Global Information Assurance Certification Paper

How Data Analytics Saved Me Money On My Digital Forensics Services

GIAC (GCFA) Gold Certification

Author: Taurean B. Dennis GCPM, GCFE, GCFA, GREM, CCE, CCFE, GMOB, GCIH
Advisor: Adam Kliarsky
Accepted: January 13th, 2017

Template Version September 2014

Abstract

Data is the building block of our modern society. The IBM Corporation estimates that we create 2.5 quintillion bytes of data every single day (Bringing big data to the enterprise.n.d). This data comes from several sources, such as personal computers, smartphones, and other types of devices (fitness bands, IoT devices). The sheer size of today's data has become much of an obstacle for many organizations that perform digital forensics and incident response due to the cost involved in storing and analyzing the data in a timeframe required to have an answer for the client. So how can data analytic solutions such as Tableau or Elasticsearch help save you lots of time and money? The following research will prove how it is done.

1. Introduction

The cost performing digital forensics investigations and incident response in cases involving data exceeding terabytes is becoming a burden for many organizations especially those in law enforcement. According to the FBI Regional Computer Forensics Laboratory, in 2013 they processed 5,973 TB of data which was 40% more than in 2011 creating an even larger backlog of data (Scanlon, 2016, pg.2). Fast forward to the present based on those past statistics, the volume of data needed to be processed for cases continue to increase significantly creating one the greatest issues to organizations performing digital forensics.

It is not uncommon in today's investigations to have evidence across many types of mediums such as workstations, laptops, mobile devices, cloud storage to name a few. In an article from Gartner on June 25, 2012, Gartner stated: "Average storage per household will grow from 464 gigabytes in 2011 to 3.3 terabytes in 2016" ("Gartner Says That Consumers Will Store More Than a Third of Their Digital Content in the Cloud by 2016," n.d.).

In 2012, Gartner believed that "the adoption of camera-equipped tablets and smartphones would drive consumer storage needs" ("Gartner Says That Consumers Will Store More Than a Third of Their Digital Content in the Cloud by 2016," n.d.). This exponential growth in storage needs will only continue to increase as trends have shown thus creating larger datasets of evidence in current and future digital investigations.

The current tools available such as EnCase, FTK, and X-Ways are not currently powerful enough to search and analyze data of this size in a reasonable amount of time. An example, based on an experience in the field of digital forensics involving a domestic crime, the prosecutor requested to have all artifacts relating to the suspect communicating to the victim provided by the end of business for review. Using a traditional forensics tool, the process took longer than expected to complete. Traditional forensic tools have a searching algorithm that is linear in nature meaning whenever a search is conducted it will always start from the beginning of the data structure and transverse until the end. To gather a better understanding of how the algorithm works, the structure of traditional magnetic-based storage, such as hard disk drives, must be examined.

Hard disk drives are storage devices that store data on rotating disks called platters. The structure of the data stored on the platters is called sectors. A sector holds user data in fixed amounts of 512 bytes (Fisher, 2016) or 4096 bytes for modern larger drives. If a user has a file that is 2000 bytes stored on a hard drive that is configured with 512-byte sectors, it will take approximately four sectors to store the file as seen in the illustration below in Figure 1

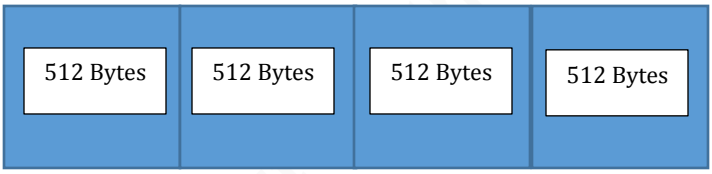

Figure 1 - A visual of a 2000 byte file that is comprised of four sectors 512 bytes

The storage of a hard drive starts at the first sector known as sector 0 and sequentially goes to the next sector until the last sector has reached the full capacity of the hard drive.

The introduction of data analytics has changed how large datasets can be searched and analyzed efficiently. Data analytics use special algorithms to search large datasets based on technologies involving machine learning, artificial intelligence, statistical analysis, and database architecture.

The benefits of using the technologies are that the tools can search and parse through very large data sets more quickly and efficiently. The IRS began using data analytics after a cyber-breach in 2014 that involved cyber criminals stealing taxpayer data. The intelligence gathered from the breach allowed the IRS to apply the information to data analytics which helped them reveal more than 600,000 additional victims of the previous breach (Thorton, 2016).

There are many solutions on the market such as Tableau, Splunk, and Google Cloud. Data analytic tools provide a broad overview of very large datasets in the form of easy to comprehend business intelligence dashboards by giving an analyst pinpoint information in the form of visualization (Wong, 2016).

