Investigation of Android Device for Discovering Hadoop Cloud Storage Artifacts

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

Abstract

Hadoop Cloud Storage 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 device. It is fast raising popularity to access Hadoop Cloud services via Android device. The widespread usage of Hadoop Cloud Storage could create the environment that is potentially conducive to malicious activities and illegal operations. Thus, the investigation of Hadoop Cloud presents the emerging challenge for the digital forensic community. Extracting residual artifacts from the cloud server is potentially difficult due to privacy policies followed by cloud providers. The attached Android device may store useful artifacts to investigate the illegal usages of Hadoop Cloud Storage. This paper utilizes Cloudera Distribution Hadoop (CDH); a popular Hadoop Cloud Storage. This paper conducts a preliminary investigation to locate and extract the residual artifacts from Android device that has accessed the CDH Cloud. The extracted artifacts can assist the forensic examiners in real world Hadoop Cloud forensics. The crime scenario which is extended the Forensic Copra's crime case is examined under the guide of CDH Forensic Investigation Framework.

Keywords*: Android Device, CDH, Criminal Activity, Forensic Investigation*

1. Introduction

Hadoop Cloud Storage is increasingly used by government, businesses, and consumers to store and access a large amount of information. According to the report of Zion Market Research, the global Hadoop market was valued at approximately USD 7.69 billion in 2016 and is expected to reach approximately USD 87.14 billion by 2022 [5]. The popularity of Hadoop Cloud 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 the data exfiltration, corruption, or modification [1]. Hadoop Cloud is subject to exploitation by criminals, who may be able to use storage for criminal purposes. Hence Hadoop Cloud is an emerging field for forensic investigators. Most accurate evidences can be extracted from cloud serves; however forensic investigators may not granted access to Hadoop Cloud servers due to privacy policies followed by cloud providers. Although it is, in principle, possible for the evidences to be obtained from the cloud service provider, this process may take a significant amount of time, or be obstructed by cross-border jurisdictional disputes. The amount of evidences can be extracted from client devices that may be of forensic investigator's interests. Android smart devices are popular client devices to access Hadoop Cloud. Thus, the focus of this research is to locate artifacts left on attached Android device, when a user has carried out different activities such as upload to and download of files and folders.

Through the attractiveness of Hadoop, a number of companies became bundle Hadoop and some related technologies into their own Hadoop distributions. 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.

This paper conducts a forensic investigation for locating and documenting the evidences by analyzing the attached Android client devices. This paper utilizes CDH Cloud for conducting forensic analysis of client devices to investigate the illegal usages of Hadoop Cloud. The crime scenario which is extended the Forensic Copra's crime case is examined.

The organization of the paper is prepared as follows: Section 2 expresses the overview of the researches conducted on the client devices investigation for cloud forensics. Section 3 describes background topics involving Android, Hadoop Cloud and CDH. Section 4 states the research questions. Section 5 expresses the preliminary forensic investigation of Android device and residual artifacts are presented. And then crime scenario is investigated in Section 6. Finally, Section 7 draws conclusion of this paper.

2. Literature Reviews

This section presents the academic publications with a technical focus on cloud forensics is relatively few compared with established disciplines. It has been pointed out that some papers discussing the forensic collection of cloud storage

have appeared, and their focus is on the client devices forensics due to the difficulties in obtaining access to a cloud provider's data center to conduct server analysis.

The researchers investigated Google Drive [9], and Skydrive [10] by extracting evidential data from data remnants on client devices when a cloud storage service has been accessed on these devices.

The identification of potential data stores is an area that can impede an investigation. The paper [15] 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 Big Data Platform.

In paper [11], the investigation was undertaken to determine the data remnants on a Windows and smart device (Apple iPhone) when a user undertakes a variety of methods to store, upload, and access data in the Dropbox.

On the knowledge of this writing, there are limited publications for preliminary investigation of Android device with the aim to conduct CDH Cloud forensics. This is the purpose of this paper is to contribute to this knowledge gap by investigating the Android client device which accessed to CDH.

3. Background

The following subsections discuss the overview of Hadoop Cloud, the popularity of Android device, why investigation of android is desirable for cloud forensics and the outline of CDH Cloud.

3.1 Overview of Hadoop

Hadoop is an open source project that seeks to develop software for reliable, scalable, distributed computing—the sort of distributed computing that would be required to enable big data.

