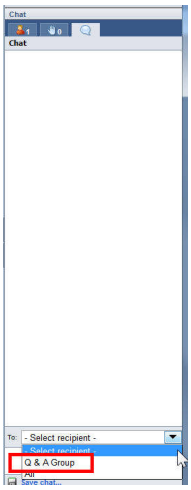
What is this thing called Hadoop and how do I secure it?

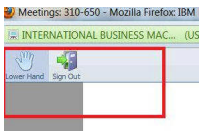
Kathy Zeidenstein, IBM InfoSphere Guardium Evangelist and Tech Talk Host
Sundari Voruganti, IBM InfoSphere Guardium QA and solutions architect
Mike Murphy, WW Center of Excellence

IBM

Logistics

- This tech talk is being recorded. If you object, please hang up and leave the webcast now.
- We'll post a copy of slides and link to recording on the Guardium community tech talk wiki page: http://ibm.co/Wh9x0o
- You can listen to the tech talk using audiocast and ask questions in the chat to the Q and A group.
- We'll try to answer questions in the chat or address them at speaker's discretion.
  - If we cannot answer your question, please do include your email so we can get back to you.
- When speaker pauses for questions:
  - We'll go through existing questions in the chat





Reminder: Guardium Tech Talks

Next tech talk: Managing SOX compliance: People, processes, and technology

Speakers: Karl Wehden and Joe DiPietro

Date &Time: Thursday, August 7th, 2014

11:30 AM Eastern Time (60 minutes)

Register here: bit.ly/1xnMMgd

Next tech talk +1: Technical deep dive for InfoSphere Guardium on z

Speakers: Ernie Mancill

Date &Time: Wednesday, August 27th, 2014

11:30 AM Eastern Time (60+ minutes)

Register here: http://bit.ly/1jhYeqZ

InfoSphere Data Privacy for Hadoop Webinar, July 24

- **Please join our webinar and learn how InfoSphere® Data Privacy for Hadoop enables you to establish a common big data business language and manage business perspectives about information, aligning those views with the IT perspective.**

  Highlights of this solution include:
  - **Automate the discovery of relationships within and across big data sources.**
  - **Extract, de-identify, mask, and transform sensitive data targeting or residing in your Hadoop environment.**
  - **Mask data on demand and in flight to your Hadoop environment.**
  - **Provide native masking support for Hadoop, CSV, and XML files.**
  - **Continuously monitor access and reduce unauthorized activities for big data environments such as Hadoop.**
  - **Create a single, enterprise-wide view of big data security and compliance across the entire data infrastructure.**

- Register here https://events.r20.constantcontact.com/register/eventReg?llr=jtqllyfab&oeidk=a07e98stk893a6e0fda

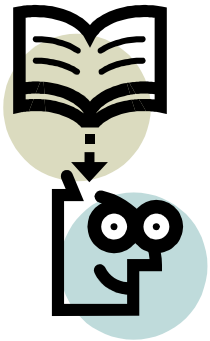
Agenda

- **Part 1**
  - –What and Why Hadoop?
  - –Security overview
  - –Guardium Activity Monitoring for Hadoop Overview
- **Part 2**
  - –Demo
- **Part 3**
  - –Implementation details and sample output
  - –Sample reports
  - –Deployment guidance (so far)
  - –Q and A

Caveat: Hadoop is a rapidly evolving space. And security within Hadoop is evolving as well. You will need to validate your own environment and capabilities with the IBM Team (for Guardium) and the Hadoop vendor you work with…

Hadoop Deployment Guide DRAFT

- **For those that are seriously looking at Hadoop deployments with Guardium, we have a draft of a deployment guide that is in constant state of change**
  - Includes much material from this tech talk plus more details and will continue to evolve
- **Cannot distribute widely because of its draft nature.**
- **If you want a draft, and agree that you understand it is a DRAFT to be used for education and planning purposes and not as official product documentation, please send Kathy Zeidenstein an email with your request.**



krzeide@us.ibm.com

Polling question

Which best describes your technical background for this topic?

1. **I know about Hadoop pretty well but not much about Guardium.**
2. **I understand Guardium but Hadoop is Greek to me (and I'm not Greek)**
3. **I understand both technologies pretty well**
4. **I don't feel comfortable with either technology. That's why I'm here!**

Big Data Scenarios Span Many Industries

Multi-channel customer sentiment and experience a analysis

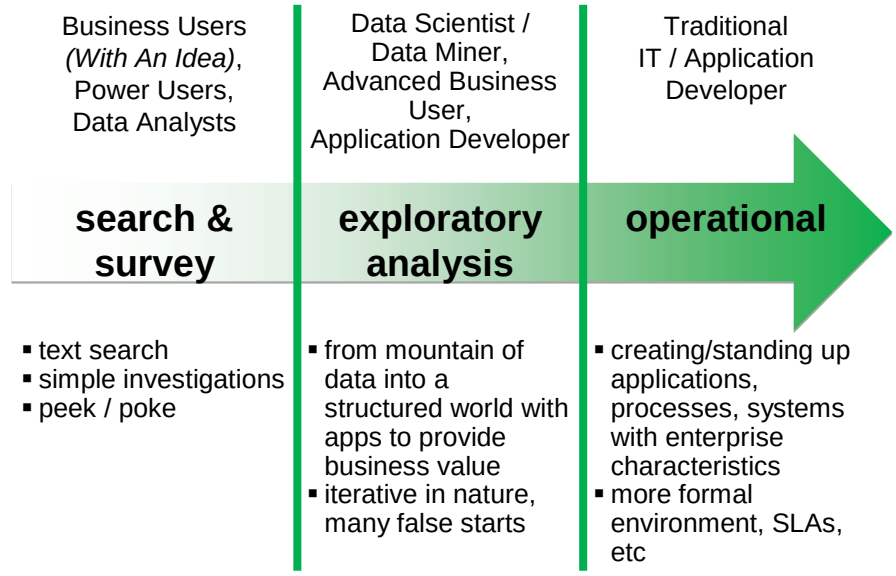
Detect life-threatening conditions at hospitals in time to intervene

Predict weather patterns to plan optimal wind turbine usage, and optimize capital expenditure on asset placement

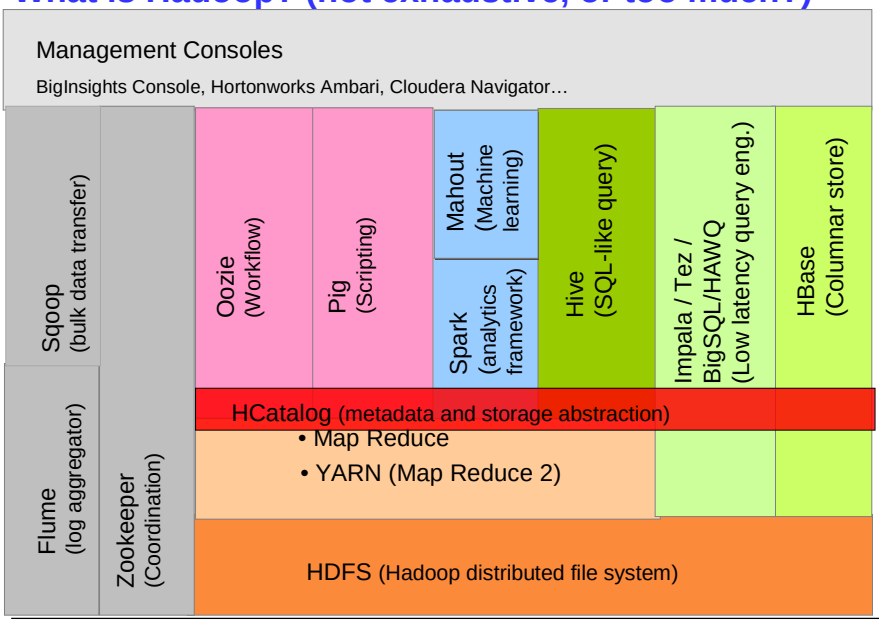
Make risk decisions based on real-time transactional data

Identify criminals and threats from disparate video, audio, and data feeds

Big Data life cycle – from raw to production



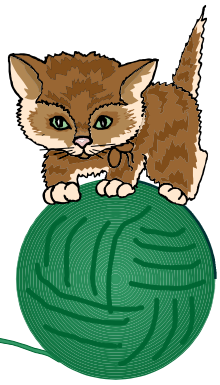
What is Hadoop? (not exhaustive, or too much?)



 Check out also http://wikibon.org/wiki/v/HBase,_Sqoop,_Flume_and_More:_Apache_Hadoop_Defined

What is Hadoop?

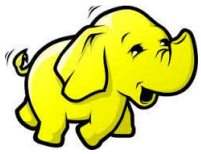
- **The framework consists of 'common,' HDFS, and MapReduce (YARN=MapReduce 2)**
  - Derived from Google papers on Apache Hadoop's MapReduce and HDFS (Hadoop Distributed File System)
  - Designed for fault tolerance and large scale processing
- **Additional Apache projects introduced to round out the 'platform, such as:**
  - Hive (warehouse queries)
  - HBase (Big Table storage)
  - And more… (introduced in many cases by vendors)
- **And other components are differentiators included by the vendors …**
  - Cloudera – Impala (open source query engine)
  - Pivotal – HAWQ (SQL-compliant)
  - IBM BigInsights – BigSQL (SQL compliant), BigSheets (spreadsheet metaphor)
  - Tez – high speed data processing app on YARN
- **According to Gartner, top 13 projects supported by major vendors include:**
  - HDFS, YARN, MapReduce, Pig, Hive, HBase, Zookeeper, Flume, Mahout, Oozie, Sqoop, Cascading and HCatalog.
  - http://blogs.gartner.com/merv-adrian/2014/03/24/hadoop-is-in-the-mind-of-the-beholder/



Yet
Another
Resource
Negotiator

 http://blogs.gartner.com/merv-adrian/2014/06/28/what-is-hadoop-now/

Hadoop releases



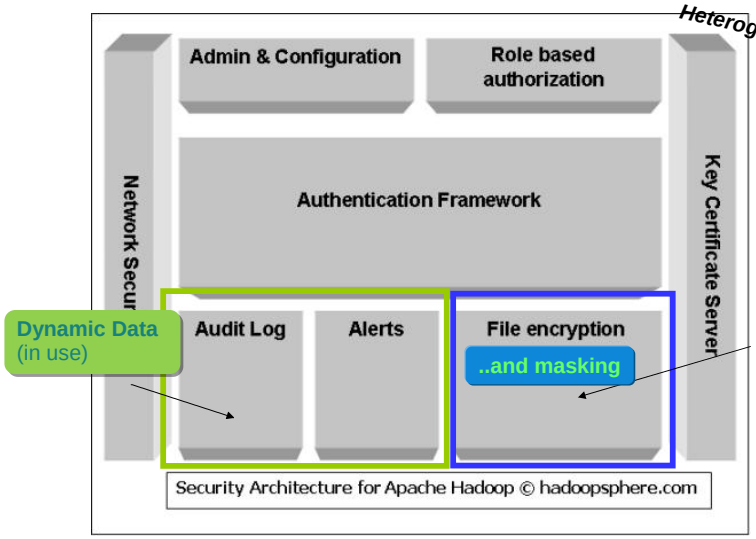
- 1.2.X - **current stable version, 1.2 release**
- 2.4.X **– current stable 2.4 version (April 2014)**
- **0.23.x –** Apache branch. Similar to 2.4 but without Name Node High Availability

