Improving the Efficiency of Big Forensic Data Analysis Using NoSQL

Md Baitul Al Sadi
Department of Information Technology
Georgia Southern University
Statesboro, GA 30458, USA
ms12508@georgiasouthern.edu

Hayden Wimmer
Department of Information Technology
Georgia Southern University
Statesboro, GA 30458, USA
hwimmer@georgiasouthern.edu

Lei Chen
Department of Information Technology
Georgia Southern University
Statesboro, GA 30458, USA
lchen@georgiasouthern.edu

Kai Wang
Department of Computer Science
Georgia Southern University
Statesboro, GA 30458, USA
kwang@georgiasouthern.edu

ABSTRACT

The rapid growth of Internet of Things (IoT) makes the task for digital forensic more difficult. At the same time, the data analyzing technology is also developing in a feasible pace. Where traditional Structured Query Language (SQL) is not adequate to analyze the data in an unstructured and semi-structured format, Not only Standard Query Language (NoSQL) unfastens the access to analyzing the data of all format. The large volume of data of IoTs turns into Big Data which just do not enhance the probability of attaining of evidence of an incident but make the investigation process more complex. This paper aims to analyze Big Data for Digital Forensic (DF) investigation using NoSQL. MongoDB has been used to analyze Big Forensic Data in the form of document-oriented database. The proposed solution is capable of analyzing Big Forensic Data in the form of NoSQL more specifically document oriented data in a cost-effective, efficient way as all the tools is being used are open source.

KEYWORDS

Digital Forensic (DF), NoSQL, Big Data, Big Data Forensic, MongoDB, Document-oriented Database, Autopsy, Internet of Things (IoT).

1 INTRODUCTION

As the use of Information Technology is rapidly progressing, the risk factor in this arena is also running in the same rhythm. Now, human civilization is depending on Information Technology significantly. Although people are enjoying the blessing of this sector, they are also experiencing difficulties when it comes to concern about Information Security. Today's tech dependent life can be annihilated within an instance if someone fails to protect the resources appropriately. Even a small backdoor may cause immense detriment. To make digital resources secure, determination of the loophole is the most signification step. When a victim loses valuable resources, it is required to investigate to determine the root cause(s). To learn the precise reason, it is the best way to conduct the digital forensic on a certain case. Digital Forensic (DF) is not only applicable to unbolt the cybercrime issues but also useful in case of regular criminal activity. This paper is intended to illustrate the digital forensic approach and to analyze them in NoSQL (Not Only SQL) database. There is a variety of tools available including Autopsy, EnCase, Foremost, FTK, Registry Recon, PTK Forensics, The Sleuth Kit, The Coroner's Toolkit, COFEE etc. to extract data from IoT devices. The extracted data will be in an unstructured format, hence NoSQL is the best solution to analyze them. Here the document-oriented database program, MongoDB has been chosen to analyze the data from Internet of Things (IoT). To our best knowledge this is pioneer work in terms of using NoSQL and MongoDB for DF.

2 BACKGROUND

2.1 Digital Forensic

There is no alternative during an investigation of a cybercrime than to perform the Digital Forensic (DF) process in order to identify the actual incident. However, sometimes regular events may demand DF investigation. Although it is directly related to cybercrime, it is useful for general crime as well. For example, the usage of IoT (Internet of things) is growing exponentially, and most of the appliances (whether it is household or official) nowadays are generated logs and connected to the internet to store them. These records may enhance the probabilities to find out some relevant information regarding an incidence.

The first-word "Digital" in DF is used to refer all sort of Digital means. It may include from small memory devices to a whole cloud system or even discrete sources like social media. On the other hand, "Forensic" leads to collect the data, preserve for further usage and analyze them for criminal investigation. Overall, DF is a field where Digital means are used to collect the data, to keep them in digital storage, to process and analyze them to use certain digital tools and to publish the results. According to the National Institute of Standard and Technology (NIST), Digital forensic can be defined as the application of science which is related to the law to identify, to collect, examine and to analyze the data by preserving the data integrity and by keeping it secure.

2.2 Big Data

Big data directs the data which are gigantic in terms of volume, which claims for very quick processing time, and there is the

MOBIMEDIA 2017, July 13-14, Chongqing, People's Republic of China

DOI 10.4108/eai.13-7-2017.2270344

inadequacy of tools to process the data. The two primary factors that describe the Big Data are:

1) The velocity of generating a variety of data;
2) preserving this huge volume of data by maintaining its veracity and enhancing the ability to analyze them by converting into a value.

Big Data can be defined with five Vs. They are:

Volume: Volume means the enormous quantity of data. According to vcloudnews.com [12], 2.5 Quintillion Byte of data create every day. In last two years, 90% of the whole world's data has been created. It is not very hard to assume how big the data volume is going to be created shortly. With the help of Big Data, it is possible to analyze them whether they are in a distributed system or located in different geographical position.

Velocity: In Big Data, velocity means the rate of new data creation and the rate of data transmission from one system to another with respect to time. Big data provides the opportunity to analyze the data while generation process takes place. It does not require to store data in databases.

