Cyber Forensic for Hadoop based Cloud System

ChaeHo Cho[1], SungHo Chin[2] and *Kwang Sik Chung[3]

[1]*Korea National Open University graduate school*
Dept. of Computer Science
[2]*LG Electronics CTO Division Software Platform Lab*
[3]*Korea National Open University, Dept. of Computer Science,*
domain@jsd.or.kr, sungho.chin@lge.com, kchung0825@knou.ac.kr

Abstract

Cloud services are so efficient and flexible to expend and manage the service so that various cloud services are commercially implemented and provided by KT uCloud, Amazon EC2, and the other companies. As could service are quickly deployed, more security problems occurs and cloud forensic procedures for cloud systems are needed. But, in multi-users serviced cloud systems, a system suspension makes serious problems to users so that collecting evidences and analysis have to be performed in the field and live analysis is important in cloud systems. Cloud system based on Hadoop distributed file system has characteristics of massive volume of data and multi-users, physically-distributed data, and multi-layered data structures. The previous forensic procedures and methodologies are not appropriate for cloud system based on Hadoop distributed file system. In order to deal with those characteristics of cloud system, we propose Hadoop based cloud forensic procedure that supports static analysis after live analysis and live collection without system suspension, and Hadoop based cloud forensic guidelines. With our proposed Hadoop based cloud forensic procedure, we can decrease the time for evidence collection and evidence volume

Keywords: *cloud distributed file system, Hadoop distributed file system, cyber forensic*

1. Introduction

Recently cloud services for companies and personals as like EC2 of Google and uCloud of KT are provided by many cloud service companies [1, 2]. Basically those cloud systems are implemented based on distributed file systems [2] and physically distributed servers so that they could guarantee reliability and scalability of service. 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, GLORY-FS of ETRI. But Hadoop distributed file system is open source and many commercial cloud service systems adopt it as their file system. Thus in near future, a lot of security problems will occur on Hadoop based cloud service system [4, 5, 6, 8].

But, previous cyber forensic methods could not be adopted to cloud system that is built based on Hadoop distributed file system [9, 10, 11]. First of all, Hadoop distributed file system is based on virtual file system and file access could be virtualized and the proof of forensic could be easily erased or un-audited. Second, cloud system assumes the multi users for various services. For that reason, cloud service halt or shut-down could be fatal to the reliability of the cloud service. But previous cyber forensic methods need the system halt or shut-down and after that forensic methods could be adopted to the cloud system. And last, cloud system cloud system manages huge file system that cloud be distributed physically at several places and the whole amount of storage volume is very large. Thus the time and

manpower for forensic proof collection cloud be so huge that the previous cyber forensic cloud not be managed.

We first compare cloud system and previous distributed system and analyze Hadoop distributed file system. After that, we propose cyber forensic for Hadoop based cloud system and cyber forensic guide line that could be used for real fields.

The rest of the paper is outlined as follows. In related works, we analyze the previous cyber forensic and show the disadvantages of the previous cyber forensic on cloud system. Hadoop based cyber forensic methods show the new cyber forensic method for cloud system and the new cyber forensic guide line for cloud system. Lastly we conclude this paper and present future works.

2. Related Works

In [12], rather than seizing the whole system, a forensic method of transferring disk images to the investigation center is proposed. This method is based on the 'snapshot' of the virtual machine in the cloud system. And, a creating an image with the volatile memory is to raise the time efficiency However, transferring the large volume of replica image on the network has the overhead. The volume has been bigger, more time and traffic has been occurred for the networking. It also has big problem of re-transferring when the network channel is failed or the hash is different from the original value.

In [20], the new concept 'Forensic Cloud' is proposed. This study is for definition of the forensic cloud framework. The objectives of the study are satisfying the requirement of forensic and providing the convenience of the user. The automation in all proposed layers can contain meaningless data and the tendency is clearer in the large-volume of file system. To find meaningful data from the real-analyzed data, information analysis and judgment of the investigators have to be involved.

In [21], the limitation of the prior works about the procedures and methodologies was identified. To solve these problems, simplification of the analysis methodology, introduce a database for analysis and applying cloud system were proposed. That is, the analysis is composed of scan, classification, searching& indexing, analysis, and documentation. To save process time, similar step are grouped and they are processed in parallel by utilizing a multi-system or a cloud. Although proposed parallel processing minimizes the processing time, the important evidence data cannot be searched in the distributed step when related evidence are processed in parallel. In the grouping phase, the big whole of analysis can be occurred since the important evidence can be split when distributing an actual storage to the several file groups

In the case of applying the original forensic methodology and processes which are based on one information processing media to the Hadoop distributed file system (HDFS), several problems can be occurred. These problems are as follows. Firstly, applications of the original forensic methodology occurs the problem of replicated image-generation time and securing storage in the HDFS since it is impossible that separated replication of each node's disk. Secondly, it is necessary that replicating with the block area of the stored data node or whole node when collecting evidence only in specific block area because HDFS stored the same replica in the different nodes. Thirdly, in general, replica is transferred with being stored in container that block external influences since replica is image file basis, but large image are not possible. Several problems such as, missing, falsification of data, and extra network traffic are also occurred when transferring the replica in the same network. Fourthly, service interruption or work stopping are occurred for other users since HDFS based cloud

service is utilized by various users. Especially in enterprise environment, system seizing results in terminating the service. Finally, a different analysis methodology is needed since HDFS physically separates master and data nodes. It is also hard to develop and apply a different analysis methodology because secondary name node also has different structure than the master node