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 the 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.

Hadoop has come to be an incredibly important technology for many big data projects and applications. Through the attractiveness of Hadoop, a number of companies became bundle Hadoop and some related technologies into their own Hadoop distributions. The three prominent Hadoop Platforms are MapR, Cloudera, and Hortonworks [12].

As Hadoop Cloud holds the sensitive data, it 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.

3.2. Android Device

Today smart mobile device market penetration and use of remote cloud services are all increasing. Cisco reports that Global mobile data traffic grew 63 percent in 2016 [2]. They also evaluated that global mobile data transmission traffic has grown 18-fold over the past 5 years [2]. Supporting these predictions, smart mobile devices have accessed cloud service providers with tremendous growth in the past years.

According to the Statista report "in the first quarter of 2017, 86.1 percent of all smart devices sold to end users were the Android devices" [6]. Android users all over the world, it shows that its popularity has no equal. In the era of cloud computing, android smart phone can be much more than a phone. It is fast raising popularity to access cloud services via Android devices. With the aim to conduct cloud forensics, the attached Android devices may store useful artifacts to investigate the illegal usages of cloud. That is why Android investigation is desirable to conduct cloud forensics.

3.3. Cloudera Distribution Hadoop (CDH)

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 [12].

4. Forensically Research Questions for Investigation of Android Device for Discovering CDH Cloud Storage Artifacts

In this paper, we conduct the initial investigation of locating artifacts on Android device that attached to CDH Cloud. This situation prompts research into the following questions:

Q 1. What artifacts can be found in Android device's storage after the user had accessed CDH Cloud via the popular web browsers?
Q 2. What artifacts can be found in Android device's storage after the user had accessed CDH Cloud via the private browsing?
Q 3. Which artifacts are resulted on Android device after undertaking the primary file operations (read, upload, download) on CDH Cloud?

5. Investigation of Android Device

This section provides an experimental environment and details the file operations undertaken on the experimental device. A preliminary forensic investigation is conducted for locating and documenting the residual artifacts that are left on Android device while accessing the CDH Cloud.

5.1. Experimental Setup

In this section, the experimental environment is set up for investigation of Android Device. The investigation scope of this paper is outlined in Figure-1. This investigation is intended to discover the residual artifacts on all storage layers of android device which has accessed to CDH Cloud via five web browsers (four popular web browsers and one private web browser). Private web browser keeps no local history the websites that are visited.

A variety of Android Emulators Virtual Machines (VMs) are created. Experimental VMs of Android are implemented using YouWave Android Emulator 3.31[16]. In order to perform a logical acquisition all experimental should be rooted before the investigation.

It is decided to examine a variety of circumstances of a user accessing CDH Cloud, and also to examine any differences when undertaking different type of file operations using different browsers.

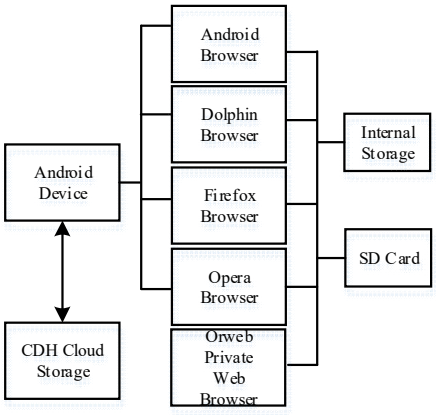

Figure 1. Block diagram for scope of investigation

Multiple Virtual Machines (VMs) of Android Emulator were created for each file operation to replicate different circumstance of usage. For each scenario, The Table-1 exposes the notable feature of Android VMs implemented in this experiment.

Table 1. Feature of experimental VMs

Feature	Software Components
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

This experiment sets up five VMs of Android Emulator. The different actions of VMs and their creation purposes are as the following;

(1) Clean VM - The first step is to install the browser software into Clean-VM's for each browser; Android Default Browser (AB), Dolphin Browser V-11.5.4 (DB) , Firefox Mobile Browser 4.4 (FMB), Opera Browser 28.0.1764.90386 (OB) and Orweb: Private Web Browser 0.7.1 (ORB).
(2) Access VM - These is used to access the CDH Cloud website using each installed browser. The sign-in option is used to log in to the user account.
(3) Read VM – In this VM, each installed browser will be used to access the CDH cloud website. The tested data-set file in the CDH is opened and read.
(4) Upload VM – In this VM, each installed browser will be used to access the CDH cloud website. The tested data-set file in the local storage is uploaded to CDH Cloud.
(5) Download VM – In this VM, each installed browser will be used to access the CDH cloud website. The tested data-set file in the CDH is downloaded to local storage of device.

