Big data security and auditing with IBM InfoSphere Guardium

Monitor and audit access for Hadoop-based systems

Sundari Voruganti
Kathy Zeidenstein

July 11, 2013
(First published October 11, 2012)

In this article, you will learn how InfoSphere® Guardium® provides database activity monitoring and auditing capabilities that enable you to seamlessly integrate Hadoop data protection into your existing enterprise data security strategy. You will learn how to configure the system and to use InfoSphere Guardium security policies and reports tailored specifically for Hadoop environments, including IBM InfoSphere BigInsights, Cloudera, Hortonworks Data Platform, and Greenplum Hadoop. You will also learn about a quick start monitoring implementation available only with IBM InfoSphere BigInsights.

This article has been updated to reflect changes in InfoSphere Guardium support for Hadoop as delivered in V9.0 GPU 50, including support for Hortonworks Data Platform and Greenplum Hadoop, and enhancements to the integrated Guardium support in InfoSphere BigInsights 2.1.

Overview

The big data buzz has been focused on the infrastructure that supports extreme volume, velocity and variety, and the real-time analytical capabilities enabled by that infrastructure. Even if big data environments like Hadoop are relatively new, the fact of the matter is that data security problems in big data environments are critical to solve up-front. Where there is data, there is the potential for privacy breaches, unauthorized access, or inappropriate access by privileged users.

What's new in V9.0 GPU 50 for Hadoop?

Actually, a lot! With this latest GPU, InfoSphere Guardium has extended its support to more Hadoop-based systems:

- Hortonworks Data Platform 1.2
- Greenplum HD 1.2

This patch also supports later releases of InfoSphere BigInsights (2.1) and Cloudera Hadoop. Notice that the Guardium Proxy in BigInsights 2.1 and later includes support for monitoring HBase events. For more information about InfoSphere Guardium support for these platforms, see the data sheet described in the Related topics section.
Compliance mandates should be enforced the same across big data environments and more traditional data management architectures, and there are no excuses for weakening data security just because the technology is young and evolving. As a matter of fact, as big data environments ingest more data, organizations will face significant risks and threats to the repositories in which the data is kept.

If you're responsible for data security at your organization, you may be required to answer questions such as:

- Who is running specific big data requests? What map-reduce jobs are they running? Are they trying to download all of the sensitive data, or is this a normal marketing query to gain insight into your customers?
- Is there an exceptional number of file permission exceptions, perhaps caused by a hacker algorithmically trying to get access to sensitive data?
- Are these jobs part of an authorized program list accessing the data? Or has some new application been developed that you were previously unaware existed?

What you need is to be able to integrate big data applications and analysis into an existing data security infrastructure, rather than relying on homegrown scripts and monitors, which can be labor-intensive, error-prone, and subject to misuse.

This article takes a look at how IBM InfoSphere Guardium V9, a comprehensive data activity monitoring and compliance solution, can be extended to include access monitoring and reporting for the Hadoop ecosystem.

Although this article covers a high-level overview of InfoSphere Guardium, it does not describe how to install and configure the InfoSphere Guardium Collector. It will describe how to configure the InfoSphere Guardium to monitor supported Hadoop activity and send it to the InfoSphere Guardium Collector for reporting by security analysts. You'll see examples of out-of-the-box reports included to help you get started quickly.

**InfoSphere Guardium in a nutshell**

The IBM InfoSphere Guardium solution continuously monitors database transactions through lightweight software probes, as shown in Figure 1.

**Figure 1. InfoSphere Guardium Data activity monitoring**
These probes (called S-TAPs, for software taps) monitor all database transactions, including those of privileged users, at the operating system kernel level without relying on database audit logs, ensuring separation of duties. The S-TAPs also do not require any changes to the database or its applications.

The probes forward transactions to a hardened Collector (an appliance) on the network, where they are compared to previously defined policies to detect violations. The system can respond with a variety of policy-based actions, including generating an alert.

InfoSphere Guardium supports a wide variety of deployments to support very large and geographically distributed infrastructures. Because this article has barely scratched the surface of what InfoSphere Guardium can do, you can review the Related topics section for links to more information about the capabilities of InfoSphere Guardium. Note that not all capabilities are available for all data sources.

**Benefits of using InfoSphere Guardium for Hadoop monitoring**

Using InfoSphere Guardium can dramatically simplify your path to audit-readiness by providing targeted, actionable information. You can imagine that if your current Hadoop audit readiness plan is based on zipping up log data and hoping that you never need it, you will probably not be able to satisfy many audit requirements from a timeliness perspective alone. Forensic analysis would no doubt be time-consuming and require homegrown scripts that are taking away resources you would rather have spending on creating business advantage around Hadoop.

