DATA SCIENCE: LA SCIENZA DEI DATI

di Germana Grazioli, Docente di Economia Aziendale

La mole di dati prodotti, memorizzati e diffusi dalla società dell'informazione ha assunto dimensioni tali da rendere inadeguati, per la loro gestione, i sistemi consolidati di database e richiedere l'utilizzo di nuove tecniche e nuove metodologie di memorizzazione, elaborazione e analisi. E' nata così la **scienza dei dati**, una disciplina trasversale, che utilizza metodi e tecniche provenienti da varie discipline quali l'informatica, la statistica e la matematica.

I Big Data

Big, alla lettera "grande". I Big Data sono infatti **grandi masse** di dati eterogenei, in parte strutturati, come i database, ma in gran parte non strutturati, come immagini, email, dati GPS, informazioni prese dai social network. L'**eterogeneità**, sia per quanto riguarda le caratteristiche dei dati sia per le fonti da cui gli stessi provengono, è il primo e principale fattore di complessità dei Big Data.

All'eterogeneità si aggiungono poi, quali ulteriori fattori di complessità, caratteristiche quali il **volume**, la **velocità di generazione** e **raccolta** e la **varietà**.

Multipli del byte					
Prefissi SI			Prefissi binari		
Nome	Simbolo	Multiplo	Nome	Simbolo	Multiplo
chilobyte	kB	10^3	kibibyte	KiB	2^{10}
megabyte	MB	10^6	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	yobibyte	YiB	2^{80}

Non esiste un valore soglia, in termini di dimensione, al disopra del quale classificare una massa di dati come Big Data. All' incremento dei dati prodotti nel mondo nell'ultimo biennio, valutato intorno al 90%, è corrisposta un' evoluzione tecnologica che ha consentito di gestire una mole misurata in **zettabyte**, ovvero miliardi di terabyte. Masse di dati di queste dimensioni richiedono una potenza di calcolo particolarmente elevata con processi di elaborazione eseguiti su decine, centinaia o anche migliaia di server.

Crescita dei Big Data : il modello delle 3 V

Studi sui Big Data degli anni 2000 hanno portato ad individuare una **crescita** nel tempo di tipo **tridimensionale**, che Douglas Laney, nel 2001 vice presidente e Service Director di Meta Group, ha sintetizzato nel modello delle **3V** (tre variabili) : [1]

[1] https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

- **Volume**, ossia quantità di dati, strutturati o non strutturati, generati ogni secondo da sorgenti eterogenee quali database, sensori, email, social media (tanto per citarne alcune);
- **Varietà**, in quanto ai dati strutturati contenuti nei database relazionali (tabelle) si sono aggiunte differenti tipologie di dati non strutturati o semi strutturati come testi, log di web server, immagini, video, audio, elementi di calcolo;
- **Velocità,** alla quale i nuovi dati vengono generati e a cui devono corrispondere tempi brevi di raccolta e analisi.

Successivamente il paradigma di Laney è stato ridefinito con l'aggiunta di altre due variabili importanti collegate alla crescita esponenziale delle masse di dati:

- **Veridicità**, indispensabile per garantire l'affidabilità dei risultati delle analisi dei dati che sono alla base di decisioni aziendali. A questo proposito sono stati elaborati sistemi di attribuzione di indici di veridicità;
- **Variabilità**, dovuta alla grande varietà di formati e provenienza, che può comportare rischi di errata comprensione dei dati al momento dell'interpretazione.

I Big Data sono da molti definiti "il nuovo petrolio" per l'economia, ovvero una fonte inestimabile di **valore**, che qualcuno definisce la **sesta V** del modello. Ma per estrarre conoscenza, e quindi ricavare valore dai Big Data, non bastano le migliori tecnologie disponibili sul mercato, occorrono progetti impegnativi, anche economicamente, per la raccolta dei dati e la loro analisi, che è opportuno far precedere da una valutazione del valore effettivo portato al business.

