Create a secure Hadoop environment with IBM InfoSphere Guardium

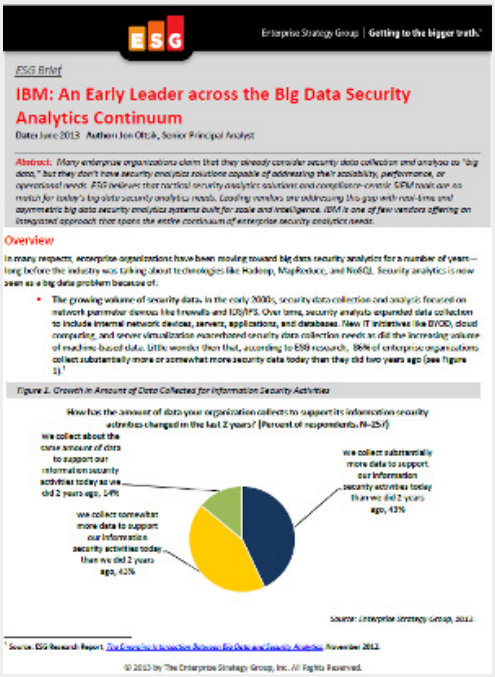
Timothy Landers (landertr@universalinet.com)
Consultant
Universalinet.com, LLC

20 January 2014

All of the benefits the Hadoop environment provides hinge on the addition of security features that are provided by an external security software solution. Just as Hadoop big data environment configurations differ, so do the security requirements for protecting that environment. All big data environments are risk-prone; therefore, they must have built-in protection against unauthorized use, threats, cyberattacks, invalid input data sources, and other challenges. To that end, IBM offers IBM® InfoSphere® Guardium®, a state-of-the-art solution for securing the Hadoop environment and protecting big data. Learn more about InfoSphere Guardium and how it can secure your Hadoop environment.

Apache Hadoop was originally designed to administer the general public's access to Google information. Hadoop's value as a pioneering technology increased when it was discovered that it might also serve as a platform for managing unstructured data across distributed nodes. Although popular among technology companies, Hadoop's programming code was never modified to support advanced security features or comply with regulatory security mandates (see). For today's uses, the Hadoop environment requires a robust security model with built-in protection against the many levels of vulnerabilities that are found when you host big data. The Core Hadoop system offers service-level authentication for tiered administration plus basic levels of protection against data breaches. Its most prevalent security feature for protecting access to big data clusters is the user password.

IBM: An Early Leader across the Big Data Security Analytics Continuum



Few products can meet the big data needs of most organizations. Learn how IBM's solution meets these requirements and more. Read this report and learn about:

- Real-time big data security analytics.
- Asymmetric big data security analytics.
- How IBM bridges the big data security analytics continuum.

Download "IBM: An Early Leader across the Big Data Security Analytics Continuum."

Big data typically uses clusters or a private cloud as storage, but a virtual environment requires data-protection features that recognize virtual machines, nodes, and networks. Big data clusters can tolerate nodes that cycle without loss of data or service interruption, but security consistency across nodes and reboots of nodes are serious performance issues in the Hadoop environment. Recognizing the industry need for providing an advanced security solution for distributed Hadoop environments, IBM has introduced an innovative security product line comprehensively referred to as **IBM Security QRadar® SIEM** (security information and event management). Security QRadar SIEM encompasses IBM QRadar, ArcSight, RSA Envision, Radius, IBM InfoSphere Guardium, Tivoli®, and HP OpenView Simple Network Management Protocol node management; Common Vulnerability and Exposures (CVE), Defense Information Systems Agency (DISA) Security Technical Implementation Guide (STIG), Center for Internet Security (CIS) Benchmark vulnerability standards compliance; the McAfee ePolicy Orchestrator security management platform; Lightweight Directory Access Protocol; Kerberos; RSA SecurID; and more.

The IBM product that was designed to protect big data environments and uses the Hadoop stack is InfoSphere Guardium, which integrates with the Hadoop framework at various Open System

Interconnection (OSI) layers. Based on the virtual or cloud environment's configuration, InfoSphere Guardium offers different sets of security features to ensure a secure Hadoop environment.