With InfoSphere Guardium, much of the heavy lifting is taken care of for you. You define security policies that specify what data needs to be kept and how to react to policy violations. Data events are written directly to the InfoSphere Guardium collector, leaving no opportunity for even privileged users to access that data and hide their tracks. Out-of-the-box reports get you up and running with Hadoop monitoring quickly, and those reports are easily customized to align with your audit requirements.

The InfoSphere Guardium S-TAP was originally designed for performance with low overhead; after all, the S-TAP is also used to monitor product database environments. With Hadoop, you will not likely see overhead above 3%, which for most Hadoop workloads will be unnoticeable.

Finally, InfoSphere Guardium provides monitoring capabilities throughout the Hadoop stack, from the user interface through to storage, as shown in Figure 2.
Why is this important? Even though much of the activity in Hadoop breaks down to MapReduce and HDFS, at that level, you may not be able to tell what a user higher up in the stack was really trying to do, or even who the user was. It is similar to showing a bunch of disk segment I/O operations instead of an audit trail of a database. So by providing monitoring at different levels, you are more likely to understand the activity as well as being able to audit on activities that come in directly through lower points in the stack.

**Hadoop activity monitoring**

The events that can be monitored include the following:

- Session and user information.
- HDFS operations – Commands (cat, tail, chmod, chown, expunge, and so on).
- MapReduce jobs - Job, operations, permissions.
- Exceptions, such as authorization failures.
- Hive/HBase queries - Alter, count, create, drop, get, put, list, and so on.

The following examples describe how some simple Hadoop commands are shown in InfoSphere Guardium reports.

**Terminology**

If you are new to InfoSphere Guardium, you may be surprised to see relational database terminology used occasionally in the reports and policy rules. Even though SQL is not used for file system data, the use of common terminology enables Guardium to provide a cross-database activity view. It is very easy to customize report column headers and contents to your preferences.

HBase: The following is a create in HBase:

```
create 'test_hbase', 'test_col'.
```

InfoSphere Guardium will show the actual command that flowed to HBase, as shown in Figure 3.
Figure 3. HBase report

HDFS: The following is a simple `ls` command in Hadoop:

```
hadoop fs -ls
```

Figure 4 is the output in an InfoSphere Guardium report.

Figure 4. HDFS ls command

You can see that under the covers, it was broken down to two different commands to get the listing and the associated file information.

Behind this seemingly simple activity monitoring is a powerful and flexible infrastructure for policy configuration and for reporting. For example, later in this article you will learn how to create a policy that will log an event to alert you whenever an unknown user accesses sensitive data. You can also create an audit report that helps you detect when new or unknown applications are accessing Hadoop data.

Quick start activity monitoring for IBM InfoSphere BigInsights

IBM InfoSphere BigInsights includes an integrated capability called the Guardium Proxy to read and send log messages to InfoSphere Guardium for analysis and reporting. With the proxy, BigInsights sends messages from Hadoop logs to the InfoSphere Guardium collector.

The advantages to the proxy include the following:

- Easy to get up and running. There is no need to install S-TAPs or configure ports. you simply enable the proxy on the NameNode, and you are ready to go.
- Because the proxy uses Apache log data as messages to send to InfoSphere Guardium, there is less noise that is required to be filtered out from those messages, such as status and heartbeat information.
- There is no delay of Guardium support for new releases of BigInsights to take advantages of message protocol changes.

Limitations: Because Hadoop is not logging exceptions to its logs, there is no way to send exceptions to InfoSphere Guardium. If you require exception reporting, you will need to implement an S-TAP. In addition, there is no support for monitoring Hive queries, although you will see the underlying MapReduce or HDFS messages from Hive.

If you are interested in getting started by using the Guardium Proxy in InfoSphere BigInsights, see Appendix A, which has the configuration instructions to enable the proxy for the Hadoop services.
Prerequisites

The following section describes the requirements for both InfoSphere Guardium and the Hadoop-based system.

InfoSphere Guardium security and compliance solution

The IBM InfoSphere Guardium solution is available as follows:

- Hardware offering – a fully configured software solution delivered on physical appliances provided by IBM.
- Software offering – the solution delivered as software images you can deploy on your own hardware either directly or as virtual appliances.

