I QUADERNI DI
Telèma

Big Data Analytics Lab: esperienza e competenza per crescere

A cura di:
Giambattista Amati
Marco Bianchi
Daniela D'Aloisi
Fondazione Ugo Bordoni

LE PERSONE SONO GENERATORI INCESSANTI DI DATI sia volontariamente, come nell'uso dei social media, sia inconsapevolmente, come nell'uso di bancomat, carte di credito, passaggi autostradali, telefono, etc. Altri dati sono prodotti dalla macchine, quali sensori, satelliti, contatori, etc. Questi dati coprono i settori più diversi, dalla meteorologia alla geografia, dai consumi energetici alla salute, dai trasporti alle statistiche. Secondo quando emerso dallo *Hadoop Summit 2014*, l'universo digitale crescerà da 3.2 a 40 zettabyte (10^{21} byte) da qui a sei anni, e l'85% di questi dati proverrà dalle nuove applicazioni sulle nuove reti.

I dati sono ormai diventati un'importante risorsa sociale ed economica al pari delle tecnologie che finora sono state alla base dei processi innovativi, anzi la coppia innovazione/Big Data è ormai considerata un motore per lo sviluppo tecnologico, la creazione di nuovi strumenti, l'insorgere di nuove professionalità. La loro stessa gestione richiede competenze che non sono attribuibili a una sola disciplina: nell'articolo parleremo della figura del *data scientist* di formazione multidisciplinare, ma già da alcuni anni nell'agenda della pubblica istruzione negli USA è entrato il termine STEM (*science, technology, engineering, mathematics*) per indicare una figura in grado di fronteggiare la complessità delle sfide tecnologiche e sociali.

La Fondazione Ugo Bordoni ha una lunga esperienza nel campo dell' Information Retrieval and Mining da testi, dati strutturati e semi-strutturati, come ad esempio pagine Web, database, blog e microblog. Attualmente le attività di ricerca hanno come fulcro il *Big Data Analytics Lab*, il nostro laboratorio di R&D per lo sviluppo di soluzioni scalabili per il monitoraggio, la ricerca e l'analisi in tempo reale delle reti sociali. Sono due i maggiori filoni di studio, l'Information Retrieval per Big Data e i Big Data analytics le cui parole chiave principali sono: dati distribuiti, ricerca e analisi in tempo reale, indicizzazione in tempo reale, *near real-time sentiment analysis*, ricerca su blog e microblog, social web analytics.

Questo quaderno di Telèma affronta il tema dei Big Data da un punto di vista tecnologico, mostrando come siano difficili da collezionare, memorizzare o processare con i sistemi convenzionali e descrivendo quali siano le metodologie usate.

Sarà inoltre mostrato come la Commissione Europea stia investendo sulla nuova economia della conoscenza, analizzando le iniziative in atto.

Quello che emerge è come con i Big Data siamo di fronte ad una rivoluzione, ma non è la quantità dei dati a essere l'aspetto più innovativo: ciò che è effettivamente rivoluzionario è quello che si può fare con questi dati.

Supplemento al numero 302 di
MEDIA DUEMILA
Novembre 2014

I QUADERNI DI Telèma

Nei numeri precedenti

Qualità e Internet mobile. Le verità nascoste? | 2	Aprile / Maggio 2011
La sostenibilità energetica non può fare a meno dell'ICT	Giugno 2011
Registro Pubblico delle Opposizioni: un'opportunità per i cittadini e le imprese	Luglio / Agosto / Settembre 2011
L'opt-out nel telemarketing è sempre più realtà: dal telefono alla posta, con uno sguardo verso Internet	Ottobre 2011
PANDORA: l'ICT per il Crisis Management	Dicembre / Gennaio 2012
Una nuova generazione di sportelli automatici accessibili e usabili da tutti	Febbraio 2012
Campi Elettromagnetici | 1	Marzo 2012
Campi Elettromagnetici | 2	Aprile / Maggio 2012
***misurainternet.it* Qualità dell'accesso ad Internet da postazione fissa**	Giugno 2012
Qualità del servizio dati in mobilità: alla partenza la prima esperienza regolamentare	Luglio / Agosto / Settembre 2012
Loudness: questa pubblicità è "troppo forte!"	Ottobre 2012
Open Government Data: una roadmap tecnica	Dicembre / Gennaio 2013
Un social network a misura della terzà età	Marzo / Aprile 2013
TV, un futuro già presente | 1	Maggio 2013
TV, un futuro già presente | 2	Luglio 2013
Smart Community: l'evoluzione sociale della Smart City	Settembre 2013
Verso una gestione unitaria dell'identità digitale	Ottore 2013
Elottromagnetismo coscienza collettiva regole e necessità	Dicembre / Gennaio 2014
Terminali pubblici accessibili per una società più inclusiva	Marzo / Aprile 2014
AGCOM - FIEG - FUB progetto informatico antipirateria. Diritti d'autore online	Maggio 2014
FUB: ricerca ed innovazione al servizio del Paese	Giugno/Luglio 2014
Esposizione personale e uso del cellulare. Campi elettromagnetici. Le norme e la scienza	Settembre 2014

IL QUADERNO DI TELÈMA È STATO REALIZZATO DALLA FONDAZIONE UGO BORDONI

Presidente: **Alessandro Luciano** | Direttore delle Ricerche: **Mario Frullone**