Variety: Variety refers to the variation of data sources and data format. There is an infinite number of data sources and a variety of data format. The Relational Database Management System (RDMS) are only capable of handling structured data, where big data have the ability to analyze all sorts of data, including structured, semi-structured and unstructured data.

Veracity: Veracity refers to the integrity and accuracy of data in Big Data. As generally in the huge volume of data, as well as a variety of data with different sources and formats, are required to handle, it is sometimes difficult to maintain its accuracy.

Value: In big data, value means the ability to convert data or information into value. Data must be quantifiable to an enterprise or an organization.

Information security experts are facing difficulties to analyze logs, network flows, system events for forensic and intrusion detection as the traditional technology does not have enough provision or tools to analyze long-term and large-scale data as: retaining large quantities of data was not economically feasible before. Big data technologies, such as the Hadoop ecosystem (including Pig, Hive, Mahout, and RHadoop), stream mining, complex event processing and NoSQL databases—are allowing the analysis of heterogeneous datasets at unprecedented scales, speeds and large-scale. These technologies enhance security analytics by facilitating the storage, maintenance, and analysis of security information. For instance, the WINE platform1 and BotCloud2 allow the use of MapReduce to process data efficiently for security analysis. Challenges are 1) Privacy: Privacy can not be possible to protect when data need to share among industries as well as with law enforcement, which goes against the privacy principle of avoiding data reuse meaning, using data only for the purposes that it collected; 2) the data provenance problem as big data lets us expand the data sources we use for processing; it is hard to be satisfied that each data source meets the trustworthiness that our analysis algorithms require to produce accurate results. As big data cannot consider as a panacea, security experts should keep investigating to find out appropriate ways to tackle sophisticated attackers from eternal arms race of attack and defense. Big data has the capability to create the world where maintaining control over the revelation of our personal information is always challenged [1].

2.3 NoSQL and Big Data

NoSQL means Not only Structured Query Language. It implies all three classes of the database: structured, semi-structured and unstructured database. The growth rate of computer storage consents the generation of a large amount of data. This exponential growth rate of data increases not only the volume of data but also its fashions. Sometimes, it becomes mandatory to analyze data for various purposes like critical business activities, environmental facts, public opinion, digital forensic, health services, social facts, etc. The conventional Relational Database Management System (RDMS) is capable of analyzing structured data only. The popularity of NoSQL inflates because of the capabilities of handling all sort of database including structured, semi-structured and unstructured. The mentionable feature which is provided by NoSQL is the scalability. NoSQL is also capable of handling big data with replication and distribution over multiple servers. NoSQL databases can categorize into four categories.

Key Value pair Database: The Key-Value Database maps a value as a key to a set of value. The set of values may contain a Null value (empty), a single value or multiple values. It does not matter what the nature of value is and how it organizes.

Document Database: Sometimes it is considered as a particular type of key-value pair database [9]. In document-oriented database, data is stored in XML (EXtensible Markup Language), JSON (JavaScript Object Notation) and BSON (Binary encoded JSON).

Column-Family Database: In column family database, it stores each table entry associated with a particular raw ID. In some cases, it is not possible to store all data in a certain sequence. Then, the data maps to a single key. Column family database is very well known because Google big table is implemented using column family structure.

Graph Database: A Graph Database or Graph-Oriented Database stores, maps and queries the relationships by using graph theory. This NoSQL database mainly consists of nodes, edges and properties. To analyze interconnection of the database that contains complex correlation like a social network or Supply Chain Management the graph database is the well-suited one.

A thorough analysis on six of the most popular NoSQL databases including MongoDB, HBase, Redis, Cassandra, CouchDB and DynamoDB is performed to highlight the features, strengths, and limitations. If data exists in the form of hierarchies and is in XML format and high level of consistency in operations is required, MongoDB would provide the best solution. If an organization majorly produces unstructured data at a high velocity and high performance in terms of processing random read/write requests is required, then Redis would be the best solution. Also if a large volume of data for each transaction is expected to be processed then

Cassandra would be able to deliver consistent performance and would be linearly scalable. [4]

The capacity of big data is beyond only storing or querying the data from a large dataset. Besides, big data has the ability to perform sophisticated analysis on the dataset and the insight value of the dataset. As discussed before, all three classes of data including structured, semi-structured and unstructured data fall into NoSQL. In a digital forensic investigation, it is very likely that it may include in an individual dataset to analyze all three classes of data at the same time. Therefore, big data is using NoSQL the appropriate platform to analyze this verity of the dataset for a digital forensic investigation.

2.4 Digital Forensics and Big Data

As the existing tools and infrastructures cannot meet the expected response time when an investigator deal with a big dataset, a conceptual model is proposed in [6] for supporting big data forensics investigations and present several use cases, where big data forensics can provide new insights to determine facts about criminal incidents. The main challenges for digital forensic investigating in big data are Identification, Collection, Organization, and Presentation. To identify cyber criminals utilizing IoT devices in criminal case investigation, Digital information is now growing beyond the capacity of current digital forensics tools and procedures.

For the new generation database NoSQL, authors [7] attempted to analyze and assess the level of maturity of NoSQL databases through the lens of CIA triad. Survey has conducted on Access Control / Authentication, Extended Security Features, Encryption of data, Concurrency Control, Domain Constraints, Structural Constraints, Cloud Security and adaptability, and Attacks to analyze the impact on CIA triad among Oracle Database, MySQL Database, Hadoop, and MongoDB. By enabling and aligning CIA triad, NoSQL Database is evolving at a great speed [7].

