Comput Syst Sci & Eng (2018) 3: 203–217

International Journal of
Computer Systems
Science & Engineering

Forensic Investigation through Data Remnants on Hadoop Big Data Storage System

Myat Nandar Oo[1], **Sazia Parvin**[2], **Thandar Thein**[3]

[1] *University of Computer Studies, Yangon*
[2] *University of New South Wales, Canberra*
[3] *University of Computer Studies, Maubin*
E-mail: myatnandaroo@gmail.com, saziap@gmail.com, thandartheinn@gmail.com

Forensic examiners are in an uninterrupted battle with criminals in the use of Big Data technology. The underlying storage system is the main scene to trace the criminal activities. Big Data Storage System is identified as an emerging challenge to digital forensics. Thus, it requires the development of a sound methodology to investigate Big Data Storage System. Since the use of Hadoop as Big Data Storage System continues to grow rapidly, investigation process model for forensic analysis on Hadoop Storage and attached client devices is compulsory. Moreover, forensic analysis on Hadoop Big Data Storage System may take additional time without knowing where the data remnants can reside. In this paper, a new forensic investigation process model for Hadoop Big Data Storage System is proposed and discovered data remnants are presented. By conducting forensic research on Hadoop Big Data Storage System, the resulting data remnants assist the forensics examiners and practitioners for generating the evidences

Keywords: Big Data Storage System, Criminal Activity, Data Remnants, Forensic Investigation, Hadoop Distributed File System

1. INTRODUCTION

The current era witnessed massive increases in data due to the increased human dependency on automated systems as well as computers. This extra generated huge data is known as, Big Data? accompanying large dataset which is characterized by velocity, volume, and variety of data. Big Data is considered one of the greatest technologies for the digital revolution of the past few centuries [18]. 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. A wide variety of applications rely on Distributed File System (DFS) based storage systems to store, process and analyze 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. 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. But HDFS is an open source and many Big Data Storage Systems adopt it.

The Hadoop version 0.1.0 is published in April, 2006 and continues to increase its versions [14]. 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 [8]. In Statista report [21], the Hadoop market was valued at 6 billion U.S. dollars worldwide in the year 2015. Hadoop Big Data Storage System can be identified as a challenge to digital forensic researchers. 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 [13]. Among them, Hortonworks is the fully open source distribution. Many organizations are doing business in the Big Data world and the criminals also find the

ways to utilize it illegally. There is a need for forensic capabilities which support investigations of crime in Big Data System. It may take time if the forensic investigation of Big Data Storage System is conducted without knowing where data remnants may reside. Forensic process models are needed to assist in the investigation of Big Data Store Systems. In this paper, a forensic investigation process model for Hadoop Big Data Storage System is proposed and it is applied to investigate two types of Hadoop infrastructure: Hadoop 2.7.1 on Ubuntu 14.04 and Hadoop Data Platform (HDP) 2.3 on Red Hat Linux hosted on Amazon EC2. This paper focuses on discovering the data remnants not only on the Hadoop Big Data Storage server but also on client machines which access to server with the aim to assist the forensic examiners for generating the effective evidences. An overview was provided in the context of forensic process models and Hadoop. This paper is organized as follows: Section 2 examines current literatures focusing on digital forensic process models and digital forensics on Hadoop Big Data System. The Section 3 describes the architecture of Hadoop, MapReduce and YARN. Furthermore, the overview of the Hadoop and the architecture of Hadoop Hortonwrok Data Platform on Red Hat Linux hosted on Amazon EC2 are also presented. In section 4, the forensically issues and the research methodology of Hadoop Big Data Storage System are presented. In addition, the proposed forensic investigation process model for Big Data Storage System is introduced. Implementation and investigation of Hadoop servers and client machines are presented in section 5 and 6; respectively. Section 7 summarizes the overall paper and the method used to answer research questions is described. Areas for future work are then highlighted.

2. LITERATURE REVIEWS

The related works of Hadoop forensic Investigation of various aspects are discussed in this section. The following literature reviews explore the procedures and approaches used by other researchers in this particular field.

2.1 Digital Forensic Process Models

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. In 2001, forensic academia held large-scale consortiums and defined a general standard digital investigation process model [20]. This model contains six stages of 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 [15]. This model includes the phases of collection, examination, analysis and reporting. The relevant data are identified, labeled and record in the collection phase and the collected data are accessed and extracted in examination phase. And then the results of the examination are analyzed to drive the useful information. Quick [24] 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 data remnants 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 data remnants 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. Cho et al. [9] 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.

2.1.1 Discussion

The forensic process models presented in [15, 10] are standard and common procedures. The model [24] is a specific model focusing on cloud storage and digital forensic investigation. The paper [9] proposed a forensic procedure for Hadoop based cloud system. 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 which are running on traditional systems have limitations on supporting forensic investigation. While addressing the active nature of this environment, the forensic investigation process model should fulfill with the following characteristics:

- the iterative nature to easily change between each phases
- the forensic data collection and analysis without system suspension
- the proactive preparation of the investigation facilities
- the integrity of the investigation
- the background knowledge of which are forensically important parts and files
- the good documentation to learn the previous lessons

The traditional process models are limited to cope with the above issues. The sound forensic process models are required. The traditional forensic analysis procedures have to be altered with the rise of Hadoop. This paper examines Big Data Storage forensic investigation process model.

2.2 Forensic Investigation of Hadoop

The evolution of Hadoop Big Data Storage System brings the challenges to forensic investigation. Forensic investigation of Hadoop Big Data Storage System is the novel field; hence there are limited publications on this area. The following section discusses the literature review of related work. A specific type of data leakage, namely Data spillage, occurs when classified or sensitive information is moved onto an unauthorized or undesignated compute node or memory media. Sensitive information spillage from the Hadoop cluster may cause information unauthorized nodes access problem. In order to remedy such data spillage challenge, Alabi et al. [2] developed the forensic framework for collecting provenance data and investigating data spillage in Hadoop. The authors aimed to provide developing tools and prevention mechanisms by analyzing data spillage events in the context of Hadoop environments and MapReduce applications. The system level metadata was utilized to map the data motion in terms of how data are stored; who, where, and how data are accessed; and how data change throughout its life cycle in a Hadoop cluster. In the paper [29], the authors discussed the Hadoop Big Data system could give to new difficulties and challenges to forensic investigators. This paper highlighted that the understanding Hadoop

internal structure is the important point for forensic investigators. They pointed out that the use of different tools and technology can do the forensics of big data. And then they demonstrated that the automated tool (Autopsy) can help finding the evidences on big data efficiently.

2.2.1 Discussion

The paper [2] investigates and protects the Data spillage from Hadoop cluster. The paper [29] highlighted that automated tools can perform forensic of big data efficiently. This paper focuses on investigation of Hadoop Big Data Storage System by analyzing data remnants on Hadoop storage server and client machines.