Una definizione trasversale per i Big Data

La nostra società è destinata a produrre una sempre crescente quantità di dati: nei prossimi anni, a quelli già prodotti dagli utenti del Web e dai social network andranno aggiunti i dati generati dai dispositivi e dai sensori dell'Internet delle Cose (Internet of Things) e quelli che saranno resi disponibili sotto forma di dati aperti (Open Data). A tal proposito Viktor Mayer-Schönberger, dell'Oxford Internet Institute dell'Università di Oxford, illustra l'impatto che l'analisi dei Big Data potrà avere a breve sul nostro stile di vita [1]:

"Siamo agli albori di un'importante nuova era nella storia dell'umanità. Se la rivoluzione di Gutenberg è stata alimentata dalla parola stampata, dal contenuto intellettuale, questa rivoluzione, ormai imminente, lo sarà dai dati e migliorerà il modo in cui prendiamo le nostre decisioni, da quali prodotti acquistare (o produrre) a quali terapie sono efficaci, come educare i nostri figli o come inventare un'autovettura senza conducente. Di conseguenza, nell'arco di un decennio, le nostre vite saranno molto diverse da oggi, sostanzialmente non perché disporremo di un nuovo strumento tecnico, bensì perché avremo una comprensione nettamente migliore della realtà".

Questa rivoluzione sarà sempre più rapida a causa della convergenza d'interessi da parte dell'intero comparto ICT e del mondo della ricerca scientifica. Infatti le aziende ICT, aprendosi alle prospettive di miglioramento del business offerte dall'analisi dei Big Data, dovranno investire di più in attività di ricerca e sviluppo e contribuiranno attivamente alla incessante produzione di nuove tecnologie abilitanti. Il mondo della ricerca scientifica ha già messo al centro le sfide per risolvere i tanti problemi da affrontare affinché le analisi dei Big Data siano rese davvero realizzabili. Vediamo in breve quali sono queste sfide secondo Gartner [2]:

"Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

Questa definizione è quella generalmente accettata quando si affronta il tema dei Big Data. Diversamente da quanto il termine "big" lasci intuire, un problema afferisce ai Big Data non solo in presenza di una enorme quantità di dati da trattare, ma anche quando la velocità con cui i dati devono essere elaborati è determinante, oppure quando è la loro eterogeneità a render necessaria l'applicazione di tecniche e l'uso di tecnologie avanzate per la gestione e il trattamento dell'informazione.

Nella definizione di Gartner si fa riferimento anche alla sostenibilità economica (economicità) delle soluzioni innovative per Big Data. Questa avviene anche attraverso la decisione di quale parco macchine dotarsi. Ad esempio, con l'affermazione di servizi di tipo cloud potrebbe aumentare sensibilmente la possibilità per aziende ed enti di ricerca di creare e gestire cluster di macchine virtuali adattabili alle loro esigenze tecniche e alle loro disponibilità economiche. Il paradigma cloud sposa, infatti, l'approccio alla scalabilità orizzontale (scale-out) tipico delle soluzioni per Big Data spesso basate su algoritmi e tecnologie in grado di sfruttare al meglio le risorse hardware composte da una moltitudine di computer non necessariamente di fascia alta (commodity computing).

La generalità del termine big, sommata alle tante variabili che caratterizzano un'applicazione di Big Data, ai nuovi paradigmi di programmazione nonché alle numerose tecnologie che ruotano attorno, rendono spesso impervia la strada a chi si avvicina alla disciplina della Scienza dei dati.

Secondo Nathan Marz, ingegnere capo presso BackType, società acquisita da Twitter nel 2011 e autore di vari progetti open-source come Apache Storm, Cascalog e ElephantDB "esistono realmente solo due paradigmi per il processamento dei dati: batch e stream". L'idea è che il processamento di tipo batch è intrinsecamente ad alta latenza: pertanto "se si deve processare un terabyte di dati tutti insieme, non sarà mai possibile eseguire la computazione in meno di un secondo adottando un pro-

cessamento di questo tipo". Diversamente un processamento di tipo stream elabora piccole porzioni di dati appena questi arrivano all'interno del sistema e prima che essi siano resi eventualmente persistenti attraverso la loro memorizzazione su file system, su database o su indici di ricerca [3].

Le piattaforme per il supporto al processamento batch sono state le prime a essere sviluppate. Tra queste, limitatamente alla comunità dell'open-source, Apache Hadoop è stata certamente una piattaforma "rivoluzionaria" perché ha messo a disposizione di tutti una struttura per la memorizzazione e per il processamento distribuito su cluster di computer a basso costo. Una curiosità storica interessante è che Hadoop nasce come costola del progetto Apache Nutch, un web crawler progettato per essere altamente scalabile. Quest'ultimo rappresenta un valido esempio applicativo dal quale risulta evidente la necessità di un sistema batch in grado di scalare orizzontalmente. In estrema sintesi un web crawler ha, infatti, l'obiettivo di scaricare e aggiornare velocemente e in modo più completo possibile i contenuti del Web realizzando una navigazione automatica dei link ipertestuali contenuti nelle pagine HTML[1]. Per far ciò in modo efficiente, è necessario sfruttare la capacità di calcolo di un elevato numero di macchine a cui distribuire le URL che vengono via via scoperte e memorizzare i risultati in strutture dati compresse e condivise. Hadoop offre tutti i meccanismi utili per eseguire queste operazioni con un'efficienza che cresce in modo lineare rispetto al numero di macchine coinvolte nel processo di crawling.

