BIG DATA

L'umanità sta accumulando insiemi enormi di dati, ma la loro gestione è un problema non banale e ancora non del tutto risolto.

Angelo Gallippi

Nel marzo 2012 il governo statunitense ha deciso d'investire 200 milioni di dollari per lo studio dei *Big Data*, ossia di quegli insiemi di dati talmente grandi che non possono venire trattati in modo adeguato ed economicamente conveniente dalla maggior parte dei sistemi informatici tradizionali. Nello studio sono coinvolte agenzie quali la National Science Foundation, i National Institutes of Health, il Dipartimento della Difesa, la DARPA (Defense Advanced Research Projects Agency), il Dipartimento dell'energia e la U.S. Geological Survey. L'autorevole gruppo di ricerca statunitense Gartner li ha inseriti tra i dieci problemi più importanti per il 2012, mentre il World Economic Forum, in un Rapporto pubblicato lo scorso anno, ha classificati i *Big Data* in una nuova categoria di risorse economiche, al pari di una materia prima.

Le "tre v"

Una loro caratterizzazione più precisa è quella che fa ricorso alle "tre v": volume, velocità e varietà, in quanto si tratta di enormi volumi di dati, che possono venire acquisiti a grande velocità e constano in una notevole varietà di tipi e formati. Alcuni aggiungono alle precedenti la quarta "v" di "variabilità", dovuta alla necessità di modificare continuamente le strutture di dati esistenti. Per quanto riguarda il volume, la stessa dicitura *Big Data* indica in realtà volumi caratterizzati da ordini di grandezza molto diversi tra loro, per i quali è opportuno usare i multipli del Gigabyte (GB): il Terabyte (TB) uguale a 1.000 GB, il Petabyte (PB) uguale a 1.000 TB, l'Exabyte (EB) uguale a 1.000 PB e lo Zettabyte (ZB) uguale a 1.000 EB.

Per esempio, mentre i dati d'interesse di una società potrebbero andare da centinaia di GB a centinaia di TB, il più capace disco fisso esistente, realizzato nel 2011 da IBM assemblando 200mila dischi tradizionali e facendoli lavorare come un'unica unità, memorizza 120 PB e gli oltre 4 miliardi di ore di video guardati ogni mese su YouTube occupano circa 25 PB. Secondo una stima di Google, i dati prodotti ogni due giorni dal genere umano ammontano a 5 EB (tanti quanti quelli prodotti fino al 2003); perciò tutti i file complessivamente memorizzati nel mondo (stimati in 500 milioni di miliardi) occupavano, alla fine del 2012, un totale di 8 ZB. In altri termini, considerato che nel dicembre 2012 l'intera Library of Congress degli USA aveva una dimensione di 330 TB (cresce al ritmo di 5 TB al mese), il pianeta memorizza attualmente l'equivalente di 24 milioni di Library of Congress. Ovviamente a tali cifre contribuiscono anche i dati prodotti dalla progressiva conversione in dati digitali dei contenuti di media tradizionali quali immagini, filmati, brani audio, testi.

La grande parte dei dati scambiati nel mondo proviene dalle reti di telecomunicazioni, il cui traffico cresce a ritmo esponenziale, anche per il rapido incremento (20 per cento annuo) del numero di utenti di smartphone: 281 PB nel 1986, 471 PB nel 1993, 2,2 EB nel 2000, 65 EB nel 2007, e si stima che il traffico su Internet raggiungerà i 667 EB annuali nel 2013.

Per quanto riguarda velocità e variabilità, i flussi di dati che provengono in modo continuo e automatico da sensori termici, microfoni in ascolto dei movimenti in un'area protetta, videocamere che scrutano un volto in una folla o registrano le targhe automobilistiche in un sistema Tutor e transazioni bancarie vanno analizzati in tempo reale per suggerire azioni che abbiano senso: attivare lo spegnimento di un incendio, allertare il personale di sicurezza, avvisare le autorità di polizia, multare gli automobilisti indisciplinati, prevenire una frode.

Dalla scienza al business

Fino a pochi anni fa i *Big Data* potevano venire trattati in modo conveniente soprattutto dalla ricerca scientifica (astronomia, fisica delle particelle, geofisica, studio dell'atmosfera, genomica) e per finalità militari (soprattutto sorveglianza ambientale).

