FORENSIC INVESTIGATION ON HADOOP BIG DATA PLATFORM

MYAT NANDAR OO

UNIVERSITY OF COMPUTER STUDIES, YANGON

JANUARY, 2019

Forensic Investigation on Hadoop Big Data Platform

Myat Nandar Oo

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

28.1.2019 

…..…… ……..……

Date Myat Nandar Oo

ACKNOWLEDGEMENTS

First of all, I would like to thank His Excellency, the Minister, the Ministry of Education for full facilities support during the Ph.D Course at the University of Computer Studies, Yangon.

I would like to express very special thanks to Dr. Mie Mie Thet Thwin, the Rector, the University of Computer Studies, Yangon, for allowing me to develop this thesis and giving me general guidance during period of my study.

I would like to express my deepest gratitude to my supervisor, Dr. Thandar Thein, Rector (acting) of the University of Computer Studies, Maubin, for providing me with an excellent atmosphere in doing research. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. Her patience and support helped me overcome many crisis situations within research and finish this dissertation. I appreciate her endless patience, positive outlook, ability to provide advice. She raises me up more than I can be.

I would also like to extend my special appreciation and thanks to Dr. Khine Moe Nwe, Professor, the University of Computer Studies, Yangon, and Dean of the Ph.D Courses for the useful comments, advice and insight which are invaluable to me.

I would like to express my respectful gratitude to Daw Aye Aye Khine, Associate Professor, Head of English Department, the University of Computer Studies, Yangon, her valuable supports from the language point of view and pointed out the correct usage in my dissertation.

I also would like to thank a lot all my teachers for mentoring, encouraging, and recommending the thesis.

In addition, I would like to thank the board of examiners for making precious comments and detailed corrections to my thesis and those who are pressing power to improve the end result.

And also, sincere thanks to all my friends for their motivating encouragement, for the stimulating discussions about research and for our fun time together. Most of the appreciations show to my friend Wint Nyein Chan for the great deal of her time.

Last but not least, I am grateful to my parents, who specifically offered strong moral and physical support, care and kindness, throughout my whole life as well as

during my Ph.D. study. Without their full support, my dissertation would not have been possible. I am very much indebted to my family for always believing in me, for their **endless love and support**. They are always supporting and encouraging me during the years of my Ph.D study.

ABSTRACT

In the era of Big Data, Hadoop Big Data Platform has been embraced by both individuals and organizations as it can offer cost-effective, large capacity storage and multi-functional services on a wide range of devices. It is fast raising popularity to access Hadoop services via client devices. The widespread usage of Hadoop Big Data Platform could create the environment that is potentially conducive to malicious activities and illegal operations. Thus, the forensic investigation on Hadoop Big Data Patform becomes the emerging field for the digital forensic community. There is also a need for a digital forensic framework relating to the forensic analysis of Hadoop Platform to guide the forensic works on Hadoop Big Data Platform to discover the potential evidences in order to identify the usages.

Hadoop produces a large amount of backlog per operation, which has led to cumulative backlogs of evidence awaiting analysis. The following major forensic challenges are arising in Hadoop Big Data Platform environment because of: complex infrastructure, the large amount of Hadoop backlog and lack of location knowledge about digital evidences. Without knowing where the evidential data may reside, it can impede an investigation.

This research proposed a forensic investigation framework to guide the forensic works on Hadoop Big Data Platform. Moreover, as the proactive research before conducting the forensics, it discovers residual artifacts (potential evidences) from Server and attached client devices of popular Hadoop Big Data Platforms: Ambari Hortonworks Data Platform (Ambari HDP), Non-Ambari Hortonworks Data Platform (Non-Ambari HDP), Cloudera Distribution of Hadoop (CDH) and MapR Hadoop Platform (MapR).

The experiments are conducted in relation to the use of popular Hadoop Big Data Platforms by accessing with the client devices of different Operating Systems (OS). The residual artifacts are also extracted from the attached client devices of differnet OS. The underlying OS of attached client devices are: Windows PC and Android Smart Phone.

It was decided to examine a user accessing Hadoop Platforms, and also to examine any differences when using different browsers: Internet Explorer, Mozilla Firefox, Google Chrome, and Android Browsers. The file operations are tested with

the different client devices for each browser to identify the different circumstance of usage.

A variety of circumstances were examined, including the different types of operation to access, upload and download data in the Hadoop. By determining the residual artifacts on server and client components, this research contributes to a better understanding of the types of artifacts that are likely to remain. The extracted artifacts can assist the forensic examiners for future forensic investigation on Hadoop Big Data Platform.

The popular crime scenarioes which are extended the Forensic Copra's crime cases and CYFOR cases are examined under the guide of proposed forensic investigation framework for Hadoop Big Data Platform.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Big Data is a large vein of technology in the 21st century. LinkedIn, eBay and Google are the first organizations which build upon the Big Data to support services. As a new entity of information technologies, Big Data can bring cost reductions actors in major improvements of computing time required to perform work or to offer new products and services. Big Data is progressively applied in areas such as health, education and finance, including the ability to examine for the hidden value of great information. For the above reason, Big Data presents the challenges relating to the security issues when illegal usages are occurred.

Many new technologies are designed for storing and managing Big Data, so called Big Data platform. Apache Hadoop is open source Big Data platform. It is designed to operate a large datasets with scalable, error-free and flexible methods. Hadoop is intended to extend the format of one device to a group of thousands of servers for storing and manipulating of Big Data. By applying the Hadoop implementation, the organizations can save IT resources for handling their organizational data. Thus, the Hadoop market is increasing sharply. However, this benefit comes with some inherent disadvantages of security concerns. It may be the target of cyber-criminal to implement it in illegal ways. Therefore, Big Data forensic is the emerging field for forensic community.

1.1 Big Data

In the era of Internet, there is no doubt that there has been an explosion of available data. For instance, in the last two years it is estimated that 2.5 quintillion bytes of data is generated par day, or 12 TB of Tweets are created every day [86]. The generated data volume is linked to several causes such faster and better network communications, cheaper storage, modernization of industries such health care industry and the usage of new devices able to generate data such mobile phones, tablets, GPS enabled devices and so on. Data statistics are often largely caused by data collection based on many mobile devices, mobile devices, camels, equipment, telephone information, electronic telecommunications and others.

The term "Big Data" has recently been used in terms of the growing volume of information that societies store, manage and analyze, because of many and most sources of information that are being used [66]. Big Data habitually includes data records that exceed the ability of using the traditional storage. The size of data is the continuous use of focus, starting with a few nodes in many databases. In many business scenarios, the data is bigger or faster than ever. Big Data technology has the potential to help the companies to consolidate their recommendations, make quick and simple solutions. Big Data is also a great factor in the context of managing both structured and unstructured data. Big Data technologies look for a form of storing, retrieving and processing this massive amount of data. Organizations have realized that there is a hidden treasure in these data sets. Identification of business trends, index searching, creating new products, better understanding of user relations or machine-learning are some of the uses for Big Data technologies. The format of Big Data varies greatly. Emails or social media posts are just two examples of unstructured data whereas word documents, spreadsheets and relational data bases are examples of semi-structured and structured data. Data is different information that is usually formatted in a specific way. There are three different formats of data. They are:

- **Structured data**: is the data which exists in a static field in a record that contains information included in spreadsheets and relational databases.
- **Semi-structured data**: refers to the information that is not resided in a relational database on the other hand there have some organizational features which can lead the easier analysis than structured data.
- **Unstructured data**: is the information that resides in the non-relational database with no pre-defined data model. It often includes text and multimedia content including e-mail, audio file, etc.

Basically, there are three classifications of Big Data sources. They are as follows:

- **Social Networking** (Information from People): This information is a record of human experiences recorded previously in books and art, and then photographs, audio and video.
- **Traditional business systems** (Information related to the process): storing data and help in providing business intelligence which is interesting business

activities such as customer registration, product production, etc. For example medical records, commercial transactions, bank records/ inventory, e-commerce and credit cards.

- **Internet of Things** (data generated by machines): derived from embedded and ubiquitous devices for measuring and controlling the situations in the physical world. For example, data from sensors, web logs, etc.

Big Data is considered one of the greatest technologies for the digital revolution of the past few centuries, and this will increase in the future. Wikibon that is an organization of experts and professionals on technology community states that Worldwide Big Data market revenues is $42 billion in 2018 [114]. Statista forecasts that the market of Big Data will be valued $55 billion in 2019 and are projected to increase to $79.5 billion in 2024, [14] as shown in Figure 1.1.

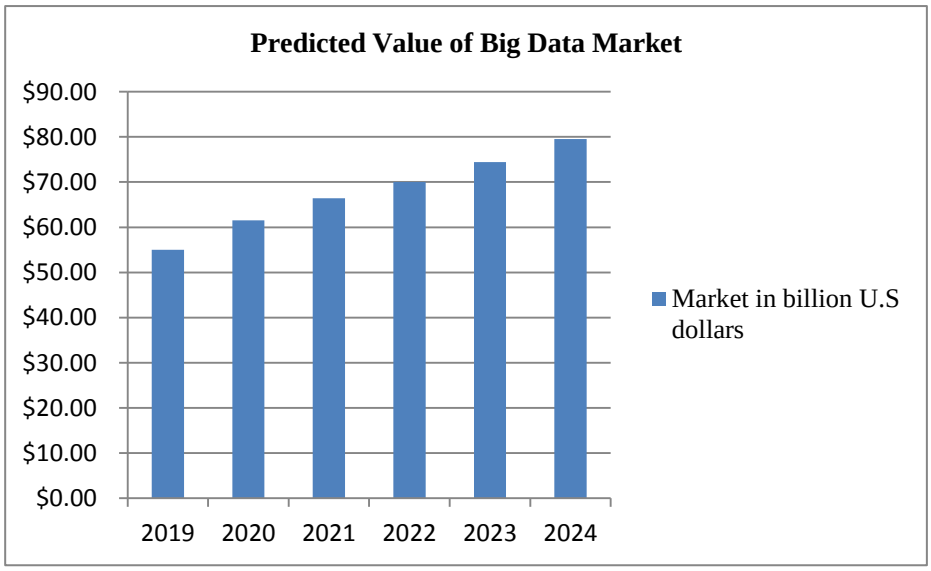

Figure 1.1 Predicted Value of Big Data Market (2019 to 2024) [14]

In order to achieve large benefit from Big Data, it is required to consider processing power, and the raw storage along with the strong analytics abilities and services. The business organizations rely on Big Data Platforms for managing Big Data to provide efficient, easy to use and consistent storage solutions by sharing multiple files with establishing a hierarchical and unified assessment of these files.

The various characteristics of Big Data: volume, velocity, variety, and complexity propose various challenges. The deviations in the data storage amount in

the diverse sectors, the generated data types are audiovisual, pictures, acoustic, or text/numeric form. Figure 1.2 shows Big Data architecture.

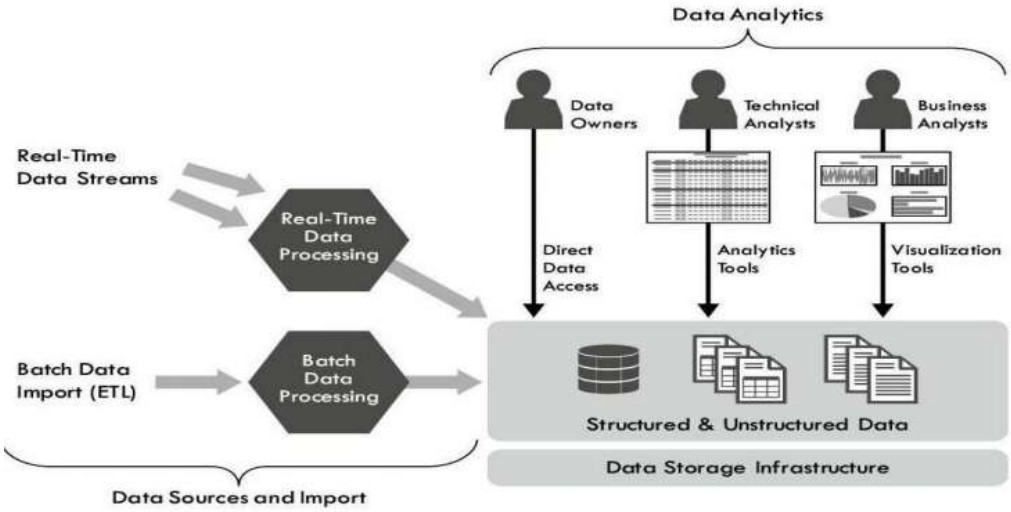

Figure 1.2 Big Data Architecture [66]

Big Data is progressively applied in areas such as health, education and finance, including the ability to examine for the hidden value of great information. With respect to the high usage of Big Data, it presents the challenges relating to the security issues when illegal usages are occurred.

1.1.1 Big Data Platform

Big Data platform is a way of technology solution that comprise with the storage, services, intelligence, applicable software packages and utilities. The custom development, querying and integration with other system can also be supported. The Big Data platform has the advantage in reducing the complexity of multiple vendors because of a cohesive solution.

Big Data platform is the technological contribution that combines the features and applications of some effective Big Data utilities as a single solution. Regardless of the types of data are incorporated, Big Data platforms are designated to collect, store, and analyze various forms of data in a single solution.

The demands for the storage of a massive amount of data make it a necessity for advanced computing infrastructure; a need to have a solution which is designed to scale out on multiple servers. This feature highlights the finest open source file systems designed to cope with the demands imposed by Big Data. The file systems:

Hadoop, Google File System, Gluster, etc. perform as the Big Data platform [113]. Hadoop is the most popular among the several technologies that have been developed to manage Big Data.

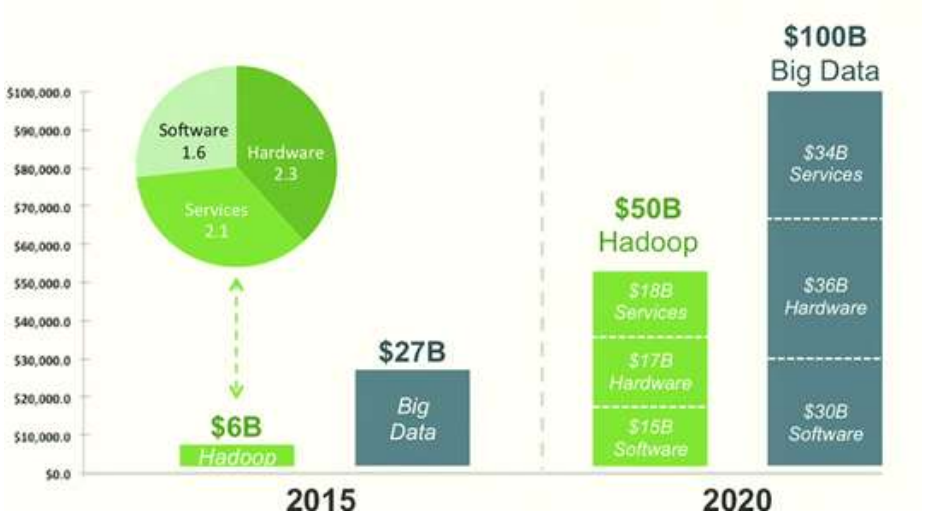

Figure 1.3 Market of Hadoop Comparing With the Whole Big Data Market [115]

Hadoop market in the field of Big Data is increasing dramatically as shown in Figure 1.3. Allied Research Organization [115] explored that the previous Hadoop market in 2015 valued only $6B while the whole Big Data market was $27B. The Global Hadoop Market reached $8.74 billion by 2016. And then it predicts that the market of Hadoop can reach $50.2 billion by 2020 while the whole Big Data market will be valued of $100B. In the end of 2017, the usage of Hadoop software (estimated by Wikibon) will be improved to $677 millions [114].

1.1.2 Hadoop Big Data Platform

The Hadoop Big Data Platform is open source, and license distribution, demonstrates important promises in dealing with necessities of the Big Data characteristics. The vision is to have the Hadoop Big Data Platform with the Map Reduce paradigm and other facility software packages to run on the HDFS cluster [42]. The Hadoop as a Big Data platform identifies important issues and ensures flexibility, scalability and capability as desired in Big Data Platform [56].

(i) Hardware Layer

(ii) DFS Layer

(iii) Resource Management Layer

(iv) Distributed Processing Layer

(v) Component Layer

(vi) Application layer

(vii) Common Service Layer

Hadoop is the JAVA open source library for distributed computing. This platform was designed to store, access and process large datasets in a scalable, fault-tolerant and flexible way. The purpose of using Hadoop is for scaling up from one machine to clusters formed by many servers which computation is done locally. The Hadoop software library includes: HDFS layer that provides redundant storage and manage data, and MapReduce which is a computer framework for the parallel processing of large data sets.

- Hadoop Distributed File System (HDFS) is the distributed storage layer and the HDFS cluster principally comprised two parts;
    - Namenode that manages the file system metadata and
    - Datanodes that store the actual data.
- MapReduce: is a programming layer created by Google. This layer applied the divide-and-conquer method for itemization of complex data into small units, and then process them in parallel.

In Hadoop version 2, a new layer for managing resource and job scheduling is used; YARN. Here is the explanation of two main components of Hadoop.

The Hadoop version 0.1.0 was published in April, 2006 and continues to increase its versions [44]. Up till now, latest released Apache Hadoop 2.7 was available in June, 2016 [3]. Hadoop is speedily mutable and new software packages are being added to Hadoop. Recently, parts of the inventive Hadoop Apache project have turned to build software, such as Avro, HBase, Pig, HCatalog, Hive, Flume, Oozie, Sqoop, and Zookeeper.

Nowadays, Hadoop Big Data Platform is increasingly adopted as the single solution of distribution. A number of companies became bundle Hadoop and related technologies into their own Hadoop distributions. The three prominent Hadoop distribution companies are MapR, Cloudera, and Hortonworks [3]. The following Table 1.1 shows the Similarities and Differences among MapR, Cloudera, and Hortonworks. Because of the wide usage of Hadoop Big Data Platform, it can be identified as a challenge to digital forensic researchers.

Table 1.1 Pros and Corns of Cloudera, Hortonworks and MapR [3]

Hadoop Distribution	Advantages	Disadvantages
Cloudera Distribution for Hadoop (CDH)	CDH has a user friendly interface with many features and useful tools like Cloudera Impala	CDH is comparatively slower than MapR Hadoop Distribution
MapR Hadoop Distribution	It is one of the fastest hadoop distributions with multi node direct access.	MapR does not have a good interface console as Cloudera
Hortonworks Data Platform (HDP)	It is the only Hadoop Distribution that supports Windows platform.	The Ambari Management interface on HDP is just a basic one and does not have many rich features.

1.1.3 Big Data Security

One of the most significant challenges of Big Data is the Security feature in Big Data that is very sensitive and, technical for legal implications. When applying to enable global access of the private databases, the personal information has been interrupted the security by combining the external large data sets. It is the cooperative term for all the procedures and tools used to guard both the data and analytics processes from malicious activities. Much like other forms of cyber-security, the big data variant is concerned with attacks that originate either from the online or offline spheres.

1.2 Digital Forensics

Forensic science is the application of science to criminal and civil laws, mainly on the criminal side during criminal investigation, as governed by the legal standards of admissible evidence and criminal procedure [33]. It means applying scientific methods and processes to solving crimes. Forensics science is a technical field. As such, much of the process requires a deep technical understanding and the use of technical tools and techniques. Depending on the nature of an investigation, forensics may also involve legal considerations, such as spoliation and how to present evidence in court [59].

Digital forensics is a reactive discipline which best practices and guidelines must be accepted after a period of experimentation where different approaches are tested. Digital forensics is the process of examining electronic evidence for legal purposes, such as criminal investigations.

Digital Forensics has grown from a relatively obscure tradecraft to an important part of many investigations [44]. Digital Forensics implements the scientifically derived and proven methods to preserve, validate, identify, analyze, interpret, document and present the digital evidences. Digital evidences are derived from digital sources to facilitate or further the reconstruction of events that are found to be criminal [21].

Nowadays, wide natured digital crimes are emerging and impacting and hence digital forensics is being acknowledged as a crucial branch. Today much of the latest contributions of technology becomes inappropriate in traditional digital forensics [44]. Digital Forensics is facing crisis due the following factors.

- growing size of storage devices: difficuilites for imaging the subjected devices,
- increasing pervasiveness of embedded flash storage and the propagation of hardware interfaces: storage devices can no longer be readily removed or imaged.
- explosing of operating systems and file formats: dramatically increasing the requirements of tool development
- using remote processing and storage: frequently data or code cannot even be found.

These problems are most obvious to forensics investigators faced with advanced computing platforms.

1.2.1 Branches of Digital Forensics

The various digital forensics branches along with the working are of the forensics. The concepts contained in the realm of digital forensics are [58]:

(i) **Media and File System Forensics**: It is the process of recovering the information from the storage media, where digital evidence can be found, as several types of file systems the medium can have, such as NTFS.

(ii) **Operating System Forensics**: OS forensics deals with OS of the computer or mobile device such as Windows, Android, and Linux.

(iii) **Network Forensics**: Network forensics is that the investigators conduct the forensic analysis by monitoring the network traffic to perceive intrusion, to collect the evidences.

(iv) **Mobile Device Forensics**: A mobile device has various locations of data storage for extracting the evidences including volatile or non-volatile memories, memory card and so on. The search and seizure rules must be followed for mobile device forensics.

(v) **Virtual System Forensics**: In this type of forensics, the inspectors could first generate an image of the host machine and then export files associated with a virtual machine.

(vi) **Application Forensics**: Application forensics contains the examination of software which are installed upon the hardware.

(vii) **Software Forensics:** Software forensics implicates the identification, discrimination, and characterization. It can assist in finding the culprit.

(viii) **Web and Email Forensics**: Browser history, cookies, registry entries on the client side, and log files on the server side can be a great source of digital evidence. The role of email forensics is to identify the scammer behind the crime. Email investigations rely heavily on email message files, email headers, and email server log files.

(ix) **Database Forensics**: It encompasses the investigation of databases and metadata for extracting the digital evidence.

(x) **Malware forensics:** It is the analysis of malicious code. The professional criminals use malware to steal confidential information from the computer.

(xi) **Hybrid and Emerging Technologies Forensics**: It contains the following branches of forensics:

- Cloud Forensics
- Social Networks Forensics
- Big Data Paradigm Forensics
- Control Systems Forensics

1.2.2 Forensic Process Model

It is claimed that digital forensics is a process that can be modeled with some reasonably established phases [75]. Most of the forensic process models have focused on “the investigative process and the different phases, they addressed the complexity

of an investigation and the features and functionality of devices, and the concrete principles of an investigation" [100].

Kruse and Heiser Model for computer forensics [63] is the earliest systematic computer forensics methodology. It was based on three fundamental phases of (i) acquiring the data evidence, (ii) checking the validity of the collected data and (iii) the analysis. It is recommended that the data integrity should be ensured by authentication process.

The National Institute of Standards and Technology (NIST) [72] defined the basic forensic process as shown in Figure 1.4. The step by step processes are:

- **Collection:** This process is relating to the identification, labeling, recording, and acquiring evidences from the potential sources.
- **Examination:** It includes forensically processing of the collected data by using either automated or manual approaches to evaluate and extract data of particular interest.
- **Analysis:** Analysis is to investigate the results of the examination, using legitimately reasonable methodologies, to develop useful information to solve the inquiries.
- **Reporting:** The final phase of reporting the results of the analysis provides the descriptions of the applied methods, explaining how tools and procedures were selected.

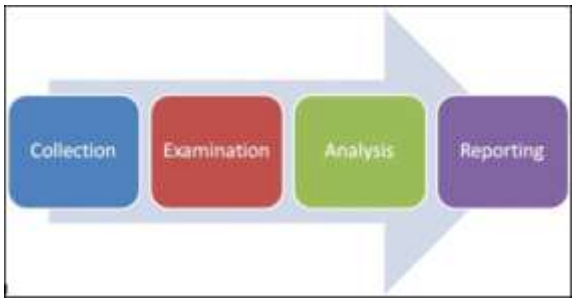

Figure 1.4 NIST Process Model for Digital Forensics [72]

1.3 Criminal Interests in Big Data

Many organizations are doing business in the Big Data Platform and the criminals also find the ways to utilize it illegally. Criminals may be able to takeover resources for illegal purposes [1]. The utilization of massive sets of data is to figure

out where criminals will strike and how they will act has given a major boost to law enforcement officials. For all the advantages Big Data has given to organizations, one that has proven especially beneficial is its use in tracking down and capturing criminals.

Many businesses have embraced this technology has well as they seek to protect their most valuable data. However, for all the advantages it provides, Big Data can also be used by the enemy. Security breaches are on the rise, and massive databases have been compromised. Considering the amount of information at risk in the internet economy, organizations have reason to be concerned. It is clear that cyber criminals view Big Data as their next effective tool, and security experts are warning that the worst may be yet to come.

Perhaps it should come as no surprise that criminals have turned to Big Data as another weapon within their arsenal [8]. Black guys gravitate towards technologies that make their goals easier to achieve, and considering that many Big Data tools are based on open source software, gaining access to them is relatively easy. Big Data is simply another part of criminals' strategy, allowing them to act with more agility and sophistication, while also executing their illegal usages more quickly.

Businesses may worry that Big Data has become the problem when it comes to cyber criminals, but Big Data may also be the solution. Using it for protection, however, requires companies and organizations to know more about it and how to adopt its solutions effectively. That includes an emphasis on protecting data through encryption and stronger access controls. Behavioral analytics can also pinpoint when a crime is likely to take place, giving ample warning before something damaging occurs. The main takeaway from this growing trend is that Big Data is merely a tool. It can be used by criminals for nefarious means, and it can be used by businesses to protect them. If the right preparation happens now, the challenge posed by cyber criminals can be mitigated.

In some cases, the use of Big Data by criminals is a necessary evolution for them. Many cyber criminals now traffic in massive information, selling valuable data to the highest bidder in the shadowy corners of the internet [8]. A simple data breach at a major organization like Target can result in millions, even tens of millions of records containing multiple data sets falling into their hands. Sorting through that

information manually is a nearly impossible task for even a small group of people, let alone one person. But with Big Data analytics, criminals can mine that data for the most useful information in a fraction of the time, effectively monetizing it for their own purposes.

Big Data environment will make it more challenging for the forensic examiners to extract the evidence than the traditional computer forensics, because of difficulty in what data is the forensically important among the large amount of data [1]. The traditional imaging approaches to digital forensics are irrelevant because full digital images of hard drives should not be carried out in Big Data environment. Big Data platform usually contains thousands of commodities machines for storing petabytes of data, therefore doing a full imaging for forensic acquisition would be extremely resource consuming [25].

When the forensic investigation of Big Data platform is conducted without knowing where residual artifacts may reside, it can impede the investigation. Therefore the forensic research on Big Data Platform for discovering the residual artifacts is necessary. There is a prerequisite for forensic capabilities which support the investigations of illegal usage in this environment. Thus, forensic investigation framework is also needed to guide the forensic work on Big Data platform.

1.4 Motivation of the Research

The motivation for conducting research into forensic investigation on Hadoop Big Data Platform is discussed with the following three points:

- Hadoop Big Data Platform is progressively used for both personals and organizations to process the large amounts of data, which can be accessed with computers or mobile electronic devices.
- Criminals are taking on the opportunity to store illicit data on Hadoop platform, which contributes to difficulties of tracing the criminal activities due to its complex platform.
- If identification of data sources and potential evidence is not able to be undertaken, the Investigations can take a considerable amount of time.

1.5 Objectives of the Research

The aim of this research is to propose forensic methodology for Hadoop Big Data Platforms and extract the artifacts for tracking the illegal usages in Hadoop implementations running on various Operating System and Amazon Web Services. This thesis is intended to help the forensic examiners for conducting the forensic works relating Hadoop Platforms. The methodology and artifacts resulting from this thesis aim to assist the generating of effective evidences in future forensic work in Hadoop Big Data Environment. Current research work in Hadoop forensic will be used as starting point to design a viable forensic methodology for Hadoop Platform implemented in Big Data Environment.

This research is intended to determine whether there are any residual artifacts that are remained on popular Hadoop Big Data Platform from both server and client portions. After studying the literature, it can be defined that there was a need to have a methodology to guide the forensic investigation on Hadoop Big Data Platform, and hence a framework is proposed and applied to conduct the forensic investigation on Hadoop. It depends on the basic procedures of digital forensic analysis for providing the framework which follows a standard digital forensics, and also supports to commence the forensic research for providing the residual artifacts of Hadoop Big Data Platform. In order to attain the aim of the research, the following objectives have to be met:

- to conduct research on popular Hadoop Platforms
- to propose a forensic investigation framework that can help the forensic practitioners, and researchers for conducting the forensic investigation on Hadoop Platform.
- to address the volume challenge in generating evidences for investigation on Hadoop Platform by locating and discovering the residual artifacts that remain on the Hadoop Platform (Storage Server and attached client machines)
- to provide the potential evidence to assist forensic practitioners in generating evidences.

