Forensic Analysis of Residual Artifacts on CDH Storage

Myat Nandar Oo[1], Thandar Thein[2]
University of Computer Studies, Yangon[1], University of Computer Studies, Maubin[2]
myatnandaroo@gmail.com[1], thandartheinn@gmail.com[2]

Abstract

Hadoop Storage is increasingly used by consumers, business, and government, and can potentially store and process large amounts of data. With the maturing and wide usage of Hadoop Storage, there are more and more crimes in this environment. The retrieval of digital evidence from Hadoop Storage can be a challenge in forensic investigation, due to its complex infrastructure and, lack of location knowledge about digital evidences. As a consequence, forensic researchers are moving towards the investigation researches of locating and documenting the residual artifacts to trace the criminal activities of Hadoop Storage. The Cloudera Distribution Hadoop (CDH) is a popular Hadoop Storage Platform, providing users a cost-effective, and in some cases free with the ability to access, store, and process data. This paper proposes a forensic investigation framework for locating and discovering the residual artifacts that remain on the CDH Storage Server and attached client machine. The residual artifacts can provide the potential evidences for forensic examiners to extract the evidences, and reconstruct the crime scene.

Keywords- CDH, crime, forensics investigation framework, residual artifacts

1. Introduction

Hadoop Storage is increasingly used by government, businesses, and consumers to store and access a large amount of information. In Statista report [12], the Hadoop market was valued at 6 billion U.S. dollars worldwide in the year 2015. 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, Cloudera, and Hortonworks [10]. CDH, Cloudera's open source Hadoop Platform, is the most popular distribution of Hadoop and related projects [17].

The popularity of Hadoop Storage enables the criminal to conduct their activities on it for exploitation. With the growing use of Hadoop to tackle processing of sensitive data, a Hadoop could be a target for data exfiltration, corruption, or modification [11].

Hadoop is subject to exploitation by criminals, who may be able to use storage for criminal purposes, thus adding to the challenge of growing volumes of digital evidence in cases under investigation.

Overcoming these investigation challenges, it is important to have a contemporary understanding of the location and type of residual artifacts left behind by file operations of storage service. The identification of potential data stores is an area that can impede an investigation. The paper [18] found out to identify potential artifacts that remain on the client devices and servers involving the use of Syncany as a private cloud storage solution supporting the Big Data Platform. The forensic researchers discovered the artifacts on client devices to identify the usage of Google Drive [6], Skydrive [7] and Dropbox [5].

It is important to have a rigorous methodology and a set of procedures for conducting forensic research on the emerging technical environments (such as Hadoop Platform and cloud computing). The forensic work of locating and documenting the forensically residual artifacts is also required to trace the criminal activities. These residual artifacts can provide the forensic practitioners in extracting the effective evidences for future forensics works.

Martin and Choo [1] 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. This paper discovered the data remnants on client devices to identify the usage of cloud storage by applying their proposed forensic framework.

As far as I know, there are no publications concerned with the forensic investigation work on CDH Storage. This paper discusses the need for Hadoop Platform forensics and proposes the forensic investigation framework for CDH with the aim to discover what artifacts can be gathered from CDH. Organization of the paper is as follows: CDH Storage is explained in Section 2. The research questions and methodology for digital forensics are defined in Section 3. The CDH Storage Forensics Framework is presented in Section 4. Section 5 draws conclusions and describes future works.

2. CDH Storage

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. 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, today announced that it is positioned as a leader in The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016 report [10]. 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.

The architecture of CDH Storage is shown in Figure 1. The targeted CDH Storage utilizes Hadoop 2.x architecture. The Resource Manager manages resources and allots the resources to the application. It 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. Cloudera Manager is an end-to-end application for managing agent on each cluster.

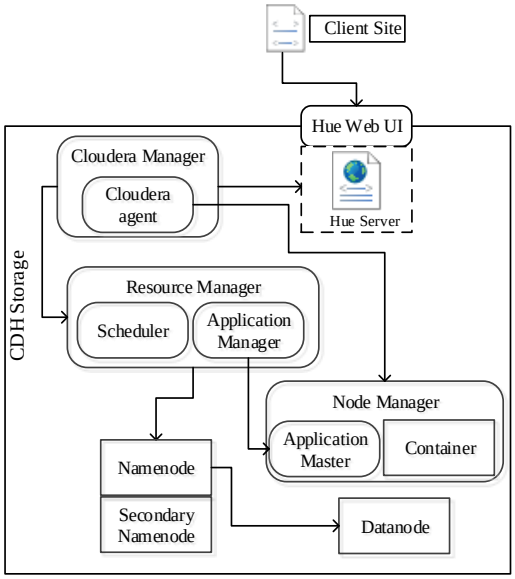

Figure 1: Architecture of CDH Storage

Conventional digital forensic methods are insufficient for investigating such composite infrastructure. Therefore, this paper proposes a forensic investigation framework for undertaking forensic research on CDH Storage. The resulting residual artifacts can provide effective evidences to the forensic examiners for future real-world CDH forensics.