Hadoop 2.x requires Guardium GPU 210 + patches to support YARN and other changes

Source: http://hadoop.apache.org/releases.html

A Hadoop Security Architecture

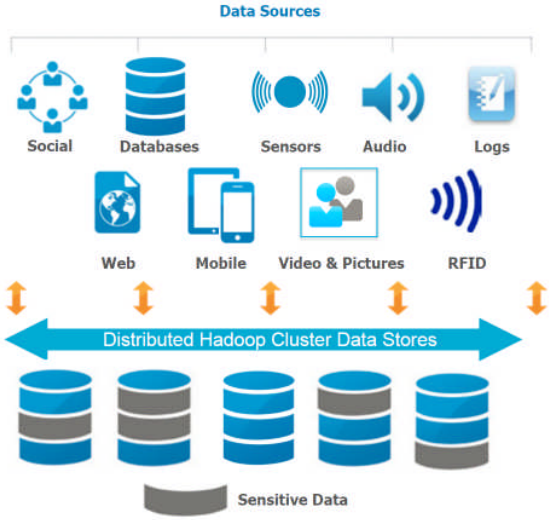
Defense in depth
Hadoop-agnostic
Heterogeneous



http://www.hadoopsphere.com/2013/01/security-architecture-for-apache-hadoop.html

Protecting Big Data at Rest –Guardium Data Encryption

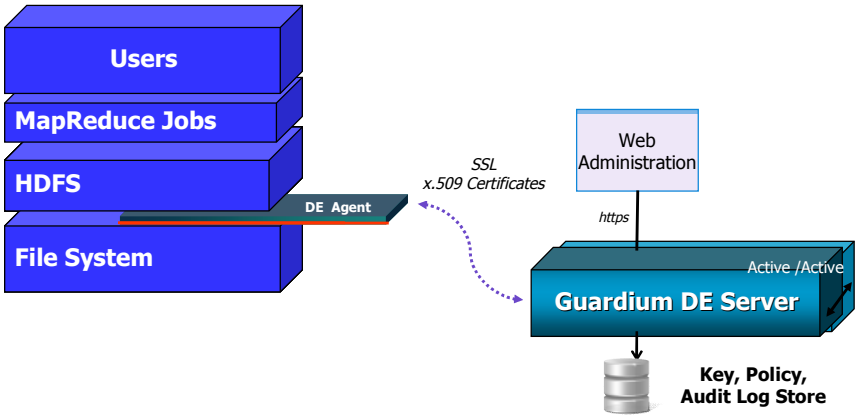
- All data sources potentially contain sensitive information
- Data is distributed as needed throughout Hadoop cluster
- Deploy IBM InfoSphere Guardium Data Encryption Agents to all systems hosting Data Stores
- Agents protect the data store at the file system or volume level



Data security best practices: A practical guide to implementing data encryption for InfoSphere BigInsights (on developerWorks)

From Vormetric

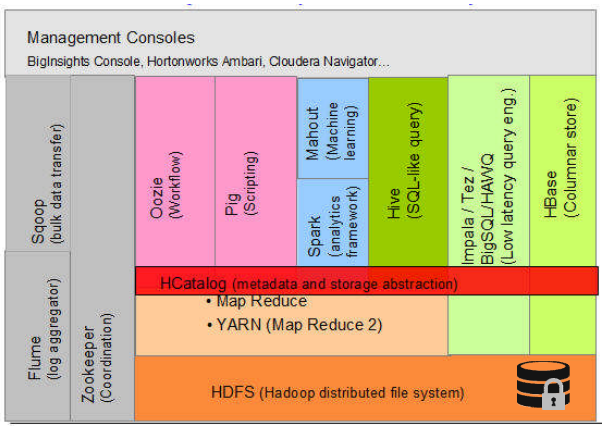
Protecting Data with Guardium Data Encryption Architecture



Data Encryption Security Server

- Policy and key management
- Centralized administration
- Separation of duties

Monitoring and auditing challenges



- **Many avenues to access**
- **Security and authentication is evolving**
- **Complex software stack with significant log data from each component**
- **Security and audit viewed in isolation from rest of data architecture**

InfoSphere Guardium Hadoop Activity Monitoring

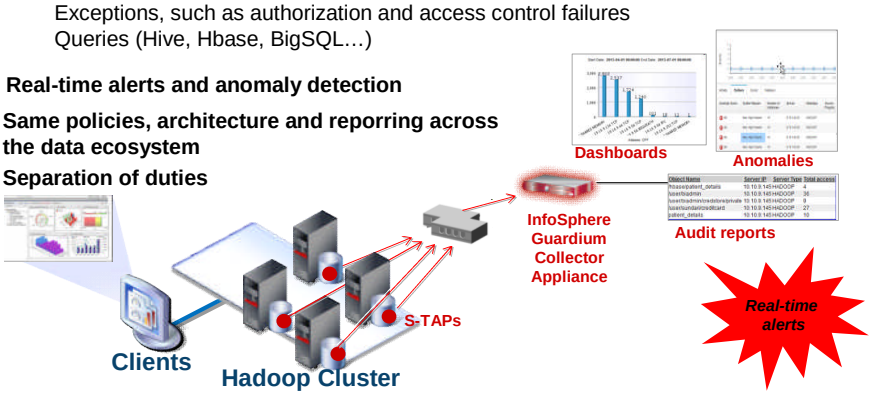
Monitors key Hadoop events:

- Session and user information
- HDFS operations and MapReduce jobs
- Exceptions, such as authorization and access control failures
- Queries (Hive, Hbase, BigSQL…)

Real-time alerts and anomaly detection

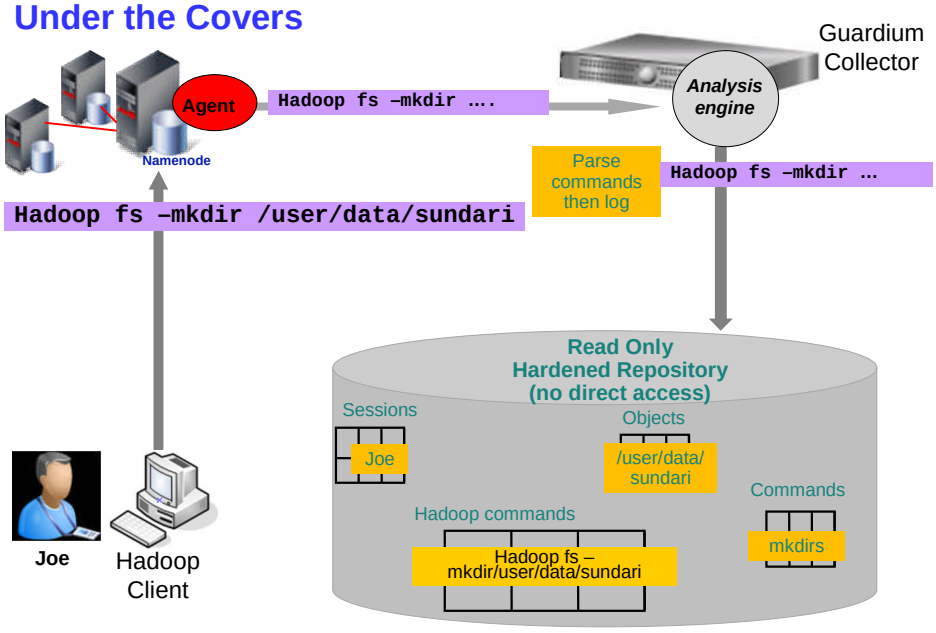
Same policies, architecture and reporring across the data ecosystem

Separation of duties



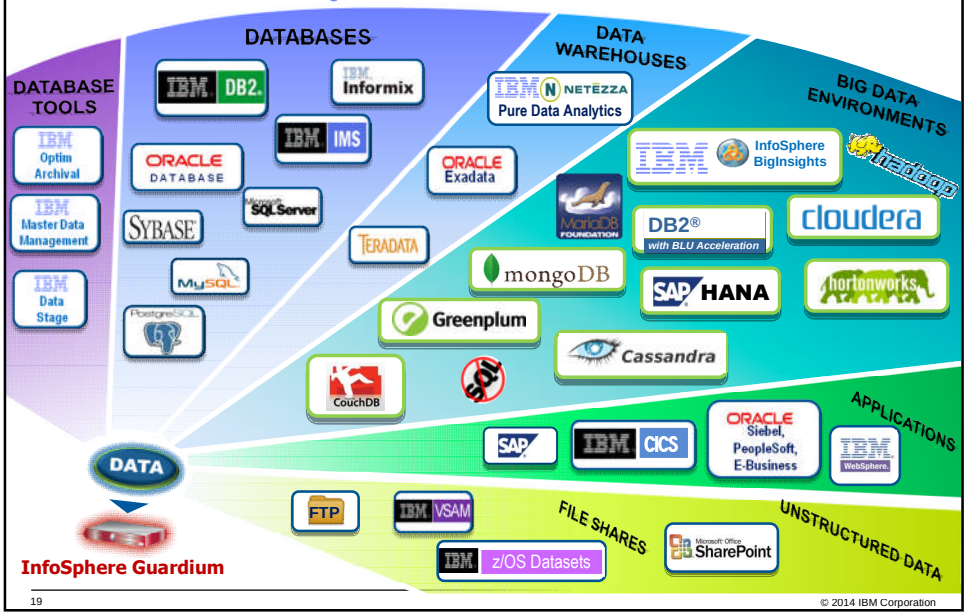
Heavy lifting occurs on Guardium collector! Very low overhead on Hadoop Cluster

Under the Covers



Comprehensive support for structured and unstructured sensitive data:

Databases, data warehouses, Big Data Environments and file shares



Data Privacy Offerings

InfoSphere Data Privacy for Hadoop

- *Secure and Protect Sensitive big data*
- *Extend Compliance Controls*
- *Promote Information Sharing*
- *Employ across diverse environments*

InfoSphere Data Security and Privacy

Purpose-Built Capabilities

Define and Share
Discover and Classify
Mask and Redact
Monitor Data Activity

InfoSphere Data Privacy and Security for Data Warehousing

- *Achieve and enforce compliance*
- *Secure and Protect sensitive data in data warehouses*
- *Reduce costs of attaining enterprise security*

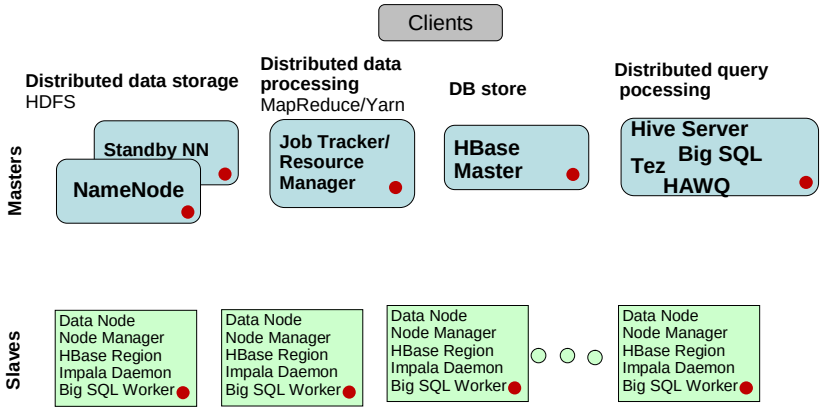
Demo

Mike Murphy

Implementation details and sample activity reports

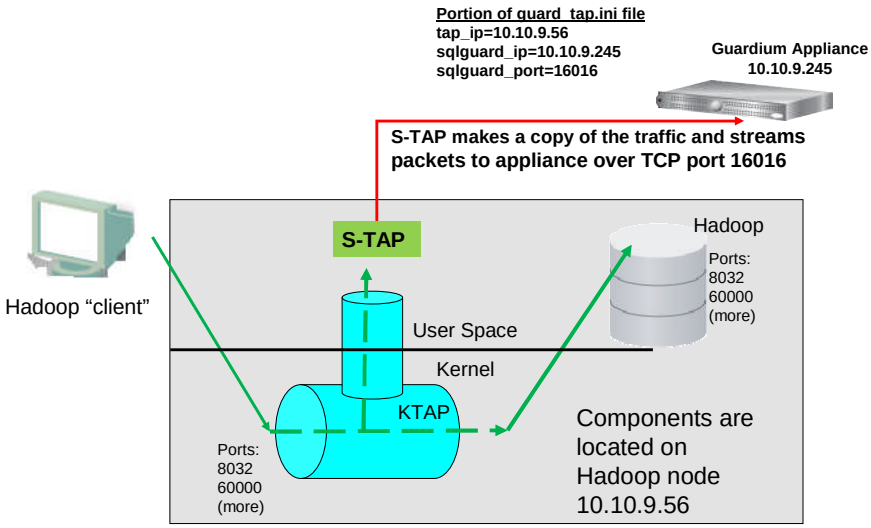
Sundari Voruganti

S-TAP placement in a Hadoop cluster



- S-TAP
  S-TAP required on the region servers/compute nodes only for monitoring HBase table operations, Impala, and Big SQL traffic.