Hadoop environment security requirements

Big data security risks can be costly, resulting in data loss, reduced productivity, decreased revenue, and lower overall value for a company. Security bottlenecks innately limit big data performance, and each component of the Hadoop environment represents potential security vulnerabilities. Distributed networks are complex to the point of requiring custom backup and recovery methods. In distributed networks, gateways load big data; web and stand-alone clients intercommunicate with nodes and application managers; and big data clusters replicate, back up, and store data. All the while, regulatory compliance mandates such as the Health Insurance Portability and Accountability Act (HIPAA) and Sarbanes-Oxley Act (SOX) in the United States, still apply. To address the multifaceted characteristics of distributed computing, more security protection is required for the Hadoop environment.

Why InfoSphere Guardium?

InfoSphere Guardium protects big data and the significant investments made in Hadoop technology and affords a comprehensive assurance of adequately protecting the company's investments. InfoSphere Guardium supplements the security deficiencies of Hadoop by integrating with the Hadoop framework to provide seamless, mission-critical security features. Providing a library of security policies, InfoSphere Guardium is the next evolution of security policy that makes (as traditional security policy making s rendered ineffective by distributed computing).

InfoSphere Guardium is packaged with numerous tools for accomplishing security monitoring and solving security issues. With its set of products, it analyzes big data traffic by using a library of security policies and strategically implements security actions against network attacks and other threats to data. InfoSphere Guardium products that create a secure Hadoop environment are shown in Figure 1 and Table 1.

Figure 1. The InfoSphere Guardium family secures Hadoop

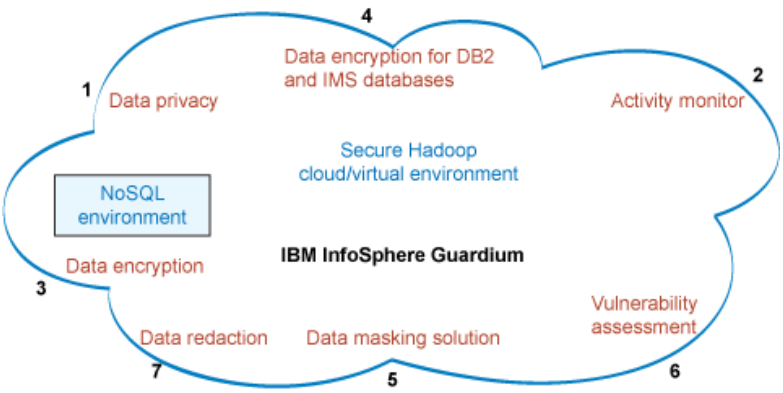


Table 1. The InfoSphere Guardium family of products

Product	Function

(1) IBM InfoSphere Data Privacy for Hadoop	Implements real-time, dynamic regulatory compliance requirements to protect sensitive data
(2) IBM InfoSphere Guardium Activity Monitor	Blocks unauthorized access to data in real time and provides alerts and notifications when security violations occur
(3) IBM InfoSphere Guardium Data Encryption and (4) InfoSphere Guardium Data Encryption for IBM DB2® and IBM IMS™ Databases	An industry-standard cryptographic utility that provides data encryption for structured and unstructured data
(5) IBM InfoSphere Optim™ Data Masking	Masks confidential data on demand
(6) IBM InfoSphere Guardium Vulnerability Assessment	Scans, detects, and recommends remedial steps to remedy database vulnerabilities
(7) IBM InfoSphere Guardium Data Redaction	Detects and removes sensitive data from displayed documents (such as PDFs, TIFF files, XML, and Microsoft® Word documents) as a data security measure

InfoSphere Guardium also provides a rich command-line interface for installations, adjustments to dynamic configurations, and to return system information.

How does InfoSphere Guardium work?

Have you ever heard the saying, "an ounce of prevention is worth a pound of cure"? It aptly applies to protecting the Hadoop environment by using the InfoSphere Guardium solution. Creating a secure Hadoop environment with InfoSphere Guardium means proactively enforcing monitoring as the number-one priority for the real-time detection and implementation of security defense measures. InfoSphere Guardium monitors systems for any unauthorized or undesired activity to provide the lead time that the system requires to mitigate, avoid, avert, or reduce the impact of a data security attack.