The focus of this research is to provide the better understanding of the type of residual artifacts which are potentially to remain. It can be used as the starting point to design a viable forensic methodology for Hadoop Platform implemented in Big Data Environment.

1.6 Research Questions and Research Methodology

This section describes the research question and methodology. The objective of this work is to propose a forensic investigation framework for Hadoop Big Data Platform and discover the residual artifacts for helping the forensics examiners. To do research for forensic investigating, the research questions and research methodology are raised as the following sub sections.

1.6.1 Research Questions

As outlined in the previous sections, Hadoop Big Data Platform stores and manages the Big Data containing the sensitive and valuable data. There is also a lack of knowledge for extracting the evidences, when conducting the forensics investigation on that platform. The primary research questions are thus defined as:

Q1: What is the forensic investigation framework to guide the forensic investigation on Hadoop Big Data Platform?

Q2: What residual artifacts result from the use of the Hadoop Big Data Platform to identify its use?

H1 - There are no residual artifacts.

H2 - There are Artifacts.

H2.a: The residual artifacts can be extracted from Server.

H2.b: The residual artifacts can be extracted from client devices.

Q3: While investigating the popular Hadoop Big Data Platforms, are the residual artifacts discovered from same locations?

Q4: Among the large amount of backlogs, which residual artifacts are useful for forensics?

Q5: How to apply forensic investigation framework in crime scenario?

1.6.2 Research Methodology

This paper proposes an investigation framework for forensic analysis of Hadoop Big Data Platform. The investigation scope contains discovering residual artifacts on popular Hadoop Platforms; Hortonworks Data Platform, Cloudera Hadoop

Server portions the attached client machines. The research questions which are stated in the above section are able to answer as the following.

M1: Forensic investigation framework for Hadoop Big Data Platform is proposed.

M2: This research discovers residual artifacts from Server and attached client devices of popular Hadoop Big Data Platforms;

- Hadoop 2.7.1
- Ambari Hortonworks Data Platform (Ambari HDP)
- Non-Ambari Hortonworks Data Platform (Non-Ambari HDP)
- Cloudera Distribution of Hadoop (CDH)
- MapR Hadoop Platform (MapR)

M3: This research presents the discovered residual artifacts resulted from the forensic investigation of popular Hadoop Big Data Platforms by accessing via the client devices of different OS.

M4: The forensically important files are extracted among the large amount of backlogs.

M5: The crime scenario of Hadoop Big Data Platforms is investigated as a case study by applying the proposed forensic investigation framework.

To answer the research questions, the experiments are conducted in relation to the use of Hadoop Big Data Platforms: Hortonworks Data Platform, Cloudera Distribution of Hadoop, and MapR Hadoop Platform. A variety of virtual machines were created to investigate a range of situations for a user accessing Hadoop Platforms, and also to examine any differences when using different browsers. Multiple access methods were explored; each made use of Hadoop Platforms with a different browser; Internet Explorer, Mozilla Firefox, Google Chrome, and Android Browsers. The file operations are tested with the different OS for each browser to identify the different circumstance of usage, as outlined in Figure 1.5.

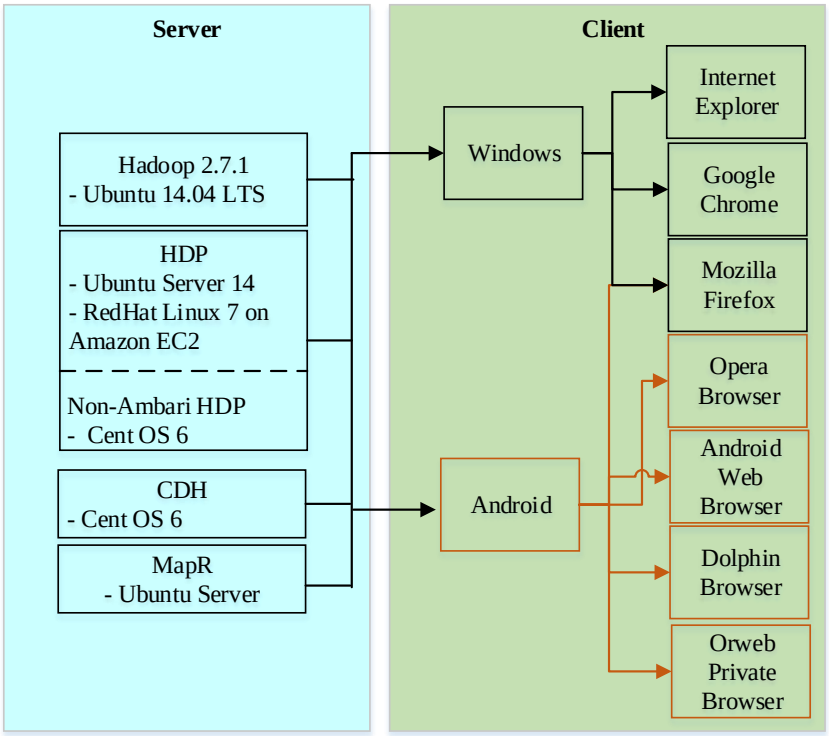

Figure 1.5 Block Diagram of Investigation Scope

1.7 Organization of the Research

This research is organized with seven chapters, including background information on Big Data, forensic investigation process model, Hadoop and then, motivations, contributions, system overview. It also provides the objective of the thesis in this chapter.

The next chapter, Chapter 2 discusses the literature reviews and related work with some existing methods are also surveyed on which the prior studies that are dealing with the dissertation.

Chapter 3 provides the proposed forensic investigation framework for Hadoop Big Data Platform. It outlines how this can be applied to forensic analysis of Hadoop Big Data Platform. Each step of the framework is explained; Scope and Identification, Prepare and Collection, Analysis, Reporting and Closing.

Chapter 4 describes two types of Hadoop Hortonworks Data Platform (HDP); Ambari HDP and Non-Ambari HDP. The structure of the HDP is discussed and the forensic investigation on HDP is also conducted. In this chapter, artifacts are extracted from both of the server and client portions of HDP. Materials presented in Chapter 3 and 4 were published in International Journal of Computer Systems Science and

Engineering (CSSE) and International Journal of Computer Science and Information Security (IJCSIS).

M.N.Oo and T.Thein, "Forensic Readiness on Hadoop Platform: Non-Ambari HDP as a Case Study", International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 9. Sept 2017.

M.N.Oo and T.Thein, "Forensic Investigation through Data Remnants on Hadoop Big Data Storage System", International Journal of Computer Systems Science and Engineering (IJCSSE) Vol. 33, Jan, 2018.

Chapter 5 expresses the infrastructure and configuration of the Cloudera Distribution of Hadoop (CDH). The investigation on CDH is also demonstrated by applying the proposed forensic investigation framework. A crime case scenario is stated to discover the artifacts on CDH Sever to with the applicant of proposed forensic investigation framework. Material presented in Chapter 5 was published in In Proceeding of the 1st International Conference on Advance Information Technology (ICAIT).

M.N.Oo and T.Thein, "Forensic Analysis of Residual Artifacts on CDH Storage", In Proceeding of the 1st International Conference on Advance Information Technology (ICAIT) Nov 1-2, 2018.

Chapter 6 expresses the infrastructure and configuration of the MapR Hadoop Platform. The investigation on MapR is also undertaken by applying the proposed forensic investigation framework. Material presented in Chapter 6 was published in In Proceeding of the IEEE International Conference on Knowledge Innovation and Invention 2018 (ICKII 2018).

M.N.Oo and T.Thein, "Forensic Investigation on MapR Hadoop Platform", IEEE International Conference on Knowledge Innovation and Invention 2018 (ICKII 2018) July 23-27 2018.

Finally, Chapter 7 presents the conclusion extracted from this research and depicts the future research lines to continue it.

CHAPTER 2

LITERATURE REVIEW

The chapter 2 states the shortcomings in the development of formal digital forensic methodology to accommodate new technologies such as cloud computing, Big Data and Hadoop. A thorough study of digital forensic investigation for the recent technologies; cloud, Big Data and Hadoop is discussed. Then, challenges posed by Hadoop and Big Data to Digital Forensics identified in recent papers are discussed. The literature review is completed by analyzing Hadoop investigation methods and current research on forensic methodologies. There are not many academic papers related to forensic investigation on Big Data platforms; in most cases they are focused on some particular technology or solution that reflect only a small part of the whole problem area. The works and efforts of previous researchers are reviewed in four portions:

(i) Cloud computing and digital forensics
(ii) Big Data and digital forensics
(iii) Hadoop forensics
(iv) Forensic investigation frameworks

2.1 Cloud Computing and Digital Forensics

Cloud computing is one of the newest innovation in computing. Cloud computing is a technology of easy to access, on-demand network access to a shared pool of computing resources, which can be scale up, and rapidly maintained with management or service delivery [15]. There are two main deployment models relating to cloud hosting: public cloud and private cloud as shown in Figure 2.1.

- Public Cloud: is the standard cloud computing model, wherein the service provider creates resources, applications or storage, available to the public through the internet.
- Private Cloud: brings similar advantages to public cloud, including scalability and self-service, but through a proprietary architecture. Unlike public clouds, which deliver services to multiple organizations, a private cloud is dedicated to the needs and goals of a single organization.

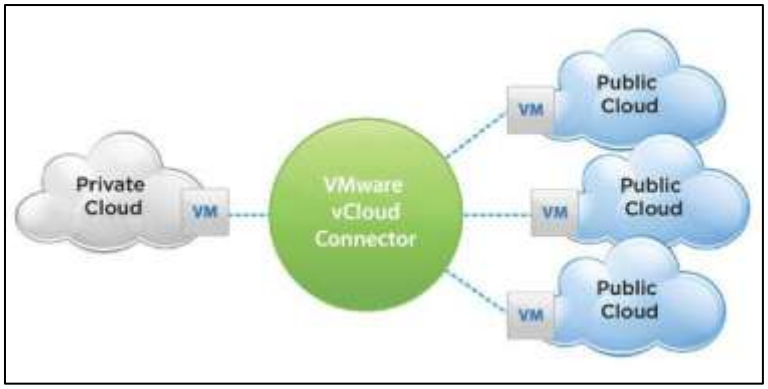

Figure 2.1 Main Deployment Models of Cloud Computing [100]

Though the private cloud can even reduce the occurrence of intellectual property theft, there are some crimes of illegal usage of sensitive data and proprietary information that the company stores on its own server. The widespread adoption of cloud applications, coupled with risky user behavior that corporations may not even be aware of, is further widening the scope for cloud-based crimes. Some of this interest is for commercial reasons, and some are fuelled by criminal intent [12].

The low matureness of cloud could face the many types of challenges in cloud. The security challenge is in the topmost rank [27]. Further, the cloud servers are moved outside the traditional security perimeter making it easy for the reach of cyber criminals. This is a growing concern particularly when cloud computing stores sensitive data about customers [4].

While mitigating cloud crime, investigators face several challenges and issues dealing with cloud forensics. In the article [12], two critical issues were discussed in the age of the cloud: regulatory challenges and privacy challenges. The focus was mainly on the trouble of defensive the confidentiality, availability, and integrity of subjected information. It provided and supported an analysis of the technological complexities of cloud computing and associated services, and found continuing resolutions to the inherent privacy challenges.

In order to recover the cloud criminal activities and maintain the security and integrity of the information stored in the cloud, cloud forensics has been introduced to help forensic investigators find potential evidence [24]. Cloud Forensics is cross-discipline between Cloud Computing and Digital Forensics. It is actually an application within Digital Forensics that supervises the crime committed over the cloud and investigates on it.

National Institute of Standards and Technology (NIST) [8] defined cloud computing as "cloud computing forensic science" is the application of scientific principles, technological practices and derived and proven methods to reconstruct past cloud computing events. This is done through identification, collection, preservation, examination, and interpretation and reporting of digital evidence.

Sophisticated interactions between Cloud Service Provider (CSP) and customers, resource sharing by multiple tenants and collaboration between international law enforcement agencies are required in most cloud forensic investigations. In order to analyze the domain of cloud forensics more comprehensively, and to emphasize the fact that cloud forensics is a multi-dimensional issue instead of merely a technical issue, the paper [91] discussed the technical, organizational and legal dimensions of cloud forensics as shown in Figure 2.2.

The technical dimension is dealings with the methodologies that are needed to perform the forensic process. The organization dimension in a cloud computing comprises of two entities: the CSP and the cloud customer. In the legal dimension, the cloud forensics involves the guidelines and contracts to ensure that forensic activities do not breach laws and regulations where the data resides.

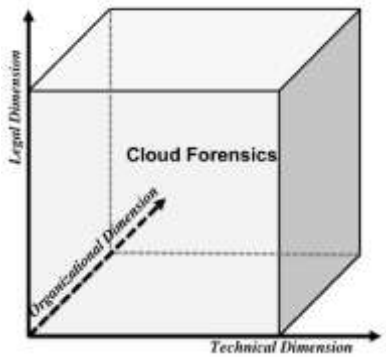

Figure 2.2 Dimensions of Cloud Forensics

The paper [60] defined cloud forensics as "the application of digital forensic science in cloud environments as a subset of network forensics". The authors highlighted the significance of cloud forensics in three different aspects, namely, technical, organizational and legal.

The paper [28] classified the literature into three dimensions: (1) survey-based, (2) technology-based and (3) forensics-procedural-based. It discussed widely

accepted standard bodies and their efforts to address the trend of cloud forensics and generate a mind map that help in identifying research gaps. Finally, it summarized existing digital forensics tools and the available simulation environments that can be used for evidence acquisition, examination and cloud forensics test purposes.

In this paper [96], the authors discussed the cloud forensic challenges that the investigators have no full control over the evidence (e.g., router logs, process logs, and hard disks). To recover this challenge, the paper presented the solution of a trust model which can add on the forensic facilities. That cloud enables to preserve the trustworthiness of evidence. The author stated that creating a secure model for cloud forensics is very important as it will lead to more trustworthy clouds, allowing their adoption in sensitive application domains such as defense, business, and healthcare.

2.2 Cloud Storage and Digital Forensics

The criminals can also misused cloud storage, and provide a distribution point for the illicit data. Cloud storage also obliges to offer a struggle in assigning ownership to find the association with illicit data [12]. Cloud storage is subject to attacks by cyber criminals, to steal the confidential information and distribute the forbidden data for criminal purposes.

As a number of researchers have pointed out, the use of cloud storage by criminals has complicated investigation and forensic examinations [47]. It is unlikely that conventional digital forensic techniques can be used to identify and seize evidential data from the cloud as data would probably be distributed worldwide and in different data centers.

As Quick and Cho [80] had posited, it is imperative that forensic examiners are knowledgeable about popular cloud storage services as well as the data artifacts and Artifacts that can be accessible from client devices in order to identify, preserve and analyses evidential data in cases involving the use of cloud storage.

The research [65] focused on discovering whether there are cloud storage data miscellanies on prevalent client devices. The proposed forensic framework was applied in analyzing widespread cloud storage services; Google Drive, and Microsoft SkyDrive to find the residual artifacts on client devices; Windows 7, and an Apple Iphone. The author pointed out that cloud storage username and password can be identified from the log file and browser information. The usages of anti-forensic

software did not eliminate the residual artifacts although full erase process can remove all data.

The forensic researchers discovered the artifacts on client devices to identify the usage of Google Drive [70], Skydrive [64] and Dropbox [80]. Martin and Choo [71] presented an integrated conceptual methodology of digital forensic framework for cloud computing that consists of (i) Evidence source identification and preservation, (ii) Collection, (iii) Examination and presentation, and (iv) Reporting and presentation phases. That paper discovered the residual artifacts on client devices to identify the usage of cloud storage by applying their proposed forensic framework.

Among cloud computing services, most consumers use cloud storage services that provide mass storage. This is because these services give them various additional functions as well as storage. It is easy to access cloud storage services using smartphones. With increasing utilization, it is possible for malicious users to abuse cloud storage services. Therefore, a study on digital forensic investigation of cloud storage services is necessary. This paper [26] proposed new procedure for investigating and analyzing the artifacts of all accessible devices, such as Windows system, Mac system, iPhone, and Android smartphone.

2.3 Cloud Services and Digital Forensics

Cloud services are massively used by both individuals and businesses as they offer cost-effective, large capacity storage and multi-functional services on a wide range of devices such as personal computers (PCs), Mac computers, and smart mobile devices (e.g. iPhones). Primarily, the type of cloud services has been divided into three major service categories as shown in Figure 2.3:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)

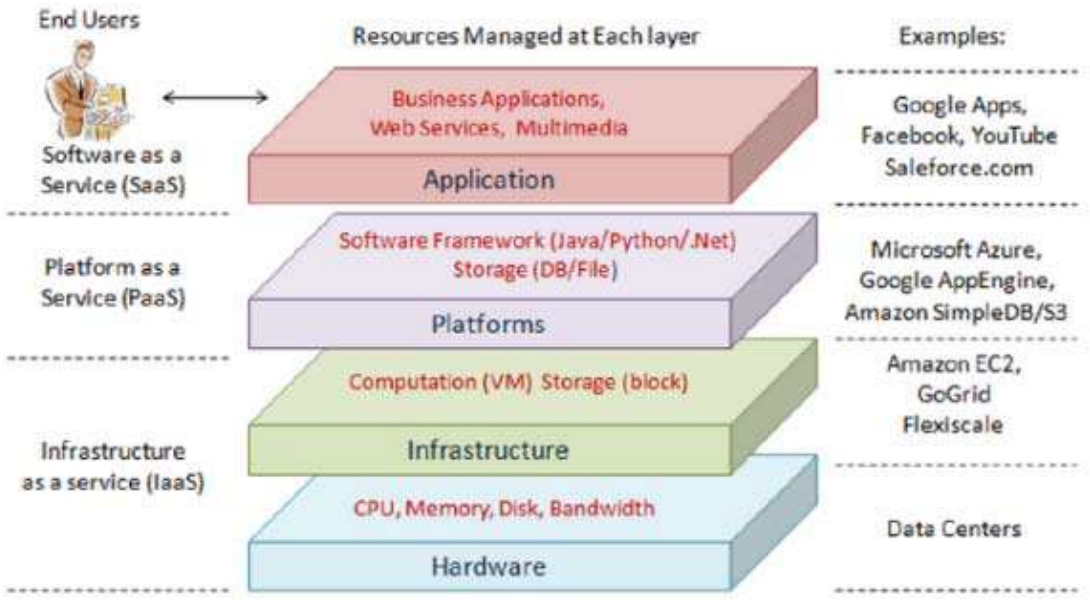

Figure 2.3 Cloud Computing Categories relating to Service Models [100]

When the Internet connected people all over the world, many did not know that they could be a victim of, what is known today as, cyber-crime. And cloud computing is one technology which cannot be utilized without using Internet. Cloud service is still evolving and it is to be exploited by criminals [6]. Cloud forensics helps in preventing and fighting the wrongful and illegal activities related to it. Digital forensics in the cloud remains a challenge, partly due to the diverse range of cloud services and devices that can be used to access such services.

This paper [100] presented a general view of cloud computing, which aims to highlight the security issues and vulnerabilities associated with cloud service models. This paper examined the three cloud service models and discussed the security challenges and issues involved with each service model along with potential solutions for each.

Due to the distributed and virtualized environment, cloud forensics is facing a great deal of challenges in comparison to traditional forensics. The paper [105] presented that the difficulty of access to evidence in logs is the most important challenge in cloud forensics. Some researchers have deal with it and have come up with solutions. Sang proposed a log-based model which can help to reduce the complexity of forensic for non-repudiation of behaviors on cloud.

He also proposed that log kept locally and synchronously, so it can be to check the activities on SaaS cloud without the CSP's interference. The local log module will use information such as unique id and timestamp on the log record locally. HASH code will be also used to detect modification on the log files. In PaaS, the CSPs should supply a log module on PaaS to the third-party in order to create a customized log module, for both of the consumer side and the cloud side.

There is a need for forensic capabilities which support investigations of crime in cyber cloud. It is needed a better secured model for cloud deployment and forensic investigation techniques to extract evidence from cloud based environments in case of any cyber-attack on cloud services.

The paper [10] discussed the comprehensive models that provides cyber Forensics capabilities on cloud computing. Based on the Virtual Machine Introspection, this paper proposed a cyber-forensic investigation system using Markov chain algorithm for Investigation. That model used the interaction of each node related with forensic actions on cloud environment and the derivation of data from the login details of the user, timestamps, event access, web page cache and logs. When user established the connection to the cloud server, then server allowed accessing the web page and requesting the web page from user. The cloud server allowed and responded to the authenticated users only. If the user authentication verified successfully then load web application to allow access web application and request wed application. This model can assume that whether the cloud consumer is the victim of the crime investigation.

2.3.1 SaaS and Digital Forensics

Software as a service (SaaS) is a software distribution model in which a third-party provider hosts applications and makes them available to customers over the Internet. With the rise of cloud computing services, the criminals also contribute to innovative behaviors of piloting cyber-crime, using SaaS as their new appliances. In the SaaS layer, web applications represent the dominant deployment model that enables the cloud users to access cloud services. Web applications represent 75% of the total reported vulnerabilities over the last three years [5]. Consequently, SaaS applications should be incessantly authenticated and scanned for vulnerabilities. Therefore, SaaS cloud forensic is in progress for rebooting the SaaS security.

In today's Internet-connected world, client devices are increasingly used to access cloud services, which allow users to access data anywhere, anytime. The client devices which use the SaaS services are targeted by cyber criminals to conduct malicious activities, such as data exfiltration, malware, identity theft, piracy, illegal trading, sexual harassment, cyber stalking and cyber terrorism. The paper [29] examined four popular cloud client apps, namely OneDrive, Box, GoogleDrive, and Dropbox, on both Android and iOS platforms (two of the most popular mobile operating systems). It identified artefacts of forensic interest, such as information generated during login, uploading, downloading, deletion, and the sharing of files.

Using a widely used open source cloud SaaS application, the paper [35] documented a series of digital forensic experiments with the aim of providing forensic researchers and practitioners with an in-depth understanding of the artefacts required to undertake ownCloud forensics. The experiments focused upon client and server artefacts, which are categories of potential evidential data specified before commencement of the experiments.

The article [100] explored the forensic work for Box; one of the popular SaaS applications. It presented that Box is much more forensically friendly cloud SaaS application. The grate logging and caching capabilities of Box help forensic practitioners a lot in their digital examinations.

The paper [102] stated the forensic analysis of Google Docs that recorded information relating to the use of the service such as IP address, number of logins, date and time access and storage usage. The investigators can get the information related faster without having to look at the information in the cloud one-by-one.

In the examination of cloud forensic especially in cloud based crime, [37] defined layers of trust based on the real situation where for example an evidence is brought to the court and the judge or jury have to decide whether they can trust the evidences presented to them in order to determine if the evidence is accurate or vice versa.

2.3.2 PaaS and Digital Forensics

PaaS is a complete and integrated Platform as a Service that allows business users and developers to cost-effectively build, deploy, and manage application workloads seamlessly on premises. Examples of PaaS systems are AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, OpenShift.

The PaaS layer provides a set of Application Programming Interfaces (APIs) that cloud users utilize to manage and interact with cloud services. The security and availability of general cloud services is dependent upon the security of these basic APIs. Appropriate security controls are mandatory requirements to enforce that only authorized cloud users can access service interfaces and make calls to authorized APIs with the right permissions.

There are two prominent types of PaaS:

- **Public PaaS** is delivered by a services provider for building applications. Examples include Salesforce Heroku, AWS Elastic Beanstalk, Microsoft Azure, and Engine Yard.
- **Enterprise PaaS** is delivered by central IT within an organization to developers and possibly partners and business customers. Enterprise PaaS sits on top of public IaaS, on-premise bare metal, and on-premise virtual machines. Some technology analysts make a distinction between the actual service that central IT is delivering (PaaS) and the software used to deliver that service.

  For example, Gartner uses the term "cloud-enabled application platform" or CEAP. Examples include Apprenda, VMware- and EMC-owned Pivotal, and Red Hat OpenShift.

Humans are a part of building and fielding the application, running on PaaS; as such, a fully secure environment can hardly be assured. The PaaS systems becoming more pervasive, an increasing number of assets, which are transmitted, manipulated, or stored digitally, are being compromised by cybercrimes. To identify and prosecute those responsible for such crimes, a digital forensic investigation aims to collect, analyse and present digital evidence necessary to demonstrate how a digital crime was committed, what harm was done, and who was responsible.

In this case, the cloud infrastructure hosts customer-developed applications and provides high-level services that simplify the development process. PaaS provides full control to customers of the application layer, including interaction of applications with dependencies (such as databases, storage, etc.), and allows customers to perform extensive logging for forensics and security purposes.

In the article [25], Salefore forensics is conducted emphasizing the login portion to identify suspicious login activity. It provides key user access data, including:

- The average number of logins to Saleforce
- Who logged in frequently
- Who logged in during non-business hours
- Who logged in using suspicious IP ranges.

The paper [39] solved the forensic challenges of accessing evidence in logs relating to the PaaS forensics. In PaaS, since the customers have full control on their application over a prepared API, system states and specific application logs can be extracted. The author proposed a logging mechanism which automatically sign and encrypt the log information before its transfer to a central logging server under the control of the customer. This mechanism can prevent potential eavesdroppers from being able to view and alter log data information on the way to the logging server.

2.3.3 IaaS and Digital Forensics

In addition to the concerns at each layer of the cloud service models described above, the architecture dependency between different layers yields some serious security concerns. The cloud computing model depends on a deep stack of inter-dependent layers where the functionality of a higher layer depends on the lower layers.

The IaaS model covers cloud physical infrastructure and virtualization. The platform interfaces and APIs in the PaaS layer depend on the virtualization of resources delivered by the IaaS.

Infrastructure as a Service (IaaS) offerings, AWS is the undisputed market leader with a 47.1% market share, followed by Microsoft Azure at 10.0% and Google Cloud Platform with 3.95% till the year of 2017 [29].

With demand for skilled security engineers at an all-time high, many organizations do not have the capability to do an adequate forensic analysis to determine the root cause of an intrusion or to identify indicators of compromise. To help organizations improve their incident response capability, the paper [35] presented the specific tactics for the forensic analysis of Amazon. Once the evidence is collected, the SIFT (the SANS Investigative Forensic Toolkit) Workstation can be used to analyze the evidence, find indicators of compromise to determine the scope of the incident, determine timelines, and perform a root cause analysis. While there is certainly much more that can be addressed on the topic of Amazon Linux EC2 Forensics, this paper provided step by step guidance for forensic analysis by using SIFT.

The focus of this paper will be on Infrastructure as a Service (IaaS) platforms for a number of reasons. IaaS consists of a large portion of the cloud services market, although not the largest, it is the fastest growing sector of the cloud services market. Accordingly, by way of Gartner news source, "The worldwide infrastructure as a service (IaaS) public cloud market grew 31 percent in 2016" [2]. IaaS is the portion of the cloud market that is largely replacing the traditional IT infrastructure. That traditional IT infrastructure would have been the focus of previous digital forensic investigations and civil litigation discovery and would have required a digital forensic expert to present evidence obtained during those examinations in court. Due to the variation and complexity of the many IaaS based services that can be built by a cloud customer, it is not likely that evidence derived from these infrastructures could be interpreted effectively in a court of law apart from the testimony of an expert witness. The expert opinion of a digital forensic analyst would be necessary to understand and communicate to the court evidence obtained from these environments. In this way, IaaS is different from SaaS, PaaS, and BPaaS which will have a common structure across all customers as well as standardized processes.

IaaS cloud service are rapidly replacing standard IT services traditionally hosted in datacenters that provide physical access to servers and data storage. The digital forensic investigator will need to be armed with new tools, techniques, and procedures in order to perform an effective investigation of various types of cloud environments. A number of researchers have made progress identifying and defining the new set of challenges [7]. The paper [67] proposed the solution with the use of

agents to regularly collect forensic data including memory images that can provide a comprehensive storage of digital forensic data.

This paper [36] explored that gap in available forensic procedures by studying the forensic acquisition of evidence from an Infrastructure as a Service (IaaS) cloud environment. This paper reviewed the general concepts of IaaS hosting and the published research with options to address the lack of established digital forensic options for acquisition. The paper also proposed a simplified methodology capable of capturing forensic data from an IaaS cloud. The forensic methodology presented in that paper was based on the usage of the AWS Command Line Interface (AWS CLI). The AWS CLI provided command-based access to all of the functions of the AWS Application Programming Interface. This was extremely beneficial to the forensic process.