To monitor Hadoop environments, you must have an InfoSphere Guardium Appliance V9.0 (hardware or software) with at least Patch 2 and preferably GPU Patch 50. That appliance should be configured as a collector. You will also need the InfoSphere Guardium Standard Activity Monitor for Hadoop software entitlement. Before attempting to monitor Hadoop, please ensure that you check the IBM support site for additional patches that may be required.

As your system grows, you can also obtain appliances configured as a Central Manager and Aggregator, which provides centralized management of multiple collectors via a single web-based console, effectively creating a federated system from multiple collectors. You can use that to centrally manage security policies and appliance settings such as archiving schedules, patch installations, user management, and so on. It also aggregates raw data and reports from multiple collectors to generate holistic, enterprise-level audit reports.

This article does not cover the installation and configuration of the IBM InfoSphere Guardium appliance and assumes that you have at least one appliance connected to the Hadoop cluster on the network.

Supported Hadoop-based systems

See the InfoSphere Guardium system requirements information on the IBM support site for updates to the supported release levels of the supported Hadoop distributions for InfoSphere BigInsights, Cloudera, Greenplum HD, and Hortonworks.

Note: For IBM InfoSphere Big Insights, support of the overlay install on Cloudera is also supported by IBM InfoSphere Guardium.

Configuring data activity monitoring

The steps required to install and configure are as follows:

1. **Plan** – Make sure you have a good understanding of the network architecture of your Hadoop cluster, including IP addresses and the relevant port numbers.
2. **Install S-TAP** and configure inspection engines on the appropriate Hadoop nodes.
3. **Validate** that activity is being monitored, by creating and reviewing activity reports.
4. **Install a security policy.**

**Plan**

The planning step is critical for a successful integration of InfoSphere Guardium with Hadoop. The following section provides a high-level overview of the architecture to provide you with the understanding you need.

Recommendation: For an initial deployment, consider just starting with the simplest configuration that supports a particular business requirement, and then expand from there. For example, start with just the requirements to monitor HDFS and MapReduce, validate the configuration, and then expand to include Hive and HBase as needed.

Figure 5 shows you where the OS-specific S-TAPs are specifically required to be installed in the cluster for complete monitoring coverage as provided by InfoSphere Guardium.

**Figure 5. STAPs required for monitoring the Hadoop stack**

IBM InfoSphere Guardium provides a centralized solution for installing and updating multiple S-TAPs using the Guardium Installation Manager to make S-TAP management simpler and more automated.

Note: For the slave nodes, S-TAP is only required for HBase Region Servers for monitoring inserts (HBase Puts).

After you install the OS-specific S-TAP on the relevant nodes, you can configure the ports that the S-TAP is monitoring by defining what is known as inspection engines for the S-TAP. These inspection engines also have specific monitoring protocols associated with them. The S-TAP intercepts the network packets, makes a copy, and does some parsing and analysis and sends the information to the InfoSphere Guardium Collector where it is further parsed, analyzed and stored in the InfoSphere Guardium Collector local database.

Before moving to the next step, review the following:

- Ensure that you are running a supported Hadoop-based system.
• Make sure that you know the IP addresses of the InfoSphere Guardium Collector(s) that will be receiving the collected traffic from your Hadoop cluster.
• Make sure that you know the IP addresses of the servers on which S-TAPs are required.
• Write down the ports to be monitored and which hosts they apply to, based on information shown in Table 1 and Table 2. This article based its port setups on the Hadoop default ports, which in general are the same across the distributions. Your configuration may differ.

Table 1. Hadoop service ports to monitor

<table>
<thead>
<tr>
<th>Service</th>
<th>Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS Name Node</td>
<td>8020, 50470</td>
</tr>
<tr>
<td>HDFS Thrift plugin for Cloudera Hue (NameNode)</td>
<td>10090</td>
</tr>
<tr>
<td>MapReduce Job Tracker</td>
<td>8021, 9290, and 50030</td>
</tr>
<tr>
<td>HBase Master</td>
<td>60000 and 60010</td>
</tr>
<tr>
<td>HBase Region</td>
<td>60020</td>
</tr>
<tr>
<td>HTTP port (for WebHDFS)</td>
<td>50070</td>
</tr>
<tr>
<td>HBase Thrift plugin</td>
<td>9090</td>
</tr>
<tr>
<td>Hive Server</td>
<td>10000</td>
</tr>
<tr>
<td>Beeswax Server</td>
<td>8002</td>
</tr>
<tr>
<td>Cloudera Manager Agent</td>
<td>9001</td>
</tr>
</tbody>
</table>

Install S-TAP and configure inspection engines

S-TAPs are operating system specific, so you'll need to install the Red Hat or SUSE Linux S-TAP for each of the appropriate nodes. This process is well-documented in the InfoSphere Guardium S-TAP help book, and can also be done using InfoSphere Guardium Installation Manager or by using a non-interactive install process that lets you install on many nodes with the same command.

