L'aumento incontrollato di dati e delle dimensioni dei dispositivi di memoria di massa e non solo, è diventato un ostacolo per gli investigatori informatici, che devono affrontare una corsa agli armamenti informatici per far fronte a questo tsunami digitale. In quest'articolo si proverà ad analizzare alcuni sistemi organizzativi che potrebbero aiutare a vincere questa battaglia.

di Nanni Bassetti

COME AFFRONTARE I "BIG DATA" NELLA DIGITAL FORENSICS

***Nanni BASSETTI**, laureato in Scienze dell'Informazione, libero professionista specializzato in digital forensics, fondatore di CFI (Computer Forensics Italy) e project manager di CAINE Linux/GNU Live distro per indagini informatiche. Docente, relatore in parecchi corsi ed eventi, autore di molti articoli tecnici e di un paio di libri.*

Ormai quando si parla di informatica si pensa a grandissime moli di dati, i terabyte sono diventati comuni, il numero di dispositivi informatici, *computer*, cellulari, *tablet*, *pendrive*, ecc. è in costante aumento, in una famiglia normale si potrebbero trovare 3 o 4 *smartphone*, 2 *tablet*, 2 *computer*, svariate *pendrive* e centinaia di cd-rom/dvd-rom.
"Si pone la questione di quanti dati digitali attualmente memorizzati su dispositivi elettronici personali o in sistemi d'archiviazione. Nessuno sa per certo, ma è stato stimato che la quantità di dati che è stata generata a livello mondiale nel 2013 ha raggiunto i 4 zettabytes. Un Zettabyte (1 triliardo di byte) di dati è probabilmente equivalente a tutti i granelli di sabbia si trovano su tutte le spiagge sulla Terra." (cfr. Streamlining the Digital Forensic Workflow by John Barbara http://www.forensicmag.com/articles/2014/10/streamlining-digital-forensic-workflow-part-1).

Tutta questa massa di dati diventa un bel problema per gli investigatori informatici, che si ritrovano a fronteggiare questo "Golia" digitale, armati peggio del celebre David. Il *blob* informatico avanza e sovrasta l'investigatore informatico, poiché quest'ultimo non riesce a vincere la corsa agli armamenti, per motivi economici, temporali, organizzativi. In aggiunta a questi problemi, gli esaminatori incontrano anche notevoli difficoltà quando si analizzano i media utilizzando gli strumenti tradizionali comuni.
Oltre ad essere costosi e *time consuming*, questi strumenti spesso hanno carenze e limitazioni:

- il volume di dati e il numero di dispositivi coinvolti nelle indagini è cresciuto molto più velocemente delle capacità dei tool forensi;
- è necessaria la formazione continua e costosa per l'uso dei tool;
- ormai ogni tool richiede quasi un intero *computer* solo per sé;
- le interfacce possono non essere facili da usare;
- non sono in grado di fare *triage* e richiedono un'analisi completa del dispositivo;
- i tempi per creare immagini forensi sono molto lunghi;
- gli investigatori devono correlare manualmente i dati provenienti da più origini;
- non sono in grado di lavorare in modo affidabile i dati che possono risiedere in sistemi aziendali, complessi sistemi di *storage*, *cloud* o su dispositivi mobili;
- ci sono difficoltà a separare i dati per consentire a più esaminatori o investigatori di eseguire analisi contemporaneamente.

Piuttosto che il tipico approccio, uno più pratico sarebbe quello di semplificare il flusso di lavoro:

- utilizzare uno strumento di *triage* per identificare le fonti di prova più probabili;
- dividere le prove tra più investigatori per ulteriori analisi;
- esportare tutti i dati rilevanti presenti in un report ben fatto.

Gli scenari, quindi, sono sempre più eterogenei, ci si trova innanzi a tantissimi dispositivi da sequestrare ed analizzare, quindi ci sono due approcci possibili: quello muscolare e quello smart.
L'approccio "muscolare" è basato sull'aumentare la forza per sostenere il carico di lavoro, quindi raddoppiare se non quadruplicare il personale addetto, acquistare più *computer*, *software*, ampliare i locali ed aumentare la formazione specifica per tutti gli

operatori. Così facendo sicuramente si riesce a creare un argine alla marea di big data che incombe, ma ha dei costi elevatissimi, che mal si coniugano con la realtà lavorativa ed economica in cui è immersa l'attività del *digital forensics specialist*.
L'approccio "smart" è basato sul *divide et impera* di Romana memoria, infatti il flusso di lavoro dovrebbe esser diviso in una parte "*triage*", ossia valutazione delle priorità ed in una parte "*deep*", ossia analisi approfondita di ciò che rimane dalla prima scrematura.

