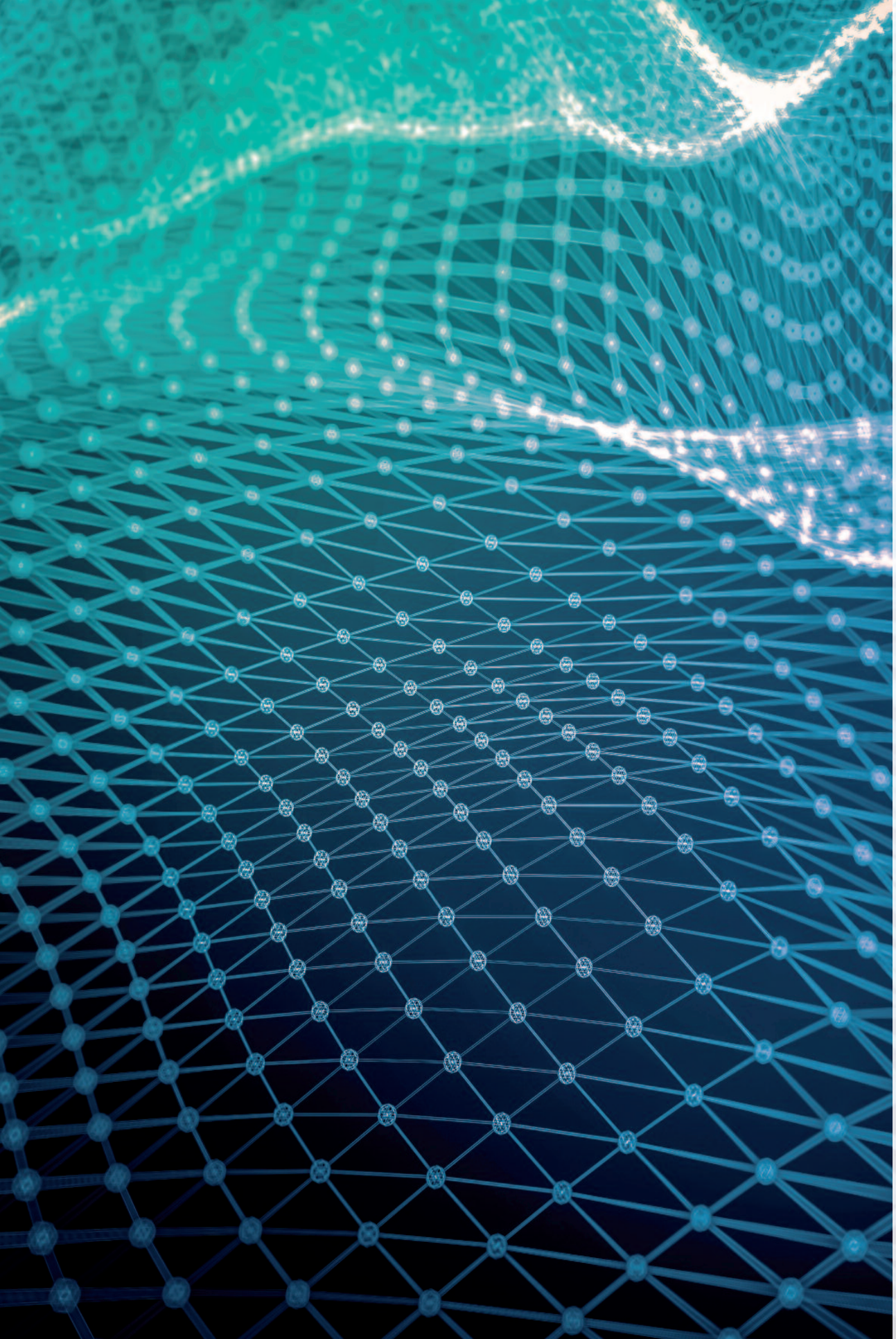


Origine dei Big Data

Cosimo Comella

Il metro quantitativo non è la chiave d'interpretazione dei Big Data, che sono metodologia, dati e paradigma computazionale. La quarta rivoluzione industriale è guidata dai dati ma abilitata da software che li organizzano e analizzano producendo valore aggiunto. Oltre ai benefici per lo sviluppo dell'economia dei dati, sono evidenti i profili d'interesse per l'intelligence, specie su fonti aperte e, in tale contesto, l'elaborazione di strategie big data efficaci in relazione ai diversi scenari operativi è un fattore competitivo per servizi d'informazione e schieramenti geopolitici. Lo sviluppo dei Big Data può avere ripercussioni su libertà e diritti fondamentali delle persone se non accompagnato da garanzie e da principi etici che li rendano compatibili con i valori delle società democratiche.

