Hadoop Forensics

Be in a defensible position.
Be cyber resilient.

June 2016
About me

Kevvie Fowler, GCFA Gold, CISSP
National Cyber Response Leader
KPMG Canada

Tel: (416) 777-3742
Email: kevviefowler@kpmg.ca

SANS Lethal Forensicator
Hadoop is a software framework for storing and analyzing data.

- We will focus on CDH 5.7
- Principles will also apply to other Hadoop versions
What is Apache Hadoop?

Hadoop overview | architecture

Web Server
Daemons
Meta Data
File System Image
FS Image files
Edit logs
Name Node logs

Web Server
Task Tracker
Logging

Data Node 1
File 1
Block 1
File 2
Block 1
Block 2

Data Node 2
File 1
Block 1
File 2
Block 2
Block 1

Data Node 3
File 1
Block 1
File 2
Block 1
Block 2

Network and remote access infrastructure

Management Tools / Infrastructure
Kerberos, LDAP, etc.
Cluster management tools

Hadoop Interfaces
MapReduce jobs
Web UI
Command line
HDFS API

Name Nodes

What type of artifacts are contained within Apache Hadoop?

- Hadoop HDFS has several artifacts
- We’ll be looking at five artifacts today
Artifacts | Cluster properties

Getting cluster information via the webUI
<address>:50070/dfshealth.html

| Live Nodes | 1 (Decommissioned: 0) |
| Dead Nodes | 0 (Decommissioned: 0) |
| Decommissioning Nodes | 0 |
Number of Under-Replicated Blocks : 0

NameNode Storage:

<table>
<thead>
<tr>
<th>Storage Directory</th>
<th>Type</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>/var/lib/hadoop-hdfs/cache/hdfs/dfs/name</td>
<td>IMAGE_AND_EDITS</td>
<td>Active</td>
</tr>
</tbody>
</table>

Via command line

```
hdfs dfsadmin -report
```

```
[cloudera@localhost -]~$ hdfs dfsadmin -report
Configured Capacity: 10079059968 (9.39 GB)
Present Capacity: 5861765120 (5.46 GB)
DFS Remaining: 5741805568 (5.35 GB)
DFS Used: 119959552 (114.4 MB)
DFS Used%: 2.05%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

Datanodes available: 1 (1 total, 0 dead)
Live datanodes:
Name: 127.0.0.1:50010 (localhost.localdomain)
Hostname: localhost.localdomain
Decommission Status : Normal
Configured Capacity: 10079059968 (9.39 GB)
DFS Used: 119959552 (114.4 MB)
Non DFS Used: 4217294848 (3.93 GB)
DFS Remaining: 5741805568 (5.35 GB)
DFS Used%: 1.19%
DFS Remaining%: 56.97%
Last contact: Sat Sep 22 12:03:06 EDT 2012
```
## What is Apache Hadoop?

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
<th>Key properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>core-site.xml</td>
<td>General Hadoop properties</td>
<td>fs.trash.interval, fs.trash.checkpoint.interval</td>
</tr>
<tr>
<td>hdfs-site.xml</td>
<td>HDFS properties</td>
<td>dfs.replication, dfs.blocksize, dfs.permissions.enabled, hadoop.tmp.dir, dfs.namenode.name.dir, dfs.namenode.checkpoint.dir, dfs.datanode.data.dir</td>
</tr>
<tr>
<td>mapred-site.xml</td>
<td>MapReduce properties</td>
<td>mapreduce.task.tmp.dir, mapreduce.jobhistory.done-dir</td>
</tr>
<tr>
<td>log4j.properties</td>
<td>Logging properties</td>
<td>hadoop.mapred.JobTracker, hadoop.mapred.TaskTracker, namenode.FSNamesystem.audit</td>
</tr>
</tbody>
</table>
The last transaction that modified the image
Offline Image Viewer (OIV) can be used to gather the current FSImage file.

