I J C T A, 9(18) 2016, pp. 8927-8933

A Study on Digital Forensics in Hadoop

Sachin Arun Thanekar[*], K. Subrahmanyam[*] and A. B. Bagwan[**]

ABSTRACT

Nowadays we all are surrounded by Big data. The term 'Big Data' itself indicates huge volume, high velocity, variety and veracity i.e. uncertainty of data which gave rise to new difficulties and challenges. Hadoop is a framework which can be used for tremendous data storage and faster processing. It is freely available, easy to use and implement. Big data forensic is one of the challenges of big data. For this it is very important to know the internal details of the Hadoop. Different files are generated by Hadoop during its process. Same can be used for forensics. In our paper our focus is on digital forensics and different files generated during different processes. We have given the short description on different files generated in Hadoop. With the help of an open source tool 'Autopsy' we demonstrated that how we can perform digital forensics using automated tool and thus big data forensics can be done efficiently.

Keyword: Big Data, MapReduce, Hadoop, HDFS, Digital Forensics

I. INTRODUCTION

Hadoop is a framework which can be used for tremendous data storage and faster processing. Moreover, it is an open-source technology. It uses large clusters of commodity hardware and thus uses distributed computing. To perform Big Data forensic investigations, the knowledge of Hadoop's internals and architecture is essential [1], [3], [4], [12-15].

1.1 The Hadoop configuration files [5-8]

The following are the Hadoop's standard configuration files,

- **hadoop-default.xml**
  General default system variables and data locations are stored in this file.
- **job.xml**
  Job-specific configuration parameters are stored in this file.
- **mapred-default.xml**
  All the MapReduce parameters are stored in this file.
- **hadoop-site.xml**
  The site-specific version of hadoop-default.xml.

1.2 The Hadoop other supporting files [2],[9]

- **Trash files**
  After the deletion of any file a subdirectory is created under the user's $HOME folder using the original file path, and the file is stored there. These files are deleted only after periodic trash deletion process run by Hadoop. User can configure the period for this.

[*] Department of Computer Science and Engineering, KL University, Vaddeswaram, Guntur District, Andhra Pradesh, India, *E-mails: sachin.thanekar@yahoo.co.in; smkodukula@kluniversity.in*

[**] Department of Computer Engineering, RSCOE, Tathwade, SSPU Pune, Maharashtra, Pune, *E-mail: abbagwan@gmail.com*

- **Log files**
  The most important files are Log files in forensic evidence. Almost every information like operations, input source, users, jobs, links, different locations etc. can be easily achieved from log files. This is indirect kind of information. Following types of logs are generated by Hadoop,

  a. Standard out and standard error
  Each Hadoop TaskTracker creates and maintains these error logs to store information written to standard out or standard error. These logs are stored in each TaskTracker node's /var/log/hadoop/ userlogs directory.

  b. Hadoop daemon logs
  This file is stored in the host operating system which contains error and warning information.

  c. Job configuration XML
  The Hadoop JobTracker creates these files within HDFS for tracking job summary details about the configuration and job run. These files can be found in the /var/log/hadoop and /var/log/hadoop/ history directory.

  d. log4j
  This file is the outcome of log4j process. The log4j application is an Apache logging interface that is used by many Hadoop applications. These logs are stored in the /var/log/hadoop directory.

  e. Job statistics
  The Hadoop JobTracker creates these logs to store information about the number of job step attempts and the job runtime for each job.

2. HADOOP FORENSICS EVIDENCE ECOSYSTEM [5], [11]

Evidence is the base of Forensics and data is the evidence in digital investigations. As hadoop stores lot of data on disk and in memory as well this data provides information for evidence. Sometimes this information may be useful or may not be useful. It totally depends on the scenario and needs. Even the definition of evidences is also changing as per different scenarios and needs. As shown in Figure 1 there are three categories of Forensic data in Hadoop:

a. Supporting information
This is not a direct information rather the information which can be used to identify evidence or may provide some clue about the operations or configurations.

b. Record evidence
This is the direct information. Any data analyzed in Hadoop comes under this category. E.g. Text files for MapReduce jobs.