ORIGINE DEI BIG DATA

Nel dibattito pubblico sulle tecnologie dell'informazione e della cosiddetta «innovazione tecnologica», da alcuni anni ha ormai piena cittadinanza il tema dei Big Data, declinato da diverse angolazioni ma coerentemente proposto come il terreno su cui si potranno sviluppare negli anni a venire (con metodi e strumenti già da tempo disponibili) delle modalità di elaborazione dei dati innovative e applicabili a raccolte d'informazione di elevatissimo volume, non confrontabili per ricchezza, varietà e aggiornamento con i dataset normalmente gestiti informaticamente da imprese e organismi di varia natura.

Proprio le caratteristiche denotate dalle quattro «V» (volume, velocity, variety, veracity) costituiscono uno dei criteri definitori di queste raccolte di dati non omogenee, distribuite, accessibili in rete, non strutturate. Le aspettative di fruizione di un valore aggiunto informativo in questo tipo di elaborazioni sono assai elevate e, di conseguenza, per i benefici attesi sia in termini economici sia in termini di utilizzabilità sociale dei risultati.

«Per la prima volta abbiamo un'economia basata su una risorsa chiave che non solo è rinnovabile, ma cresce con il suo utilizzo. Non si corre il rischio che si esaurisca, ma di esserne sopraffatti».
Con questa efficace metafora John Naisbitt ha voluto rendere il carattere di giacimento informativo dei Big Data che, a differenza degli idrocarburi, non si esauriscono mai e si autoalimentano con l'uso che ne viene fatto. Mentre non è nuova la situazione in cui gli *usage data* di un sistema informativo entrino a far parte del patrimonio informativo dell'organizzazione, al pari dei normali dati di business, in questa visione i dati d'uso e gli altri metadati costituiscono il principale asset e contribuiscono in modo considerevole ad accrescere il carattere della raccolta dati nel senso delle quattro «V» sopra richiamate.
Fin qui abbiamo solo evocato il concetto di Big Data senza alcuna pretesa definitoria e, proseguendo in tal senso verso una maggiore comprensione del fenomeno, occorre capire quale sia la provenienza di questi dati. Certamente i servizi di rete e il traffico che comportano hanno una cospicua parte nella loro generazione: un noto rapporto di Cisco Networks, basato sull'osservazione del traffico di rete globale, porta a ritenere che negli ultimi due anni si siano generati più dati che nell'intera storia dell'umanità[1] e che nel 2019 verranno prodotti e trasmessi più dati dell'intero periodo 1984-2013, corrispondente all'era internet successiva all'evoluzione dell'originaria Arpanet e all'apertura della rete a operatori commerciali.
Nello stesso tempo, Google gestisce oggi più di 60.000 ricerche al secondo[2], ognuna delle quali, oltre a produrre dei risultati, genera anche ulteriori dati destinati a essere successivamente riutilizzati o a costituire la base per altre ricerche.
Nell'ambito delle reti sociali, la sola piattaforma Facebook ha un numero complessivo di utenti superiore a un miliardo e ottocento milioni, ma il dato più impressionante è quello relativo al numero medio di utenti giornalmente collegati da dispositivi mobili (in grado quindi di fornire informazioni arricchite da posizione, immagini, valori acquisiti da sensori di vario tipo): si tratta di più di 1,2 miliardi di persone (stime di febbraio 2017)[3], con una produzione di più di 3,5 miliardi di messaggi al giorno, mentre su YouTube vengono caricati ogni minuto video per la durata complessiva di 400 ore.

1. Cisco 2016.
2. Livestats 2017.
3. Facebook 2017.

La tecnologia I*nternet of Things* (IoT) è poi un'altra fonte di un volume elevatissimo di dati acquisiti da sensori di ogni tipo, spesso incorporati in dispositivi di varia funzionalità. Analogamente, la capacità di produrre dati si estende a tutti i dispositivi digitali come i sistemi *wearable*, i sistemi di trasporto intelligente, le smart house e le smart city. Occorre chiarire che non tutti i dati prodotti da sistemi embedded o gli usage data dei servizi di rete diventano automaticamente disponibili, dopo la loro generazione, per l'uso da parte di terzi. Tuttavia, v'è un crescente movimento verso la condivisione di risorse informative, nella consapevolezza che nell'economia della rete i comportamenti eccessivamente chiusi e protettivi non paghino e che sia più conveniente la cessione (di parte) del bagaglio conoscitivo delle aziende e delle organizzazioni, per metterlo a disposizione di altri in modo da innescare un circuito virtuoso che funga da stimolo a nuovi servizi.