Although research has been conducted to investigate a forensically sound acquisition of evidence from cloud systems, it has not, up to this point, reached a defined standard that can be referenced and agreed upon by digital forensic practitioners today. This paper explores that gap in available forensic procedures by studying the forensic acquisition of evidence from an Infrastructure as a Service (IaaS) cloud environment. IaaS cloud service are rapidly replacing standard IT services traditionally hosted in datacenters that provide physical access to servers and data storage. This paper reviews the general concepts of IaaS hosting and then reviews the published research with options to address the lack of established digital forensic options for acquisition. The paper then proposes a simplified methodology capable of capturing forensic data from an IaaS cloud. Cloud based services are rapidly replacing the traditional Information Technology services built on physical servers. The change is impacting both the platforms that services are hosted on as well as the makeup of organizations that are providing these services.

Many small organizations can now quickly move from their initial concept of a service to sell into the rollout of a large and complex technology stack in a matter of days. This is impacting both the technology used where physical servers are no longer purchased for building a technical infrastructure as well as the makeup of the personnel supporting this infrastructure where a team of one or two people can now replace what traditionally required a larger IT team. These small organizations supporting Internet based services are unprepared for performing large scale incident response or digital forensics across the infrastructure they have built and segmented.

Traditional methods of digital forensics that required physical access to devices are no longer possible with the cloud-based technology model. However, with a number of changes in methodology, it is possible to leverage the benefits of cloud-based infrastructure to rapidly perform a broad based forensic acquisition. In order to accomplish this, some of the more traditional foundations of digital forensics require evolution to work within the bounds of this new technology.

2.4 Big Data and Digital Forensics

The term Big Data regularly occurs in scientific and practical discussions [18]. One standard rule is Gartner, which reports Big Data; "high-volume, high-velocity, and high-variety information assets that demand cost-effective innovative forms of information processing for enhanced insight and decision-making" [45].

Big Data dramatically changes trade routes and activities. Because Business organizations want to benefit from their competitors, the organizations need more relevant information for the better, rational and realistic decision-making processes. In order to make the decision with fact-based theory, they are trying to be able to analyze raw data faster and easier to access the in-depth and valuable information. It drives for the aggressively practice of Big Data tools and platforms in business environment [58].

Big Data are usually the difficulties that necessitate instantaneous attention. The privacy and security are the most important challenges in Big Data that is sensitive and embraces conceptual, technical and legal significance. The business environment is becoming satisfy the advantages of Big Data tools and platform; consequently the involvement of Big Data in business environment is more and more powerful. The sensitive information including future plan of the organization are also store on the Big Data platform. In the never-ending race to stay ahead of the business competition, the malicious actors are keenly stealing this type of intellectual data through the illegal usages [60]. The risk of information leakage and criminal cases are increased in Big Data environment due to its high applicants.

As a new research area, digital forensics is a subject in a rapid development for Big Data environment is getting attention more than ever. Computing breach requires digital forensics to seize the digital evidence to locate who done it and what has been done maliciously and possible risk/damage assessing what loss could leads

to. In particular, for Big Data criminal cases, Digital Forensics has been facing even more challenge than original digital breach investigations [39].

Evidence data is crucial for Digital Forensics. Big Data evidence collecting is challenging. The Big Data-Digital Forensics issue is difficult due to some issues. One of them is physically identify specific wanted device. Data are distributed, customer or the digital forensics practitioner cannot have a fully access control like the traditional investigation does. This digital forensics challenges with Big Data is crucial. Almost all the traditional way of investigation is not appropriate any longer.

The volume of data in a case to be forensically processed is a major concern. While care must be taken not to exclude relevant data, all reasonable steps to "whittle down" the mountain of potential evidence should be taken. Many forensic tools provide the functionality needed to efficiently reduce the "noise" in a case and examine the files of most interest to the examiner. Failing to adequately address the volume of data in a case can exact significant costs in terms of time and money.

Digital Forensic has been an emerging discipline which has occasioned from the improperly use of computer and digital devices by lawbreakers to commit some types of crime [73]. The new technology and trends (such as Big Data and cloud computing) are also enabling the digital forensics become so extensive. As Big Data platform grows more sophisticated, so must the field of modern digital forensics.

2.4.1 Big Data Platforms and Digital Forensics

Big Data demands the storage of a massive amount of data. This makes it a necessity for advanced storage infrastructure; a need to have a storage solution which is designed to scale out on multiple servers. This feature highlights the finest open source file systems designed to cope with the demands imposed by Big Data. The article [114] explores the 6 best file systems which can act as the Big Data platform.

- Quantcast File System
- Hadoop
- Ceph
- Lustre File system
- GlusterFS
- PVFS Designed to scale to petabytes of storage

There are many kinds of distributed file systems such as network file systems (NFS) of SUN, Google File System (GFS) of Google, and HDFS (Hadoop distributed file system) of Apache and GLORY-FS of ETRI.

GFS cluster is simply a network of computers for managing the large amount of data. GFS is a proprietary distributed file system developed by Google to provide efficient, reliable access to data using large clusters of commodity hardware. Each cluster might contain hundreds or even thousands of machines. In each GFS clusters there are three main entities as shown in Figure 2.4:

- Clients
- Master servers
- Chunk servers.

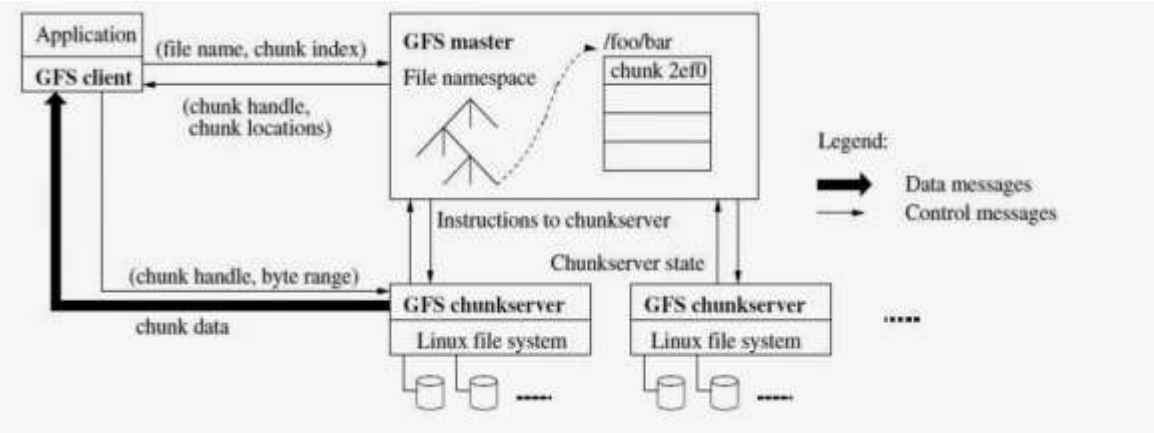

Figure 2.4 Architecture of GFS [99]

This paper [99] examined the feasibility of developing a forensic acquisition tool in a distributed file system. Using GFS and KFS distributed file systems as vehicles and through representative scenarios and examples, the authors develop forensic acquisition processes and examine both the requirements of the tool and the distributed file system must meet in order to facilitate the acquisition. The authors conclude that cloud storage has features that can be leveraged to perform acquisition (such as redundancy and replication triggers) but also maintains a complexity, which is higher than traditional storage systems leading to a need for forensic-readiness-by-design.

Perhaps it should come as no surprise that hackers have turned to Big Data as a weapon within their arsenal. Hackers gravitate towards technologies that make their

goals easier to achieve, and considering that many Big Data tools are based on open source software, gaining access to them is relatively easy. Big Data is simply another part of criminals' strategy, allowing them to act with more agility and sophistication, while also executing their cyber-attacks more quickly.

Many cyber criminals now traffic in massive information, selling valuable data. With Big Data analytics, criminals can mine that data for the most useful information in a fraction of the time, effectively monetizing it for their illegal purposes. Consequently, the Big Data platforms are the underlying source to trace the criminal activities and illegal usage of upper layer data set.

The paper [47] states the forensic challenges of Big Data platform. As mentioned in the previous section, there are the occurrences of the large amount of backlog per operation which are waiting for forensic analysis and have potential to be evidences. Generating evidences in such circumstances without prior knowledge of evidence location especially design for this environment may be like looking for a needle in a haystack. Logs are useful tools to troubleshoot and get evidence but in today's networks the number of logs to analyses makes this a daunting task. The velocity of these logs also makes harder to work of a digital forensic investigator.

In Big Data platform forensics, data volume in increasing at a much faster rate leading to backlogs in examination and analysis of many investigations [1]. The paper [73] found out to identify potential artifacts that remain on the client devices and servers involving the use of Syncany [102] as a private cloud storage solution supporting the Big Data Platform. It is one example of a number of products available with similar feature sets (other examples include GlusterFS [107], BeeGFS [43] and Ceph [50]). It is important to make the distinction between backend and frontend storage systems in the cloud computing environment as both are commonplace.

2.5 Hadoop Forensics

Hadoop is the most popular open source solution for deploying as Big Data platforms [39]. It is the implementation of the GoogleTM MapReduce parallel computing program framework. There are two major components of a Hadoop system: the HDFS file system for data storage, and parallel computing data processing framework [48].

The HDFS file system is among a number of distributed file system such as PVFS, Lustre, and Google File System (GFS). Unlike PVFS and Lustre, RAID is not used as part of the data protection mechanism . Instead, HDFS replicates data over multiple nodes, called Datanodes, to ensure reliability [31].

The HDFS file system architecture is designed after the Unix file system which stores files as blocks. Each block stored in a Datanode can be composed of data of size 64MB or 128MB as defined by system administrator. Each group of blocks consists of metadata descriptions that are stored by the Namenode. The Namenode manages the storage of file locations and monitors the availability of Datanodes in the system.

As the Hadoop forensic is the just emerging field for innovative technology, there is not many academic paper relating to Hadoop forensics. The basic structure and data volume of Hadoop Big Data Platform makes forensic investigation harder [104]. Analyzing the evidence that is logs provided by different applications, equipment and processes is vital but the non-standardization of the log format.

Evidence can be exposed to future science advances as the bit-by-bit copy generated it will be available to future investigations if the chain of custody is properly followed evidence would be still admissible in court. However, indexing speeds decrease as the amount of data raise [24] which seems to point to an unavailability of this method to Hadoop forensics.

Along with these, Hadoop forensics also presents the data size issue. In the case of a case in Hadoop clusters, the acquisition stage could consist acquisition of Petabytes of information. It takes 28 days produce a bit by- bit copy of a Petabyte image considering the current speed transfer of 6 GB/s [24]. Nowadays can be found Hadoop clusters storing data sets larger than 20 PB [39].

Although there are number of system level configurations that system administrators can implement to help secure Hadoop systems, they do not eliminate illegal incidents related to malicious user. Hadoop backlog can be the clue to trace the usages of cybercriminals. It is important that the prior knowledge are needed when determining what potential digital evidence should be gathered from data sources [65]. It leads to the obligation to conduct the forensic investigation researches relating to the new technical paradigms such as Hadoop Big Data Platforms.

Cho et al. [24] highlighted that the preceding forensic procedures are not suitable for HDFS based cloud system because of its characteristics; gigantic volume of distributed data, multi-users, and multi-layered data structures. These characteristics can generate two problems in the gathering evidences phase. One problem is that file blocks are replicated on different nodes while the other is the excessive time increase and storage of the original copying. They proposed a general forensic procedure and guideline for Hadoop based cloud system. In this proposed procedure, the authors added live analysis and live collection to the original forensic procedure to avoid the system suspension. By conducting the static and live collection simultaneously, the Hadoop forensic analysis can diminish the time for proof collection. However, they did not present a case study or specific scenario to illustrate their proposals.

2.6 Forensic Process Models and Frameworks

There are guidelines and procedures in relation to conducting digital investigations which are used by practitioners to reduce the risk of evidence being excluded from a legal process [64], [74] and [93]. When an investigation involves a new technology environment, the application of the guidelines may be difficult [16, 111]. It is important to note that there are several types of cloud services and each type with its potential different use in criminal activity, there will be variation in the way criminal investigation is carried out in each type of cloud service [71].

Some research works are conducted as references relating the forensic process framework. Digital forensics is the practice of collecting, analyzing and reporting on digital data in a way that is legally admissible. Along the digital forensic history, several process models were proposed for forensic investigation.

Forensic academia held large-scale consortiums and defined a general standard digital investigation process model [39]. This model contained six stages: planning, incident response, collect data, data analysis, presentation of finding and instance closure. This process model covers not only computer but also network forensics.

The National Institute of Standards and Technology (NIST) described the original forensic process model [8]. This model included the four phases: collection, examination, analysis and reporting. The relevant data were identified, labeled and record in the collection phase and the collected data were accessed and extracted in

examination phase. And then the results of the examination were analyzed to drive the useful information.

Quick [65] described that there are numerous types of cloud services that have a hypothetically different use in criminal actions. A need of sound digital forensic framework related to the client devices forensic analysis for identifying probable data holding is highlighted. The authors proposed a forensic investigation framework for cloud storage and then applied it for cloud forensic to solve the crime stories. The use of proposed framework was also beneficial to guide the research and applicable in digital forensic investigation.

Forensics investigators will face challenges while identifying necessary pieces of evidence from a Big Dataset, and collecting and analyzing that evidence [82]. In this article [113], the first working definition of Big Data forensics systematically analyzed the Big Data forensics domain to explore the challenges and issues in this forensics paradigm. This paper proposed a conceptual model for supporting Big Data forensics investigations and present several use cases, where forensics can provide new insights to determine facts about criminal incidents [113].

2.7 Chapter Summary

Through the reviewing of the previous literatures, this chapter is able to prove that the new technology and trends (such as Big Data and cloud computing) can make the digital forensics become so extensive. This chapter describes the theory background of this research and reviews the efforts of previous researchers in four portions. All of the previous works are the academic research that proposed forensic work and methodology for innovative technologies such as cloud computing, cloud storage, Big Data, Big Data platforms and Hadoop.

As Big Data platform grows more sophisticated, it becomes the field of modern digital forensics. As the forensic investigation for Hadoop Big Data Platform is just the emerging field for forensic community, there is not much research work in this environment. It can bring the challenges to forensic investigation as like it does in other research and technical areas. Therefore, today's forensic frameworks which are running on traditional systems have limitations on supporting forensic investigation. The sound forensic process frameworks are required. The traditional forensic analysis procedures have to be altered with the rise of Hadoop. In the next chapter, this

research will describe the proposed forensic investigation process framework for Hadoop Big Data Platform.

CHAPTER 3

FORENSIC INVESTIGATION FRAMEWORK FOR HADOOP BIG DATA PLATFORM

As highlighted in the literature review of previous chapter, a methodology and framework is compulsory to guide forensic investigations relating to the application of Hadoop Big Data Platform. This chapter outlines a proposed forensic investigation process flow of framework undertakings to enlarge upon the traditional process models of computer forensic analysis. This serves to expand the common framework, to be appropriate while dealing with Hadoop Big Data Platform.

3.1 Traditional Forensic Investigation Frameworks

Over the years, a number of digital forensics models have been proposed [83]. There are various digital forensic models occupied in digital investigative processes. However, these existing digital forensics methods may not be fit-for-purpose in the Big Data environment. Developments of the suitable Digital Forensic Framework were required for presenting the digital evidence in a better way. Digital forensic investigation process is needed to search digital devices directed for relevant evidence. There is a need for a forensic analysis framework to guide investigations, which is flexible enough to be able to work with future technology trend. This section proposes a forensic investigation framework for Hadoop Big Data Platform which expands the common framework.

A high-level conceptual framework to notify the problems (scalability and comprehension) of digital evidence has been suggested [75] how digital evidence can be presented in a good manner. The framework [75] consists of three main phases: explore, investigate, and correlate. The ‘explore’ phase is a starting phase for any digital forensics framework. The main goal of this phase is to provide a general overview the data of digital evidence and also enable the investigator to be more focus on the related information. In the ‘investigate’ phase, the investigators focused and visualized with greater information and details links.

An integrated (iterative) conceptual digital forensic framework was proposed [75] based on NIST, which emphasizes the differences in the preservation of forensic

data and the collection of cloud computing data for forensic purposes. Cloud computing digital forensic issues are discussed within the context of that framework.

The new model for a computer forensic investigation has been identified [68] and this model has the four stages of 'identification of digital evidence, preservation of digital evidence, analysis of digital evidence, and presentation of digital evidence'. The practice of intelligence analysis concerns itself with data analysis, and has been refined over the years. As outlined by in the paper [84], the intelligence process is a continuous cycle of tasking, collection, analysis, dissemination and feedback.

Digital forensics is the practice of collecting, analyzing and reporting on digital data in a way that is legally admissible. Quick [78] described that there are numerous types of cloud services that have a hypothetically different use in criminal actions. A need of sound digital forensic framework related to the client devices forensic analysis for identifying probable data holding is highlighted. This research focused on discovering whether there are cloud storage data miscellanies on prevalent client devices. The proposed forensic framework was applied in analyzing widespread cloud storage services; Google Drive, and Microsoft SkyDrive to find the residual artifacts on client devices; Windows 7, and an Apple IPhone. The paper [26] pointed out that cloud storage username and password can be identified from the log file and browser information. The usages of anti-forensic software did not eliminate the residual artifacts although full erase process can remove all data. The use of proposed framework was also beneficial to guide the research and applicable in digital forensic investigation.

3.2 Big Data and Hadoop Forensic Methodology

The evolution of Hadoop Big Data Storage System brings the challenges to forensic investigation as like it does in other research and technical areas. Therefore, today's forensic process models and frameworks which are running on traditional systems have limitations on supporting forensic investigation.

The authors [24] also highlighted that the preceding forensic procedures are not suitable for HDFS based cloud system be-cause of its characteristics; gigantic volume of distributed data, multi-users, and multi-layered data structures. These characteristics can generate two problems in the gathering evidences phase. One problem is that file blocks are replicated on different nodes while the other is the

excessive time increase and storage of the original copying. They proposed a general forensic procedure and guideline for Hadoop based cloud system. In this proposed procedure, the authors added live analysis and live collection to the original forensic procedure to avoid the system suspension. By conducting the static and live collection simultaneously, the Hadoop forensic analysis can diminish the time for proof collection. However, they did not present a case study or specific scenario.

While addressing the active nature of this environment, the forensic investigation process framework should fulfill with the following characteristics:

- Iterative nature to easily change between each phases
- Forensic data collection and analysis without system suspension
- Proactive preparation of the investigation facilities
- Integrity of the investigation
- Background knowledge of which are forensically important parts and files
- Documentation to learn the previous lessons

The traditional process models are limited to cope with the above issues. The sound forensic process frameworks are required. This section describes a forensic investigation framework that can guide the forensics analysis for Hadoop Big Data Platform. This proposed process model is based on NIST forensic process model. The next section describes the proposed forensic investigating process framework for Hadoop Big Data Platform as shown in Figure 3.1.

3.3 Proposed Forensic Investigation Framework for Hadoop Big Data Platform

As the contribution of this process model which is shown in Figure 3.1, there is a cycle on the phases. If the forensically sound data cannot be analyzed in the phase of analysis, the investigation can go back to preparation and collection phase to arrange the usable tools and techniques for efficient collection. Likewise, if there is a difficulty in reporting phase, re-operate the analysis phase. Throughout the process, detailed documentations of every step should be retained. These documents are applied to reconstruct the event in generating investigation report, which can be used by investigators. The investigator can prepare the important things for the next investigation by regarding the previous documentations.

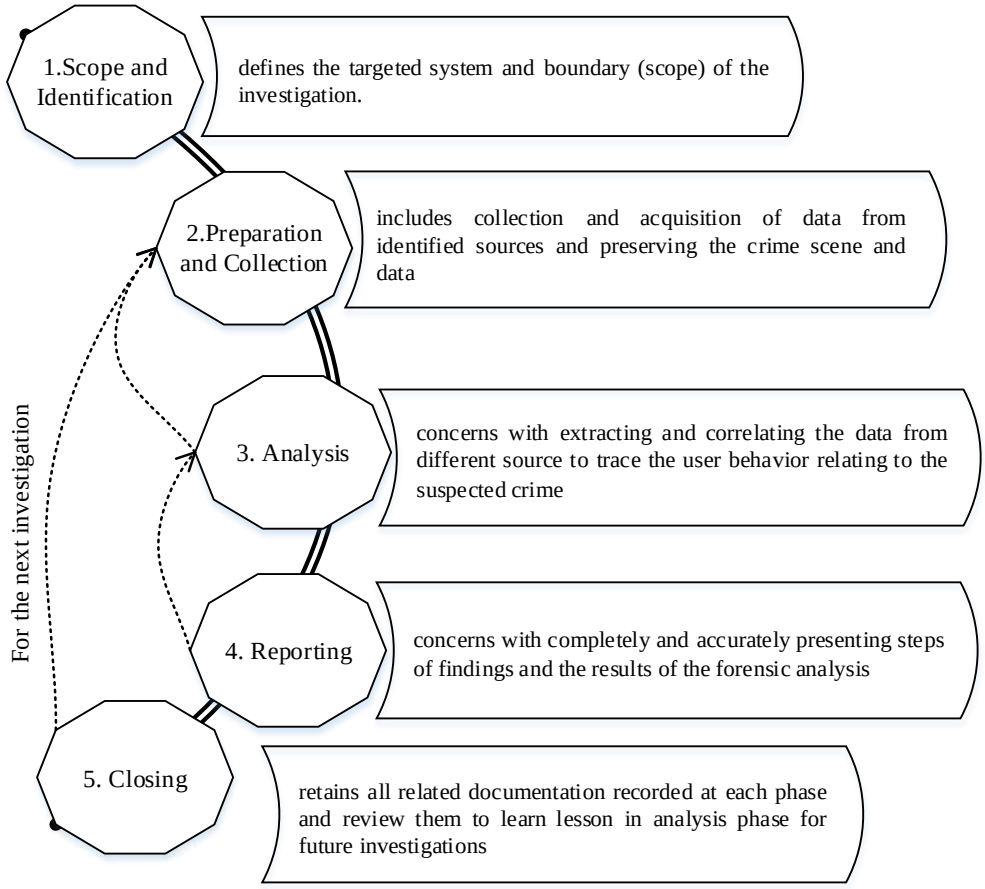

Figure 3.1 Proposed Forensic Investigation Framework for Hadoop Big Data Platform

3.3.1 Scope and Identification

It is the very first important phase to start the investigation. The investigator needed to survey the physical area of the system to set the edges of the investigation. This phase demonstrates the edges of the forensic investigation; the targeted system, the purpose of the investigation, what methods should be applied, when it is taken out, how long it may take, and who will conduct the investigation. During the identification, the following steps are taken into considerations:

- recognizing the possible data source
- locating the data sources
- Identifying the physical sources.

3.3.2 Preparation and Collection

It is the proactive measure that enables to maximize the ability as well as minimize the effort and unexpected risk associated with the investigation. Thus, the

investigators prepare a set of requirements for ongoing phases. This phase is operated based on the prior experiences or studies the documentations of previous investigations. This phase depicted the materials needed to prepare for the next phases and compares tasks and their required materials. The necessary resources for collecting data are Forensic Server, backup devices, or blank media. In multi-user storage server, the system suspension makes the serious problem to users. It makes to change the original data files. In emphasizing the integrity of the investigation, the data are collected remotely. The Forensic server is a facility machine to support remote collection and forensic analysis task. The investigator should setup one similar system environment with the identified system for studying the infrastructure of the targeted system.

Data collection is a critical phase in a digital investigation. The data analysis phase can be rerun and corrected, if needed. However, improperly collecting data may result in serious issues later during analysis, if the error is detected at all. If the error goes undetected, the improper collection will result in poor data for the analysis. For example, if the collection was only a partial collection, the analysis results may understate the actual values. If the improper collection is detected during the analysis process, recollecting data may be impossible. This is the case when the data has been subsequently purged or is no longer available because the owner of the data will not permit access to the data again. In short, data collection is critical for later phases of the investigation, and there may not be opportunities to perform it again. The goals of the collection phase are as follows:

- Forensically sound collection of relevant sources of evidence utilizing technical best practices and adhering to legal standards
- Full, proper documentation of the collection process
- Collection of verification information (for example, MD5 or control totals)
- Validation of collected evidence
- Maintenance of chain of custody

3.3.3 Analysis

The analysis phase is the process by which collected and validated evidence is examined to gather and assemble the facts of an investigation. Many tools and techniques exist for converting the volumes of evidence into facts. In some

investigations, the requirements clearly and directly point to the types of evidence and facts that are needed. These investigations may involve only a small amount of data or the issues are straightforward.

The process for analysis is dependent on the requirements of the investigation. Every case is different, so the analysis phase is both a science and an art. Most investigations are bounded by some known facts, such as a specific timeframe or the individuals involved. The analysis for such bounded investigations can begin by focusing on data from those time periods or involving those individuals. From there, the analysis can expand to include other evidence for corroboration or a new focus. Analysis can be an iterative process of investigating a subset of information.

Analysis can also focus on one theory but then expand to either include new evidence or to form a new theory altogether. Regardless, the analysis should be completed within the practical confines of the investigation.

After collecting the data, the relevant pieces of information are assessed and extracted from the collected data. The important task is to attach a copy of the collected data to the environment in a read-only manner. And then forensics analysis tools and techniques are applied. Among the analysis methodologies including; data mining, data correlation, anomaly detection, profiling, timeframe, data hiding, application and file, and ownership and possession, the suitable analysis methods for this environment are described as follows:

- **Keyword Searching:** Big Data investigations can contain both structured and unstructured data source. This information can contain keywords of wanted information. The simplest method is matching with keywords. The data is gathered and extracted from the metadata layer of the file system and then parsed to sort in order to be analyzed.
- **Timeline Analysis:** The end goal is to embody the incident activity done in the system including its date, the artifact involved, action and source.
- **Media and Artifact Analysis**: In this step, the investigator is overwhelmed with the amount of information that the investigator could be looking at. The investigator should be able to answer questions such as what programs were executed, which files were downloaded, which files were clicked on, which directories were opened, which files were deleted, where did the user browsed to

and many others. This analysis method emphasizes on registry and log files to trace the footages of the criminals or illegal usages. Hence, the knowledge of file systems, configuration Artifacts and registry Artifacts should be exposed to reduce the amount of data to be analyzed.

At the end of the analysis phase, the output is hanged over to the next phase to draw the event reconstruction and reporting.

3.3.4 Reporting

This phase presents the findings as the outcome of the investigation. The results obtained from above phases are organized to draw a conclusion. This phase is the presenting strategic for exposing the incidence (case); this must be full of clarity, completeness, and accuracy of the findings. The findings should be presented in a clear and understandable way that is accessible to non-technical spectators.

Nowadays, investigators typically use multiple possible evidence items. For that reason, as shown in Figure 3.2, investigators may end up with multiple reports on digital evidence items, generated using different tools [11]. The lack of standards in the reporting function of computer forensic tools may hinder the computer investigation process. When an investigator uses different forensic tools, he/she may face difficulties in exposing evidence items from into the official investigation report that could be presented to attorneys or clients.

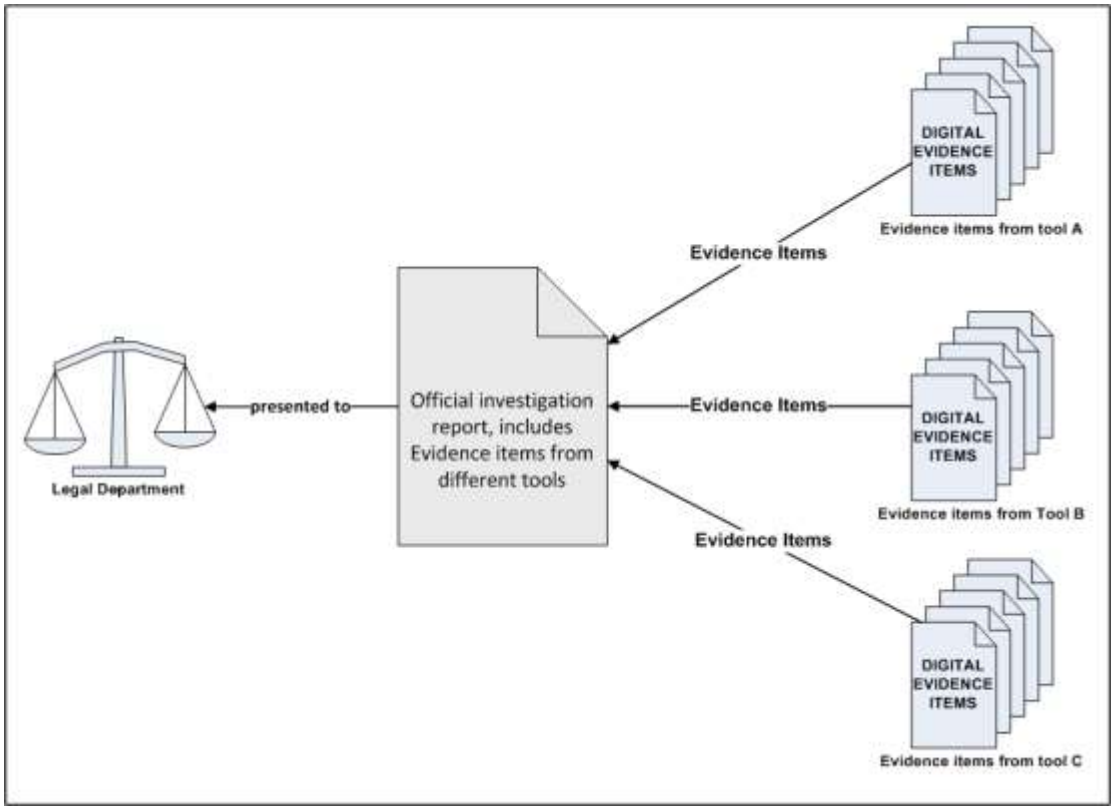

Figure 3.2 Forensic Reporting of Evidences from Variety of Sources

The report structure typically includes one or more sections detailing the evidence considered and the steps the investigator took to arrive at his findings. This is typically done by identifying the name, type, and characteristics of the evidences. There are many report formats relating to specific case type. A standard approach is to describe the process in chronological order, from identification through analysis. They perform to draw the event line with a specific feature (time, sequence).

