CERIAS

The Center for Education and Research in Information Assurance and Security

Data Spillage In Hadoop Clusters

Oluwatosin Alabi[1], Joe Beckman[2], Dheeraj Gurugubelli[3]
[1]Purdue University, oogunwuy@purdue.edu; [2]Purdue University, beckmanj@purdue.edu; [3]Purdue University, dj@purdue.edu

Problem Statement

- Data spillage is the undesired transfer of classified information into an unauthorized compute node or memory media
- The loss of control over sensitive and protected data can become a serious threat to business operations and national security (NSA Mitigation Group, 2012).

Research Question

Can classified data leaked, by user error, into an unauthorized Hadoop Distributed File System (HDFS), be located, recovered, and removed completely from the server?

Specific Goals

- Apply a deletion protocol
- Determine can data be recovered in HDFS & to what extent?

1 HDFS

Load Tagged Document

- Retrieval of NameNode Metadata
- Recovery of DataNode Data Remnants

Delete Tagged Document

- Retrieval of NameNode Metadata
- Recovery of DataNode Data Remnants

Apply digital forensics procedure to locate data remnants

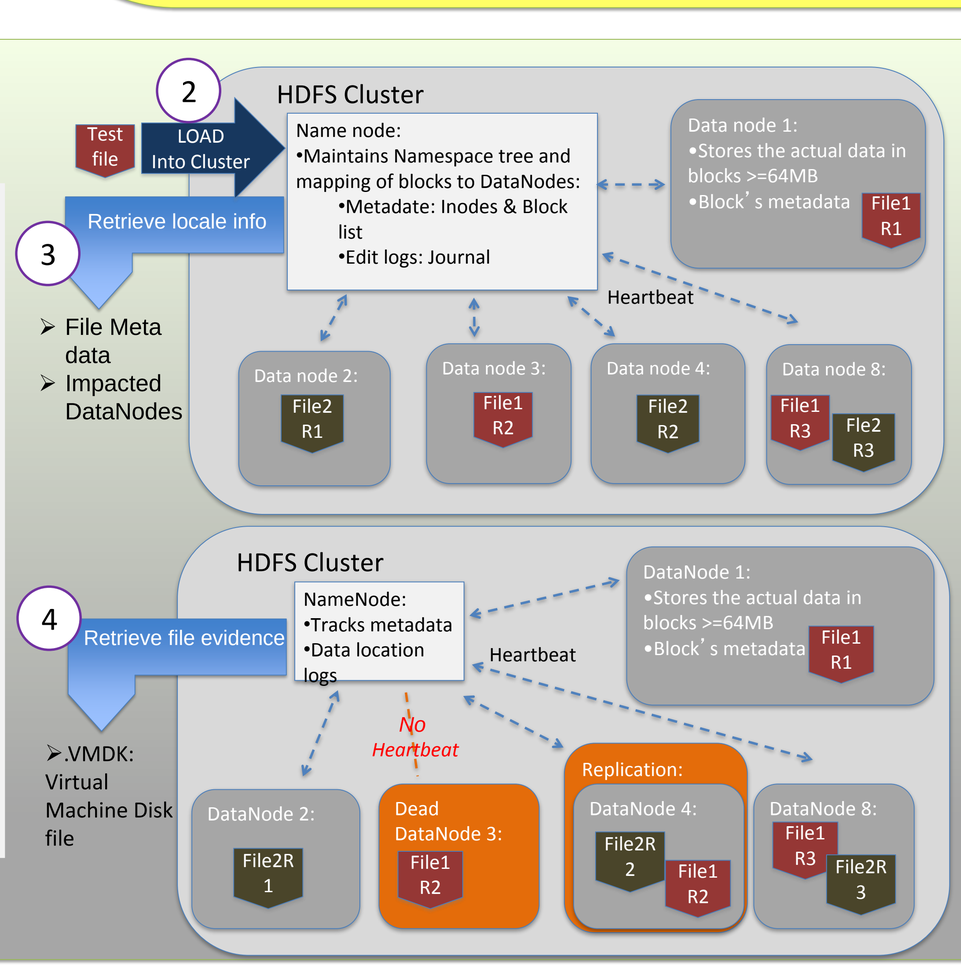
Assess data sanctification level

Compare results using the delete file method to a more secure method (e.g. NIST 800-88)

Research Process

1. Scope: Determine the investigation boundaries and limitation
2. Preparation: Prepare cluster environment
3. Identification: Identify the spilled data location on data nodes using the metadata on name node
4. Collection & Preservation: Acquire the .vmdk image files of impacted nodes.
5. Examine & Analyze: Conduct forensic procedure to find any spilled data remnants on the disks
6. Presentation: Document and Report Findings

Iterative



5 Examine & Analyze Disk Files

Forensic Acquisition	• Mount disk images	Forensic Tools: FTK Imager; Forensic Tool Kit 5.3
Log Analysis (Name Node)	• HDFS Event Logs • Disk ID's (Before Delete) • Moved to ./trash (After Delete)	
Forensic Analysis (Data Node)	• Block-Based Carving • Header/Footer Carving • File structure based Carving • Live Search • Index Search	

6 Lessons Learned

- ❖ Imaging retrieval and process time is proportional to the disk size: Larger the disk -> Longer processing time
- ❖ File types require different recovery procedures: Text file searching using FTK tool cannot be done doing
- ❖ Research process is iterative and each iteration may require different steps/tools

6 Future Work

- ❖ Exploration of the efficacy of various secure data sanitation methods for data removal in a virtual HDFS cluster
- ❖ Extension of this process to include the minimization of cluster downtime during the removal process
- ❖ Extension of this process to include detection and removal of other file types
- ❖ Automation of data removal process in HDFS

References:
[1] Mitigations NSA Group. (2012). Securing Data and Handling Spillage Events [White Paper]. Retrieved fromhttps://www.nsa.gov/ia/_files/factsheets/final_data_spill.pdf
[2] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* (pp. 1-10). IEEE.
[3] Lim, S., Yoo, B., Park, J., Byun, K., & Lee, S. (2012). A research on the investigation method of digital forensics for a VMware Workstation's virtual machine. *Mathematical and Computer Modelling*, *55*(1), 151-160.

CERIAS

PURDUE UNIVERSITY