Quest'ultimo approccio ha l'indubbio vantaggio di esser più economico, flessibile e scalabile, ma come metterlo in pratica?
Il senso generale è quello di scremare immediatamente i dispositivi contenitori, poi i dati ed infine l'approfondimento, ma per far questo si dovrebbe avere una squadra specializzata nella fase di perquisizione e sequestro, già in questo momento gli specialisti dovrebbero saper riconoscere e scegliere ciò che è il più pertinente possibile all'oggetto d'indagine.

Primo step: scegliere ciò che si vuole subito analizzare. Dividere i compiti con più operatori.

Secondo step: avere un *kit* di *software* che possa fare un'analisi degli elementi più importanti ed evidenti che servono all'indagine. Quindi questi *software*, a seconda dell'oggetto d'indagine, dovrebbero tirar fuori subito i dati più rilevanti, ad esempio se si parla di truffe, fatturazioni false, insomma cose connesse a documenti e carteggi, allora il sowtware fa subito un'estrazione dei *file* documento tipo: doc, docx, xlsx, xls, cvs, pdf, odt, ecc. Poi e-mail, *file* cancellati, ecc. Se il caso riguarda pedopornografia, allora si lancerà un *software* in grado di estrarre subito tutte le immagini e video.

Dopo questo secondo *step*, si riduce ancor di più il *set* di dati da analizzare, escludendo quei reperti contenenti nessuna informazione utile e passando ad un'analisi approfondita di quelli rimasti che contenevano tracce più attinenti all'investigazione.
Effettuare copie forensi dei dispositivi interessanti. In questo modo si ha una specie di filtro a insiemi successivi sempre più piccoli, in modo da ridurre il *set* di dati da analizzare e rendendoli compatibili con i tempi e la struttura del laboratorio preposto all'analisi.

Volendo immaginare un *work flo*w si potrebbe progettarlo così:

Polizia Giudiziaria addestrata all'identificazione ed acquisizione forense dei reperti digitali
- *Analisi veloce sui reperti identificati come importanti.*
- *Estrazione dati identificati come importanti.*
- *Ulteriore scrematura dei reperti in base all'analisi dei primi dati estratti.*
- *Copie forensi, ossia duplicazione tramite file immagine dei dischi.*
- *Scrivere tutto quello trovato in dei report.*
- *Utilizzare software per incrociare i dati provenienti dai vari dispositivi.*

↓

Laboratorio forense di specialisti esterni
- *Analisi profonda sui reperti scremati, utilizzando sempre strumenti ad-hoc per estrarre i dati più pertinenti all'indagine.*
- *Report finale.*

I *tool* di pre-analisi per identificare le priorità dovrebbero essere specifici e di facile uso e forensi per limitare al massimo l'impatto sul reperto originale. Un arsenale di *software* che estrae i dati utili ad identificare il reperto come fonte utile all'indagine può essere commerciale o open source/freeware come le live distro forensi.
Se si avvia un *computer* con una *live distro* forense (Caine, Deft, ecc.) si possono lanciare degli strumenti di ricerca come: grep e strings, The Sleuthkit, Autopsy, NBTempo, regripper, photorec, ecc..
Oppure i *software* come Win-Ufo/Win-Fo o Dart, sempre presenti nelle live distro forensi, senza contare molti *software* commerciali creati appositamente per il triage, che permettono una preview di informazioni utili. Dopo quest'ispezione si può decidere se effettuare la copia forense del disco e quindi metterlo nel paniere dei dispositivi candidati ad un'analisi approfondita.
Quando il *file* immagine arriva al consulente o laboratorio esterno, questo potrà esser analizzato con più calma e correlato ad altri reperti, magari già collegati tra loro dalla prima fase d'analisi, per esempio dall'esame del registro dei dispositivi USB collegati a quel *computer*.

L'approccio pratico

In un'abitazione nella quale abitano un uomo, una donna e due bambini, si trovano 3 *computer*, 2 hard disk esterni, 1 *tablet*, 3 *smartphone* e 5 *pendrive* da 16Gb. Ipotizziamo che il reato sia quello di detenzione e diffusione di immagini illegali. In prima istanza la P.G. procede con una live distro ad analizzare i dispositivi in uso all'uomo, riducendo il set a 1 *computer*, 2 hard disk esterni, 3 *pendrive* e 1 telefono. Avviando il *computer* con una *live distro* forense o con altri *software* di pre-analisi, si cercano tracce di: navigazione su siti illegali; fotografie e video; *file* aperti di recente; parole chiave; *data carving* su immagini e video; eventuali attività di chat; registro dei dispositivi USB che sono stati collegati a quel *computer*; *file* cancellati; ecc.