3. ARCHITECTURE OF HADOOP

Understanding the Hadoop internal structure is the important point for forensic investigators. This section presents the overview of Hadoop Big Data system and architecture of HDP 2.3 on Red Hat Linux on Amazon EC2.

3.1 Hadoop Big Data System

The HDFS and MapReduce are the main Hadoop modules. HDFS allocates the files across the cluster to offer fault tolerant access and high-throughput. For distributed data processing, MapReduce is considered and efficient programming model. 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 64 MB or 128 MB as defined by system administrator [3]. 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. Hadoop offers a MapReduce framework for applications writing for large amounts of structured/semistructured data processing across large clusters of machines in a consistent and fault tolerant way. It uses a MapReduce implementation engine for fault-tolerant distributed computing system along the large stored datasets in the cluster's DFS. This MapReduce technique has been popularized by the fact that Google uses this technique on its clusters and licensed to Apache. In the separate Map and Reduce steps, each step is performed in parallel, where each operates on sets of key-value pairs. Therefore, program execution is divided into a Map and a Reduce phases, divided by data transfer between nodes in the cluster. A node completes a Map function in the first step on a section of the input data. The Map output is a set of records in the form of key-value pairs, stored on that node. The records for a key are aggregated at the node to run the Reducer for that key. This includes data transfer between machines. The second

Reduce step is congested until the Map step data is transferred to the suitable machine. The Reduce step generates another set of key-value pairs for final output. This programming model is controlled to the use of key-value pairs. However, a surprising number of tasks will be adequate for this framework. The Hadoop architecture is changed from Hadoop 1.x to Hadoop 2.x. YARN (Yet Another Resource Negotiator) is a new module added in the Hadoop 2.x. It is employed for Cluster Resource Management. Figure 1 shows the architecture of Hadoop version 2.

Figure 1 illustrates the layers of Hadoop 2.x architecture: storage layer HDFS and processing layer YARN. MapReduce 2 is a distributed application type that run MapReduce framework on top of YARN. The Resource Manager manages resources and allots the resources to the application. The Resource Manager has Scheduler and Application Manager components. The Scheduler executes the scheduling function using the client applications? resource requirements. The application Manager employs to accept job-submissions, exchanging-container to execute the specific Application Master and provides the service for restarting the Application Master container on failure. The Application Master has the responsibility of negotiating suitable resource containers from the Scheduler, tracking their status and monitoring. For launching containers, the Node Manager is engaged, where each can house a map or reduce task.

3.2 Hadoop HDP 2.3 on Red Hat Linux on Amazon EC2

Hortonworks distribution provides Hadoop system based on Apache Hadoop to analyses, sort and manage Big Data. Hortonworks is the simply commercial vendor that allocate complete open source Apache Hadoop without additional proprietary software. Hortonworks is easier learning curve to provide IT friendly tools for users.

According to Gartner [16], 2014 Iaas Magic Quadrant, Amazon Web Service is the irresistible market share leader, with more than 5 times of the compute capacity in the use than the combined total of the other 14 providers. Amazon Web Services provides cloud computing services to build, secure, and organize Big Data applications. In order to meet the companies requirements for

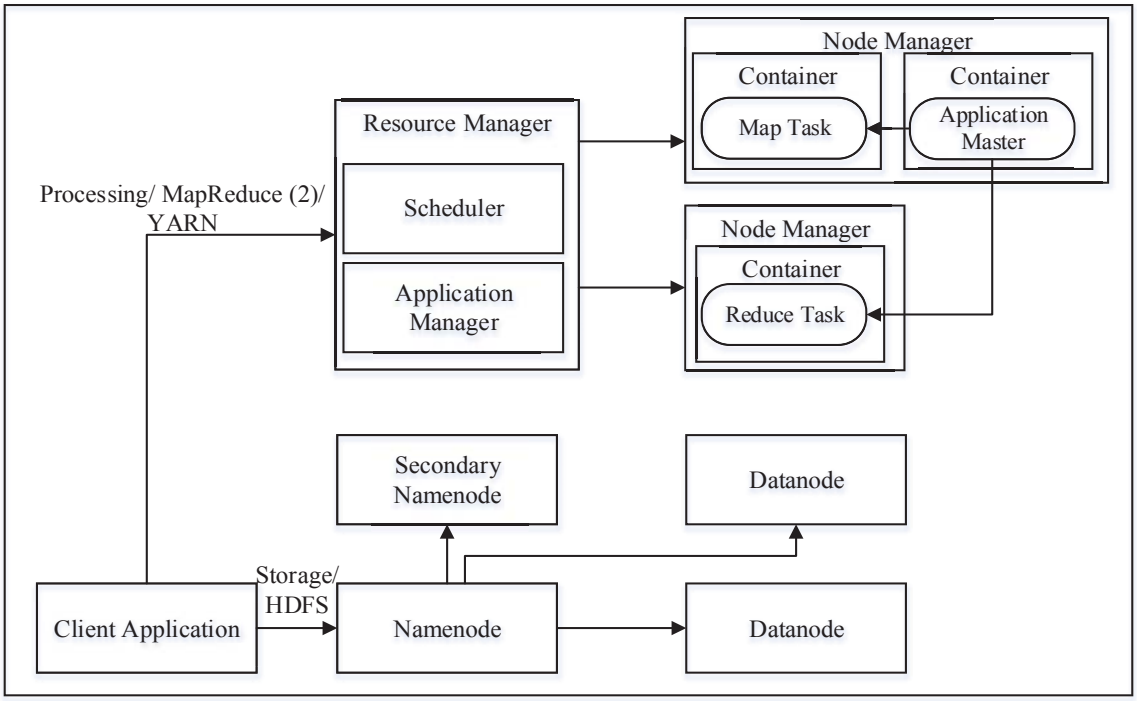

Figure 1 Hadoop 2.x architecture.