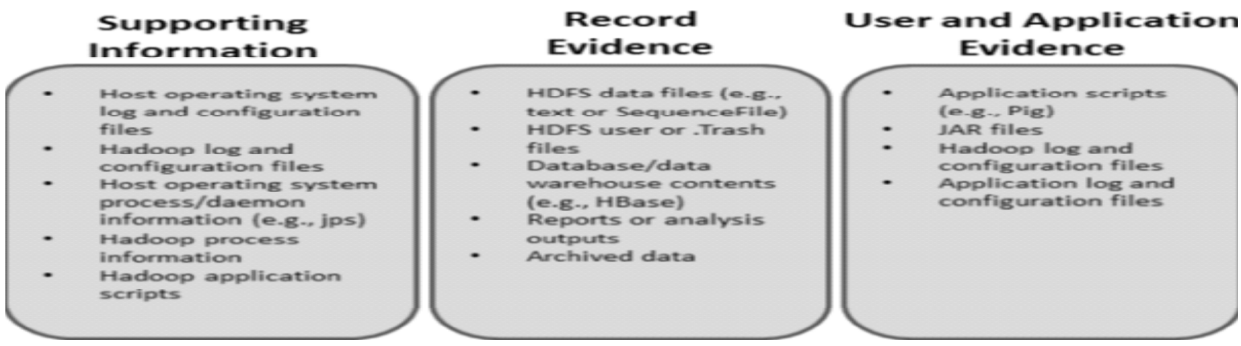


Figure 1: Types of forensics evidences

c. User and application evidence

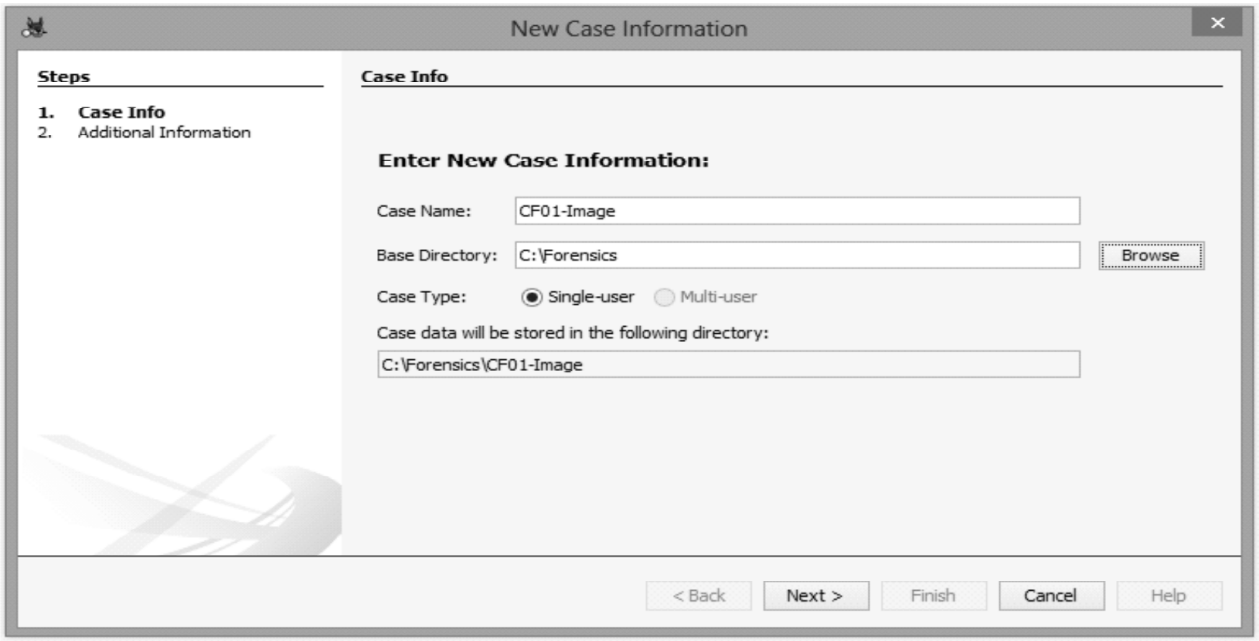
This evidence includes the configuration files, log files, analysis scripts, metadata, MapReduce logic and any change / logic that performed on data. It gives information about how the data was analyzed or generated.

