IBM InfoSphere Guardium Data Activity Monitor for Hadoop-based systems

Proactively address regulatory compliance requirements and protect sensitive data in real time

Highlights

- Monitor and audit data activity for Hive, MapReduce, HBase and HDFS
- Build upon proven database monitoring technology
- Enforce separation of duties with a nonintrusive architecture
- Scale across the enterprise using a federated architecture

The proliferation of data from endpoint devices, growing user volumes, and new computing models like cloud, social business and big data have created demands for data access and analytics that can effectively handle staggering amounts of data. Hadoop-based systems (a set of open-source components) have emerged to address the challenges of analyzing data and turning it into actionable insight. Hadoop consists of a distributed file system and a MapReduce application framework, as well as additional optional components, such as Hive for data warehouse queries and HBase, a NoSQL database, for fast record lookups.

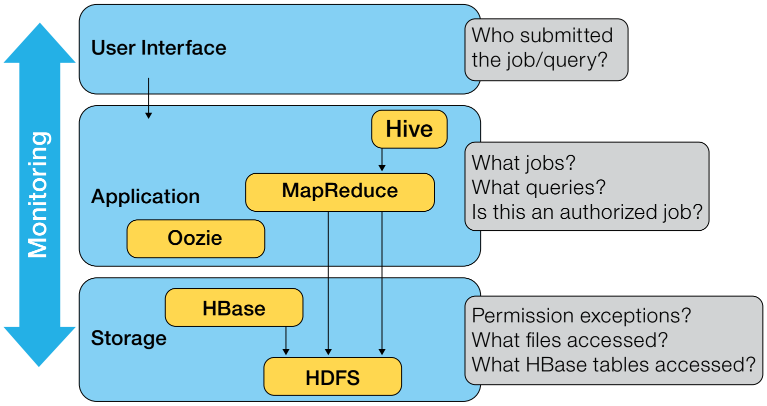


Figure 1: InfoSphere Guardium monitors activity throughout the Hadoop stack

IBM® InfoSphere® Guardium® is the leader in addressing data security and compliance concerns by delivering the first data monitoring and auditing solution for multiple Hadoop-based systems.

Addressing data security and protection challenges in Hadoop

To address some fundamental security issues, Hadoop-based systems have recently delivered some enhancements for authorization and privileges; there is nothing here that relational databases haven't had for years. Audit and compliance requirements around the world require more robust accountability in terms of being able to log and verify who did what, and when, for a database transaction. This information must be stored for a defined period of time, sometimes years.

Beyond simple compliance, however, organizations have a responsibility to do whatever they can to avoid embarrassing or damaging data breaches. Demonstrating compliance gets you part of the way there, but when breaches do occur, being able to detect and react quickly, within minutes or hours rather than days or weeks, can mean the difference between a hugely damaging loss and a minor inconvenience.

For this reason, organizations using any combination of relational databases, Hadoop, or NoSQL databases still need to implement best practices for auditing and compliance:

- Continuous real-time monitoring to ensure data access is protected and audited.
- Policy-based controls based on access patterns to rapidly detect unauthorized or suspicious activity and alert key personnel.
- Protection of sensitive data repositories against new threats or other malicious activity.
- Demonstrate compliance to pass audits: It's not enough to develop a holistic approach to data security and privacy; organizations must also demonstrate and prove compliance to internal and external auditors.

Understand the hidden costs and security risks of homegrown security solutions

How are organizations handling the auditing and compliance requirements for Hadoop? It is likely that many organizations have not yet come to terms with the problem, or may be considering custom solutions based on Hadoop log data.

This approach is problematic in several ways:

- Logs from the various Hadoop components are primarily used for debugging and for operational purposes, not for auditing. Therefore, you are unlikely to get the level of granularity required for audit purposes.
- The native Hadoop audit log will significantly increase overhead and traffic in the Hadoop cluster as more data gets written to the log.
- Any approach that relies on log data does not comply with separation-of-duties (SOD) requirements because logs stored in the system can be deleted or tampered with by any privileged user or hacker to cover their tracks.
- Real-time alerting is not supported; any compliance infraction or data breach could take weeks or months to discover using custom solution approaches.