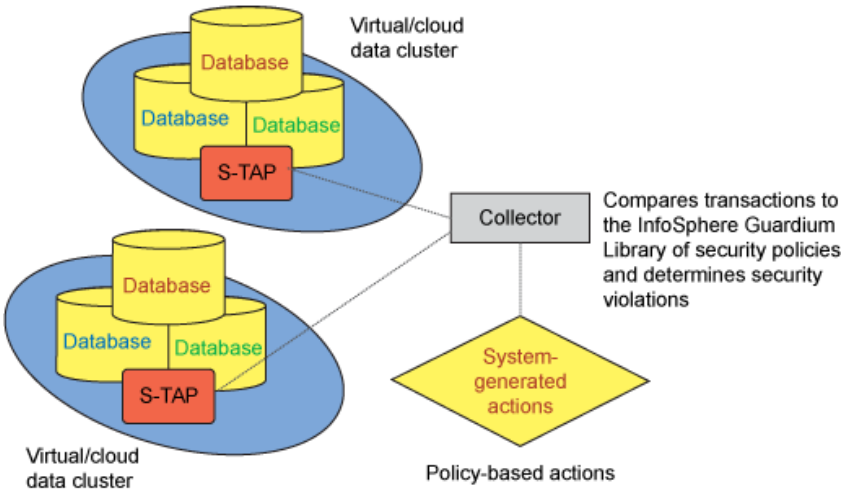
InfoSphere Guardium uses application programming interfaces (APIs) to orchestrate maintenance operations (for example, auditing transactions) or to generate reports. IBM offers Hadoop with web-based compatibility, allowing users to browse Hadoop Distributed File System (HDFS) files through a web browser. Because web-based applications can also be in the form of a stand-alone software application that is coded to support web browsers, the web functionality of Hadoop and IBM applications in serving big data now becomes twofold. From authentication to authorization when you access big data clusters, big data characteristics must be preserved along with the integrity of the environment and cluster functionality.

Monitoring prompts security policy actions

Security monitoring refers to a continuous analysis of database transactions. InfoSphere Guardium monitors the database transactions of all users by using software taps (S-TAPs) as probes, but it also integrates with other IBM security solutions and infrastructures.

As a result, InfoSphere Guardium simplifies the enormously complex task of providing consistent security across the Hadoop distributed environment. A snapshot of S-TAP performance is provided in Figure 2. An S-TAP is placed in every cluster and forwards a copy of every database transaction to the InfoSphere Guardium **Collector**. The Collector is an appliance or device for logging, storing, auditing, and analyzing database transaction audits for security violations.

Figure 2. S-TAP performance in InfoSphere Guardium



The system-generated actions reflect the InfoSphere Guardium **Policy Engine**, which provides the policies for security compliance that is used in identifying security violations. An InfoSphere Guardium **Aggregator** is the appliance that consolidates analyses from multiple collectors to generate enterprise-wide security reports. In this way, companies get the benefits of preliminary warnings about security violations. InfoSphere Guardium meticulously monitors database transactions to detect unauthorized usage, rogue entities, data breaches, and other security attacks and threats.

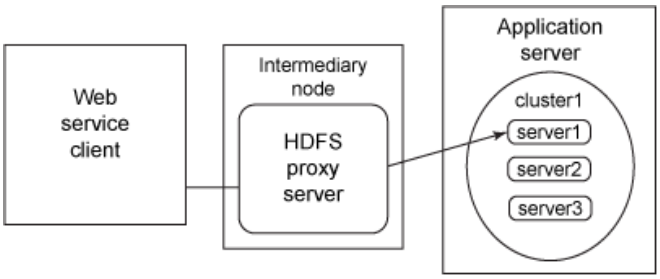
The security levels companies want and Hadoop needs

Big data network owners want a network that can securely process data quickly and speedily. This functionality relies on *parallelism*—the spreading out of processing over more computer processors to speed up the data processing process. Amdahl's Law of speed constraints proves that the number of processors that can be used effectively is limited before no faster speed can be accomplished. Distributed computing's parallelism still uses a massive number of processors and physical or virtual nodes to perform concurrent computations. In addition, big data environments of different types can coexist within the same virtual or cloud platform. For example, environments such as Hadoop and NoSQL are mismatched, which weakens the effectiveness of their built-in security tools. Therefore, creating the secure Hadoop environment requires the InfoSphere Guardium solution to scale to big data.