Tuttavia, alla fine dello scorso decennio i progressi compiuti da diversi settori della Information Technology (IT) e la diffusione del *cloud computing* hanno abbassato i costi della potenza di calcolo e della memorizzazione, consentendo il trattamento dei *Big Data* anche a pubblica amministrazione e grandi aziende: la prima allettata dalla promessa di risultati interessanti soprattutto nei settori strategici dell'evasione fiscale e delle prestazioni sanitarie. L'analisi approfondita

dei Big Data potrebbe consentire di recuperare miliardi di euro dalla scoperta di chi non paga le tasse e di chi le paga poco, dal controllo di quanti richiedono indennità di disoccupazione e risarcimenti per infortuni sul lavoro, dal passaggio da un sistema sanitario che interviene per curare a uno in grado di prevedere e prevenire le malattie, attraverso una valutazione dell'efficacia delle cure mediche. Si stima che, se usati in modo opportuno, i Big Data potranno fare risparmiare al sistema sanitario americano 300 miliardi di dollari l'anno e al settore pubblico europeo 250 miliardi di euro.

Per le grandi aziende la capacità di gestire i *Big Data* potrebbe accrescere la produttività e aumentare i margini operativi fino al 60 per cento, migliorando notevolmente i vantaggi offerti da tecniche commerciali applicate da decenni quali *data mining* (estrazione e analisi dei dati), *business intelligence* (raccolta e analisi di informazioni aziendali strategiche) e soprattutto *predictive analytics*, una metodologia di analisi che cerca schemi significativi nei dati per ottenerne una conoscenza di valore, quindi suggerisce le azioni o le decisioni da prendere. Impiega modelli descrittivi e predittivi e, comportando calcoli pesanti, utilizza approcci e algoritmi tipici di statistica, programmazione informatica e ricerca operativa.

Mentre sono decenni che le aziende prendono decisioni gestionali basandosi sui dati transazionali memorizzati in database relazionali, e quindi strutturati, negli anni recenti si sono venute formando autentiche miniere di dati non tradizionali e poco o per nulla strutturati, generati da blog e social media e consistenti in e-mail, testi, immagini, video, audio, *spreadsheet*. A questi si aggiungono i dati generati da macchine, contatori intelligenti e reti di sensori, i quali hanno consentito di collegare a Internet gli oggetti più svariati: non solo smartphone e tablet, ma anche stampanti, sveglie, impianti di elettricità, gas e trattamento dell'aria, caffettiere e serbatoi di automobili, tanto per citare i principali oggetti che già comunicano dati su se stessi e utilizzano informazioni aggregate da altri, prefigurando il cosiddetto "Internet delle cose". D'altra parte, i sensori hanno trasformato molti produttori in società di servizi, dato che consentono di monitorare un prodotto per stabilire se ha bisogno di riparazioni prima di rompersi: per esempio, la BMW usa i dati dei sensori per avvertire i suoi clienti quando devono fare la manutenzione delle loro vetture. Altra fonte di *Big Data* sono i dispositivi a radio frequenza RFID, che consentono l'identificazione automatica di oggetti, persone e animali, e sono stati venduti fino al 2012 in oltre 15 miliardi di esemplari, dei quali circa 4 miliardi nel solo 2012. Gli impieghi prevalenti sono come etichette in capi di abbigliamento (1 miliardo), ticket di transito (500 milioni) e identificazione di animali (294 milioni), con una crescita annua del 30 per cento.