3. AUTOPSY

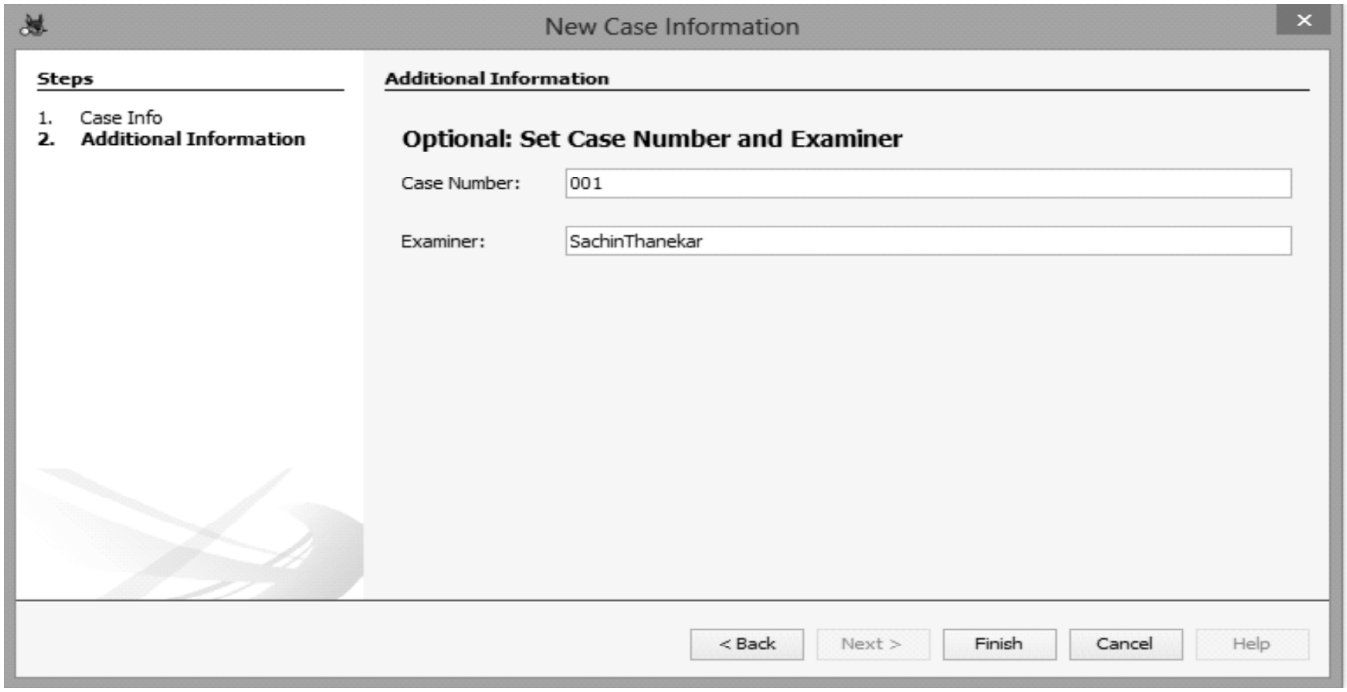
Autopsy is a freeware forensic tool. Efficient file carving, data carving, keyword searching are the main features. It provides good GUI that's why it is widely used in forensic investigations. Following are the steps for file or data carving [11],

1. Download and install the latest version of Autopsy from, http://www.sleuthkit. org/autopsy/.
2. Start Autopsy. Select New Case.