CARATTERISTICHE DEI BIG DATA

La caratteristica peculiare dei Big Data non è la tipologia dei dati in sé, quanto la loro fruibilità per elaborazioni massive anche in contesti caratterizzati da elevata velocità di aggiornamento. Sono dati e metodologie assistiti da specifiche tecnologie informatiche per la raccolta e la costruzione dei bacini informativi.
Mentre a livello terminologico fanno la loro comparsa a metà degli anni Novanta, è dal 2000 in poi che si consolida una loro definizione che, seppur abusata, non è priva di efficacia, facendo riferimento a un insieme di caratteristiche sintetizzate dalle quattro «V».
A differenza di quanto avviene nelle tradizionali attività di *business intelligence* e di *data ware housing*, in cui tipicamente si fa ricorso a raccolte di dati strutturati, nel paradigma Big Data è fondamentale la possibilità di sfruttamento dei dati non strutturati da cui è possibile ricavare valore aggiunto informativo grazie a opportuni strumenti di machine learning.
L'assimilazione dei Big Data a grandi raccolte, cogliendone esclusivamente l'aspetto quantitativo (ancorché rilevante) come sola o prevalente connotazione, è fuorviante: d'altra parte, strumenti tipici del paradigma Big Data possono essere applicati anche a raccolte di dati di minor volume e disponibili pure su scala locale, seppure l'efficacia di una tale strategia sarà fortemente condizionata dal ridotto volume di informazioni disponibili.