STAP Architecture



Configuring inspection engines (know your ports)

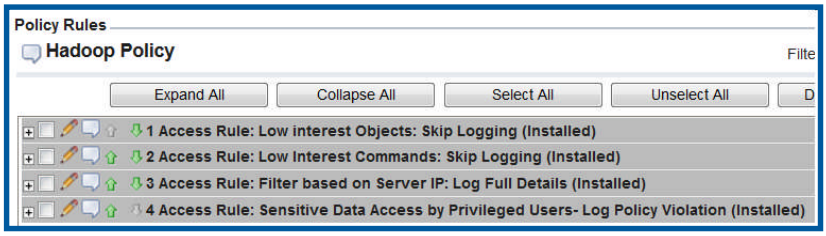
Hadoop Node	Service	Default Ports	Inspection engine protocol
Namenode	HDFS Name Node	8020, 50470	Hadoop
Namenode	HDFS Thrift plugin for Cloudera Hue	10090	Hadoop
Namenode	HTTP port (for WebHDFS)	50070	HTTP
Namenode	Resource Manager (YARN only)	8032	Hadoop
Namenode	HTTP port (HCatalog)	50111	HTTP
Job Tracker	MapReduce Job Tracker	8021, 9290, 50030	Hadoop
HBase Master	HBase Master	60000 and 60010	Hadoop
HBase Master	HBase Thrift plugin	9090	Hadoop
HBase Region	HBase Region	60020	Hadoop
Hive Master	Hive Server (Thrift protocol messages)	10000	Hadoop
Hue Server	Beeswax Server (with MySQL as Hue metastore)	8002	Hadoop
Hive Server	Hive Metastore (Thrift protocol messages)	9083	Hadoop
Impala daemons	Impala	21000	Hadoop
Management node	BigSQL server	51000	DB2
Compute node	BigSQL worker	51000	DB2
Impala daemons	Impala	21000	Hadoop

Configure only what you need!
Use Guardium APIs to speed deployment

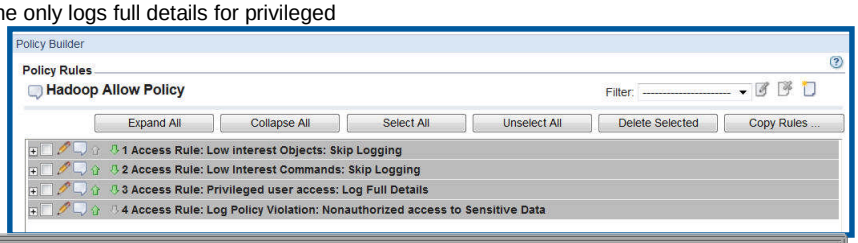
Decrypting user names from Kerberos…

- **If you are using built-in integration such as BigInsights GuardiumProxy, this is handled before traffic is sent to Guardium**
- **For Big SQL, you will need to configure DB2_exit with S-TAP.**
- **For other Hadoop distributions, you will need to do special configuration with keytabs.**
- **If you are planning an implementation that uses Kerberos, please work with IBM to get this information.**
- **Also included in deployment guide draft that is available from the authors upon request.**

Sample Hadoop Policies

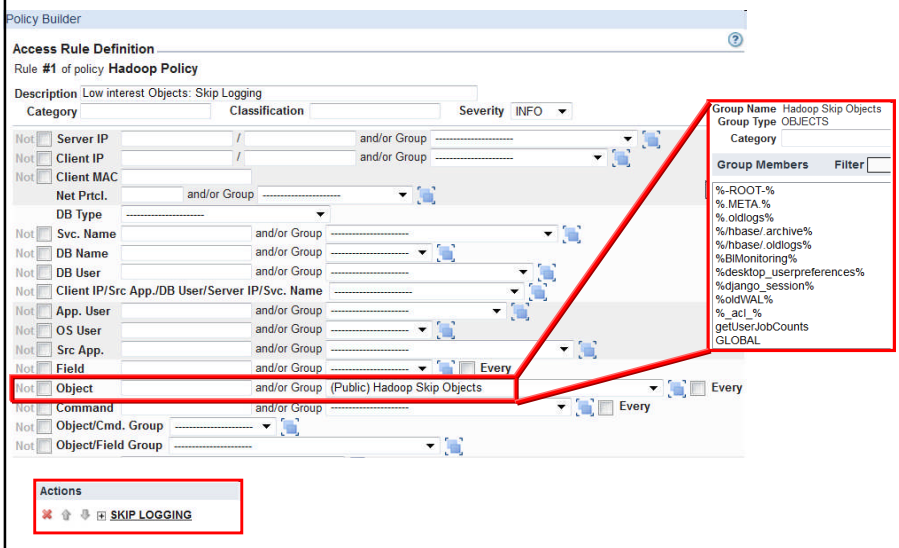


This one only logs full details for privileged users

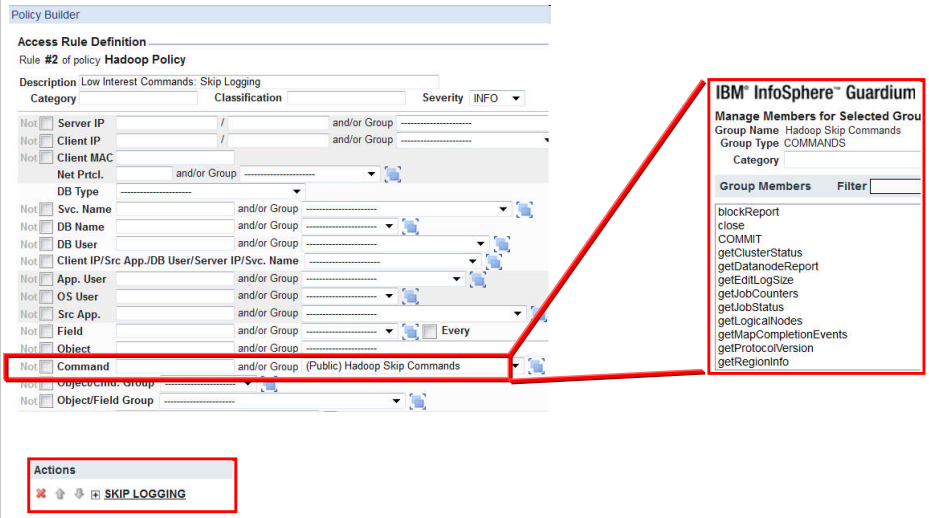


Tip: Log Full details is NOT recommended for production environments except when *exact* timestamp is needed for *each* statement. (Default behavior will aggregate occurrences of exact same command within an hour.)

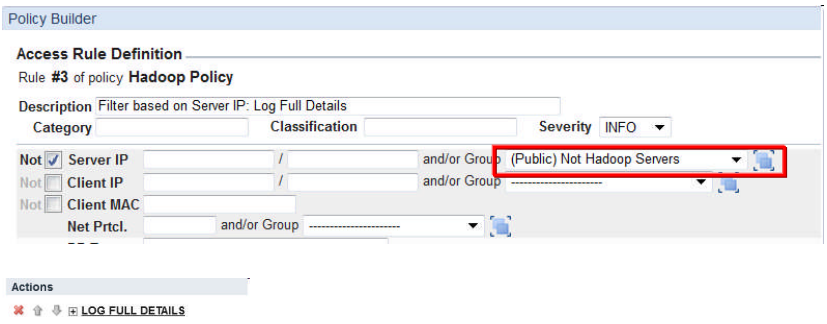
Hadoop Policy Filter out uninteresting objects



Hadoop Policy: Filter out uninteresting commands

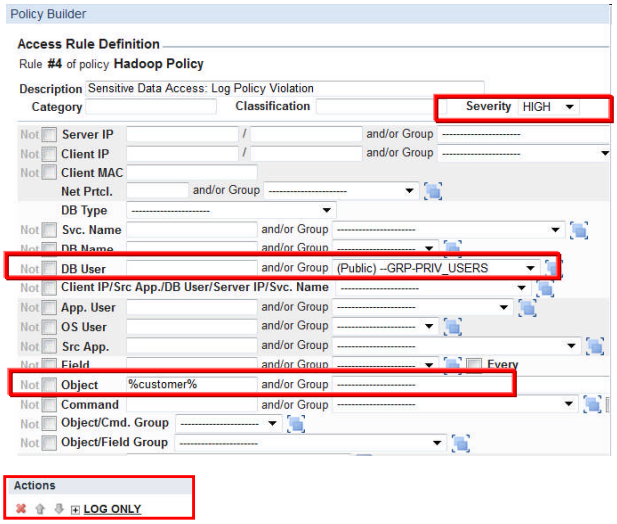


Hadoop Policy (filter out non hadoop servers)



Tip: You MUST put something in the Not Hadoop Servers group, even if it's a dummy IP or you will not collect traffic. (You can remove this condition if it's not necessary).

Hadoop Policy: Sensitive Data Access Incident



HDFS

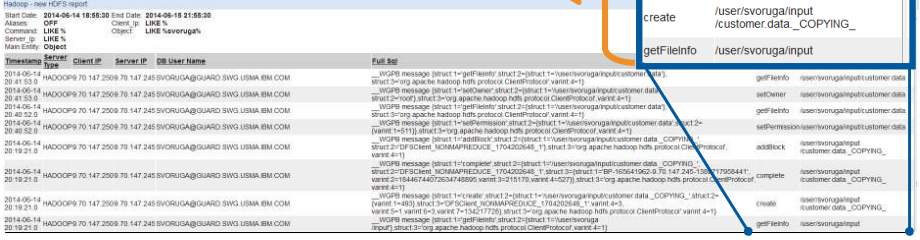
A Java-based file system that spans all the nodes in a Hadoop cluster. It links together the file systems on many local nodes to make them into one big file system.