Next, you need to configure inspection engines appropriate for the node and services being monitored. Inspection engines are where you indicate which protocol to use for monitoring (Hadoop), and where you define which ports to monitor. Table 1 showed an excerpt of the ports used by default and that InfoSphere Guardium can monitor. Your ports may differ.

Table 2 shows you the information that was used to configure the Hadoop cluster for this article, and is based on default Hadoop ports.

Table 2. Hadoop service ports to monitor to configure Hadoop cluster

<table>
<thead>
<tr>
<th>Inspection engine for...</th>
<th>Protocol</th>
<th>Port range.</th>
<th>KTAP DB Real port</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS, Job Tracker, Beeswax server</td>
<td>HADOOP</td>
<td>8000-8021</td>
<td>8021</td>
</tr>
<tr>
<td>MapReduce Master and Thrift plug-in</td>
<td>HADOOP</td>
<td>9000-9291</td>
<td>9291</td>
</tr>
<tr>
<td>Hive Server and HDFS Thrift Plug-in for Hue</td>
<td>HADOOP</td>
<td>10000-10090</td>
<td>10090</td>
</tr>
</tbody>
</table>
Recommendation: You can specify multiple inspection engines per server; you should do this when the protocol is the same, and you want to avoid configuring too large a port range for each inspection engine. It is a best practice to not configure many ports you don’t need since this places additional overhead on the InfoSphere Guardium collector components, which would need to analyze traffic that is not relevant. However, for simplicity, you may want to include port ranges on some of the inspection engines where it makes sense.

You can add inspection engines from the user interface: Administration Console > Local Taps > S-TAP Control > Add Inspection Engine.

Or you can use an API, create_stap_inspection_engine. See Appendix B for example API commands that you can use to create the inspection engines using default ports.

Figure 6 shows a few examples of some of the inspection engines after they were defined.

Figure 6. A sample of some inspection engines for Hadoop

You can read more about the inspection engine configuration fields in the S-TAP help book, which you can find online. However, the following is a summary of some of the key fields.

- **Protocol**: The type of data source being monitored (Hadoop). The options are available as a pull-down menu.
- **Port Range**: The range of ports monitored for this inspection engine. As mentioned previously, keep this range as limited as possible. For this article, The applicable ports were divided up into closely corresponding groups, such as the 9000 range or the 50000 range.
- **K-TAP real port**: This parameter should just be set as the last port in the range for that inspection engine. If just one port is defined, then set K-TAP real port to be the same.
- **Client IP Addresses/Masks**: Each inspection engine monitors traffic between one or more client and server IP addresses. This field acts as a filter to define and restrict clients to be monitored. For example, you might have some trusted clients that don’t require auditing,
and you can filter out those clients ahead of time, which can reduce the overall load on the collector. The IP address is a single location and the mask works as a wildcard to let you define a range of IP addresses. A mask 255.255.255.255 (which has no zero bits) identifies only the single address specified by IP address. In the case of article, it is using 0.0.0.0 for both client and mask so all clients will be monitored.

- **Connect to IP**: The IP address for S-TAP to use to connect to the monitored data source. For Hadoop, you can use the default, 127.0.0.1.
- **Process name**: For a Hadoop configuration, you do not need this.

**Validate that activity is being monitored**

As an Administrator, navigate to the **System View** tab of the InfoSphere Guardium web console and ensure that the S-TAPs for your Hadoop cluster are active and showing green, which indicates that the S-TAP is connected to the InfoSphere Guardium collector. Figure 7 shows what this might look like for one host.

**Figure 7. S-TAP status monitor**

![S-TAP status monitor](image)

After you validate that the S-TAPs are configured properly on all applicable nodes, you should already be capturing any work that is running on the system. You can run a shell command or the sample wordcount job to validate that you are seeing data. In either case, you will need to use InfoSphere Guardium drill-down reports (available from the View tab for users), or create your own reports to view the activity.

More detail on the Hadoop reports is described in the [Hadoop reports included with InfoSphere Guardium](#) section. For the purposes of validation, this article will describe how to use the drill-down reports that are available for security administrators who are assigned user roles in the system.

When you log in as a user and click the View tab, you will see a graph much like what is shown in Figure 8. Double-click the graph to drill down into the details.
Figure 8. Drill down into details

There are many paths through the data. Figure 9 shows one sample drill down.