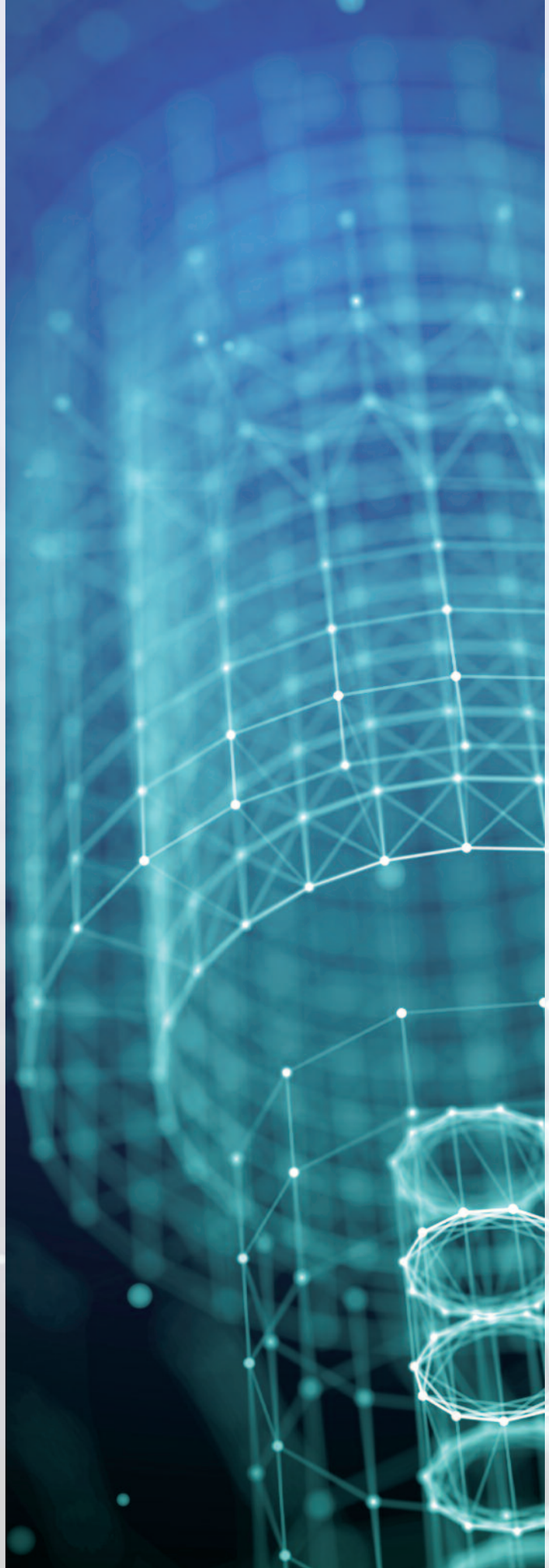
LA BIG DATA STRATEGY

Appare evidente che i Big Data non siano solo i dati o le raccolte di dati in sé, ancorché siano proprio questi a guidare le attività di analisi e ricerca: in risalto è la metodologia che, coniugando l'applicazione di strumenti di raccolta, analisi e di Data Mining alle vaste collezioni d'informazione digitale e dinamicamente generata, disseminata e accessibile tramite la rete, permette di sfruttarne il valore anche quando i dati di base siano dei meri sottoprodotti (in forma di metadati o usage data) derivati da altri processi di elaborazione dell'informazione.
Mentre il dibattito pubblico sui temi delle tecnologie e della società dell'informazione non è esente da semplificazioni, occorre chiarire come una Big Data strategy si debba basare necessariamente su metodologie e strumenti specifici, fondati su architetture di elaborazione peculiari per questo tipo di applicazioni: non bastano 'tanti dati' per fare i Big Data. In proposito, Victor Mayer-Schönberger e Kenneth Cukier, nel loro noto best seller sull'argomento hanno definito con il termine Big Data «l'analisi di dati su larga scala per estrapolare nuove indicazioni e creare nuove forme di valore [...] la vera rivoluzione non sta nelle macchine che elaborano i dati, ma solo nei dati in sé e nel modo in cui li usiamo»[4].
Le aspettative riposte sui Big Data e, più in generale, sull'elaborazione di dati pubblicamente disponibili perché messi a disposizione da fonti aperte debbono essere rapportate ai limiti che il quadro normativo contempla relativamente ai trattamenti di dati personali, nell'assunto che le grandi raccolte di informazioni, in molti casi, costituiscano delle vere e proprie banche dati nel senso attribuito a tale termine dalle norme in materia; oppure, che raccolte di notizie ritenute anonime, e quindi, non soggette alla disciplina di protezione dei dati personali, in realtà rechino informazioni suscettibili, se raccordate ad altri dati, di consentire l'identificazione di persone e lo svelamento di loro caratteristiche anche tra le più intime e riservate.

NUOVI STRUMENTI, NUOVE ARCHITETTURE

Le tecnologie Big Data sono costituite da nuovi strumenti e architetture finalizzate all'estrazione di valore da grandi volumi di dati connotati da varietà, grazie a capacità di raccolta veloce, ricerca e analisi.

4. Mayer-Schönberger – Cukier 2013.



Le ricerche sono guidate dai dati e sfruttano informazioni relative al contesto, potendo pervenire a tassi di accuratezza più elevati rispetto agli ordinari strumenti di business intelligence. In più, le informazioni di contesto via via arricchite con altri dati o con i risultati delle precedenti analisi comportano miglioramenti in termini di performance, analogamente a come le ultime tessere di un puzzle diventano più facili da collocare avendo davanti il mosaico quasi completamente composto, come osservano Ann Cavoukian e Jeff Jonas in un testo monografico[5] in cui affrontano il tema della *Privacy by Design* applicata ai Big Data.
Preliminarmente all'analisi dei dati, questi devono essere raccolti ed elaborati per estrarne informazione utile, in una fase di preprocessing e di preparazione dei dati. La loro caratteristica peculiare è la non omogeneità di tipo e la mancanza di assunzioni sulla loro struttura: i dati raccolti possono essere infatti di tipo strutturato, come tabelle di fogli elettronici o di data base, oppure semi-strutturati, come nel caso degli Xml in cui il formato dei dati può essere specificato insieme ai dati cui si applica. Ancora, possono individuarsi dati 'quasi-strutturati', come le raccolte di link ipertestuali oppure dati del tutto privi di struttura, come i contenuti testuali documentali o i file multimediali.
Tale varietà richiede nuovi strumenti per essere gestita secondo il paradigma dei Big Data: tra quelli più significativi occorre far menzione dell'ambiente software open-source Apache Hadoop – utilizzato per la memorizzazione di elevate quantità di dati tramite un filesystem ad alte prestazioni come Hadoop Distributed File System (Hdfs) – che consente elaborazioni parallele e distribuite grazie al modello di programmazione MapReduce, di cui Apache Hadoop fornisce una propria implementazione. Le origini di Hadoop sono rintracciabili nel motore di ricerca Nutch, sviluppato da Doug Cutting e Mike Caffarella nel 2004, che integrava MapReduce e il Google File System descritto in alcuni articoli pubblicati nel 2003[6]. Hdfs è dotato della capacità di distribuire i dati, allo scopo di sfruttare la capacità di elaborazione parallela di MapReduce; non si sostituisce ai filesystem tradizionali come ext3, ext4 o Xfs, ma si basa sui filesystem di ciascun volume di storage per gestire i dati memorizzati, suddividendoli in blocchi distribuiti su un cluster e dando luogo a una registrazione ridondante (almeno 3 copie di ciascun blocco distribuite nel cluster) per garantire la funzionalità del sistema in caso