S-TAP installed on…	Ports	Protocol
Name Node Secondary Name Node	8020	Hadoop
	50070	HTTP (WebHDFS)

```
hadoop fs -put customer.data /user/svoruga/input
```

```
hadoop fs -ls /user/svoruga/input
```

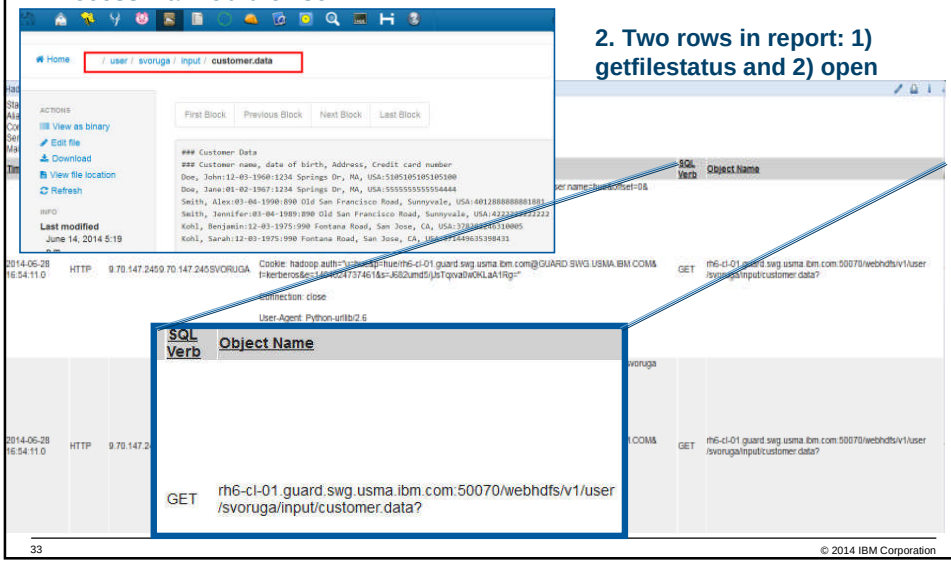
SQL Verb	Object Name
getFileInfo	/user/svoruga/input/customer.data
setOwner	/user/svoruga/input/customer.data
getFileInfo	/user/svoruga/input/customer.data
setPermission	/user/svoruga/input/customer.data
addBlock	/user/svoruga/input /customer.data._COPYING_
complete	/user/svoruga/input /customer.data._COPYING_
create	/user/svoruga/input /customer.data._COPYING_
getFileInfo	/user/svoruga/input



WebHDFS

WebHDFS is an HTTP Rest server bundle.

1. Access via Web browser



2. Two rows in report: 1) getfilestatus and 2) open

Logical MapReduce Example: Word Count

```
map(String key, String value):
// key: document name
// value: document contents
for each word w in value:
  EmitIntermediate(w, "1");

reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values:
  result += ParseInt(v);
Emit(AsString(result));
```

Content of Input Documents

Hello World Bye World

Hello IBM

Map 1 emits:
< Hello, 1>
< World, 1>
< Bye, 1>
< World, 1>

Map 2 emits:
< Hello, 1>
< IBM, 1>

Reduce (final output):

< Bye, 1>
< IBM, 1>
< Hello, 2>
< World, 2>

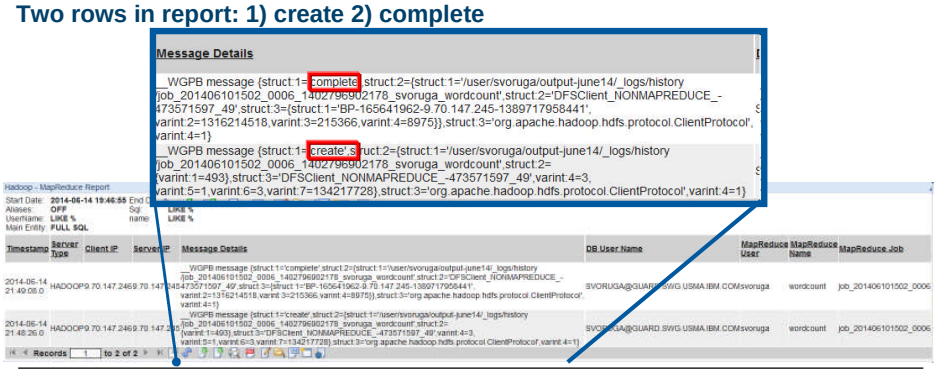
Map Reduce 1

STAP installed on..	Ports	Protocol
Job Tracker (MR1)	8021 50030	Hadoop

Notes and restrictions
- **MapReduce Report filters out noise between create and complete.**
- **HDFS report will show file level accesses**
- **MapReduce name is computed attribute in Hadoop 1.x**
- **MapReduce name is an object (can be used in policies) with Guardium support for Hadoop 2.x**

```
hadoop jar ./wordcount.jar WordCount /user/svoruga/input /user/svoruga/output-june14
```

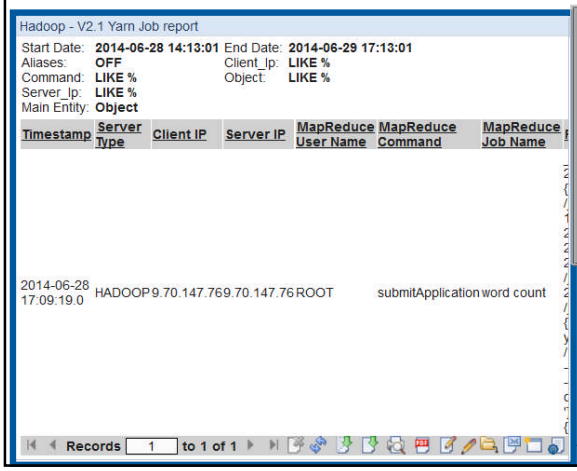
MapReduce 1 example.
Two rows in report: 1) create 2) complete



Mapreduce 2 (YARN)

STAP installed on..	Ports	Protocol
Resource Manager (YARN)	8032	Hadoop

```
hadoop jar hadoop-mapreduce-examples-2.4.0.2.1.1.0-237.jar wordcount
/user/HWtest/input /user/HWtest/wc1
```



Tip on IE configuration.
Check the file:
/etc/hadoop/conf/yarn-site.xml and look for the following:

```
<property>
  <name>yarn.resourcemanager.address</name>
  <value>hw-v2-
03.guard.swg.usma.ibm.com:8050</value>
 </property>
```

Extend the 8000-8021 IE on the machine above to 8050 (or whatever the port is)
OR
Create a new Hadoop IE with the port (if it does not fall in any defined port ranges)

HBase

A column-oriented database management system that runs on top of HDFS.

STAP installed on..	Ports	Protocol
HBase Master	60000 60010	Hadoop
Region Server	60020	Hadoop

Notes:
Nicely formatted list of Hbase shell commands
http://learnhbase.wordpress.com/2013/03/02/hbase-shell-commands/

1. Create an HBase table with two column families

```
create 'Blog', {NAME=>'info'}, {NAME=>'content'}
```

2. Corresponding activity in report

Hadoop - new HBase report

Start Date: **2014-06-29 13:27:34** End Date: **2014-06-30 14:27:34**
Aliases: **OFF** Client_Ip: **LIKE %9.70.147.76%**
Command: **LIKE %** Object: **LIKE Blog%**
Server_Ip: **LIKE %9.70.147.76%**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	DB User Name	SQL Verb ▲	Object Name
2014-06-29 14:17:21.0	HBASE	9.70.147.250	9.70.147.246	SVORUGA@GUARD.SWG.USMA.IBM.COM	createTable	Blog

HBase, continued

3. Add values

```
put 'Blog', 'Matt-001', 'content:post', 'Do elephants like monkeys?'
```

4. Activity report (PUT is sent across as multi)

Hadoop - new HBase report

Start Date: **2014-06-29 13:28:24** End Date: **2014-06-30 14:28:24**
Aliases: **OFF** Client_Ip: **LIKE %9.70.147.76%**
Command: **LIKE %** Object: **LIKE Blog%**
Server_Ip: **LIKE %9.70.147.76%**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	DB User Name	SQL Verb	Object Name
						#
						0208 Matt-001
						7FFFFFFFFF
						FFFFFFFFFF
						FFFFFFFFFF
						FF01000000
2014-06-29 14:17:56.0	HBASE	9.70.147.250	9.70.147.247	SVORUGA@GUARD.SWG.USMA.IBM.COM	multi	0107 content
						0000000100
						0000 A
						000000 A
						0000001F00
						00001A0008Ma

Tip: Extraneous characters in Object Name are cleaned up with Guardium support for Hadoop 2.x..

HBase, continued

5. Retrieve data

```
get 'Blog', 'Michelle-004'
get 'Blog', 'Matt-001'
```

6. Each get in this case has two rows. 1) table object and 2) row object

Hadoop - new HBase report

Start Date: **2014-06-29 13:37:19** End Date: **2014-06-30 14:37:19**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %Blog%**
Server_Ip: **LIKE %**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	DB User Name	SQL Verb	Object Name
2014-06-29 14:35:25.0	HBASE	9.70.147.250	9.70.147.247	SVORUGA@GUARD.SWG.USMA.IBM.COM	get	Blog
2014-06-29 14:35:25.0	HBASE	9.70.147.250	9.70.147.247	SVORUGA@GUARD.SWG.USMA.IBM.COM	get	Matt-001
2014-06-29 14:35:13.0	HBASE	9.70.147.250	9.70.147.247	SVORUGA@GUARD.SWG.USMA.IBM.COM	get	Blog
2014-06-29 14:35:13.0	HBASE	9.70.147.250	9.70.147.247	SVORUGA@GUARD.SWG.USMA.IBM.COM	get	Michelle-004

Hive

Hive allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL. HQL statements are broken down by the Hive service into MapReduce jobs and executed across a Hadoop cluster.

Notes and restrictions
- Hive CLI will be captured as MapReduce, Thrift, and HDFS.

STAP installed on..	Ports	Protocol
Hive Master (Hive Server 2)	10000	Hadoop

1. Query data

hive> `select * from credit_card;`

Ended Job = **job_201407031507_0002**
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Cumulative CPU: 3.26 sec HDFS Read: 3434932 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 260 msec

2. Corresponding MapReduce job

MapReduce Report

Start Date: **2014-07-03 21:09:24** End Date: **2014-07-05 00:09:24**
Aliases: **OFF** Sql: **LIKE %**
UserName: **LIKE %** name: **LIKE %**
Main Entity: **FULL SQL**

Timestamp	Server Type	Client IP	Server IP	DB User Name	MapReduce Job
2014-07-04 00:01:17.0	HADOOP	9.70.147.250	9.70.147.245	SVORUGA@GUARD.SWG.USMA.IBM.COM	job_201407031507_0002
2014-07-04 00:01:17.0	HADOOP	9.70.147.250	9.70.147.245	SVORUGA@GUARD.SWG.USMA.IBM.COM	job_201407031507_0002
2014-07-04 00:01:17.0	HADOOP	9.70.147.250	9.70.147.245	SVORUGA@GUARD.SWG.USMA.IBM.COM	job_201407031507_0002

3. Thrift traffic against Hive Metastore

Hadoop - THRIFT Report

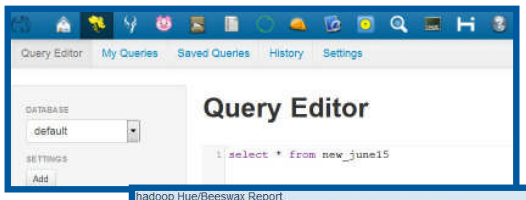
Start Date: **2014-07-03 21:00:55** End Date: **2014-07-05 00:00:55**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %credit_card%**
Server_Ip: **LIKE %**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	SQL Verb	Object Name	Thrift User ▼
2014-07-03 23:39:33.0	HADOOP	9.70.147.250	9.70.147.245	get_table	credit_card	svoruga@GUARD.SWG.USMA.IBM.COM
2014-07-03 23:52:57.0	HADOOP	9.70.147.250	9.70.147.245	get_table	credit_card	svoruga@GUARD.SWG.USMA.IBM.COM

Tip: This report requires configuration of a computed attribute to extract user. See backup slides.

Hive example from Hue/Beeswax

Hue is a web interface to Hadoop. Beeswax is the Hive user interface in Hue.