Il grande valore potenziale dei dati non strutturati, la cui crescita contribuisce per l'80/90 per cento al totale complessivo, interessa un numero crescente di imprese, che pensano di includerli nella propria analisi di *business intelligence* accanto a quelli tradizionali per prendere decisioni gestionali. Una recente inchiesta di Capgemini ha evidenziato che nei processi in cui hanno applicato un approccio analitico sui *Big Data* le aziende hanno ottenuto un miglioramento medio del 26 per cento delle prestazioni rispetto ai tre anni precedenti, e prevedono un ulteriore miglioramento del 41 per cento nei prossimi tre anni. Le informazioni sono ormai ritenute il quarto fattore di produzione, importante quanto il capitale, la forza lavoro e le materie prime. Ne consegue che in futuro le organizzazioni in grado di sfruttare i Big Data, per esempio applicando tecniche avanzate di analisi predittiva in tempo reale, sopravanzeranno quelle che non lo sono. Analogamente, in una ricerca di GigaOm Pro, il 77 per cento degli intervistati dichiara di avere allocato un budget per progetti Big Data, e il 61 per cento che potrebbe utilizzare un fornitore di servizi per progetti di questo tipo nei prossimi 12/18 mesi, anche se il 51 per cento degli interpellati si dimostra comunque preoccupato per la sicurezza dei dati, e il 34 per cento esprime perplessità sui possibili costi. Per due terzi degli intervistati la raccolta e l'analisi dei dati è alla base della strategia aziendale e del processo decisionale quotidiano, in particolare per le aziende dei settori dell'energia e delle risorse naturali, dei servizi finanziari, farmaceutici e biotecnologici. Nove dirigenti su dieci ritengono che le decisioni prese negli ultimi tre anni sarebbero state migliori avendo a disposizione tutte le informazioni necessarie.

Vengono così ripensate le strategie di memorizzazione, gestione e *analytics* dei dati, affidando agli esistenti sistemi basati su tecnologie meno recenti (cosiddetti *legacy*) specifici carichi di lavoro ad alto valore e basso volume, ma affiancandoli progressivamente con prodotti specifici per alti volumi, che ottimizzano la gestione dei dati ponendo i carichi di lavoro *Big Data* nei giusti sistemi. Tuttavia le risorse necessarie per catturare e organizzare una grande varietà di dati da differenti sorgenti e analizzarli con facilità in modo da ottenerne un effettivo valore commerciale non sono ancora adeguate. Un rapporto McKinsey del maggio 2011 indica che entro il 2018 gli Stati Uniti si troveranno ad affrontare una carenza di personale con competenze approfondite in tema di analisi dei dati che riguarderà da 140mila a 190mila posizioni di lavoro, mentre mancheranno circa 1,5 milioni di manager e analisti con le competenze adeguate per analizzare i *Big Data* e ricavarne decisioni efficaci. D'altra parte, non sempre la gestione dei *Big Data* è considerata una priorità dai massimi livelli decisionali aziendali, e in molte aziende, soprattutto manifatturiere, non esiste ancora una "cultura *Big Data*".

Analisi dei comportamenti

I dati considerati più preziosi sono quelli relativi alle attività di *business* (vendite, acquisti, costi) e all'andamento dei punti vendita; sono ritenute utili anche le informazioni sui clienti, quali e-mail e profili sui social media. La loro analisi viene applicata correntemente per descrivere, prevedere e migliorare le prestazioni aziendali, in particolare per quanto riguarda gestione delle decisioni, vendite al dettaglio, scorte di magazzino, ottimizzazione dei numeri identificativi degli articoli e del marketing, dimensionamento e ottimizzazione della forza vendita, modellizzazione di prezzi e promozioni.

Altri importanti ambiti di *analytics* di *Big Data* sono il Web, i rischi nel credito, la previsione di frodi e la scienza predittiva. In particolare, l'analisi dei sentimenti degli utenti dei social media e l'estrazione di informazioni commerciali dettagliate dai contenuti sono i due impieghi dei *Big Data* responsabili della percentuale maggiore delle nuove spese nel settore IT: ben il 45 per cento annuo ossia, secondo le stime di Gartner, 28 miliardi di dollari nel 2012 e 34 miliardi nel 2013. Cifre spese in massima parte per adattare i sistemi informativi tradizionali alle nuove esigenze poste dalla elaborazione dei *Big Data*: dati generati da macchine, dati sociali, dati largamente variati, velocità imprevedibili, eccetera, mentre nel 2012 per l'acquisto di software sono stati spesi appena 4,3 miliardi di dollari.