As the amount of the data is rapidly increasing, cloud-like intensive computing process and gigantic storage enable to facilitate Digital Forensic as a Service. A design proposed in [8] details server and client architecture where the client application could be a cloud-based system and server uses Apache Hadoop platform for creating distributed processing environment in master-slave forms. Prior to archiving evidence data in NAS (Network Accessed Storages), it is required to accomplish an ETL (Extract, Transform, and Load) process. Hbase based index database provides facility to the investigator to indexed forensic search from remote to index terms and meaningful pattern of digital evidence [8].

2.5 Digital Forensic Process and Tools

The methodology of data collection from Android-based devices describes which enables a minimal data corruption as well as omission by special boot mode for extraction of evidence comprehensively. The methodology of evidence collection process must adhere following criteria: Data preservation, Atomic Collection, Correctness, Determinism, Evidence Preservation, Usability, Vetting Ability and Reproducibility. The evidence collection process, which uses Android Debug Bridge (ADB) is able to duplicate data from Android to Computer. The recovery method of Android devices has two key components: 1) Recovery image (generalized component: Header, Kernel, and Ramdisk) and 2) Flashing tools. Different operating systems have altered names and data structures for the recovery image and various techniques for image flashing, but the general methodology of using the recovery mode for collection will work for all modern mobile devices. [9]

Two popular public cloud service providers (Microsoft One Drive and Amazon Cloud Drive) have been used to perform digital forensic Investigation. Probable evidence may be found from timestamps, file hashes, client software log files, memory captures, link files and other evidence from OS, Hardware, RAM and application data. Most significant artifacts are stored in log files, database files, system configuration and setup files, which give information about every action performed within Mega Cloud Drive and One Drive [7].

Authors in [10] were intended to design a novel model to find out the best approach from a digital forensic investigation from the IoT. For IoT forensic, the most popular models are- a) Digital Forensic Investigation Model (DFIM), b) The Hybrid Model and c) 1-2-3 Zones of Digital Forensic. The proposed model in [10] is based on Machine-to-Machine (M2M) communication. Digital forensic procedure would be a chain of custody, lab analysis, result, proof and defense and archive & storage as all these stages are more to the existing method of conducting digital forensic. The potential autonomy of IoT or lack of control over IoT by those it impacts will be doubtless to generate IoT adoption resistance potentially manifested by public protests, negative publicity campaigns, and actions by governments. [10]

According to US Department of Justice, to deal with electronic means as evidence in forensic, there are some principles to follow. They are: 1) for any action taken to keep evidence secure and to collect data, the information of proof shouldn’t be changed, 2) only well-trained digital forensic personnel are permitted to lead the examination of digital measures, 3) proper documentation needs during seizing, storing, transferring and examine the electronic evidence and it should be stored securely for review anytime.

In every step, there is no predefined rule or protocol to follow but some general rule. Actions need to be performed as it demands. To extract data from IoT devices, a variety of tools listed in Table-1 are available, including Autopsy, EnCase, Foremost, FTK, Registry Recon, PTK Forensics, The Sleuth Kit, The Coroner's Toolkit, COFEE, etc. The extracted data will be in an unstructured format, hence NoSQL is the best solution to analyze them.

Autopsy is a very useful and easily manageable digital forensic tool. One is required to create a new case to investigate a device using Autopsy. A new case may involve multiple data sources. Autopsy can take data from three type of data sources: 1) image or VM file, 2) local disk, and 3) logical file. The format of disk images supported by Autopsy are listed below:

Raw Single (For example: *.img, *.dd, *.raw, *.bin)
Raw Split (For example: *.001, *.002, *.aa, *.ab, etc)
EnCase (For example: *.e01, *.e02, etc)
Virtual Machines (For example: *.vmdk, *.vhd)

There is a range of Ingest Module in Autopsy which is responsible for analyzing specific activities. They are as following

Recent Activity
Hash Lookup
File Type Identification
Embedded File Extractor
Exif Parser
Keyword search
Email Parser
Extension Mismatch Detector
E01 verifier
Android Analyzer
Interesting File Identifier
PhotoRec Carver
Virtual Machine Extractor

Autopsy can generate six particular type of report format including HTML, EXCEL, TEXT, GOOGLE EARTH / KML, STIX and TSK BODY FILE.

3 SYSTEM ARCHITECTURE

The System Architecture involves the following steps (figure-1):

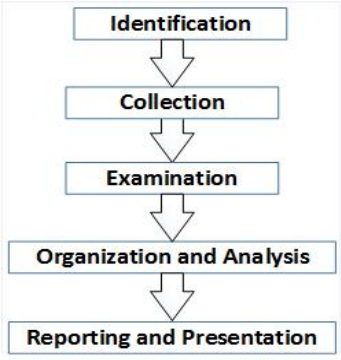

Figure 1: Steps in a Digital Forensic (DF) Investigation

Step-1: Identification

The most important and critical part is to identify the digital data sources from which data is to be collected. The identification process depends on the incident. The investigator must make decision according to the type of event and available resources. In the process, it is very critical to keep documentation of every single move. Identification of wrong devices may lead the whole investigation process in a wrong direction, and it can affect the time frame of the entire process.