Notes and restrictions
- This report only valid when MySQL is used as the Hive metastore.
- This report uses computed attributes for user name and dbname, which cannot be used in policies.

hadoop Hue/Beeswax Report

Start Date: **2014-06-16 06:33:40** End Date: **2014-06-17 18:33:40**
Aliases: **OFF** Sql: **LIKE %**
Main Entity: **FULL SQL**

Timestamp	Server Type	Client IP	Server IP	Full SQL ID	Hive User	Hive Command	Hive Database
2014-06-16 09:54:22.0	HADOOP	9.70.147.245	9.70.147.254	44268539	svoruga	'SELECT * FROM `default.new_june15` LIMIT 100',	default
2014-06-16 09:54:15.0	HADOOP	9.70.147.245	9.70.147.254	44268049	svoruga	'CREATE TABLE `default.new_june15` 0A (0A `name` string) 0A ROW FORMAT DELIMITED 0A FIELDS TERMINATED BY 275C 001 270A COLLECTION ITEMS TERMINATED BY 275C 002 270A MAP KEYS TERMINATED BY 275C 003 270A STORED AS TextFile',	default

Pig

A programming language and runtime environment for processing data sets

Notes.
- Captured through HDFS and MapReduce. No special ports required.

1. Load and transform data

```
A = LOAD 'ssn.txt' USING PigStorage() AS (id:int, name:chararray,
SSN:chararray);  --loading
B = FOREACH A GENERATE SSN; -- transforming
DUMP B; -- retrieving
```

Counters:
Total records written : 3
Total bytes written : 15
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201407031507_0006

2. Excerpt from MapReduce report (jobid)

MapReduce Name	MapReduce Job
Job7049598498710416054.jar	job_201407031507_0006

3. HDFS report showing file name as object

Hadoop - new HDFS report

Start Date: **2014-07-07 15:21:45** End Date: **2014-07-08 18:21:45**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %ssn%**
Server_Ip: **LIKE %**
Main Entity: **Object**

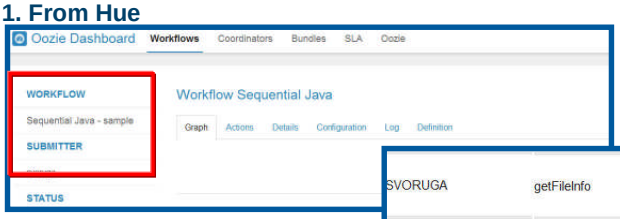
Timestamp	Server Type	Client IP	Server IP	DB User Name	SQL Verb	Object Name
2014-07-07 18:14:14.0	HADOOP	9.70.147.250	9.70.147.245	SVORUGA@GUARD.SWG.USMA.IBM.COM	getBlockLocations	/user/svoruga /input/ssn.txt
2014-07-07 18:14:13.0	HADOOP	9.70.147.250	9.70.147.245	SVORUGA@GUARD.SWG.USMA.IBM.COM	getFileInfo	/user/svoruga /input/ssn.txt

Oozie

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Notes and restrictions
- No special ports required. Traffic will be captured as HDFS and MapReduce.

1. From Hue



2. HDFS report excerpt

SVORUGA	getFileInfo	/user/svoruga/oozie-oozi/0000000-140710110905718-oozie-oozi-W
SVORUGA	getBlockLocations	/user/hue/oozie/deployments/_svoruga_-oozie-8-1405045171.14/workflow.xml
SVORUGA	getFileInfo	/user/hue/oozie/deployments/_svoruga_-oozie-8-1405045171.14
SVORUGA	getFileInfo	/user/hue/oozie/deployments/_svoruga_-oozie-8-1405045171.14/config-default.xml
SVORUGA	getFileInfo	/user/hue/oozie/deployments/_svoruga_-oozie-8-1405045171.14/lib
SVORUGA	getFileInfo	/user/hue/oozie/deployments/_svoruga_-oozie-8-1405045171.14/workflow.xml

Sqoop

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

Notes and restrictions
- No special ports required. Traffic will be captured as HDFS, MapReduce, and Thrift.
- Need Hadoop IE on ports 10000 and 9083

1. Import data from SQL Server into Hive Table

```
sqoop import --connect
"jdbc:sqlserver://9.70.148.102:1433;database=sample;
username=sa;password=Guardium123" --table CC_CALL_LOG --hive-import
```

2. MapReduce report shows jar of same name as table.

DB User Name	MapReduce User	MapReduce Name	MapReduce Job
SVORUGA@GUARD.SWG.USMA.IBM.COM	svoruga	CC_CALL_LOG.jar	job_201407031507_000

3. Thrift message shows activity on the Hive Metastore to create the table

2014-07-03 23:23:20.0	HADOOP	9.70.147.250	9.70.147.245	create_table	CC_CALL_LOG	

Tip: This report requires configuration of a computed attribute to extract user name into separate column. Examine the 'SQL' or 'full sql' column to determine where user name is.

Impala - Cloudera

Cloudera Impala is an open source (MPP) SQL query engine for data stored in hadoop.

STAP installed on..	Ports	Protocol
Impala Daemon nodes	21000	Hadoop

Notes and restrictions:
- Impala sends the statement over the wire in a a 'query' message.
- For now, the actual command is parsed by Guardium as an object.
- User is not sent over the wire on the connection so user name apears as an object rather than in DB User.

```
[rh6-cl-03.guard.swg.usma.ibm.com:21000] > select tab1.col_1, MAX(tab2.col_2), M
IN(tab2.col_2) FROM tab2 JOIN tab1 USING (id) GROUP BY col_1 ORDER BY 1 LIMIT 5;
Query: select tab1.col_1, MAX(tab2.col_2), MIN(tab2.col_2) FROM tab2 JOIN tab1 U
SING (id) GROUP BY col_1 ORDER BY 1 LIMIT 5
+-------+-----------------+-----------------+
| col_1 | max(tab2.col_2) | min(tab2.col_2) |
+-------+-----------------+-----------------+
| false | 243423.325      | 1243.5          |
| true  | 12789.123       | 123.123         |
+-------+-----------------+-----------------+
Returned 2 row(s) in 1.08s
```

In the 'Object' domain report, the same SQL appears twice.

One row where the SQL is the object

SQL	SQL Verb	Object Name
__THRIFT message={method='query',struct:1={string:1='select tab1.col_1, MAX(tab2.col_2), MIN(tab2.col_2) FROM tab2 JOIN tab1 USING (id) GROUP BY col_1 ORDER BY 1 LIMIT 5',list:3={},string:4='svoruga'}}	query	select tab1.col_1, MAX(tab2.col_2), MIN(tab2.col_2) FROM tab2 JOIN tab1 USING (id) GROUP BY col_1 ORDER BY 1 LIMIT 5
__THRIFT message={method='query',struct:1={string:1='select tab1.col_1, MAX(tab2.col_2), MIN(tab2.col_2) FROM tab2 JOIN tab1 USING (id) GROUP BY col_1 ORDER BY 1 LIMIT 5',list:3={},string:4='svoruga'}}	query	svoruga

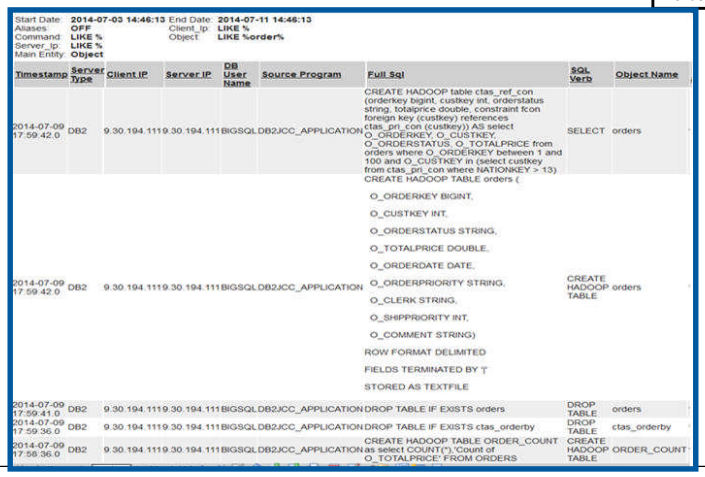
One row where the user is the object

Big SQL V3 (IBM)

IBM's SQL interface to its Hadoop-based platform, InfoSphere BigInsights.

STAP installed on..	Ports	Protocol
Management Nodes and Compute Nodes	51000	DB2

Notes and restrictions:
- Big SQL prior to V3 is not supported.
- Requires 9.1 GPU 210 + patch
- Requires use of DB2 Exit for encrypted connections, Kerberos, GPFS



Start Date: 2014-07-03 14:46:13 End Date: 2014-07-11 14:46:13
Aliases: OFF Client_Ip: LIKE %
Command: LIKE % Object: LIKE %order%
Server_Ip: LIKE %
Main Entity: Object

Timestamp	Server Type	Client IP	Server IP	DB User Name	Source Program	Full Sql	SQL Verb	Object Name
2014-07-09 17:59:42.0	DB2	9.30.194.111	9.30.194.111	BIGSQL	DB2JCC_APPLICATION	CREATE HADOOP table ctas_ref_con (orderkey bigint, custkey int, orderstatus string, totalprice double, constraint fcon foreign key (custkey) references ctas_pri_con (custkey)) AS select O_ORDERKEY, O_CUSTKEY, O_ORDERSTATUS, O_TOTALPRICE from orders where O_ORDERKEY between 1 and 100 and O_CUSTKEY in (select custkey from ctas_pri_con where NATIONKEY > 13)	SELECT	orders
2014-07-09 17:59:42.0	DB2	9.30.194.111	9.30.194.111	BIGSQL	DB2JCC_APPLICATION	CREATE HADOOP TABLE orders (O_ORDERKEY BIGINT, O_CUSTKEY INT, O_ORDERSTATUS STRING, O_TOTALPRICE DOUBLE, O_ORDERDATE DATE, O_ORDERPRIORITY STRING, O_CLERK STRING, O_SHIPPRIORITY INT, O_COMMENT STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE	CREATE HADOOP TABLE	orders
2014-07-09 17:59:41.0	DB2	9.30.194.111	9.30.194.111	BIGSQL	DB2JCC_APPLICATION	DROP TABLE IF EXISTS orders	DROP TABLE	orders
2014-07-09 17:59:36.0	DB2	9.30.194.111	9.30.194.111	BIGSQL	DB2JCC_APPLICATION	DROP TABLE IF EXISTS ctas_orderby	DROP TABLE	ctas_orderby
2014-07-09 17:58:36.0	DB2	9.30.194.111	9.30.194.111	BIGSQL	DB2JCC_APPLICATION	CREATE HADOOP TABLE ORDER_COUNT as select COUNT(*),'Count of O_TOTALPRICE' FROM ORDERS	CREATE HADOOP TABLE	ORDER_COUNT

Other components

Through HDFS

- **Mahout – machine learning**
- **Avro – data serialization**
- **Flume – log aggregator**
- **Spark - advanced DAG execution engine that supports cyclic data flow and in-memory computing.**

Through http

- **SOLR – Messages are large.**

Tip: New projects and components are being added for Hadoop constantly. Before allowing usage in a secure environment in which auditing is required, validate that you can get correct information from HDFS-level auditing/reporting after upgrading to latest Guardium patches.

Reports and their associated queries

Planning for reports

- **Examine the default hadoop and non-Hadoop groups included in your version of Guardium and augment appropriately**
- **Sample queries in this section use groups we created for Read commands, Write commands, etc.**
- **Some reports you see in this presentation use Full SQL as a column. Most of the time this is not needed. It can be overwhelming for auditors and and Guardium will not collect this information anyway if you are not using log full details. You can get what you need from SQL (rather than FULL SQL)**
  - And, yes, columns can be renamed so you are not stuck with the relational metaphor.