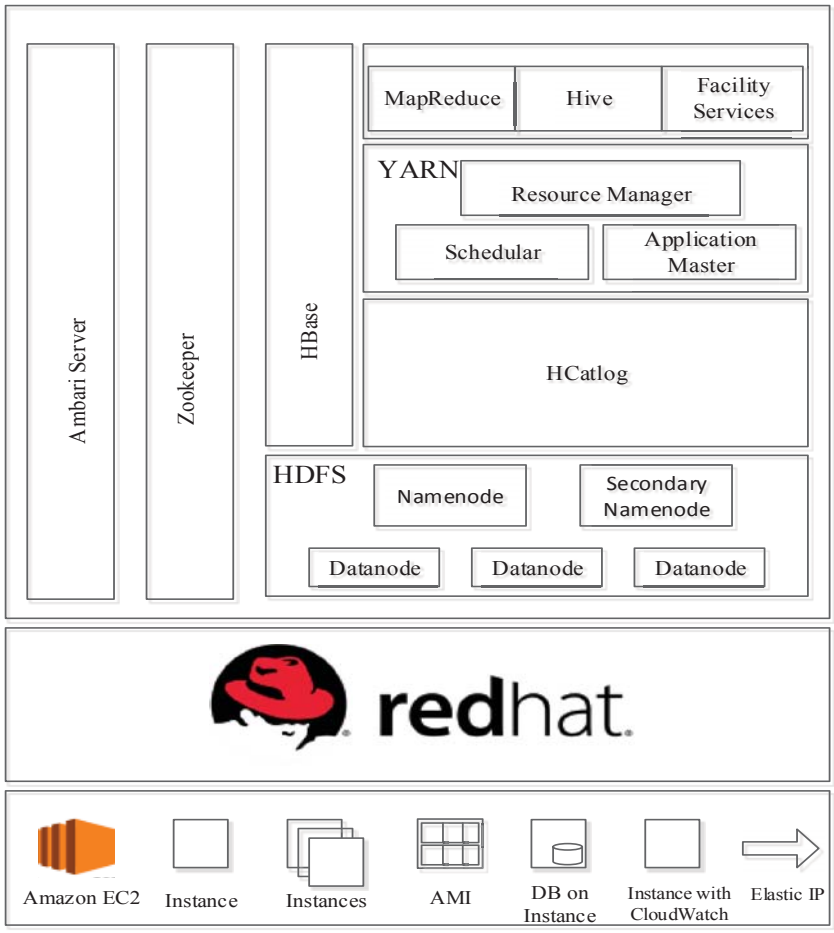

Figure 2 HDP 2.3 on Red Hat Linux7 server hosted on Amazon EC2.

intensive treatment of very large data volumes, Amazon Web Services offers Elastic Compute Cloud infrastructure (EC2) intending to help the Big Data Storage System through additional computing power. The effective usage of HDP 2.3 on Red Hat

Linux7 server hosted on Amazon EC motivates us to investigate on this environment.

Figure 2 presents the HDP 2.3 Red Hat Linux7 server hosted on Amazon EC2. Each instance on Amazon EC2 is a virtual server in the cloud. An Amazon Machine Image (AMI) offers the compulsory information to launch an instance. The instances can be monitored using Amazon CloudWatch that 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 Linux7 server is deployed as the instance of EC2 and Hadoop HDP is installed.

4. HADOOP BIG DATA STORAGE FORENSIC INVESTIGATION RESEARCH QUESTIONS AND METHODOLOGY

In this section, the research methodology for this paper is discussed, and then a proposed forensic investigation process model is outlined, which is applied in investigation to Hadoop Big Data Storage server and attached client machines.

4.1 Hadoop Big Data Storage Forensic Investigation Research Questions

This research focuses on the investigation of Hadoop 2.7.1 on Ubuntu 14.04 LTS and HDP 2.3 on Red Hat Linux7 server hosted on Amazon EC2. Without knowing the information where the data remnants are remained, it may take time for forensic investigation. For solving the issues; whether there are any data remnants to identify the use of Hadoop Big Data Storage System, and where the remnants may reside, the research questions are raised. Table 1 shows the research questions of this environment. These research questions are associate with the investigation of two Hadoop infrastructures.

Table 1 Research question for forensic investigation of Hadoop Big Data Storage System.

Q 1. What data are remained on sever site resulting from the use of the Hadoop Big Data Storage System to identify its use?
The sub questions are raised from the above primary question. Q 1.2. What data remnants can be discovered on Ubuntu 14.04 when Hadoop 2.7.1 run on it?
Q 1.1. What data remnants can be discovered on Red Hat Linux7 server hosted on Amazon EC2 if the Hortonworks HDP 2.3 is running on it?
Q 2. What data are remained on Client site resulting from the access of Hadoop Big Data storage server to identify its use?
Q 2.1. What data can be found on Windows 7 computer hard drive and memory after accessing the Hortonworks HDP 2.3 on Amazon EC2 via web browser and SSH access?
This work proposes a new process model for forensic investigation of Hadoop Big Data Storage System. The investigation scope contains discovering data remnants on Hadoop Server and the attached client machines.

4.2 Proposed Forensic Investigation Process Model for Hadoop Big Data Storage System

Over the past few years, a number of forensic process models have been proposed. However these existing models may not be fit-for-purpose in the Big Data Storage System environment. A sound forensic process model for investigation in this environment is required. This section describes a new investigation process model that is adaptable for Hadoop Big Data Storage System. This proposed process model is based on NIST forensic process model. Figure 3 illustrates the proposed investigating process model for Hadoop Big Data Storage System.

As the contribution of this process model, there is a cycle on the steps of requirements preparation, collection, and analysis. If the forensically sound data cannot be collected in the phase of collection, the investigation can go back to requirements preparation phase to arrange the usable tools and techniques for efficient collection Likewise, if there is a difficulty in analysis phase, re-operate the requirements preparation 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.

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. Figure 4 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.

Requirements Preparation: 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 study the documentations of previous investigations. Figure 4 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. The preparing of forensic tools to be adaptable

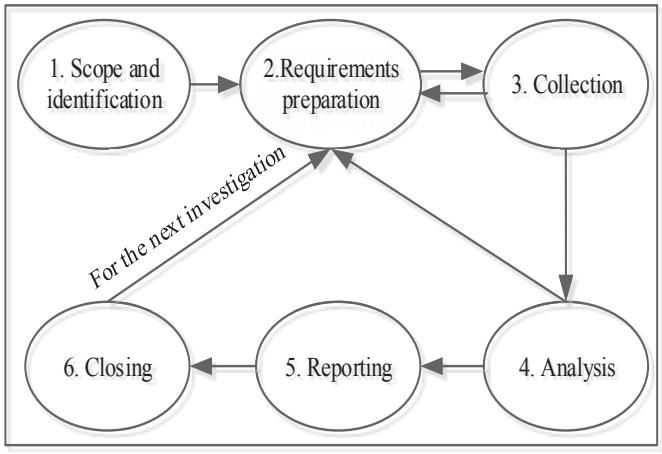

Figure 3 Proposed forensic investigation process model for Hadoop Big Data Storage System.