Step-2: Collection

The data collection process is the most sensitive step. The investigator must be very careful that the process to extract data from the source is not making any change to the source and the investigator has to preserve the data for future use. There are a lot of tools as shown in table-1 to extract data from identified sources. The data collection process is entirely dependent on the type of the sources and how it is going to be analyzed. As it is a very sensitive process, only well-trained persons should be allowed to collect data from identified sources. To collect data from mobile devices, backing up of all Android, iPhone and Windows phone. The procedure to make the backup may vary vendor to vendor or even model to model of the device.

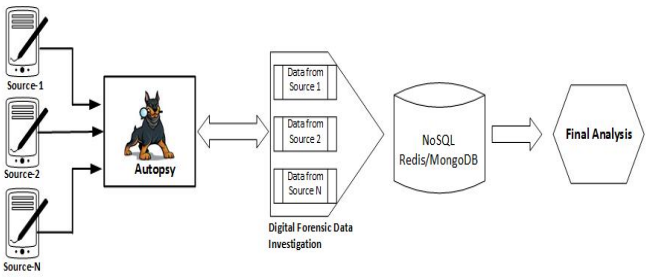

Figure 2: System Architecture of Digital Forensic (DF) Process

Step-3: Examination

When an investigator has all the acquired data, which are required to investigate in a correct format, it is then ready for examination. There is a good range of tools to analyze the collected data. Tools have to be selected according to the nature of the data and the examination process. In our case, the tool Autopsy has been used to extract and examine data from mobile devices. Autopsy enables an extensive data lookup process, which was discussed in the earlier section. The data, which are acquired by the Autopsy, is fallen on the NoSQL (Not only Structure Language) category.

Step-4: Organization and Analysis

After examining, the NoSQL data are required to be analyzed. Several NoSQL database packages are available to analyze the data. According to the orientation of the data and the analyzing process, appropriate NoSQL package has to determine. Before the analyzing process, data might require for being organized to compatible the analyzing data with the NoSQL package. In this case, MongoDB or Redis is used.

Step-5: Reporting and Presentation

The goal for the Forensic Investigation is to finalize a concrete result. Following the preceding steps, an investigator will have a result. The results are needed to present in a way by which anyone,

even who has limited knowledge in the field of digital forensic, will understand every step of the process. As this report and presentation will represent the whole investigation, it should be conveyed every small steps and information in a precise way.

4 PERFORMANCE ANALYSIS

Big data forensics based on HDFS (Hadoop Distributed File System) and Cloud Correlate different data sets enable the opportunity for identifying many new insights that were not possible before [6]. The dimension of digital evidence has grown exponentially with heterogeneous data. Traditional relational database management systems (RDBMS) typically expose a query interface based on SQL (Structured Query Language). However, these mainly require for management of structured data and it is difficult to scale to the ever growing size of data sets. Authors in [2] mainly investigate the performance of two NoSQL application MongoDB and Riak. Primary investigations conducted in [2] by Read Operations, where it tests the scalability of read operations with the variation of the size of the dataset. For read operation, when data set is less than 22.5GB the performance of MongoDB is better then Riak but when the dataset is larger than 22.5GB the performance of the Riak goes better. This performance is measured in terms of average read latency in millisecond (ms). For balanced read and update operations (which also measures in terms of latency in ms) MongoDB is good while the volume of the dataset is less than 16GB to 20GB and Riak performs better in dealing with large datasets, but MongoDB outperforms Riak on reasonably small datasets due to it's in-memory processing [2].

A thorough analysis on six of the most popular NoSQL databases, including MongoDB, HBase, Redis, Cassandra, CouchDB and DynamoDB is performed to highlight the features, strengths, and limitations [4]. If data exists in the form of hierarchies and is in a semi-structured format like XML format and high level of consistency in operations is required, MongoDB would provide the best solution [4]. If an organization majorly produces unstructured data at a high velocity and high performance in terms of processing random read/write requests is required, then Redis would be the best solution [4]. Also if a large volume of data for each transaction is expected to be processed then Cassandra would be able to deliver consistent performance and would be linearly scalable [4]. MongoDB is best to use when it is required to have high scalability and caching operation, as a replacement of web application which uses RDMS and for managing the contact for the semistructured database [4].

To compare read, write, delete, and instantiate operations on key-value stores implemented by NoSQL and SQL databases, consider the properties like scalability, consistency, support for data models, support for queries, and Management tools. Of the NoSQL databases, RavenDB and CouchDB do not perform well in read, write and delete operations [5]. Casandra is slow on read operations, but is reasonably good for write and delete operations [5]. Couchbase and MongoDB are the fastest two overall for read, write and delete operations. Couchbase, however, does not support fetching all the keys (or values) [5]. If iterating through keys and values is not required for an application, then Couchbase will be a right choice. Otherwise one may choose MongoDB who comes the close second to Couchbase in the read, write, and delete operations [5].