Oggi, a quasi 10 anni di distanza dalla data del suo primo rilascio (datato 2005), un intero ecosistema di strumenti per il trattamento di Big Data è stato costruito intorno ad Hadoop. Tra questi alcuni software, come ad esempio Apache HBase, Apache Hive, Apache Pig e Apache Zookeeper sono da considerarsi già sufficientemente maturi per utilizzi industriali; altri, come Apache Mahout e Apache Giraph, si stanno lentamente affermando.

Nell'ambito del processamento di tipo stream, la controparte di Hadoop è Apache Storm.

Storm è un framework per la definizione e l'esecuzione, in ambiente distribuito, di code per il processamento di flussi potenzialmente infiniti di dati. Questo implica che, su piattaforma Storm, la quantità e la frequenza di arrivo dei dati influenzano significativamente le scelte algoritmiche e architetturali. Storm è stato concepito per favorire lo sviluppo di applicazioni *near real time* e anche in questo caso un breve excursus storico ci permette di presentare un esempio applicativo particolarmente significativo: Storm nasce, nel 2011, come strumento utile alla realizzazione di prodotti finalizzati a misurare l'impatto delle aziende sui social network sia attraverso analisi di dati storici, sia per mezzo di analitiche real-time [4]. Più precisamente, la necessità era quella di sviluppare una piattaforma scalabile e facilmente programmabile per intercettare e garantire l'elaborazione di tutti i messaggi generati dagli utenti di Twitter (firehose). Nel giro di pochi anni Storm è stato adottato da numerosissime aziende[2] tra cui Twitter, Weather Channel, Groupon e Yahoo! e insieme ad altri framework, come Splunk, si è ritagliato un importante ruolo nel contesto dell'analisi dei Big Data in tempo reale.

La descrizione delle modalità di processamento dei dati appena trattata fornisce solo una delle numerose prospettive che dovrebbero essere considerate per presentare in modo esauriente il tema dei Big Data. Ciò nonostante dovrebbe essere sufficiente per lasciare intuire il livello di complessità dell'argomento, nonché il grado di maturità di alcune tecnologie che potrebbero contribuire a breve a migliorare la nostra comprensione della realtà attraverso l'analisi dei Big Data.

---

1 I risultato prodotto da un crawler rappresenta, tipicamente, l'input della fase di indicizzazione di un motore di ricerca grazie alla quale è possibile costruire strutture ad-hoc utili per eseguire ricerche sui contenuti indicizzati.

2 L'elenco delle principali aziende e le relative modalità di impiego di Storm è riportato all'indirizzo: https://storm.apache.org/documentation/Powered-By.html

Distribuire e Accedere ai Big Data

Quando i dati sono voluminosi, è importante che gli algoritmi definiti per la loro analisi si sforzino di mantenere i dati nella memoria principale perché accedere ai dischi comporterebbe tempi di calcolo troppo elevati. Secondo la legge di Moore, le prestazioni dei processori raddoppiano ogni 18 mesi circa, e dunque ciò che oggi è considerato "voluminoso" potrebbe non esserlo più in futuro. Però occorre ancora quasi un giorno per trasmettere in rete una decina di terabyte di dati letti da un singolo disco, quindi per elaborare anche pochi terabyte di dati occorre distribuirli su un numero molto elevato di macchine. Ad esempio per indicizzare tutto il Web, che secondo una stima approssimativa è costituito da più di 650 milioni di pagine attive, occorre costruire un indice compresso di almeno 300 terabyte che, per essere interrogato efficientemente, deve essere distribuito e replicato su un numero di macchine che non può essere inferiore alle centinaia di migliaia. Già nel 2003 il cluster di Google era composto da circa 15.000 macchine [6] con un numero di siti Web 15 volte più piccolo che nel 2014.

Occorre quindi distribuire opportunamente i dati su più macchine per diminuire i tempi di lettura e trasmissione, allo scopo di ridurre il più possibile i costi di comunicazione tra i nodi serventi e quello centrale. L'elaborazione dei dati deve essere però pensata ed effettuata il più possibile localmente, cioè sulle macchine in cui i dati risiedono. Anche gli algoritmi si devono pensare e realizzare seguendo uno stile, o meglio un modello di programmazione adeguato, introdotto da Google e chiamato MapReduce, che tenga cioè conto del problema della località dei dati [7]. La distribuzione dei dati in ingresso sulle macchine, e anche la loro duplicazione nel caso di eventuali rotture dei server, avviene in modo trasparente per l'utente. In fase di elaborazione dell'algoritmo, si dichiarano quali risultati intermedi dovranno essere prodotti localmente (operazioni di Map) e quali operazioni invece si dovranno effettuare per aggregare tali dati e produrre i risultati finali (operazioni di Reduce). Non tutti gli algoritmi si prestano a essere distribuiti così facilmente, ma le operazioni statistiche più semplici come quelle basate sulle frequenze, o anche alcuni algoritmi iterativi di Data Mining basati su semplici operazioni algebriche applicate a grandi matrici di dati, come ad esempio molti i modelli predittivi che fanno uso di tecniche di regressione possono essere descritti in termini di operazioni di MapReduce.

Sia il già citato Hadoop e il più recente Spark sono progetti open source di Apache, e sono utilizzati per gestire operazioni di tipo MapReduce su file distribuiti, Spark riesce a migliorare le prestazioni di Hadoop da 10 a 100 volte, a seconda che l'elaborazione dei dati avvenga su disco o in memoria centrale. Grazie a questi miglioramenti nelle prestazioni di Spark, il progetto open source di Data Mining per Big Data più promettente, Mahout di Apache, è già migrato da Hadoop a questa nuova piattaforma di calcolo distribuito.