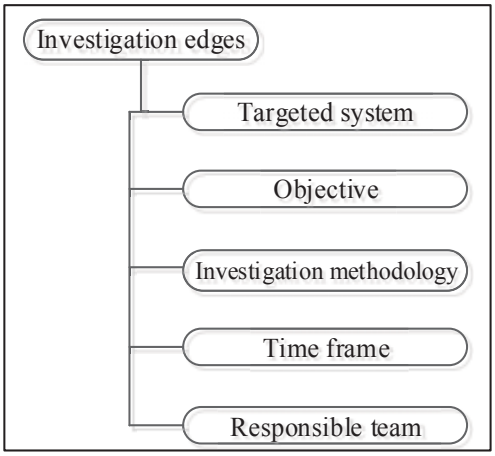

Figure 4 Edges of the investigation.

with this environment is one of the responsibilities of this phase.

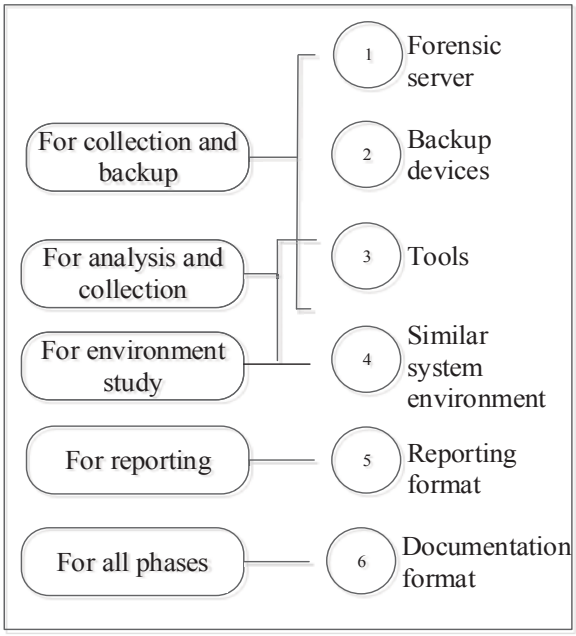

Figure 5 Required materials for investigation.

Figure 5 describes the material requirements to undertake the investigation. Forensic Server is needed to set up to collect and analyze the forensic data. To study the background knowledge of the infrastructure, a system which is similar configuration with the targeted system is equipped. The function of the Forensic Server is depicted in Figure 6. The Forensic Server requires the high access right to collect forensic data from targeted machines. The collected data are duplicated in other backup media for emergency use. The forensic analysis is also done in this server. Forensic imager tools and analysis tools are setting up on it. The responsibility of Forensic Server is

- to perform remote collection
- to store and make backup the forensic copy of disk images, memory dump and registry files which are collected from each machine
- to mount these file and explore in read only mode
- to conduct investigation and analysis.

Analysis: 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 follow:

Keyword Searching: Big Data investigations can contain both structured and unstructured data source. These data may contain keywords of wanted information. The simplest method is matching with keywords. The data is gathered and mined from the file system's metadata layer and then parsed to sort for further analysis.

Timeline Analysis: The end goal is to embody the incident activity performed in the system comprising its date, the involved artifact, action and source.

Media and Artifact Analysis: The investigator is overawed with the large amount of information that he has to check, he

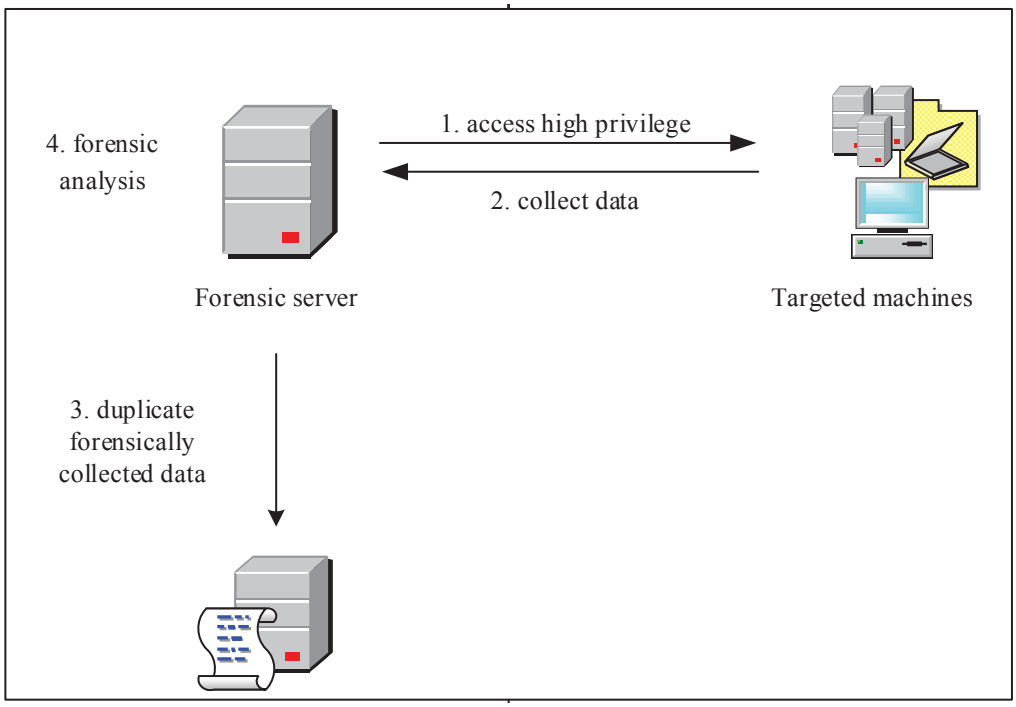

Figure 6 Functions of Forensic Server.

should examine some points such as which files were downloaded, what programs were executed, which directories were opened, which files were clicked on, which files were deleted, where did the user browsed to and so on. In this analysis way, each file is emphasized on to trace the footages of the criminals or illegal usages on it. Hence, knowledge of file systems is required; configuration remnants and registry remnants to take advantage of this procedure that reduce the data amount to be analyzed. At the end of this phase, the output is hanged over to the next phase to draw the event reconstruction and reporting.

Reporting: This phase presents the findings as the outcome of the investigation. The results obtained from above phases are organized to draw a conclusion. The overall view established that the associations between individual results may provide a picture. It is the presenting strategic for exposing the incidence (case); this must be full of clarity, completeness, and accuracy of the findings. This means the findings have to be presented in a comprehensible way that is available to a non-technical audience. 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.

Closing: This phase retains all related documentation recorded at each phase of the investigation. Each phase is reviewed so that the lesson can be learnt and applied for future investigations. The Figure 7 describes the tasks to accomplish closing phase.

In this phase, the conclusion is drawn by deciding upon the result of reporting phase. All collected data through the process and resulting data remnants are stored and archived. The document in this phase is finalized document and that contains the summarizing the activities and occurrences of the whole process. The resulting document is stored together with previous ones. The documentation file of whole process allows the investigators to prepare the required materials and methodologies for the future investigations.