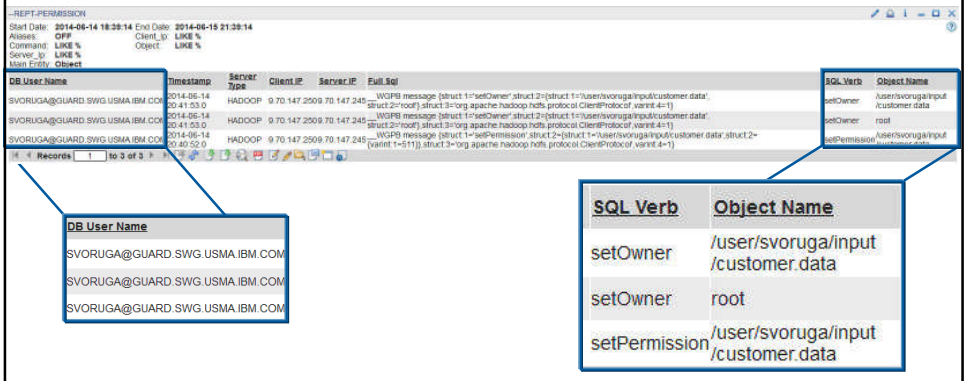
Tip: If you are not familiar with reporting and the impact of policies on data collected for reporting, please see the following resources (log into developerWorks first)

•4-part video series on policies

•Tech Talk: Reporting 101

Permissions report

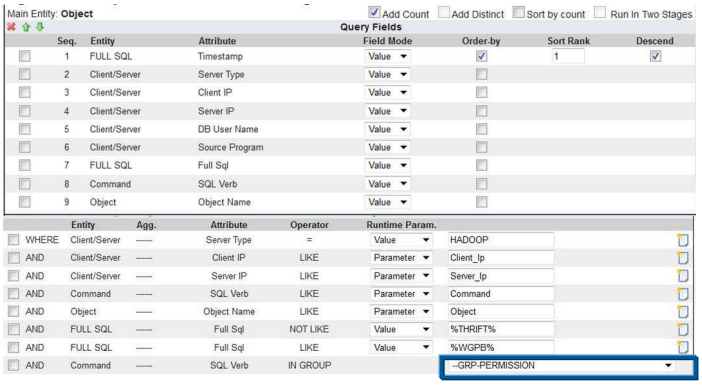
Sundari is granting herself root on customer.data



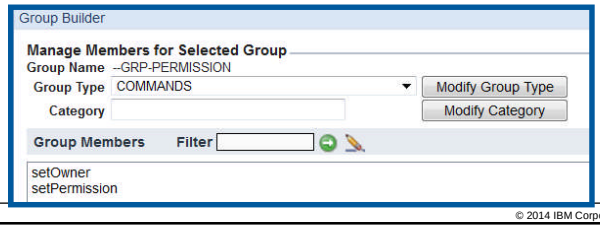
DB User Name
SVORUGA@GUARD.SWG.USMA.IBM.COM
SVORUGA@GUARD.SWG.USMA.IBM.COM
SVORUGA@GUARD.SWG.USMA.IBM.COM

SQL Verb	Object Name
setOwner	/user/svoruga/input /customer.data
setOwner	root
setPermission	/user/svoruga/input /customer.data

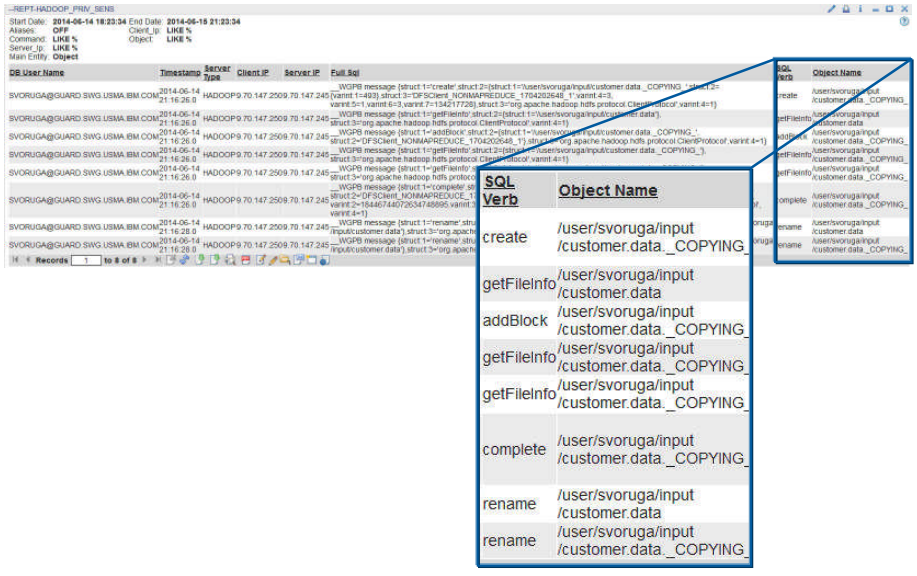
Permissions report query



Group of permission commands



Privileged users accessing sensitive data



Privileged users accessing sensitive data query

Main Entity: **Object** — Add Count | Add Distinct | Sort by count | Run In Two Stages

Query Fields

Seq.	Entity	Attribute	Field Mode	Order-by	Sort Rank	Descend
1	FULL SQL	Timestamp	Value	✓	1	✓
2	Client/Server	Server Type	Value			
3	Client/Server	Client IP	Value			
4	Client/Server	Server IP	Value			
5	Client/Server	DB User Name	Value			
6	Client/Server	Source Program	Value			
7	FULL SQL	Full Sql	Value			
8	Command	SQL Verb	Value			
9	Object	Object Name	Value			

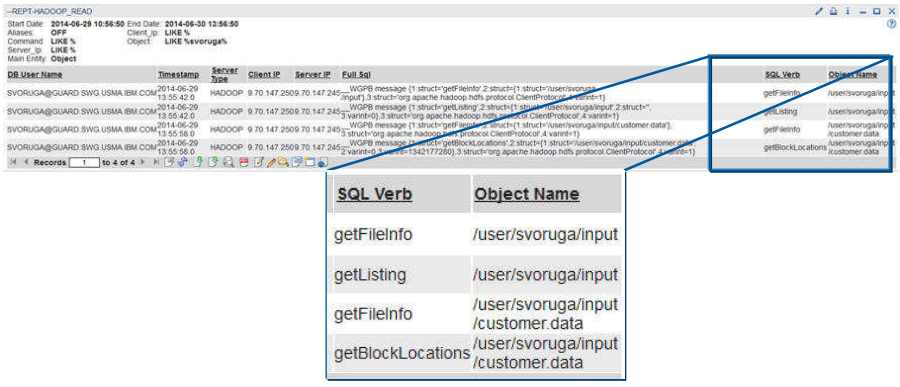
Addition mode: AND | OR | HAVING

Query Conditions

	Entity	Agg.	Attribute	Operator	Runtime Param.	
WHERE	Client/Server	------	Server Type	=	Value	HADOOP
AND	Client/Server	------	Client IP	LIKE	Parameter	Client_Ip
AND	Client/Server	------	Server IP	LIKE	Parameter	Server_Ip
AND	Command	------	SQL Verb	LIKE	Parameter	Command
AND	Object	------	Object Name	LIKE	Parameter	Object
AND	FULL SQL	------	Full Sql	NOT LIKE	Value	%THRIFT%
AND	FULL SQL	------	Full Sql	LIKE	Value	%WGPB%
AND	Client/Server	------	DB User Name	LIKE GROUP		--GRP-PRIV_USERS
AND	Object	------	Object Name	LIKE GROUP		--GRP-SENSITIVE_OBJ

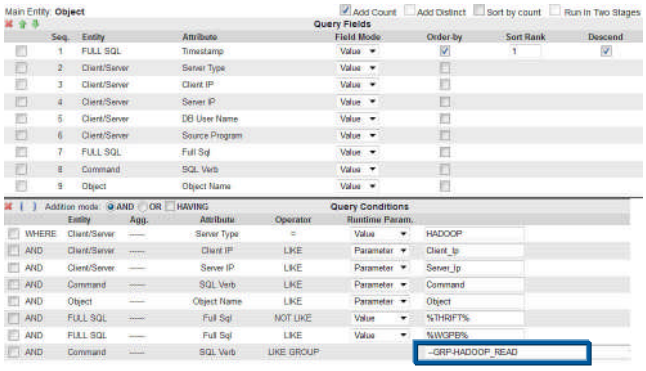
Hadoop Read Activity

```
hadoop fs –cat /user/svoruga/input/customer.data
```

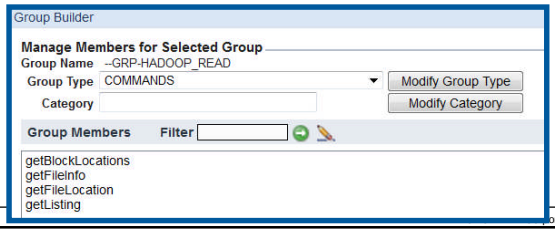


SQL Verb	Object Name
getFileInfo	/user/svoruga/input
getListing	/user/svoruga/input
getFileInfo	/user/svoruga/input /customer.data
getBlockLocations	/user/svoruga/input /customer.data

Hadoop Read Activity Query

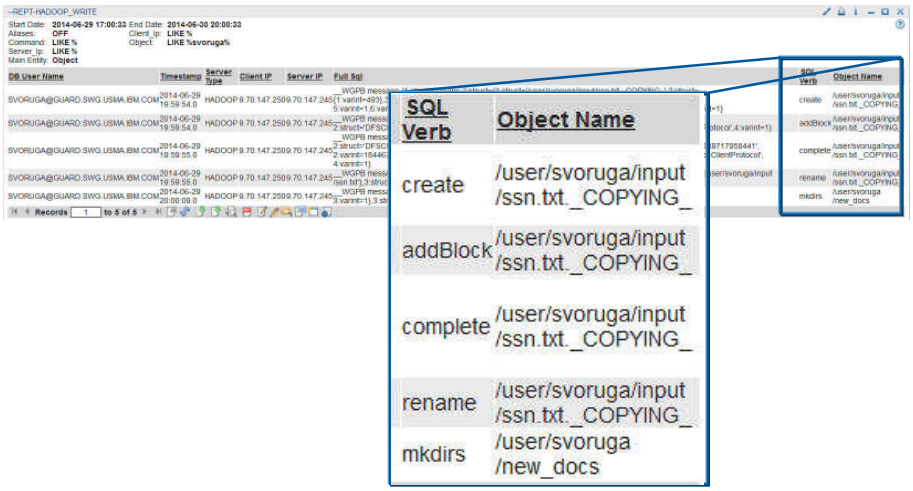


Group of read commands



Hadoop Write Activity

```
hadoop fs -put ssn.txt /user/svoruga/input
hadoop fs -mkdir /user/svoruga/new_docs
```



Hadoop Write Activity query (Conditions only)

Addition mode: AND OR HAVING — Query Conditions

	Entity	Agg.	Attribute	Operator	Runtime Param.	
WHERE	Client/Server	------	Server Type	=	Value	HADOOP
AND	Client/Server	------	Client IP	LIKE	Parameter	Client_Ip
AND	Client/Server	------	Server IP	LIKE	Parameter	Server_Ip
AND	Command	------	SQL Verb	LIKE	Parameter	Command
AND	Object	------	Object Name	LIKE	Parameter	Object
AND	FULL SQL	------	Full Sql	NOT LIKE	Value	%THRIFT%
AND	FULL SQL	------	Full Sql	LIKE	Value	%WGPB%
AND	Command	------	SQL Verb	LIKE GROUP		--GRP-HADOOP_WRITE