Big data network owners want to use the Hadoop stack's built-in security tools to provide security to the entire Hadoop environment, but big data environments can be massive. In fact, big data is synonymous with the Hadoop framework—a grouping of HDFS, Apache Hadoop NextGen MapReduce (YARN), MapReduce, and other components into a tailored, open source solution. Optional components such as Dremel, chef, Apache Hive, Puppet, and Percolator further enhance the Hadoop environment to offer features such as graphs, XML data, custom data access, and management and processing. Using various technologies with built-in security features can prove to be less secure than implementing InfoSphere Guardium, which provides a comprehensive security solution for the Hadoop environment. Moreover, aside from service-level authorization and the web proxy capabilities of YARN, no built-in security features exist in Core Hadoop to

protect Hadoop big data stores and applications across distributed nodes. Providing security to the entire Hadoop environment means using the most trusted security solutions available, but the HDFS proxy connects web browser clients to nodes by using remote procedure call over TCP/IP. Although this behavior is adequate to perform a database transaction, the transaction is less secure than using TCP/IP-to-TCP/IP connectivity. InfoSphere Guardium can supplement the required security that TCP/IP-to-TCP/IP connectivity provides because in actuality, a bidirectional security provision is in place for this proxy configuration that defends data that is transmitted from the HDFS proxy and defends data that is transmitted to the node. See Figure 3.

Figure 3. A secure Hadoop architecture



Big data network owners want secure big data accessibility by various user types, as well. From a security perspective, database transactions are monitored for authorized security settings down to the user role level to provide this level of security. User roles and passwords are the primary forms of big data security, but database relationships can also be distributed. In a complex mesh of schemas, the database model offers only limited access and denies protection against user and system access. In contrast, InfoSphere Guardium offers groundbreaking protection at a granular level that allows administrators to protect big data at the distributed node, field, and even user role levels.

Security violations can be costly

To better understand the value that InfoSphere Guardium provides, you must look at the potential harm small or indirect security violations can cause.

Big data security risks include potential random security attacks; data breaches; ineffective security policies; and rogue nodes, users, or applications that gain access to the cluster. A Hadoop environment with the appropriate security provisions can result in compromised data, which can in turn result in the sending of malicious data or links to either service. Also, some nodes are self-organizing, which means they require a "choke-point" not available in a peer-to-peer "mesh" cluster. As a result, they cannot benefit from gateways, firewalls, or monitoring security tools. Big data stacks build in almost no security because they are premised on the web services model and the Open Web Application Security Project (OWASP) Top Ten list.

Another area of security is the access to administrative data. At least one administrator administers nodes. Full access to a node requires constraints to ensure the facilitation of separating duties among different administrators. Similarly, relational database platforms require security constraints to facilitate security. Big data platforms lack their array of built-in facilities, documentation, and third-party tools to address this requirement.

Detecting whether a big data cluster was breached is a hidden necessity that requires a proactive and efficient approach. The constant monitoring of transaction logs is a practical solution. Add logging to the existing cluster, use the shared web features for managing log files, or add on an SIEM or other log-management product. Logging adds security for detecting attacks, diagnosing failures, or investigating unusual behavior by tracing events to their root cause. MapReduce requests are an event that can be logged, for example. Hadoop offers partial solutions for authorization.

InfoSphere Guardium cryptographic security

An extra deciding point in favor of InfoSphere Guardium for creating a secure Hadoop environment is that it has ample provisions for the sufficient horizontal scalability and transparency that is needed to work with big data. The Data Encryption feature provides cryptographic encryption and decryption without interruption to the Hadoop environment. The InfoSphere Guardium centralized policy and key management services are a bonus to protect unstructured and structured data by requiring that users who attempt to access encrypted files have the required encryption key or certificate to do so. Policy specifications against and appliance interference are a must have for semantic webs and guard against security and privacy violations that stem from inference.

An industry best practice for protecting data at rest is encryption, which guards against attempts to access data outside established application interfaces. Encryption serves to protect the big data that is replicated, transferred, and transmitted to and from the cluster. Also, replication offers rogue administrator users an opportunity to steal or otherwise harm big data. In addition, only few NoSQL variances provide encryption for data at rest.