5. FORENSIC INVESTIGATION OF HADOOP STORAGE SERVER

The Hadoop characteristics; low cost, computing power, scalability and storage flexibility makes the Hadoop to deploy as organizational storage server. As the use of Hadoop storage server continues to grow rapidly, it becomes the target or facility to commit crime. In this work, the proposed process model is applied to investigate two infrastructures of Hadoop storage systems with different installation methods.

5.1 Forensic Investigation of Hadoop Infrastructure I

In this section, the forensic investigation is implemented on Hadoop storage sever infrastructure I by applying the proposed forensic investigation process model.

5.1.1 Scope and Identification Phase

The scope of the investigation system is Hadoop Big Data Storage System. The targeted machines to be investigated are Hadoop Big Data Storage Server, version 2.7.1, on Ubuntu 14.04 LTS. Hadoop 2.7.1 is released by Apache Foundation. This investigation intends to discover the data remnants when the ser-

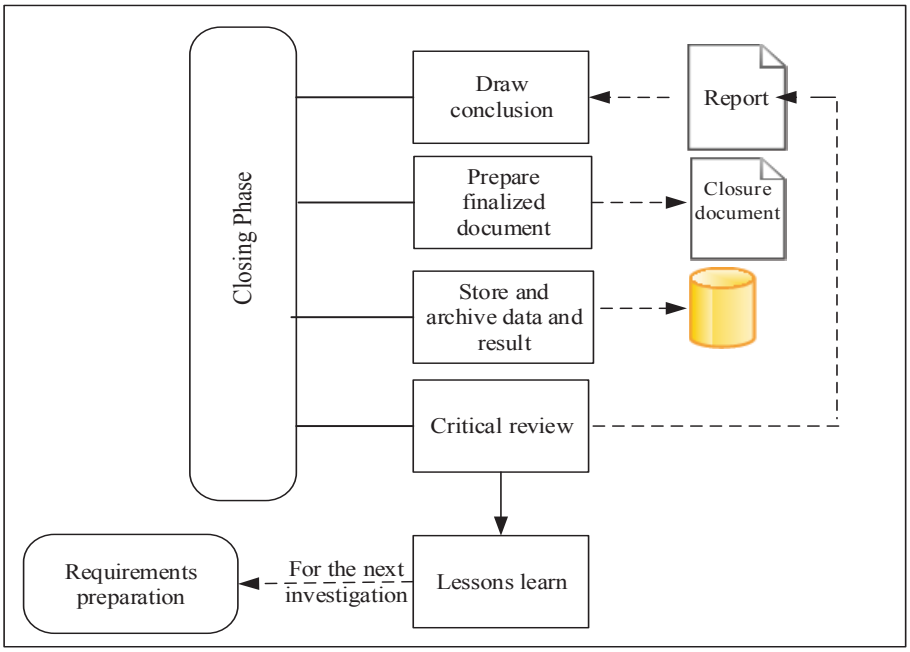

Figure 7 Tasks of closing phase.

vices; file uploading, downloading and MapReduce processing are operated on the Hadoop infrastructure I.

5.1.2 Requirements Preparation Phase

In order to support the investigation, it is desirable to prepare the tools, software and methods which are compactable with this system infrastructure. Firstly, the investigator should prepare the system environment which has the same infrastructure with this current targeted environment. This similar system allows the investigator to study the nature of targeted system, test the tools and techniques. The following is step by step installation and configuration of Hadoop Storage Server.

1. Installing Java on Ubuntu 14.04
2. Installing and Configuring SSH
3. Disabling IPv6
4. Installing the Hadoop Package And then, a Forensic Server is setting up. It connects to the targeted server machines to conduct investigation and analysis. In addition, backup devices and blank media such as external storage devices are prepared to duplicate the data. The forensic tools and techniques should be prepared. The PuTTY 0.67[23], WinSCP 5.7.7 [31] and FileViwerPlus 2.01 [12] are prepared to collect data and change the format into human readable form.

5.1.3 Collection Phase

Before conducting the forensic analysis, the data should be forensically collected for analysis. For the effective collection, the prioritizing the data sources; likely value and volatility should be implemented. The forensically important parts and files are collected in first priority. Through the study of the Hadoop infrastructures, the forensically important categories in Hadoop are:

- Hadoop Daemon Logs created by the Hadoop daemons end with .log, and others end with .out.
  - The .log and .out file names are constructed as follows:
  - Hadoop-<user-running-Hadoop> – <daemon> – <hostname>.log
    For example: Hadoop-Hadoop-Datanode-IP-xxxx.log
- JobTracker Logs
  - are created by the jobtracker.
  - are stored in two places:/var/log/Hadoop and /var/log/Hadoop/history.
    or home/Hadoop/logs/history/done/version x/<hostjob_id>/<year>/<month>/<day>/<serial>
- TaskTracker Logs (for a particular task attempt)
  - are created by each tasktracker.
  - are captured when a task attempt is run.
  - /var/log/Hadoop/userlogs/attempt_<job-id>_<map-or-reduce>_<attempt-id>
  - home/Hadoop/logs/userlog/
- 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.

After collecting the important files, collecting the volatile data should takes precedence over nonvolatile data. To get the memory image, dump the memory with "dd" command.

" 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 "

5.2 Analysis Phase

By analyzing the collected data, this investigation can track the footage on Hadoop storage server to identify the usages. The usages include the primary operation services; uploading, downloading and MapReduce function. Tables 2 through 4 reports the data remnants related to each operation service. In the tables, the IP address is expressed as xxx. The data remnants are:
As shown in Table 2, the uploaded file name, source IP, destination IP and file operation name; upload = WRITE are left on Hadoop server when the uploading service is operated.

As shown in Table 2, the uploaded file name, source IP, destination IP and file operation name; upload = WRITE are left on Hadoop server when the uploading service is operated.

In addition, Table 4 presents the remnants related to MapReduce task. When the MapReduce task is operated on a data set, a new folder is appeared in the metadata level log file. The name of this folder is composed of the program name (eg. WordCount) and processing time of this task. In Syslog, the output file name is recorded.

5.2.1 Reporting Phase

The investigators arrange the finding evidences to embody the event. They draw the event line with a specific feature; time and sequence. The output is documented to present in court of law. This phase relates to the legal presentation of the collected evidence and investigation. The presentation of findings can be demonstrated in many formats of documentation. Regardless of the format, one important point to notice is to clearly present the findings and collected evidences. The following data remnants are discovered on Hadoop Big Data Storage Server, version 2.7.1, on Ubuntu 14.04 LTS.

- Source IP
- Destination IP
- File operation
- Operated file name.