Modalità di acquisizione dei dati

A seconda del contesto da cui provengono e degli strumenti di rilevazione utilizzati i dati sono distinti in :

- **human generated,** provenienti da piattaforme di social network, blogging e micro-blogging, social news, social bookmarking, multimedia sharing, wiki, siti di domande e risposte, siti di recensioni, portali di e-commerce, click stream da siti web gestiti tramite cookies;
- **machine generated,** prodotti da sensori, GPS, IoT, RFID, centrali di monitoraggio di eventi meteorologici, strumenti scientifici, sistemi di HFT (High Frequency Trading) dei mercati finanziari, dispositivi biomedicali ed altri.

Processi di gestione e analisi dei dati

La gestione e l'analisi di masse di dati che crescono rapidamente in volume e varietà richiedono la definizione di adeguati **processi**, riconducibili a due tipologie :

- **Big Data Management**, per l'acquisizione, la memorizzazione ed il recupero dei Big Data;
- **Big Data Analytics,** per l'analisi e l'acquisizione di informazioni utili da grandi dataset. A seconda degli obiettivi di analisi perseguiti si parla di:

- **Descriptive Analytics,** per descrivere la situazione attuale e passata dei processi aziendali, in formato testuale e grafico;
- **Predictive Analytics**, ossia analisi dei dati finalizzata a delineare scenari di sviluppo futuro utilizzando tecniche matematiche e statistiche;
- **Prescriptive Analytics**, che unitamente all'analisi dei dati propongono soluzioni operative (di breve periodo) e/o strategiche (di medio-lungo periodo) sulla base dei risultati della stessa;
- **Automated Analytics,** in grado di implementare autonomamente l'azione da intraprendere in base al risultato delle analisi svolte.

IoT (Internet of Things)

Emblema dell'industria 4.0, **Internet of Things (IoT)** - Internet delle Cose **-** è un neologismo utilizzato per indicare un insieme di tecnologie che permettono di **collegare a Internet qualunque tipo di apparato**, allo scopo di monitorarne il funzionamento e ricevere informazioni, a cui far seguire azioni di vario tipo.

Evoluzione "intelligente" della semplice sensoristica, i dispositivi IoT possono svolgere funzioni di diverso grado di complessità come :

- rilevare e comunicare dati in formato digitale attraverso la connessione in rete;
- effettuare un primo livello di elaborazione e selezione dei dati a livello locale, per trasferire solo quelli rispondenti a determinati requisiti;
- raccogliere dati, effettuare un primo livello di selezione ed eseguire immediatamente azioni in funzione delle informazioni elaborate.

Le tecnologie IoT hanno trovato applicazione in ogni ambito della vita sociale, da quello domestico (controllo a distanza di elettrodomestici) a quello cittadino (controllo del traffico, delle condizioni atmosferiche), sanitario (monitoraggio dei comportamenti), ambientale (controllo dei fattori inquinanti) fino all' ambito industriale, nelle cosiddette aziende 4.0. Nelle aziende 4.0 le tecnologie IoT consentono, ad esempio, di monitorare telematicamente ogni fase della lavorazione dei prodotti, per individuare eventuali criticità del processo, o di collegare direttamente ogni macchinario a dispositivi diagnostici, in grado di segnalare tempestivamente eventuali malfunzionamenti e di ricevere istruzioni per l'autoriparazione.

Immagazzinamento, modellazione e analisi

L'immagazzinamento dei Big Data comporta la necessità di sviluppare, in parallelo, **infrastrutture hardware** per l'archiviazione di enormi dataset non strutturati o semi-strutturati e **strumenti per la gestione** dei dati archiviati.

L'organizzazione logica dei dati è dunque affidata a basi di dati NoSQL - **Not Only SQL** – in grado di gestire i Big Data superando le restrizioni del modello relazionale e del linguaggio SQL grazie a caratteristiche quali, ad esempio, l'essere schema-less, quindi in grado di memorizzare

informazioni eterogenee, con diversa struttura e/o formato, all'interno della stessa entità.

Alla fase di immagazzinamento viene affiancata una fase di trasformazione dei dati per prepararli alla successiva fase di analisi. Lo scopo dell'analisi è quello di estrarre conoscenza dai Big Data, individuando negli enormi dataset correlazioni, trend, pattern ed indici statistici di vario tipo.