InfoSphere Guardium Data Encryption feature also protects data from malicious users or administrators that gain access to data nodes and directly inspecting files; it renders stolen files or copied disk images unreadable. Encrypted files block attacks that might otherwise circumvent application security controls. InfoSphere Guardium file-layer encryption provides consistent protection across different platforms. It is not apparent to both Hadoop and calling applications, and it scales to big data as the cluster grows. In fact, most of the security that the Hadoop environment requires can be accomplished through encryption controls. But encryption keys must be protected at the OSI Layer 2 to provide effective cryptographic security of big data clusters. In addition, storing encryption keys on local disk drives is convenient, but distribute the encryption keys and certificates to provide security to every user, group, and application. Such operations might require APIs, which must be secure enough to execute the programming code commands that maneuver big data without compromising big data. InfoSphere Guardium provides each security feature that the Hadoop environment requires and more.

Conclusion

You can achieve a secure Hadoop environment by enabling the right set of IBM security tools. Often, big data is subject to legal regulatory compliance, as mandated by law. Create a secure Hadoop environment by using IBM features both data and infrastructure protection to ward off would-be attackers and defend the weaknesses and vulnerabilities of big data applications.

When it comes time to secure your big data environment, remember these key points:

- The best Hadoop security solution scales with the environment's big data.
- Use third-party security tools because they are built into the Hadoop framework.
- Policy specifications address more granular levels of security requirements.
- Security controls are architecturally and environmentally consistent with the cluster architecture.

Hadoop environments face the same big data security and privacy challenges. To protect big data is essentially to create a secure Hadoop environment, and IBM provides a strong set of security mechanisms as an industry best practice secure Hadoop environment solution.

Resources

Learn

- Hadoop, the open source system for scalable distributed computing, is a top-level Apache project. At the Hadoop site, you can learn about Hadoop and the collection of other projects that extend and enhance big data by processing with Hadoop.
- Be sure to check out Hadoop Poses a Big Data Security Risk: 10 Reasons Why for more information about security in Hadoop environments.
- Learn more about InfoSphere Guardium Data Redaction, and learn how internal security features complement the external security of the Hadoop environment.
- Visit the Security On developerWorks blog to learn about new security-related how-to guides, articles, and demo videos.
- Sign up for the weekly Security On developerWorks newsletter for the latest security headlines.
- Follow @dwsecurity to get updates from the developerWorks security zone in real time.
- Read Introduction to Parallel Computing by Ted G. Lewis and Hesham El-Rewini (Prentice-Hall, 1992—especially pages 31-32 and 38-39).
- Sensitive data comes in many forms, but a growing list of regulations exists to ensure that data is kept private. These regulations include HIPAA for medical data and SOX for financial data.
- Kerberos is a network authentication protocol that is designed to provide strong authentication for client/server applications that use secret-key cryptography.
- Learn more about vulnerability standards compliance with DISA STIGs, and CIS Benchmarks.
- Learn more about the OWASP Top Ten Project.
- Find out how industry experts see the Hadoop security model and what it requires in Big Data Security: The Evolution of Hadoop's Security Model.
- Get more information on security topics in the Security site on developerWorks.
- Follow developerWorks on Twitter.
- Watch developerWorks on-demand demos that range from product installation and setup demos for beginners to advanced functionality for experienced developers.

Get products and technologies

- InfoSphere Guardium encompasses many technologies, including:
  - InfoSphere Data Privacy for Hadoop
  - InfoSphere Guardium Activity Monitor
  - InfoSphere Guardium Data Encryption
  - InfoSphere Guardium Data Encryption for IBM DB2 and IBM IMS Databases
  - InfoSphere Optim Data Masking Solution
- Learn about IBM Security QRadar SIEM as a state-of-the-art technology and industry-standard security management solution.

Discuss

- Browse the InfoSphere Guardium 9.1 TechTalk forum to see updates and revisions to the InfoSphere Guardium feature set.
- Join the developerWorks community, a professional network and unified set of community tools for connecting, sharing, and collaborating.

About the author

Timothy Landers

Timothy Landers, a principal at Universalinet.com, LLC, is a practice lead in an independent consultancy. He has an MBA in Technology Management and is a Project Management Institute-certified Project Management Professional with more than 15 years in increasingly more responsible roles within the IT field. He has written more than 28 technical courses for corporate training, vocational training, and higher education plus new product manuals, professional certification exams, and commercial sales catalogs (such as SkillSoft).