Table-2 shows the summary configurations of VMs.

Table 2. Summary configuration of experimental VMs

Devices	Software Components
Clean VM	AB, DB, FMB, OB, ORB
Access VM	AB, DB, FMB, OB, ORB
Read VM	AB, DB, FMB, OB, ORB
Upload VM	AB, DB, FMB, OB, ORB
Download VM	AB, DB, FMB, OB, ORB

5.2. Residual Artifacts on Android Device

These experiments are performed by a series of read, upload, and download operations on the sample data-set. The important parts and files of each browser are listed in Table-3. The investigation of each operation on each experimental VM is exposed in the later tables. 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 cannot be traced in non-rooting the Android device, browser history and artifacts are able to be located in rooted Android device.

Table 3. Important parts and files of web browsers

Android Default Browser	
Data	Path
Cache	/data/data/com.android.browser/cache/webviewCacheChromium
History	/data/data/com.android.browser/databases/browser2.db
Cookie	/data/data/com.android.browser/databases/webviewCookiesChromiumPrivate.db
Password	/data/data/com.android.browser/databases/browser/webview.db/password
Dolphin Browser V-11.5.4	
Cache	/data/data/mobi.mgeek.TunnyBrowser/dolphin_webviewCache.db
History	/data/data/mobi.mgeek.TunnyBrowser/databases/browser.db
Cookie	/data/data/mobi.mgeek.TunnyBrowser/databases/browser/webviewCookiesChromium.db
Password	/data/data/mobi.mgeek.TunnyBrowser/databases/browser/webview.db/password
Firefox Mobile Browser 4.4	
Cache	/data/data/org.mozilla.firefox/cache/webviewCache
History	/data/data/org.mozilla.firefox/databases
Cookie	/data/data/org.mozilla.firefox/databases/webview
Password	/data/data/org.mozilla.firefox /databases/browser/webview.db
Opera Browser 28.0.1764.90386	
Cache	/data/data/com.opera.browser/cache/webviewCache
History	/data/data/ com.opera.browser /databases
Cookie	/data/data/ com.opera.browser/databases/webview
Password	/data/data/com.opera.browser /databases/browser/webview.db
Orweb: Private Web Browser 0.7.1	
History	/data/data/info.guardianproject.browser/webview.db

Firstly, we tested that "Access VM" access the CDH Cloud via the listed web browsers and then examined and analyzed. The residual artifacts are shown in Table- 4.

Table 4. Artifacts of web browsers for accessing CDH Cloud

File name	**Artifacts**
Android Default Browser	
databases/browser2.db/	- URL, web page , date
webviewCacheChromium.db/cookies	- URL, date
/data/data/com.android.browser/databases/browser/webview.db	- user name, password, host
/data/data /com.android.browser/cache /data.file (Hex file)	- URL
Dolphin Browser V-11.5.4	
/data/data/mobi.mgeek.TunnyBrowser/dolphin_webviewCache.db	- URL, date
/data/data/mobi.mgeek.TunnyBrowser/databases/browser.db	- URL, date
/data/data/mobi.mgeek.TunnyBrowser/databases/browser/webview.db/ password	- password
Firefox Mobile Browser 4.4	
/data/data/org.mozilla.firefox/databases	- URL, date
/data/data/org.mozilla.firefox/databases/webview	- URL, date
Opera Browser 28.0.1764.90386	
data/data/ com.opera.browser /databases	- URL, date
Orweb: Private Web Browser 0.7.1	
/data/data/info.guardianproject.browser/webview.db	- URL, date, user name

Secondly, a sample data-set (tested file) on CDH Cloud is read. This operation is tested on "Read VM" then examined and analyzed.

And then, we download the data-set from the CDH Cloud by testing on "Download VM".

Finally, the data-set is uploaded to the CDH Cloud from "Upload VM" and then examined and analyzed.