Sui dati immagazzinati , strutturati, semi-strutturati e non strutturati, possono essere eseguite :

- **analisi di testi** (text mining)**,** per l'estrazione di informazioni da testi non strutturati contenuti in documenti, email, pagine Web e post su blog e social network. Per l'analisi di testi sono stati sviluppati algoritmi per il riconoscimento di argomenti (topic modeling), la ricerca delle migliori risposte ad una domanda (question answering), l'individuazione delle opinioni degli utenti su determinate notizie (opinion mining) ed altri ancora;
- **analisi di dati multimediali,** mediante algoritmi di machine learning che permettono di estrarre informazioni da file multimediali, classificati e resi accessibili utilizzando algoritmi d'indicizzazione (multimedia indexing) e raccomandazione (multimedia recommendation);
- **analisi del Web**, per ricavare informazioni e conoscenza sui contenuti, la struttura e l'utilizzo del Web. La topologia può essere ricostruita mediante algoritmi di crawling che seguono i collegamenti ipertestuali per individuare relazioni tra pagine o siti Web. Il profiling dell'utilizzo del Web da parte degli utenti, finalizzato a personalizzare le loro esperienze di navigazione, comporta l'esame di un numero elevato di log di server, sessioni, transazioni, ricerche e visite.

La Data Science e i nuovi profili professionali

L' «Osservatorio Big Data analytics & Business intelligence» del Politecnico di Milano ha svolto nel 2018 un'analisi delle offerte di lavoro pubblicate su Linkedin individuando le figure professionali più richieste dalle aziende nel settore delle Scienze dei dati. Da quest'analisi è stato possibile individuare almeno tre tipi di profili professionali: il Data Scientist, il Data Engineer e il Data Analyst.

Il **Data Scientist** è il profilo professionale di livello più alto, a cui vengono richieste forti competenze interdisciplinari
Compito del Data Scientist è gestire la trasformazione dei Big Data in conoscenza utilizzabile da parte dell'azienda. Per trasformare dati grezzi in informazioni utili il Data Scientist deve saper sovraintendere a processi di :

- mappatura, organizzazione e controllo delle fonti e dei flussi dei dati aziendali;
- modellizzazione dei dati mediante algoritmi matematico-statistici e trasformazione degli stessi in informazioni

- analisi del valore dei dati per ogni area di business, in particolare per le attività che costituiscono il core business dell'azienda, in funzione dello sviluppo di nuovi servizi o nuovi prodotti.

Al Data Scientist si richiedono pertanto competenze in materia di Informatica **(**conoscenza dei linguaggi di programmazione e di strumenti per la gestione dei Big Data), Project Management (capacità di gestione dei progetti, di coordinamento e di controllo di tutte le risorse impiegate nel raggiungimento di specifici obiettivi), Matematica, Analisi e Statistica (capacità di analisi dei dati, di creazione di modelli, di utilizzo di strumenti e tecniche statistiche e di interpretazione dei risultati).

Il Data Scientist deve necessariamente possedere una piena conoscenza del core business aziendale, delle principali dinamiche dell'impresa e dei mercati nei quali essa opera. Fondamentale, inoltre, vista l'interdisciplinarietà delle sue competenze, la capacità di gestire le relazioni e di operare in teamwork.

Il **Data Engineer** ha il compito di realizzare e amministrare strutture in grado di gestire i Big Data. Si tratta di un ruolo molto tecnico, per il quale si richiede competenza nell' utilizzo di strumenti informatici e dei linguaggi di programmazione.

E' compito del Data Engineer gestire l'infrastruttura, la cosiddetta Data Pipeline, mediante la quale i dati vengono trasferiti dai punti di raccolta agli strumenti di immagazzinamento, elaborazione e analisi. In questo contesto, è di sua competenza il controllo di integrità dei dati a seguito di ogni operazione di trasmissione o di memorizzazione .

Il **Data Analyst**, infine, ha il compito di trasformare i dati in una serie di informazioni da comunicare al management aziendale. Al Data Analyst sono richieste abilità analitiche, propensione per il ragionamento matematico e statistico, competenze di programmazione e doti comunicative, necessarie per presentare i risultati dell'analisi dei dati in forma chiara e comprensibile. Il Data Analyst non deve avere necessariamente forti competenze tecniche, deve però saper utilizzare le tecnologie disponibili per sviluppare le analisi che gli sono richieste.