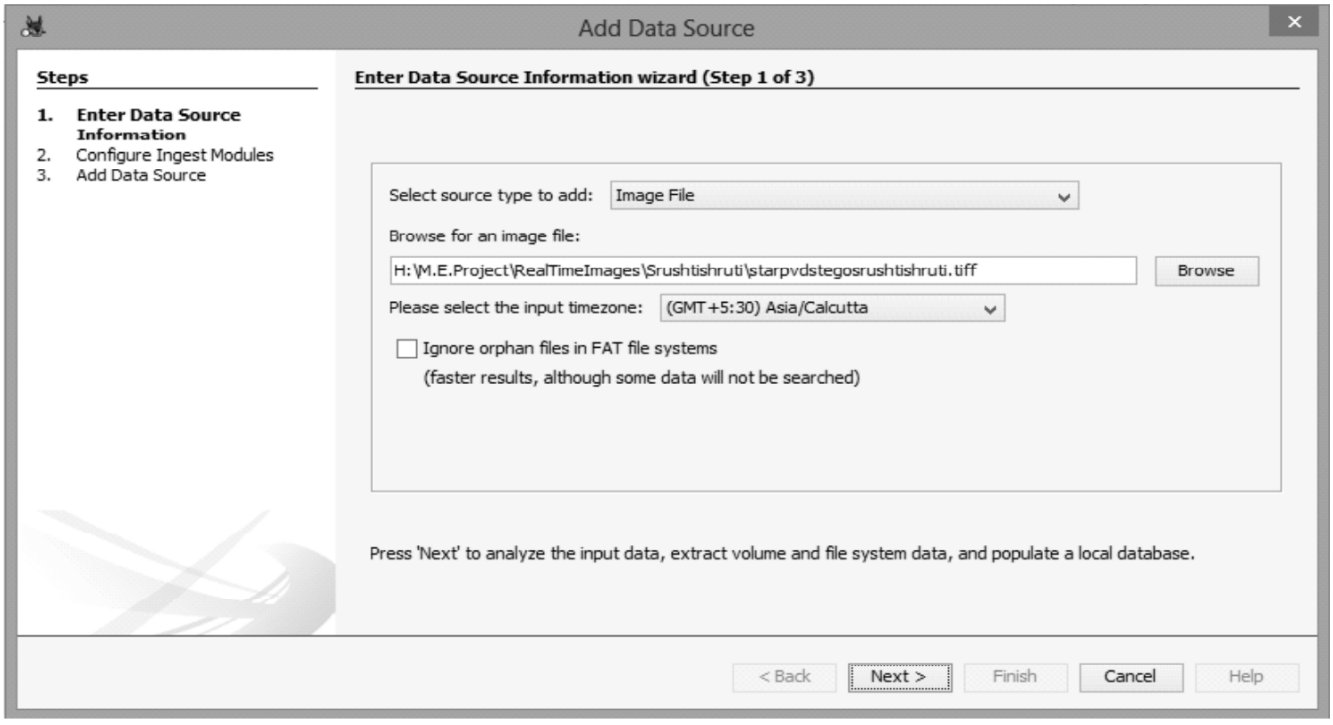
3. Enter the Case Name and Base Directory. Output will be stored in this directory. Click on next.



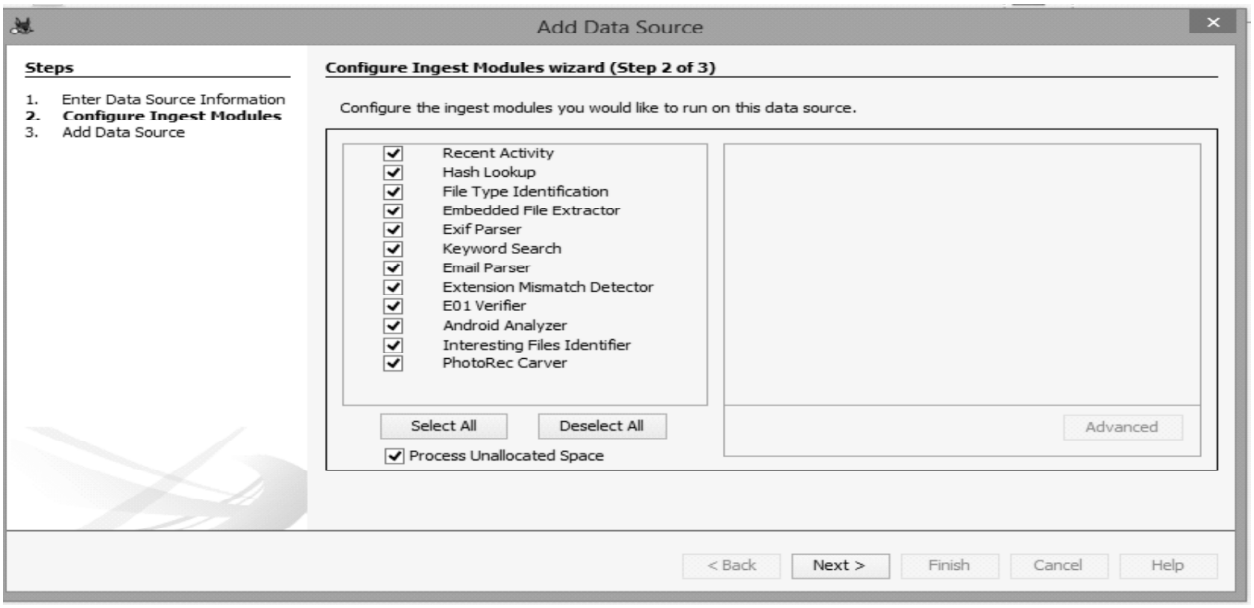
4. Enter the Case Number and Examiner information, and click Next. Here case number should be numerical sequential number like 001.