3. HADOOP Based Cyber-Forensic Method

Our proposed Hadoop forensic step is for generally and widely used forensic procedure. We redefine the each step as a common tool used by the machinery of law or a private organization

Table 1. Comparison between original forensic and Hadoop file system based forensic

Forensic procedure of Hadoop file system	
Preparation	
Identification	
1st step Collection and Analysis	Collection
	Live Analysis
2nd step Collection and Analysis	Collection
	Transport
	Static Analysis
Reporting	

To address problems caused by applying original forensic procedure to Hadoop and to be a Hadoop based forensic, we redefine the procedure as follows. From the first preparation step to the last step of reporting, the overall structure is similar to the forensic structure proposed by the National Police Agency or NIST. However, there are some features when considering the characteristic of the cloud service and Hadoop file system.

①Preparation

A team should be composed of experts in research, technical analysis, law, etc., for forensic. To be an efficient collection, equipment and preparation for evidence collection must be ready. Therefore, in this step, it is important that explaining the necessity and appropriateness of investigate to the Hadoop file system administrator, network administrator, service administrator, planner and legal counsel to cooperate with them.

②Identification

Through this step, the load and time for collecting evidence can be minimize. In this step, since it is impossible that generating replica image of the whole large file system, pre-analysis is performed to minimize the scope of data collection. The identification is the most important step in Hadoop forensic and it is base of the shortening the overall

analysis time, the minimizing the scope of analysis and the reducing the damage of other users or organization.

③Live Collection

Since the name node has the information about where file and block are located, the evidence collection and analysis about the name node are performed in the field. The reason why the real time analysis is necessary in the field is that finding the specific block area for real evidence collection. It can be minimizing the time for generation replica and evidence collection.

④Live Analysis

This step is to analysis the structure of Editlog and FsImage files in the namenode. The real-time analysis of these two files is performed to identify the location of the block and the composition of the cluster which is needed for collecting real evidence. This analysis is for the next static analysis. Sometimes, the secondary namenode can be included for the analysis.

⑤Statistic Collection

This step is to collect evidence by the file or block from the live collection and live analysis steps. Based on the scope ruled in the live collection step, this step guarantees the integrity with the original one and generates the replica. Sometimes, replicating or seizing the whole node and taking the picture of the damaged system are processed. It must be prepared for the proof of integrity through taking picture or video recording of the all equipment, office and output of the monitor screen.

⑥Transport

After the collection, as in the original forensic, the transport step delivers the replicated image with being stored in the container or directly transfers the image to the investigator's computer. Not only in the collection and transport steps, the information form the live analysis also included in this step. The proof of integrity is the most important in the transport step. Lawful method and procedure must be processed to ensure it. If the physical equipment is seized, all process must be taken by picture or recoded by video. All process must be monitored by the observer and his statement also be submitted with the evidence.

⑦Static Analysis

Files or images collected in the field or seized equipment are analyzed by the various forensic tools and methods. Damaged files are recovered in this step. Sometimes, several forensic methods such as, description of the encoding file, extracting hidden information from Slack Space or file, finding usage trace of malicious equipment or program are adopted.

⑧Reporting

After the all collection and analysis steps are completed, final report is organized for submitting evidence to the court based on the results from the above steps. This step is similar with the reporting step in other domain and it must prevent analysis and reporting time from getting longer

4. Hadoop Based Cyber-Forensic Guidelines

We consider characters of Hadoop distributed file system and cloud services for Hadoop based cyber forensic guidelines and try to help detectives and managers in real fields. We think Hadoop based cyber forensic guidelines could be field manual. Especially Hadoop based cyber forensic guidelines has focus on forensic actions plan and live collection and analysis doe volatile memory and evidences, since Hadoop based cyber forensic guidelines should be referenced in the field. Especially collected and analyzed evidences would be used as proof for cyber crimes at the court. In each step of Hadoop based cyber forensic guidelines, legal process should be followed and the originality and proof integrity should be guaranteed for evidence collection and analyse[16].

①Preparation

At this stage, in order to make a live forensic team and cooperation, the necessity of cyber forensic should be explained to the Hadoop distributed file system administrator, network administrator, service administrator, planner and legal counsel.

②Identification

At this stage, in order to complete rapid and accurate evidence analysis, evidence collection and analysis should be narrowed and simplified. Maximized limitation of Hadoop distributed file system should be verified, collected and analyzed, since data is distributed over the whole cloud system. Location and structure of name node, secondary name node, backup name node should be confirmed. And network structure of machine and computer should be analyzed.

