A New Technique for File Carving on Hadoop Ecosystem

Esraa Alshammari, Ghazi Al-Naymat, Ali Hadi
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
israa88h@gmail.com, g.naymat@psut.edu.jo, a.hadi@psut.edu.jo

***Abstract*— using file carving techniques is one of most recent techniques that is used to retrieve the important data from unallocated space in a corrupted file system. In the traditional operating systems, such as Windows or Linux that have a small size of hard disk to store data, the researchers implemented many file carving techniques to carve a specific type of files (e.g. PDF, JPIG... etc.). However, with the presence of a specially designed file system that stores a huge volume of data, namely Hadoop Distributed File System (HDFS), the carving techniques should be established to recover the minimum amount of data corrupted by attackers considering the HDFS capabilities. In this research, we propose a framework, which encompasses different steps working together as a new carving technique to retrieve the most possible pieces of files that are corrupted by 10%, and to ensure that the files are successfully carved.**

***Keywords*—**Carving Technique, Data Recovery, File System, Hadoop Ecosystem.

I. INTRODUCTION

The carving techniques are widely used in digital forensics and digital investigation due to the importance of file carving in retrieving the minimum amount of data stored in an unallocated space or stored in a corrupted file system.

File carving is the process of collecting data or extracting large dataset from unallocated space in any file system. The carving techniques conduct in every digital investigation in order to carve every data that can be useful in the investigation process and used in court as an evidence [1, 2].

The process of carving any files is based on the type of methods that are used to carving file from unallocated space. These types are as follows [3]:

- Metadata carving based method: the carving depends on the existing of metadata even if the file system is corrupted.
- File Header based method: the carving process depends on the header of the file, which includes the unique identifiers that identify the file type and the beginning of the file
- Header-Footer based method: uses the header or the identifier of the beginning of the file and the footer that identifies the tail of the same file. This carving method extracts the blocks of space between the header and the footer and considers that as one file.
- File Structure Carving based method: an advance carving technique used to recover a fragmented file that has not header or footer identifiers

In a traditional file system, the carving techniques are used to recover a specific type of data stored combined with its metadata which is provided by the file system itself. Then, the investigators will use these files to present to the court as evidence which will affect the judgment for the case. In contrast, when the distributed file system (HDFS) appeared and became widely used, the huge data that stored within such system will be used as a good evidence in any digital investigation due to its huge volume if it isn't corrupted.

Hadoop [4] is the new trend of distributed file systems that is used to store, access and analyze data that is large in volume, variety, and velocity). The HDFS ecosystem is a combination of the Hadoop file system and other complex and reliable components or projects to provide an ecosystem that are employed to solve problems. Hadoop ecosystem is alienated into main four layers, these layers are data storage, data processing, data accessing, and data management. Also, some researchers combine the data processing and data accessing into one layer as data processing layer [5-8].

- Data Storage: this layer is responsible for storing the data in the distributed file system. The main component of this layer is HDFS; which is the Java - based file system that stores a huge volume of data. And HBase; which is Apache - based that support the large table to store data.
- Data Processing: this layer is responsible for processing, managing and scheduling the data, resources, and clusters. Also, this layer has two main component, Map Reduce, and YARN. The Map Reduce is the big data processing framework proposed by Google that is used to process a huge dataset. While the Yet Resource Negotiator (YARN) is another framework for scheduling and managing the resources, and Map Reduce used YARN parallel processing mechanism on the huge dataset.

978-1-5386-0527-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICTCS.2017.16

- Data Access Layer: the layer that provides a simple access to the data. In this layer, there are many tools used to accessing the data. Hive is an example of the tools that are used in this layer to provide a summarization of the data. Another example is Pig; it is high-level language for parallel computation. In addition, Mahout is a machine learning, and data mining library.
- Data Management Layer: the contact layer between the system and the user. There are many tools in this layer that the user may use easily, Chukwa; is a system that responsible for collecting data, Zookeeper is a coordinator services for distributed application, to mention a few.

In a digital crime scene, the data stored in any file system is the most important evidence that the investigators search for to be presented to the court. Hadoop file system is the same as other file systems since it can store any type of files, such as (PDF, CSV and JPEG files). Carving JPEG files as digital evidence in the court is very useful evidence that may help the court to take a decision in the crime case [9].

JPEG file is the widest compression formats that used in digital cameras. This type of format depends on two type of compression format; JPEG File Interchange Format (JFIF), and Exchangeable Image File Format (EXIF). Each JPEG file has a unique magic value (Start of Image SOI and End of Image EOI) that help the operating system to determine the type of file. These markers not only help the operating system, but also help the investigators to retrieve the image based on it, and it's referred as the header and footer of the image [13].