5.2.2 Closing Phase

By viewing the resulting remnants, the conclusion can be drawn that the usage of Hadoop 2.7.1 server can be identified. The investigator checks the documentations of each phase to extract which factors should be noticed for the next investigation. The finalized documentation is created and the whole documents are organized. The collected data are stored in archived format.

5.3 Forensic Investigation of Hadoop Infrastructure II

In this section, the forensic investigation is implemented on Hadoop storage sever infrastructure II by applying the proposed forensic investigation process model.

5.3.1 Scope and Identification Phase

In this investigation, the target system for investigation is Hadoop HDP 2.3 on Red Hat Linux 7 server which is hosted on Amazon Web Services EC2. The objective of this investigation is to trace the operation of file uploading, downloading, just opening a file on server and uninstalling the HDP2.3. This section focuses on discovering whether there are any data remnants on this storage server.

5.3.2 Requirements Preparation Phase

Initially, 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 their website to be installed. EC2 storage space is rent to install the Red Hat Linux7. HDP 2.3 is deployed on the top of Red Hat. In order to setup the Hadoop via Ambari, the installation steps are:

1. Lunching an EC2 instance
2. Pre-requisites for setting up Hadoop in Amazon Web Services
3. Hadoop cluster installation (via Ambari) Afterward, 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. The infrastructure study and testing the tools are done in this similar environment. In addition, the Forensic Server is implemented. Forensic tools, methods and other facility software for collection and analysis are also prepared as in section 5.1.3.

5.3.3 Collection Phase

For data collection, the Forensic Server connect the EC2 instance via PuTTY 0.67[23] and WinSCP 5.7.7. The forensic data are duplicated in other media. The prioritizing of the data sources; likely value and volatility is implemented. The forensically important files and volatile data are collected in first priority.

5.3.4 Analysis Phase

The exporting VM files are opened in the Forensic Server. This collected VM are analyzed to identify the usage and discover the remnants. Tables 5 through 7 list the data remnants by tracking the upload, download and read operation.

Table 2 Data remnants for file uploading.

Location	File Names	Remnants	Remarks
Home/Hadoop/logs	Hadoop-xxx-Namenode–xxx.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	src: xxx.xxx.xx.xxx Dest:xxx.xxx.xx.xxx OP: WRITE	Source IP, Dest IP, File operation

Table 3 Data remnants for file downloading.

Location	File Names	Remnants	Remarks
Home/Hadoop/logs	Hadoop-xxx-Namenode–xxx.log.2016-02-05	NameSystem.allocateBlock: /user/xxx/file_name.csv (original path of uploaded data set)	uploaded file name
Home/Hadoop/logs	Hadoop-xxx-Datanode–xxx.log.2016-02-05	src: xxx.xxx.xx.xxx Dest:xxx.xxx.xx.xxx OP: WRITE	Source IP, Dest IP, File operation

Table 4 Data remnants for MapReduce task.

Home/Hadoop/Hadoop 1.2.7/logs/userlogs	job-2016042511_0001	Folder name is given by processing time	New folders appear
Home/Hadoop/Hadoop 1.2.7/logs/userlogs/job-2016042511_0001/attampt_2016042511_0001_r_00000	Syslog	save output to file_name.csv	
Logs/history/done/ version1/xxx/461552092389_/2016/04/25/00000/	job_201604250911_0001_1461554117485_xxx_word+count	JOBNAME="word count"USER="xxx" SUBMIT_TIME="1461554117485"	New folders appear
var/opt/Hadoop/cluster/dfs/data/current	blk_498030556106732159	JOBNAME="word count" USER="xxx" SUBMIT_TIME="1461554117485"	

Table 5 Data remnants for file uploading.

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

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

The data remnants in Hadoop 2.7.1 on Ubuntu 14.04 LTS and the remnants on HDP 2.3 on Red Hat Linux7 server are same, but the two artifacts are the exception. These additional remnants are presented in Tables 5 through 7. Table 7 presents the remnants which are left by the process of opening and reading the file on the Hadoop storage server.

The Hadoop HDP is uninstalled from Red Hat Linux7 sever with the command: yum remove

Hadoop*, yum remove hdp*. The remaining remnants are:

Table 6 Data remnants for file downloading.

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

Table 7 Data remnants for reading file.

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

- 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

5.3.5 Reporting Phase

For the full presenting the forensic report, the investigator observes the data remnants and reconstructs the event to explore in law court. For the investigation of both Hadoop Storage Server infrastructures, the remaining data remnants are the same, however, the parts and files which contain these remnants are different.

5.3.6 Closing Phase

The investigator needs to catch up the process in every step to notice which factors are important for the next investigation.

6. FORENSIC INVESTIGATION OF CLIENT MACHINES

The storage and processing services can be supported by connecting the storage server from the client machines. In this section, the investigation is conducted on client machines that are accessed to the Hadoop server.

6.1 Scope and Identification Phase

The investigation objective of this section is to discover what data remnants are left on client machines when accessing the server. While identifying the client machines, we found that there are two type of accessing methods; via web access and SSH access. In web access, the client machine is connected to Hadoop server with the link "http://ec2-50-112-211-185.us-west2.compute.amazonaws.com:8080/". Table 8 illustrates the targeted area of client machines. In SSH access, the client machines access the Hadoop server with PuTTY 0.67 through the listing steps as shown in below:

- Generate the key .ppk from .pem with PuTTYKeyGen 0.6.7.11830 [22]
- Run the PuTTY0.67 software
- Open Connection >> SSH >> Auth
- Browse the private key file from authentication
- Load .ppk file
- Login with user name and password

6.2 Requirements Preparation Phase

The tools which are compactable with the targeted machines are prepared for both static and live analysis. Afterward, for studying the infrastructure, the three type of client machines are prepared as stated in Table 8. A Forensic Server is also prepared to collect and analysis forensic data.

Table 8 Targeted client machines.

Client Machine Type	Access Method	Accessed software	Underlying OS
Client machine 1	Web Access	Mozilla Firefox 49.0.1	Windows 7, 64bit
Client machine 2	Web Access	IE 8.0.7601	Windows 7, 64bit
Client machine 3	SSH access	PuTTY 0.95	Windows 7, 64bit

6.3 Collection Phase

In this phase, the investigator collects disk image, memory dump and protected registry file of current machine by using forensic tools and file viewer software. To create the forensic image of hard disk, the write blocker is used to ensure that no data is written back to hard drive. We use the AccessData FTK imager 3.0.0.143 [1] for imaging by blocking write mode. After imaging the hard drive, the image file is collected in a Forensic Server. The memory dump files of each client machine are also collected for the live analysis.

6.4 Analysis Phase

The collected data from each VM are conducted to analyze. The acquiring image files are mounted to Forensic Server and open in read only mode to discover the data remnants.