Figure 9. Sample drill down

Whenever you click on a row in the report, you have a menu of options to choose from in terms of the next level of reporting you would like to see.

Hadoop reports included with InfoSphere Guardium

InfoSphere Guardium includes several out-of-the-box reports for Hadoop, including the following:

- MapReduce activity.
• Unauthorized MapReduce jobs.
• Hue/Beeswax activity.
• HDFS, HBase, and Hive activity.
• Exception report.

If you are logged in as a user, you can find the predefined reports by clicking the View tab. From the left navigation pane, click Hadoop, and the reports are listed there.

If you are logged in as an Admin, you will need to add the reports to your console. The following steps assume you have a My New Reports tab already defined on your console, and that you are logged in as administrator.

2. In the report title section, use the pull-down menu to locate one of the reports, such as the Hadoop - Hue/Beeswax report, and then click Search.
3. In the report search results window, click the Add to My New Reports button, as shown in Figure 10.

**Figure 10. Add the report to a pane called My New Reports**

4. Now you can run a command in Beeswax using Hue and see the report. For example, in this article, the following Hive command was entered, as shown in Figure 11.
5. Go to the *Hue/Beeswax* report, you will likely see **No data found**. This is because you need to specify some runtime parameters to tell the system what to display. To do this, click on the pencil icon to customize the report query as shown in Figure 12.

**Figure 12. Submitting a query in Beeswax**

6. Add a time period for the query from and to dates (depending on your workload, you may want to pick a smaller value, perhaps hours or a day) and the percent sign or other search parameters for the LIKE field for the SQL and Table_Name fields, as shown in Figure 13.

**Figure 13. Specify runtime parameters for Hue/Beeswax report**

7. You should now see some data appearing in the report, as shown in Figure 14.

**Figure 14. Hue/Beeswax report**
8. Now do the same steps for the **MapReduce** report (if you are an admin):
   b. Search for **MapReduce** report.
   c. Add to a report pane.
   d. Edit the report to add runtime parameters.