Algoritmi di Data Mining

Secondo Forrester Research [8], il termine big per i dati significa semplicemente riuscire a far scalare gli algoritmi esistenti di Data Mining su grandi volumi di dati, ovvero:

"Per Big Data si intende l'insieme delle soluzioni software e hardware che permettano alle organizzazioni di scoprire, valutare e realizzare modelli predittivi, analizzando sorgenti informative molto grandi di dati al fine di migliorare le proprie performance e mitigare i rischi". Le soluzioni alle quali Forrester Research si riferisce sono quelle di Data Mining fornite dalle piattaforme di Business Intelligence e Analytics (BI&A). Gartner, come Forrester, ogni anno fornisce il quadrante magico delle migliori piatta-

forme di BI&A secondo prestazioni e visione del mercato. I leaders nel 2014 sono stati Tableau, Qlik, Microsoft, IBM, SAS, SAP, Tibco, Oracle e MicroStrategy.

Con l'affermarsi dei Big Data sta emergendo anche una figura professionale molto complessa, lo "scienziato dei dati" (Data Scientist). Il Data Scientist svolge un ruolo chiave nel processo d'interpretazione e analisi dei dati, e quindi di definizione e di realizzazione dei modelli predittivi. Uno scienziato dei dati è un'evoluzione estrema dell'analista dei dati. La sua formazione è multidisciplinare, in informatica, statistica e matematica, ma soprattutto deve essere in grado di conoscere bene e interpretare i problemi aziendali per poter proporre i modelli giusti per l'organizzazione.

La capacità di realizzare i modelli predittivi da parte di una grande organizzazione dunque richiede due componenti essenziali, non ancora percepite chiaramente distinte: da un lato deve esistere un team di IT che sia in grado di fornire tutte le soluzioni software e hardware per gestire ed elaborare i Big Data (la piattaforma di BI&A eventualmente su cloud), dall'altro occorre il Data Scientist necessario a realizzare e valutare i modelli elaborati e a comunicare i risultati all'organizzazione.

L'idea principale di una piattaforma di BI&A è che un Data Scientist scriva le sue funzioni matematiche o statistiche, lavori cioè su un piano strettamente logico, si disinteressi di come queste vengano distribuite operativamente, proprio come avrebbe fatto se il suo programma fosse stato eseguito su un singolo computer. Al momento il linguaggio di dichiarazione preferito dai Data Scientist si chiama R, ma anche Python ha una buona penetrazione d'uso. Ad esempio Tableau che è la piattaforma più performante tra quelle di BI&A, secondo il rapporto annuale di Gartner, ha realizzato un'integrazione forte con R. Purtroppo al momento non esiste una versione open source distribuita di R.

Il Data Scientist dunque scrive i suoi programmi nello stile di R. L'idea è cioè di disaccoppiare il linguaggio dichiarativo (front-end) dal problema della distribuzione dei dati (back-end). Questa idea è anche alla base del modello di business usata da EMC con la sua piattaforma proprietaria GreenPlum e il suo linguaggio di specificazione (Pivotal R) che implementa in modo distribuito alcune librerie di R (la versione su singola macchina è disponibile come open source). Questa piattaforma a differenza delle altre non è presente nel rapporto Gartner perché consiste in una piattaforma cloud (Platform-as-a-Service, PaaS) e compete con quella di Amazon, AWS.

R ha moltissime librerie ed è possibile trovare un qualsiasi algoritmo di Data Mining. Per questa ragione non è facile implementare tutte le librerie di R in forma distribuita. Se una piattaforma BI&A si potesse integrare fortemente con R, tutti gli script, una volta descritti in R, potrebbero essere facilmente eseguiti sulla piattaforma distribuita. Inoltre queste soluzioni potrebbero valere per sempre e non andrebbero riscritte. Dunque per chi è nella fase di investire sulla tecnologia, la scelta ora è tra:

- seguire il Magic Quadrant di Gartner, cioè investire su chi al momento si posiziona meglio;
- investire al suo interno sulle nuove tecnologie open source, valutando se sviluppare tecnologie proprietarie anche in attesa che il mondo open source diventi più maturo e ricco.

Del resto il quadrante magico di Gartner cambia molto rapidamente di anno in anno (a parte la presenza costante di alcuni big player). C'è da aggiungere che il giudizio degli utenti interpellati, che definiscono il quadrante magico, è molto spesso basato sull'usabilità e la visualizzazione dei risultati piuttosto che dalla reale dimensione dei dati elaborati. I volumi dei dati esistono come le piattaforme di BI&A, ma non ci sono ancora Data Scientist che siano in grado di indicare quali piattaforme e tecnologie di Data Mining utilizzare e cosa estrarre da tutti questi dati. È questo quello che in sostanza emerge nelle raccomandazioni del rapporto Gartner del 2013:

1. Valutate la vostra strategia sui Big Data, l'adozione e le relative priorità effettuate da altri operatori del settore, ma esplorate al di fuori del settore in cerca di casi d'uso innovativi e nuove applicazioni.

2. Allineate le iniziative sui Big Data alle decisioni di business. Individuate le priorità di business e le fonti di dati che possano portare una più profonda comprensione delle opportunità.
3. Cercate esperti di Big Data quando avete bisogno di aiuto per superare gli ostacoli su competenze, tecnologia, strategia o obiettivi di business. Investite in competenze di Data Scientist e in quelle sulle nuove tecnologiche.