Group of write commands

Group Builder

Manage Members for Selected Group

Group Name --GRP-HADOOP_WRITE

Group Type COMMANDS — Modify Group Type

Category — Modify Category

Group Members — Filter

addBlock
complete
create
delete
mkdirs
rename

Permissions exceptions report

```
svoruga@rh6-cli-06:>Hadoop fs –mkdir /user/dgundam/test
```

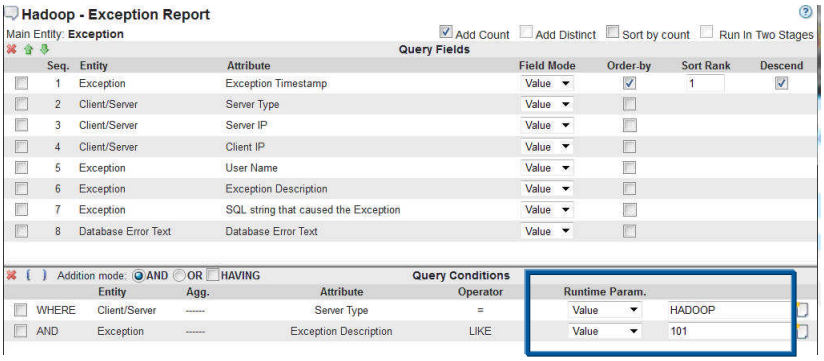
mkdir: Permission denied: user=svoruga…

Hadoop - Exception Report

Start Date: 2014-07-10 09:55:54 End Date: 2014-07-11 12:55:54
Aliases: OFF ExceptionNo: LIKE %
Main Entity: Exception

Exception Timestamp	Server Type	Server IP	Client IP	User Name	Exception Description	SQL string that caused the Exception	Database Error Text
2014-07-10 12:53:52.0	HADOOP	9.70.147.245	9.70.147.250	SVORUGA@GUARD.SWG.USMA.IBM.COM	101	__WGPB message {1:struct='mkdirs',2:struct={1:struct='/user/dgundam /test',2:struct={1:varint=493}, 3:varint=1},3:struct='org.apache.hadoop.hdfs.protocol.ClientProtocol',4:varint=1}	AccessControl Exception

Permissions exceptions query builder



Hadoop Exceptions (General)

-Big Data Exceptions

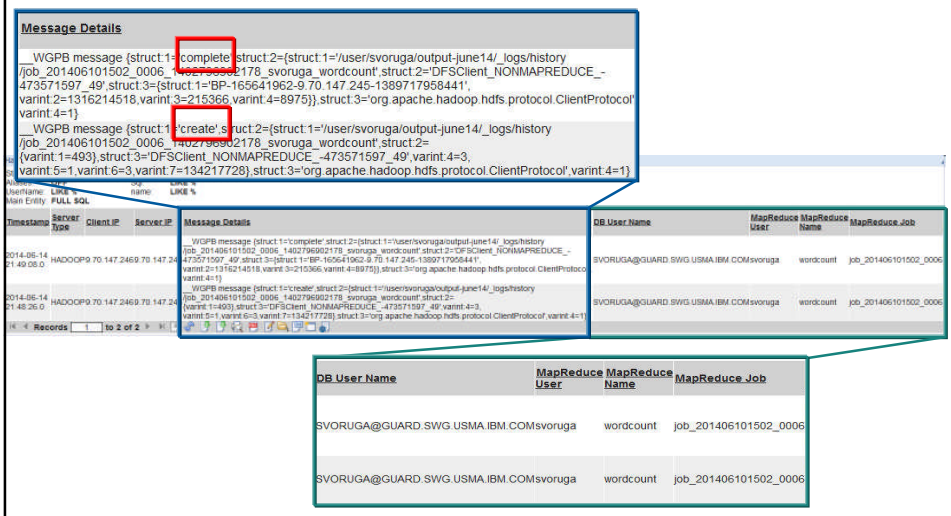
Start Date: **2014-06-09 00:00:00** End Date: **2014-06-12 00:00:00**
Aliases: **ON** EnterExceptionLike: **LIKE %**
ServerIPLike: **LIKE %**
Main Entity: **Exception**

Client IP	Server IP	Server Type	DB User Name	Database Error Text	Count of Exceptions
10.10.9.3	10.10.9.3	HADOOP	BIADMIN	IO Exception	1
10.10.9.3	10.10.9.3	HADOOP	BIADMIN	LeaseExpired Exception	1
10.10.9.3	10.10.9.3	HADOOP	BIADMIN	NotServingRegion Exception	99
10.10.9.3	10.10.9.3	HADOOP	BIADMIN	Severe error	6
10.10.9.3	10.10.9.3	HADOOP	JOE	AccessControl Exception	1

Records 1 to 5 of 5

Hadoop MapReduce (V1)

The prebuilt MapReduce Report includes only a record for the create and one for completion of the job.



Mapreduce 2 (YARN)- not predefined

```
hadoop jar hadoop-mapreduce-examples-2.4.0.2.1.1.0-237.jar wordcount
/user/HWtest/input /user/HWtest/wc1
```

Hadoop - V2.1 Yarn Job report

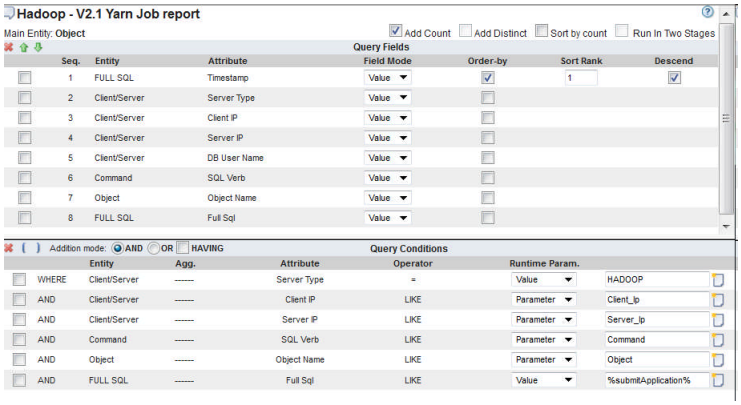
Start Date: **2014-06-28 14:13:01** End Date: **2014-06-29 17:13:01**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %**
Server_Ip: **LIKE %**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	MapReduce User Name	MapReduce Command	MapReduce Job Name
2014-06-28 17:09:19.0	HADOOP	9.70.147.76	9.70.147.76	ROOT	submitApplication	word count

Records 1 to 1 of 1

In contrast to the MapReduce 1 predefined report, YARN output has only row. See query builder on next page.

Mapreduce 2 (YARN) – Query Builder



One possible strategy for testing/deployment

This will evolve as we get feedback from clients and partners, like you!

- **Mandatory to monitor HDFS- you'll need to determine what else you want to monitor based on your auditing requirements**
- **HDFS is simplest to configure and validate (STAP only needed on NameNode and secondary NN)**
- **Add workload and monitor impact on the collectors. Use filtering when possible.**
  - No Ignore Session policy rule for Hadoop! .
- **Add more inspection engines gradually, validate results, and monitor impact on the collectors.**

Ordered steps

1. **Start by monitoring the basics - HDFS on ports 8020, 50070 (HTTP), and HCatalog 50111**
   - Run commands and validate using the HDFS report.
   - Consider adding more work to more closely simulate production and monitor load on the appliance(s).
2. **Add Mapreduce (don't forget to restart S-TAP after adding the new IE)**
   - Run mapreduce jobs and validate using the MapReduce report.
   - Again monitor load on appliance
3. **Add Hiveserver2 (if using)**
   - Run Hive commands and validate using the HDFS and MapReduce reports
4. **Add HBase (if using) – S-TAPs must go on RegionServer nodes!**
   - Run HBase commands and validate using the HBase reports
5. **Add Impala (if using) – S-TAPs on all Impala daemon nodes**
   - Validate using MapReduce and HDFS reports – S-TAPs on all nodes where Impala Daemon is.
6. **Add BigSQL (if using) – S-TAP on management and compute nodes!**
   Validate using database activity reports

Questions?

Reminder: Guardium Tech Talks

Next tech talk: Managing SOX compliance: People, processes, and technology

Speakers: Karl Wehden and Joe DiPietro

Date &Time: Thursday, August 7th, 2014

11:30 AM Eastern Time (60 minutes)

Register here: bit.ly/1xnMMgd

Next tech talk +1: Technical deep dive for InfoSphere Guardium on z

Speakers: Ernie Mancill

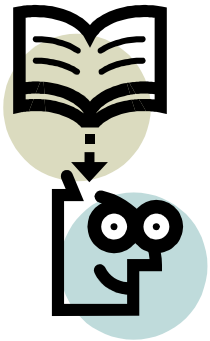
Date &Time: Wednesday, August 27th, 2014

11:30 AM Eastern Time (60 minutes)

Register here: http://bit.ly/1jhYeqZ

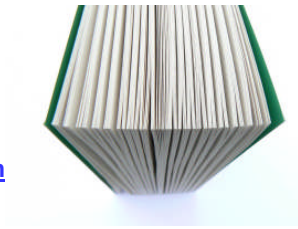
Hadoop Deployment Guide DRAFT

- **For those that are seriously looking at Hadoop deployments with Guardium, we have a draft of a deployment guide that is in constant state of change**
  - Includes much material from this tech talk plus more details and will continue to evolve
- **Cannot distribute widely because of its draft nature.**
- **If you want a draft, and agree that you understand it is a DRAFT to be used for education and planning purposes and not as official product documentation, please send Kathy Zeidenstein an email with your request.**



krzeide@us.ibm.com

Resources



- **Information Governance Principles and Practices for a Big Data Landscape (Redbook)** http://www.redbooks.ibm.com/abstracts/sg248165.html?Open
- **E-book "Planning a security and auditing deployment for Hadoop"** http://www.ibm.com/software/sw-library/en_US/detail/I804665J74548G31.html
- **developerWorks article: "Big Data Security and Auditing with InfoSphere Guardium"** http://www.ibm.com/developerworks/data/library/techarticle/dm-1210bigdatasecurity/

(note: this is Hadoop 1.x)

For more information

- **InfoSphere Guardium YouTube Channel – includes overviews and technical demos**
- **developerWorks forum (very active)**
- **Guardium DAM User Group on Linked-In (very active)**
- **Community on developerWorks (includes content and links to a myriad of sources, articles, etc)**
- **Guardium Info Center**

InfoSphere Guardium Virtual User Group.
Open, technical discussions with other users.

Send a note to bamealm@us.ibm.com if interested.

Existing Hadoop Security Review

History

- Hadoop originally conceived for a trusted environment
- Security infrastructure is growing

Existing Hadoop Distributions – General Capabilities

- Uses LDAP or Kerberos network authentication protocol
- Nodes /Servers In Hadoop cluster prove identity to each other and with users
- Access control for specific MapReduce andHive – more provided by vendors
- Access control to HDFS files or directories
  - In Hadoop 2.4, a POSIX-style ACL list can be used.