9. Run a MapReduce job. This article used the sample word count program in Cloudera. The syntax to run wordcount is: 
   ```text
   bin/hadoop jar hadoop-*-examples.jar wordcount \n   ```
   in-dir \n   out-dir.

10. For this article, the following was run: 
    ```text
    hadoop jar hadoop-0.20.2-cdh3u4-examples.jar \n    wordcount /user/svoruga /user/svoruga/wc100
    ```
    You can see a report much like what is shown Figure 15.

**Figure 15. MapReduce report**

(For a larger version of Figure 15.)

As you can see, for this article, the query parameters were customized to specify that only activity in which `svoruga` and `word%count` appear in the message (*Full SQL*) is to be returned on the report.

**Troubleshooting**

The InfoSphere Guardium Hue/Beeswax reports assume the use of the Thrift message format and the MySQL database. If you are using MySQL and your Hue/Beeswax report still doesn't show data, you may need to configure Beeswax to use port 8002 as follows, which was the port used by Thrift for the system example in this article.

1. Navigate to the Hue .ini file:
   - For CDH3: `/etc/hue/hue-beeswax.ini`.
   - For CDH4: `/etc/hue/hue/ini`, where the `-hadoop "examples.jar"` is in the `/user/lib/hadoop` directory. Replace with the correct jar file. 
   - `in-dir` is the HDFS directory where the input file is.
   - `out-dir` is the HDFS directory where the output file will be placed.

2. Uncomment the following line:
   ```text
   beeswax_server_port=8002
   ```

3. Stop and restart hue using the following commands:
   ```text
   /etc/init.d/hue stop
   ```
Install a security policy

In InfoSphere Guardium, a security policy contains an ordered set of rules to be applied to the observed traffic between clients and servers. One or more rules are combined to create a policy. For the Hadoop security policy in this article, access rules were defined, which are rules to help reduce the amount of traffic to be logged to the InfoSphere Guardium collector.

**Recommendation:** Do not modify the sample policy. Instead, create a clone and use that as the basis for your modification.

To access the Hadoop Policy and create a clone, do the following.

1. Log in as an administrator and navigate to Tools>Config & Control>Policy Builder.
2. From the Policy Finder, select Hadoop Policy and then click the Clone button.
3. Enter a new name for the policy and then click Save.

To install a policy, do the following.

1. Log in as an administrator and go to Administration Console > Configuration > Policy Installation.
2. Select the Hadoop policy clone you created and choose the appropriate install action. See the online help for more information about policy installation and the implications of having more than one policy.

The rules for the Hadoop Policy are shown in Figure 16. Click on the plus to see more details. You can edit the rule by clicking on the pencil icon.

**Figure 16. Rules in the sample Hadoop policy**

The following is a summary of each of the rules in the policy.

- Access Rule: Low interest objects: Allow
  Figure 17 shows the rule definition.
The following are the two main items of interest in this policy.

- A definition of a group of objects, such as user preferences, that is unlikely to be of interest. If you click on the group builder icon, you can see the objects that are part of the HadoopSkipObjects group, as shown in Figure 18.
Figure 18. Low interest objects rules for Hadoop

You can modify this group as needed.

- The Allow action means that a policy violation will not be logged for these objects, and they will not be considered for further analysis.

- Access Rule: Low Interest Commands: Allow
  Similar to the rule above, but specifically for commands.

- Access Rule: Filter based on Server IP: Log Full Details
  This rule enables you to filter out activity from any non-Hadoop servers that are using this same Guardium Collector.

**Important:** You *must* modify the **Not Hadoop Servers** group to include all the IPs of any servers you want to filter out. If there are no such servers, then enter a dummy IP, but not 0.0.0.0. If you do not have something in that group, then your reports will not work.

**Cool things you can do**

The following are a few key things you can do with InfoSphere Guardium to help you meet your audit and compliance requirements for Hadoop. This section describes ways to answer the questions that were posed at the beginning of the article.

- Is someone who is not previously authorized to do so accessing sensitive data?
- Is there a new application/job accessing the system?
- Is there an exceptional number of file permission errors?
Tell me when an unauthorized user accesses sensitive data

There are many different rules you can use to create policies that can help you enforce your auditing requirements.

**Tip:** If you add any rules to your Hadoop policy clone, make sure the previous rule has **Continue to next rule** selected. Otherwise, your new rule may never be evaluated.

Figure 19 shows a rule in which two groups are defined as follows.

- Known Hadoop users
- Known sensitive data objects/files

**Figure 19. Example policy rule for access to sensitive files**

The rule has a negation for the known users, which means that if a user who is not part of that known group accesses those sensitive files, that information will be logged, and you can see those occurrences in an incident report for further investigation. If it turns out that the access is legitimate, you can add that user to the known group.

Tell me when new MapReduce jobs use the system

Many enterprises are concerned about keeping track of new applications that access their data, and an automated report can help you do that. InfoSphere Guardium provides an unauthorized MapReduce job report that you can customize to help you identify when new MapReduce jobs enter the system.
You can schedule this report to run regularly as part of an audit process that runs in the background. This enables you to be notified when new jobs enter the system, so they can be properly reviewed and added to the authorized job list as appropriate.

Setting up this report takes a bit of configuration. You need to create and customize a group called Hadoop Authorized Job List. You will need to:

1. Create and Populate that group with the list of known and approved jobs in your system. (Note: With 9.0 GPU 50, the Hadoop Authorized Job List is included with the system. You will just need to populate it.)
2. Assign roles to the group so appropriate people in your organization can see and use that group in building reports.
3. Customize the Hadoop-Unauthorized MapReduce Jobs report to include that group as a runtime parameter.

Here are the detailed steps on how to configure the group:

1. From the administration console, go to Tools > Config and Control > Group Builder. Or, if you are logged in as a user, go to the Monitor/Audit > Build Reports > Group Builder, and then click Next.
2. In the Create New Group fields, specify Public as the Application Type, give it the name you want (such as Hadoop Authorized Job List), and from the dropdown list for Group Type Description, select OBJECTS, as shown in Figure 20. Click the Add button.