Analisi o eliminazione della coda lunga

Finora abbiamo parlato dei Big Data in generale delle piattaforme e tecnologie abilitanti e cosa dovrebbero fare le organizzazioni per sfruttare queste nuove opportunità. Ora vediamo un aspetto più specifico legato al mining su Big Data, cioè quello della gestione delle vendite, dello stoccaggio e dell'esposizione fisica o online dei prodotti. In realtà questo problema di ottimizzazione delle vendite si ripresenta in modo del tutto simile in altri settori e ad altri problemi applicativi, quali la genetica, la biologia, la medicina o più in generale in settori dove l'informazione è molto dispersa tra moltissime entità, categorie e relazioni. Supponiamo che le frequenze delle categorie dei nostri dati si distribuiscano secondo una coda lunga, ovvero che molti dati confluiscano in poche categorie, mentre molte categorie si riempiano di pochi dati esemplificativi. Pensiamo, come esempi, alla distribuzione delle specie biologiche, alla distribuzione delle malattie in medicina, agli acquisti online o nei supermercati dei prodotti, alla memorizzazione dei risultati delle interrogazioni nei motori di ricerca (cache).

Lo spazio fisico limitato offerto dagli scaffali impone che una qualsiasi catena di supermercati selezioni alcuni beni che possano soddisfare i bisogni della maggior parte dei clienti e che allo stesso tempo siano in grado di massimizzare i profitti. In generale i prodotti che corrispondono alla parte alta della domanda sono quelli che vengono esposti. Al contrario i grandi siti di vendita online possono servire anche la coda lunga della domanda. In realtà, la coda lunga della domanda può avere un volume di affari equivalente o anche più remunerativo di quella alta, ma l'elaborazione di milioni di prodotti per milioni di utenti richiede una gestione dei dati e dei modelli predittivi di vendita molto complessi. I modelli dei prodotti sono sostituiti da nuovi e inoltre esistono in rete sul sito stesso o in altri le recensioni che possono modificare e orientare le decisioni stesse di acquisto dei clienti.

Anche problemi classici di Data Mining possono raggiungere dimensioni di complessità notevoli anche per i supermercati convenzionali, come ad esempio studiare la composizione dei carrelli della spesa per definire la composizione ottimale di prodotti da esporre negli scaffali e stabilire regole associative tra prodotti. Sarebbe sbagliato eliminare del tutto certi beni magari più ingombranti, meno costosi o meno richiesti se si stabilisse che molti clienti sono fedeli a molti prodotti di nicchia.

In medicina, in un modo del tutto simile, è importantissimo studiare le comorbidità, ovvero stabilire la presenza contemporanea di più malattie o di più condizioni croniche in un paziente. Il rilevamento delle comorbidità presenti in un paziente hanno ad esempio un impatto elevatissimo sulla salute dei cittadini e anche sui tempi medi di degenza negli ospedali.

Se si vuole analizzare la coda lunga allora non possiamo pensare di avere dei tempi di risposta ottimali e nello stesso forniti in tempo reale. Occorre indicizzare e elaborare tutti i dati e questo richiede di avere i dati memorizzati su file e dunque di effettuare un'elaborazione di tipo "batch".

Se non si vuole analizzare la coda lunga allora è possibile non elaborare o memorizzare tutti i dati. Qui si possono intraprendere due strade: ridurre drasticamente le dimensioni dei file eliminando i dati della coda oppure mantenere una struttura dati in memoria che si aggiorni real-time e contenga le frequenze delle categorie e delle relazioni più numerose.

Nel caso di un trattamento batch, essendo i dati molto sparsi tra le categorie ammissibili, lo spazio di occupazione della memoria e il tempo di calcolo sarebbero troppo dilatati rispetto ai dati realmente necessari. Occorre dunque ridurre la sparsità dei dati con tecniche di riduzione della matrice di rappresentazione. Gli algoritmi di riduzione però sono distribuibili con una programmazione di tipo MapReduce e sono tutti facilmente riconducibili al calcolo di algebra lineare delle matrici. Questo tipo di approccio batch è però praticabile se i dati già acquisiti non si modificano sostanzialmente nel tempo. Se volessimo ad esempio sapere quali siano gli argomenti più "trend" su Twitter nell'ultimo minuto allora non potremmo più seguire questo approccio e affrontare direttamente il problema dello streaming.

Velocità ovvero analizzare i dati in modalità streaming

"Gli investimenti negli ultimi 10 anni sono stati guidati principalmente da aziende orientate al cliente o dall' "Internet delle persone." I prossimi 10 anni saranno guidati da investimenti in applicazioni che utilizzano "l'Internet delle cose". La più rapida crescita del tipo di dati sono i flussi in tempo reale degli eventi, dei sensori e delle macchine, gli eventi generati dalle persone e dalle transazioni dei sistemi di business. Queste nuove applicazioni, combinate con intuizioni provenienti da altri nuovi tipi di dati e insieme a nuovi tipi di analisi, genereranno la prossima grande ondata di analisi degli investimenti e la trasformazione del business [7].