3. Research Questions and Research Methodology

The identification of data remnants provides a better understanding of the types of artifacts that are likely to remain, and the access point(s) for digital forensics examiners to assist in the 'identification' stage of an investigation, which then follows with preservation and analysis. The ability to identify relevant data in a timely fashion can impact on an investigation by not including data that may be crucial to accurate findings in relation to circumstances; as such, the identification of data is an important part of the digital investigation process. The focus of this paper is to discover the residual artifacts left on CDH Storage and client machine.

3.1. Research Questions for Forensic Investigation on CDH Storage

For undertaking forensic research on the CDH storage environment, the following questions are examined:

Q 1. What data remnants are likely to remain after the use of CDH?

Q 2. What artifacts are created during the file operation on CDH storage?

Q 3. What data are remained on client machine that are resulting from the use of CDH to identify its use?

3.2. Forensic Investigation Framework for CDH Storage

The proposed forensic investigation framework is based on the National Institute of Standards and Technology [3]. It comprises five phases; preparation, collection, analysis, and documentation and presentation, and closing as shown in Figure 2.

In the framework, the analysis phase can be cyclic and iterative as it is common that during an investigation a forensic examiner may need to return to a previous step.

1. Preparation: concerns with preparation of tools, techniques, research methodology, training, acquisition, and management support

2. Collection: includes collection and acquisition of data from identified sources and preserving the crime scene and data
3. Analysis: concerns with an in-depth systematic search, focuses on identifying and locating potential evidence
4. Presentation: concerns with completely and accurately documenting of findings and the residual artifacts
5. Closing: retains all related documentation recorded at each phase and review them to learn lesson for future real-world forensics

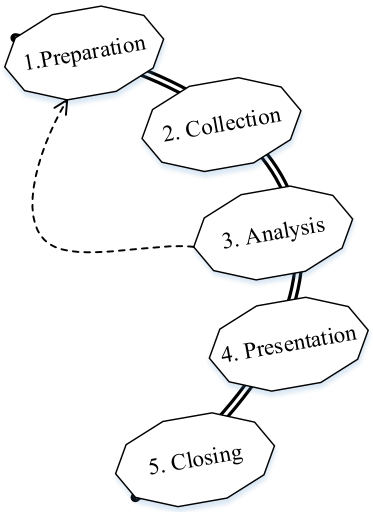

Figure 2: Forensic Investigation Framework for CDH Storage

4. Forensic Investigation Research on CDH Storage

This proposed framework finds out residual artifacts that are likely to remain after the use of CDH. These residual artifacts are useful to identify the action of a criminal. The remained artifacts can provide the potential evidences for forensic examiners to extract the evidences in exploring the illegal usages; what did the criminal do with CDH.

The investigation scope of this paper is to trace whether the suspected person connects the CDH server, and operate the primary file operations (upload, read and download) on the confidential data. Therefore, this section locates and discovers the residual artifacts on CDH Storage Server and attached the client machine to trace the file operations.

4.1. Preparation Phase

The targeted infrastructure is implemented and forensic tools are prepared to do the forensic research for gathering the artifacts that are likely to remain after the usage of CDH.

Table 1: Tools Prepared for Forensic Investigation on CDH Storage

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 [13]	To parse the file system, produce directory listings, as well as extracting/analysing the files, Windows registry, swap file/partition, and unallocated space from the forensic images.
SQLite Browser Version 3.4.0 [14]	To view the contents of SQLite database files.
Browser History Spy V-3.0 [15]	all-in-one software to instantly recover or view the browsing history from popular web browsers
WebBrowserPassView v1.56 [16]	the password recovery tool that reveals the passwords stored by the web browsers
File Viewer Plus 2 [8]	View, edit, save, and convert over the Hex files

The forensic tools for investigation both CDH Storage Server and the client machine are prepared as shown in Table 1.Testing environment and summary configurations of server and client are described in Table 2.