Authors in [3] reviewed the features of NoSQL database technologies and performed a comparison between RDBMS (i.e., MySQL) and NoSQL (i.e., MangoDB and Riak) by considering Data Replication and Data Sharding, Consistency according to CAP Theorem and Quorums. The goal of the tests was to investigate the scalability of the three databases as the size of the dataset increases for a cluster with a constant number of nodes and established that, all databases were loaded step by step until a total load of 50 GB reached. In Insertion operation test MongoDB and MySQL perform much better than Riak. For insertion operation, MongoDB and MySQL take 3 to 5 milliseconds where Riak takes more than 10 milliseconds. In Read Operation, both the NoSQL (MongoDB and Riak) performs far better than MySQL [3].

As data volume getting significantly large, the analysis for the investigation become more difficult. Forensic Cloud Environment (FCE) may facilitate forensic analysis of Big Data case. An experiment has performed in [11] which consists of Ingester, Interchange Framework, Hadoop cloud, Worker, Intelligent sharing framework and concludes feasible and it can be used to improve the turnaround time for investigation [11].

Table-1: Available Digital Forensic (DF) Tools

Tools	Websites	Free or Commercial
Autopsy	www.autopsy.com	Open Source
EnCase	www.guidancesoftware.com/encase-forensic	Need Subscription
Foremost	foremost.sourceforge.net/	Open source
FTK	accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk	Need Subscription
REGISTRY RECON	arsenalrecon.com/apps/recon	Need to purchase
The Sleuth Kit (TSK)	www.sleuthkit.org/	Open source
PTK	wiki.sleuthkit.org/index.php?title=PTK	Commercial
The Coroner's Toolkit	www.porcupine.org/forensics/tct.html	IBM Public License
Microsoft COFEE	cofee.nw3c.org/	Commercial

Table-1 shows the available tools for DF. Most of the tools including EnCase, FTK, REGISTRYRECON and Microsoft

COFEE are required to be purchased or subscribed. Where Autopsy is a GUI based open source program, and it allows efficiently analyze the hard drive and smartphone backup. Its plugin architecture supports to find add-on modules or develop custom modules in Java or Python. Foremost is an open source tool with programming difficulties which limits on processing the file within 2 gigabytes.

Guymager: Guymager is an open source software which enables to make images of electronic storage devices with a high speed in various formats like EWF (Expert Witness Disk Image Format), AFF (Advance Forensic Format) and most popular dd format. This QT based image maker is specially designed for forensic investigator with a built-in hash calculator (both MD5 and SHA256). Guymager is known as a forensic media acquisition program. Fig-3 shows a glance of the process of preparing an image using Guymager.

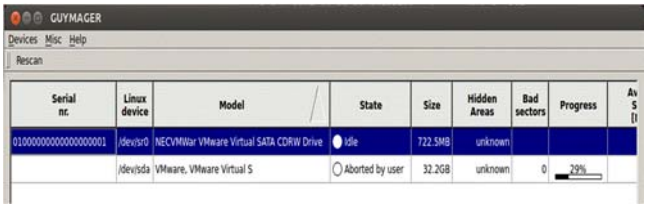

Figure 3: Process of preparing image using Guymager

Foremost: Foremost, a Linux-based open source console forensic application, is a powerful file carving tool which recover the file based on the file header, file footer and the data structure, works on dd, EnCase, Safeback or on the drive directly. It is developed by the Air Force of United States for a special Investigation, and later on it is open for public use. Foremost mainly works based on a customizable configuration file. According to the parameters set in configuration file, it looks up the header, footer or the structure of the data. Foremost recover files of different formats including .gif, .jpg, .bmp, .pnp, .exe, .avi, .wav, .mpg, .html, .pdf, .mov, .zap, .rar and so on from an image. When the size of the storage medium becomes gigantic, automatic file processing tools like foremost enriches the efficiency of the DF process. Fig-4 is showing the output after performing the file curving using foremost.

```
sadi@ubuntu:~$ ls out_put_01
audit.txt  dll   exe  jar  mov  ole  ppt   rif  sxc  vis  xls
avi        doc   gif  jpg  mp4  pdf  pptx  sdw  sxi  wav  xlsx
bmp        docx  htm  mbd  mpg  png  rar   sx   sxw  wmv  zip
```

Figure 4: The output of the Foremost after the process of file carving

Compared to Foremost, Autopsy is more convenient to use as it has a powerful GUI. At the same time, Autopsy can be used in both Windows and Linux Operation Systems. Where, on the other hand, Foremost has only Command Line Interface (CLI), and it can only be operated in Linux environment. The influential module based system like TSK (The Sleuth Kit) adds influential potential to Autopsy. Although foremost doesn't have such module, its robustly customizable configuration file enables dynamic feature during the file carving process. In a forensic Investigation process, the final and comprehensive part should be the documentation and the presentation. Autopsy is proficient in generating a wide range of report which is already described in the section 2.7. On the other hand, foremost has only one form of a report which is in text format. Autopsy is not only a decent tool for preparing a report, but also it enables a platform for a high-level analysis with a deep level of raw data including hex, strings, file metadata, indexed text, media content and so on. Some of the relevant data are illustrated in Fig-5(a-d).