An evidence custody form usually contains the following information [66]:

• Case Number

• Investigation Organization

• Investigator

• Nature of Case

• Location evidence was obtained

• Description of evidence

• Evidence recovered by

• Date and time

• Change of custody log

At this point, the authors were able to gather the data requirements to define the standard that could be used in reporting digital evidence items in computer forensic tools.

3.3.5 Closing

This phase retains all related documentation recorded at each phase of the investigation process. Review of the investigation process should be done so that the lesson can be learnt and used for future investigations. In this phase, the conclusion is drawn by deciding upon the result from reporting phase. All collected data through the process and resulting residual artifacts are stored and archived. The documentation should be created as the previous phases do. One point to notice is that this document is the finalized document; summarizing the activities and occurrences of the whole process. The resulting document is stored in the documentation file together with previous ones. The documentation file of whole process allows the investigators to prepare the required materials and methodologies for the future investigations.

3.4 Chapter Summary

This chapter stated some traditional forensic process frameworks and why they are not suitable for the Hadoop Big Data Platform forensics. It defined the proposed framework, which serves to extend the process of traditional forensic frameworks with the addition of the initial steps of Scope and Identification and preparation. In addition, final step of Closing is also offered. The framework is in the iterative nature, and the forensic practitioners can return to previous steps, whilst the overall investigation progresses. This serves to expand common digital forensic analysis frameworks to be applicable when dealing with the Hadoop Big Data Platform.

The proposed framework can give the answer of the research question 1 which is described in previous chapter 1; section 1.6. This framework is applied in the following chapters dealing with the forensic research of the popular Hadoop Platforms to provide a guiding framework to step the process through, as would be the case in a digital forensic investigation, and test the application of the proposed framework with the crime sceneries. This follows the process of a common digital

forensic examination, enabling forensic examiners to apply the proposed framework to real-world investigations and also help to extract the forensic evidences.

CHAPTER 4

FORENSICS INVESTIGATION ON HORTONWORKS DATA PLATFORM

Hadoop platforms are progressively implied by individuals, and organizations to process the large amount of information. A number of companies became bundle Hadoop and related technologies into their own Hadoop distributions as the Hadoop Platforms. The three prominent Hadoop Platforms are MapR Distribution of Hadoop, Cloudera Hadoop Distribution, and Hortonworks Data Platform [51]. The Hortonworks Data Platform (HDP) embraces Apache Hadoop and facility software packages [52] to store, process and analyze the Big Data sets. HDP is designed to which including core Hadoop technology such as the HDFS, MapReduce, Yarn and also additional software packages. This chapter describes the discovering the residual artifacts (artifacts) of two types HDP: Ambari Hortonworks Data Platform (Ambari HDP) and Non-Ambari Hortonworks Data Platform (Non-Ambari HDP).

4.1 Ambari Hortonworks Data Platform

The installation of HDP is with the help of Ambari management system: Ambari HDP. The package Ambari is the open source installation and management system for Hadoop Platforms. In Ambari HDP, Ambari can install core Hadoop and other facility software packages [55].

4.1.1 Architecture of Ambari HDP

Apache Ambari helps for managing the complexity of large-scale, distributed data storage and processing infrastructure using clusters of commodity hosts networked together [51]. It collects a wide range of information from the cluster's nodes and services. It can express as an easy-to-use and centralized interface: Ambari Web.

Ambari Web presents the service information which is used to construct and administrate the HDP cluster to execute basic operations of the responsibility to start and stop services, construct additional hosts to cluster, and update configurations. Any user can view Ambari Web features. Users with administrator-level roles can access more options of operator-level. For example, an Ambari administrator can

manage cluster security, an administrator can monitor the cluster, but a view-only user can only access features to which an administrator grants required permissions.

The Ambari Server collects information from all clusters [30]. Every host has a replica of the Ambari Agent, which permits the Ambari Server to control the host. The following Figure 4.1 is a simplified representation of internal architecture of Ambari.

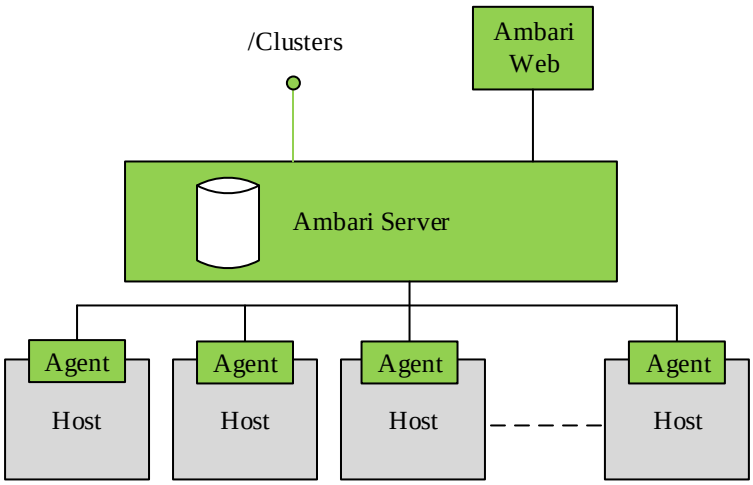

Figure 4.1 Internal Structure of Ambari Architecture [54]

Ambari Web is a client-side application written with JavaScript that requests the Ambari REST API of Ambari Server to contact cluster and performs cluster procedures [54]. After authenticating to Ambari Web, the application authenticates to the Ambari Server. Communication between the browser and server occurs asynchronously using the REST API. The Ambari Web UI periodically accesses the Ambari REST API, which resets the session timeout. Therefore, by default, Ambari Web sessions do not timeout automatically.

The Figure 4.2 shows the architecture of Ambari HDP 2.3. The Namenode manages the storage of file locations and monitors the availability of Datanodes in the system. The ResourceManager manages resources and allocates the resources to the application. The ResourceManager has Scheduler and manages the Node Manager of other hosts in cluster. The Scheduler performs the scheduling function based the resource requirements of the client applications. NodeManager is responsible for launching containers, each of which can house a map or reduce task. Ambari Server manages the Ambari Agents, and then connection is set up between the Ambari Server host and all other hosts in the cluster.

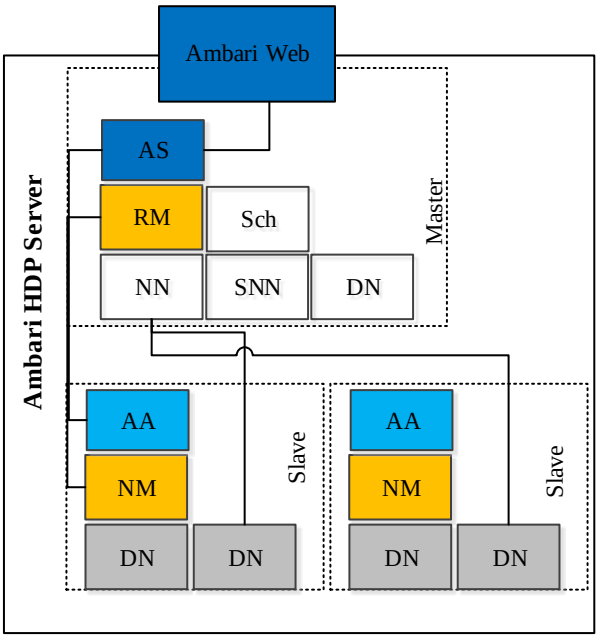

Figure 4.2 Architecture of Ambari HDP Server

AS= Ambari Server

AA= Ambari Agent

RM= Resource Manager

NM= Node Manager

Sch= Scheduler

NN= Namenode

SNN= Secondary Namenode

DN= Datanode

4.1.2 Installation and Configuration of Ambari HDP Server

Ambari manages and monitors the end-to-end solution of cluster. The Ambari user interface and REST APIs deployes configuration updates, and display services of all nodes in cluster as a pivot [9].

Ambari server is firstly installed by performing the following tasks in order to set up the HDP cluster [54].

1. Get Ready for an Ambari Installation
2. Download the Ambari Repository
3. Install the Ambari Server
4. Set up the Ambari Server
5. Start the Ambari Server

Ambari Server is installed with Ambari Agents on server host, and then password-less SSH connections is set up between the Ambari Server host and all other hosts in the cluster. The Ambari Server host uses SSH public key authentication to remotely access and install the Ambari Agent. The following steps are needed to be configured for setting up Password-less SSH. Generate public and private SSH keys on the Ambari Server host.

ssh-keygen

Copy the SSH Public Key (id_rsa.pub) to the root account on the target hosts. ssh/id_rsa, ssh/id_rsa.pub

Add the SSH Public Key to the authorized_keys file on the target hosts.

cat id_rsa.pub >> authorized_keys

Depending on the version of SSH, the user may need to set permissions on the .ssh directory (to 700) and the authorized_keys file in that directory (to 600) on the target hosts.

chmod 700 ~/.ssh

chmod 600 ~/.ssh/authorized_keys

From the Ambari Server, make sure it can connect to each host in the cluster using SSH, without having to enter a password.

ssh root@<remote.target.host> where <remote.target.host> has the value of each host name in the cluster.

After the Ambari installation, the HDP can be downloaded, installed and configured using Ambari as shown in Figure 4.3. These following steps are for configuring HDP using Ambari.

(i) Log in the Apache Ambari UI and start the Cluster Installation wizard. The default Ambari user name and password are admin and admin.

(ii) In the Select Version page of the wizard, remove all base URLs that do not apply to the operating system. Change the HDP Base URL to the URL appropriate for the provided operating system.

(iii) In the Choose Services page, select the following services needed to run an HDP cluster with full capabilities.

(iv) HDFS

(v) YARN + MapReduce2

(vi) ZooKeeper

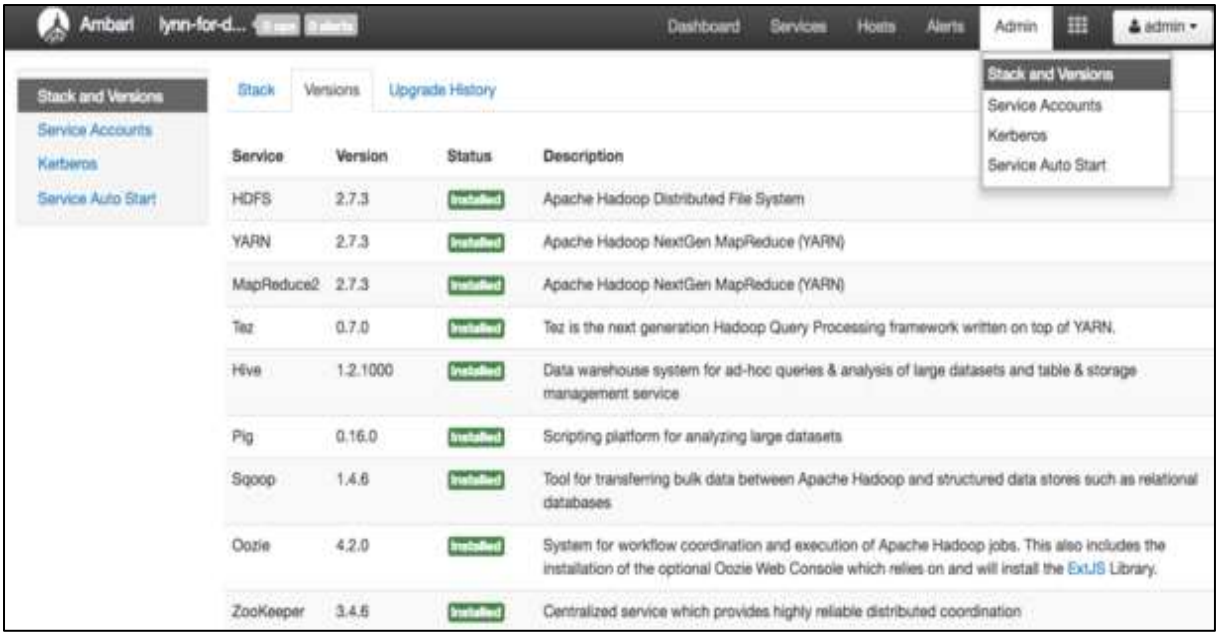

Figure 4.3 Ambari to Configure the HDP

4.1.3 Ambari HDP2.3 on Red Hat 7 Server Hosted on Amazon EC2

This sub section expresses the investigation of Hortonworks HDP 2.3 on Red Hat 7 server which is setting up on Amazon Web Services EC2. Hortonworks distribution provides Hadoop system based on Apache Hadoop for analyzing, storing and managing Big Data. Hortonworks is the only commercial vendor to distribute complete open source Apache Hadoop without additional proprietary software. Hortonworks is easier learning curve to provide IT friendly tools for users.

Gartner [46] reported that, AWS earned highest placement for ability to execute and furthest for completeness of vision in 2017 Gartner's Magic Quadrant IaaS for the 7th consecutive year. AWS is the overwhelming market share leader, with more than 5 times of the compute capacity in the use than the aggregate total of the other 14 providers. Amazon Web Services provides the IaaS services to build, secure, and deploy Big Data applications [2]. The needs of businesses for intensive treatment of very large volumes of data is solved by Amazon Web Services by providing the Elastic Compute Cloud infrastructure (EC2) which is intending to help the Big Data storage system through additional computing power. The effective usage of HDP 2.3 on a HDP 2.3 on Red Hat 7 server hosted on Amazon EC2mazon EC2 motivates us to investigate on this environment. The Figure 4.4 presents the HDP 2.3 Red Hat 7 server hosted on Amazon EC2.

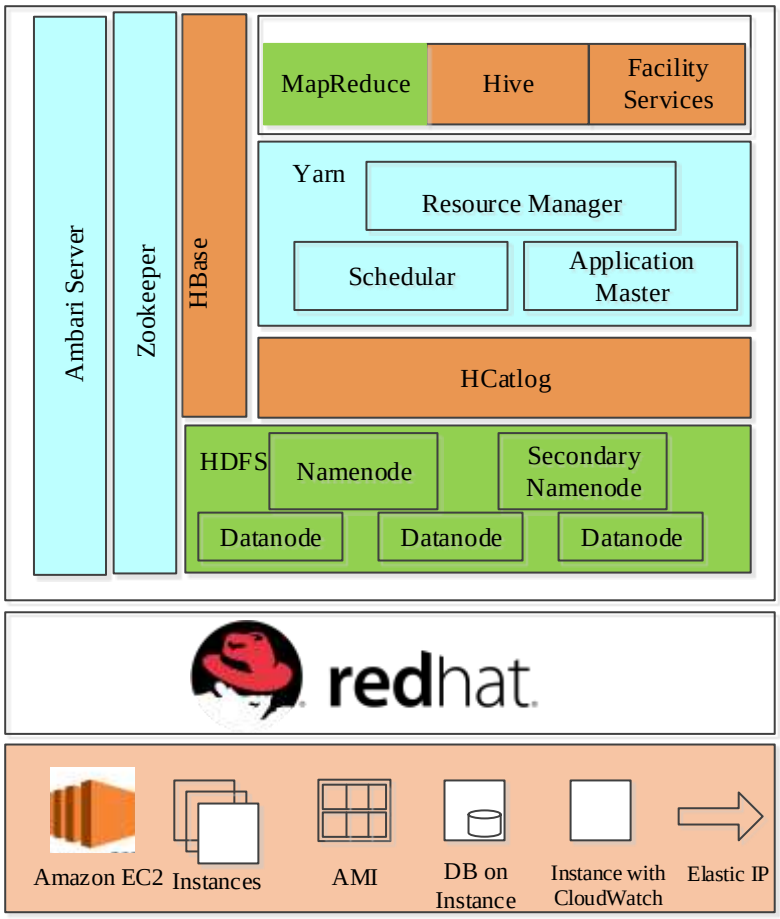

Figure 4.4 HDP 2.3 on Red Hat 7 Server Hosted on Amazon EC2

Each instance on Amazon EC2 is a virtual server in the cloud. An Amazon Machine Image (AMI) provides the information required to launch an instance. The instances can monitored using Amazon CloudWatch, which collects and processes raw data from Amazon EC2 into readable, near real-time metrics. A DB instance is an isolated database environment running in the cloud. In this system Red Hat 7 server is deployed as the instance of EC2 and Hortonworks Hadoop is installed.

4.1.4 Backlogs of HDP

In order to trace the criminal activities on Hadoop platform, the backlogs are able to reconstruct the crime actions. But, as the nature of Hadoop, many log files are updated per processing; there is a large amount of backlogs [97].

However, most of them are not forensically valuable. The following types of logs can be found on machines which are running Hadoop:

- Hadoop daemon logs: These are stored in the host operating system; these .log files contain error and warning information. By default, these log files will have a Hadoop prefix in the filename.
- log4j: These logs store information from the log4j process.
- Standard out and standard error: Each Hadoop TaskTracker creates and maintains these error logs which logs are stored in each TaskTracker node's /var/log/hadoop/userlogs directory.
- Job configuration XML: Hadoop JobTracker creates for tracking job summary details about the configuration and job run which can be found in the /var/log/hadoop and /var/log/hadoop/history directory.
- Job statistics: The Hadoop JobTracker creates these logs to store information about the number of job step attempts and the job runtime for each job.
    - Some noticeable log files which are update per Hadoop operation are shown in the following.
- Hadoop-
    - <user-running-Hadoop>-<daemon>-<hostname>.log
    - For example: Hadoop-Hadoop-Datanode-IP-xxxx.log
- JobTracker Logs
    - are created by the jobtracker.
    - home/Hadoop/logs/history/done/versionx/<host-job_id>/<year>/<month>/<day>/<serial>
- TaskTracker Logs
    - are produced by each tasktracker
    - are captured when a task attempt is run
    - /var/log/Hadoop/userlogs/attempt_<job-id>_<map-or-reduce>_<attempt-id>
- HDFS metadata
- fsimage
    - contains the file system complete state at a point in time
    - allocates a unique, monotonically increasing transaction ID
- edits
    - is a log listing each file system change (file creation, deletion or modification)

- is made after the most recent fsimage.

4.2 Non-Ambari Hortonworks Data Platform

The Hortonworks Data Platform, contributed by Apache Hadoop, is a immensely scalable and fully open source platform for storing, processing and analyzing large volumes of data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner. The HDP contains the necessary set of Apache Hadoop projects including MapReduce, Yarn, Hadoop Distributed File System (HDFS), and other facility software packages. These projects have been integrated and tested as part of the Hortonworks Data Platform release process and installation and configuration tools have also been included [53]. In Ambari-HDP, Ambari can automatically install core Hadoop and other facility software packages. The HDP installation excludes Ambari, and all installations are by manual. That type of HDP is so called Non-Ambari HDP.

4.2.1 Architecture of Non-Ambari HDP

HDP will be a fully open source distribution, including all the components used for a typical Hadoop deployment, including Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase and Zookeeper. The architecture of targeted Hadoop Platform HDP 2.3 of Non-Ambari HDP is shown in Figure 4.5. Hadoop Distributed File System (HDFS) is the core technology for the efficient scale-out storage layer, and is designed to run across low-cost commodity hardware. Apache Hadoop YARN is the prerequisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels. Hue 2.6.12 is applied as the web UI and gateway for all operations on HDP. With the Resource Manager and Node Manager, YARN provides much better resources utilization, which also adds to spinning up a cluster. YARN allows for running different applications that share a common pool of resources. There are no pre-defined Map and Reduce slots, which helps to better utilize resources inside a cluster. The ability to run non-MapReduce tasks inside Hadoop turned YARN into a next-generation data processing tool. Hadoop 2.0 features additional programming models, such as graph processing and iterative modeling, which extended the range of tasks that can be solved using this tool. Hue (Hadoop User Experience) is an open-

source web interface that supports Apache Hadoop and its ecosystem, licensed under the Apache v2 license.

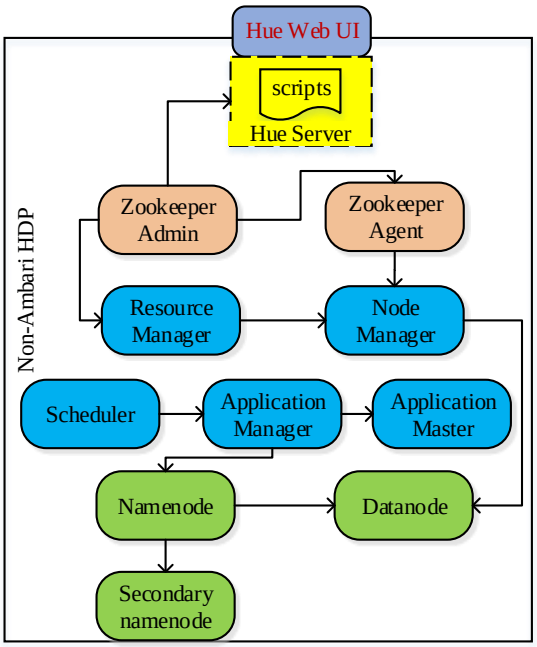

Figure 4.5 Architecture of Non-Ambari HDP

4.2.2 Installation and Configuration of Non-Ambari HDP

This sub section describes installation of the Non-Ambari HDP that exclusive of the help of Ambari. Use the following instructions to deploy the HDP.

(i) Prerequisites

(a) Software requirements are

- yum
- zypper
- php_curl
- reposync
- apt-get (for Ubuntu and Debian)
- rpm (for RHEL, CentOS, or SLES)
- scp

(b) For JDK (Java Development Kit) requirements, the correct JDK is needed to be installed on all cluster nodes. The supported JDKs for HDP are:

- Oracle JDK 1.7 64-bit or higher
- Oracle JDK 1.8 64-bit or higher

(c) Meta stored database requirement, new instance of PostgreSQL is installed on the host machine.

(ii) Remote Repositories Configurations

A remote yum repository is configured as the standard HDP installation fetches over the Internet. Access to the remote repository is set up for each of the hosts.

(iii) Deployment Type

While it is possible to deploy all of HDP on a single host, this is appropriate only for initial evaluation.

(iv) Information Collection

- The following command is used to check for the fully qualified domain name (FQDN) for each host

  ```
  hostname -f
  ```

- The hostname, database name, username, and password for the metastore instance. If an existing instance is used, the database user is created for HDP must be granted ALL PRIVILEGES on that instance.

(v) Environment Preparation

The synchronize of the clocks of each node in the cluster must be constructed. If the system does not have access to the Internet, set up a master node as an NTP xserver. The Security-Enhanced (SE) Linux feature are disabled and necessary ports must be open and available during the installation process by temporarily disabling the iptables.

(vi) Companion Files Downloading

HDP has provided a set of companion files, including script files and configuration files. These files are download and used as a reference point.

(vii) Environment Parameters Defination

The directories for install, configuration, data, process IDs, and logs based on the Hadoop Services which are planned to install. The directories for Hadoop Core are defined to set up the environment.

(viii) System Users and Groups Creation

In general Hadoop services should be owned by specific users and not by root or application users.

(ix) HDP Memory Configuration Settings

- Run the Yarn Utility Script
- Calculate the YARN and MapReduce Memory Configuration Settings.

(x) Hadoop Configuration

This section describes how to set up and edit the deployment configuration files for HDFS and MapReduce. It must be set up several configuration files for HDFS and MapReduce. Hortonworks provides a set of configuration files that represent a working HDFS and MapReduce configuration.

The configuration files are provided to set up the HDFS and MapReduce environment, complete the following steps:

1. Extract the core Hadoop configuration files to a temporary directory.

The files are located in the configuration_files/core_hadoop directory where decompressed the companion files.

1. Modify the configuration files.
2. On the node, create an empty file named dfs.exclude inside $HADOOP_CONF_DIR. Append the following to /etc/profile:

And then, HDP is validated. In order to start the HDP, (1) format and start HDFS (2)Smoke test HDFS (3) Configure YARN and MapReduce (4) Start YARN (5) Start the MapReduce JobHistory Server.

4.3 Discovering Residual Artifacts on Sever and Client Portions for Forensic Investigating on HDP 2.3

Without knowing the information where the artifacts are remained, it takes the considerable amount of time for forensic investigation. This section is intended to discover the residual artifacts on HDP as the proactive work before conducting the forensics. The discovered artifacts could help the future HDP forensics. Meant for conducting the forensic investigating research on this environment, the research questions are raised as follows:

- What data is remained on sever portion resulting from the use of the HDP to identify its use?
- The sub questions are raised from the above primary question.
    - What artifacts can be discovered on Hortonworks HDP 2.3 is running on it?

- What artifacts are remained on client portion resulting from the operation on HDP 2.3?

The investigation scope contains discovering residual artifacts on HDP server the attached client machines.

4.3.1 Experimental Setup for Discovering Residual Artifacts

The testing environment of Ambari HDP on AWS EC2 and Non-Ambari HDP are prepared for extracting the residual artifacts.

4.3.1.1 Experimental Setup of Ambari HDP on AWS EC2

The Non-Ambari HDP on Centos 6.7 is prepared the system environment which may be the same infrastructure with targeted environment. This similar system allows the investigator to study the nature of targeted system, test the tools and techniques. The step by step installation and configuration of Hadoop Server is as follows:

(i) Installing Java on Centos
(ii) Installing and Configuring SSH
(iii) Disabling IPv6
(iv) Installing the Hadoop Package

Secondly, the Ambari HDP is setting up on RedHat7 Server on AWS EC2. The environment of the same infrastructure with the targeted system is set up with the aim to study the targeted system. HDP 2.3 can be directly downloaded from the Hortonworks website [51]. EC2 storage space is rent to install the RedHat7. Ambari HDP 2.3 is deployed on the top of Red Hat. In order to setup the Hadoop via Ambari, the installation steps are:

A. Lunching an EC2 instance
B. Pre-requisites for setting up Hadoop in Amazon Web Services
C. Hadoop cluster installation (via Ambari)

Hadoop HDP is called by address ‘http://ec2-16……:8080/’ via the web browsers. The default sign in name is ‘admin’ and password is also ‘admin’.

4.3.1.2 Experimental Setup of Non-Ambari HDP on Centos 6.7

To perform the experiments, the Centos 6.7 is implemented as the underlying OS for setting up the Non-Ambari HDP. The testing environment and summary configurations of server are described in Table 4.1.

Table 4.1 System Configuration for Testing Environment of Non-Ambari HDP

Non-Ambari HDP Server Configuration
Operation system - Cent OS 6.7 Virtual memory size - 2GB Virtual HD size - 16GB HDP Version - Non-Ambari HDP 2.3 IP address/ URL - http://hostname/8888

4.3.1.3 Experimental Setup of Client Devices

The summary configurations of client for testing environment are described in Table 4.2 in order to discover the residual artifacts on client portion of HDP. The attached devices may be Windows PCs and Android smart phones.