5. Cavoukian – Jonas 2012.
6. Ghemawat – Gobioff – Leung 2003.

di guasti anche molteplici. I blocchi vanno registrati su sistemi fisicamente separati, e Hdfs ha la visibilità della struttura del cluster, riconoscendo le aggregazioni di sistemi a livello di rack, in modo da mantenere e ripristinare il livello di ridondanza appena rilevi un guasto, evitando anche che un guasto esteso a un intero rack possa avere effetti sulle performance complessive del filesystem. Oltre a garantire la *fault-tolerance*, il meccanismo di replica consente di determinare dinamicamente quale nodo debba elaborare (*map step*) un certo blocco di dati, assicurando pertanto un'efficace gestione delle risorse di elaborazione e un'ottimizzazione della performance.
Il paradigma MapReduce permette, sfruttando l'Hdfs, di suddividere un task complesso in una serie di task più piccoli eseguibili in parallelo, consolidando gli output parziali in un risultato finale. Il paradigma si basa su due distinte fasi: M*ap*, in cui si esegue un'operazione su una porzione di dati e si produce un risultato intermedio; R*educe*, in cui si aggregano (consolidano) i risultati parziali per produrre il risultato finale. A ogni passo di MapReduce si utilizzano coppie ordinate (key, value) come input e come output, che possono avere anche forme complesse. Per la creazione di programmi MapReduce si può ricorrere al linguaggio Java, usando ambienti interattivi di sviluppo (Ide). Ciascun programma consiste in tre file Java, uno ciascuno per il codice driver, il codice map e il codice reduce. Il programma Java viene poi eseguito ricevendo in input gli opportuni file Hdfs.

I RISCHI DERIVANTI DAI BIG DATA

L'elaborazione informatizzata dei dati pubblicamente disponibili è in grado di incidere in modo significativo sui diritti e sulle libertà delle persone e, perciò, va eseguita in un contesto di consapevolezza e di responsabile utilizzo anche in quei casi in cui, tramite le elaborazioni dei dati, si perseguano interessi pubblici di elevato livello.
I volumi di dati da trattare in elaborazioni Big Data sono enormi e largamente eccedenti, in genere, le capacità di storage di una singola organizzazione. Per questo motivo la predisposizione di strumenti di analisi presuppone quasi obbligatoriamente un ricorso a servizi di tipo cloud computing, che permettono una calibrazione delle risorse ottimizzata sui tempi di svolgimento di campagne di elaborazione e analisi e sui volumi di storage da riservare alla raccolta dei dati. Benché sia possibile organizzare delle attività *on premise*, non sempre si possono giustificare gli investimenti necessari e appare perciò più conveniente ricorrere a risorse *on demand*, sia per la fase di scouting che di raccolta, preprocessing e analisi dei dati.

A partire da dati disaggregati e non strutturati, i risultati delle elaborazioni sono tendenzialmente dei dati strutturati e organizzati, che rappresentano il valore aggiunto informativo ottenuto. Essi rivestono particolare delicatezza, sia per il soggetto che ha organizzato la campagna di analisi, che potrebbe non avere interesse a divulgarli, sia per la qualità delle informazioni prodotte, che possono avere in molti casi il carattere di dato personale in tutte le sue declinazioni (ivi compreso quello sensibile o giudiziario).

LE CAUTELE NELLA PUBBLICAZIONE DEI DATI

Il titolare di un trattamento secondo il paradigma Big Data deve per questo porsi il problema di proteggere adeguatamente almeno i dati personali ricavati dalle elaborazioni contro i rischi di accessi abusivi, di illecito trattamento, di comunicazione a terzi non legittimati e gli altri rischi impliciti in ogni trattamento informatizzato.
Il fatto che dati disponibili in forma di open data o ad altro titolo acquisibili da parte di un soggetto, ma ritenuti anonimi da chi li mette a disposizione, possano andare incontro a effetti di reidentificazione anche in presenza di accorgimenti presuntivamente mitiganti il rischio è cosa nota e scientificamente dimostrata: si pensi ai lavori seminali di Pierangela Samarati e Latanya Sweeney[7] sulla k-anonimity del 1998, preceduti nel 1997 dal noto esperimento in cui Sweeney dimostrò l'agevole possibilità di reidentificare un individuo (nel caso specifico, il governatore Weld dello stato del Massachusetts) attingendo a dei data set ritenuti anonimi e relativi a raccolte di record sanitari, rilasciati dalla Massachusetts Group Insurance Commission (Gic), unitamente a dati elettorali della città di Cambridge[8].
D'altra parte, è noto che anche la mera pubblicazione dei risultati di un'analisi con tecniche di Data Mining può lasciare spazio allo svelamento di dati personali, anche nel caso in cui i dati sorgente originari siano mantenuti riservati[9].