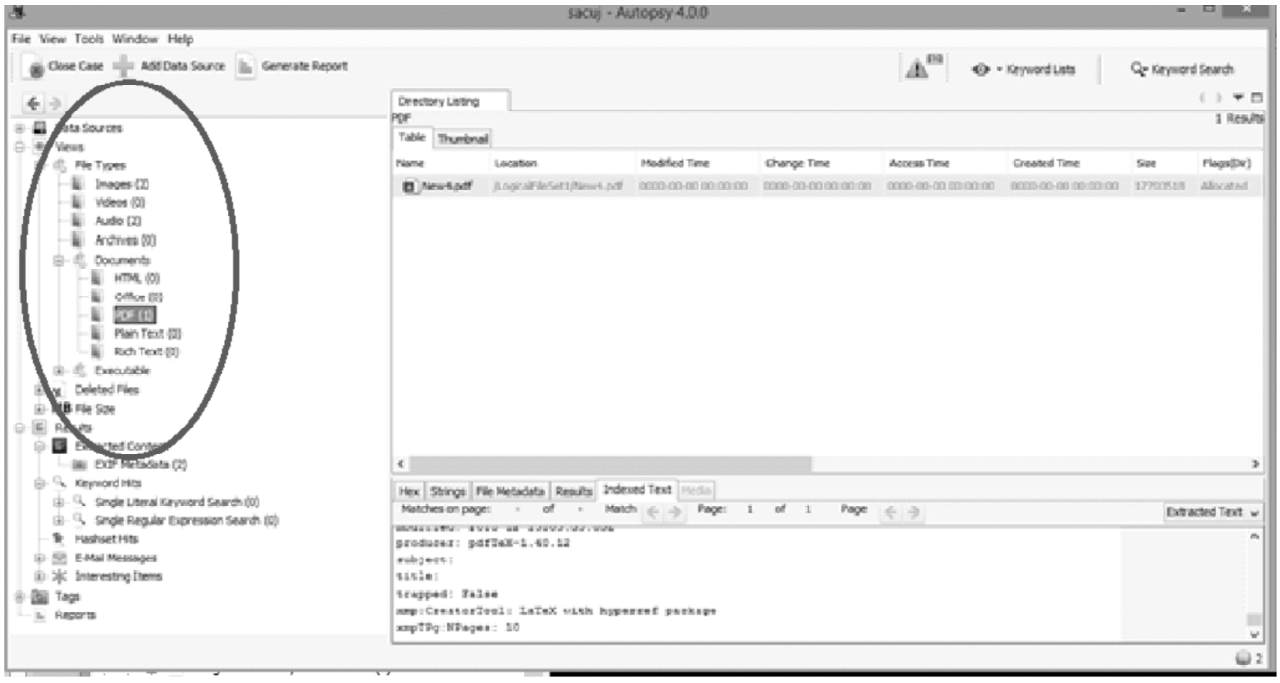
5. Thus new case is opened. To add evidence, click on Add Data Source. Select either Image File or Logical Files, and enter the path to the image file or logical files. Click Next:



6. Select all ingest modules that apply to the investigation. For testing purposes, select all options and ensure that Process Unallocated Space is checked. That provides slack space analysis capabilities, including deleted file recovery:



The investigator can view picture, audio, video, document and other file types from the source file system and slack space. The investigator can also view the files under the Views menu. Several different types of views are provided like File Types, Recent Files, Deleted Files, MB File Size (grouped by file size) etc.

