprise di grosse dimensioni.
Si prevede che l'internet delle cose possa potenzialmente contribuire con 15.000 miliardi di dollari al PIL mondiale entro il 2030, ed è certo che il 2018 sarà il punto di partenza per raggiungere questo traguardo.

In questo scenario, i Big Data superano la nozione iniziale di necessità per diventare parte essenziale dell'infrastruttura di ogni moderna organizzazione, industriale e non: i Big Data stanno contribuendo in positivo alla battaglia contro la malaria nei paesi africani, grazie al lavoro dell'organizzazione non governativa PATH, ed esistono intere piattaforme come DrivenData.org, dedicate al (Big) Data for good, ovvero all'impiego dei dati per il bene comune.

I trend emergenti: real-time analytics e Interactive SQL

Hadoop è certamente un riferimento fondamentale ma quando parliamo di dati è inevitabile menzionare SQL. A che punto siamo con i tool che permettono di sfruttare le potenzialità dei Big Data e dei database più tradizionali?
Società come Exasol e MemSQL stanno spingendo per le real-time in-memory analytics in tempi rapidissimi, e abbiamo a disposizione una vasta scelta di motori SQL-on-Hadoop (Apache Impala, Hive LLAP, Presto, Phoenix e Drill) e OLAP-on-Hadoop (AtScale, Jethro Data e Kyvos Insights), che stanno azzerando i confini tra DWH e Big Data.
SQL Server di Microsoft nella sua ultima versione, supporta i dati JSON.[5]
In alcuni casi, l'adozione di Hadoop sta superando quella dei database tradizionali, come negli applicativi di analisi dei carichi di lavoro, che operano secondo logiche giornaliere.
Così come si sviluppano le tecnologie, così diventa sempre più cruciale la possibilità di ingerire dati diversi, memorizzati in maniere diverse e che necessitano dello sviluppo di connettori che contribuiscono a rendere il cosmo dei Big Data sempre più maturo.

Tecnologie come Spark e Azure stanno permettendo di superare i limiti di Hadoop sopratutto per le applicazioni che prevedono un'interazione in tempo reale o i dati in streaming.
Il primo è spinto dal suo stesso ecosistema, in costante ed esponenziale sviluppo (che comprende componenti specifici per AI e ML come MLlib), il secondo dalla facilità di integrazione con le tecnologie Microsoft preesistenti.

[5] Per restare al passo con il rilascio di nuove tecnologie e software Big Data, suggerisco www.datatau.com e i report offerti da Tableau

Stiamo osservando una convergenza tra Cloud e Big Data (spinta sopratutto dall'IoT), con costi sempre minori e readiness maggiore del primo, con l'ambizione di fornire sia i servizi di storage che una parte dei sistemi di analytics in soluzioni gestite.

Un nuovo filone che emergerà sempre più sarà quello dei metadati, o meglio dei "Meta Big Dati": sistemi come Alatione e Waterline promettono di occuparsi di categorizzare automaticamente l'informazione contenuta nei nostri Data Lake.

Big Data diventa tecnologia abilitante in prodotti innovativi come Looker: un tool per la collaborazione in team sulle dashboard di qualsiasi natura, con dati di ogni tipo; Cluvio, che abbatte i confini tra le query eseguite tramite SQL e le advanced analytics e dataviz messe a disposizione da R.

Sul fronte data cleaning si fa notare Forestpin, un tool dedicato alla ricerca di anomalie dei dati, e che aggiunge la possibilità di creare dei job per le operazioni routinarie di data cleaning, in maniera intelligente.

L'"Elastic Stack", ecosistema cresciuto attorno all'open source Elasticsearch, è il motore di ricerca più utilizzato dalle aziende dal 2017.

Tornando a Hadoop e Spark… cosa scelgo per il mio business?

Ho promesso di orientarvi all'inizio dell'articolo, per cui ecco alcune considerazioni per meglio comprendere in cosa Hadoop differisce da Spark, perché sono due elementi fondamentali di ogni progetto Big Data.

Hadoop 2 e 3 sono motori di elaborazione dati sviluppati open source in Java, rilasciati rispettivamente nel 2013 e a dicembre 2017. Hadoop è stato creato con l'obiettivo primario di mantenere l'analisi dei dati su disco (per la logica del cost-effective riportata all'inizio dell'articolo), secondo un paradigma noto come e*laborazione batch*. Pertanto, Hadoop da sè non supporta l'analisi e l'interattività in tempo reale.

Spark 2.X è un motore di elaborazione e analisi dati open source sviluppato in Scala e rilasciato nel 2016. Spark arriva in un momento in cui l'analisi in tempo reale delle informazioni diventa cruciale, poiché molti importanti servizi si affidano alla capacità di elaborare i dati immediatamente. Di conseguenza, Apache Spark è stato costruito per l'elaborazione real-time dei dati ed è ora molto popolare perché può gestire in modo efficiente i flussi di informazioni, consentendo di elaborare i dati in modalità interattiva.

Hadoop lavora su disco, quindi non ha bisogno di molta RAM per funzionare. Questo può determinare un TCOS (Total Cost of Ownership) minore rispetto a Spark. Hadoop 3 richiede ancora meno spazio su disco (è stata rivisto e ottimizzato il meccanismo di fault tolerance). Spark ha per contro bisogno di molta RAM, il che determina costi maggiori.

Hadoop tende puntualmente a perdere la sfida della velocità contro Spark. Si stima che a parità di requisiti, Spark sia 100 volte più veloce di Hadoop in modalità in-memory e 10 volte più veloce nelle operazioni su disco. Hadoop 3 è in media il 30% più veloce di Hadoop 2.

Hadoop tende a essere più maneggevole con Java ma può essere utilizzato con diversi linguaggi. Spark supporta Scala, Java, Python e R.
Hadoop è più sicuro di Spark che usa una password: per l'autenticazione si affida a Kerberos e alle ACL (Access Control Lists)

YARN, il resource manager di Hadoop, ha subito un aggiornamento che rende separabili i processi di lettura e scrittura, fattore che migliora inevitabilmente la scalabilità.

È fondamentale ribadire che Spark gira su Hadoop e che può godere di un ampio parco di librerie in perenne crescita, che rendono disponibili facilmente diversi strumenti.

Tra le più celebri ed essenziali non si possono non citare MLlib e Spark SQL.

Sperando di aver reso più chiari a voi lettori l'argomento e la terminologia, resto a disposizione e in attesa dei vostri commenti.

COSIMO D'AMICIS: l'aiuto per comprendere un mondo circondato di dati

Via Amendola 13, 40121 - Bologna
Telefono (+39) 320-7529385
e-mail info@cosimodamicis.it
www.cosimodamicis.it