In further discussion of this topic, a use case will be conducted to show the time and cost difference of using data analytic tools compared to traditional digital forensics tools. A sample size of 250 GB forensic image will be used that consists of pictures, documents, compressed archives (ZIP and RAR), and more. The time it will take to index and search the digital evidence will be recorded between several digital forensic and data analytic tools. The cost formula is an example of one that might be used in the industry.

2. Indexing

The experiment setup consists of the following components listed below; please note there was no available licensing for FTK and X-Ways at the time of this experiment.

Hardware

HP Laptop configured with the following:
CPU: Intel Core i7-4600U @2.10GHZ
RAM: 16GB
Hard Disk: 500GB Solid State

Software
Windows 7 Enterprise
VMware Workstation 10.0.7
EnCase 7 (SAFE License)
Tableau Desktop (Free License)
ELK (Elasticsearch, Logstash, and Kibana)
Ubuntu Virtual Machine

2.1 EnCase Enterprise

The first test was conducted using EnCase Enterprise on a 250GB image called backup3_3_16.E01 in Expert Witness Format (EWF). In this test, the time was recorded to see how efficient in terms of time it will take for a traditional forensic tool such as EnCase to index a 250 GB forensic image compared to a data analytic solution. The following steps were taken to setup the experiment.

1. A new case was setup with the name Forensic Test # 1

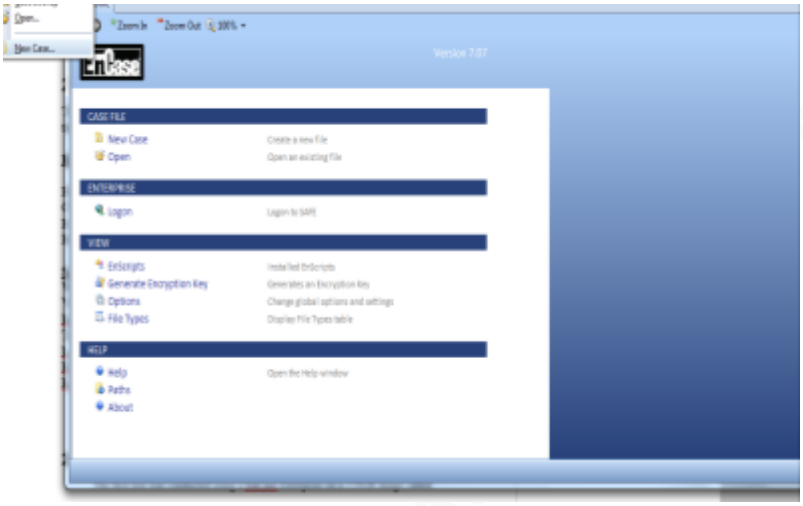

Figure 2 - The main menu of EnCase 7 to create a new case

2. On the next screen, the "ADD EVIDENCE" button under the EVIDENCE tab is clicked to add the image backup3_3_16.E01.

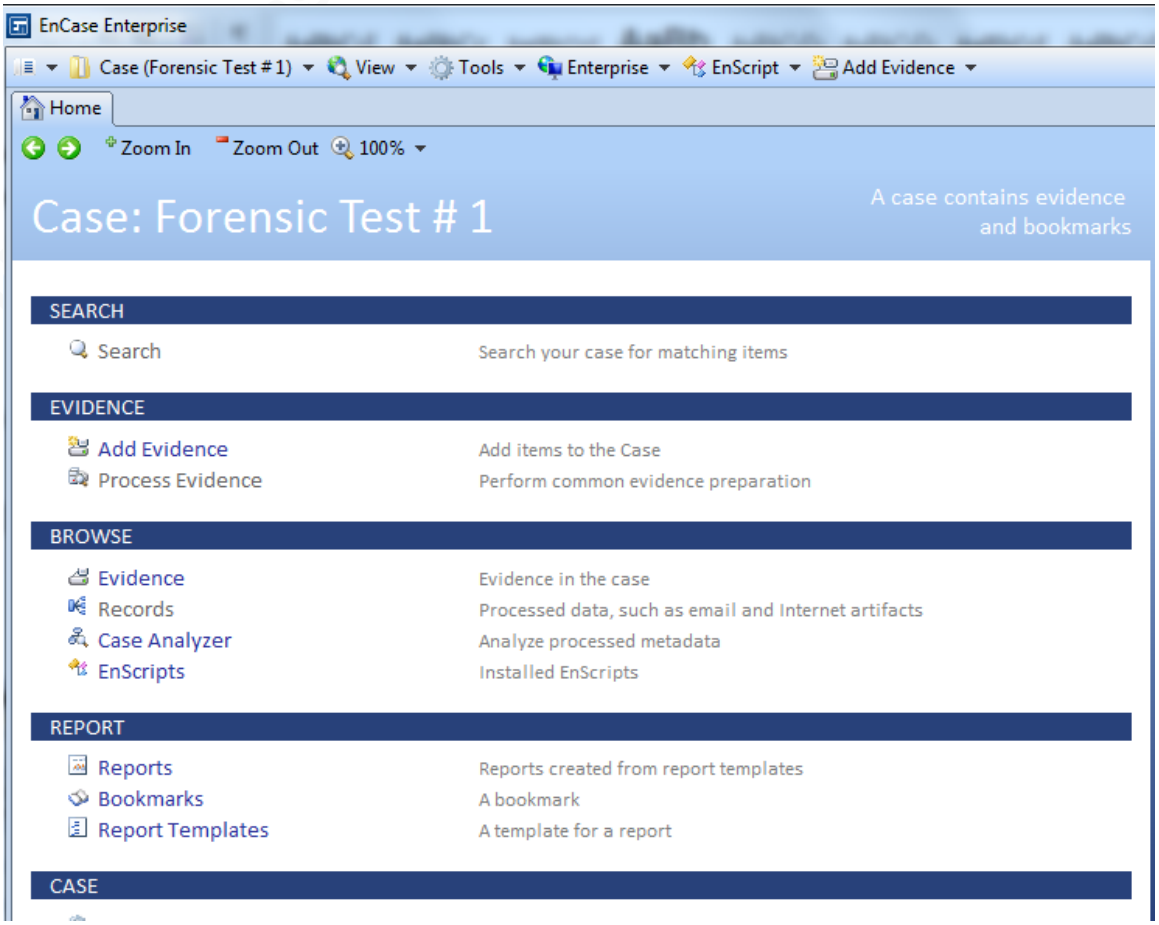

Figure 3 - Adding Evidence backup3_3_16.E01 in the Forensic Test #1 case that will be indexed

3. The following settings were used to index the image in the Case Processor, after clicking ok, the digital timer was started