**Figure 20.Naming the new group**

3. In the Manage Members pane, enter a MapReduce job name in the Create & add new members field, and then click Add to add that member to the group. Continue adding names, as shown in Figure 21. When you are done adding MapReduce job names, click the Back button.
Figure 21. Populate the group with authorized jobs

4. In the Group Builder, find your group in the Modify Existing Group list and then click the Roles button as shown in Figure 22.

Figure 22. Associate roles with the group

5. Select the roles you want to be able to use this group. We have simply selected All Roles, as shown in Figure 23. Click the Apply button.
Figure 23. Indicate which roles can use this group

Now you have finished with the task of creating the Hadoop Authorized Job List group, and you are ready to move to the next task, which is to associate it with the report.

1. As described in the Hadoop reports included with InfoSphere Guardium section, if you are logged in as a user, you can find the predefined reports by clicking the View tab. From the left navigation pane, click Hadoop, and the reports are listed there.

2. Click on Hadoop – Unauthorized MapReduce Jobs. It will likely show No data found. Click on the pencil icon to customize this report, as shown in Figure 24.

Figure 24. Customize the report

3. Select the group name from the list, as shown in Figure 25. Make sure the date parameters cover a time period when you know you will see at least a small set of results to validate that the report is working. Then click the Update button.
Figure 25. Add the group to the report runtime parameters

4. From the left navigation, click on the **Hadoop – Unauthorized MapReduce Jobs** report again. It should be populated with data from any reports that are not in your authorized job group. An extract of the report is shown in Figure 26, where you can see that a job named PiEstimator is shown because it was not on the authorized list of jobs.

**Figure 26. Report includes activity from jobs not in the authorized group**
Tell me if there is an exceptional number of file permission errors

InfoSphere Guardium includes out-of-the-box exception reporting for Hadoop. For example, if you are logged in as a user, you can go to View > Hadoop > Hadoop - Exception Report to see the out-of-the-box report, similar to what is shown in Figure 27.

Figure 27. Sample Hadoop exception report

You can also create an alert based on the same query that is used for the report. With an alert, you can have an email sent whenever a threshold for a specific condition, such as file permission exceptions, goes above a certain limit.

You can also choose to log the alert as a policy violation, which will put this alert on the Incident Management tab of the InfoSphere Guardium web console.

Here are the high-level steps to create the exceptions query and to enable it in an alert.

1. Navigate to the Alert Builder:
   • For an administrator, go to Tools > Config and Control > Alert Builder.
   • For a user, go to Protect > Correlation Alert > Alert Builder.
2. From the Alert Finder, click New.
3. In the Query Definition section of the Add Alert screen, select Hadoop – Exception Report from the pull-down menu, as shown in Figure 28, and fill out the rest of the alert requirements.

Figure 28. Use exception report query to build your alert

Figure 29 is an example of an alert that was created for this article that specifies an exception of 101 for file permission exceptions.
Notice that the alerts are logged as a policy violation so that any alerts that are triggered also appear from the Incident Management tab. Also, notice at the bottom of the example, the administrator named David Roz will get at least one email when the alert is triggered.

**Conclusion**

We hope you’ve enjoyed this tour through InfoSphere Guardium for securing Hadoop environments. If you are using or evaluating Hadoop and are considering a security strategy around its deployment, we think the information provided in this article can help you think about what you need and how InfoSphere Guardium can help. Existing Guardium users can easily extend their current data security and audit processes to include Hadoop.

**Acknowledgements**

The authors would like to extend their gratitude to the following people without whom this article would never have seen the light of day:
Appendix A: Configuring the Guardium proxy in IBM InfoSphere BigInsights

This appendix describes the steps to enable the Guardium Proxy in IBM InfoSphere BigInsights to send copies of relevant log messages to InfoSphere Guardium.

Figure 30 shows you the architecture of the solution.

Figure 30. Log messages are sent to the Guardium Proxy and then forwarded to the Guardium Collector

Enabling the integration between InfoSphere BigInsights and InfoSphere Guardium is much simpler as of BigInsights 2.0. You enable the Guardium Proxy at BigInsights installation time (BigInsights 2.0 and later). Logging events are sent over a socket connection. Port 16015 is used for this socket connection. The proxy then forwards those messages to the InfoSphere Guardium collector (default port 16016) which parses and stores those messages in the Guardium internal tables for reporting, alerting, and so on.

The screenshot below is an excerpt from the InfoSphere BigInsights 2.1 installation panel in which you specify the port addresses of the proxy, the Guardium collector, and the host names for the collector and the node on which you run the proxy (usually the name node).
You can find details of the integration in the IBM InfoSphere BigInsights Information Center (see Related topics).

**Validate the configuration**

You can test the configuration by submitting a job, including a sample wordcount job, and seeing the results in the InfoSphere Guardium reports.

Through your BigInsights web console, submit a wordcount job. See the BigInsights information center in the Related topics section for more information about how to do this.

Log in to the InfoSphere Guardium web console as a user and select one of the Hadoop reports, such as BigInsights - MapReduce. Figure 26 shows you an excerpt from a MapReduce report for BigInsights when the proxy is used.

**Figure 32. Partial MapReduce report for BigInsights**

(View a larger version of Figure 26.)

You can see information about permissions in the Full SQL section of the report. You can also see that the report includes information about the name of the job, the user name who submitted the job, and even the jar file name of the job. This information is parsed out for you from the full message, and because it appears as a field in the report, you can do things such as create alerts on those fields. See this section of the article for more details on customizing reports.
Appendix B: Sample GuardAPI command to configure inspection engines

The GuardAPI provides access to InfoSphere Guardium functionality from the command line to enable you to automate repetitive tasks. To run these commands you must log in with one of the CLI (command line interface) accounts and have been granted the role of admin or CLI. For more information about the API, see the InfoSphere Guardium Appendices online help book.

Listing 1 shows the commands that were used to create the inspection engines via the API in this article.

Listing 1. Sample grdapi commands to configure inspection engines in our sample environment

