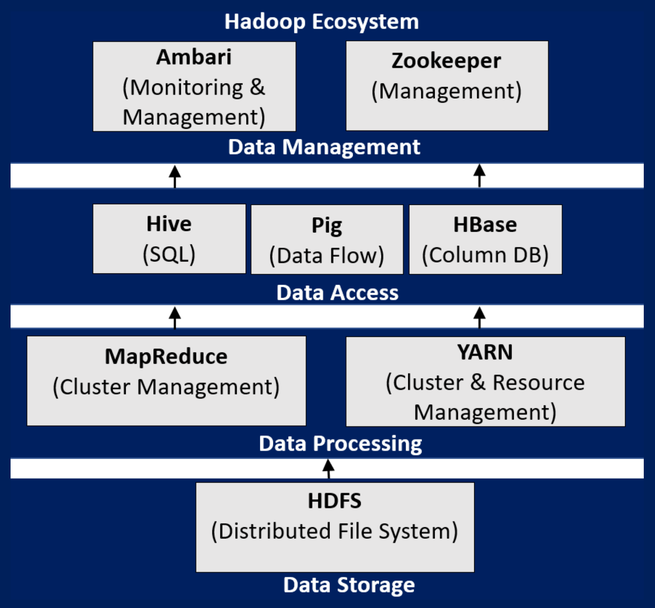
Big Data Forensics: Hadoop 3.2.0 Reconstruction

UNIVERSITY OF SOUTH ALABAMA
SCHOOL OF COMPUTING

Edward Harshany[1], Ryan Benton[1], David Bourrie[1], William Glisson[2]
[1] *University of South Alabama,* School of Computing: Data Science Research Forum
[2]*Sam Houston State University, College of Science and Engineering Technology*

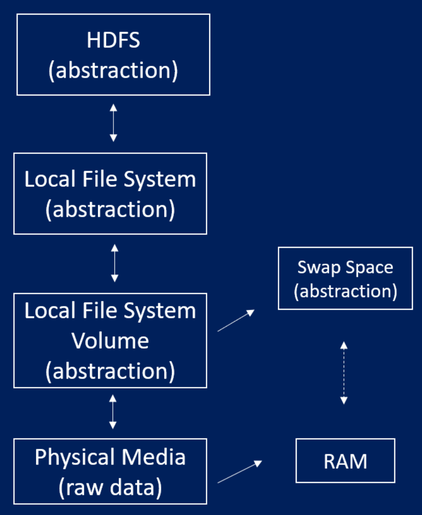
Introduction

Conducting digital forensic investigations in a big data distributed file system environment presents significant challenges to an investigator given the high volume of physical data storage space. Metadata management is vital to the Hadoop Distributed File System (HDFS). HDFS is designed to centrally manage all distributed file system metadata through the master server called the Namenode.



Goal

This study aims to investigate the effectiveness of utilizing a subset of metadata generated at the HDFS abstraction layer to reconstruct file system operations and map data to physical data location

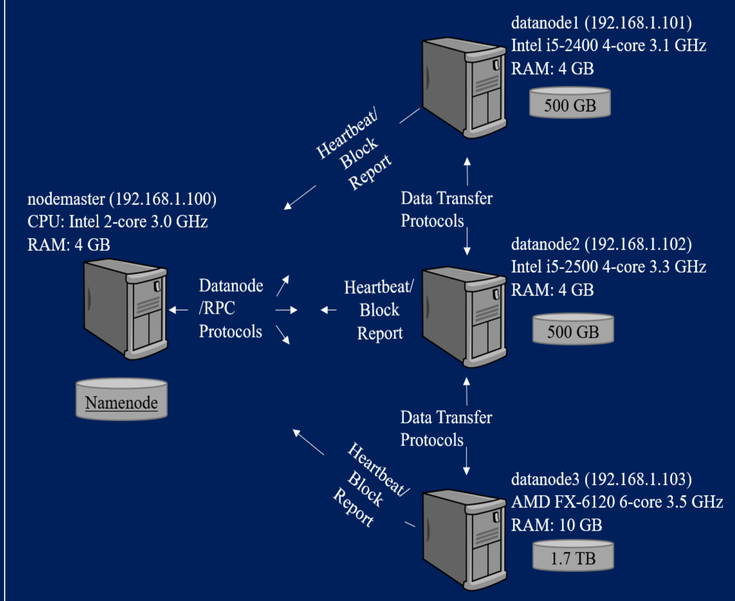


Approach

Our approach is based on determining the footprint left behind by HDFS as it executes on a system. HDFS metadata is collected at the central server and utilized to reconstruct events in time series. Journal file entries are then used to map HDFS block data to physical data location.

Methodology

Methods were confined to the construction of directories and file operations addition and deletion in a specific order. The timeline creates a directory structure within the HDFS namespace, adds files to the HDFS namespace, and deletes specific files from the HDFS namespace. The goal is to reconstruct the sequence of operations over this time period and discover file locations from the HDFS metadata. The cluster was configured in a fully distributed mode with Hadoop 3.2.0. on available commodity hardware, each running 64-bit Ubuntu 18.0.4.2 operating system with version 4.15 Linux kernel.



Analysis and Findings

HDFS metadata files were recovered from the live system and converted to .XML files. The XML files are read by a parser developed to extract attributes. HDFS audit logs were converted to text files and parsed for analysis purposes. The metadata files contain structure to deduce partial event sequencing. Converting metadata modification times and searching Namenode master audit log file for the converted date and time reveals block datanode destinations including deleted files.

Time Seriess	File	Block ID	Dnode1	Dnode2	Dnode3
1	file1.txt	1073741840	X		X
2	file2.txt	1073741841		X	X
3	file2.txt	1073741842		X	X
4	file2.txt	1073741841		X	X
5	file2.txt	1073741842		X	X
6	file3.txt	1073741843	X		X
7-43					
44	file4.txt	1073741881	X		X
45	file4.txt	1073741882	X		X
46	file1.txt	1073741840	X		X

Future Work

Future work entails expanding the scale to a large clustered environment with varying replication and erasure coding schemes, incorporating metadata from additional layers of the Hadoop ecosystem, and automating the reconstruction process.