The residual artifacts which are left on the experimental Android devices while performing the operation are shown as the following Table-5.

Table 5. Artifacts of web browsers for primary file Operation

File name	Artifacts
Android Default Browser	
Read Operation	
databases/browser2.db/ histroy databases/browser2.db/ images	URL, web page, file name, Date
Download Operation	
webviewCacheChromium. db/cookies	- URL, date
/data/data/com.android.bro wser/databases/browser/we bview.db	- File name, directory
Upload Operation	
/data/data/com.android.bro wser/databases/browser/we bview.db	- File name, directory
Dolphin Browser V-11.5.4	
Read Operation	
/data/data/mobi.mgeek.Tu nnyBrowser/dolphin_webv iewCache.db	- URL, date
Download Operation	
/data/data/mobi.mgeek.Tu nnyBrowser/databases/ download/	- file name - downloaded - directory
Upload Operation	
/data/data/mobi.mgeek.Tu nnyBrowser/databases/do wnload/	- file name - directory
Firefox Mobile Browser 4.4	
Read Operation	
/data/data/org.mozilla.firef ox/databases	- URL, date
Download Operation	
/data/data/org.mozilla.firef ox/databases	- URL, date
Upload Operation	
/data/data/org.mozilla.firef ox/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.guardianpro ject.browser/webview.db	- URL, date, user name

6. Investigation with Example Crime Scenario

In this section, a crime scenario is presented and the investigation is conducted. An example crime scenario in CDH Cloud environment is described as the following. This scenario is extended the Digital Forensic Corpus "M57 Jean" crime case study [3].

Background:

The Company, DEF organization uses the services of CDH Cloud provided by third party. Every authorized person in this organization can access it from their PC and Android Device via web browsers. They use this for obtaining the service of uploading, downloading and opening the files on it.

Case: document exfiltration

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 that Mr. Felix, download the data from his Android device. The investigator has been given:

- Mr. Felix Android Device
- A copy of the targeted data-set file

Question to answer:

Did Mr.Felix commit the crime (downloading the file)?

6.1. CDH Cloud Forensic Investigation Framework

It is also common practice that a forensic framework be used to guide the investigation. In the context of our paper, we implement the CDH Cloud forensic investigation framework which has been proposed in my previous published paper "Forensic Analysis of Residual Artifacts on CDH Storage" [8].

This is, perhaps, the first digital forensic framework designed to be adaptable for CDH.

It comprises five phases; preparation, collection, analysis, and documentation and presentation, and closing.

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 of Collection phase.

1. Preparation: concerned 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

a. Preparation

Tools, techniques, research methodology, training, acquisition, and management support are prepared. The targeted devices are identified.

A forensic server is also arranged. The responsibility of forensic server is

- to extract the forensic data from targeted devices
- to store the forensic copy
- to mount these file and explore in read only mode
- to conduct investigation and analysis

Forensic imager tools and analysis tools are setting up on it.

b. Collection

In this phase that we collect the forensic data from identified devices.

This paper examines the approach of acquiring and extracting the history data from Android by Android Software Development Kit (SDK) and Android Debug Bridge (ADB) [13].

c. Analysis

In analysis phase, the located artifacts described in above section assist as the prior knowledge of which are the important parts and files for forensic analysis. The evidences found in log and metadata files are exposed through Figure- 2, and 3.

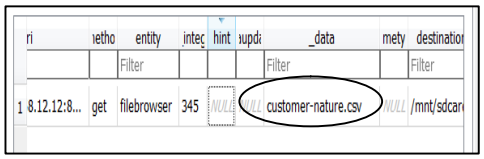

Figure 2: Artifacts in "databases. db/ download"

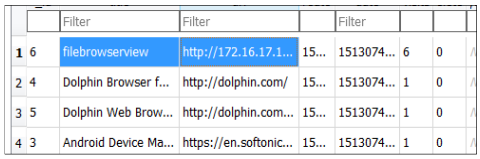

Figure 3: Artifacts in "databases.db/history"

d. Presentation

The investigator arranges the found evidences to embody the crime and reconstruct the event line. The Table-6 shows the forensics report for Investigation of Android for CDH forensics.