③Live Collection and Live Analysis

At this stage, live collection and analysis should be applied to name node, secondary node and backup node in the field. The system clock should be recorded and would be used as live collection and analysis process time. The system MAC time could be used as file history tag, evidence access time, and modification time. After that, network connection information, open port information and protocol, user audit information, network routing table, process and file information should be collected and analyzed.

④Static Collection

Disk replica of damaged system should be minimized. Hadoop distributed file system basically replicates and manages three replicas. All digital information should be collected by the unit of file or replica image. And all collected information has hash value for the integrity between the original one and replica.

⑤Transport

Whole disk of name node or only d\hard disk replica should be transport. In the case of transporting name node servers, fenders and safety boxes should be equip for safety against external shock and electromagnetic waves. Replicated image files should be transferred to image record disks and put into safety transport box.

⑥Static Analysis

Accesses trials and modification evidences of servers and clients should be carefully detected. And the original evidences should be carefully managed so that the replicas would be analyzed and used in order to track the cyber crime evidence. If the original evidences should be analyzed, the original evidences would be write-locked[24].

⑦Reporting

Reports should be clear and detailed. And it would be treated as evidence at the court and managed carefully

5. Conclusion

There are many difficulties in cyber forensics on the IT environment that is rapidly changing. But previous methods and tools of cyber forensics are still insufficient to satisfy new requirements of cloud services. Previous cyber forensics for mass distributed file system based on the Hadoop distributed file system, suggested in this paper, has many portions to be complement and improved.

In this paper, we presented the issue and the reason that the previous cyber forensics technologies and procedures could not be appropriate to apply to the Hadoop distributed file system. There are two problems in the stage of gathering evidences. One is confirming the file blocks replicated by nodes that are different each other. And the other is the excessive increase of time of copying the original. In the case of replicating the entire file system on the event of transferring the evidences, there have been some difficulties of acquiring more storage for moving these evidences. In the multi user environment of the Hadoop distributed file system, the entire service could be stopped on the case of halting the system or seizing evidences. Rising demands and interests for mass storage need a new way of forensics technologies and methodologies.

In order to solve these problems, the Hadoop based cyber forensics have to be simultaneous with live analysis of master node on the crime scene and static analysis after acquisition of evidences. And the ways of acquirement and analysis of evidences without stopping cloud services are request because of the characteristic of multi user environment

References

[1] H. Kim, Y. Lee, “Cloud computing service present condition and prospects”, Korea Information and Communications Society Magazine, vol. 27, no. 12.
[2] J. H. Yun, Y. H. Park, S. J. Lee, S.-M. Jang, J. S. Yoo, H. Y. Kim, Y.-K. Kim, “A Non-Shared Metadata Management Scheme of Large Distributed File Systems”, The Korea Institute of Information Scientists and Engineers Journal, vol. 36, no. 4.
[3] S. Ghemawat, H. Gobioff and S.-T. Leung, “The Google File System”, Proceeding SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles, **(2003)**.

[4] Y. S. Min, K. S. Jin, H. Y. Kim and Y. K. Kim. "A Trend to Distributed File Systems for Cloud Computing", ETRI Electronics and Telecommunications Trends, vol. 24, no. 12, **(2009)**.
[5] Hadoop: OSS Based Massive Data Management Technology. http://cloud.mk.co.kr/. 2010. 10. 20.
[6] HDFS Architecture Guide, http://Hadoop.apache.org, **(2012)**.
[7] Hadoop: The Definitive Guide, 2nd Edition, O'Reilly Media.
[8] Apache HBase Reference Guide. Apache Software Foundation, http://hbase.apache.org/book/book.html, **(2012)**.
[9] Analyzing Big Data with Hadoop, decorated a new history. Bloter.net 2011.7.20. http://www.bloter.net/archives/68650.
[10] Y. H. Yoo, B. G. Song and S. J. Park, "The necessity and training plans for Digital Forensics expertise", Korean Association Of Police Science Review, vol. 11, no. 4, **(2009)**.
[11] G. Y. Na, "A Study on the Gathering of Forensics in Cloud Computing Environment", **(2011)**.
[12] C.-S. Park, "Study on Security Considerations in the Cloud Computing", The Korea Academia-Industrial cooperation Society, vol. 12, no. 3, **(2011)** March.

Authors

ChaeHo Cho

Mr. Chaeho cho is a graduate student for Korea National Open University, dept. of computer Science. He is interested in cloud computing, and Hadoop file system.

SungHo Chin

Dr. Chin received the B.S. degree, the M.E. degree, and the Ph.D. degree in Computer Science Education from Korea University, in 2002, 2004 and 2010, respectively. He is currently a Senior Research Engineer at LG Electronics since 2011. His research interests are in Cloud computing, Grid computing, Distributed computing and Service Delivery Platforms.

Kwang Sik Chung

Dr. Chung received B.S. degree, the M.E. degree, and Ph.D. from Korea University, in 1993, 1995 and 2000, respectively. His major was distributed mobile computing. Currently he has interesting in M-learning and cloud computing for smart learning. He is an assistant professor at Korea National Open University.