Processing Large Volume Data Sets in Criminal Investigations

ACFE Sydney Chapter

Dr Allan Watt, CFCE, CFE
Centre for Policing, Intelligence and Counter Terrorism,
Macquarie University
Sydney, Australia

Introduction

- Digital Forensics constantly evolves and will for the years to come
- Most practitioners use industry-standard software
- Coupled with knowledge and experience, conduct investigations on digital devices
- The vast majority of cases until the recent few years, involved end-user devices: PCs, Laptops, Apple Mac products and more recently Smartphones and Tablets

The Current Horizon

- Moore's Law and other new developments has seen, the growth in hard disk drives
- Not uncommon now to find 3TB drives
- 1 TB drives with laptops
- The problem the Input/Output (IO) has not dramatically increased
- Non compressed an image of 250GB in many cases at best takes about one hour
- 3 TB uncompressed could take 12 hours to image

Time Overhead

- This creates a huge time overhead for digital forensic examiners
- Some cases noncriminal breaches of regulations in some countries prohibit seizure and require acquisitions to be made onsite
- This therefore becomes an issue and staff may either have to be subjective as to what they take and run the risk of losing valuable evidence
- Even conducting physical acquisitions of Smartphones like the iPhone4 can take hours

Acquisitions

- Digital forensics involves Collection, Preservation, Analysis and Presentation of evidence
- Nearly all start with collection and preservation
- Only in a few cases where an acquisition does not occur
- Acquisition is the process of collecting the digital evidence and preserving it in a format which prevents the original data being modified

Drive Acquisition

- Involves making a copy of the computer's internal hard disk drive
- In earlier days this was a clone
- Last 14 or so years, involved removing the drive from the computer, attaching a write-blocking device to it, then with forensic software making a physical acquisition
- Once the image has been created it is considered preserved and is encapsulated
- The most common formats are E0, AD, DD

Network Devices

- With servers numerous options are also available, that include, starting the server with a bootable disc and with a cross-over cable image the server over a network
- Alternatives, image the drives individually and then logically rebuild the RAID within the forensic software so it can again be seen as a working entity
- The problem with all of these is time
- Nearly all cases want to take a full physical acquisition

The Cloud

- What really is the cloud?
- Cloud storage is simply a term that refers to online space that you can use to store your data
- Cloud storage provides a secure way of remotely storing your important data
- From a forensic perspective cloud storage is any network storage in which the examiner cannot directly get his hands on
- Most Big Data will reside in the cloud

Cloud Remoteness

- Add to the equation, size and remoteness and that all the data has to be sucked down through a network cable, further compounds the situation
- The Cloud storage facility could be anywhere in the world and its physical location may be unknown
- Even if physical access to the cloud device exists, chances are you will not be able to take it offline, as it could shared by tens or hundreds of unrelated users
- Then there are the user's credentials, username and password to gain access to it

Enterprise Solutions

- The most preferred option is to use an Enterprise tool, Encase Enterprise; FTK also has one and there are also specific tools like F-Response
- Deploy an Agent to the remote device, presents the remote storage as a drive letter on the computer which is accessing the cloud and has the enterprise tools on in
- The problem now occurs again, with firstly size, the cloud capacity may be in TBs and the network speed is also going to have a major impact on the time
- If there are many users at the same time accessing that storage, then the overhead on that could result in very low acquisition times

Risk of Alteration

- Criminals involved in drug dealing, fraud and child abuse material matters, may store data in a cloud or clouds, rendering acquisitions not impossible, but cumbersome
- It should be not excluded that someone could access the cloud storage facility and take that user's data offline, rendering the acquisition potentially worthless
- Alternatively, someone else could login at the same time and start deleting data whilst it is still acquiring, again rendering some or all of the acquisition worthless

Software

- Over the last 15 years, there have been many industry tools produced to assist with conducting digital forensic examinations
- Prior to this examiners would simply interact with the clone or some created their own tools
- Tools need to be constantly updated and new needs arise continually as vendors appear with new and much needed products
- An investigator will have a collection of tools of choice they use during an investigation

Tools

- Some of these tools are as follows:
- *General all round Forensic*
  - Guidance Software - Encase Forensic
  - AccessData – Forensic Took Kit (FTK)
  - X-Ways Forensics
- Many other tools for specific actions

Limitations

- Many have data limitations and are still primarily one instance one or two images analysed through a one on one forensic device
- Most digital forensic investigations, this works well as the majority of cases there is one or two Persons of Interest (POI) and one to three devices
- Beyond that, the overhead on hardware is too great, especially with larger capacity devices

Fraud

- Accounting data is often within an accounting package and cannot be directly accessed through the forensic software
- Data files need to be exported and mounted directly with the software, or emulate the image as an external drive or run Virtual Machine
- Often a client server or even cloud environment, this would have to be emulated
- Also volumes of data, which singularly can still be analysed on one by one or two by two basis; however, there is a lot of duplication/replication
- Having all the data available to analyse in one environment at one time is of great benefit; however, duplicate data can also be removed

Processing

- In the earlier days more data was directly accessibly and searchable immediately within an image
- Nowadays, much more data is no longer directly accessible without first processing it into a searchable and intelligible format for analysis
- Some data files would be: MS Exchange data, MS Outlook PST data. Also OLE streams, since Office 2007 can also include, backup and other compressed files
- Windows Registry, link files, event log, plist files and many others
- Encase 6 and earlier, were using scripts that processed what was needed for a given investigation as needed and processed on the fly
- Searching would also be done during the analysis as and when required, however providing a time overhead in the process

Later Processing

- FTK and Encase 7, provides the software coupled with a Database back engine that allows pre-processing of any or all data as required
- Can also be indexed, so once indexed, searches of data can be done instantly, data bookmarked and eventual exported into a report
- Initial time overhead and can be hours or days or even months, but once completed allows an investigator to quickly conduct the analysis and complete a job
- This is a new and efficient way, but they still have hardware overheads
- In complex fraud investigations with numerous devices, there will be duplicates and cause replication in the investigation process and these need to be removed
- These new tools also have the ability to de-duplicate

Lets look at it in summary

- What is this thing called Big Data?
  - Big Data is basically lots of data, but now with the addition of multimedia, growth is becoming exponential.
- How big is big and will it get bigger?
  - Currently anything over about 3 TB is becoming Big Data.
  - It will get bigger, cost of cloud storage is getting cheaper and it is cheaper to buy more than archive it off.

Lets look at it in summary

- How do we process this data in Fraud Cases, e.Discovery v Digital
  - The cost of processing large datasets is becoming excessive, the standard forensic tools on a PC can no longer cut it.
  - One option is use data analytics, but this will not pick up false invoicing, false fabrication or false documents.
  - Random sampling of data is still an option.

Lets look at it in summary

- Forensics?
- What are the solutions to tackling the problem, in-house/outsource/Other?
  - In-house new equipments $350 to $500k plus annual licensing, $50 k
- What is the cost of processing large datasets outsourced?
  - 100 GB of email containers $100K

Summary

- The face of digital forensics is changing and the likelihood of large datasets will occur more frequently
- The processing costs coupled with hardware, software and licenses will make in-house processing difficult and the feasibility of outsourcing has to seriously be taken into consideration within a PPP
- LE/Watchdog Agencies could outsource to their e-discovery/large volume processing jobs to commercial agencies or a centralised government owned facility could be established to undertake this work for all agencies within their jurisdiction