L'esempio più recente e clamoroso di analisi dei comportamenti è costituito dal software di *analytics* estrema Riot (*Rapid Information Overlay Technology*) della statunitense Raytheon, il quinto fornitore mondiale del settore della difesa. Il programma, non in commercio, "traccia" gli utenti dei social network quali Facebook, Twitter e Linkedin, integrando i dati archiviati – che comprendono amicizie, post, foto e luoghi visitati – con quelli di localizzazione GPS ricavati dalle applicazioni Latitude e Foursquare. Ne risulta un quadro dettagliato della vita di un soggetto che permette anche di fare previsioni sulle sue azioni future, per esempio dove si troverà e chi incontrerà (ricordiamo che una ricerca del Massachusetts Institute of Technology ha mostrato che l'informazione sulle persone con cui un soggetto comunica con maggiore frequenza può fornire indicazioni sul suo orientamento sessuale). Riot ha subito suscitato le preoccupazioni dei difensori della privacy on line, anche se non è certo il primo prodotto del genere *spyware*: l'anno scorso l'FBI annunciò lo sviluppo di un'applicazione per scoprire azioni di aggiotaggio attraverso i social media, mentre la National Security Agency statunitense sicuramente usa e progetta software avanzati per analizzare l'enorme mole di dati - l'equivalente di una Library of Congress ogni sei ore - che raccoglie attraverso Echelon, il controverso sistema mondiale d'intercettazione satellitare sviluppato dai governi di Australia, Canada, Nuova Zelanda, Regno Unito e Stati Uniti. Altre applicazioni sono meno invasive della sfera privata: la città tedesca di Colonia ha avviato un progetto pilota di previsione del traffico, che raccoglie in tempo reale i dati da più di 150 stazioni di monitoraggio e 20 videocamere sulle strade, autostrade e punti di confluenza notoriamente problematici, riuscendo a prevedere il volume e il flusso dei veicoli con un'accuratezza di oltre il 90 per cento e un anticipo fino a 30 minuti.

Ma l'analisi dei comportamenti attraverso lo studio dei *Big Data* è effettuata anche con finalità commerciali: l'applicazione Gateway di MicroStrategy consente di combinare la visione aziendale di un'impresa con il profilo di un cliente su Facebook, mentre il rivenditore al dettaglio americano Williams-Sonoma usa la conoscenza dei suoi 60 milioni di clienti per produrre differenti versioni del suo catalogo. I "motori di raccomandazioni" suggeriscono prodotti complementari a quelli acquistati in base all'analisi predittiva di vendite incrociate, aumentando la dimensione media degli ordini: la società di e-commerce Amazon dichiara che il 30 per cento delle vendite sono generate dal suo motore di raccomandazioni ("Suggerimenti dell'editore"). Le compagnie di assicurazione cominciano a monitorare gli stili di guida dei clienti per offrire tariffe basate sulla loro prudenza (o imprudenza) anziché sull'età o il sesso, mentre la catena di supermercati britannica Tesco raccoglie ogni mese 1,5 miliardi di dati sugli acquisti e li usa per calibrare prezzi e promozioni, esattamente come da anni molti siti Web di e-commerce raccolgono i click dei visitatori e correlano le loro informazioni demografiche con i comportamenti di acquisto per proporre offerte commerciali vantaggiose.

Il "marketing basato sulla localizzazione" è attuato, tra gli altri, dalla catena britannica Starbuck, una delle 130 società che usano la piattaforma di localizzazione mobile Placecast per tracciare il milione di propri clienti che hanno accettato di ricevere offerte personalizzate quando si trovano nelle vicinanze di un negozio della catena. Ma sono innumerevoli le iniziative commerciali che usano gli smartphone come "sensori" della posizione dei proprietari per suggerire articoli di probabile interesse acquistabili nelle vicinanze, proporre sconti su articoli venduti in zona, inviare un voucher utilizzabile in un negozio.

I prodotti

Per quanto riguarda l'hardware, la elaborazione dei *Big Data* può sfruttare diverse evoluzioni tecnologiche fondamentali conseguenti a innovazioni quantitative che, per gli ordini di grandezza coinvolti, hanno comportato differenze qualitative. Ne sono esempi le CPU costituite da numerosi *core*, capaci di operare ad alta velocità in parallelo e spesso con significative riduzioni nel consumo energetico; la quantità di memoria RAM disponibile a basso costo; la capacità dei dischi fissi di ultima generazione (i cui tempi di accesso non sono però cambiati in modo apprezzabile); la diffusione delle *flash memory*; la larghezza di banda e la latenza ridotta delle nuove infrastrutture di rete, che superano le prestazioni di un disco rigido locale. Tutto ciò offre un potenziale enorme per innovazioni nel settore del *file serving*, dove però il software non è riuscito a stare al passo con i progressi dell'hardware, ed è necessaria una vera rivoluzione nel modo in cui vanno progettate le tecnologie dei *file system* distribuiti. Gli attuali prodotti sono sviluppati soprattutto da grandi società: si calcola che Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC e HP abbiano investito finora oltre 15 miliardi di dollari in società di software specializzate nella gestione dei dati e in *analytics*. Questo settore industriale vale nel suo complesso oltre 100 miliardi di dollari e cresce di almeno il 10 per cento l'anno: circa il doppio dell'intero comparto del software.

