BIG DATA: VULNERABILITÀ E RISCHI EMERGENTI

COME HANNO COSTRUITO I LORO FILE SYSTEM E QUALI ATTACCHI HANNO SUBITO I GRANDI UTILIZZATORI DEI GRANDI DATI.

Marco Bozzetti, OAI founder

Il termine Big Data è sempre più d'uso comune, da un lato per l'enorme crescita delle informazioni multimediali, dall'altro per il consolidamento di fornitori di social network, di soluzioni terziarizzate e di cloud, che concentrano nei loro data center quantità di dati nell'ordine degli Zettabyte, ossia di miliardi di Terabyte.
Come da tempo sappiamo nei Big Data abbiamo a che fare con grandi volumi, alta velocità e varietà di tipi di informazione, ma questi sono tutti concetti 'relativi' che individuano collezioni di dati complesse e che diventano non gestibili con i tradizionali strumenti di archiviazione e di interrogazione.
La terziarizzazione e il cloud, con la concentrazione di infrastrutture e applicazioni ICT nei data center dei provider, enfatizzano ulteriormente il fenomeno dei Big Data. Ma proprio la centralizzazione e/o la gestione centralizzata di grandi quantitativi di dati diviene un obiettivo per gli attaccanti. Obiettivo che tende a essere perseguito sia con i ben noti attacchi ai database quali SQL Injection, cross site scripting, furto dell'identità degli utenti e degli amministratori dei sistemi e dei DB, sia con attacchi specifici che sfruttano le eventuali vulnerabilità dei sistemi ICT a supporto dei Big Data.
Infatti l'archiviazione e la gestione di Big Data, centralizzata o distribuita, il più delle volte non si basa sulle 'tradizionali' logiche di Rdbms e di SQL, ma su nuove tecniche che cercano di rispondere efficacemente alle esigenze di volumi, velocità e varietà del tipo di dati trattati.
Sono in corso a livello mondiale significative ricerche per trovare algoritmi e strumenti per il trattamento dei Big Data, che coinvolgono tecniche quali l'intelligenza artificiale, algoritmi genetici, sistemi che auto-apprendono (learning machine), analisi topologiche, matematica dei sensori, e altro. Basati su Rdbms o su logiche NoSQL, l'elaborazione di Big Data richiede normalmente logiche di MPP, Massive Parallel Processing, più o meno distribuite.

Nuove strutture e nuovi modelli per i Big Data

Al di là degli ambiti di ricerca di cui sopra, le realizzazioni operative più significative per i Big Data includono a oggi le soluzioni Amazon, Google e Apache Hadoop.
La soluzione S3, Simple Storage Service, di Amazon è un file system on line interfacciabile via web services Soap e/o Rest. Gli oggetti archiviati sono file, ciascuno fino a 5 TB e ciascuno associato a un 'meta dato' descrittivo di 2 KB. Questi oggetti sono organizzati in 'bucket' (traducibile in 'cesti') assegnati a un 'account' dei servizi Amazon.
La soluzione Google Cloud Storage è concettualmente simile a S3: è un file system on line interfacciabile via web services Rest. Gli oggetti archiviati in Google Storage (GS) sono file fino a 5 TB (ini-

zialmente erano fino a 100 GB, ma la concorrenza fa mettere le ali ...), anch'essi organizzati in bucket e identificati da un'unica chiave assegnata dall'utente. Ogni oggetto è indirizzabile da URL Http. Nell'ottica dei Big Data, a fianco delle capacità di archiviazione, Google offre due interessanti servizi:

- Google Big Query consente di analizzare grandi quantità di dati in pochi secondi; si basa su un linguaggio tipo SQL ed è interfacciabile da Rest API, Json-RPC e Google apps;
- Google Prediction API: praticamente un motore 'predittivo' con capacità di autoapprendimento nel cloud.

Google ha anche introdotto un framework, chiamato MapReduce, per l'elaborazione di grandi quantità di dati in parallelo su cluster di computer distribuiti. MapReduce è ispirato alla programmazione funzionale e si basa sulle funzioni Map e Reduce. La prima filtra i dati in input e li distribuisce sui computer del cluster. La seconda raccoglie i risultati delle elaborazioni della fase Map e li correla e combina per fornire il risultato finale richiesto. Per esempio la funzione Map effettua un sort di una grande anagrafica basandosi sul cognome, ed estraendo delle liste per ciascun cognome. E la funzione Reduce calcola il numero di occorrenze dei cognomi nelle singole liste, stabilendo la classifica dei cognomi più diffusi.
Apache Hadoop è una libreria di programmi software che costituisce un framework open source per l'elaborazione di grandi insiemi di dati, i Big Data appunto, operante su cluster di sistemi distribuiti usando dei semplici modelli di programmazione. Hadoop è usata sovente come framework per MapReduce. Apache Hadoop è già in uso in ambienti quali Google, Yahoo, IBM, Facebook, NewYork Times.

Vulnerabilità ed attacchi per i Big Data

Storicamente gli attacchi già occorsi per i Big Data, in particolare su Amazon e Google, erano basati sui tradizionali attacchi di saturazione delle risorse, DoS (Denial of Service) e DDoS (Distributed DoS). E' assai probabile che nel prossimo futuro vengano portati attacchi più sofisticati, partendo da vulnerabilità riscontrabili nei nuovi framework. Un primo esempio è la violazione di diritti d'accesso con Kerberos su Hadoop per cluster operanti in MapReduce (si veda CVE 2012-1574 e 3376).
I possibili rischi includono tutti quelli tipici di un database e delle sue interrogazioni, sia con SQL sia con web services, tipicamente SQL Injection, LFI/RFI (Remote File Inclusion), cross-site scripting, i furti d'identità digitale (soprattutto degli amministratori di sistema e dei data base), la non verifica e convalida da parte dei programmi applicativi, in particolare di quelli web, dei dati di input dell'utente, l'uso di algoritmi di hash deboli per la memorizzazione delle password, la violazione dei controlli e dei diritti d'accesso. A questi si possono aggiungere specifiche vulnerabilità dei nuovi ambienti quali Apache Hadoop e i file system, lo scambio di dati tra i vari nodi coinvolti nel MPP, e così via.

Come proteggersi?

I Big Data sono nella grande maggioranza dei casi gestiti da provider, in particolare di cloud: sono loro ad attuare e gestire gli strumenti e le politiche di sicurezza. Il normale cliente finale non può che controllare la loro esistenza ed efficacia, per esempio rispetto al contratto sottoscritto, e gestire correttamente il proprio account, in primis l'identificativo d'utente e le password.
A livello di controllo del provider, al di là del sistematico controllo dei report e della consolle forniti e di eventuali audit, può essere utile la pubblicazione: The Cloud Security Alliance Consensus Assessments Initiative Questionnaire. L'utente ha inoltre la possibilità di rinforzare alcuni strumenti che da lui dipendono o che può opzionalmente scegliere: uso di password di una certa complessità, cambiate con una certa frequenza e mai riusate, uso di tecniche di autenticazione forte e/o di autenticazione multi fattore, uso di connessioni crittate (SSL) e crittazione dei dati critici archiviati.

Marco Bozzetti
marco.bozzetti@malaboadvisoring.it