(i) Testing Environment I, Analysis on Client Machine 1

For testing environment 1, the Client VM 1 is investigated. The specification of client machine1 is the Windows 7 64 bit which is accessing the server via Mozilla Firefox 49.0. The disk image file of this Client VM 1 mounted to the Forensic Server. The mounting drive is explored and the data remnants are discovered. The data remnants such as web address, access date, uploaded file name, upload date are found as shown in the Table 9.

(ii) Testing Environment II, Analysis on Client Machine 2

For testing environment 2, the Client machine 2 is investigated. The specification is Windows 7 64 bit OS that access the server via IE 8.0.7601. The data remnants are discovered as shown in Table 10.

(iii) Testing Environment III, Analysis on Client Machine 3

Table 11 depicts the data remnants of analyzing on client machine 3. When the PuTTY is used to connect the server, the remaining remnants are:

- the session name under the hierarchy of %regedit% > Simon Tatham > PuTTY
- port number :name(or)IP under the hierarchy of %regedit% > Simon Tatham > PuTTY > SshHostKeys.

6.4.1 Memory Analysis on Client Machines

Figure 8 illustrated the reading .mem file with File Viwer Plus. The memory dump file with extension .mem is the type of Hex file so, the FileViewerPlus 2.0.1.36 [12] is conducted to view the .mem file. The discovering remnants are file name, bowering web address and date.

6.5 Reporting Phase

The investigator arranges the finding evidences to embody the event that could be. They draw the event line with a specific feature (time, sequence). The output is documented to present in court of law. This phase relates to the legal presentation of the collected evidence and investigation. The presentation of findings can be demonstrated in many format of documentation. Regardless of the format, one important point to notice is to clearly present the findings and collected evidences. The following data remnants are discovered on client site resulting from the use of the Hadoop Big Data Storage System:

While accessing the Hadoop server via web browser,

- URL
- Accessed date
- Operated file names are also discovered in client site to identify the use of Hadoop.

While accessing the Hadoop server by SSH access, the discovering data remnants are

- session name
- port number.

6.6 Closing Phase

By viewing the resulting remnants, the conclusion can be drawn that we can discover data remnants on client machines to trace the usage of Hadoop server. The whole documentations are organized for later use. The collected data are stored in archived format. The investigator reviews the tasks of each phase to extract which factors should be notice for the next investigation.

7. CONCLUSION AND FUTURE WORKS

The popularity and advent of Big Data technology persuade the criminals to focus on it. It leads to forensically investigate the Big Data Storage System by tracking the illegal usage. Big Data Storage System is identified as an emerging challenge to digital forensic researchers and practitioners. The traditional forensic process models are not fit for this environment. There is a need for a process model to guide forensic investigation where the Big Data Storage System is involved. In addition, without knowing where data remnants may reside can take the considerable amount of time for forensic analysis. This paper proposes an investigation process model for Hadoop Big Data Storage System. This six-phased iterative model is based on

Table 9 Data remnants on investigating Client machine1.

File Location	Remnants	Remarks
\AppData\Local\Mozilla\ Firefox\Profiles\rnllugnp.default-1456638777411\cache2\ entries\	000BB33FB49A474BFDB7 010734E1D094AE390271	web address, Access Date
C:\Users\User\AppData\ Roaming\Mozilla\Firefox\ Profiles\rnllugnp.default-1456638777411\datareporting\ archived\2016-10	1475316058064.c9aa4ed3-091c-4f71-996d-17955a8ac7ca.main.jsonlz4	Uploaded file name, date

Table 10 Data remnants on investigating Client machine2.

File Location	File Name	Remnants
%Profile%\AppData\Local\ Microsoft\Windows\Temporary Internet Files\Content.IE\index.dat	000BB33FB49A474BFDB7010 734E1D094AE390271	Web address, access Date
%Profile%\AppData\Local\ Microsoft\Windows\Temporary Internet Files\Content. IE\ <Random>\ <All of the files>	1475316058064.c9aa4ed3-091c-4f71-996d-17955a8ac7ca.main.jsonlz4	Uploaded file name, date

Table 11 Data remnants of connecting via PuTTY.

File Location	File Name	Remnants
Registry editor	%regedit%user_name/PuTTY	Session name
Registry editor	%regedit%user_name/PuTTY/SshHostKeys	IP address

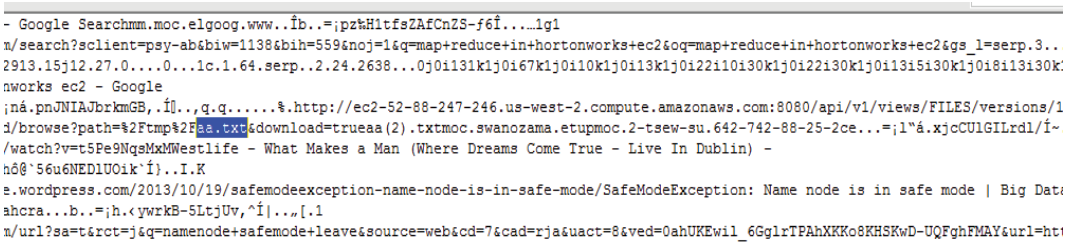

Figure 8 Reading .mem file with File Viwer Plus.

NIST standard model. The proposed forensic process model guides the investigation from the scope to closing phase. The model is applied to investigate on two infrastructures of Hadoop Big Data Storage System which are residing on different OS;

(i) Hadoop 2.7.1, on Ubuntu 14.04 LTS

(ii) Hadoop HDP 2.3 on Red Hat Linux 7 server which is hosted on Amazon Web Services EC2. The usage of Hadoop can be identified by discovering data remnants on the Hadoop servers and also on client machines. The identification of the log and metadata files is important in the investigation of Hadoop Big Data Storage System. Browser history, registry, and memory captures are also important in an investigation of client machines. Also of note that even the uninstallation of Hadoop HDP leaves the remnants. The resulting data remnants of Hadoop Big Data Storage System can assist the forensic examiners in generating evidence to implement effective Hadoop Big Data Storage System forensics. Future research opportunities include conducting forensic research on other DFS such as GLORY-FS, NFS, and GFS. This paper can also be extended to the investigation of other client devices. The developing of automatic live forensic tools for forensic investigation of DFS based Big Data Storage System will be a future work.

Acknowledgement

The authors would like to acknowledge the anonymous referees for their valuable comments in improving the quality of the paper.