Nel caso dei *Big Data* strutturati cominciano oramai a prendere piede diverse applicazioni che offrono soluzioni quanto meno soddisfacenti. Le attuali tendenze per coprire le necessità crescenti evolvono, al di fuori dei tradizionali sistemi di gestione di database transazionali, verso tecnologie alternative quali NoSQL (nella quale rientra il database open source MongoDB) e NewSQL, che consentono di fare crescere la scala dei dati analizzabili e gestibili, aiutando ad esaminare larghi insiemi di dati in strutture non tradizionali (alberi, grafi o coppie a valore di chiave anziché tabelle). Altri cambiamenti stanno avvenendo nel trattamento dei dati delle "cose", dove emergono architetture a elevate prestazioni e altamente scalabili; essi comprendono elaborazione parallela, collegamenti in rete ad alta velocità e memorizzazione a elevata velocità di accesso, che aiutano ulteriormente a elaborare grandi volumi di dati a sostenuti tassi di MB al secondo.

Molto più delicata è invece l'area, a più rapida crescita, dei dati non strutturati (file), perché i produttori di sistemi dominanti (EMC, NetApp) sono ancorati a prodotti e modalità di vendita, installazione e servizio tradizionali, e hanno difficoltà a spostarsi in direzioni più consone alle nuove esigenze. Tali fornitori offrono sistemi monolitici che presentano diversi svantaggi: si possono fare crescere solo moltiplicandone il numero, il che complica la gestione e l'accesso ai dati, che finiscono in contenitori separati; inoltre non coprono i bisogni di chi parte da dimensioni piccole e vuole crescere gradualmente, mentre ancorano l'utente al singolo fornitore. In generale si adattano soluzioni progettate trent'anni fa a problemi di scala enormemente superiore, facendo aumentare la scala dei problemi. In definitiva, sono rari i prodotti innovativi realmente capaci di risolvere problemi

di scala inimmaginabile in passato e oggi comuni. Ne deriva che le difficoltà aumentano e le soluzioni, quando esistono, sono parziali.

In particolare, le infrastrutture relative ai *Big Data* non strutturati dovrebbero potersi estendere in modo da acquisire massicce quantità di dati, come nel caso dei video e foto dei social network. Sensori e monitor sono oramai parte integrante di apparecchiature di qualsiasi genere e la capacità di memorizzare, catalogare e analizzare i dati da essi forniti è carente, anche per quanto riguarda la videosorveglianza. Sarebbe poi necessario memorizzare enormi sequenze di dati da correlare e analizzare, come nei settori farmaceutico e della genomica, e servire video in tempo reale, sostituendo le tradizionali modalità di erogazione di contenuti via etere, cavo e satellite con quelle basate su Internet, per utenti fissi e mobili.

La memorizzazione di enormi volumi di dati non relazionali con un debole schema e la loro elaborazione a ritmi estremamente elevati, necessarie per l'analisi dei *Big Data*, hanno stimolato la comparsa di tecnologie quali Hadoop, che pre-elabora al volo in modo parallelo e distribuito grandissime quantità di dati provenienti dai sistemi più disparati e non correlati, indipendentemente dal formato nativo, ed effettua una veloce *analytics* esplorativa, rompendo l'approccio tradizionale che esegue le transazioni usando un codice procedurale e la gestione degli stati. Hadoop è un software open source lanciato nel 2008 da Apache, basato su una struttura dati internamente ridondante, installato su economici server standard industriali che possono scalare entro ampi limiti, anziché su costosi hardware proprietari specializzati. Ispirato dai due sistemi progettati da Google per gestire le pagine Web, il File System e il MapReduce, Hadoop è diventato subito lo standard di fatto per memorizzare, elaborare e analizzare volumi da centinaia di TB fino a PB di dati strutturati e no, quali file audio e di log, immagini, registrazioni di comunicazioni, e-mail e via dicendo.