Drastically, Autopsy is a very efficient open source tool comparing any other open source Digital Forensic tools because of its competencies to analyze the evidence of DF investigation objects. From very low-level hexadecimal data, to metadata in an extended form, it provides the visualization of the multimedia data. In Autopsy, the Indexed Text provides subtle details about a file. For example, the image file which is shown in Fig-5d provides the information like aperture value of the lens or the speed of the device during the shooting of the photo and so on which is shown in Fig-5c. This information can play a dynamic role during a DF Investigation.

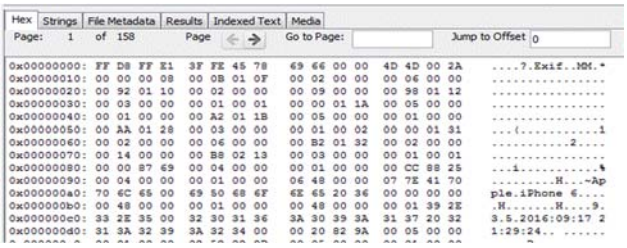

Figure 5a: Autopsy allows to view intensive level (HEX) of the data of a file

Figure 5b: Autopsy provide the opportunity access detail metadata of file

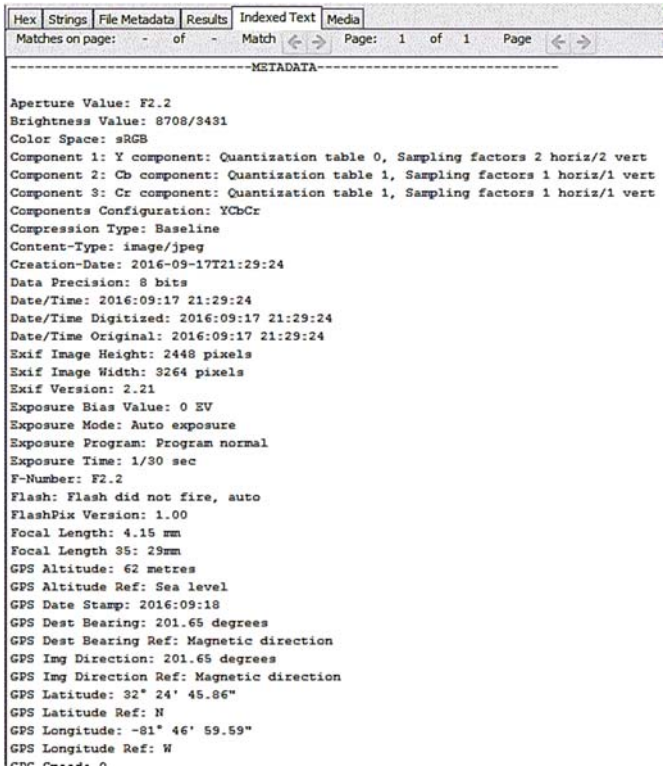

Figure 5c: The Indexed text of Autopsy provide details information about a file.

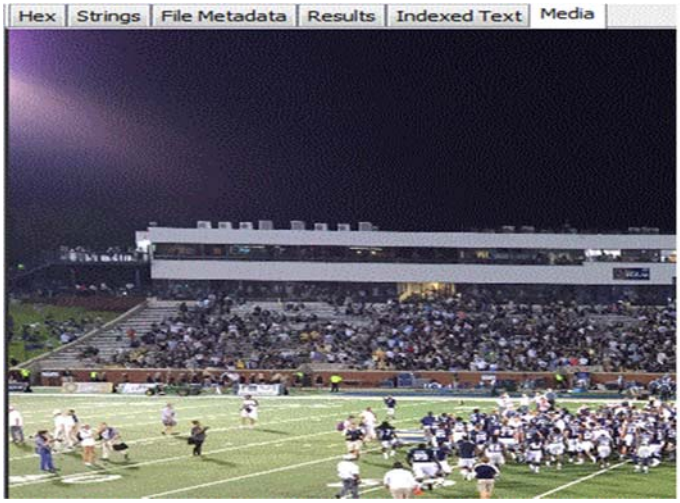

Figure 5d: If the file is a multimedia file Autopsy enables a preview with the appropriate codec.

In the proposed architecture, the two most primary tools have been used are Autopsy and MongoDB. Both tools are open source and efficient when compared to other tools and their respective tasks. The combination of these two tools makes the proposed architecture most competent concerning the efficiency of analyzing data, presenting the result of the analysis and finally, finding co-relation among them which will be emphasized more in the next research work. And of course, as both of the tools are open source, there is no cost to using this architecture.

5 DEMONSTRATION

Once data has been collected and examined from Autopsy data is ready for analysis. After the completion of data examination, Autopsy can generate report in different format, including HTML (HyperText Markup Language), Excel, Text, Google Earth/KML, STIX (Structures Threat Information Expression) and TSK body file format. This paper is intended to analyze the data using MongoDB database. For this case, the result exports in Excel format. MongoDB is an appropriate platform for analyzing document-oriented NoSQL Database. In this research, raw data of individual mobile devices has been selected to examine. Here raw data refers to the data including application data, multimedia data, call history, contact information, EXIF (Exchangeable Image File Format) metadata, email addresses, email messages, etc. Data of the given formats are best to explore in the document-oriented database. Hence, MongoDB has been selected to analyze them. MongoDB uses JSON (JavaScript Object Notation) format in its database. Therefore, Autopsy generated a report in Excel format (Fig-6) and has to be converted into JSON format. As the data is taken from the mobile devices, the generated report contains the metadata of photos taken from the mobile device. As a result, the report contains fields like the date taken when the photos was captured, device manufacturer, device model, latitude, longitude, altitude of the location where the photo was captured and source file.