Si scoprono immagini d'interesse per l'indagine tra i primi risultati di questa pre-analisi, in seguito si rileva che dei 2 *hard disk*

esterni solo uno è stato collegato più volte a quel *computer*, poi si evidenzia che il telefono non è mai stato collegato al *computer*, infine si scopre che solo una *pendrive* è stata collegata. Stesso trattamento per gli altri dispositivi, che danno tutti esito negativo.
Chiaramente sarebbe sempre consigliabile effettuare le copie forensi di tutti i dischi in ogni caso, ma questo è da valutare di volta in volta. A questo punto si effettua la copia forense con verifica hash solo di: 1 *hard disk* del *computer*; 1 *hard disk* esterno; 1 *pendrive;* si sequestra e/o si analizza il telefono.
I 4 reperti vanno nella fase d'analisi approfondita. Si è quindi risparmiato tempo e denaro perché si è passati da 14 reperti a 4, riducendo il numero di GigaByte da analizzare.

L'approccio probabilistico

Il Dr. Simson Garfinkel ha immaginato un approccio di tipo probabilistico basato su una raccolta casuale di settori di un *hard disk*, al fine di capire se quel dispositivo contenga o meno delle informazioni importanti. In sintesi il metodo è basato su catturare pochi Gigabyte di settori in modo casuale e analizzare il campione. Per un discorso probabilistico, il campione potrebbe già rivelare se ci sono informazioni importanti sul disco: con questo approccio un disco da 1Tb può esser classificato in 5 minuti rispetto ai 208 minuti previsti (per approfondire il metodo http://simson.net/ref/2012/2012-10-01%20Forensics%20Innovation.pdf).

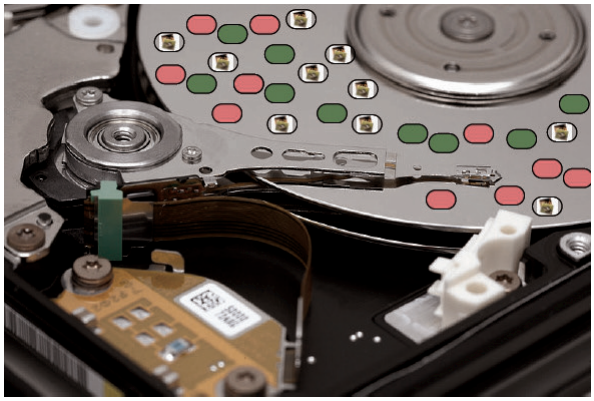

Figura 1 tratta da http://simson.net/ref/2012/2012-10-01%20Forensics%20Innovation.pdf

Vediamo un esempio. Assumiamo che un disco da 1 TB abbia solo 10 MB di dati.

- 1TB = 2 miliardi di settori da 512 byte.
- 10MB = 20.000 settori da 512 byte.

Se leggiamo solo 1 settore, la probabilità che esso sia vuoto è:

(2.000.000.000 – 20.000) / 2.000.000.000 = 0.99999

quindi prossimo al valore 1, il che significa certezza che sia vuoto.

Se prendiamo 2 settori, la formula diventa:

((2.000.000.000 – 20.000) / 2.000.000.000)*((1.999.999.999 – 20.000) / 1.999.999.999) = 0.99998

In questo caso la probabilità che il campione sia vuoto è leggermente più bassa, quindi più settori prendiamo e più aumenta la probabilità di prenderli pieni di dati. La probabilità di un campione di soli settori vuoti tende quindi a zero:

$$P(X = 0) = \prod_{i=1}^{n} \frac{((N - (i - 1)) - M)}{(N - (i - 1))}$$

Conclusioni

I metodi qui descritti consentono un approccio più mirato alla ricerca di prove digitali e un uso più efficiente dei laboratori e consulenti forensi. Questi flussi di lavoro sfruttano il fatto che gli investigatori, che meglio conoscono il caso, possono scegliere ed ordinare le evidenze, lasciando agli esperti forensi informatici l'analisi approfondita ed il *reporting*, ottenendo così in tempi e costi relativamente contenuti, i risultati necessari e sufficienti per concludere il lavoro senza i "backlogs", cioé gli accumuli di lavoro che rischiano di esser svolti male e frettolosamente.
Quest'articolo vuole solamente descrivere alcuni metodi che negli ultimi anni si stanno pensando al fine di rendere più snella, veloce e meno costosa la *digital forensics*, soverchiata dalla crescita abnorme dei *terabyte* di dati, quindi è solo uno spunto di riflessione, che può aprire però varie discussioni ed approfondimenti sul tema. ©

BIBLIOGRAFIA

- http://www.forensicmag.com/articles/2014/10/streamlining-digital-forensic-workflow-part-1
- http://www.forensicmag.com/articles/2014/09/streamlining-digital-forensic-workflow-part-2
- http://www.forensicmag.com/articles/2014/12/streamlining-digital-forensic-workflow-part-3
- http://simson.net/ref/2012/2012-10-01%20Forensics%20Innovation.pdf
- https://www.tracksinspector.com/?p=167
- http://www.studioag.pro/wp-content/uploads/2013/10/DigitalForensicsBigData.pdf