Hadoop permette di collegare tra loro computer standard di fascia alta (del costo di 1.500/4.000 euro l'uno), detti nodi, in un cluster che può contenere fino a 4mila nodi. Per esempio, se si prevede l'ingresso di 1 TB di dati al giorno e una crescita mensile inferiore al 5 per cento, servono 61 nodi per un anno di utilizzo; se la crescita mensile è del 5 per cento il numero di nodi sale a 81 e a 109 se la crescita è del 10 per cento. Le prestazioni variano con il numero di nodi: per ordinare 9 TB di dati con un cluster di 900 nodi servono 108 minuti; per 14 TB con un cluster di 1.400 nodi servono 2,2 ore e per 20 TB con un cluster di 200 nodi, 2,5 ore. Prima di memorizzare i dati non è necessario predisporre uno schema rigido né conoscere la modalità d'interrogazione, che può essere decisa in seguito e comprendere domande non previste *a priori*. In tal modo si possono vedere relazioni altrimenti nascoste e ottenere risposte in precedenza fuori portata, e quindi prendere un numero maggiore di decisioni basate su dati oggettivi anziché sull'intuizione. Con il suo costo vantaggioso, Hadoop ridefinisce l'economia dei dati. Infatti sistemi *legacy*, sebbene adatti per determinati carichi di lavoro, non sono progettati per le necessità dei *Big Data*, e risultano di gran lunga troppo costosi per venire usati per scopi generali con insiemi di dati enormi. La scalabilità, l'architettura efficiente e la convenienza economica di Hadoop renderanno questa tecnologia sempre più attraente.

Le prospettive offerte dai *Big Data* sono state fiutate anche dai capitalisti di ventura statunitensi: un gruppo di questi, con alla testa Federico Faggin, ha finanziato nel settembre 2011 con 2,5 milioni di dollari la nascita di Peaxy, Inc. di Francesco Lacapra, fondata a San Jose di California. Nell'arco di nove mesi Lacapra, con un gruppo di quindici ingegneri, ha realizzato il sistema di gestione e archiviazione dati Hyperfiler™, un software di *file serving* svincolato da un hardware specifico e in grado di aggregare computer eterogenei all'interno di un unico *file system* distribuito a scalabilità illimitata. Hyperfiler™ gestisce volumi da 1 TB a molti PB, costa una frazione rispetto ai concorrenti e sarà disponibile alla fine dell'anno.

Prospettive future

Secondo il parere degli analisti, a causa della pervasività dei loro effetti i *Big Data* diventeranno ben presto un requisito standard nelle procedure informatiche d'avanguardia, rendendo obsolete le procedure e tecnologie precedenti. Verso la fine del 2015 le principali organizzazioni cominceranno a usare l'esperienza acquisita con i *Big Data* in qualche forma incorporata nelle proprie architetture e procedure. A partire dal 2018 le soluzioni basate sui *Big Data* offriranno sempre minori vantaggi rispetto a quelle tradizionali che hanno incorporato nuove caratteristiche e funzioni per supportare una maggiore agilità per quanto riguarda volume, varietà e velocità. Tuttavia le competenze, le procedure e gli strumenti attualmente considerati soluzioni *Big Data* sopravvivranno finché le grandi organizzazioni avranno incorporato i principi di progettazione e acquisito le competenze necessarie ad affrontare le problematiche *Big Data* come una flessibilità di routine. Di conseguenza intorno al 2020 i *Big Data* diventeranno semplici *data*, nel senso che non si differenzieranno più per caratteristiche e funzionalità e la loro gestione rientrerà nelle normali offerte dei venditori, esattamente come un supercomputer degli anni Ottanta è diventato un normale computer venti anni dopo, o un pc multimediale di metà anni Novanta è diventato un semplice pc dieci anni dopo. Per contro gli approcci architetturali, le infrastrutture e i sistemi hardware/software che non si adatteranno a questa nuova realtà usciranno dal mercato e le organizzazioni che tenteranno di resistere al cambiamento subiranno pesanti conseguenze economiche. ■

Angelo Gallippi è studioso d'informatica e saggista scientifico.