In presenza di un flusso massivo di dati non si ha tempo per indicizzare, recuperare dagli indici le informazioni aggregate, o analizzare i dati. In base alla memoria disponibile si stabilisce quale informazione in forma aggregata è possibile mantenere filtrando opportunamente i dati. Se non si fa in tempo a analizzare tutti i dati si stabilisce di prelevare un campione il più grande possibile dei dati. Dobbiamo ricordarci che è un flusso e quindi questo campione assume l'aspetto di una serie temporale piuttosto che di un insieme incrementale. Per questa ragione occorre mantenere in memoria una struttura di tipo scorrevole (sliding window) nella quale si contano gli elementi dell'insieme campionato nel tempo e dal quale è possibile chiedere con un certo errore statistico quali e quanti elementi ci siano da un certo istante di tempo in poi.

Per collocare i dati entranti nelle categorie monitorate o effettuare un confronto di similarità, non si ha tempo di analizzare il dato in dettaglio e per questo si devono usare quelle che in informatica di chiamano signature (cioè matrici ridotte dei dati realizzate mediante funzioni di hashing).

Per fare un esempio, con circa un Giga di memoria centrale sarebbe possibile mantenere una timeline giornaliera contenente 7 milioni di parole, quelle più frequenti, in un flusso di 150 milioni di tweets. È anche possibile esporre in tempo reale oltre agli andamenti, i grafici che mostrino il loro trend passato e futuro estratto con opportuni modelli predittivi.

Per altre applicazioni non aspettiamoci che ci sia qualcosa di già pronto per affrontare il nostro problema specifico, ma sarà il Data Scientist a indicare la strada migliore da intraprendere.

Orizzonte temporale

Fino al decennio scorso ogni innovazione tecnologica aveva sempre prodotto più lavoro a lungo termine. Negli ultimi anni il tasso di disoccupazione è cresciuto in tutto il mondo del 10% e quello della sottoccupazione del 20% e questo fenomeno è dovuto a un andamento che molti economisti, come Cowen Tyler, pronosticano come irreversibile: "Quello che sta accadendo è un incremento nell'abilità delle macchine di sostituire il lavoro intelligente umano, comunque vogliamo chiamare

queste macchine “AI”, “software”, “smartphone”, “disponibilità superiore di hardware e capacità di memorizzazione”, “sistemi integrati migliori” o una qualsiasi combinazione di queste”. La scuola economica della Oxford Martin School, come Carl Benedikt Frey, ha pronosticato che circa il 45% dei lavori attualmente esistenti negli Stati Uniti si estingueranno a breve e verranno sostituiti dalle macchine. La domanda chiave è quanto il nostro lavoro sia oggi complementare a quello della macchina, o se al contrario il computer possa fare meglio senza di noi. Anche la scienza, secondo Cowen, non sarà immune da questa rivoluzione tecnologica.

La scienza serve a fare predizioni, controllare l’ambiente, comprendere il nostro mondo, e dunque subirà i cambiamenti maggiori. Cominceremo a capire sempre meno la scienza che governa le nostre vite e il nostro lavoro. La capacità di analisi massiva e esaustiva dei dati, la possibilità di creare rapidamente associazioni e correlazioni esatte (senza errore statistico) sostituirà di fatto i modelli matematici più complessi. Ad esempio Google può utilizzare metriche molto semplici (edit distance) su sequenze anche molto lunghe di parole per suggerire o correggere quello che scriviamo e non usando più i linguaggi formali, come ad esempio le grammatiche generative di Chomsky. Analogamente quando si ha la possibilità di elaborare tutti i dati non si ricorre più a un campione ma si ha già l’universo, e dunque basta vedere cosa è più frequente, non calcolare ciò che sia più probabile.

Come i Big Data possono fare la differenza: una visione europea

Da diversi anni la Commissione Europea è convinta che dati saranno il centro della futura economia della conoscenza ritenendo che un buon uso dei Big Data possa portare nuove opportunità anche in settori tradizionali come i trasporti, la salute e i processi manifatturieri.

La Figura 1 mostra la visione della Commissione di come i Big Data possano effettivamente fare la differenza grazie al miglioramento di processi di elaborazione e di analitiche, in particolare saranno in grado di:

- generare nuovi prodotti e servizi basati sull’elaborazione della conoscenza in molti e diversi campi;
- aumentare la produttività attraverso nuove metodologie di business intelligence;
- dare impulso alla ricerca;
- fornire un contributo rilevante ad alcuni sfide sociali;
- incrementare l’efficienza del settore pubblico.

La Figura 1 ci dice che l’applicazione a settori della vita reale degli strumenti e metodi dei data analytics, insieme al cloud, alle reti veloci a larga banda e alla computazione ad alte prestazioni creerà centinaia di migliaia di nuovi posti di lavoro e con una crescita di 17 miliardi di euro nel 2015: per esempio, nella sola UK è previsto nel 2017 un aumento del 240% di specialisti di Big Data corrispondente a circa 69 mila posti di lavoro.

Come descritto nelle pagine dell’agenda digitale europea, le attività della commissione sono orientate sia alle strategie politiche che di ricerca.

In primo luogo ha elaborato una strategia, contenuta nella comunicazione sull’economia basata sui dati (Towards a thriving data-driven economy, COM(2014)442 final) in cui sono descritte le caratteristiche della futura economia dei dati con le conclusioni operative su come supportare e accelerare la transizione verso questo nuovo scenario.

L'attenzione è anche verso gli open data con:

- misure legislative sul riutilizzo delle informazioni del settore pubblico, come ad esempio la direttiva 2003/98/EC del Parlamento Europeo e del Consiglio con regole sia a livello nazionale che europeo sullo sfruttamento di tali dati;
- misure non legislative per stimolare la piena disponibilità dei dati pubblici,
- apertura del portale degli open data della EU (http://open-data.europa.eu/en/data/) e catalogazione dei portali disponibili a livello mondiale (http://open-data.europa.eu/en/data/).