7. Samarati – Sweeney 1998.
8. Sweeney 1997.
9. Giannotti – Pedreschi, 2008.

Il tema della reidentificazione tramite l'interconnessione di dati pubblicamente disponibili è stato negli ultimi anni costantemente all'attenzione delle autorità europee di protezione dati, di organismi internazionali e anche del gruppo di coordinamento europeo art. 29 Data Protection Working Party (Wp29), istituito in conformità all'articolo 29 della direttiva 95/46/Ce che ha introdotto a livello comunitario la protezione dei dati personali. In particolare, il Wp29 ha stilato nel 2014 un interessante parere[10] sulle tecniche di anonimizzazione che ha influentemente ispirato i produttori di dati destinati alla pubblicazione online, o comunque indirizzato la diffusione e la ricerca di idonee modalità per mitigare i rischi di *single outing* di interessati, anche in presenza di accorgimenti basilari, quali la deprivazione dei data set pubblicati delle informazioni anagrafiche e di quelle comunque immediatamente identificative.
Pur nella complessità del compito, molto può essere compiuto da chi detiene le raccolte dati da cui vengono estratti gli open data spesso offerti alla fruizione in rete e, perciò, prestatisi naturalmente a elaborazioni di tipo Big Data. Tra le tecniche e gli accorgimenti per la mitigazione del rischio (di identificazione), che vanno sempre correlati al contesto e alle finalità, si annoverano la *randomization*, la *noise addition*, la *differential privacy*, la generalizzazione.
Per la riduzione della *linkability* dei dati è presa in considerazione la pseudonimizzazione, che successivamente è entrata a far parte delle definizioni e degli strumenti previsti dalla General Data Protection Regulation. Le tecniche di pseudonomizzazione più usate includono: la cifratura con chiave segreta; l'applicazione di funzioni hash anche arricchite da quantità di sicurezza aggiuntive per realizzare delle *salted-hash function* o delle *keyed-hash function* con chiave segreta memorizzata; la *tokenizzazione*, che comporta la sostituzione di porzioni identificative dei dati con valori non derivabili matematicamente dai dati originari.
Sempre nel 2014, il Wp29 è intervenuto con uno S*tatement* sui Big Data[11].
In linea con il lavoro del Wp29 è il rapporto pubblicato da Enisa nel 2015 sull'introduzione del paradigma *privacy by design* nel contesto dei Big Data[12], sfruttando le privacy enhancing technologies disponibili.
Sui rischi derivanti dai Big Data di notevole interesse, anche per i profili previsionali di scenario, è il Rapporto pubblicato nel 2015 dello European Data Protection Supervisor (Edps)[13], in cui si analizzano le sfide poste dalle nuove tecnologie di elaborazione massiva di informazioni, evidenziando che l'innovazione e la difesa dei diritti fondamentali non possono essere percepite come antitetiche e che le tecnologie non devono contrastare con i valori e i diritti alla base delle moderne democrazie.