```bash
#hdfs job tracker, hdfs name node beeswax server
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
ktapDbPort=8021 portMax=8021 portMin=8000 stapHost=<My Hadoop Node IP>

#Mapreduce job tracker, cloudera agent and thrift plugin
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
ktapDbPort=9291 portMax=9291 portMin=9000 stapHost= <My Hadoop Node IP>

#hive server, thrift plugin
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
ktapDbPort=10090 portMax=10090 portMin=10000 stapHost= <My Hadoop Node IP>

#HDFS name node ports
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
ktapDbPort=50069 portMax=50069 portMin=50010 stapHost= <My Hadoop Node IP>

#HDFS name node ports
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
ktapDbPort=50470 portMax=50470 portMin=50071 stapHost= <My Hadoop Node IP>

#WebHDFS
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HTTP
KtapDbPort=50070 portMax=50070 portMin=50070 stapHost=<My Hadoop Node IP>

#HBase region servers
grdapi create_stap_inspection_engine client=0.0.0.0/0.0.0.0 protocol=HADOOP
KtapDbPort=60010 portMax=60010 portMin=60000 stapHost= <My Hadoop Node IP>
```

You will need to ensure that your inspection engine maps appropriately to the Hadoop node that has the corresponding services installed on that node. In this case, it was a simple one-node configuration, so the inspection engines were grouped by like port number. Your configuration will likely be more complex than this.

Appendix C. Using Guardium command line interface (CLI) to filter Hadoop noise

InfoSphere Guardium has a rich command line interface. You can use the CLI to directly configure the Collector's analyzer component to filter out Hadoop noise rather than using the security policy by using the `store gdm_analyzer_rule new` command to specify a specific Hadoop application and pattern to exclude. The example in Listing 2 shows use of the command to filter out HBase getServerRegion messages.

```bash
store gdm_analyzer_rule new
```

Big data security and auditing with IBM InfoSphere Guardium
Listing 2. CLI command to modify the collector’s filtering

```
store gdm_analyzer_rule new
Please enter rule description (optional): HDP
Please enter rule type (required): 5
Please enter rule action (optional. Default to 0):
Please enter active flag (optional. Default to 1):
Please enter DB protocol (required): 25
Please enter server IP (optional):
Please enter server IP mask (optional. Default to 255.255.255.255):
Please enter service name (optional):
Please enter pattern (optional): getServerRegion
Please enter format (optional): 1
```

The options of interest include the following.

- **Rule type**: Specify 5 for Hadoop exclusion rule.
- **Rule action**: Keep the defaults.
- **DB Protocol**: Specify ‘25 for Hadoop.
- **Pattern**: Enter the exact name and case of the message pattern you would like to exclude.
- **Format**: Enter the code for the Hadoop service to exclude. Values are:
  - 0 - HDFS
  - 1 - HBase
  - 2 - Hadoop IPC
  - 3 - Job Tracker
Related topics

- Visit the InfoSphere Guardium web site for links to white papers, demos, and more.
- Visit the InfoSphere Guardium Information Center to help you make the most of InfoSphere Guardium data activity monitoring solution.
- Read the topic entitled Running applications to learn how to submit a job.
- Read the topic entitled InfoSphere BigInsights installed components for the list of default ports used by components of InfoSphere BigInsights.
- Read the topic entitled Sending auditing information to InfoSphere Guardium for more details on configuring the integration.
- Watch videos on the InfoSphere Guardium YouTube channel, including demos of support for SAP, DB2 for z/OS, and others. Here is a short overview video and a somewhat longer more technical video on the Big Data monitoring solution described in this article.
- For a high level overview of the support for monitoring Hadoop-based in InfoSphere Guardium, read this data sheet.
- Read the ebook entitled Planning a security and auditing deployment for Hadoop for ideas to help you plan a deployment.
- Evaluate IBM products in the way that suits you best: Download a product trial, try a product online, use a product in a cloud environment, or spend a few hours in the SOA Sandbox learning how to implement Service Oriented Architecture efficiently.