Table 4.2 System Configuration for Testing Environment of Attached Client Devices

Client Device (Windows)	**Client Device (Android)**
Operation system - Windows 7 64 bit - Windows 10 64 bit Virtual memory size - 2GB Virtual HD size - 16GB Browsers - Mozilla Firefox 33.0.2 - IE 9.10.9200.16384, - Google Chrome 38.0.2125.111 m	Operation system - Windows 7 64 bit - Windows 10 64 bit Android Version - Android 4.0.4 Ice Cream Sandwich (ICS) Model Number - GT-P3113 Kernel Version - 3.0.8-android-x86+ehaung@u64 #2 Build Number - RomsWell_V1.1 Browsers - Mozilla Firefox 33.0.2 - IE 9.10.9200.16384, - Google Chrome 38.0.2125.111 m

4.3.2 Residual Artifacts of Ambari HDP on AWS EC2

The Tables 4.3, 4.4 and 4.5 are explanation of the remained artifacts on Ambari HDP 2.3 on AWS EC2 by tracing the usages. The usages include the primary operation services; uploading, downloading and reading functions. Tables 4.3, 4.4 and 4.5 reports the artifacts related to each operation service. In the Table 4.3, the discovered artifacts for the operation (uploading) are exposed.

Table 4.3 Residual Artifacts of Ambari HDP (File Uploading)

Location	File Names	Artifacts	Remarks
Home/Hadoop/logs	Hadoop-Namenode.log.2016-02-05	/user/IP/file_name.csv (original path of uploaded data set)	uploaded file name
Home/Hadoop/logs	Hadoop-xxx-Datanode--xxx.log.2016-02-05	Dest:192.168.32.34 OP: HDFS-WRITE	Dest IP, File operation
Home/Hadoop/logs	hdfs-audit	cmd=create src=/folder_name/file_name.csv	uploaded file name

The artifacts which expresses the source IP is like that 'admin (auth:PROXY) via user_name (auth:SIMPLE)' because the client machine accesses the sever via web browser. So the user name is stated as 'admin'. When the file is downloaded from server to local machines, the remained artifacts are represented in the Table 4.4.

Table 4.4 Residual Artifacts of Ambari HDP (File Downloading)

Location	File Names	Artifacts	Remarks
Var/log/Hadoop/hdfs/	hdfs-audit	2016-10-04 00:18:48, allowed=true ugi=admin (auth:PROXY) via xxx (auth:SIMPLE) IP=192.168.56.121	Source IP date File operation
Home/Hadoop/logs	hdfs-audit	cmd=getfileinfo src=/folder_name/file_name.csv	file name

The remained artifacts for the file operation (Reading) are shown in the Table 4.5. The artifacts are the source IP, the date of the operation, and the type of operation. These artifacts are located in the directories of /var/log/Hadoop/hdfs and /home/Hadoop which are exposed in the column "Location" of the Table 4.5.

Table 4.5 Residual Artifacts of Ambari HDP (File Reading)

Location	File Names	Artifacts	Remarks
Var/log/Hadoop /hdfs/	hdfs-audit	2016-10-04 00:18:48, allowed=true ugi=admin (auth:PROXY) via xxx (auth:SIMPLE) IP=/x192.168.56.121	Source IP date File operation
Home/Hadoop/ logs	hdfs-audit	cmd=open src=/folder_ame/file_name.csv	file name

The Hadoop HDP is uninstalled from Red Hat Linux7 sever with the command: yum remove Hadoop*, yum remove hdp*. The remaining Artifacts are:

- HDP 2.3 file under the link var/cache/yum/ x86_64/7 Server
- A sentence of public-repo-1.hortonworks.com 11864270 0 1474353396 in the timedhosts under /var/cache/yum/x86_47/7 server/HDP-2.3
- etc/yum.repos.d/hdp.repo

4.3.3 Residual Artifacts of Non-Ambari HDP

This sub section analyzes to identify the usage and discover residual artifacts. Among the large amount of log files, the contents in Namenode.log, syslog, blk_#####, job-######, hdfs-audit.log, hue-access.log files are useful for forensic environment. The contents of log file are able to uncover the criminal activity. Among them, the forensically complete logs which can embody the crime scene are hue-access.log and hdfs-audit.log. The criminal activity can be embodied by applying only these two log files even excluding other backlogs. In the configuration files for handling the logs, we can adjust the maximum amount of log. If the maximum amount is occupied, the old files are made as backup log files. The residual artifacts are found in the same files as the previous Ambari HDP investigation except hue-access.log because Ambari HDP configuration has no hue portion.

When accessing a Hadoop file operation via Hue Web UI, hue-access log maintain a history of requests. Information about request, including requested date/time, log level, client IP address, name of user operating the file, accessed method, HTTP protocol. More recent entries are typically appended to the end of file. Tables 4.6, 4.7 and 4.8 reports the artifacts related to each operation service. In the Table 4.6, the discovered artifacts for the operation (uploading) are exposed.

Table 4.6 Residual Artifacts of Non-Ambari HDP (File Uploading)

Location	File Names	Artifacts	Remarks
/var/log/hue	Hue-access	[04/Jan/2019 04:42:21 +0000] INFO 192.168.1.101 Mr.A - "POST /filebrowser/upload/file HTTP/1.0"	Source IP Date Operation User name
/usr/hdp/version /hadoop/log/	Datanode	2019-01-04 13:42:20,577 INFO bytes: 6496, op: HDFS_WRITE, offset: 0,	Date Operation
usr/hdp/version/ hadoop/log/	hdfs-audit	2019-01-04 13:42:20,405 ugi=Mr.A (auth:PROXY) via hue (auth:SIMPLE) cmd=create src=/user/Mr.A/brndlog.txt.tmp	UserName Opeartion File name

When the file is downloaded from server to local machines, the remained artifacts are the IP of source machine, the downloaded date, the type of operation and the user name as shown in Table 4.7. These artifacts are existed in the hue –access file and hdfs-audit file in the directories of /var/log/hue and /usr/hdp respectively.

Table 4.7 Residual Artifacts of Non-Ambari HDP (File Downloading)

Location	File Names	Artifacts	Remarks
/var/log/hue/	hue-acess	04/Jan/2019 05:54:43 +0000] INFO 192.168.1.101 Mr.A - "GET /filebrowser/download//user/	Source IP Date Operation User name
/usr/hdp/version /hadoop/log/	hdfs-audit	2019-01-04 14:54:43,827 ugi=Mr.A (auth:PROXY) cmd=getfileinfo src=/sample.csv	Date Operation
usr/hdp/version/ hadoop/log/	Namenode	2019-01-04 14:54:43,722 ugi=Mr.A (auth:PROXY) cmd=open src=/sample.csv	UserName Opeartion File name

The remained artifacts for the file operation (Reading) are shown in the Table 4.8. The artifacts are the source IP, the date of the operation, and the type of operation. These artifacts are located in the directories of /var/log/Hadoop/hdfs and /home/Hadoop which are exposed in the column "Location" of the Table 4.8.

Table 4.8 Residual Artifacts of Non-Ambari HDP (File Reading)

Location	File Names	Artifacts	Remarks
Var/log/Hadoop/ hdfs/	hdfs-audit	2016-10-04 00:18:48, allowed=true ugi=admin IP=/xxx.xx.xx.xxx	Source IP Date Operation User
Home/Hadoop/ logs	hdfs-audit	cmd=open src=/folder_ame/file_name.csv	Date Operation

The Figure 4.6 and 4.7 show the parameters and values of a record (entry) in hue-access.log and hdfs-audit.log file.

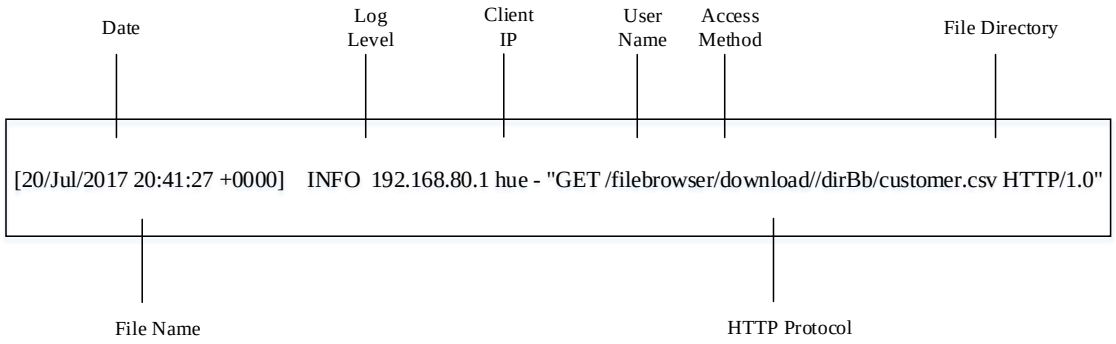

Figure 4.6 Parameters and Values of a Record in 'hue-access.log'

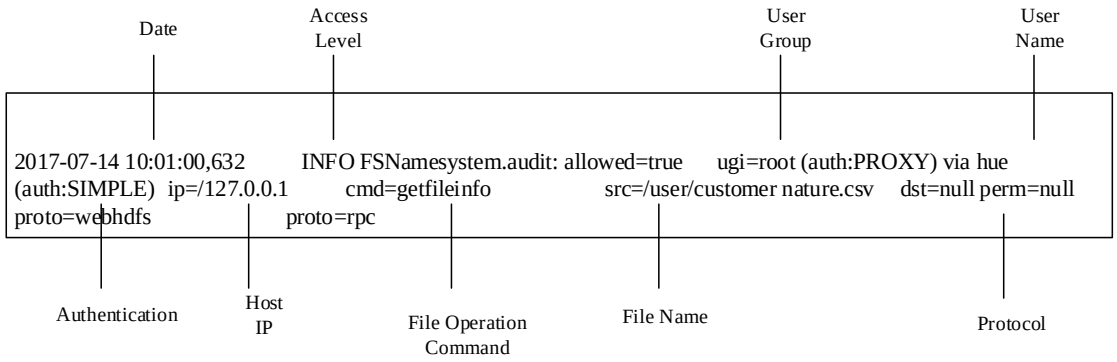

Figure 4.7 Parameters and Value of a Record in 'hdfs-audit.log'

4.3.4 Residual Artifacts of Client Devices

The experiments of discovering the residual artifacts on Windows and Android devices are performed by a series of read, upload, and download operations. The important parts and files of each browser are listed in Table 4.9 and 4.10. The artifacts found on Windows 7 PC are URL, date, time, and file name. Moreover, the browser, log file, the accessed web URL, website title, visited date and time can be identified. The residual artifacts of popular web browsers; Mozilla Firefox, IE and Google Chrome on Windows 7 PC are shown in Table 4.9.

Table 4.9 Artifacts of Windows Browsers for Primary File Operations

Mozilla Firefox 33.0.2	
File name	Path
Cache	%LocalAppData%\Mozilla\Firefox\profile\xxxxx.default\cache2\entries
History	%AppData%\Mozilla\Firefox\profile\xxxxx.default\places.sqlite %AppData%\Mozilla\Firefox\profile\xxxxx.default\formhistory.sqlite
Cookie	%AppData%\Mozilla\Firefox\profile\xxxxx.default\cookies.sqlite %AppData%\Mozilla\Firefox\profile\xxxxx.default\ permissions.sqlite
IE 9.10.9200.16384	
Cache	%LocalAppData%\Microsoft\Windows\TemporaryInternet Files\Low
History	%LocalAppData%\Microsoft\Internet Explorer
Cookie	%LocalAppData%\Microsoft\Windows\Cookies
Google Chrome 38.0.2125.111 m	
Cache	%LocalAppData%\Google\Chrome\user data\default\cache
History	%LocalAppData%\Google\Chrome\user data\default\history
Cookie	%LocalAppData%\Google\Chrome\user data\default\cookie

This experiment found that the majority of artifacts are stored in database files of the storage layer of Android. File Viewer plus [4] HHHexEditorNeo [7] and SQLite DB Browser [14] are used to decrypt the encrypted databases and to view the contents of the DB file. Although private web browsing like Orweb Browser cannot be traced in non-rooting the Android device, browser history and artifacts are able to be located in rooted Android device. The residual artifacts of popular web browsers of Android device are shown in Table 4.10.

Table 4.10 Artifacts of Android Browsers for Primary File Operations

File name	Artifacts
Android Default Browser	
Read Operation	
databases/browser2.db/ history databases/browser2.db/ images	URL, web page, file name, Date
Download Operation	
webviewCacheChromium.db/cookies	- URL, date
/data/data/com.android.browser/databases/browser/webview.db	- File name, directory
Upload Operation	
/data/data/com.android.browser/databases/browser/webview.db	- File name, directory
Dolphin Browser V-11.5.4	
Read Operation	
/data/data/mobi.mgeek.TunnyBrowser/ dolphin_webviewCache.db	- URL, date
Download Operation	
/data/data/mobi.mgeek.TunnyBrowser/databases/ download/	- file name - downloaded - directory
Upload Operation	
/data/data/mobi.mgeek.TunnyBrowser/databases/download/	- file name - directory
Firefox Mobile Browser 4.4	
Read Operation	
/data/data/org.mozilla.firefox/databases	- URL, date
Download Operation	
/data/data/org.mozilla.firefox/databases	- URL, date
Upload Operation	
/data/data/org.mozilla.firefox/databases/webview	- URL, date
Opera Browser 28.0.1764.90386	

Read Operation	
data/data/ com.opera.browser /databases	- URL, date
Download Operation	
data/data/ com.opera.browser /databases	- file name - downloaded directory
Upload Operation	
data/data/ com.opera.browser /databases	- file name - directory
Orweb: Private Web Browser 0.7.1	
Read Operation	
/data/data/info.guardianproject.browser/webview.db	- URL, date, user name

4.4 Case Study: Forensic Investigation on HDP 2.3

In this section, a crime scenario is presented and the investigation is conducted to this HDP. An example crime scenario in HDP 2.3 which is extended from the M57 Case [77] is described as follows.

Background:

The Company, DEF organization uses the services of Hortonworks Hadoop which is built on Centos 6. Every authorized person in this organization can access this through 'http://192.168.32.34/8080/' from their own web browsers. They use this for obtaining the service of uploading, downloading and opening the files on it. One suspected case is that valuable data set had been leaked to a competitor. The file name is "customer nature.csv".

Case: Document Exfiltration Case of Digital Corpora: M57 Case [38]

The data-set containing confidential information named "customer nature.csv" was posted as an attachment in the forum of a competitor's website. In the initial investigation, the prime suspect was Mr. Felix, who managed credential files. He had used a personal computer for business purposes.

The investigator has been given:

- Mr. Felix's PC and Android Device
- A copy of the targeted data-set file

Question to answer:

Did Mr.Felix commit the crime?

4.5 Forensic Investigation on Client and Server Portions of HDP 2.3

This section expresses the investigation of Ambari HDP 2.3 by applying the proposed Forensic Investigation Framework.

4.5.1 Scope and Identification of Investigating HDP 2.3

In this investigation, the target system for investigation is Red Hat 7 server on which HDP 2.3 is installed and hosted Amazon EC2. The objective of the investigation is to discover the residual artifacts that Mr. Felix connects the HDP Server on Amazon EC2, opened and downloaded the dataset to his PC. Therefore, the identified sources are Red Hat 7 server, hosted on EC2 and Window 7 64 bit PC.

4.5.2 Preparation and Collection of Investigating HDP 2.3

Forensic tools, methods and other facility software for collection and analysis are also prepared.

(i) Forensic Data Collection at Server Site

The remote data collection from server instance on EC2 cannot secure because the forensic server needs to connect the internet. For the secure data collection, EC2 instance can be exposed to the virtual version of Citrix Xen, Microsoft Hyper-V, or vmware, vSphere. Exporting an instance is useful to deploy a copy of EC2 instance in on-site virtualization environment. The server instance is exposed to vmware by using the Amazon EC2 API tools.

For the volatile data collection, imaging the memory of the server “dd if=/dev/mem of=media/usb/memory.image”

The non-volatile data is collected by imaging the hard drive of the machine.

“dd if=/dev/sda | /media/usb/disk.image ”

The resulting imaging disks are mounted on forensic server.

(ii) Forensic Data Collection at Client Site

To create the forensic image of hard disk of Windows, the write blocker must be used to ensure that no data is written back to hard drive. The AccessData FTK

imager 3.0.0.143 [1] is used for imaging by blocking write mode. After imaging the hard drive, the image file is collected in a forensic server.

For the investigation of Android devices, the approach of acquiring and extracting the history data from Android by Android Software Development Kit (SDK) and Android Debug Bridge (ADB) [115].

4.5.3 Analysis for Investigating HDP 2.3

In analysis phase, the residual artifacts described in above section assist as the prior knowledge of which are the important parts and files for forensic analysis. The important residual artifacts found in log and metadata files (Server site) are client IP, date, time, file name, file allocate path and file operation as exposed through Figure 4.8 and 4.9 Figure 4.10 illustrates the downloaded file name which is found in a metadata file; edits-inprogress_0000000083.hex.

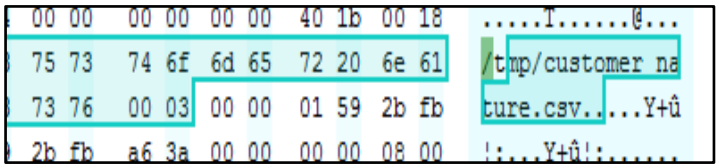

Figure 4.8 The metadata file of HDP Server

The following Figure 4.9 shows the client IP, operation type, file name and its path which are remained in hdfs-audit.log. Hex Editor Neo and FileViewerPlus [41] are applied to open Hex files and log files.

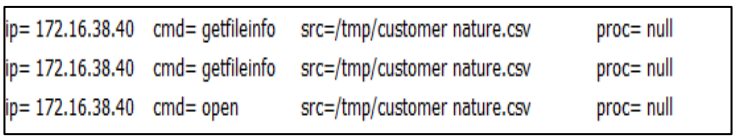

Figure 4.9 The 'hdfs-audit.log' of HDP Server

In order to analyze the collected data, variety of forensic analysis and recover tools are tested to apply. However among the various tools, this section depicts some effective tools that is adaptable with this case and targeted system Browserhistoryspy [111] to view web browser history, Recova [85] to recover deleted files, the residual artifacts found in client devices are URL, date, time, file name as shown in Figure 4.10. It illustrates the browser log file, in that file, the accessed web URL, website title, visited date and time can be identified. The web address of server machine is

also found in browser cache entries file of client machine as shown in Figure 4.11. Because of the hex file, it is viewed by the hex file reader; FileViewerPlus [41].

Figure 4.10 Browser Log File of Mozilla Firefox 33.0.2 of Windows 7 PC (viewed by browser spy tool)

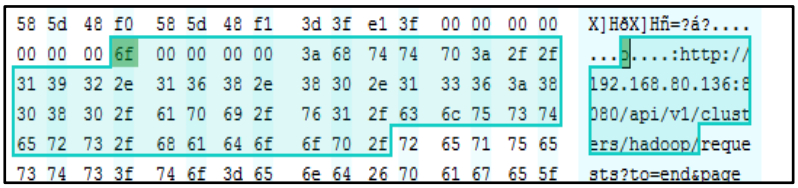

Figure 4.11 Browser 'cache_entries' File Windows 7 PC

While investigating the android smart phone of client device, the web address of server machine and the file name are history of databases.db file of Dolphin Browser V-11.5.4 as shown in Figure 4.12. These artifacts can be extracted by the sqlite.

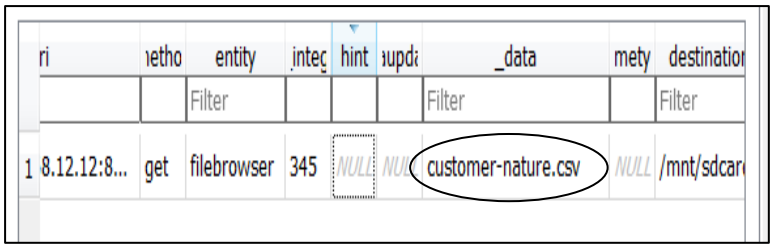

Figure 4.12: Artifacts in "databases. db/ download" of Dolphin Browser V-11.5.4 in Android Device

4.5.4 Reporting of Investigating HDP 2.3

The investigator arranges the finding evidences to embody the crime and reconstruct the event line. The forensic report for the investigation of 'Document Exfiltration Case' on HDP Platform is presented in Table 4.11.

Table 4.11 Forensic Report for the Document Exfiltration Case

FORENSIC REPORT FOR HADOOP HDP

INVESTIGATOR : Mr. Adam

Case Type : Suspect

Case Number : 1

1. Status: Complete

2. Summary of finding: To find the related information of "customer nature.csv".

Server Side Evaluation

Step 1 : Finding the residual artifacts that log and metadata files on Hadoop

Step 2 : Copying disk by imaging commands

Step 3 : Checking integrity with MD5

Step 4: Mounting on forensic server

- Client Side Evaluation

Step 1 : Finding the Internet History of PC and Android Device that are access Hadoop server via web browser

Step 2: Extracting/parsing dump file of disk by analysis tools.

Step 3 : Calculating hash value by using MD5 hasher

3. Items Analyzed:

TAG Number:	ITEM DESCRIPTION:
0100	HDP Server
0166	Personal Computer (PC), Serial# 123457
0167	Android 4.0.4 Ice Cream Sandwich (ICS) GT-P3113

4. Finding for item 0100

i. The specification of item 0100: CPU-8core, 20GHz, 32GB RAM

ii. The examined hard drive was found to contain a Centos 6.7 operating system

iii. Among the metadata and log file, the discovered residual artifacts in audit.log and edits-inprogress_00000000000083.hex :

a. Login time is starting from 23/Dec/2016 20:46:44

b. Open the "Customer nature.csv" at 23/Dec/2016 21:18:43

c. Download again the file "/tmp/ Customer nature.csv" at 23/Dec/2016 21:18:49.

5. Finding for item 0166 and 0167

i. The examined hard drive was found to contain windows 7 ultimate 64 bits operating systems

The examined device was found to Android 4.0.4 Ice Cream Sandwich (ICS) "customer nature.csv" and date are found in residual artifacts of Dolphin Browser mobi.mgeek.TonnyBrowser\

i. "customer nature.csv" and date are found in internet history of Mozilla Firefox at "rnllugnp.default-1456638777411\places.sqlite"

ii. When using browsing history tools, user Mr. Felix access the http://172.16.38.32:8080 at 23/Dec/2016 starting from 20:46:44

7. Provided Items: Along with this hard copy, the report is also submitted with one CD for electronic copy report. The CD contains the same information with this hard copy.

4.5.5 Closing of Investigating HDP 2.3

As the result of the following of report Mr. Felix stole the "customer nature.csv" by using Mozilla Firefox web browser from his PC and Android Device. The documentation in each phase is stored. The difficulties, solutions, usage of tools and all experiences of each step are reviewed for the preparation phase of the next investigations.

4.4. Chapter Summary

This chapter discusses the architecture of two types of HDP; Ambari HDP and Non-Ambari HDP depending on the installation method. In Ambari HDP, the Hadoop

cluster and facility software packages can be installed using Ambari and deployed automatically. In non- Ambari HDP is the manual installation of Hadoop cluster by step by step configuration. The forensically important files and residual artifacts of HDP 2.3 are discovered by applying the proposed forensic investigation framework. The forensic investigation framework for Big Data Platform is used in the investigation of HDP on both server and client portions.

The artifacts of HDP server and client devices of Windows and Android devices are exposed. The artifacts found on HDP server are client IP, accessed timestamp, type of operation and file name. These artifacts are mostly located in .meta and .log files.The forensically important files and residual artifacts of popular web browsers on client devices are extracted. The artifacts found on client machine are URL, date, time, and file name. Moreover, the browser, log file, the accessed web URL, website title, visited date and time can be identified. The web address of the server machine is also found in the browser cache entries file of the client machine. As outlined, there are a wide range of investigation points for an examiner to determine the use of HDP, such as; directory listings, prefetch files, registry, browser history, and memory captures.

A popular forensic case study of "Document Exfiltration Case of Digital Corpora: M57 Case" is investigated to demonstrate the application of the research findings. The evidences have been extracted from both HDP server and client device to expose who committed the crime. From the perspective of finding residual artifacts, the different finding between Ambari and non-Ambari is that "hue-access" log can be found in the non-Ambari HDP. The "hue-access" log can maintain the information about request, including requested date/time, log level, client IP address, name of user operating the file, accessed method, HTTP protocol.

CHAPTER 5

THE FORENSIC INVESTIGATION ON CLOUDERA DISTRIBUTION OF HADOOP

In this chapter, the focus is the forensic investigation on Cloudera Distribution of Hadoop (CDH) Server and client devices. The proposed framework is also used to step through the process in a logical manner. The forensic investigation is conducted for tracing the file operations on CDH Platform. The residual artifacts are discovered on CDH Server and client devices of Windows PC and Android device. The real-world crime scenario provided by CYFOR is investigated as the case study by applying the proposed forensic investigation framework.

5.1 Cloudera Distribution of Hadoop

The retrieval of digital evidence to embody the crime scenes on storage service of Hadoop Platform can be a challenge in forensic investigation, due to its complex infrastructure and, lack of knowledge on location of digital evidence. Accordingly, forensic researchers are moving towards the investigation researches of locating and documenting the residual artifacts to trace the criminal activities on Hadoop Platform. Cloudera Distribution for Hadoop (CDH) is a popular Hadoop Platform, providing users a cost-effective, and in some cases free with the ability to access, store, and process data.

Cloudera was the first vendor to offer Hadoop as a package and continues to be a leader in the industry. Its Cloudera CDH distribution, which contains all the open source components, is the most popular Hadoop distribution Cloudera is the best known player and market leader in the Hadoop space to release the first commercial Hadoop distribution. Today announced that it is positioned as a leader in The Forrester Wave™: Big Data Hadoop Distributions, Q1 2017 report [86].

CDH is the most complete, tested, and popular distribution of Apache Hadoop and related projects. CDH delivers the core elements of Hadoop – scalable storage and distributed computing – along with a Web-based user interface and vital enterprise capabilities. CDH is Apache-licensed open source and is the only Hadoop solution to offer unified batch processing, interactive SQL and interactive search, and role-based access controls.

The Hadoop backlogs of CDH are useful to trace illegal usages and embody the crime scene. Obtaining these artifacts from log files could provide forensic examiners with valuable evidence.

5.2 Architecture of CDH

Cloudera, the global provider of the fastest, easiest, and most secure data management and analytics platform built on Apache Hadoop and the latest open source technologies. The infrastructure of CDH Storage server and client is shown in Figure 5.1.

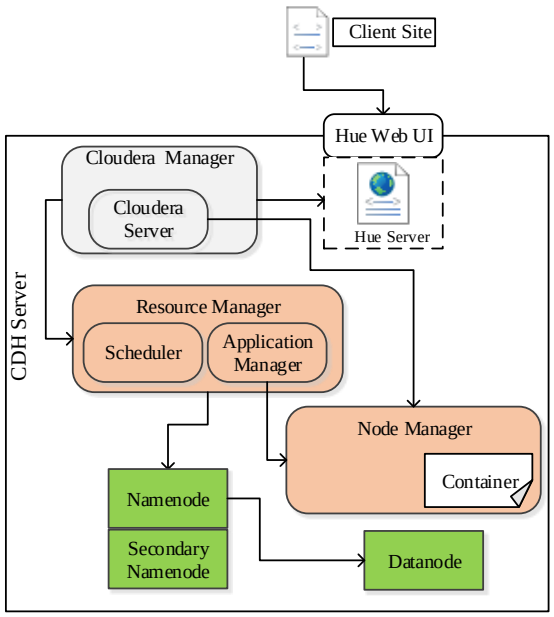

Figure 5.1 Infrastructure of CDH Server

Cloudera Manager is an end-to-end application for managing CDH clusters. Cloudera Manager sets the standard for enterprise deployment by delivering granular visibility into and control over every part of the CDH cluster—empowering operators to improve performance, enhance quality of service, increase compliance and reduce administrative costs.

With Cloudera Server Manager, the complete CDH stack can easily deploy and centrally operate and other managed services. The application automates the installation process, reducing deployment time from weeks to minutes; gives a cluster-wide, real-time view of hosts and services running; provides a single, central console to enact configuration changes across the cluster; and incorporates a full range of reporting and diagnostic tools to optimize performance and utilization. This

primer introduces the basic concepts, structure, and functions of Cloudera Server Manager Resource Manager is the configuration of resources and a policy for scheduling the resources among YARN applications running in the pool.

Cloudera Manager provides granular visibility into and control over every part of the CDH cluster—empowering operators to improve performance, enhance quality of service, increase compliance, and reduce administrative costs. With Cloudera Manager, it can be easily deployed and centrally operated the complete CDH stack and other managed services. The application automates the installation process, reducing deployment time from weeks to minutes; gives a cluster-wide, real-time view of hosts and services running; provides a single, central console to enact configuration changes across the cluster; and incorporates a full range of reporting and diagnostic tools to help to optimize performance and utilization. Cloudera Manager also provides an API can use to automate cluster operations.