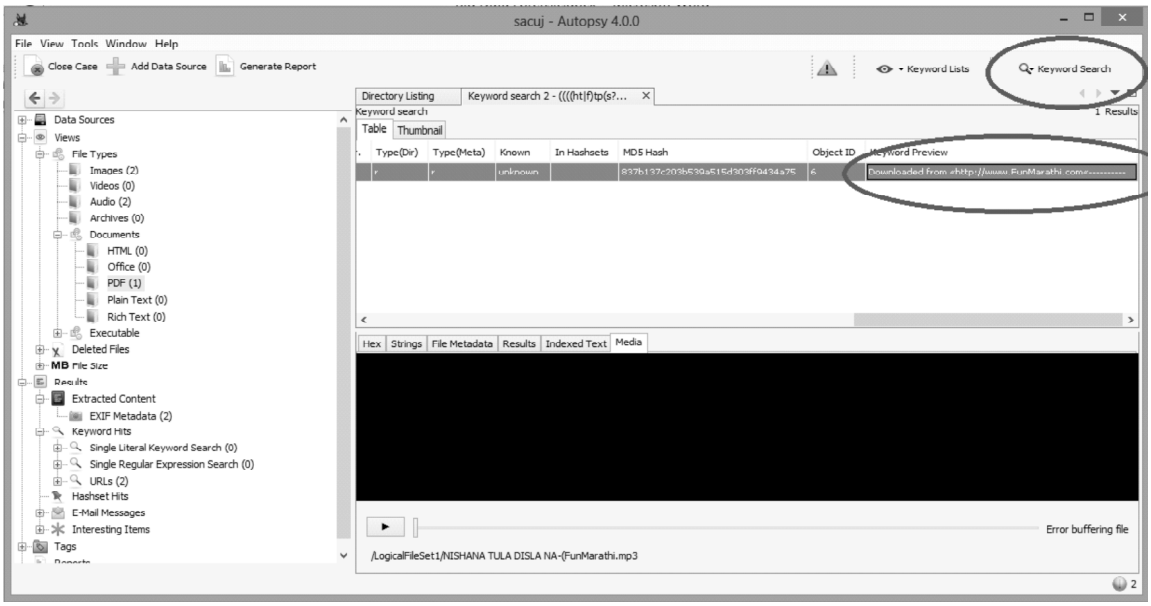


Investigator can extract files to required location by right-clicking on the file and selecting Extract File(s). Any type of file can be extracted using this method no matter whether it is deleted or present. Files can also be tagged in Autopsy for later extraction or further analysis. The investigator can right-click on a selected file and select Tag to save the file with tag information for later analysis.

Autopsy also provides keyword searching capabilities. In the top-right corner, Autopsy has the Keyword Search and Keyword Lists menus. Click on Keyword Search to run a one-time search. The search can run the following three types of searches:

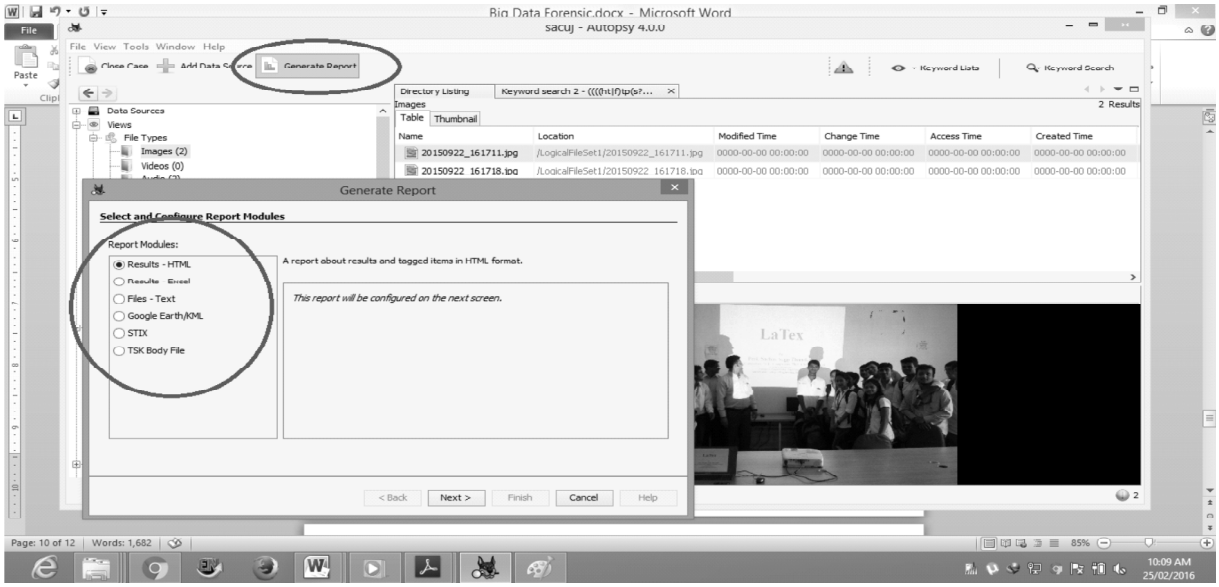
- An exact word match
- A substring match
- A regular expression match