10. Wp29 2014a.
11. Wp29 2014b.
12. Enisa 2014.
13. Edps 2015.

BIG DATA E INTELLIGENCE

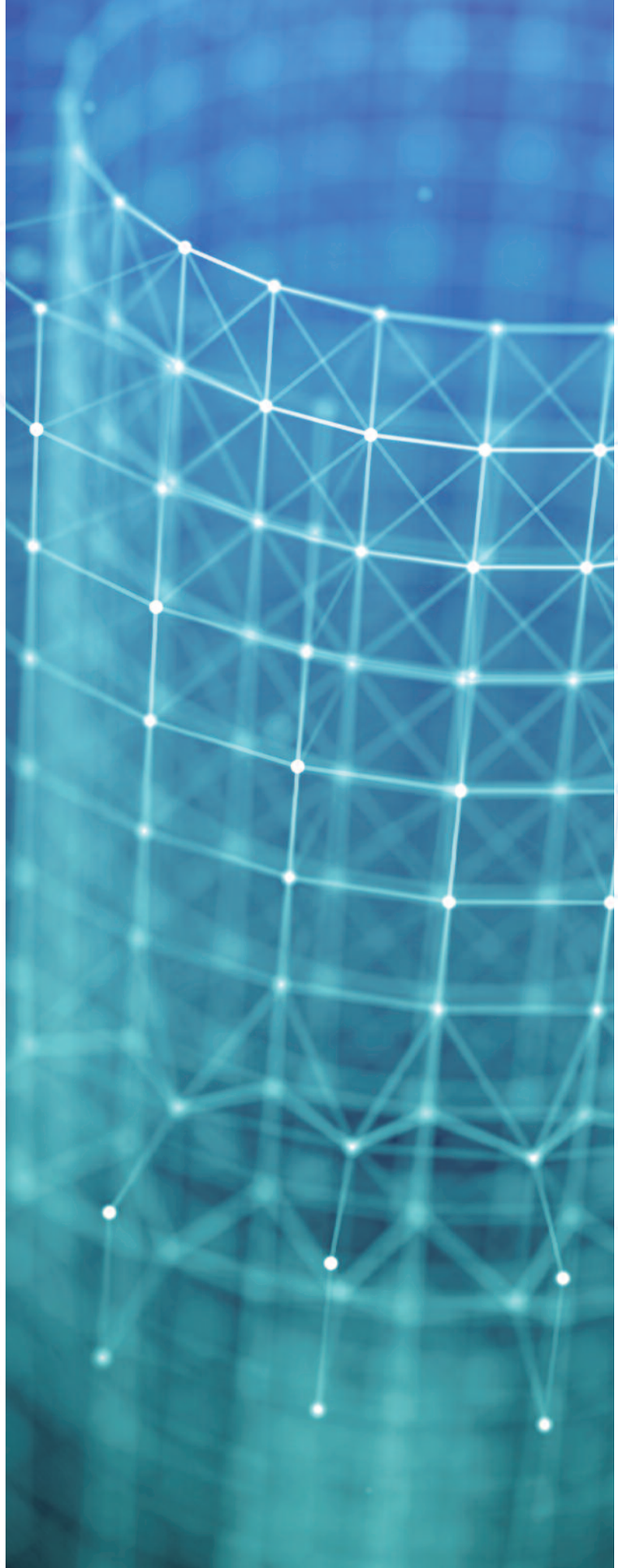
Le attività d'intelligence su fonti aperte trovano nel paradigma Big Data prospettive esaltanti: ciò che prima andava faticosamente cercato nelle pagine interne dei giornali, nell'ascolto di remote trasmissioni radio o tramite partecipazione diretta a eventi e riunioni, per restare a esempi immediatamente comprensibili, viene oggi offerto e 'portato in casa' dell'analista che si confronta, semmai, coll'arduo compito di filtrare, depurare, raccogliere e analizzare una quantità di flussi dalle più disparate fonti. Si è realizzato un passaggio da uno scenario del tipo *the analyst finding the data* a un altro molto differente, del tipo *the data finding the analyst*. Le più grandi aziende di informatica stanno facendo cospicui investimenti per lo sviluppo di strumenti Big Data anche per i settori della sicurezza e dell'intelligence, in cui sono evidenti i benefici di disporre di una tecnologia così potente[14], in particolare per attività di T*hreat* P*rediction and* P*revention*.
Le attività di E*xtract*, T*ransform and* L*oad* (Etl) delle informazioni sono grandemente facilitate dal ricorso agli strumenti di elaborazione dei Big Data, così come quelle di *real-time analytics* applicabili a flussi continui di dati, talvolta soggetti a vincoli di tempo reale, costituiti da immagini video, segnali elettromagnetici o acustici. Questi flussi di dati si caratterizzano per l'estrema volatilità e il loro valore informativo, per non perdere di efficacia, deve essere estratto e consumato in un ciclo di vita molto breve, passando in poche ore dall'utilità estrema all'irrilevanza.
Casi d'uso tipici sono i sistemi di analisi alimentati da reti di videosorveglianza, da sistemi d'intercettazione di comunicazioni elettroniche, da dati di traffico telefonico e telematico: in queste applicazioni il software di analisi può individuare pattern ricorrenti o comunque d'interesse evidenziando all'analista le relazioni tra i flussi di dati e fornendo le relative priorità, a vantaggio del decision-making.
Tutte le elaborazioni citate appaiono non innovative in quanto a finalità, ma lo sono del tutto se condotte nell'ambito di una Big Data strategy, in cui tutte le fasi di analisi e correlazione possono essere eseguite in uno stadio precocissimo di acquisizione dei dati, senza necessità che questi siano strutturati in data base e sistemi informativi tradizionali, la cui consultazione dovrebbe av-