REFERENCES

1. "AccessData Forensic Toolkit (FTK)," Available: http://accessdata.com/solutions/digitalforensics/forensic-toolkit-ftk. Accessed: Oct. 29, 2016.
2. O.Alabi et al., "Toward a Data Spillage Prevention Process in Hadoop using Data Provenance," Proceedings of the 2015 Workshop on Changing Landscapes in HPC Security, pp.9–13, 2015.
3. "Apache Hadoop," Available: https://archive.apache.org/dist/Hadoop/core/. Accessed: Oct. 15, 2016.
4. "Amazon EC2 -Virtual Server Hosting," Available: https://aws.amazon.com/ec2/. Accessed: Nov. 8, 2016.
5. S.S.Baboo and S.M.Megalai, "Cyber Forensic Investigation and Exploration on Cloud Computing Environment," Global Journal of Computer Science and Technology, vol. 15, no.1, 2015.
6. D.Conroy, "Forensic Data Analysis Challenges in Large Scale Systems." Intelligent Distributed Computing IX. Springer International Publishing, pp. 451–457, 2016.
7. N.Beebe, J.Clark, "A Hierarchical, Objectives-Based Framework for the Digital Investigations Process," in Digital Forensic Research Conference, Baltimore, MD., Aug. 2004.
8. V. Chemitiganti, "What is Apache Hadoop?," in Business Values of Hadoop, Hortonworks, 2016. Available: http://hortonworks.com/apache/hadoop.
9. C.H.Cho et al., "Cyber Forensic for Hadoop Based Cloud System," International Journal of Security and its Applications, vol 6.3 , pp.83–90 , July, 2012
10. H.Chung et al., "Digital Forensic Investigation of Cloud Storage Services" Digital investigation vol. 9, no.2, pp. 81–95, Nov. 30, 2012.
11. B. Cusack and R Lutui, "Up-Dating Investigation Models for Smart Phone Procedures," Australian Digital Forensics Conference, Perth, WA, 2014.
12. "FileViewerPlus," Available: http://fileviewerplus.com.siterankd.com/. Accessed: Sept. 8, 2016.
13. M. Gualtieri and N. Yuhanna, "The Forrester Wave: Big Data Hadoop Solutions, Q1 2014," Forrester, 2014.
14. "Welcome to apache Hadoop," Available: http://hadoop.apache.org/. Accessed: Nov. 8, 2016.
15. 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.
16. L. Leong, "The 2014 Cloud Iaas Magic Quadrant–Lydia Leong," in Infrastructure, 2014. Available: http://blogs.gartner.com/lydia_leong/2014/05/30/the-2014-cloud-iaas-magicquadrant.
17. U.Mohan and S.Salisu "The Use of Big Data in the Field of Digital Forensic Investigations (Comparative Study between Digital Forensics in UK and Nigeria)," International Journal of New Technologies in Science and Engineering, vol. 2, no. 4, Oct, 2015.
18. A.McAfee and E.Brynjolfsson "Big Data: The Management Revolution" in Harvard Business Review, Available: https://hbr.org/2012/10/big-data-the-management-revolution. Accessed: Nov. 1, 2016.
19. "MooseFS 3.0 User's Manual," Available: https://moosefs.com/Content/Downloads/moosefs3-0-users-manual.pdf
20. G. Plamer "A Road Map for Digital Forensic Research" The MITRE Corporation., Tech. Rep. No. DTR -T001-01, Nov.6, 2001.
21. S. V. President, "Hadoop/Big Data Market Size Worldwide 2015-2020 | Statistic," Statista, 2016. Available: https://www.statista.com/statistics/587051/worldwide-Hadoop-bigdatamarket/. Accessed: Nov. 8, 2016.
22. "Putty Keygen Download" Available: http://PuTTY.org/downloads/PuTTYkeygen. Accessed: Nov. 8, 2016.
23. "Putty download" Available: www.putty.org. Accessed: Nov. 8, 2016.
24. D. Quick, "Cloud Storage Forensic Analysis," M.S.thesis, School of Computer & Information Science, University of South Australia, Adelaide SA, 2012.
25. D.Quick and K.R. Choo. "Data Reduction and Data Mining Framework for Digital Forensic Evidence: Storage, Intelligence, Review And Archive." Trends & Issues in Crime and Criminal Justice, no. 480, pp. 1–11, Sept, 2014.
26. R.S.Satti and F.Jafari, "Reviewing Existing Forensic Models to Propose a Cyber Forensic Investigation Process Model for Higher Educational Institutes," International Journal of Computer Network and Information Security, vol. 7, no. 5, pp. 16–24, Apr. 2015.
27. J.Sremack, "Big Data Forensics–Learning Hadoop Investigations," Birmingham, UK: Packt Publishing Ltd, Aug. 2015.
28. "Statistics and Facts about Big Data," Available: https://www.statista.com/topics/1464/bigdata. Accessed: Nov. 1, 2016.
29. S.A.Thanekar, et al., "A Study on Digital Forensics in Hadoop" International Journal of Control Theory and Applications. Published By: International Science Press, vol. 9, no. 18, pp. 8927–8933, 2016.
30. S. A. Weil, "Ceph: Reliable, Scalable, and High-performance Distributed Storage," Ph.D. dissertation, Santa Cruz, CA, USA, 2007.
31. "WinSCP 5.7.7" Available: https://winscp.net. Accessed: Nov. 8, 2016.

Biographical notes

Myat Nandar Oo is a tutor of Information and Communication Technology Research Center, Yangon, Myanmar. She got the Bachelor of Computer Science (B.C.Sc), the Bachelor of Computer Science (Honours), and the Master of Computer Science (M.C.Sc) in 2006, 2007, and 2010 respectively. Now she is a Ph.D candidate in University of Computer Studies, Yangon (UCSY), Myanmar. Her research interests include digital forensics, big data, cloud computing and distributed system.

Dr. Sazia Parvin is a data and system security researcher at the School of Business, UNSW, Canberra. Her research interests include network security, trust management, cyber systems, cloud computing, big data analytics, system software and intelligent information systems. Her research is published in various top ranked publications. She has published over 38 research papers in her fields of interest as journals and international conferences. She is an Associate Editor for International Journal of Computer System Science and Engineering (IJCSSE) and International Journal of Engineering Intelligent Systems (IJEIS). She has more than 7 years of experience in information system?s design and development in various business environments. She holds 7 years of extensive teaching experience in Software and Computer Engineering discipline. She has achieved several prestigious Research Grants from Australia and South Korea. She is also the recipient of the 'Gold Medal' Bachelor of Computer Science and Engineering Award from Jahangirnagar Univeristy in 2004 for her outstanding performance (First Class First Position) in that academic year.

Dr. Thandar Thein received her M.Sc. (computer science) and Ph.D. degrees in 1996 and 2004, respectively from University of Computer Studies, Yangon (UCSY), Myanmar. She did post doctorate research in Korea Aerospace University. She is currently a Pro-Rector of University of Computer Studies, Maubin. Her research interests include cloud computing, mobile cloud computing, big data, digital forensic, security engineering, and network security.