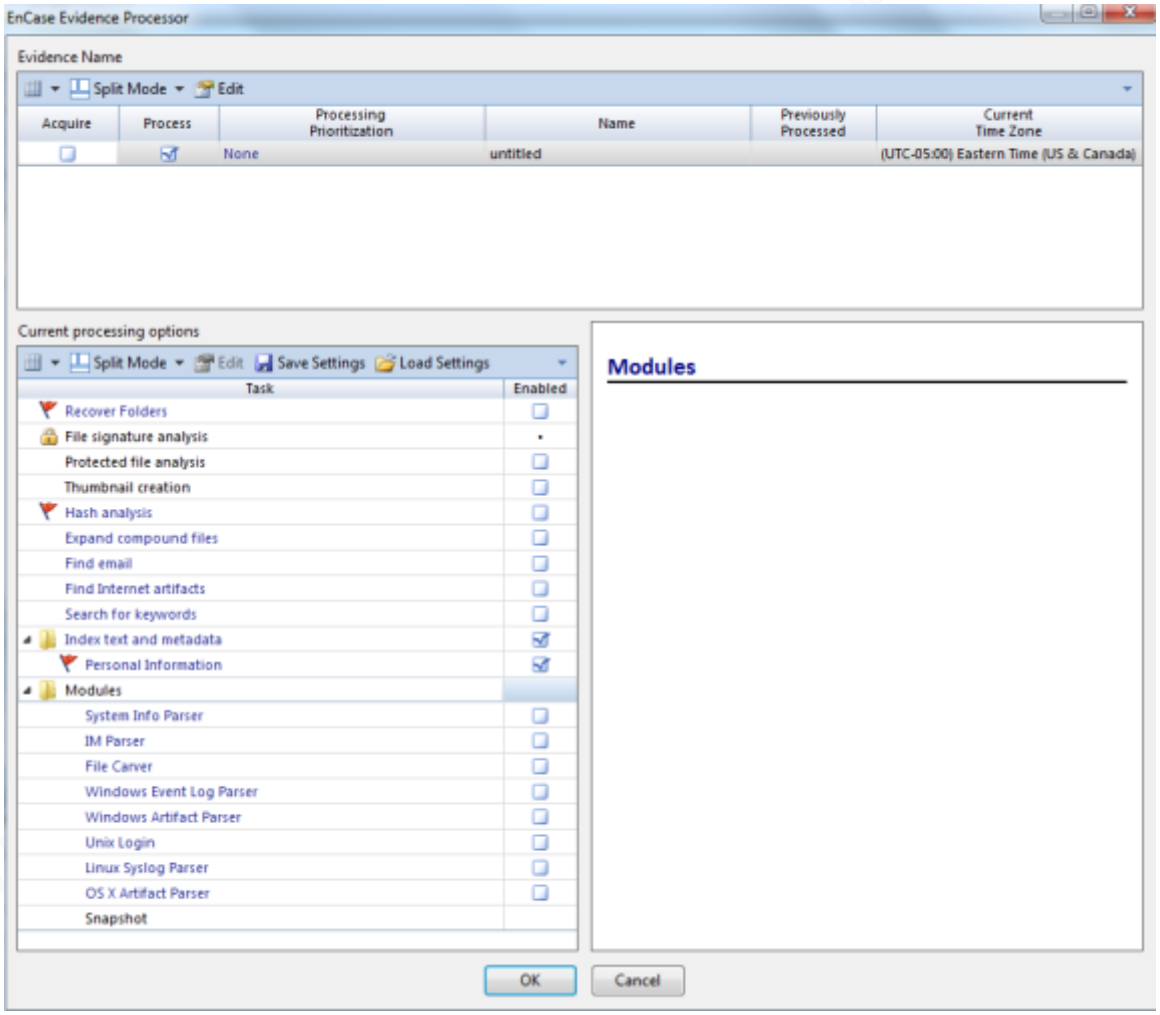

Figure 4 - The Processing Options used in EnCase 7 to index the forensic image backup3_3_16.E01

4. The indexing process completed in approximately three hours.

Many firms use a cost per hour ratio when determining the cost of digital investigations. Digital investigations are broken down into four main phases: imaging, processing, analysis, and reporting. The imaging phase consists of procuring digital evidence using forensically sound procedures and tools while protecting the integrity of the evidence. The processing phase involves working on a secondary copy of the master copy that was created from the original evidence using an automated forensic tool such as EnCase to break down the evidence into a form for analysis.

After the evidence is processed, the analysis phase begins whereby the practitioner looks for related artifacts that provide the answers for the client. The final phase is reporting which involves the examiner writing a report of his or her findings for the client. The report gives the client the necessary information to determine what the next step will be in the investigation. An example, an investigation involving intellectual property theft, if the report provided shows that the suspect was indeed the culprit, the next decision would be to pursue prosecution.

The firm's policies determine the cost of each phase. In some firms that provide digital forensic services, imaging is priced at a flat rate depending on the size and type of media; however, the other phases processing, analysis, and report writing are charged at an hourly rate. Using the following formula. The focus will be on the cost per hour for indexing.

Total Indexing Cost = Cost Per Hour X Number of Hours

Referencing back to the EnCase Enterprise index experiment, the total amount of time to index took 3 hours. If the firm theoretically charges $350/hr to index data. The total cost to the client so far in the investigation process would be

Cost Per Hour = $350

Number of Hours = 3

Total Processing Cost = $350 * 3 = $1050

2.1.1 Tableau Desktop and ElasticSearch

In the midst of this experiment, there was no knowledge of data analytic tools being not able to index evidence directly from a forensic image file or mounted filesystem. In order to ingest evidence from a forensic image, it is converted into a format compatible with the data analytic tools available as of present.

Most of the data analytic tools such as Tableau Desktop or ElasticSearch take in data in the form of text files, comma separated values(CSV), or SQLite databases to name a few. The following steps were taken to prepare the experiments based on the research conducted from Michael Maurer(Maurer, 2016).

1. Downloaded the log2timeline binary **plaso-1.5.1-win-amd64-vs2010.zip** from https://github.com/log2timeline/plaso/releases
2. In the Plaso the following command was run to create an index of the forensic image mounted in FTK Imager

log2timeline.exe output.plaso E:

3. After the plaso file was created the file was converted into a CSV file to ingest into Tableau using the following command:

Psort –o l2tcsv –w output.l2tcsv output.plaso

4. Using the file output.l2tcsv to import into Tableau Desktop, the following error was displayed

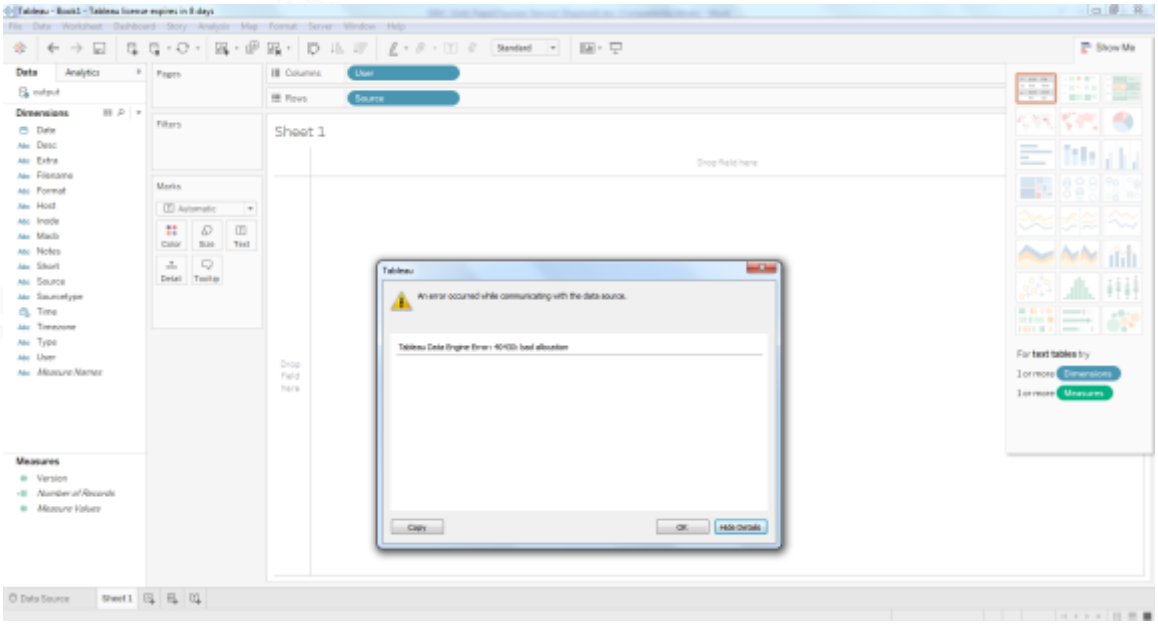

Figure 5 - Error displayed in Tableau Desktop when trying to import the output.l2tcsv file

The error could not be resolved during this experiment to show the speed comparison of indexing data between a traditional forensic tool and data analytic tool An alternative solution was used in the form of Elasticsearch.

Using the same plaso file, this was repeated using ElasticSearch**.** An Ubuntu virtual machine was used to prepare the environment.

1. The following command converted the plaso file into a format Elasticsearch can ingest

```
test@ubuntu:~$
test@ubuntu:~$
test@ubuntu:~$ psort.py -o elastic  --raw_fields --index_name forensictest1 '/home/test/Desktop/output.plaso'
```

Figure 6 - The psort command was run with the elastic switch to convert the plaso file into a format Elasticsearch can ingest

After the conversion completed approximately in two hours, the indexed data could be searched within Elasticsearch by accessing the Kibana dashboard interface on **"localhost:5601"**.

2. The previous psort command indexes the data before it is ingested into Logstash which explains the speed of the indexing being very fast.

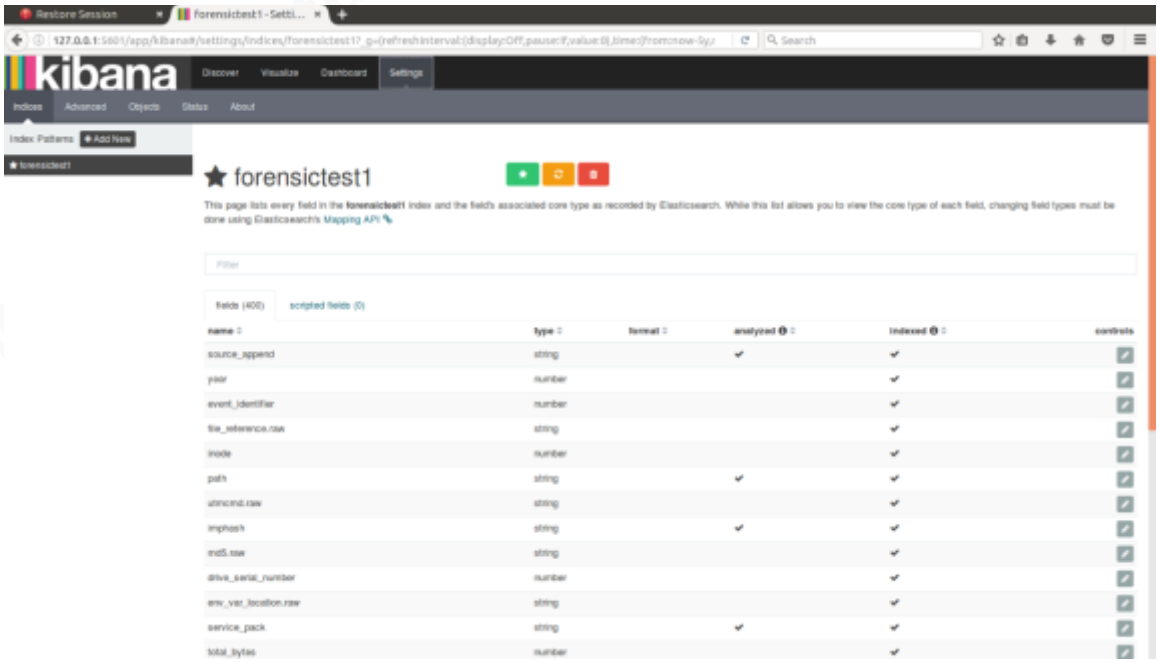

Figure 7 - This is Kibana dashboard showing the ingested data from the index created in Fig. 6

2.1.2 Indexing Speed (Cost Comparison Ratio)

After conducting the indexing experiment between a traditional forensics tool compared to a data analytic tool. It can be observed that the data analytic tool was faster than the traditional forensics tool.