When the search is run, Autopsy searches all extracted text from the evidence and returns the list of files with one or more matches and all associated metadata. Multiple search terms can be searched in a batch via the Keyword Lists menu. Clicking on Keyword Lists | Manage Lists brings up the search settings, where the investigator can add a list of multiple search strings. A list is created and saved after all terms have been entered. Then, the investigator can search the evidence using that list by clicking on the dropdown menu next to Keyword Lists and selecting the list to use. The results of keyword searches can be exported as a report file that shows the file that matched and the search term.



With the help of Autopsy we can generate results by following way,

1. Click Tools | Generate Report.
2. Select Results | Excel or Results | HTML.
3. Select All Results or Tagged Results, and click Finish. A report file is generated that displays which files match the search terms.



4. CONCLUSION

To find the evidences on big data we need to understand the basic structure of the big data, through which we can find the evidences more efficiently. By using the different tools and technology we can do the forensic of big data. As in big data the volume of data to be considered is very big, automated tool can help us to do it efficiently. We have shown it with the help of 'Autopsy'.

REFERENCES

[1] Intel, "Planning guide: Getting started with big data", http://goo.gl/ jyU73v, Tech. Rep., 2013.

[2] K, Chitharanjan, and Kala Karun A, "A review on hadoop — HDFS infrastructure extensions", *JeJu Island: 2013*, pp. 132-137, 11-12 Apr. 2013.

[3] S. Sagiroglu and D. Sinanc, "Big data: A review," *International Conference on Collaboration Technologies and Systems (CTS). IEEE, 2013*, pp. 42–47, 2013.

[4] S. Madden, "From databases to big data", *IEEE Internet Computing*, vol. 16, no. 3, pp. 4–6, 2012.

[5] Joe Sremack, "Big Data Forensics-Learning Hadoop Investigation".

[6] www.tutorialpoint.com/hadoop/hadoop_big_data_solution.html.

[7] www.edureka.co/blog/hadoop-cluster-configuration-files/

[8] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed ûle system," 26*th IEEE Symposium on Mass Storage Systems and Technologies (MSST). IEEE, 2010*, pp. 1–10, 2010.

[9] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas, and J. McPherson, "Ricardo: integrating r and hadoop", ACM *SIGMOD International Conference on Management of data. ACM, 2010*, pp. 987–998, 2010.

[10] www.archive.ahrq.gov/download/pub/evidence/pdf/autopsy/autopsy.pdf

[11] S. Garfinkel, Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus, Digital Investigation, 9, 2012, S80–S89, 2012.

[12] S. Gowri, G. S. Anandha Mala, "Effective Retrival of Data from E-mail Corpus for Digital Investigations", *Indian Journal of science and technology*, Volume 8, Supplementary Issue 9, May 2015.

[13] Sungjin Lee, Sunghyuck Hong, "Analysis of Time Records on Digital forensics", *Indian Journal of science and technology*, Volume 8, Supplementary 7, April 2015.

[14] Deevi Radha Rani, S. Venkateswarlu, " Security against Timing Analysis Attack ", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 5, No. 4, pp. 759~764, August 2015.

[15] Hamid Bagheri , Abdusalam Abdullah Shaltooki, "Big Data: Challenges, Opportunities and Cloud Based Solutions", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 5, No. 2, pp. 340~343, April 2015.