Table 2: System Configuration in Testing Environment

(CDH Storage)Server Configuration	Client Configurations
Operating System - Cent OS 6 Virtual memory size - 2GB Virtual HD size - 16GB Cloudera Version - Cloudera- quick-start VM - CDH 5.7.0 [2] IP address/ URL - http://hostname/8888	Operating System - Windows 7 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

4.2. Collection Phase

A digital forensic investigation is the ability to conduct analysis on a forensic copy, rather than interacting with or altering the original source. With the aim to adapt the nature of CDH, live forensic data collection is also needed. Data is copied in a forensic manner, using write-protection and creating a bit-for-bit copy. Secure collection of evidence is important to guarantee the evidential integrity and security of information. Files were identified which would contain the information needed to conduct the analysis; the virtual hard drives (VMDK files) in each Virtual Machine (VM) folder, each memory instance (VMEM files). These were identified for each of the VMs.

For the volatile data collection, imaging the memory of the server

"dd if=/dev/mem of=media/usb/memory.image"

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

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

MD5 hash value of each file is calculated and verified with each forensic copy to ensure the integrity of the duplicated file.

4.3. Analysis Phase

For this research, each of the forensic copies of the hard drives, memory and VM image captures were examined using the prepared forensic analysis tools. This paper locates the forensically valuable files and extracts the residual artifacts from the large amount of backlogs, metadata, registry and image. The residual artifacts on server and client machine can trace the criminal activities.

4.3.1. Evidential Analysis on CDH Storage Server

By analyzing the collected data, this investigation can track the footage on the CDH Storage Server to identify the usages. The usages include the primary file operations; upload, read and download. In this research, the residual artifacts related to each file operations are reported in Tables 3 through 5.

Among the large amount of backlogs and metadata of CDH Storage Sever, the most important artifacts which can completely explore the primary file operation are discovered in the hdfs-audit.log and access.log files. The residual artifacts are the operated date, user name, file name, source IP, destination IP and file operation name.

Table 3: Residual Artifacts (File Uploading)

File name - hdfs-audit.log	Path - /var/log/hadoop-hdfs/ Artifacts - [1]2017-6-20 00:18:48,allowed = true ugi=admin[2] src = /home/file.pdf [3] cmd=create[4] Type of artifacts - [1]Date - [2]User name - [3]File name - [4]Operation (create)
File name - access.log	Path - /var/log/hue Artifacts - [1]20/Jul/2017 20:41:27 [2]172.16.38.24 [3]admin – POST[4] [5]/filebrowser//dirBb/file.pdf[6] Type of artifacts - [1]Date - [2]Source IP - [3]User name - [4]Access method (POST for upload) - [5]File Path - [6]File name

Table 4: Residual Artifacts (File Reading)

File name - hdfs-audit.log	Path - /var/log/hadoop-hdfs/ Artifacts - [1]2017-6-20 00:18:48,allowed = true ugi=admin[2] src = /home/file.pdf [3] cmd=getfileinfo[4] Type of artifacts - [1]Date - [2]User name - [3]File name - [4]Operation (getfileinfo)
File name - access.log	Path - /var/log/hue Artifacts - [1]20/Jul/2017 20:41:27 [2]172.16.38.24 [3]admin – GET[4] [5]/filebrowser//dirBb/file.pdf[6] Type of artifacts - [1]Date - [2]Source IP - [3]User name - [4]Access method (GET for read) - [5]File Path - [6]File name

Table 5: Residual Artifacts (File Downloading)

File name - hdfs-audit.log	Path - /var/log/hadoop-hdfs/ Artifacts - [1]2017-6-20 00:18:48,allowed = true ugi=admin[2] src = /home/file.pdf [3] cmd=open[4] Type of artifacts - [1]Date - [2]User name - [3]File name - [4]Operation (open)
ile name - access.log	Path - /var/log/hue Artifacts - [1]20/Jul/2017 20:41:27 [2]172.16.38.24 [3]admin – GET[4] [5]/filebrowser//dirBb/file.pdf[6] Type of artifacts - [1]Date - [2]Source IP - [3]User name - [4]Access method (GET for download) - [5]File Path - [6]File name

4.3.2. Evidential Analysis on Client Machine.

In order to analyze the vmdk image and collected data on the client machine, the prepared forensic analysis and recover tools are tested and applied. The aim of analysis is to explore what residual artifacts are left to identify whether CDH Storage Server was accessed via the web browser on the client machine.