In additional to that, the JPEG file format consists of other important markers that are used as identifiers, and help to identify other useful information such as the used compression format. These markers are shown in Table 1 [3].

Table 1: Important Markers in JPEG File Format [3].

Name	Byte Value in Hexadecimal	Description
SOI	0xFF 0xD8	Start of Image
APP0	0xFF 0xE0	JFIF metadata
APP1	0xFF 0xE1	ExIF metadata
DHT	0xFF 0xC4	Define Huffman Table
COM	0xFF 0xFD	Comment
EOI	0xFF 0xD9	End of Image

The contribution of this research is to implement a new carving technique on a specially designed file system, namely Hadoop ecosystem. The proposed carving framework will cover the main two scenarios related to carving JPEG files, one that undertakes into consideration the presence of file system metadata and headers, and the other supposes the absence of the file system metadata and headers to recover the minimum number of fragments related to the same JPEG file.

The rest of this paper is organized as follows: Section 2 presents the related work. The proposed framework will be presented in Section 3. Finally, Section 4 present the conclusion of the research.

II. Related work

In last decade, many carving techniques have been implemented and developed to be used in digital investigation and digital forensics. These techniques are varying based on the type of files used, such as focusing on the whole forensic image (like hard disk, USB), file system (like HFS), database (like SQL), text files like (Doc, PDF), or multimedia file (like video, JPEG) which are the most important files in any digital forensics case. In this section, we shed light on some of the important work performed in this area

1. Metadata Based

In term of metadata, the carving techniques rely on the useful information that each file system provides which is metadata for each type of file. This information presents a basic information about any type of files such as creation and modification date, file size, location, the owner of the file and other useful information. In [10], the authors presented a method for carving a JPEG file based on the thumbnail that the file system provides besides the metadata while browsing the image. The first step of this method is extracting the thumbnails from a hidden file (thumb.db) and then enlarges the image based on the header information for the original image and reordering / reassembling the fragments to obtain the original image.

Other techniques that used the thumbnail to recover the files is mentioned in [11] which used a unique hex pattern (UHP) to recognize the thumbnails or embedded JPEG files. The method is used the (UHP) pattern to developed a tool called PattrecCarv that automatically distinguishes the JPEG files. The authors implemented a fixed pattern to determine if the image is a thumbnail in a JFIF format, EXIF format, or embedded. The process of PattrecCarv tool is based on searching for the start of image (SoI), searching for the embedded image header; if they found the header, they determine that it is an embedded image, else they are searching for thumbnail header.

2. File Header/Footer and Structured Based

Rather than using thumbnails to recover a JPEG file that limits two cases: first is the existence of file system, second is the discontinuity of the fragments will occur between lines, the authors used other techniques that can handle the distributed of the fragments and the corruption of file system that will delete the metadata.

The authors in [12] proposed a reconstruction and recovery system that takes into consideration the realistic and the complexity of fragmentation ignoring the type of files in a real-life scenario. The Progressive Joint Carver (Pro-joint carver) system extracts the raw data and reorders the fragments if it is unordered in the correct sequence, and then uses the header that are obtained from the raw data as a road map to linked all fragments with each other for the specific file. The system used a weight computation and path-based sorting for the linking process to link the fragments with its proper header. The result shows that the Pro-Joint Carver is achieved a higher efficiency and accuracy comparing with Adroit Photo Forensics system. Even more, the system caned recover the deleted fragments.

Related to the fragmentation issues, the authors in [13] considered JPEG file format using bit sequence pattern matching to recover a JPEG file fragments. They addressed the most important issue related to the fragments which is the missing of fragment header that will cause a problem to the decoder to determine the linked fragments and they proposed their methods to solve the problem. Furthermore, they proposed the construction of pseudo header for the fragments that have a missing header or a stand-alone fragment. In addition to that, the authors used the starter markers that provided and found in the header of the fragment for detection and recovery after bit-stream errors to recover the disrupted fragments.

Another technique that deals with missing a fragments header in JPEG files presented in [14]. The authors proposed an advance JPEG carving method that studied the structural characteristics of the JPEG file and recover the fragments when the headers and/or the markers of the fragments are missing. They used the characteristics to decompress the incomplete file and determined the parameters to produce a meaningful image by reconstruction the headers of an orphaned fragment. They obtained a large number of JPEG files that were available on Flickr photo sharing website and studying the characteristics of a different digital camera to extract information about the parameters and settings that the digital camera used. The results show the reliability of defining the decoder setting for fragments of size 4 KiB or above. In additional to that, they tested the efficiency of the proposed carving techniques using the JPEG files that have the Adobe Photoshop marker in their headers, and they found that the proposed techniques will not be affected.