Anche l'accesso aperto (open access) alle pubblicazioni scientifiche e dei risultati di ricerca fanno parte di questa strategia. Le proposte presentate in H2020 devono avere una sezione dedicata a come sono gestiti i dati prodotti dai progetti.

La fusione dei concetti Big Data e open data conduce agli open Big Data, anche rispetto ai quali la Fondazione Ugo Bordoni conduce attività di ricerca.

Big Data e H2020

L'altro versante di attenzione è quello dei bandi Horizon 2020 (H2020) che pongono grande enfasi sui Big e Open Data: nell'ambito del programma di lavoro ICT 2014-2015, esiste una call specifica per i Big Data (ICT 16, Big Data - research) anche se sono un tema ampiamente trasversale, presente anche nei bandi relativi alle Societal Challenges.

A conferma dell'interesse per il tema, la sessione sulla call ICT 16 allo scorso ICT Proposer Day (9-10 ottobre a Firenze) per la seconda call di H2020 è stata così affollata, con persone in piedi e sedute in ogni angolo disponibile, che hanno dovuto chiudere la sala: oltre le presentazioni da parte della Commissione, ci sono stati ben 26 interventi, tra i quali quello della FUB, per la presentazione di proposte e/o offerta di competenze.

L'impatto atteso dichiarato nel bando è molto significativo per spiegare cosa l'Europa si aspetta dai Big Data:

- possibilità di monitorare pubblicamente e quantitativamente il progresso nelle prestazioni e nell'ottimizzazione delle tecnologie per "very large data analytics" in un ecosistema europeo costituito da centinaia di aziende, fondamentale per la pianificazione industriale e lo sviluppo strategico;
- tecnologie per data analytics che siano realmente innovative per permettere analisi predittive e in tempo reale, con la possibilità di potere validare i risultati per mezzo di esperimenti rigorosi atti a testare la loro scalabilità, l'accuratezza e la fattibilità. Queste tecnologie devono essere pronte per gli innovatori e sviluppatori di sistemi su larga scala;
- le tecnologie sviluppate devono dimostrare di tenere il passo con la crescita dei volumi di dati e varietà attraverso esperimenti di validazione;
- dimostrazione del potenziale tecnologico e di valore degli open data europei che documentino il miglioramento delle posizioni di mercato e la creazione di posti di lavoro.

Gli interventi dei rappresentanti della Commissione (Francesco Barbato e Kimmo Rossi) hanno offerto alcuni spunti interessati su come indirizzare le attività di ricerca in questo campo seguendo le direttrici principali di H2020, ossia il passaggio da attività di ricerca e sviluppo ad attività di ricerca e innovazione per avvicinare il momento del passaggio al mercato.

PPP sui Big Data

Lo schema di Figura 2 suggerisce quattro passi per fare leva sul potenziale dei Big Data:

1. Investire nelle idee: trovare degli strumenti per stimolare la ricerca e la collaborazione tra tutti gli attori per aumentare l'impatto innovativo reale.
2. Infrastrutture per un'economia dei dati: mettere in connessione tutte le reti e infrastrutture disponibili per aumentare la sinergia e ottenere risultati effettivi.
3. Sviluppare le basi costruttive: individuare e incrementare tutti gli elementi che sono i blocchi costruttivi della nuova economia dei dati.
4. Fiducia e sicurezza (trust & security): rendere sicuro l'accesso, la memorizzazione e la condivisione dei dati basandosi su regole certe e condivise, garantendone l'affidabilità, la proprietà e il buon uso.

Senza entrare nel merito dei singoli aspetti, vogliamo soffermarci sul secondo punto del primo passo: la creazione di un PPP (Public-Private Partneship) che sono partenariati tra la Commissione (la parte pubblica) e aziende, università ed enti (la parte privata) con investimenti delle due parti che producono un effetto moltiplicativo per i processi di co-finanziamento e per l'impatto che generano. Per i Big Data, il 13 ottobre scorso è stato firmato un memorandum d'intesa per la costituzione di un partenariato pubblico-privato, il cui avvio è previsto per il 1° gennaio 2015, tra la EU e l'industria europea dei dati con un investimento di 2.5 miliardi di euro: l'obiettivo è quello di portare l'Europa in prima linea nella competizione globale sulla gestione dei dati.

L'accordo è stato firmato da Neelie Kroes, allora Vicepresidente della Commissione Europea, e Jan Sundelin, presidente della Big Data Value Association, formata da società quali ATOS, Nokia Solutions and Networks, Orange, SAP, Siemens e da istituti come il Fraunhofer e il centro di ricerca tedesco sull'intelligenza artificiale.

L'UE ha stanziato più di 500 milioni di euro di fondi del programma Horizon 2020 per 5 anni (2016-2020), cui dovrebbero corrispondere investimenti dei partner privati pari ad almeno il quadruplo (2 miliardi di euro).

Neelie Kroes ha dichiarato che "I dati sono il motore e il cardine dell'economia futura. Qualsiasi tipo di organizzazione ha bisogno di elementi costitutivi per migliorare i propri risultati, dalle aziende agricole alle fabbriche, dai laboratori alle officine".