	A	B	C	D	E	F	
1	Date Taken	Device Manufacturer	Device Model	Latitude	Longitude	Altitude	Source File
2	2016-04-25 04:48:24 EDT	Apple	iPhone 6	23.73012222	90.41490278	17.16372	/LogicalFileSet3/Backup/c1804a9057254079e34
3	2016-04-25 04:48:55 EDT	Apple	iPhone 6	23.73037222	90.41500833	17.14544	/LogicalFileSet3/Backup/c1804a9057254079e34
4	2016-05-06 11:31:59 EDT	Apple	iPhone 6	23.74476944	90.39998611	16.77457	/LogicalFileSet3/Backup/c1804a9057254079e34
5	2016-05-06 11:32:14 EDT	Apple	iPhone 6	23.74478611	90.39995556	16.77457	/LogicalFileSet3/Backup/c1804a9057254079e34
6	2016-05-06 11:32:15 EDT	Apple	iPhone 6	23.74478611	90.39995556	16.77457	/LogicalFileSet3/Backup/c1804a9057254079e34
7	2016-05-06 11:32:28 EDT	Apple	iPhone 6	23.74480833	90.39979444	16.97429	/LogicalFileSet3/Backup/c1804a9057254079e34
8	2016-05-06 11:32:36 EDT	Apple	iPhone 6	23.74482222	90.39977222	16.97429	/LogicalFileSet3/Backup/c1804a9057254079e34
9	2016-05-06 11:33:01 EDT	Apple	iPhone 6	23.74485556	90.39958056	17.12072	/LogicalFileSet3/Backup/c1804a9057254079e34
10	2016-05-06 11:33:02 EDT	Apple	iPhone 6	23.74494444	90.39942222	17.12072	/LogicalFileSet3/Backup/c1804a9057254079e34
11	2016-05-06 11:33:06 EDT	Apple	iPhone 6	23.74494444	90.39942222	17.12072	/LogicalFileSet3/Backup/c1804a9057254079e34
12	2016-05-06 11:33:09 EDT	Apple	iPhone 6	23.74501389	90.39933056	17.12072	/LogicalFileSet3/Backup/c1804a9057254079e34
13	2016-05-06 11:33:23 EDT	Apple	iPhone 6	23.74501389	90.39933056	17.12072	/LogicalFileSet3/Backup/c1804a9057254079e34
14	2016-05-06 11:33:40 EDT	Apple	iPhone 6	23.74486944	90.39954444	17.23105	/LogicalFileSet3/Backup/c1804a9057254079e34

Figure 6: Report from Autopsy in EXCEL format

The converted form in JSON format of Autopsy report is given below:


```
{
    "Date Taken" : "2016-04-25 04:48:24 EDT",
    "Device Manufacturer" : "Apple",
    "Device Model" : "iPhone 6",
    "Latitude" : 23.73012222,
    "Longitude" : 90.41490278,
    "Altitude" : 17.16372392,
    "SourceFile":
"/LogicalFileSet3/Backup/c1804a9057254079e34
```

```
af70c253059b9c986e4bb/d46c91af097d8ea8c13f56
bb0ea882325dc7d0e9",
    "Tags" : null
}
```

Next, the data in JSON format is inserted into MongoDB NoSQL database. During the data insertion process in MongoDB, a unique ID called object ID is added. Object ID consists of a 12-bit hexadecimal number. The four first bytes represent the time stand; the next three bytes is a computer generated a unique identifier, following 2 bytes represent another unique ID based on the process and last three bytes is an incremental ID.The structure of the Object ID is shown in Fig-7.

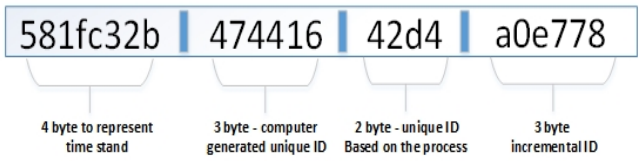

Figure 7: Structure of Object ID

After the data insertion, the data in MongoDB looks like below, including insertion of some important commands of MongoDB:

In MongoDB, a database is created with the name of "autopsydb." The following command in Fig-8 is used to switch from one database to another.

```
use autopsydb
```

```
MongoDB Enterprise > use autopsydb
switched to db autopsydb
MongoDB Enterprise >
```

Figure 8: Use of specific database or switch database in MongoDB

The data insertion command in MongoDB is given in below:

```
db.device1.insert([
{"Date Taken":"2016-04-25 04:48:24 EDT",
 "Device Manufacturer":"Apple",
 "Device Model":"iPhone 6",
"Latitude":23.73012222,
"Longitude":90.41490278,
"Altitude":17.16372392,
"SourceFile":"/LogicalFileSet3/Backup/c1804a
9057254079e34af70c253059b
9c986e4bb/d46c91af097d8ea8c13f56bb0ea882325d
c7d0e9","Tags":null
}])
```