In addition to the fragmentation issues that dealing with the heavily fragmented JPEG file, the authors in [15] presented carving technique to identify the pattern of file fragments distributed and segmented in the memory. They used a statistical byte frequency program to retrieve patterns by implementing some statistical parameter to recognize a unique pattern for each JPEG files. The authors proposed 17 clustering attributes used in the statistical byte frequency program to link the fragments with each other to obtain a proper JPEG files. The limitation of this method is that they are evaluating the proposed method only for four type of JPEG files.

In the same topics, the authors in [16] provided a framework that solves the way of reordering the fragments with missing/or not enough of decoding pieces of information. The framework consists of two main component: the new similarity metric to determine if there are two blocks sequentially related to the same JPEG file, and the other one is the algorithm that is used to recover the whole fragments for the JPEG file. Additionally, they used a new similarity measure to identify fragmentation points more accurately, which is the robustness of the algorithm. Finally, the proposed algorithm have two limitations, one is the speed of the algorithm, and another is that if the image has a sharp change in the content, then the fragmentation points will changes too, so the algorithm can't deal with such case.

By reviewing the literature, despite the importance of carving techniques for any user's data stored, however, it does not exist in Hadoop file system or any distributed file system. Even more, there is no literature shows how the importance of using the data that the users stored as an evidence to solving any digital forensics investigation [17, 18].

There are a few researchers focused on the digital forensics investigation in distributed file system, most of them concentrate to the cloud environment which is very similar to the Hadoop ecosystem in term of the four Vs (Volume, Variety, Velocity, and Veracity.)

The authors in [19] proposed a hybrid or an iterative digital forensics framework based on the main two frameworks in the literature. The proposed framework presents and ensures the difference in preserving data and acquiring or collecting it from the cloud environment for investigation purpose. They are focusing on the combination of identification and preserving of the evidence from the cloud environment and proposed a collection stages for the challenges that faced the cloud environment in term of the time-consuming process.

In addition, the authors in [20] utilized an open source software to studying the structure of the cloud client devices to extract the evidence. They addressed some challenges related to acquiring the evidence from the

cloud environment, such as acquiring the logs from the cloud servers or cloud service provider, which is not easy to obtain, or losing the data in the cloud, which will need the carving techniques to obtain the minimum amount of data in order to presented to the court in a criminal case.

Related to the cloud forensics, the authors in [21] presented a case study using ownCloud an open source tool, and conducted experiments for the storage as a service (StaaS) cloud computing to provide an in-depth understanding of the artifacts and evidence to undertake a cloud forensics. They focused on the client and server artifacts, which they determine as useful evidences. Their experiment presented a technical summary of artifacts type and categories provided to the practitioners. They also provided guidelines and recommendation for future direction in the StaaS products

In contrast to the cloud environment, the authors in [22] proposed a case study in Hadoop file system. They focused on digital forensics and the files that the HDFS generated during the different process. They also provide a description for these files using open source tools namely Autopsy, which is a forensic purpose tool that helps the investigators to search and collect evidence in a specific file system. Additionally, they provide the capability of using an automated forensics tools in a distributed file system, which is Hadoop.

To the best of our knowledge, there are two main differences between our proposed method and all techniques in the literature: First of all, the use Hadoop file system itself, which is the newest trend of distributed file system that stores a huge amount of data across a clusters or number of data nodes efficiently. Second is the fact that there is a lack of research that focuses on such file system in forensics and digital investigation. Even more, there is no research in literature put up the carving technique for users' files that stored in Hadoop to retrieve it and used in the digital forensics and investigation.

III. THE PROPOSED FRAMEWORK

The proposed framework (Fig. 1) presents all steps that should be followed in order to carving JPEG file in Hadoop ecosystem.

Step 1: The input of the framework is raw data taken from Hadoop storage media.
Step 2: Search inside raw data for a proper structure (File System), which provides metadata.
Step 3: If the structure is found:

- Then the road map for each file will be used to obtain all fragments for the same file.

Step 4: Else, if the structure (file system) is corrupted, even the metadata will be corrupted:

- Then look for the raw data once again and search for more information (Step 5-9).