The indexing procedure of EnCase 7 was approximately 3 hours to complete on a 250GB image with specs available for this experiment. The procedure for indexing in the two data analytic solutions Tableau and ELK completed within 2 hours which includes the data being indexed and converted into a format that is compatible. Observing the formula again to compare the indexing time and cost between the traditional forensic tool versus the data analytic tool, the choice to use a data analytic tool appears lower in cost.

ELK and Tableau Desktop Cost:

Total Indexing Cost = Cost Per Hour X Number of Hours

$350/hr X 2 = $700

EnCase 7 Cost:

Total Indexing Cost = Cost Per Hour X Number of Hours

$350/hr X 3 = $1050

3. Evidence Searching

The next series of experiments will examine the time it takes a traditional forensic tools to sift through a large amount of data versus a data analytics tools.

3.1 EnCase Enterprise

As discussed previously, traditional forensic tools are designed to search through data in a linear method. The linear method of searching through data involves starting from point A until point B is reached. Using this analogy in comparison to searching through data storage, the traditional forensic tool would start from the beginning sector and transverse to the ending sector.

Using EnCase Enterprise the follow procedures were conducted to test approximately how long it would take to search through 250 GB image using a keyword list search.

1. A new case was created called “Forensic Test 2.”

The following keywords were used

“forensics,” “security,” “Tableau,” “SANS,” “645-689-2000”, Test#!$

2. On the evidence tab, the option “Raw Search Selected”

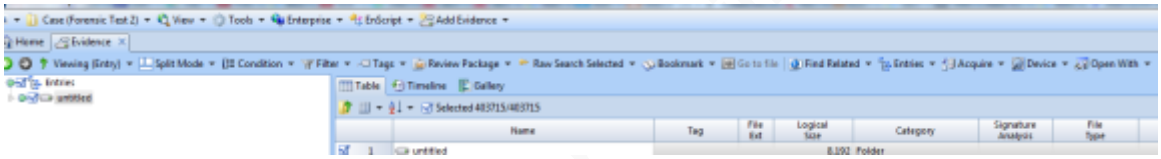

Figure 8 - Adding the Keyword List to EnCase 7 to search for hits in the forensic image backup3_3_16.E01