5.3 Installation and Configuration of CDH

This section provides instructions for installing CDH and using the command line interface (CLI). In a deployment Cloudera Manager is responsible for installing, configuring, and managing CDH and other services. Cloudera strongly recommends using Cloudera Manager, which simplifies the installation, configuration, and maintenance of CDH and other services. CDH and Cloudera Manager (CM) Supported Operating Systems. CDH provides 64-bit packages for select versions of RHEL-compatible, Centos, and Ubuntu operating systems. This procedure is recommended for installing CM and CDH for production environments. The general steps in the installation procedure are as follows:

- **Step 1: Configure a Repository**

CM is installed using package management tools such as yum for compatible systems. These tools depend on access to repositories to install software. The internal repository can be created for hosts. To use the Cloudera repository:

- Download the cloudera-manager.repo file to directory on the CM server with the following commands

- /etc/yum.repos.d/
- sudo-rpm-import https://archive.cloudera.com/cm5/redhat/6/x86_64/cm/RPM-GPG-KEY-cloudera

Step 2: Install JDK

For the JDK, install the Oracle JDK version provided by Cloudera using Cloudera Manager, a different Oracle JDK directly from Oracle, or OpenJDK. Most Linux distributions supported by Cloudera include OpenJDK, but manual installation instructions are provided below if needed. OpenJDK is supported with Cloudera Enterprise 5.16.0 and higher. Requirements are:

- The JDK must be 64-bit. Do not use a 32-bit JDK.
- The installed JDK must be a supported version as documented in CDH and CM Supported JDK Versions.
- The same version of the JDK must be installed on each cluster host.
- It is installed at /usr/java/jdk-version.

- **Step 3: Install Cloudera Manager Server**

Install the Cloudera Manager Server packages either on the host where the database is installed, or on a host that has access to the database. The health status and configuration of server can be viewed by Cloudera Manager as shown in Figure 5.2.

sudo yum install cloudera-manager-daemons cloudera-manager-server

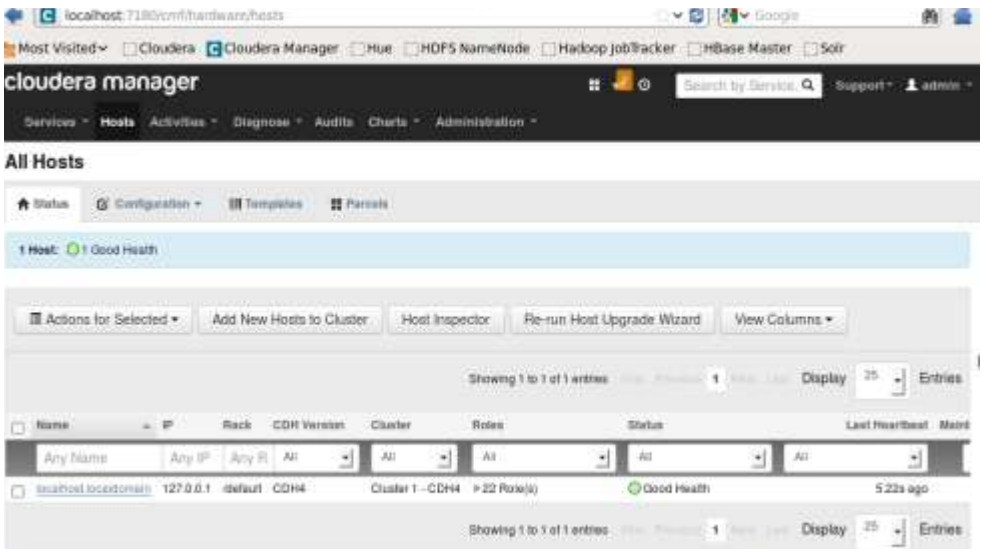

Figure 5.2 Server Status shown in Cloudera Manager

- **Step 4: Install Databases**

  CM uses various databases and datastores to store information about the CM configuration. Cloudera recommends installing the databases on different hosts than the services. Separating databases from services can help isolate the potential impact from failure or resource contention in one or the other. It can also simplify management in organizations that have dedicated database administrators. PostgreSQL, or MariaDB, or MySQL, or Oracle database can be used for the CM Server.

- **Step 5: Set up the CM Database**

  CM Server includes a script that can create and configure a database for itself. The script can:

  - Create the CM Server database configuration file.
  - Create and configure a database for CM Server to use.
  - Create and configure a user account for CM Server.

- **Step 6: Install CDH and Other Software**

  After setting up the Cloudera Manager database, start CM Server, and log in to the CM Admin Console and start CM Server:

  - sudo service cloudera-scm-server start

  Other services including Hadoop are installed via the CM Console as shown in Figure 5.3.

- **Step 7: Install Hue Server**

  The Hue Server is a container web application that sits between CDH installation and the browser.

  - sudo yum install hue
  - sudo yum install hue-plugins

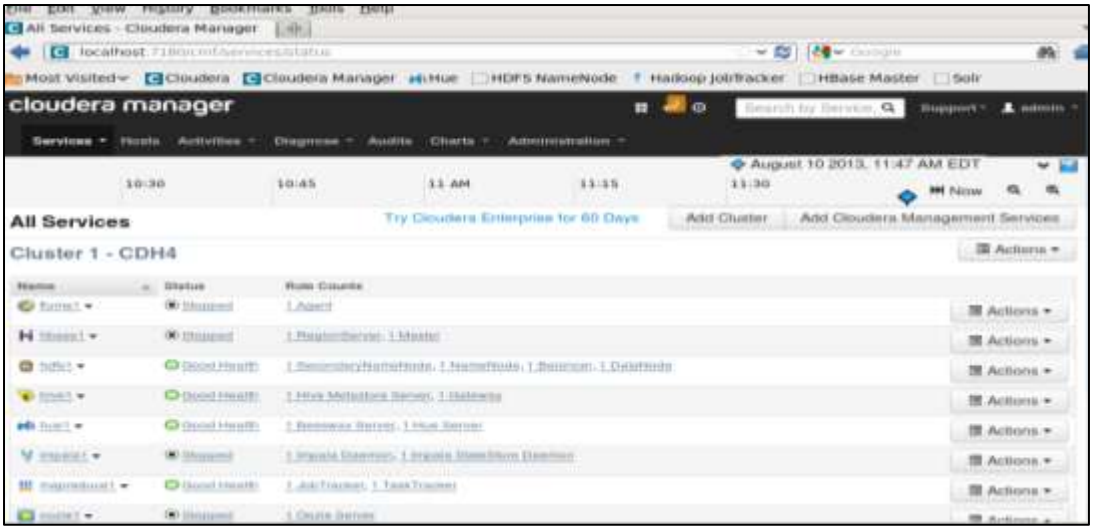

Figure 5.3 Installations of Services with CM Console

5.4 Discovering Residual Artifacts for Forensics Investigation on CDH

This section states the proposed forensic investigation framework for locating and discovering the residual artifacts that remain on the CDH Server and attached client devices. The residual artifacts can provide the potential evidences for forensic examiners to extract the evidences, and reconstruct the crime scene. The resulting residual artifacts can provide effective evidences to the forensic examiners for future CHD forensics.

5.4.1 Experimental Setup for Discovering Residual Artifacts

The experimental environment is set up for investigation of CDH by accessing various client devices. This investigation is intended to discover the residual artifacts on both server and client portion by accessing via popular web browsers. The testing environment and summary configurations of server and client devices are described in Table 5.1.

Table 5.1 System Configuration of CDH Platform for Testing Environment

(CDH Storage)Server Configuration	**Client Configurations**
Operation system - Cent OS 6 Virtual memory size - 2GB Virtual HD size - 16GB Cloudera Version - CDH 5.7.0 IP address/ URL - http://hostname/8888	Operation system - Windows 7 64 bit - Android 4.0.4 Ice Cream Sandwich (ICS) Virtual memory size - 2GB Virtual HD size - 16GB Browsers - Mozilla Firefox 33.0.2 - IE 9.10.9200.16384, - Google Chrome 38.0.2125.111 m - Android Default Browser - Dolphin Browser V-11.5.4 - Opera Browser 28.0.1764.90386

5.4.2 Residual Artifacts of CDH Server

When the file operations (upload, read, download) are processed, the residual artifacts on CDH Server are listed through Table 5.2 to 5.4. The residual artifacts for client devices are already described in section 4.3.4 of Chapter 4.

The remained artifacts for the file operation (Uploading) are shown in the Table 5.2. The artifacts are the source IP, the date of the operation, and the type of operation. These artifacts are located in the directories of hdfs/namenode/current/, var/log/Hadoop/hdfs and var/log/hue which are exposed in the column “Location” of the Table 5.2.

Table 5.2: Residual Artifacts of CDH Server (File Uploading)

Location	**File Names**	**Artifacts**	**Remarks**
/hdfs/namenode/ current/	fsimage	/user/IP/file.pdf	- file name

/var/log/hadoop -hdfs/	hdfs-audit.log	12017-6-20 00:18:48,allowed = true ugi=admin src = /home/file.pdf cmd=create	- Date - User name - File name - Operation (create)
/var/log/hue	access.log	120/Jul/2017 20:41:27 172.16.38.24 admin – POST /filebrowser//dirBb/file.pdf	- Date - Source IP - User name - Access method (POST for upload) - File Path - File name

The residual artifacts for download operation are shown in the Table 5.3. The artifacts are the source IP, the timestamp of the operation, and the operation name. The operation name for downloading is “open” and access method is “GET”.

Table 5.3: Residual Artifacts of CDH Server (File Downloading)

Location	**File Names**	**Artifacts**	**Remarks**
/hdfs/namenode/ current/	Fsimage	/user/IP/file.pdf	- file name
/var/log/hadoop -hdfs/	hdfs-audit.log	12017-6-20 00:18:48,allowed = true ugi=admin src = /home/file.pdf cmd=open	- Date - User name - File name - Operation (open)
/var/log/hue	access.log	120/Jul/2017 20:41:27 172.16.38.24 admin – GET /filebrowser//dirBb/file.pdf	- Date - Source IP - User name - Access method - File Path - File name

The artifacts for reading operation are the source IP, the timestamp of the operation, and the operation name. The operation name for downloading is “getfileinfo” and access method is “GET” as shown in Table 5.4.

Table 5.4: Residual Artifacts of CDH Server (File Reading)

Location	File Names	Artifacts	Remarks
/hdfs/namenode/ current/	Fsimage	/user/IP/file.pdf	- file name
/var/log/hadoop -hdfs/	hdfs-audit.log	12017-6-20 00:18:48,allowed = true ugi=admin src = /home/file.pdf cmd= getfileinfo	- Date - User name - File name - Operation (getfileinfo)
/var/log/hue	access.log	120/Jul/2017 20:41:27 172.16.38.24 admin – GET /filebrowser//dirBb/file.pdf	- Date - Source IP - User name - Access method - File Path - File name

5.5 Case Study: CDH Investigation

The following case study is extended from the exposed crime case of CYFOR which is a leading authority in digital evidence and has over a decade of expertise in digital forensics.

Background

The Company, DEF organization uses the services of CDH which is built on Centos 6. Every authorized person in this organization can access this through 'http://192.168.32.34/8888/' from their own web browsers. They use the CDH platform as the file server.

Case: Employee Data Theft Case Study

CYFOR; the investigating organization were contacted directly by a manufacturer through the recommendation of a solicitor. The business owners suspected that an employee had just leaving, stolen company's important data including business plan and new product's design in order to start up a competing business. The former employee marched lawsuit proceedings and took him to court. During the former employee's defense, he revealed information that he could not lawfully been aware of during his employment tenure. CYFOR were instructed by the client to expose the evidence that the former employee had committed the crime.

5.5.1 Forensic Investigation on CDH

This proposed framework finds out residual artifacts which remained after the file operations on CDH. The remained artifacts can provide the potential evidences for forensic examiners to extract the evidences, and reconstruct the crime scene. The conventional digital forensic methods are insufficient for investigating such composite infrastructure. Therefore, this chapter undertakes the forensic investigation on CDH by applying the proposed Hadoop Forensic Investigation Framework.

5.5.1.1 Scope and Identification of Forensic Investigation on CDH

The identified sources are

- CentOS 6 on which CDH 5 is installed
- Client PC of Ubuntu 14.04 LTS
- The copy of stolen files (business_plan.pdf)

5.5.1.2 Preparation and Collection of Forensic Investigation on CDH

The traditional forensic collection is performed as the bit-by-bit copy generated. However, indexing speeds decrease as the amount of data raise [24] which seems to point to an unavailability of this method to the current environment. Along with these, Hadoop forensics also presents the data size issue. In the case of a case in Hadoop clusters, the acquisition stage could consist the collection of Petabytes of information. It takes 28 days produce a bit by- bit copy of a Petabyte image considering the current speed transfer of 6 GB/s [24]. The Hadoop clusters generate the large amount of log files per operation. Thus, an SSH and telnet client tool; PUTTY [76] is used to connect and collect the forensic data from the server portion.

The documentation of residual artifacts resulting from the previous section is applied as the prerequisite analysis, to assist the current forensic investigation. By consuming the previous knowledge located the residual artifacts; the investigator could know the specific location of artifacts. So it can lead the cost effective way for forensic works.

5.5.1.3 Analysis of Forensic Investigation on CDH

This collected forensic data is analyzed to identify the usage and discover residual artifacts.

(a) Evidential Analysis on CDH Server

The content of 'hue-access.log' and 'hdfs-audit.log' files shows that 'business_plan.pdf' is downloaded by Mr.A as presented in Figure 5.4 and Figure 5.5.

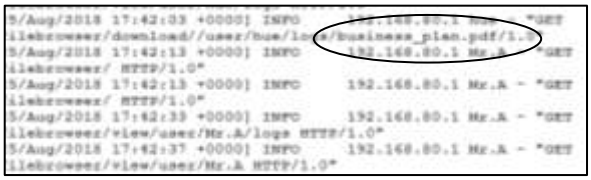

Figure 5.4 The 'hue-access.log' File of CDH Server

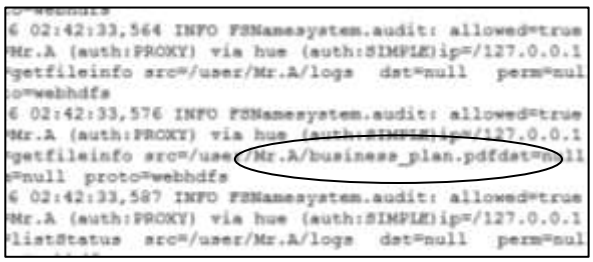

Figure 5.5 The 'hdfs-audit.log' File of CDH Server

(b) Evidential Analysis on Client Device

The command for locating history data of Mozilla firefox is

- ~/.mozilla/firefox/c1ryki12.default$ sqlite3 places.sqlite

The evidencential data is extracted by the following sqlite query and the resulting file is as shown in Figure 5.6.

- SELECT datetime(a.visit_date), b.url FROM moz_historyvisits AS a JOIN moz_places AS b ON a.place_id=b.id WHERE 1 ORDER BY a.visit_date ASC;

Figure 5.6 The "places.sqlite" File of Ubuntu 14.04 Client Machine (viewed by sqlite3)

5.5.1.4 Reporting for Forensic Investigation on CDH

The investigator arranges the finding evidences to embody the crime and reconstruct the criminal activity. The forensic report for the investigation of 'Employee Data Theft Case' on CDH Platform is presented in Table 5.5.

Table 5.5 Forensic Report for the Employee Data Theft Case

FORENSIC REPORT FOR CDH
INVESTIGATOR : Mr. Adam
Case Type : Require Evidences to prove the criminal activity
Case Number : 2
1. Status: Complete
2. Summary of finding: To find the related information of "business plan". Server Side Evaluation Step 1 : Finding the residual artifacts that log and metadata files on CDH Server Step 2 : Setting up forensic server Step 3 : Connecting the Server Step 4: Collecting the data • Client Side Evaluation Step 1 : Finding the Internet History of PC that are access CDH server via web browser Step 2: Extracting/parsing the history data by analysis tools Step 3 : Reconstruct the event

3. Items Analyzed:

TAG Number:	ITEM DESCRIPTION:
0100	CDH Server
0166	Personal Computer (PC), Serial# 123457

4. Finding for item 0100

iv. The specification of item 0100:

v. The examined Server was found to contain a Centos 6.7 OS

vi. Among the metadata and log file, the discovered residual artifacts in access.log and audit.log files :

a. Login time is starting from 23/Dec/2016 20:46:44

b. Open the "business_plan.pdf" at 23/Dec/2016 21:18:43

c. Download again the files "business_plan.pdf" at 23/Dec/2016 21:18:49.

5. Finding for item 1066

iii. The examined hard drive was found to contain Ubuntu 14.04 operating systems

iv. "business_plan.pdf"

7. Provided Items: Along with this hard copy, the report is also submitted with one CD for electronic copy report. The CD contains the same information with this hard copy.

As the result of the following of report Mr.A stole the "business plan" by using Mozilla Firefox web browser from his PC. The documentation in each phase is stored. The difficulties, solutions, usage of tools and all experiences of each step are reviewed for the preparation phase of the next investigations.

5.5.1.5 Closing of Forensic Investigation on CDH

The whole documentations are organized for later use. The collected data is stored in archived format. The forensic researcher reviews the tasks of each phase to

extract which factors should be notice for the next investigation. The difficulties, solutions, usage of tools and all experiences of each step are reviewed for the preparation phase of the next investigations.

5.6 Chapter Summary

This chapter describes the forensic investigation of a popular Hadoop Platform: CDH. The residual artifacts are extracted from the server and client portions of CDH and the forensic case study is investigated by applying the proposed forensic investigation framework for Hadoop Big Data Platform. This was the first experiment conducted for the forensic investigation on CDH. The forensically important files are exposed and the residual artifacts are extracted from the server the client portions of CDH. The attached client devices are PC and smart phones of Windows, Linux and Android Oss. Most of the forensically important files among the Hadoop backlogs are .log and .meta files. The important artifacts of client devices are located in the browser histories, caches, cookies, memory images, and registries. In order to extract the evidences on Android devices, the investigator should have the root permission. The private web browsing in Android cannot permit to extract the artifacts if this device is non-rooted.

As a case study for CDH forensic, a popular forensic case of "Employee Data Theft Case Study" provided by CYFOR is investigated by applying the proposed forensic investigation framework. The artifacts can help to extract the evidence which can prove the criminal activity in CYFOR crime case. The outcomes of this research proved to be beneficial for the real-word forensic investigation when information was located that identified the use of CDH to trace the illegal usages.

CHAPTER 6

THE FORENSIC INVESTIGATION ON MapR HADOOP PLATFORM

In this chapter, the focus is forensic investigation and discovering residual artifacts on MapR Hadoop Platform. The proposed framework is applied to investigate the crime scenarios and trace the illegal usages on MapR.

6.1 MapR Hadoop Platform

MapR is a complete enterprise-grade distribution for Apache Hadoop. The MapR Distribution for Apache Hadoop has been engineered to improve Hadoop's reliability, performance, and ease of use. The MapR distribution provides a full Hadoop stack that includes the MapR File System (MapR-FS), MapReduce, a complete Hadoop ecosystem, and the MapR Control System user interface. The MapR distribution provides several unique features that address common concerns with Apache Hadoop as shown in Table 6.1.

Table 6.1 Characteristics of MapR Hadoop Platform

Issue	Addressed by MapR Feature	Apache Hadoop
Data Protection	MapR Snapshots provide complete recovery capabilities. MapR Snapshots are rapid point-in-time consistent snapshots for both files and tables. MapR Snapshots make efficient use of storage and CPU resources, storing only changes from the point the snapshot is taken. MapR Snapshots can configure schedules with easy to use but powerful scheduling tools.	Snapshot-like capabilities are not consistent, require application changes to make consistent, and may lead to data loss in certain situations.
Security	With wire-level security, data transmissions to, from, and within the cluster are encrypted, and strong authorization mechanisms enable to tailor the actions a given user is able	Permissions for users are checked on file open only.

	to perform. Authentication is robust without burdening end-users. Permissions for users are checked on each file access.	
Disaster Recovery	MapR provides business continuity and disaster recovery services out of the box with mirroring that's simple to configure and makes efficient use of cluster's storage, CPU, and bandwidth resources.	No standard mirroring solution. Scripts based on distcp quickly become hard to administer and manage. No enterprise-grade consistency.
Enterprise Integration	With high-availability Direct Access NFS, data ingestion to cluster can be made as simple as mounting an NFS share to the data source. Support for Hadoop ecosystem projects like Flume or Sqoop means minimal disruptions to existing workflow.	
Performance	MapR uses customized units of I/O, chunking, resync, and administration. These architectural elements allow MapR clusters to run at speeds close to the maximum allowed by the underlying hardware. In addition, the Direct Shuffle technology leverages the performance advantages of MapR-FS to deliver strong cluster performance, and Direct Access NFS simplifies data ingestion and access. MapR-DB tables, available with the M7 license, are natively stored in the file system and support the Apache HBase API. MapR-DB tables provide the fastest and easiest to administer NoSQL solution on Hadoop.	Stock Apache Hadoop's NFS cannot read or write to an open file.

Scalable Architecture (without single points of failure)	The MapR Converged Data Platform provides High Availability for the Hadoop components in the stack. MapR clusters don't use Namenodes and provide stateful high-availability for the MapReduce JobTracker and Direct Access NFS. Works out of the box with no special configuration required.	Namenode HA provides failover, but no failback, while limiting scale and creating complex configuration challenges. Namenode federation adds new processes and parameters to provide cumbersome, error-prone file federation. The High-Availability JobTracker in stock Apache Hadoop does not preserve the state of running jobs. Failover for the JobTracker requires restarting all in-progress jobs and brings complex configuration requirements.

The MapR Hadoop permit to spool back to a documented upright data set. The version 5.1 of MapR Hadoop distribution offers certification, permission, and encryption facilities to shield the data in cluster. The MapR users can be certificated through Kerberos, LDAP/AD, NIS, or any other service. For permission, MapR provides Access Control Lists (ACLs) for job queues, volumes, and the cluster all together. MapR-FS executes permission checks on each file access as MapR provides POSIX permissions on files and directories. Other Hadoop platforms only check permissions on file open.

The encrypted data transmission for traffic within the MapR clusters integrate wire-level security (WLS), along with traffic between the server hosts and client hosts. MapR influences the Hadoop Fair Scheduler to guarantee fair allocation of resources to diverse operators, and contains provision of SELinux. The MapR File System routines volumes as a unique administration entity. A logical unit: volume builds to relate procedures to a set of files, directories, tables, and sub-volumes. The volumes can be created for each user, department, or project.

Volumes can administer disk usage restrictions, set replication levels, establish ownership and control acceptable actions, by different clusters. When policies are set on a volume, all files contained within the volume inherit the same policies set on the volume. MapR Hadoop distribution can manage volume authorizations through ACLs in the MapR Control System. The different level of permissions are set on a file or

directory for users by using standard UNIX commands, after that volume has been mounted through Network File System (NFS).

The MapR direct access file system allows real-time access data by means of the NFS protocol. By NFS, standard operations can directly access the MapR-FS while the traditional file I/O operations can access data in a conventional UNIX FS. A remote client can easily mount a MapR cluster over NFS to transmit data among the cluster. MapR Hadoop platform can support High Availability (HA) management data processing services for spontaneous stability all over the cluster. The MapR Control System (MCS), can implements REST API to monitor services for host and cluster level. MapReduce, management services and data access services such as NFS deliver incessant service through any system break down. MapR provides HA improvements in shuffle phase for all HDFS services. With MapR, Hadoop services are enable to configure for implementation on several nodes even failure.

6.1.1 MapR Control System

The MCS supports a graphical user interface for cluster management by means of all the operations of the command-line and REST APIs. The job monitoring metrics and troubleshooting are provided by MCS, such as which jobs requisite the maximum memory within a specific time or which procedures triggered work and task failures. The MCS Dashboard offers the cluster information, cluster utilization and the health of attached node in cluster, alerts an alarms, and displays a cluster heat map. The following Figure 6.1 shows the MCS Dashboard:

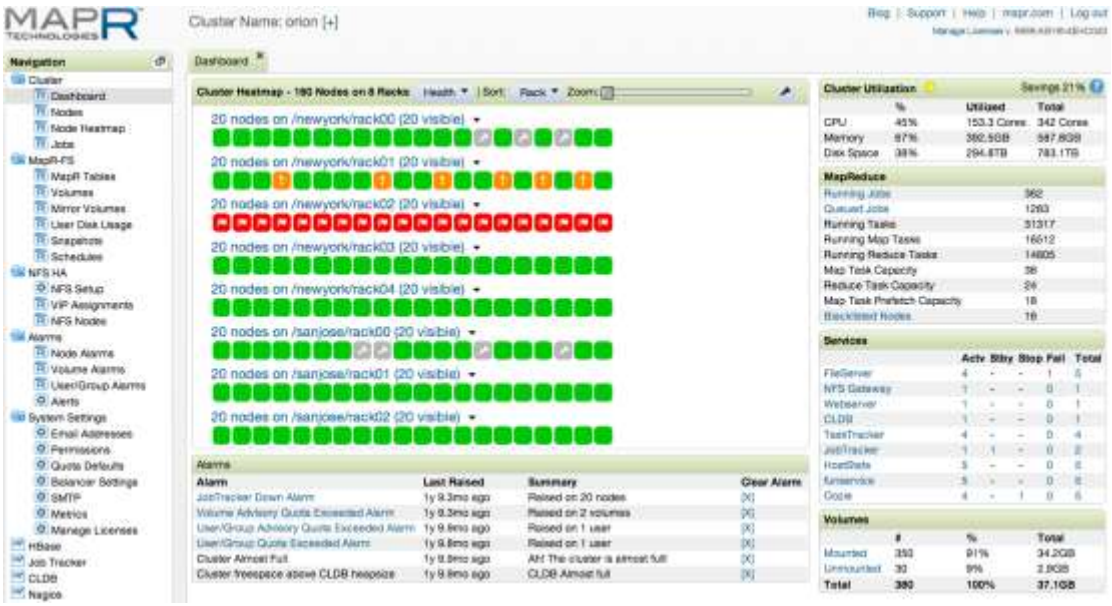

Figure 6.1 MCS Dashboard

6.2 Architecture of MapR Hadoop Platform

MapR is a comprehensive, industry-standard, Hadoop Platform with the improvement of the Hadoop FS abstraction interface and expands the HA of the distributed file system, eliminating the Namenode. MapR Hadoop Platform comprises the completed Hadoop ecosystem factiliy packages. MapR Hadoop enables real time executions; enlightening data load/unload procedures. The version 4.0.1 of MapR Hadoop Platform presents the Hadoop 2.x architecture. Hadoop 2.x and YARN constructs a resource scheduling and management framework.

This following Figure 6.2 shows the architecture of MapR Hadoop Platform. The MapR Hadoop Platform is the easiest, most dependable, and fastest Hadoop Platform. It is the one and only Hadoop Platform which contribute the MapR Direct Access NFS to permits direct data input and output for realtime analytics, with the improvement of HA. MapR-FS contains several storage pools existed in each host of the cluster. A storage pool which is default number three is prepared with numerous disks assembled by MapR-FS. The data containers are stored in and duplicated amongst the storage pools [69]. MapR has a more distributed approach for storing metadata on the processing nodes which is the dissimilarity with Cloudera and Hortonworks. MapR Hadoop Platform implements on a new file system known as MapR-FS. MapR Hadoop Platform eliminates Namenode architecture and does not rely on the Linux FS. The MapR refined the HDFS architecture to offer elasticity,

increase HA, and assist exceptional structures for data administration and performance.

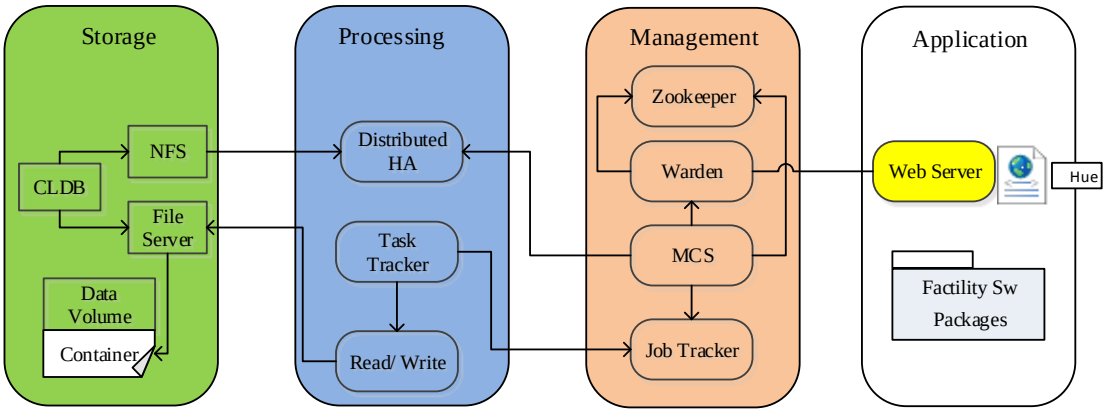

Figure 6.2 Architecture of MapR Hadoop Platform

The following Table 6.2 provides a list of six features of MapR-FS namely storage pools, containers, Container Location Database (CLDB), volumes, snapshots and direct access NFS.