Table 6. Forensic Report of CASE 001/17

FORENSIC REPORT of CASE 001/17
INVESTIGATOR : Mr.Jean
Case Type : Suspect
Case Number : ####### **001/17**
1. Status: Complete
2. Summary of findings: To find the related information of "**customer nature.csv**". Step 1 : Finding the history and metadata of Android that are accessed CDH via web browsers Step 2: Extracting/parsing metadata directly off the device using analysis tools. Step 3: Examine the extracted data.
3. **Items Analyzed:**

Item #####	
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

5. Finding for item ####

i. The OS of examined device is Android OS.

-"**customer nature.csv**" and date are found in residual artifacts of Dolphin Browser mobi.mgeek.TonnyBrowser\

ii. The user Mr. Felix access the CDH Cloud at 10/Dec/2017

7. Items Provided: In addition to this hard copy report, one compact disk (CD) was submitted with an electronic copy of this report. The report on CD contains hyperlinks to the above-mentioned files and directories.

e. Closing

As the result of the following of report Mr. Felix stole the "customer nature.csv" by using the Dolphin web browser from his 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.

7. Conclusion

The usage of Hadoop Cloud is becoming more widespread. The fast raising popularity of accessing Hadoop Cloud services via Android device could lead to an increase in the rate of crime. Consequently, the Hadoop Cloud is identified as an emerging challenge to digital forensic researchers and practitioners. Hadoop Cloud environment managed by a third-party is potentially difficult to investigate. The artifacts extraction from cloud server may be restricted by the service level agreement. It is anticipated that the attached Android device may store useful artifacts to investigate the illegal usages of Hadoop Cloud. This paper utilizes Cloudera Distribution Hadoop (CDH); a popular Hadoop Cloud. This paper conducts a preliminary forensic investigation of Android device which attached to the CDH Cloud. It locates the residual artifacts left on Android device when the user has performed the primary operations. In the preliminary investigation, the investigation is tested with five Android VMs; each of which accesses the CDH Cloud via four popular web browsers and a private web browser of Android. This paper conducts the preliminary investigation on Android clients of CDH Cloud and highlights the residual artifacts.

The extracted artifacts resulting from the initial investigation can assist the forensic examiners for generating of evidences in real world Hadoop Cloud forensics. The crime scenario which is extended the Forensic Copra's crime case is examined under the guide of CDH Forensic Investigation Framework.

References

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

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021," [online] Available: https://www.cisco.com/service providers/ , 2017, [Accessed: 9 -Dec- 2017].

[3] "DigitalCorpora- M57-Jean Scenarios," [online] Available: https://digitalcorpora.org/corpora/scenarios/m57-jean, [Accessed: 9 -Sept- 2017].

[4] "FileViewerPlus," [online] Available: http://fileviewerplus.com.siterankd.com, [Accessed : 9 -Sept- 2017].

[5] "Global Hadoop Market Share Will Hit USD 87.14 billion by 2022 : Zion Market Research," [online] Available: https://www.zionmarketresearch.com/inquiry/hadoop-market [Accessed :1 -Dec- 2017].

[6] "Global Market Share Held By the Smartphone, " [online] Available: https://w5ww.statista.com/statistics/266136/global-market- share-held- by- smartphone- operating systems [1 -Dec- 2017].

[7] "HHNeoEditor," [online] Available: http://neo.com [Accessed: 1 -Dec- 2017].

[8] 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, 2017.

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

[10] D.Quick and K.-K. R. Choo, "Digital Droplets: Microsoft SkyDrive Forensic Data Remnants," Future Generation Computer System, vol. 29, no. 6, pp. 1378–1394, 2013.

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

[12] "Report of HadoopBigDataDistribution," [online] Available: https://www.forrester.com/report/The+Forrester+Wave+Big+Data+Hadoop+Distributions+Q1+2016, 2016 [Accessed : 9 -Dec- 2017].

[13] "SDK-Platform-Tool and ADB," [online] Available: http://www.teamandroid.com/2017/05/06/download-adb-fastboot-android-sdk-platform-tools, 6. May, 2017 [Accessed : 9 -Sept- 2017].

[14] "SQLiteDB Browser," [online] Available: http://sqlitebrowser.org [Accessed : 9 -Dec- 2017].

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

[16] "YouWave for Android 3.3.1," [online] Available: https://www.youwave.com/downloads[Accessed : 7 -Sept- 2017].