3. Under “Raw Search Selected,” the option “Add Keyword List” was used

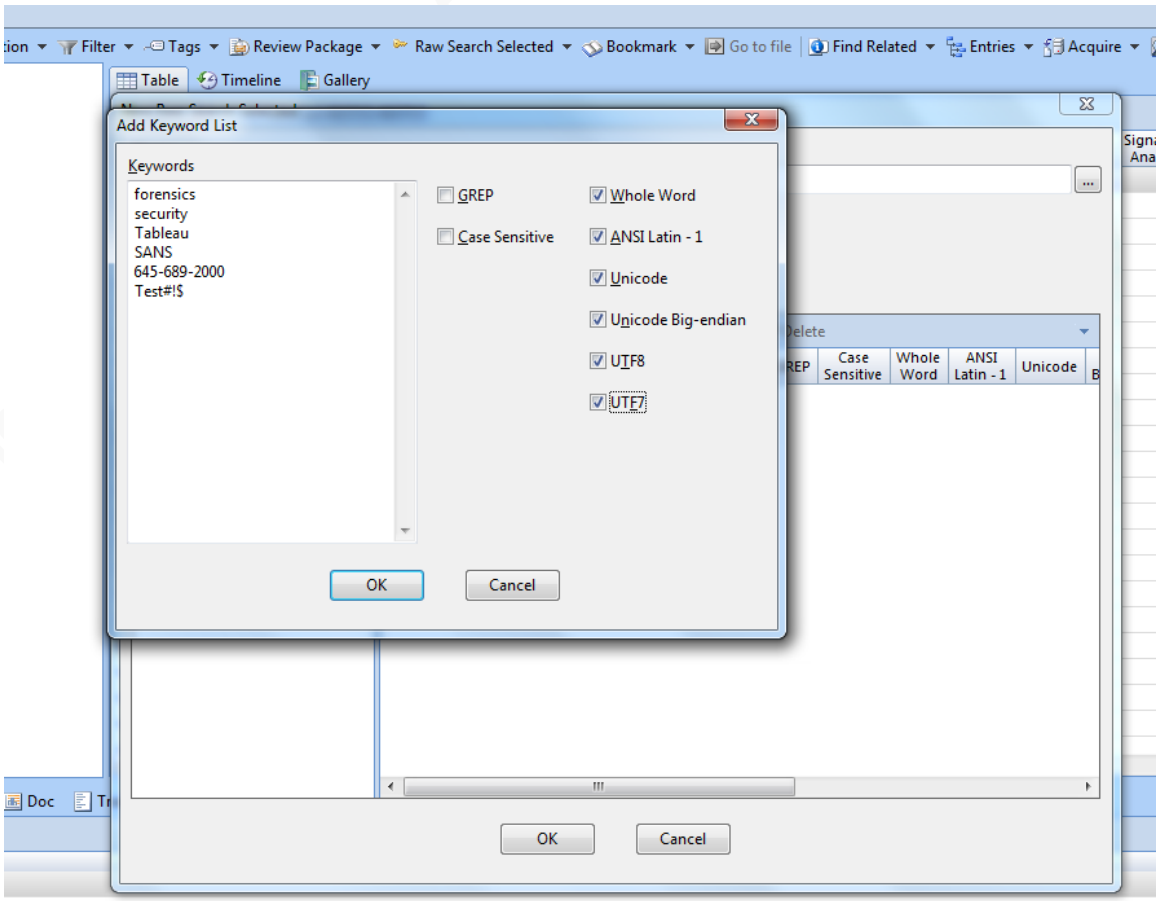

Figure 9 - Keyword Search Options

The search time duration was several hours because EnCase, as well as, other traditional forensic tools use a linear method for searching through data. The other complications with searching data in this manner are compressed artifacts, encrypted

data, and the way some artifacts are encoded. Finally, the available hardware used to run the search also contributed to duration.

EnCase 7 Cost:

Total Searching Cost = Cost Per Hour X 7 hrs (approximately)