Once the data has inserted into the database, the data could be found by the command "db.device1.find()" shows in Fig-9(a) and the data will arrange in a better way if the command db.device1.find().pretty()is inserted shows in Fig-9(b)

```
db.device1.find()
```

```
MongoDB Enterprise > db.device1.find({"Date Taken" : "2016-04-25 04:48:24 EDT"})
{ "_id" : ObjectId("5828def247441642d4a8e799"), "Date Taken" : "2016-04-25 04:48:24 EDT", "D
e Manufacturer" : "Apple", "Device Model" : "iPhone 6", "Latitude" : 23.73012222, "Longitude
90.41490278, "Altitude" : 17.16372392, "SourceFile" : "/LogicalFileSet3/Backup/c1804a9057254
34af70c253059b9c986e4bb/d46c91af097d8ea8c13f56bb0ea882325dc7d0e9", "Tags" : null }
MongoDB Enterprise >
```

(a)

```
db.device1.find().pretty()
```

```
MongoDB Enterprise > db.device1.find({"Date Taken" : "2016-05-06 11:38:04 EDT"}).pretty()
{
        "_id" : ObjectId("581fc32b47441642d4a0e78b"),
        "Date Taken" : "2016-05-06 11:38:04 EDT",
        "Device Manufacturer" : "Apple",
        "Device Model" : "iPhone 6",
        "Latitude" : 23.74497222,
        "Longitude" : 90.39989444,
        "Altitude" : 17.59643436,
        "Source File" : "/LogicalFileSet3/Backup/c1804a9057254079e34af70c253059b9c986e4bb/b1
0603700bdf7719c3a65b22ca2f12715d37",
        "Tags" : null
}
```

(b)

Figure 9: Commands for data searching

6 CONCLUSION AND FUTURE DIRECTION

The primary challenge for digital forensic investigation is to identify the appropriate data source for evidence from plenty of suspected sources. After the identification of the appropriate source, data are needed to be extracted from the source. The tool Autopsy has been used to extract data. After the extraction of data in Excel (.xls) format, data were converted into JSON format. MongoDB has been chosen to analyze the data. Both of the tools Autopsy and MongoDB are open source and free to use. Hence, the approach that used throughout this research is cost-effective. The technical task starting from data extraction by Autopsy to data insertion to MongoDB is straight-forward. At this point, data collection and insertion of data into NoSQL have been completed. This research work will be extended to analyze more data to identify the evidence from digital sources that lead to completing a case study which will involve all essential steps of Digital Forensic.

REFERENCES

[1] A. Cárdenas, P. K. Manadhata and S. P. Rajan. 2013. Big Data Analytics for Security. In *IEEE Security & Privacy*, vol. 11, no. 6, 74-76, doi: 10.1109/MSP.2013.138

[2] M. Qi. 2014. Digital forensics and NoSQL databases. In *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Xiamen, 734-739. doi: 10.1109/FSKD.2014.6980927

[3] M. Qi, Y. Liu, L. Lu, J. Liu and M. Li. 2014. Big Data Management in Digital Forensics. In 2014 IEEE 17th International Conference on Computational Science and Engineering, Chengdu, 238-243. doi: 10.1109/CSE. 2014.74

[4] P. P. Srivastava, S. Goyal and A. Kumar. 2015. Analysis of various NoSql database. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 539-544. doi: 10.1109/ICGCIoT.2015.7380523

[5] Y. Li and S. Manoharan. 2013. A performance comparison of SQL and NoSQL databases. In 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Victoria, BC, 15-19. doi: 10.1109/PACRIM.2013.6625441

[6] S. Zawoad and R. Hasan. 2015. Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities. In 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, New York, NY, 1320-1325. doi: 10.1109/HPCC-CSS-ICESS.2015.305

[7] S. Srinivas and A. Nair. 2015. Security maturity in NoSQL databases - are they secure enough to haul the modern IT applications?. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 739-744. doi: 10.1109/ICACCI.2015.7275699

[8] J. Lee and S. Un. 2012. Digital forensics as a service: A case study of forensic indexed search. *In* 2012 International Conference on ICT Convergence (ICTC), Jeju Island, 499-503. doi: 10.1109/ICTC.2012.6387185

[9] D. Votipka, T. Vidas and N. Christin. 2013. Passe-Partout: A General Collection Methodology for Android Devices. In IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, 1937-1946, doi: 10.1109/TIFS.2013.2285360

[10] S. Perumal, N. M. Norwawi and V. Raman. 2015. Internet of Things (IoT) digital forensic investigation model: Top-down forensic approach methodology. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC), Sierre, 19-23. doi: 10.1109/ICDIPC.2015.7323000

[11] O. Tabona and A. Blyth. 2016. A forensic cloud environment to address the big data challenge in digital forensics. In 2016 SAI Computing Conference (SAI), London, 579-584. doi: 10.1109/SAI.2016.7556039

[12] 2.5 Quintillion Bytes of Data Created Daily, retrieved from online http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/ on March 27, 2017