The artifacts found on the client machine are URL, date, time, and file name. Moreover, the browser, the log files, the accessed web URL, the title of the website, the visited date and time are also identified. The web address of the server machine is also found in the browser cache file entries on the client machine. The forensically important files and residual artifacts of the popular web browsers; Mozilla Firefox, IE and Google Chrome are shown in Table 6.

Table 6: Important files and paths of Web Browsers

Mozilla Firefox 33.0.2	
Data	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

4.4. Presentation Phase

According to the experiment, we found that a variety of data remnants were located when the user makes file operations on CDH.

The important files, paths, artifacts of storage server and attached client machine are documented. This information enables a practitioner to conduct forensic analysis and will assist to embody the criminal activity.

4.5. Closing Phase

The whole documentations are organized for later use. The collected data are 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. Conclusion and Future Works

The usage of Hadoop Storage is becoming more widespread. It is possible for malicious users to handle the illegal usages and the number of crimes on them has increased rapidly. This paper proposes a forensic investigation framework for locating and discovering the residual artifacts on CDH Storage Server and attached client machine. The residual artifacts can identify the use of CDH, trace the file operations and explore the illegal usages. The remained artifacts can provide forensic examiners in generating the effective evidences, and embody the criminal activity. The in-depth forensic analysis with crime scenario on CDH will also be presented in our later research. Future research opportunities also include conducting forensic research and exploring forensic methodologies for other Hadoop Distributions and Big Data solutions. And then, the development of log analysis model for forensic investigation of Hadoop Storage Platforms will also be a future work.

6. References

[1] B.Martini and K.K.R. Choo, An integrated conceptual digital forensic framework for cloud computing. Elsevier-Digital Investigation, volume 9(2), pp. 71-80, 2012.

[2] Cloudera Hadoop Distribution, Available: www.Cloudera.com, Accessed: September 3, 2017.

[3] K. Kent, "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.

[4] Dcfldd Available: http:// stefanoprenna.com/ blog/2014/03/02/ tutorial-how-to-use-dcfldd-instead-of-dd/ Accessed: September 3, 2017.

[5] D.Quick and K.K.R.Choo, "Dropbox Analysis: Data Remnants on User Machines," Digital Investigitation, vol. 10, no. 1, pp. 3–18, 2013.

[6] D.Quick and K.K.R.Choo, "Google Drive: Forensic Analysis of Data Remnants," J Netw Comput Appl, vol. 40, pp. 179–193, 2014.

[7] D.Quick and K.K.R.Choo, "Digital Droplets: Microsoft SkyDrive Forensic Data Remnants," Future Gener. Comput. Syst., vol. 29, no. 6, pp. 1378–1394, 2013.

[8] FileViewerPlus Available: http://fileviewerplus.com.siterankd.com/. Accessed: Sept. 8, 2017.

[9] FTK Imager, Available: accessdata.com/product-download/ftk-imager-version-3.2.0, Accessed: Sept. 8, 2017.

[10] Report of Hadoop Big Data Distribution, Available: "https:// www.forrester.com/ report/The+Forrester+Wave+Big+Data+Hadoop+Distributions+ Q1+2016. Accessed: September 3, 2017.

[11] S.Acharya, J.Cohen "Towards a More Secure Apache Hadoop HDFS Infrastructure," in Network and System Security, Lecture Notes in Computer Science Volume 7873, 2013, pp 735-741.

[12] S. V. President, "Hadoop/Big Data Market Size Worldwide 2015-2020 | Statistic," Statista, 2016. Available: https://www.statista.com/statistics/587051/worldwide-Hadoop-bigdata-market/. Accessed: Nov. 8, 2016.

[13] Sleuthkit Autospy, Available: https:// www.sleuthkit.org/autopsy/download.php, Accessed: Sept. 8, 2017.

[14] SQLite Database Recover, Available: https://www.stellarinfo.com/sqlite-repair.php , Accessed: Sept. 8, 2017.

[15] Web Browser Pass View, https:// www securityxploded.com/ browser-password-decryptor.php, Accessed: Sept. 8, 2017.

[16] Webbrowserhistoryspy: Availabe: http:// www.nirsoft.net, Accessed: August. 4, 2017.

[17] Top 6 Hadoop Vendors providing Big Data Solutions in Open Data Platform Available:https://www.dezyre.com/article/top-6-hadoop-vendors-providing-big-data-solutions-in-open-data-platform/93, Accessed: August. 4, 2017.

[18] 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 ", in Security, Pravicy and Anonymity in Computation, Communication and Storage, Springer International Publishing, 2016, pp.342-357.