Organizations would need to spend significant IT resources working around these issues. In most cases, organizations are using Hadoop to give them a competitive edge for business insights, a much better use of IT resources than managing and processing huge amounts of log data.

Rely on a scalable enterprise-wide database security and compliance platform

InfoSphere Guardium has extended its market-leading data activity monitoring solution to include leading-edge platforms, such as Hadoop, to help your organization meet compliance requirements while exploiting new innovations in data processing and analytics.

With a nonintrusive architecture (see Figure 2) that requires no configuration changes to the Hadoop servers, InfoSphere Guardium provides full visibility into data activity through the

Hadoop software stack, and provides full separation of duties. Operating system software taps (S-TAPs) are installed on key servers in the Hadoop cluster, such as the NameNode. These software taps rapidly stream network packets over to a separate, tamper-proof software or hardware appliance, known as a collector, for parsing, analysis and logging into its internal repository. Because processing of the network traffic occurs on the collector, overhead on the Hadoop cluster is very low.

The InfoSphere Guardium repository is the heart of the system and enables rich reporting, real-time alerting, and automated workflow management.

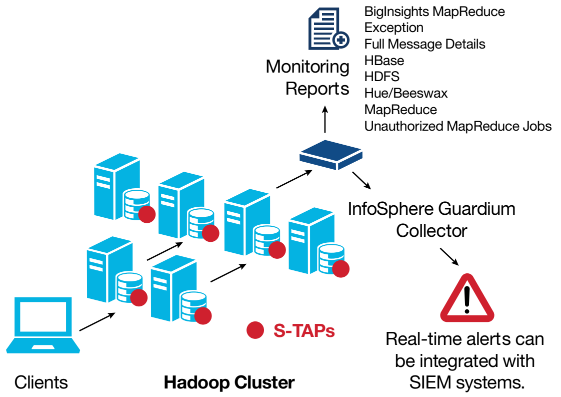


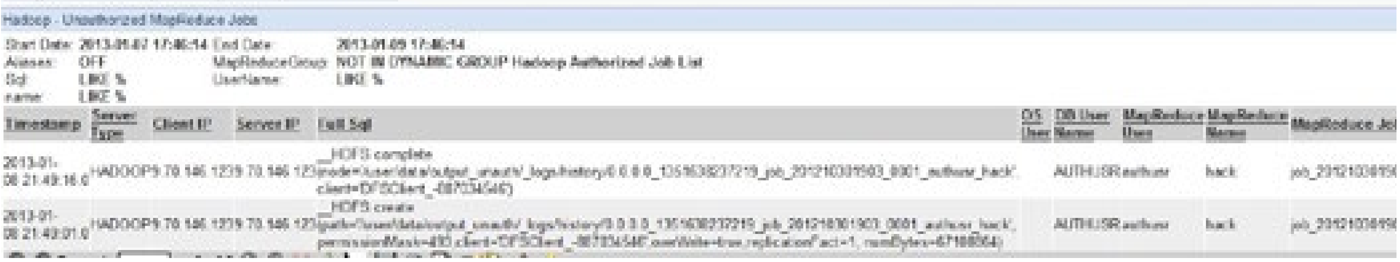
Figure 2: Scalable, nonintrusive architecture enforces separation of duties

Streamline compliance validation using automated, policy-based monitoring and auditing

The InfoSphere Guardium web console provides centralized management of alerts, report definitions, compliance workflow processes, and settings (such as archiving schedules) without the involvement of Hadoop administrators, thus providing the SOD required by auditors and streamlining compliance activities. A broad range of management functions can be executed across the entire database infrastructure, including:

- Defining granular security policies, using indicators of possible risk (appropriate for the particular environment), including the file or data object, type of access (reading, updating, deleting), user ID, MapReduce job name, and more
- Defining actions in response to policy violations, such as generating alerts and logging full incident details for investigation
- Automating compliance workflow for routine activities, such as remediation or approval of new MapReduce jobs, including steps such as sign-offs, commenting and escalation
- Ready-to-use reports (see Figure 3) tailored for Hadoop and a full and customizable reporting capability

InfoSphere Guardium provides full visibility into the Hadoop data environment, making it possible to identify unauthorized activities, like data tampering or hacking, address real time. Automation of the entire security and compliance lifecycle reduces labor costs, facilitates communication throughout the organization, and streamlines audit preparation.



Hadoop - Unauthorized MapReduce Jobs

Start Date:	2013-01-07 17:48:14	End Date:	2013-01-09 17:48:14
Aliases:	OFF	MapReduceGroup:	NOT IN DYNAMIC GROUP Hadoop Authorized Job List
Sql:	LIKE %	UserName:	LIKE %
name:	LIKE %		

Timestamp	Server Type	Client IP	Server IP	Full Sql	OS User	DB User Name	MapReduce User	MapReduce Name	MapReduce Job
2013-01-08 21:49:16.0	HADOOP	9.70.146.123	9.70.146.123	__HDFS complete (node=/user/data/output_unauth/_logs/history/0.0.0.0_1351638237219_job_201210301903_0001_authuser_hack', client=DFSClient_-087034646)		AUTHUSR	authuser	hack	job_20121030190
2013-01-08 21:49:01.0	HADOOP	9.70.146.123	9.70.146.123	__HDFS create (path=/user/data/output_unauth/_logs/history/0.0.0.0_1351638237219_job_201210301903_0001_authuser_hack', permissionMask=420 client=DFSClient_-087034646' overWrite=true replicationFact=1, numBytes=67108864)		AUTHUSR	authuser	hack	job_20121030190

Figure 3: Unauthorized MapReduce job report

Supported Hadoop distributions* and capabilities

	IBM InfoSphere BigInsights 1.4, 2.0	Cloudera CDH 3, Update 2,3,4, CDH 4.0, 4.1, 4.2	Horton-works Data Platform 1.2	Greenplum HD 1.2
Supports separation of duties, including role-based interface and storing audit data in a separate hardened appliance	√	√	√	√
Activity monitoring, including privileged users and sensitive data access	√	√	√	√
Monitor Kerberos-authenticated users	√	√	√	√
Integrate audit results with other monitored databases for enterprise-wide reporting	√	√	√	√
Real-time alerts	√	√	√	√
Hadoop security policy	√	√	√	√
Hadoop ready-to-use reports	√	√	√	√
Federated architecture	√	√	√	√
Compliance workflow and automation	√	√	√	√
Full set of administration APIs	√	√	√	√

**For an updated list of supported data platforms for monitoring, see http://www.ibm.com/support/docview.wss?&uid=swg27035836*

About IBM InfoSphere Guardium

InfoSphere Guardium is part of the IBM InfoSphere integrated platform and the IBM Security Systems Framework. The InfoSphere Integrated Platform defines, integrates, protects and manages trusted information in your systems. The InfoSphere Platform provides all the foundational building blocks of trusted information, including data integration, data warehousing, master data management, and information governance, all integrated with a core of shared metadata and models. The portfolio is modular, so you can start anywhere and mix and match InfoSphere software building blocks with components from other vendors, or choose to deploy multiple building blocks together for increased acceleration and value. The InfoSphere platform is an enterprise-class foundation for information-intensive projects, providing the performance, scalability, reliability and acceleration needed to simplify difficult challenges and deliver trusted information to your business faster.

For more information

To learn more about IBM Guardium, visit
ibm.com/guardium

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
April 2013

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Please Recycle