Step 5: Search for a JPEG file header
Step 6: Search for JPEG file footer
Step 7: If JPEG file header and footer found,

- The retrieving method will extract all fragments between the header and the footer
- Considering the file that retrieved as one file.

Step 8: else, only the header of JPEG file is found,
Step 9: calculate the 7 statistical features for each fragment to group them based on the similarity.

These features are: mean, median, and mod attributes, standard deviation, variance, standard deviation frequency, and entropy. We focused currently at only 7 features, but we may consider more statistical feature later on. It should be also noted that this framework is proposed as part of a master degree and this paper demonstrates a work on progress findings.

IV. CONCLUSION

The carving technique is the art of retrieving files regardless of its type. In digital forensics and investigation, the need for carving and retrieving the users' files are very important due to the significance of the evidence that will be obtained and delivered to the court. Additionally, the huge volume of data will help the investigators and the judicial system to take a proper decision based on the evidence obtained from the big data system. In this study, the focus will be on Hadoop which is a specially designed file system that stored the huge amount of data across different data nodes, with variety and velocity of these data. Data on Hadoop can be treated as the data stored on traditional file systems like Windows or Linux, where investigators can use the evidence stored to support the decision. In this research, a new carving framework was proposed for Hadoop ecosystem to retrieve the minimum amount of data that was corrupted intentionally by attackers with the corruption of the file system itself. To the best of our knowledge, there is no other researchers mentioned the carving or retrieving method for users' files, using Hadoop ecosystem, to be used in the digital forensics and investigation.

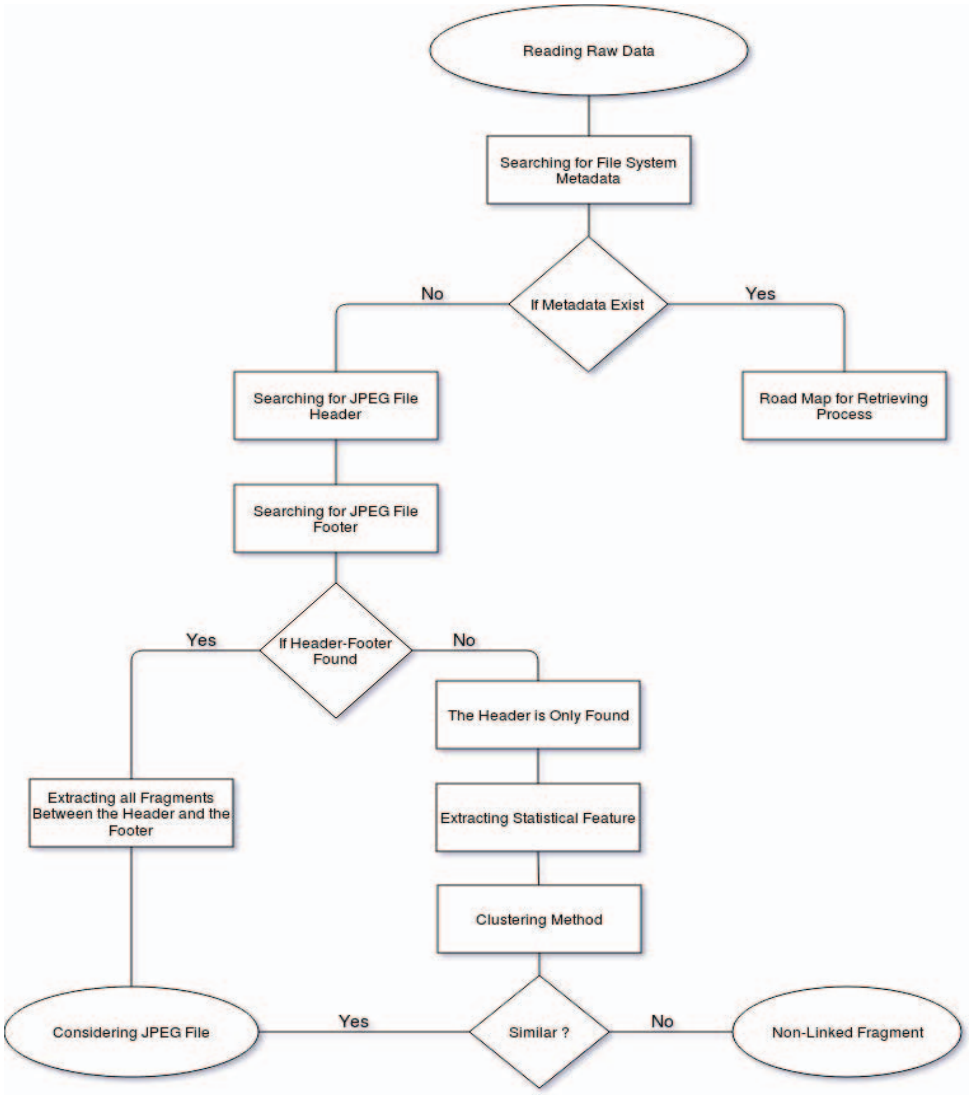

Fig1: The main steps for the carving method

REFERENCES

[1]Principal Security Consultant, "Introduction to File Carving", McAfee ® Foundstone ® Professional Services, 2011.