$350/hr X 7 = $2450

3.1.1 ElasticSearch

Using the EnCase search experiment as a baseline, the same type of search was conducted under ElasticSearch using the same keywords:

1. Using the Ubuntu virtual machine to access ElasticSearch through Kibana, the following syntax was used

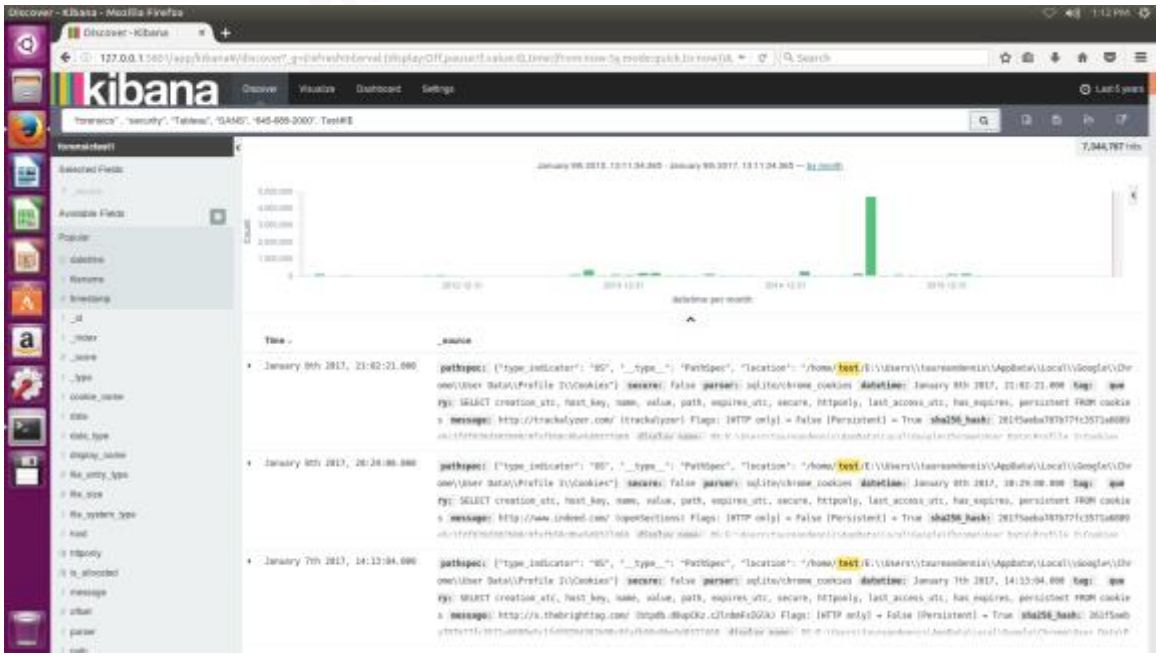

Figure 10 - The results in ElasticSearch through Kibana dashboard from the keyword search

2. The result of the search completed with 7,344,767 hits in under a minute; however, this is due to the data being indexed previously and the method Elasticsearch utilizes for searching through large datasets.