Table 6.2 Features of MapR-FS

Feature	Description
Storage pools	A group of disks that MapR-FS writes data to.
Containers	An abstract entity that stores files and directories in MapR-FS. A container always belongs to exactly one volume and can hold namespace information, file chunks, or table chunks for the volume the container belongs to.
CLDB	A service that tracks the location of every container in MapR-FS.
Volumes	A management entity that stores and organizes containers in MapR-FS. Used to distribute metadata, set permissions on data in the cluster, and for data backup. A volume consists of a single name container and a number of data containers.
Snapshots	Read-only image of a volume at a specific point in time used to preserve access to deleted data.
Direct Access NFS	Enables applications to read and write data directly into the cluster.

6.3 Installation and Configuration of MapR Hadoop Platform

The following instructions are to install and configure the MapR Hadoop Platform.

Step 1. Prerequisites

Install the Java. (JDK and JRE)

Check the Java versions:

Create a text file /tmp/disks.txt listing disks and partitions for use by MapR.

Step 2. MapR Repository Supplement

Established internal repository.

Download and install the repo.

Step 3. Software Installation

The file /tmp/disks.txt to know the holding a list of disks and partitions is created before installation.

Access with the root level and a MapR cluster is set up by using the commands:

```
yum install mapr-single-node
/opt/mapr/server/disksetup -F /tmp/disks.txt
/etc/init.d/mapr-zookeeper start
/etc/init.d/mapr-warden start
/opt/mapr/bin/maprcli acl edit -type cluster -user <user>:fc
```

Step 4. MCS Checking

MCS is navigated in a browser on the MapR server with the URL

```
https://<host>:8443
Example:https://localhost:8443
```

The username and password are the administrative user in Step 3.

The following commands can be used to avert from starting automatically cluster.

```
chkconfig --del mapr-warden
/etc/init.d/mapr-zookeeper start
/etc/init.d/mapr-warden (as root or with sudo)
```

6.4 Discovering Residual Artifacts for Forensic Investigation on MapR

This section discusses locating and discovering the residual artifacts which endure on the MapR Server and client devices. The residual artifacts can provide the potential evidences for forensic examiners to extract the evidences, and reconstruct the crime scene.

6.4.1 Experimental Setup for Discovering Residual Artifacts

The summary configurations of server and client for testing environment are described in Table 6.3. The experimental setup for discovering the residual artifacts, to identify the use of MapR Hadoop Platform, creates 4 MapR Hadoop Server Virtual Machines (VMs) as shown in Figure 6.3. VMware Workstation 12 Player was used to construct the experimental VMs. For instance, on "Access VM", login to the Hadoop Server with the default user login name. With "Download VM", the sample data is downloaded to client devices. And then, the circumstances are compared with the Clean VM to estimate which artifacts are left while running particular operations.

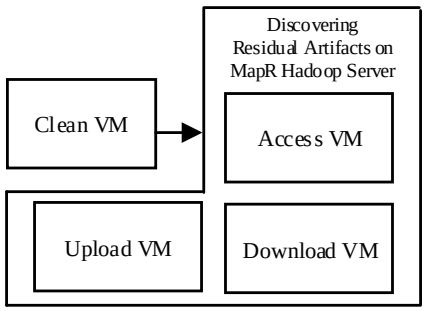

Figure 6.3 VMs Creation for Experiment

This section presents the findings, particularly residual artifacts discovered on MapR Hadoop Server to trace the diversity of operations. After performing the access, upload, and download operations, the snapshot of each VMs are taken to obtain Virtual Machine Memory (VMEM) and Virtual Machine Disk (VMDK) files available in each VM folder. This experiment mounts and analyzes the VMEM and

VMDK to the analysis machine using FTK Imager version 3.2.0 [4]. This analysis conducts keyword searches to locate MapR identifications, the sample data, and default login name. FileViwerPlus 6.4[3] is used to access and recover evidential data from metadata and log files in the experiments.

Table 6.3 System Configuration of MapR for Testing Environment

Specification	Value
Operating system	- Cent OS 7
Virtual memory size	- 8GB
Virtual HD size	- 20 GB
MapR Version	- MapR 6.0 - Hadoop 2.7.0
IP address/ URL	- https://hostname: 8443 - http://hostname:8888
Default User Name and Password	- mapR - mapR

6.4.2 Residual Artifacts on MapR Server

Firstly, this section tests the access (login) operation with "Access VM". This experiment can locate the artifacts on MapR Hadoop Server. The residual artifacts are shown in the following Figure 6.4, 6.5, 6.6 and 6.7.

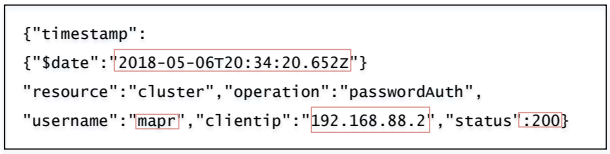

Figure 6.4 Artifacts for 'Login' in /opt/mapr/log/authaudit.json of MapR Server

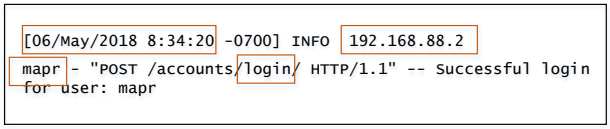

Figure 6.5 Artifacts for 'Login' in opt/mapr/hue/access.log MapR Server

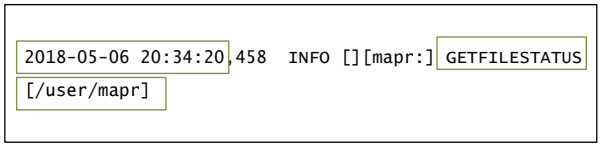

Figure 6.6 Artifacts for 'Login' in /opt/mapr/httpfs/httpfs-1.0/logs/httpfs-audit.log of MapR Server

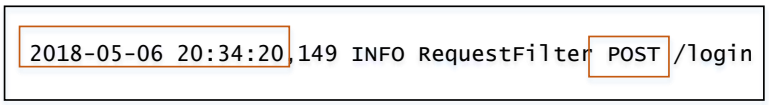

Figure 6.7 Artifacts for 'Login' in /opt/mapr/apiserver/logs/apiserver.log of MapR Server

The summary description of the extracted residual artifacts for login operation is described in the following Tables 6.4. The artifacts are discovered in the authaudit.json, access.log, httpfs-audit.log and apiserver.log. The authaudit.json contains information of every login name, client machine IP and timestamp that had accessed the MapR server. The access.log file includes the exact time of login, user name, client IP and operation. In httpfs-audit.log and apiserver.log, the timestamp and operation are discovered. The format of timestamp is UTC date/time. The extracted artifacts could present

- user login name
- client machine's IP
- accessed date and time.

Table 6.4 Residual Artifacts of MapR (Login)

Location	File Names	Artifacts	Remarks
/opt/mapr/log	authaudit.json	2018-05-06T20:34:20.65 passwordAuth mapr 192.168.88.2 200	- Date (UTC) - Operation (passwordAuth= login) - User name - IP address - Httpstatus (200 = OK)
/opt/mapr/ httpfs/httpfs-1.0/logs	hdfs-audit.log	Artifacts 2018-05-06 0:34:20 192.168.88.2	- Date (UTC) - Client IP - Operation

		GETFILESTATUS Mapr	(GETFILESTATUS) - User name
opt/mapr/hue/	access.log	106/May/2018 8:34:20 2192.168.88.2 3mapr login4	- Date - Source IP - User name - Access method - File Path - File name
/opt/mapr/ apiserver/logs	apiserver.log	2018-05-06 20:34:20 POST	- Date (UTC) - Client IP - Operation

The "Upload VM" is examined and analyzed for uploading operation. The artifacts and forensically important files are shown in the Figure 6.8, 6.9, and 6.10. The user name, operation (upload in this section) and time stamp can be discovered in access.log, httpfs-audit.log and apiserver.log. In httpfs-audit.log, the parameter for upload operation is 'getfilestatus'.

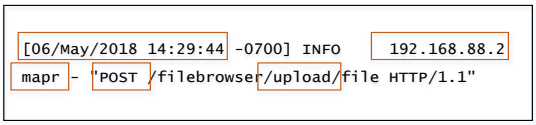

Figure 6.8 Artifacts for Upload in opt/mapr/hue/access.log of MapR Server

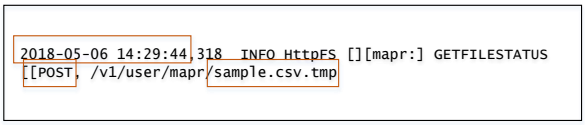

Figure 6.9 Artifacts for Upload in /opt/mapr/httpfs/httpfs-1.0/logs/httpfs-audit.log of MapR Server

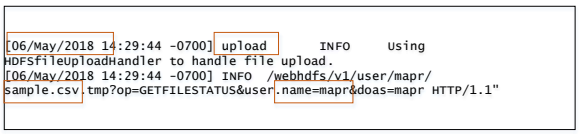

Figure 6.10 Artifacts for Upload in /opt/mapr/apiserver/logs/apiserver.log MapR Server

The summary description of the extracted residual artifacts of the uploading a file to MapR server is described in the following Table 6.5.

Table 6.5 Residual Artifacts of MapR (File Uploading)

Location	File Names	Artifacts	Remarks
/opt/mapr/ httpfs/httpfs-1.0/logs	hdfs-audit.log	Artifacts 2018-05-06 0:34:20 192.168.88.2 POST Mapr	- Date (UTC) - Client IP - Operation - User name
opt/mapr/hue/	access.log	106/May/2018 8:34:20 192.168.88.2 mapr POST, upload	- Date - Source IP - User name - Access method - File name, file path
/opt/mapr/ apiserver/logs	apiserver.log	2018-05-06 20:34:20 POST, upload	- Date (UTC) - Client IP - Operation

The file is downloaded from the MapR Hadoop Server by testing on "Download VM". The residual artifacts while performing the operation are shown as the following Figure 6.11, 6.12, 6.13 and 6.14.

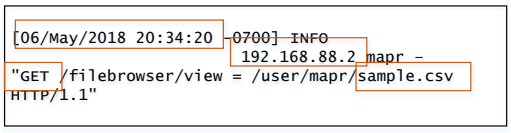

Figure 6.11 Artifacts for Download in opt/mapr/hue/access.log of MapR Server

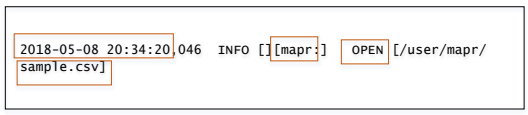

Figure 6.12 Artifacts for Download in /opt/mapr/httpfs/httpfs-1.0/logs/httpfs.log of MapR Server

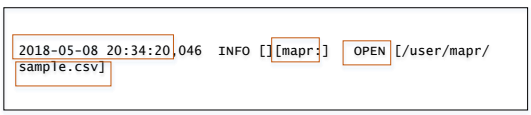

Figure 6.13 Artifacts for Download in /opt/mapr/httpfs/httpfs-1.0/logs/httpfs-audit.log of MapR Server

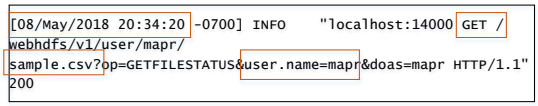

Figure 6.14 Artifacts for Download in /opt/mapr/hue/runcpserver.log of MapR Server

The summary description of the extracted residual artifacts from above figures is described in the following Table 6.6.

Table 6.6 Residual Artifacts of MapR (File Downloading)

Location	File Names	Artifacts	Remarks
/opt/mapr/ httpfs/httpfs-1.0/logs	hdfs-audit.log	Artifacts 2018-05-06 0:34:20 192.168.88.2 POST Mapr	- Date (UTC) - Client IP - Operation - User name
opt/mapr/hue/	access.log	106/May/2018 8:34:20 192.168.88.2 mapr POST, download	- Date - Source IP - User name - Access method
/opt/mapr/ apiserver/logs	apiserver.log	2018-05-06 20:34:20 POST, download	- Date (UTC) - Client IP - Operation

6.4.3 Residual Artifacts of Client Devices

This information of this section is already described in section 5.4.3 in chapter 5.

6.5 Case Study: MapR Hadoop Investigation

The circumstances of the case study are greatly simplified for the purposes of this case study;

- Every authorized person in this organization can access this through "http://192.168.32.34/8888/" from web browsers.

Case: Gossip Spreading

- Someone is spreading some rumor by posting the article (gossip) to speak ill of the organization
- The executive team of the organization wants to know who the criminal is.

6.5.1 Forensic Investigation on MapR

This section discusses locating and discovering the residual artifacts that remain on the MapR Server and attached client machines. As described in literature,

conventional digital forensic methods are insufficient for investigating such composite infrastructure. Therefore, this chapter undertakes the forensic investigation on MapR by applying the proposed Hadoop Forensic Investigation Framework. The conducted forensic investigation can provide effective evidences to the forensic examiners for future real-world CHD forensics.

6.5.1.1 Scope and Identification of Forensic Investigation on MapR

The identified sources are

- MapR Hadoop server which is installed on Centos 7
- PC with Windows 10 OS

6.5.1.2 Preparation and Collection of Forensic Investigation on MapR

The traditional forensic collection preforms as the bit-by-bit copy generated. However, indexing speeds decrease as the amount of data raise which seems to point to an unavailability of this method to the current environment. Along with these, Hadoop forensics also presents the data size issue. In the case of a case in Hadoop clusters, the acquisition stage could consist of the collection of petabytes of information. It takes time to produce a bit by- bit copy of a Petabyte image. The Hadoop clusters generate the large amount of log files per operation. Thus, an SSH and telnet client tool; PUTTY [76] is used to connect and collect the forensic data from the server portion. The forensic tools for investigation are prepared as shown in Table 6.7. The documentation of residual artifacts resulting from the previous section is applied as the prerequisite analysis, to assist the current forensic investigation. By consuming the previous knowledge located the residual artifacts; the investigator could know the specific location of artifacts. So it can lead the cost effective way for forensic works.

Table 6.7 Forensic Tools for MapR Investigation

Tool	Usage
FTK Imager Version 3.2.0.0 [9]	To create a forensic image of the .VMDK files.
dcfldd, dd version 1.3.4-1 [4]	To produce a bit-for-bit image of the .VMEM files.
Autopsy 3.1.1 [17]	To parse the file system, produce directory listings, as well as extracting/analyzing the files, Windows registry, swap file/partition, and unallocated space from the forensic images.
SQLite Browser Version 3.4.0 [18]	To view the contents of SQLite database.
Browser History Spy V-3.0 [19]	the all-in-one software to instantly recover or view the browsing history from popular web browsers
WebBrowserPassView v1.56 [20]	password recovery tool that reveals the passwords stored by the web browsers
File Viewer Plus 2 [8]	View, edit, save, and convert over Hex files

6.5.1.3 Analysis of Forensic Investigation on MapR

This collected forensic data is analyzed to identify the usage and discover residual artifacts.

(a) Evidential Analysis on MapR Server

In order to analyze the collected data, variety of forensic analysis and recover tools are tested to apply. As shown in Figure 6.15, the residual artifacts found in client devices are URL, date, time, file name.

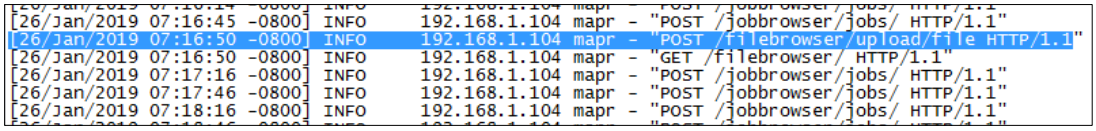

Figure 6.15 The 'hue-access.log' File of MapR Server

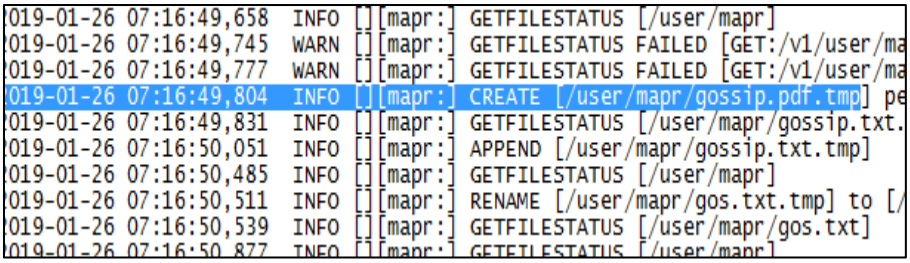

Figure 6.16 The 'hdfs-audit.log' File of MapR Server

The Figure 6.16 illustrates the browser log file, in that file, the accessed web URL, website title, visited date and time can be identified. The web address of server machine is also found in browser cache entries file of client machine as shown in Figure 6.17. Because of the hex file, it is viewed by the hex file reader; FileViewerPlus [41].

Figure 6.17 The 'runcp-server.log' File of MapR Server

(b) Evidential Analysis on Client Machine

For extracting the artifacts on client machine of Windows 10 PC, the collected disk images and memory images of client machine are mounted on the forensic machine. The 'browser spy' tool can be used to explore the thumbnails, cache and history of all type of Windows browsers. The following Figure 6.18 shows that the 'gossip.pdf' file was uploaded from the targeted client machine via the Mozilla browser.

Figure 6.18 Browser Log File of Windows 10 Client Machine (Viewed by 'browser spy' Tool)

6.5.1.4 Reporting for Forensic Investigation on MapR Hadoop Platform

Conferring to the experiment, the residual artifacts were remained when user makes file operations on MapR. The important files, paths, artifacts of storage server and attached client machine are documented. This information supports the investigaters in conducting the forensic work and will assist to embody the criminal activity. The forensic report for the investigation of 'Rumor Spreading Case' on MapR Hadoop Platform is presented in Table 6.8.

Table 6.8 Forensic Report for the Rumor Spreading Case

FORENSIC REPORT FOR MapR
INVESTIGATOR : Mr. Adam
Case Type : Finding the Criminals
Case Number : 3
1. Status: Complete
2. Summary of finding: To find the related information of "Rumor Spreading Case". Server Side Evaluation Step 1 : Finding the residual artifacts that log and metadata files on MapR Server Step 2 : Setting up forensic server Step 3 : Connecting the Server Step 4: Collecting the data • Client Side Evaluation Step 1 : Finding the Internet History of PC that are access MapR server via web browser

Step 2: Extracting/parsing the history data by analysis tools

Step 3 : Reconstruct the event

3. Items Analyzed:

TAG Number:	ITEM DESCRIPTION:
0100	MapR Server
0166	Personal Computer (PC), Serial# 123457

4. Finding for item 0100

The specification of item 0100:

The examined Server was found to contain a Centos 6.7 OS

Among the metadata and log file, the discovered residual artifacts in audit.log and edits-inprogress_00000000000083.hex :

d. Login time is starting from 23/Dec/2016 20:46:44
e. Open the “business plan” at 23/Dec/2016 21:18:43
f. Download again the file “/tmp/ gossip.pdf” at 23/Dec/2016 21:18:49.

5. Finding for item 1066

The examined hard drive was found to contain windows 10 64 bits OS

“gossip.pdf” and date are found in internet history of Mozilla Firefox at “rnllugnp.default-1456638777411\history”

When using browsing history tools, user Mr. Felix access the http://172.16.38.32:8080 at 23/Dec/2016 starting from 20:46:44

7. Provided Items: Along with this hard copy, the report is also submitted with one CD for electronic copy report. The CD contains the same information with this hard copy.

6.5.1.5 Closing of Forensic Investigation on MapR Hadoop Platform

The whole documentations are organized for later use. The collected data is stored in archived format. The forensic researcher reviews the tasks of each phase to extract which factors should be notice for the next investigation. The difficulties,

solutions, usage of tools and all experiences of each step are reviewed for the preparation phase of the next investigations.

When the Putty is connected to the server to collect the data, unlike the other platforms, one notable point for the next investigation is that the file permission is needed to be upgraded for the MapR before the investigation.

6.6 Chapter Summary

This chapter describes the MapR Hadoop Platform, architecture of MapR and installation methods. This was the first experiment conducted for the forensic investigation on MapR. The Forensic Investigation Framework is applied to be assistance for all aspects of an investigation, from the beginning of outlining the scope to endure on-track, with the facility to go back to a previous step. Furthermore, in closing phase review was important to ensure any new processes or information was documented and reported, and ensured the data used throughout the process was stored appropriately. It is important to develop and use a consistent digital forensic framework, such as the one examined in this research, to ensure investigations are thorough, all issues are encompassed, and the process is flexible enough to apply in different situations. The outcomes of this research proved to be beneficial for the real-word forensic work when the case of relating the use of MapR Hadoop Platform to trace the illegal usages. The details of this investigation research are important to highpoint the real-world forensics in intensifying the information and methodology for forensic practices.

CHAPTER 7

CONCLUSION AND FURTHER EXTENSIONS

In the age of Big Data, Hadoop is an entirely open source system for handling Big Data. As the Hadoop Big Data Platform is quickly mature, more illegal usages are occurring. For this reason the more forensic investigation researches on Hadoop Big Data Platform should be conducted. The traditional forensic process models are not fit for this environment. There is a need for an investigation framework to guide forensic work where the Hadoop Big Data Platform is involved. In addition, without knowing where artifacts may reside can take the considerable amount of time for forensic analysis. This research proposes a forensic investigation framework to guide the forensic work for Hadoop Big Data Platform. Moreover, as the proactive work before forensic work, this research discovers the residual artifacts that are remained from the usage on Hadoop Big Data Platform. Finally, the forensic works are conducted to investigate the popular crime scenarios by applying the proposed forensic investigation framework.

This research discovers the residual artifacts on the Hadoop server portion and client devices of popular Hadoop platforms. It is undertaken by investigating on the popular Hadoop Big Data Platforms which are residing on different OS. The usage of Hadoop can be identified due to the remaining data which is discovered on the Hadoop servers and also on client machines.

The proposed forensic investigation framework guides the investigation from the scope to the closing phase. The residual artifacts are extracted over a variety of circumstances including a sequence of file operations; access, upload, and download the data on the Hadoop Platform. This research could demonstrate that Hadoop Platform can provide the significant number of useful artifacts for forensic practitioners. The analysis of Hadoop Server states that the potential information sources of artifacts includes core log files, ecosystem log, and metadata files. The residual artifacts of Hadoop Big Data Platform could assist the forensic examiners in generating evidence to conduct the Big Data platform forensics.

As outlined, there is a variety of investigation results for an examiner to determine the use of Hadoop Platform, such as; directory listings, prefect files, registry, browser history, and memory images. By determining the residual artifacts, this research provides a better considerations of the location of artifacts, and the

access point for forensics examiners to assist an investigation. The usage of the proposed Forensic Investigation Framework for Hadoop Platform is also applicable to guide the forensic work relating to emerging technology of Big Data, and could be helpful in a real world forensics from commencement to completion.

7.1 Dissertation Summary

This section has summarized the research work in previous chapters in order to tie up the overall dissertation book for better understanding of the proposed system.

In Chapter 1, Big Data, and Big Data platform, are introduced and briefly discussed Hadoop Big Data Platform. After that, it presents the Hadoop Big Data Platform, digital forensics, and braches of forensics. Then, motivational factors that lead to the necessity of the research work are explained. And then, the research questions are described and the methodologies are also exposed to answer the above questions. Finally, accomplished objectives which are the significant part of the research are presented.

Chapter 2 reviews the works and efforts of previous researchers of forensic methodology for innovative technologies; cloud and Big Data. The literature is reviewed in four portions: cloud computing and digital forensics, Big Data and digital forensics, Hadoop forensics and forensic investigation frameworks.

Chapter 3 explained the proposed Forensic Investigation Framework for Hadoop Big Data Platform and how this can be applied to forensic analysis of Hadoop Big Data Platform. Each step of the framework was explained; Scope and Identification, Preparation and Collection, Analysis, Reporting and Closing. The cyclic nature of the framework was then outlined.

Chapter 4 shows the forensic investigation on Hortonworks Data Platform: Ambari HDP and Non-Ambari HDP. The detailed architecture is described and the installation and configuration of HDP is also presented. The residual artifacts for both server and client portion of HDP are discovered and exposed. Finally, the HDP forensic is conduct to investigate the real-world case study.

Chapter 5 shows the forensic investigation on Cloudera Distribution of Hadoop. The detailed architecture is described and the installation and configuration of CDH is also presented. The residual artifacts for both server and client portion of

CDH are discovered and exposed. Finally, the CDH forensic is conducted to investigate the real-world case study.

Chapter 6 has shown the forensic investigation on Map Hadoop platform. The detailed architecture is described and the installation and configuration of MapR is also presented. The residual artifacts for both server and client portion of MapR are discovered and exposed. Finally, the MapR forensic is conduct to investigate the real-world case study.

7.2 Results and Discussion

The focus of this research is to discover whether there are any residual artifacts on popular Hadoop Big Data Platforms. Through this research work, if any operation occurred on Hadoop platform, the residual artifacts remain on both server and client portions. The usages of Hadoop platform can be traced by extracting the residual artifacts either from Hadoop server or client machines.

When conducting the current research, it had determined that there was a need to have a methodology to guide the investigation, and hence a process flow of the investigation framework was proposed and considered using the practical work to observe the benefits of using this methodology when undertaking Hadoop forensics. This was built upon the guidelines for digital forensic analysis to provide for a process which follows a common digital forensic examination, enabling forensic examiners to apply the methodology and expenditure the artifacts to real-world investigations, and also serve to research the residual artifacts using a real-world process of analysis.

By applying the proposed Forensic Investigation Framework for Big Data Platform, the most popular Hadoop Big Data Platforms are investigated and potential evidences are located. The forensically important files are exposed and the residual artifacts are extracted from the server the client portions of HDP, CDH and MapR Hadoop Platforms. The attached client devices are PC and smart phones of Windows, Linux and Android OS. Most of the forensically important files among the Hadoop backlogs are .log and .meta files. The types of artifacts of HDP and CDH are similar but the MapR has a little different types of artifact as MapR architecture is a quite different from the others. The artifacts of each client devices are same for all three popular Hadoop Platforms. The important artifacts of client devices are located in the browser histories, caches, cookies, memory images, and registries. In order to extract

the evidences on Android devices, the investigator should have the root permission. The private web browsing in Android cannot permit to extract the artifacts if this device is non-rooted.

The real word crime cases of Document Ex-Filtering Case, Employee Data Theft Case and Gossip Spreading Case are investigated to conduct the forensic investigation on popular Hadoop Platforms by applying the proposed forensic investigation framework. The discovered residual artifacts as well as proposed forensic investigation framework will guide practitioners, examiners, and researchers when undertaking forensic analysis of Hadoop Platform. The residual artifacts are able to provide as the potential evidence to assist forensic practitioners in generating evidences.

7.3 Future Research

This research has proposed the forensic investigation on the popular Hadoop Big Data Platforms; Hortonworks Data Platform, Cloudera Distribution of Hadoop and MapR Hadoop Platform for extracting the residual artifacts from both server portion and attached client devices with different OS (Ubuntu, Windows 7, Windows 10 and Android).

For the further extension, the future research will be conducted as the followings:

- forensic investigation and discovering residual artifacts on other Big Data Platforms such as Gluster, GFS, etc..
- forensic investigation on other client devices; IPhone, mobile devices etc..
- extension of the proposed framework for other Big Data Platforms
- forensic investigation of the crime such as child abduction, money laundering etc.. by analyzing the data set on BDP

ACRONYM

API	Application Programming Interfaces
AWS	Amazon Web Service
CDH	Cloudera Distribution of Hadoop
Centos	Community Enterprise Operating System
CLI	Command-line Interface
CSP	Cloud Service Provider
CSV	Comma Separated Value
DFS	Distributed File System
HBDP	Hadoop Big Data Platform
HDP	Hadoop Data Platform
EC2	Elastic Compute Cloud
GFS	Google File System
Gluster	Gluster File System
HDFS	Hadoop Distributed File System
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
SaaS	Software as a Service
IaaS	Infrastructure as a Service
iOS	iPhone OS
MapR-FS	MapR File System
MapR	MapR Hadoop Platform
NFS	Network File System
NIST	National Institute of Standards and Technology
PVFS	Parallel Virtual File System

RHEL	Red Hat Enterprise Linux
SIFT	SANS Investigative Forensic Toolkit
UI	User Interface
VM	Virtual Machine
VMDK	Virtual Machine Disk

AUTHOR'S PUBLICATIONS

[p1] M.N.Oo and T.Thein, "Forensic Investigation on Hadoop Hortonworks Data Platform", In Proceeding of the 15th International Conference on Computer Application (ICCA) Feb 16-17, 2017.