[2]R. Poisel and S. Tjoa, "A Comprehensive Literature Review of File Carving", 2013 International Conference on Availability, Reliability and Security, 2013.

[3]E. Alshammary and A. Hadi, "Reviewing and Evaluating Existing File Carving Techniques for JPEG Files - IEEE Xplore Document", Ieeexplore.ieee.org, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7600210/. [Accessed: 09- May- 2017].

[4]"Apache™ Hadoop", Apache™ Hadoop, 2017. [Online]. Available: http://hadoop.apache.org/]. [Accessed: 14- May- 2017].

[5]S. Mehta and V. Mehta, "Hadoop Ecosystem: An Introduction", International Journal of Science and Research (IJSR), vol. 5, no. 6, pp. 557-562, 2016.

[6]J. Dittrich and J. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce", Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2014-2015, 2012.

[7] D. P and A. Raman G R, "A Study of Hadoop - Related Tools and Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 9, pp. 160-164, 2015.
[8] S. Landset, T. Khoshgoftaar, A. Richter and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal of Big Data, vol. 2, no. 1, 2015.
[9] N. Alherbawi, Z. Shukur and R. Sulaiman, "A Survey On Data Carving In Digital Forensics", Asian Journal of Information Technology, vol. 15, no. 24, pp. 5137-5144, 2016.
[10] H. Guo and M. Xu, "A Method for Recovering JPEG Files Based on Thumbnail", 2011 International Conference on Control, Automation and Systems Engineering (CASE), 2011.
[11] M. Sarfraz, M. Hussain and M. Ishaq, "Carving Thumbnail/s and Embedded JPEG Files Using Image Pattern Matching", Journal of Software Engineering and Applications, vol. 06, no. 03, pp. 62-66, 2013.
[12] V. Thing, T. Chua and M. Cheong, "Design of a Digital Forensics Evidence Reconstruction System for Complex and Obscure Fragmented File Carving", 2011 Seventh International Conference on Computational Intelligence and Security, 2011.
[13] H. Sencar and N. Memon, "Identification and recovery of JPEG files with missing fragments", Digital Investigation, vol. 6, pp. S88-S98, 2009.
[14] E. Uzun and H. Sencar, "Carving Orphaned JPEG File Fragments", IEEE Transactions on Information Forensics and Security, vol. 10, no. 8, pp. 1549-1563, 2015.
[15] N. Abdul Kadir, S. Abd Razak and H. Chizari, "Identification of fragmented JPEG files in the absence of file systems", 2015 IEEE Conference on Open Systems (ICOS), 2015.
[16] Y. Tang, J. Fang, K. Chow, S. Yiu, J. Xu, B. Feng, Q. Li and Q. Han, "Recovery of heavily fragmented JPEG files", Digital Investigation, vol. 18, pp. S108-S117, 2016.
[17] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey", Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, 2014.
[18] u. Sheloadkar and H. Joshi, "SURVEY PAPER ON BIG DATA ANALYTICS AND HADOOP", International Journal OF Engineering Sciences & Management Research, vol. 4, no. 1, pp. 58-63, 2017.
[19] B. Martini and K. Choo, "An integrated conceptual digital forensic framework for cloud computing", Digital Investigation, vol. 9, no. 2, pp. 71-80, 2012.
[20] G. Al, "Extracting Potential Forensic Evidences from Cloud Client Device using own Cloud as a Case Study", International Journal of Computer Applications, vol. 132, no. 7, pp. 15-21, 2015.
[21] B. Martini and K. Choo, "Cloud storage forensics: ownCloud as a case study", Digital Investigation, vol. 10, no. 4, pp. 287-299, 2013.
[22] S. Thanekar, K. Subrahmanyam and A. Bagwan, "A Study on Digital Forensics in Hadoop", Indonesian Journal of Electrical Engineering and Computer Science, vol. 4, no. 2, p. 473, 2016.