ElasticSearch Cost:

The client in this particular example, would not be charged a cost for searching due to how quick the procedure completed.

4. Overall Comparison

Analyzing the conducted experiments between the traditional forensic tool EnCase and the data analytic tool ElasticSearch, there are positives and negatives to take into account. The following categories must be taken into account for thorough comparison: Data Preparation, Indexing, Processing, and Search Method

4.1 Data Preparation

EnCase and many other traditional forensic tools were designed to ingest forensic images in formats that have been acceptable in the digital forensics community and as admissible evidence in court systems worldwide. Some of the formats that EnCase accepts are Expert Witness(E01, Ex01, and DD). These formats require no preparation to use with EnCase.

Comparing the data preparation phase with a data analytic tool, in this case, Tableau and ElasticSearch, the forensic image format must first be converted into a log format that is compatible with these tools using a procedure like Plaso that can extract the metadata and event information from the forensic image. The conversion process, unfortunately, adds time and cost to the preparation phase when using the data analytic tool.

4.1.1 Indexing

Indexing speed varies between tools for both traditional and data analytic tools. Focusing on EnCase again, the indexing speed took approximately 3 hrs to complete versus the two hours for ElasticSearch. When EnCase indexes data, it has to parse through data that is compressed, encrypted, or contain complex encoding. The type of hardware available ultimately determines the time it takes to complete.

ElasticSearch was faster at indexing because the file format was already prepared by the tool Psort into a text-based format that can be seamlessly ingested. The process completed in 2 hours.The indexing process for data analytic tools is much faster.

4.1.2 Processing

This phase involves traditional forensic tools parsing the actual artifacts in the evidence such as unzipping compressed artifacts, decrypting files, and opening compound files. This process can take a very long time depending on the size of the data, type of data, and the hardware the procedure is running on. Data Analytic tools also have to process evidence; however, the type of files involved are usually less complex log files that do not involve obstacles such as encryption thus making this procedure faster compared to traditional forensic tools.

4.1.3 Searching

As mentioned previously, this is where the vast difference shows when using data analytic tools versus traditional forensic tools. EnCase took approximately 7 hours to complete the search of the 250 GB image using the keywords provided due to the linear method used to search the data.

ElasticSearch was able to search the same amount of data in under a minute due to the format of the data it is compatible with along with the advanced algorithms that it utilizes. Elastic Search is the optimal choice for performing searches.

5. Conclusion

There is so much more that can be tested to determine the cost effectiveness of using data analytics over traditional forensics tools. The decision to use either of the two or both comes down to the particular situation.

An example, an E-Discovery case involving encrypted documents would need to involve the use of both solutions considering the company wants to cut costs. The decryption of the data would need to be run through a traditional forensics tool that has the capability remove the encryption. After decryption, the data is converted into a format that a data analytic tool can utilize for performing quick and powerful search capabilities saving an organization time and money.

The field of data analysis is continuing to evolve making data analytic tools even more attractive to choose over traditional forensic tools when dealing with large datasets.

Technologies such as artificial intelligence, predictive coding, voice recognition, and even augmented reality are changing the way large unstructured data is analyzed. The decreasing costs of faster computational power in regards to CPU speed, RAM Size, and storage make these emerging and evolving technologies easily accessible and affordable.

As data continues to become exponentially larger as the years progress, there will likely be a convergence of data analytic technologies implemented into the traditional forensic suites provided by corporations such as Guidance Software, Nuix, AccessData, and much more.

References

Fisher, T. (2016, April 26). What is a Sector? (Disk Sector Definition). Retrieved from https://www.lifewire.com/what-is-a-sector-2626003

Gartner Says That Consumers Will Store More Than a Third of Their Digital Content in the Cloud by 2016. (n.d.). Retrieved from http://www.gartner.com/newsroom/id/2060215

IBM - What is big data?. (n.d.). Retrieved from https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Maurer, M. (2016, June 26). *Distributed Forensic Timeline (DiFT)*. Retrieved from http://diftdisk.blogspot.com/

Scanlon, M. (2016, October 2). *Battling the digital forensic backlog through data deduplication*. Retrieved from https://arxiv.org/pdf/1610.00248.pdf

Thorton, D. (2016, November 22). *How agencies use big data to improve health, security, save money - FederalNewsRadio.com*. Retrieved from http://federalnewsradio.com/big-data/2016/11/agencies-use-big-data-improve-health-security-save-money/

Wong, W. (2016, January 6). *Schools Tap Big Data To Understand Trends | EdTech Magazine*. Retrieved from http://www.edtechmagazine.com/k12/article/2016/01/schools-tap-big-data-understand-trends