14. Ibm 2013.

venire con interrogazioni e linguaggi strutturati. L'efficacia dell'approccio Big Data diventerà ancora più marcata al crescere della proporzione tra dati strutturati e non, a vantaggio di questi ultimi. Basti pensare all'esplosiva generazione d'informazioni dinamiche provenienti dalle piattaforme di social media e dai sensori in rete, che già surclassano quelle presenti in data base strutturati.
In questo immane lavoro – che nella metodologia Big Data vede una possibilità concreta di perseguire salti di livello nella capacità analitica e di discernimento relativi a persone, fatti e situazioni d'interesse – reali e concreti sono i rischi, da una parte, di far prevalere la 'volontà di sapere', la tentazione di oltrepassare certe soglie d'invalicabilità, e dall'altra di attivare eccessi e malcondotte anche in un contesto tecnico-sociale particolarmente qualificato e controllato. La protezione degli asset informativi elaborati giustifica in questi ambienti il ricorso a misure di sicurezza di livello elevato per limitare al minimo i pericoli di accesso incontrollato alle informazioni da parte di soggetti non legittimati. Anche in questa circostanza, la metodologia Big Data si presta a fornire soluzioni, permettendo la realizzazione di sistemi di *anomaly detection* per mitigare le minacce interne, rappresentate dai *data leakage*, dal sabotaggio di infrastrutture o di asset informativi, o per rilevare i percorsi di propagazione del malware o di infiltrazione ed esfiltrazione d'informazioni in seguito ad attacchi ai sistemi e alle reti.

CONCLUSIONI

Lo sfruttamento delle possibilità offerte dai Big Data richiede lo sviluppo di una vera e propria Big Data strategy che permetta alle organizzazioni di focalizzare gli obiettivi e di individuare gli strumenti, pianificando investimenti in mezzi e risorse umane. Accanto alle attività programmate e mirate a target specifici, la tecnologia favorisce la definizione di campagne di esplorazione di dati e flussi dinamici con un approccio ispirato alla *serendipity*, in cui le finalità ultronee nell'utilizzo dei dati si affiancano a quelle primarie, permettendo la scoperta di relazioni non sospettate tra i dati e l'emergere di nuovi fenomeni possibilmente d'interesse. In generale, una Big Data strategy deve essere predisposta nella piena consapevolezza dei limiti, fissati anche dalle norme in materia di protezione dei dati personali, alle analisi conducibili e ai risultati ottenibili, per evitare di ledere diritti e libertà fondamentali delle persone.



Nel contesto delle attività di intelligence, l'approccio Big Data può fornire strumenti di elevatissima efficacia, il cui utilizzo deve avvenire in un contesto di rafforzate garanzie e guidato da elevato senso etico, anche etimologicamente inteso, e perciò ispirato al migliore perseguimento dei fini istituzionali nel rispetto dei valori e delle libertà che caratterizzano il nostro sistema democratico.

BIBLIOGRAFIA

Article 29 Data Protection Working Party, *Opinion* 05/2014 *on Anonymisation Techniques*, Brussels 2014a.
Article 29 Data Protection Working Party, *Statement on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the* Eu, Brussels 2014b.
A. Cavoukian – J. Jonas, *Privacy by Design in the Age of Big Data*, Ontario 2012.
Cisco 2016, *10th annual Cisco Visual Networking Index* (Vni) *Forecast*: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf> [02-05-2017].
Enisa, *Privacy and Data Protection by design*, 2014.
European Data Protection Supervisor (Edps), *Opinion* 7/2015, *Meeting the challenges of Big Data*, Brussels 2015.
European Data Protection Supervisor (Edps), *Opinion* 8/2016, *Coherent Enforcement of Fundamental Rights in the Age of Big Data*, Brussels 2016.
Facebook 2017, *Facebook Reports Fourth Quarter and Full Year* 2016 *Results*: <https://s21.q4cdn.com/ 399680738/files/doc_financials/2016/Q4/Facebook-Reports-Fourth-Quarter-and-Full-Year-2016-Results.pdf> [02-05-2017].
S. Ghemawat – H. Gobioff – S. Leung, *The Google File System*, *19th Acm Symposium on Operating Systems Principles*, Lake George, New York 2003.
F. Giannotti – D. Pedreschi, *Mobility, Data Mining and privacy*: A *Vision of Convergence*, in *Mobility, Data Mining and Privacy*, Springer-Verlag Berlin Heidelberg 2008, pp. 1-11.
Ibm, *Big Data for the intelligence community*, 2013.
Livestats 2017: <http://www.internetlivestats.com/one-second/#google-band> [02-05-2017].
V. Mayer-Schönberger – K. Cukier, *Big Data*: A *Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt, 2013.
P. Samarati – L. Sweeney, *Protecting Privacy when Disclosing Information*: *k-Anonymity and its Enforcement through Generalization and Suppression*, Technical Report, Computer Science Laboratory, Sri International 1998.
L. Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, «Journal of Law, Medicine and Ethics» (1997) 25, pp. 98-110.