Questo accordo di partenariato, a parità di altri già attivati nell'ambito dei programmi FP7 e H2020 come ad esempio quello per le reti di nuova generazione 5G e con il quale ha forti punti di contatto, dovrebbe convogliare in un'unica direzione gli sforzi del pubblico, dei privati e del mondo accademico: ricerca e l'innovazione per rivoluzionare la gestione dei Big Data in tutti i settori e per offrire servizi avanzati.

Tra i suoi obiettivi il PPP ha anche quello di sostenere "spazi d'innovazione" per mettere a disposizione ambienti sicuri per la sperimentazione sui dati sia privati che aperti, e serviranno da incubatori di imprese e da piattaforme per lo sviluppo di competenze e migliori pratiche.

Useremo le parole di John Quackenbush, professore di biologia computazionale e biologia a Harvard, che ha dichiarato: *"From Kepler using Tycho Brahe's data to build a heliocentric model of the solar system, to the birth of statistical quantum mechanics, to Darwin's theory of evolution, to the modern theory of the gene, every major scientific revolution has been driven by one thing, and that is data."* [9]

HOW CAN BIG DATA MAKE A DIFFERENCE?

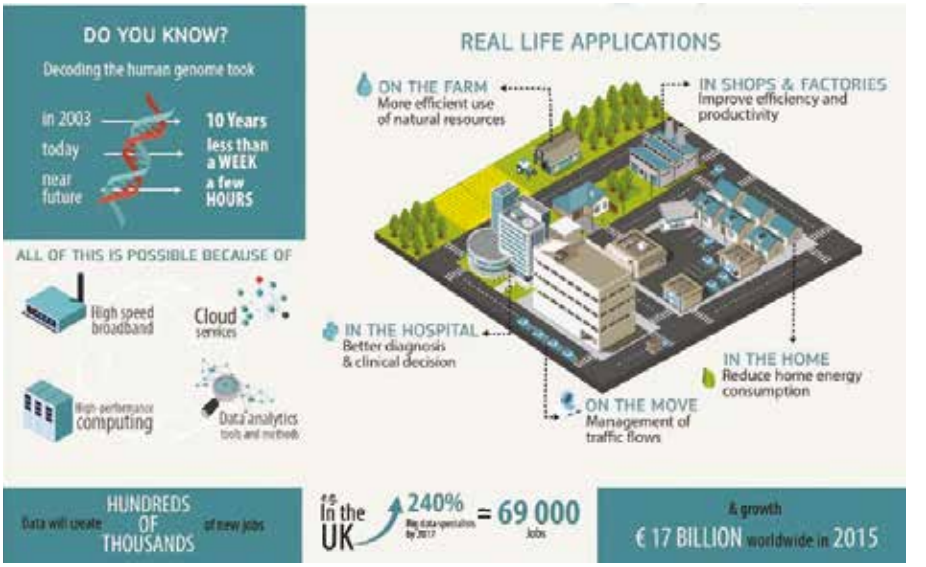

Figura 1. Come i Big Data possono fare la differenza? (dal sito EU Digital Agenda).

4 STEPS TO LEVERAGE THE POTENTIAL OF BIG DATA

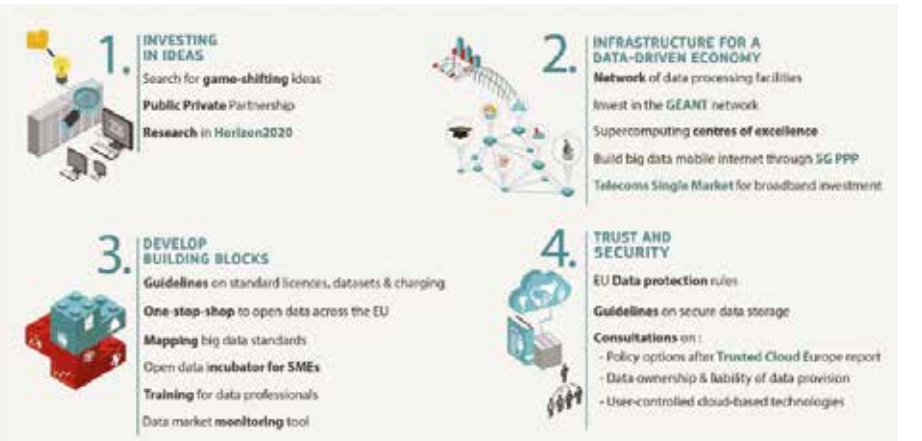

Figura 2. Come sfruttare il potenziale dei Big Data in quattro passi (dal sito EU Digital Agenda).

Bibliografia

[1] "Big Data, big era, big change!" **http://www.ilsole24ore.com/art/notizie/2013-11-03/big-data-big-era-big-change-083808.shtml** (2013)

[2] Gartner Web Site: http://www.gartner.com/it-glossary/big-data/

[3] Mike Barlow. Real-Time Big Data Analytics: Emerging Architecture. O'Reilly Media (2013).

[4] Nathan Marz. History of Apache Storm and lessons learned. http://nathanmarz.com/blog/history-of-apache-storm-and-lessons-learned.html

[5] Twitter Web Site: **https://about.twitter.com/company**

[6] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. 2003. Web Search for a Planet: The Google Cluster Architecture. IEEE Micro 23, 2 (March 2003), 22-28.

[7] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters

[8] Mike Gualtieri, The Forrester Wave: Big Data Predictive Analytics Solutions, Q1 2013

[9] Jonathan Shaw, Why "Big Data" Is a Big Deal. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal, March-April 2014.