Example command:

```
hdfs oiv -i "<file and path>" -o "<file and path>" -p
```

Delimited: `-delimiter "|"`
Artifacts | Metadata (continued)

Metadata | Edit log
- All writes to files on disk are first completed within the edit log
- 1M Edit operations are retained by default
Operations within the Hadoop editlog

- OP_INVALID
- OP_ADD
- OP_RENAME_OLD
- OP_DELETE
- OP_MKDIR
- OP_SET_REPLICATION
- OP_DATANODE_ADD
- OP_DATANODE_REMOVE
- OP_SET_PERMISSIONS
- OP_SET_OWNER
- OP_CLOSE
- OP_SET_GENSTAMP
- OP_SET_NS_QUOTA
- OP_CLEAR_NS_QUOTA
- OP_TIMES
- OP_SET_QUOTA
- OP_RENAME
- OP_CONCAT_DELETE
- OP_SYMLINK
- OP_GET_DELEGATION_TOKEN
- OP_RENEW_DELEGATION_TOKEN
- OP_CANCEL_DELEGATION_TOKEN
- OP_UPDATE_MASTER_KEY
- OP_REASSIGN_LEASE
- OP_END_LOG_SEGMENT
- OP_START_LOG_SEGMENT
- OP_UPDATE_BLOCKS
Artifacts | Edit (transaction) log

Metadata | Edit log (continued)

- Offline Edits Viewer (OEV) with the `stats` processor dumps a summary of operations within the edit log
- OEV can be run online

```
hdfs oev -i <file and path> -o <file and path> -p stats
```
Artifacts | Edit (transaction) log

**Metadata | Edit log (continued)**

Offline Edits Viewer (OEV) without the `stats` processor dumps operations performed after the last FSImage file

```
hdfs oev -i "<edit filepath>" -o "<out file>" -v
```
Artifacts | Edit (transaction) log

Edit log structure

Example of file copied from local system to hadoop
Artifacts | Retrieving files from Hadoop

Fsck command can be used to list the blocks associated with a given file

```
hdfs fsck /<file> -files -blocks -racks
```
Artifacts | Retrieving files from Hadoop

Current state of Hadoop | File retrieval (continued)

Navigate the local FS of the specified data node to the block location

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Type</th>
<th>Date Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>blk_-4780450617297924898</td>
<td>4 B</td>
<td>blk_-4780450617297924898</td>
<td>Thursday</td>
</tr>
<tr>
<td>blk_-4780450617297924898_2584.meta</td>
<td>11 B</td>
<td>meta document</td>
<td>Thursday</td>
</tr>
<tr>
<td>blk_-5410253256845230727</td>
<td>52 B</td>
<td>plain text document</td>
<td>Thursday</td>
</tr>
<tr>
<td>blk_-5410253256845230727_2588.meta</td>
<td>11 B</td>
<td>meta document</td>
<td>Thursday</td>
</tr>
<tr>
<td>blk_5487976805417073172</td>
<td>2.3 MB</td>
<td>executable</td>
<td>Today</td>
</tr>
<tr>
<td>blk_5487976805417073172_2590.meta</td>
<td>18.7 k</td>
<td>meta document</td>
<td>Today</td>
</tr>
<tr>
<td>blk_-7567954003590675439</td>
<td>2.3 MB</td>
<td>executable</td>
<td>Today</td>
</tr>
<tr>
<td>blk_-7567954003590675439_2592.meta</td>
<td>18.7 k</td>
<td>meta document</td>
<td>Today</td>
</tr>
</tbody>
</table>

Blocks (files) can be imaged directly from the local FS
Artifacts | Retrieving deleted files

Hadoop Trash
- Often enabled by organizations
- When enabled – deleted files are not really deleted
- Files are moved to Trash location (./.user/<username>/.Trash) where they are scheduled for deletion at a later date/time

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Size</th>
<th>Replication</th>
<th>Block Size</th>
<th>Modification Time</th>
<th>Permissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CreditCardNumbers09212012.txt</td>
<td>file</td>
<td>2.34 MB</td>
<td>1</td>
<td>64 MB</td>
<td>2012-09-23 12:05</td>
<td>rw-r--r--</td>
</tr>
</tbody>
</table>
Hadoop threats to look out for

A key Hadoop risk - HiveQL includes many operators, functions and expressions commonly abused by SQL by injection attacks

- Count
- Union
- Distinct
- Wait for
- Sub queries
- Expressions joined by OR in a WHERE clause
- Comparisons between two constants

<table>
<thead>
<tr>
<th>Type of injection</th>
<th>SQL/ASP .NET</th>
<th>HIVEQL/ HUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dynamic SQL Injection</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Blind SQL Injection</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stacked queries</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Questions?
Thank you

Contact me

Kevvie Fowler
Partner, National Cyber Response Leader
KPMG Canada

T: 416-777-3742
E: kevviefowler@kpmg.ca