Hadoop Security Gaps

- Limited auditing & alerting capabilities (evolving) – not all projects even write audit logs – requires processing to get 'audit-like' information
- Role based access controls to data (evolving)
- Data masking and data encryption (at rest)

> ***"… most are complex to setup and there is still no reference standard across deployments."***
>
> **Maloy Manna, The Business Intelligence Blog, April 2014**
>
> http://biguru.wordpress.com/2014/04/13/basics-of-big-data-part-2-hadoop/

Authentication and Authorization by Default

- **No authentication by default–** A malevolent user can sudo to user HDFS (member of Hadoop supergroup) and delete everything because the cluster doesn't know who he really is.
- **Authorizations:**
  - HDFS file permissions are UNIX-like
  - Access to files and map reduce jobs are controlled via Access Control Lists (ACLs)
  - In Hadoop 2.4, a POSIX-style permission model can be used to enable much finer level control and reduce the overhead of managing hierarchical authorizations.
- **Default in HBASE -** everyone is allowed to read from and write to all tables available in the system.
  - **There are table and row access controls, with cell level being added in 0.98.**
- **Default Hive authorization is to prevent good users from doing bad things…Three methods now include:**
  - Client-side authorization – it is possible to bypass these checks
  - Metastore authorization (server side, to complement HDFS security)
  - SQL-Standards based security (Hive 0.13)

Nice blog from Cloudera explains it in a clear way. https://blog.cloudera.com/blog/2012/03/authorization-and-authentication-in-hadoop/

More info about Hadoop 2.4 enhancements: http://hortonworks.com/blog/hdfs-acls-fine-grained-permissions-hdfs-files-hadoop/

Hive authorization: https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Authorization

Recommended: http://www.slideshare.net/oom65/hadoop-security-architecture

Security is hot topic in Hadoop space

The Apache Software Foundation
http://www.apache.org/

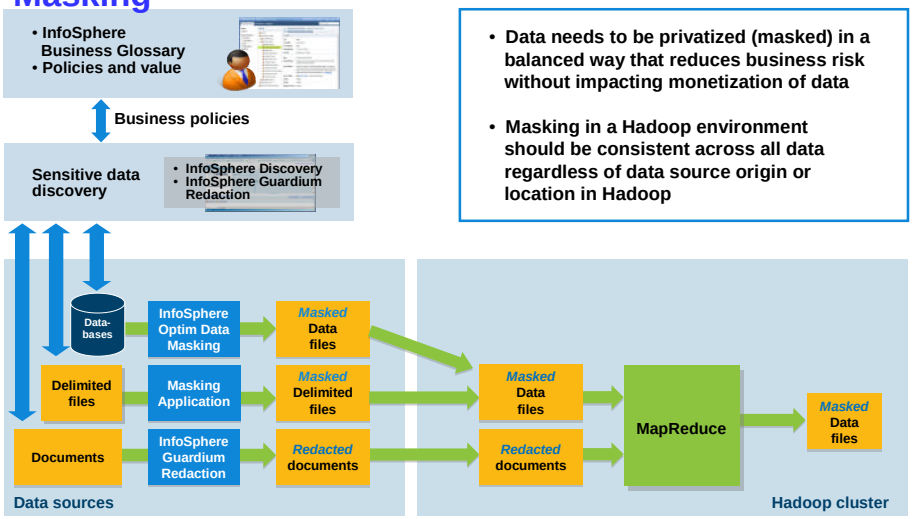
- **Project Rhino (launched by Intel)**
  - https://github.com/intel-hadoop/project-rhino/#apache-projects
- **Sentry for authentication and role-based access control (contributed to open source by Cloudera)**
- **Apache Knox gateway**
  - Provide perimeter security by providing a gateway that exposes access to Hadoop clusters through a Representational State Transfer (REST)-ful API. (incubator- contributions from Hortonworks)
- **Hadoop Cryptographic File system (part of Project Rhino- Intel) –**

"… most are complex to setup and there is still no reference standard across deployments."

Maloy Manna, The Business Intelligence Blog, April 2014

http://biguru.wordpress.com/2014/04/13/basics-of-big-data-part-2-hadoop/

Building a secure Hadoop environment -Masking

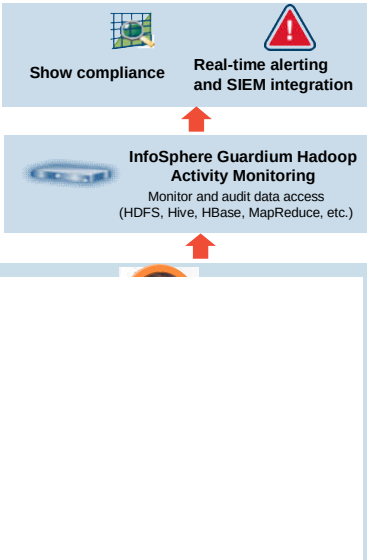


- Data needs to be privatized (masked) in a balanced way that reduces business risk without impacting monetization of data
- Masking in a Hadoop environment should be consistent across all data regardless of data source origin or location in Hadoop

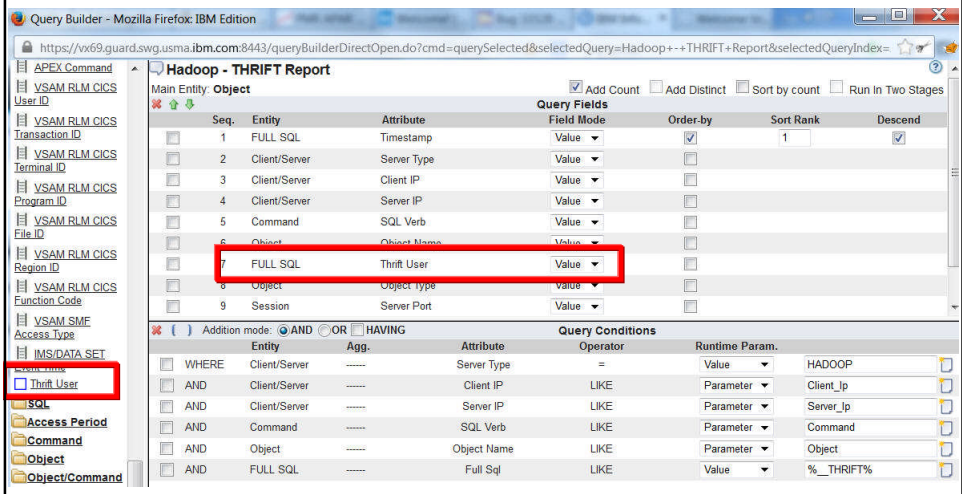
Monitoring a Hadoop environment

Reduce risk by detecting suspicious behavior in real time

- Who is accessing what data?
- Detect and alert on anomalous behavior
- Detect and alert on anomalous connections
- Detailed reporting for compliance and forensics
- Don't let fox watch the henhouse (SOD)



Thrift report query used in our presentation



Thrift user is a computed attribute. For more information on creating a computed attribute for Thrift, see the Hadoop Deployment Guide draft. Email Kathy Zeidenstein for a copy.

Tez - Hortonworks

Tez is an application framework built on Hadoop YARN that can execute complex directed acyclic graphs of general data processing tasks.

Notes and restrictions:
- No special ports
- Activity will be captured via normal mechanisms such as HDFS or YARN

1. Specify Tez as execution engine and run query

```
hive> set hive.execution.engine=tez;
hive> select h.*, b.country, b.hvacproduct, b.buildingage, b.buildingmgr
    > from building b join hvac h
    > on b.buildingid = h.buildingid;
...
Status: Running (application id: application_1396475021085_0018)
Map 1: -/-    Map 2: -/-
Map 1: 0/1    Map 2: 0/1
....
6/18/13 3:33:07 68    69    10    4    3    Brazil JDNS77 28    M3
6/19/13 4:33:07 65    63    7    23    20    Argentina    ACMAX2219    M20
6/20/13 5:33:07 66    66    9    21    3    Brazil JDNS77 28    M3
Time taken: 18.963 seconds, Fetched: 8000 row(s)
```

2. This report shows activity against the Hive metastore (Thrift message)

Spark

Apache Spark is a is a fast and general engine for large-scale data processing.

Notes and restrictions
- No special ports required. Traffic will be captured as HDFS and MapReduce,

```
SPARK_JAR=/opt/cloudera/parcels/CDH-5.0.0-
1.cdh5.0.0.p0.47/lib/spark/assembly/lib/spark-assembly_2.10-0.9.0-cdh5.0.0-
hadoop2.3.0-cdh5.0.0.jar \
/opt/cloudera/parcels/CDH-5.0.0-1.cdh5.0.0.p0.47/lib/spark/bin/spark-class
org.apache.spark.deploy.yarn.Client \
   --jar /opt/cloudera/parcels/CDH-5.0.0-
1.cdh5.0.0.p0.47/lib/spark/examples/lib/spark-examples_2.10-0.9.0-cdh5.0.0.jar \
   --class org.apache.spark.examples.SparkPi \
   --args yarn-standalone \
   --num-workers 3
```

YARN report

Hadoop - V2.1 Yarn Job report

Start Date: **2014-07-11 10:41:49** End Date: **2014-07-12 13:41:49**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %**
Server_Ip: **LIKE %**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	MapReduce User Name	MapReduce Command	MapReduce Job Name
2014-07-11 13:37:55.0	HADOOP	9.70.150.18	9.70.150.14	SVORUGA@CLOUDERA	submitApplication	Spark

Spark, continued

HDFS report

Hadoop - new HDFS report

Start Date: **2014-07-11 10:45:30** End Date: **2014-07-12 13:45:30**
Aliases: **OFF** Client_Ip: **LIKE %**
Command: **LIKE %** Object: **LIKE %1405004851056_0003%**
Server_Ip: **LIKE %**
Main Entity: **Object**

Timestamp	Server Type	Client IP	Server IP	DB User Name ▼	SQL Verb	Object Name
2014-07-11 13:37:54.0	HADOOP	9.70.150.18	9.70.150.12	SVORUGA@CLOUDERA	mkdirs	/user/svoruga/.sparkStaging /application_1405004851056_0003
2014-07-11 13:37:54.0	HADOOP	9.70.150.18	9.70.150.12	SVORUGA@CLOUDERA	setPermission	/user/svoruga/.sparkStaging /application_1405004851056_0003
2014-07-11 13:37:54.0	HADOOP	9.70.150.18	9.70.150.12	SVORUGA@CLOUDERA	getFileInfo	/user/svoruga/.sparkStaging /application_1405004851056_0003 /spark-examples_2.10-0.9.0-cdh5.0.0.jar
2014-07-11 13:37:55.0	HADOOP	9.70.150.18	9.70.150.12	SVORUGA@CLOUDERA	getFileInfo	/user/svoruga/.sparkStaging /application_1405004851056_0003 /spark-examples_2.10-0.9.0-cdh5.0.0.jar
2014-07-11 13:37:54.0	HADOOP	9.70.150.18	9.70.150.12	SVORUGA@CLOUDERA	create	/user/svoruga/.sparkStaging /application_1405004851056_0003 /spark-examples_2.10-0.9.0-cdh5.0.0.jar

Common Hadoop core in most Hadoop Distributions

Component	BigInsights 3.0	HortonWorks HDP 2.1	MapR 3.1	Pivotal HD 2.0	Cloudera CDH5
Hadoop	2.2	2.4	1.0.3	2.2	2.3
HBase	0.96.0	0.98.0	0.94.13	0.96.0	0.96.1
Hive	0.12.0	0.13	0.12	0.12.0	0.12.0
Pig	0.12.0	0.12	0.11.0	0.10.1	0.12.0
Zookeeper	3.4.5	3.4.5	3.4.5	3.4.5	3.4.5
Oozie	4.0.0	4.0.0	3.3.2	4.0.0	4.0.0
Avro	1.7.5	X	X	X	1.7.5
Flume	1.4.0	1.4.0	1.4.0	1.4.0	1.4.0
Sqoop	1.4.4	1.4.4	1.4.4	1.4.2	1.4.4

***Note. MapR uses MapR-FS, which is not currently supported by Guardium DAM**

81