[p2] M.N.Oo and T.Thein, "Forensic Readiness on Hadoop Platform: Non-Ambari HDP as a Case Study", International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 9. Sept 2017, pp.193-203.

[p3] M.N.Oo and T.Thein, "Forensic Analysis of Residual Artifacts on CDH Storage", In Proceeding of the 1st International Conference on Advance Information Technology (ICAIT) Nov 1-2, 2018, pp. 31-38.

[p4] M.N.Oo and T.Thein, "Forensic Investigation through Residual artifacts on Hadoop Big Data Storage System", International Journal of Computer Systems Science and Engineering (IJCSSE) Vol. 33, Jan, 2018.

[p5] M.N.Oo and T.Thein, "Investigation of Android Device for Discovering Hadoop Cloud Storage Artifacts", In Proceeding of the 16th International Conference on Computer Application (ICCA) Feb 22-23, 2018, pp.50-57.

[p6] M.N.Oo and T.Thein, "Forensic Investigation on MapR Hadoop Platform", IEEE International Conference on Knowledge Innovation and Invention 2018 (ICKII 2018) July 23-27, 2018.

BIBLIOGRAPHY

[1] A. Adedayo, M. Oluwasola. "Big Data and Digital Forensics." In the Proceeding of IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), 12-14, June 2016, Vancouver, BC, Canada, pp. 1-7, doi: 10.1109/ICCCF.2016.7740422.

[2] "Amazon EC2 - Virtual Server Hosting", [Online] Available at: https://aws.amazon.com/ec2/. [Accessed: 8/11/2016].

[3] "Apache Hadoop", [Online] Available at: https://archive.apache.org/dist/Hadoop/core/. [Accessed 15/10/ 2016].

[4] A. Alex, M. Edington, and R. Kishore. "Forensics framework for cloud computing", Published in the Journal of Computers & Electrical Engineering, Volume 60, pp.193-205, 2017.

[5] M. Almorsy, J.Grundy, and A.Ibrahim, "Supporting automated vulnerability analysis using formalized vulnerability signatures", In Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering, pp.100-109, 2012.

[6] Almulla, Sameera, Youssef Iraqi, and Andrew Jones. "A state-of-the-art review of cloud forensics", Published in the Journal of Journal of Digital Forensics, Security and Law, Volume 9, Issue 2, 2014.

[7] Alqahtany, Saad, N.Clarke, S.Furnell, and C.Reich, "A forensic acquisition and analysis system for IaaS." Published in the Journal of Cluster Computing Volume 19, Issue 1, pp. 439-453, 2016.

[8] Arora, Minit, and Himanshu Bahuguna. "Big Data Security–The Big Challenge." Published in the Journal of Scientific & Engineering Research, Volume 7, Issue 12, December 2016.

[9] "Automatic Installation with Ambari", [Online] https:// teradata.github.io/presto/docs/141t/installation/installation-ambari.html. [Accessed: 8/11/2016].

[10] Baboo, Capt Dr S. Santhosh, and S. Mani Megalai. "Cyber Forensic Investigation and Exploration on Cloud Computing Environment." Published in the Global Journal of Computer Science and Technology, 2015.

[11] Bariki, Hamda, Mariam Hashmi, and Ibrahim Baggili. "Defining a standard

for reporting digital evidence items in computer forensic tools." In Proceedings of the International Conference on Digital Forensics and Cyber Crime. Springer, Berlin, Heidelberg, 2010.

[12] Basu, Subhajit. "Cloud Crimes: Understanding the Privacy Challenges " [Online] https://works. bepress. com/subhajitbasu/94/, [Accessed: 8/11/2016].

[13] N.Beebe, J.Clark, "A Hierarchical, Objectives-Based Framework for the Digital Investigations Process," in Proceedings of the Digital Forensic Research Conference, Baltimore, MD., August 2004.

[14] "Big Data revenue worldwide from 2016 to 2027 by major segment (in billion U.S.dollars)", [Online] Available at: https://www.statista.com/statistics/254266/global-big-data-market-forecast/, [Accessed: 8/11/2016].

[15] H. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big Data A Fashionable Topic with (out) Sustainable Relevance for Research and Practice?", Business & Information Systems Engineering, Volume 5, Issue 2, 2013, pp.65-69.

[16] D. Birk, (2011), 'Technical Challenges of Forensic Investigations in Cloud Computing Environments', paper presented at the Workshop on Cryptography and Security in Clouds, IBM Forum Switzerland, Zurich.

[17] Biggs, Stephen, and Stilianos Vidalis. "Cloud computing: The impact on digital forensic investigations." Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for. IEEE, 2009

[18] Computing, Cloud. "Business Benefits With Security, Governance and Assurance Perspectives. White Paper." Information Systems Audit and Control Association, 2009.

[19] H. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big Data", [Online] Available at: https://link.springer.com/article/10.1007/s12599-013-0249-5, [Accessed: 8/11/2016].

[20] "Case Studies", https://cyfor.co.uk/

[21] Casey, Eoghan, and Aaron Stanley. "Tool review–remote forensic preservation and examination tools." Published in the Journal of Digital Investigation: Volume 1, Issue 4, pp. 284-297, 2014.

[22] Casey, Eoghan. "Digital evidence in the courtroom." Digital Evidence and Computer Crime: Forensic Science, Computer and the Internet, Academic Press, May 2011.

[23] D.Conroy, "Forensic Data Analysis Challenges in Large Scale Systems." Publish in the Journal of Intelligent Distributed Computing Springer, Volume 9, pp. 451-457, 2016.

[24] V. Chemitiganti, "What is Apache Hadoop?" in Business Values of Hadoop, Hortonworks, 2016. [Online] Available at: http://hortonworks.com/apache/hadoop. [Accessed: 8/11/2016].

[25] C.H.Cho , "Cyber Forensic for Hadoop Based Cloud System," Published in the International Journal of Security and its Applications, Volume 6, Issue 3 , pp.83-90 , July, 2012

[26] Chung, Hyunji, et al. "Digital forensic investigation of cloud storage services." Digital investigation Volume 9, Issue 2, pp. 81-95, 2012.

[27] H.Chung et al., "Digital Forensic Investigation of Cloud Storage Services" Digital investigation vol. 9, no.2, pp. 81-95, Nov. 30, 2012.

[28] "Cloud Vision 2020: The Future of the Cloud Study, Logic Monitor" , [Online] https://www.logicmonitor.com/resource/the-future-of-the-cloud-a-cloud-influencers-survey/

[29] "Cloud Forensics: Box", [Online] Available at: https://cyberforensicator.com/2018/04/21/cloud-forensics-box/. [Accessed: 8/11/2016].

[30] C. Coles, "AWS vs Azure vs Google cloud market share 2017" [Online] Available at: https://www.skyhighnetworks.com/cloud-security-blog/microsoft-azure-closesiaas-adoption-gap-with-amazon-aws/. [Accessed: 9/12/2017]

[31] "Collectin information", [Online] Available at: https://ambari.apache.org/1.2.2/installing-hadoop-using-ambari/content/ambari-chap1-4.html. [Accessed: 8/11/2016].

[32] Degree Hub, "Computer Science and Digital Forensics", [Online] https://www.computersciencedegreehub.com/faq/what-is-digital-forensics/

[33] "DFRWS, Research Road Map", [Online] Utica, NY, 2001, http://www.dfrws.org/2001/.

[34] “Digital Forensic Analysis of Amazon Linux EC2 Instances”, [Online] Available at: https://dash.harvard.edu/handle/1/24829568. [Accessed: 8/11/2016].

[35] Douglas A Orr and Peter White, “Current State of Forensic Acquisition for IaaS Cloud Services,” Published in the Journal of Forensic Science & Criminal Investigation, Volume 10 Issue 1, 2018.

[36] J. Dykstra and A. T. Sherman, “Acquiring Forensic Evidence form Infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques." Published in the Journal of Digital Investigation, Volume 9, 2012.

[37] B. Dominik, C.Wegener, “Technical issues of forensic investigations in cloud computing environments”, In the Proceeding of IEEE Sixth International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE), 2011.

[38] “Document Exfiltration Case of Digital Copera: M57 Case” [Online] Available at: https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario/. [Accessed /12/1/2018]

[39] “Facebook has the world's largest Hadoop cluster!”, [Online] https://hadoopblog.blogspot.com/2010/05/facebook-has-worlds-largest-hadoop.html.

[40] Feng, Xiaohua, and Yuping Zhao. "Digital forensics challenges to Big Data in the cloud." In the Proceeding of 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 21-23 June 2017, Exeter, UK, doi: 10.1109/ 2017.132.

[41] “FileViewerPlus”, [Online] Available: http://fileviewerplus.com.siterankd.com/. [Accessed 8/8/ 2016].

[42] Fowler K. “Hadoop forensics: Tackling The Elephant In The Room”, [Online] Available at: http://2012.video.sector.ca/video/51118145. [Accessed /12/1/2018]

[43] Garfinkel, Simson L. "Digital Forensics Research: The Next 10 Years." Published in the Journal of Digital Investigation, Volume 7, 2010.

[44] Gartner, “IT Glossary: Big Data” [Online].Available at: http://www.gartner.com/it-glossary/big-data/. [Accessed /12/1/2018]

[45] Gartner, “AWS Named as a Leader in Gartner’s Infrastructure as a Service (IaaS) Magic Quadrant for 7th Consecutive Year”, https://aws.amazon.com/blogs/aws/aws-named-as-a-leader-in-gartners-infrastructure-as-a-service-iaas-magic-quadrant-for-7th-consecutive-year, Accessed: January 2018.

[46] G. Grispos, T. Storer and W. B. Glisson, "Calm Befor the Storm: The Challenges of Cloud Computing in Digital Forensics," Published in the International Journal of Digital Crime , Volume 4, Issue 2, pp. 28-48, 2012.

[47] Gupta, Aman, and Pranita Jain. "A Map Reduce Hadoop Implementation Of Random Tree Algorithm Based On Correlation Feature Selection", Published in the International Journal of Computer Application, Volume 160, pp. 41-44, 2017.

[48] M. Gualtieri and N. Yuhanna, "The Forrester Wave: Big Data Hadoop Solutions, Q1 2014," Forrester, 2014.

[49] Gudu, Diana, Marcus Hardt, and Achim Streit. "Evaluating The Performance And Scalability of The Ceph Distributed Storage System." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014.

[50] “Hadoop Big Data,” [Online] https://www.gartner.com/it-glossary/big-data/

[51] Hortonwork Inc. “Apache Hadoop Basics”, [Online] Available at: http://hortonworks.com, [Accessed] Nov, 2016.

[52] Hortonworks Data Platform, “Non-Ambari Cluster Installation Guide”, [Online] https://docs.hortonworks.com/HDPDocuments/Ambari-2.6.2.0/bk_ambari-installation.

[53] InfoSec Institude, “Computer Forensics: Areas of Study”, [Online] https://resources.infosecinstitute.com/category/computerforensics/introduction/areas-of-study

[54] “Install the HDFS using Ambari” [Online] https://docs.hortonworks.com/HDPDocuments/HDF3/HDF-3.1.2/bk_installing-hdf/content/ch_install-hdf.html

[55] Irons, Alastair, and Harjinder Singh Lallie, "Digital Forensics To Intelligent Forensics", Published in the Journal of Future Internet, Volume 63, pp.

584-596, 2014.

[56] John Walker, Saint. "Big Data: A revolution that will transform how we live, work, and think." Published in the International Journal of Advertising, Volume 33 Issue 1, pp. 181-183, January 2014, doi: 10.2501/IJA-33-1-181-183.

[57] Kessel, P. V. "Big Data changing the way businesses compete and operate." Insights on governance, risk and compliance, EY Report, 2014.

[58] Kumari, Noble, and A. K. Mohapatra. "An insight into digital forensics branches and tools." In the Proceeding of the 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT),New Delhi, India, 18 July, 2016.

[59] Kim, Jang-Hee "The Characteristics Of Incidental Pituitary Microadenomas In 120 Korean Forensic Autopsy Cases", Published in the Journal of Korean Medical Science, Volume 27, 2007.

[60] Kindervag, John, "Control And Protect Sensitive Information In the Era of Big Data." For Security & Risk Professionals, Technical Report, 2012.

[61] K. Kent, et al., "Guide to Integrating Forensic Techniques into Incident Response," Special Publication 800-86, Computer Security Division Information Technology Laboratory National Institute of Standards and Technology, Gaithersburg, Maryland, 2006.

[62] Kolthof, Daan. "Crime in the cloud: An analysis of the use of cloud services for cybercrime." Student Conference on IT, Enschede, The Netherlands, 2015.

[63] W. G. Kruse and J. G. Heiser, "Computer Forensics – Incident Response Essentials", Pearson Education, September 26, 2001.

[64] Lallie, Harjinder S. and Pimlott, Lee, "Applying the ACPO Principles in Public Cloud Forensic Investigations," Published in the Journal of Digital Forensics, Security and Law, Volume 7, Issue 1, Article 5, 2012 doi:https://doi.org/10.15394/jdfsl.2012.1113. [Online] Available at: https://commons.erau.edu/jdfsl/vol7/iss1/5

[65] Larry, D., and D. Lars. "Digital Forensics for Legal Professionals." Syngress, 16 Sept 2011.

[66] Li, Shuyu, et al. "A sticky policy framework for Big Data security." In the

Proceeding of IEEE First International Conference on Big Data Computing Service and Applications (BigDataService), 2015 IEEE, Redwood City, CA, USA, 30 March-2 April 2015.

[67] Liu C, Singhal A, Wijesekera D, “Identifying Evidence For Implementing A Cloud Forensic Analysis Framework”, In the proceeding of IFIP International Conference Digital Forensics, Orlando, Florida, 2018.

[68] R. McKemmish, “What Is Forensic Computing?”, Published in the Journal of Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology, Volume 118, pp. 1-6, 1999.

[69] MapR Inc, “MapR Hadoop,” [Online] Available at: http://www.mapr.com

[70] Martini, Ben, and Kim-Kwang Raymond Choo. "Cloud Storage Forensics: Owncloud As A Case Study." Published in the Journal of Digital Investigation, Volume 10 Issue 4, pp.287-299, 2013.

[71] Martini, Ben, and Kim-Kwang Raymond Choo. "An Integrated Conceptual Digital Forensic Framework for Cloud Computing." Published in the Digital Investigation, Volume 9, Issue 2, pp. 71-802012.

[72] Mell, Peter, and Tim Grance. "The NIST Definition Of Cloud Computing", Published in the Journal of National Institute of Standards And Technology Volume 53, Issue 6, 2009.

[73] Nelson, Bill, Amelia Phillips, and Christopher Steuart. "Computer Forensics And Investigations As A Profession", Guide to Computer Forensics and Investigations, Fourth Edition, Boston, Course Technology, May 2010.

[74] NIJ, Electronic Crime Scene Investigation: “A Guide for First Responders”, 2008. [Online] Available at: http://www.nij.gov/pubs-sum/219941.htm.

[75] G. Plamer “A Road Map for Digital Forensic Research” The MITRE Corporation, Tech. Rep. No. DTR - T001-01, Nov.6, 2001.

[76] “PUTTY”, [Online] Available at: https://www.putty.org/ [Accessed: 10/2/2017]

[77] Quick, Darren, and Kim-Kwang Raymond Choo. "Big Digital Forensic Data: Data Reduction Framework and Selective Imaging”. Published in the Journal of Springer, 2018.

[78] D. Quick, “Cloud Storage Forensic Analysis,” M.S.thesis, School of Computer & Information Science, University of South Australia, Adelaide

SA, 2012.

[79] D. Quick and K.-K. R. Choo, “Google Drive: Forensic Analysis of Data Remnants,” Published in the Journal of Network Computing Application, Volume 40, pp. 179–193, 2014.

[80] D. Quick and K.-K. R. Choo, “Digital Droplets: Microsoft SkyDrive Forensic Residual artifacts,” Published in the Journal of Future Generation of Computing System, Volume 29, Issue 6, pp. 1378–1394, 2013.

[81] D. Quick and K.-K. R. Choo, “Dropbox Analysis: Residual artifacts on User Machines,” Published in the Journal of Digital Investigation, Volume 10, Issue 1, pp. 3–18, 2013.

[82] D. Quick and K.-K. R. Cho, "Big Forensic Data Management In Heterogeneous Distributed Systems: Quick Analysis Of Multimedia Forensic Data," Published in the Journal of Software: Practice and Experience, Volume 47 Issue 8, pp.1095-1109, 2017.

[83] Rahim, Nordiana, et al. "Digital Forensics: An Overview of the Current Trends.", Published in the International Journal of Cryptology Research Volume 4 Issue 2, 2014.

[84] J. Ratcliffe, “Intelligence-Led Policing”, Published in the Journal of Trends and Issues in Crime and Criminal Justice, Volume 248, pp. 1-6, 2003.

[85] “Recova”, [Online] Available at: http://www.filehippo.com. [Accessed: 10/1/2017]

[86] Report of Hadoop Big Data Distribution, [Online] Available at: "https:// www.forrester.com/ report/The+Forrester+Wave+Big+Data+Hadoop+Distributions+Q1+2017. [Accessed: 10/1/2017]

[87] “RightScale 2018 State of the Cloud Report”, [Online] Available at: https://www.rightscale.com/press-releases/rightscale-2018-state-of-the-cloud-report. [Accessed: 10/1/2017]

[88] B.Roshan, “General Architecture of Google File System”, [Online] Available at: http://programming-project.blogspot.com/2014/04/general-architecture-of-google-file.html [Accessed: 10/1/2017]

[89] Roussev, Vassil. "Digital Forensic Science: Issues, Methods, And Challenges." Published in the Synthesis Lectures on Information Security, Privacy, & Trust, Volume 8, Issue 5, pp.1-155, 2016.

[90] Ruan, Keyun, et al. "Cloud forensics." In the Proceeding of the IFIP International Conference on Digital Forensics. Springer, Berlin, Heidelberg, 2011.

[91] Ruan K , Carthy J , Kechadi T , Crosbie M “Cloud Forensics”, Published in the Journal of IFIP Advances In Information And Communication Technology Advances In Digital Forensics, Volume 361, pp. 35–46, January 2011.

[92] SAI, Hb 171-2003 Guidelines for the Management of It Evidence, Standards Australia, Sydney, Australia, 2003.

[93] Shvachko, Konstantin, et al. "The Hadoop Distributed File System." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. Ieee, 2010.

[94] Simou, and Stavros. "A Survey On Cloud Forensics Challenges And Solutions." Published in the Journal of Security and Communication Networks, Volume 9, Issue 18, pp. 6285-6314, 2016.

[95] Simou, Stavros, et al. "Cloud Forensics Solutions: A Review", Published in the Lecture Notes in Business Information Processing, Volume 178. Springer, Cham, 2014.

[96] Singh, Shalini, and Meena Sharma. "The Prototype For Implementation of Security Issue In Big Data Application Using Hadoop Server." Published in International Journal of Computer Applications, Volume 145 Issue 13, 2016.

[97] J.Sremack, “Big Data Forensics–Learning Hadoop Investigations,” Birmingham, UK: Packt Publishing Ltd, Aug. 2015.

[98] Spyridopoulos, Theodoros, and Vasilios Katos. "Requirements for A Forensically Ready Cloud Storage Service." Published in the International Journal of Digital Crime and Forensics (IJDCF), Volume 3, Issue 3, pp. 19-36, 2011.

[99] Subashini, Subashini, and Veeraruna Kavitha. "A Survey On Security Issues In Service Delivery Models Of Cloud Computing", Published in the

Journal of network and computer applications, Volume 34, Issue 1, pp.1-11, 2011.

[100] A. Tanner and D. Dampier, “An Approach for Managing Knowledge in Digital Forensics Examinations”, Published in the Journal of Computer Science Security, Volume 4, Issue 5, 2010.

[101] M. Taylor, J.Haggerty, D. Gresty and R. Hegarty, “Digital Evidence in Cloud Computing Systems,” Published in the Journal of Elsevier- Computer Law and Security Review, Volume 26, pp.304-308, 2010.

[102] Y.Y.Teing, A.Dehghantan, K.K.R.Choo, Z.Muda, M.T.Abdullah and W.C.Chai, “A, Closer Look at Syncany Windows and Ubuntu Clients’ Residual Artifacts ”, Published in the Security, Pravicy and Anonymity in Computation, Communication and Storage, Springer International Publishing, pp.342-357, Zhangjiajie, China, , 16-18 November 2016.

APPENDIX A

CONFIGURATION AND SOFTWARE SETUP

The forensic analysis of Big Data Hadoop Platform on commodity is performed on Linux VMs. The VMs are interconnected via a 1-gigabit Ethernet. The host machine runs Windows 7 and has Intel Core i7-3.40GHz processor, 4GB physical memory, and 950-GB disk. As software components, Hadoop 2.7 is used. The following is the installation and configuration of Hadoop cluster.

Installing Java on Ubuntu 14.04

Hadoop is a framework written in Java for running applications on large clusters of commodity hardware so Hadoop requires a working Java.The usr/lib/jvm is the default installation location of the Java JDK and the Java JRE. Enter the following command in the console to create this folder, if it does not already exist:

- sudo mkdir -p /usr/lib/jvm

(The –p option ensures that all folders in the mkdir path are created.)

- sudo mv jdk-7u21-linux-i586.tar.gz /usr/lib/jvm
- sudo mv jre-7u21-linux-i586.tar.gz /usr/lib/jvm

Navigate to the “installation folder”.

- cd /usr/lib/jvm

Unpack the tarball archives.

- sudo tar zxvf jdk-7u21-linux-i586.tar.gz
- sudo tar zxvf jre-7u21-linux-i586.tar.gz
- sudo rm jdk-7u21-linux-i586.tar.gz
- sudo rm jre-7u21-linux-i586.tar.gz

Display the contents of the installation folder.

- ls -l

Response:

jdk1.7.0_21

jre1.7.0_21

Inform Ubuntu where Java installation is located.

- sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.7.0_21/bin/javac" 1

- sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jre1.7.0_21/bin/java" 1

Inform Ubuntu that this is default Java installation.

- sudo update-alternatives --set "javac" "/usr/lib/jvm/jdk1.7.0_21/bin/javac"
- sudo update-alternatives --set "java" "/usr/lib/jvm/jre1.7.0_21/bin/java"

Update system-wide PATH.

- sudo echo "JAVA_HOME=/usr/lib/jvm/jdk1.7.0_21" >> /etc/profile
- sudo echo "PATH=$PATH:$JAVA_HOME/bin" >> /etc/profile
- sudo echo "export JAVA_HOME" >> /etc/profile
- sudo echo "export PATH" >> /etc/profile

Reload system-wide PATH.

- . /etc/profile

Test the new installation.

- java –version

java version "1.7.0_21"
Java(TM) SE Runtime Environment (build 1.7.0_21-b11)
Java HotSpot(TM) 64-Bit Server VM (build 23.21-b01, mixed mode)

- javac –version

javac 1.7.0_21

Installing and Configuring SSH

Hadoop control scripts rely on SSH to perform cluster-wide operations. It requires SSH access to manage its nodes.

- apt-get install ssh
- which ssh
- which sshd
- which ssh-keygen

- root@ubuntu:/home/hadoop# pwd

Response:

/home/hadoop

Generate an RSA key pair

- root@ubuntu:/home/hadoop# ssh-keygen –t rsa –P “”

Copy the public key (~/.ssh/id_rsa.pub) content and append to the file ~/.ssh/authorized_keys

- hadoop@ubuntu:~# cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

Try ssh on localhost

- root@ubuntu:/home/hadoop# ssh localhost

In hadoop, the master’s public key should be added to all the slaves’ ~/.ssh/authorized_keys file, so that master can easily communicate to all the slaves.Master public key file is id_rsa.pub and open it with text editor and copy the contents.

- root@slave:/home/hadoop# nano ~/.ssh/authorized_keys

Paste the contents of id_rsa.pub into it.

- root@master:/home/hadoop# ssh slave

Installing Hadoop

- root@ubuntu:/home/hadoop# pwd

Response:

/home/hadoop

Extract Hadoop tar file and change the owner and mode of folder.

- root@ubuntu:/home/hadoop# tar –xvzf hadoop-1.1.2.tar.gz
- root@ubuntu:/home/hadoop# chown –R 777 hadoop-1.1.2
- root@ubuntu:/home/hadoop# chmod –R 777 hadoop-1.1.2

Set JAVA_HOME in /home/hadoop/hadoop-1.1.2/conf/hadoop-env.sh

- export JAVA_HOME=/usr/lib/jvm/jdk1.7.0_21

Set JAVA_HOME and HADOOP_HOME as environment variable

- root@ubuntu:/# cd /home/hadoop/hadoop-1.1.2
- root@ubuntu:/home/hadoop/hadoop-1.1.2# nano ~/.bashrc
- export JAVA_HOME=/usr/lib/jvm/jdk1.7.0_21
- export HADOOP_HOME=/home/hadoop/hadoop-1.1.2
- export PATH=$PATH:$HADOOP_HOME/bin
- root@ubuntu:/home/hadoop/hadoop-1.1.2# bin/hadoop version

Response:

Hadoop 1.1.2

Create a base directory (/var/opt/hadoop/cluster) for hadoop to store dfs and mapreduce data.

- root@ubuntu:/# cd /var/opt
- root@ubuntu:/# mkdir hadoop
- root@ubuntu:/# cd hadoop
- root@ubuntu:/# mkdir cluster

Creating MapReduce+HDFS cluster

server1 (VM1)	192.168.43.53	NameNode, Secondary NameNode, JobTracker, DataNode, TaskTracker

server2 (VM2) 192.168.43.199 DataNode, TaskTracker

server3 (VM3) 192.168.43.51 DataNode, TaskTracker

To change the name of virtual machine

root@ubuntu:/#nano /etc/host/hostname

delete localhost

change master (or) slave1 (or) slave2

In MapReduce+HDFScluster, server1 serves as master and slave, and the other servers serve as slaves. ssh all slaves from the master

root@ubuntu:/# ssh localhost

root@ubuntu:/# ssh slave1

root@ubuntu:/# ssh slave2

Edit config file /home/hadoop/hadoop-1.1.2/conf/masters (only in master e.g server1 that run NameNode)

This file is used for Secondary NameNode

- root@ubuntu:/# gedit /home/hadoop/hadoop-1.1.2/conf/masters

localhost

Edit config file /home/hadoop/hadoop-1.1.2/conf/slaves (only in master e.g server1 that run NameNode and JobTracker)

- root@ubuntu:/# gedit /home/hadoop/hadoop-1.1.2/conf/slaves

master

slave1

slave2

Edit config file /home/hadoop/hadoop-1.1.2/conf/core-site.xml (all machines)

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://192.168.43.53:9000</value>

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>/var/opt/hadoop/cluster</value>

</property>

</configuration>

Edit config file /home/hadoop/hadoop-1.1.2/conf/mapred-site.xml (all machines)

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

<name>mapred.job.tracker</name>

<value>192.168.43.199:8021</value>

</property>

</configuration>

Edit config file /home/hadoop/hadoop-1.1.2/conf/hdfs-site.xml (all machines)

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

</configuration>

Formatting the HDFS

- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/hadoop namenode –format

To start the HDFS and MapReduce daemons,

- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/start-dfs.sh
- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/start-mapred.sh

To stop the HDFS and MapReduce daemons,

- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/stop-dfs.sh
- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/stop-mapred.sh

(or)

To start all the HDFS and MapReduce daemons,

- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/start-all.sh

To stop all the HDFS and MapReduce daemons,

- root@ubuntu:/# /home/hadoop/hadoop-1.1.2/bin/stop-all.sh

A nifty tool for checking whether the expected Hadoop processes are running is jps

(Master Machine)

- root@master:/# jps

Response:

2287 TaskTracker

2149 JobTracker

1938 DataNode

2085 Secondary NameNode

2349 jps

1788 NameNode

(Slave1 Machine)

- root@slave1:/# jps

Response:

2287 TaskTracker

1938 DataNode

2349 jps

(Slave2 Machine)

- root@slave2:/# jps

Response:

2287 TaskTracker

1938 DataNode

2349 jps

To run a sample MapReduce application

- root@master:/# /home/hadoop/hadoop-1.1.2/bin/hadoop fs –put /home/hadoop/test.txt test.txt
- root@master:/# /home/hadoop/hadoop-1.1.2/bin/hadoop jar hadoop-1.1.2-examples.jar wordcount test.txt output.