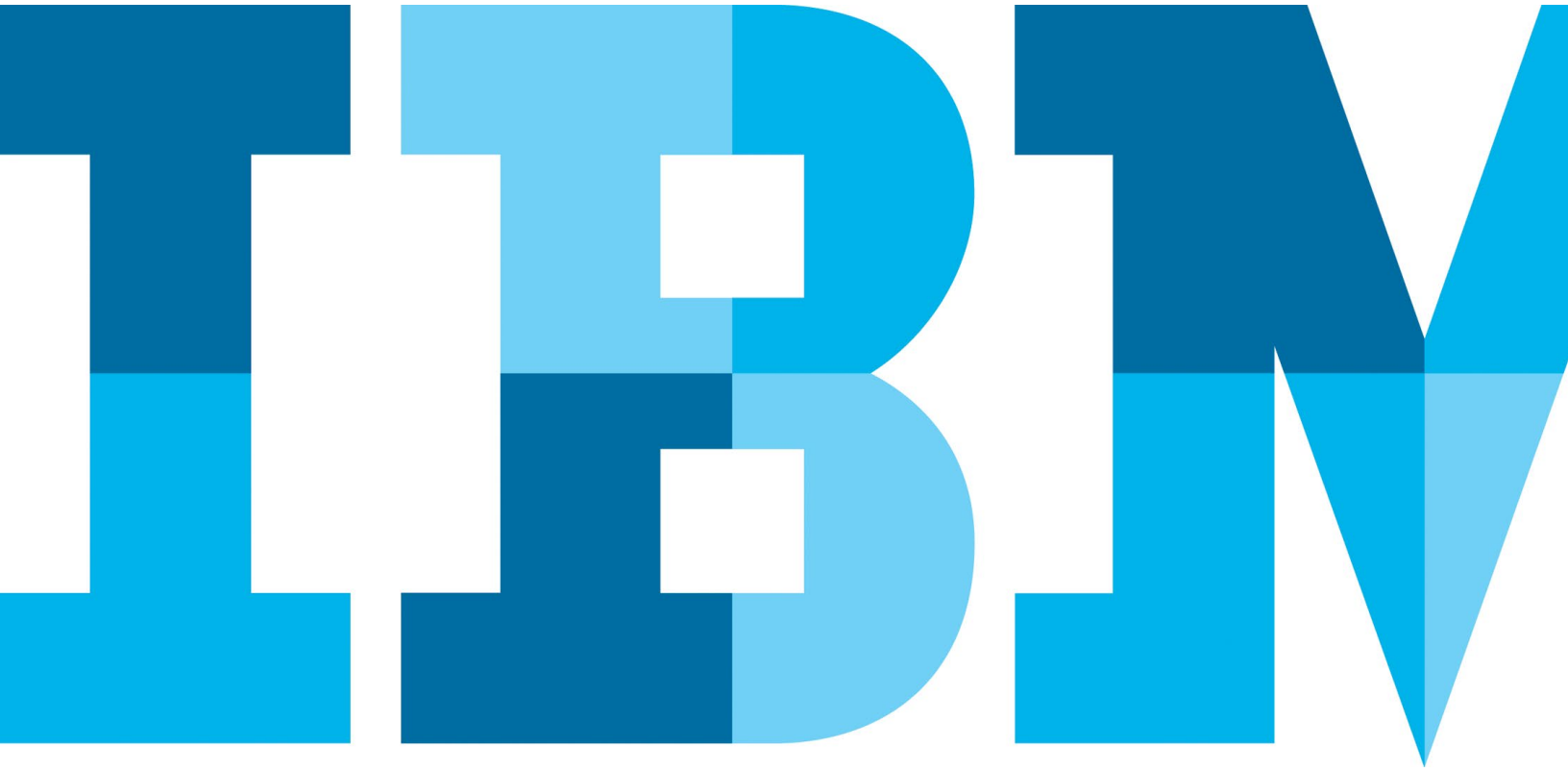
Security
Thought Leadership White Paper

November 2016

Finding the path to security in the big-data landscape

Six best practices to safeguard sensitive data in Hadoop and NoSQL environments



IBM

Introduction

The promise of big data holds great allure. By using analytics to extract hidden insights from the volume, velocity and variety of information that characterizes big data, organizations can create value and build competitive advantage. Using big-data analytics, they can find new opportunities, drive product innovation and grow revenue. They can achieve faster, more accurate decision-making, more agile operations, and more personalized relationships with customers.

All of these are very real—and very achievable—benefits organizations can derive from harvesting the results of running big-data analytics on accrued information from customer transactions, machine sensors, mobile communications platforms, social media, traditional sources and other sources. But in their haste to spin this raw data into business gold, many organizations take shortcuts. Blinded by big data's shiny allure, they frequently skip an essential step: In the rush to use big data to drive business forward, organizations all too often neglect data security.

And as more and more organizations use big data to enhance their business, some observers—including IBM[1]—are beginning to predict a wave of data breaches targeted specifically at big-data resources.

This white paper examines how some of the ways organizations use big data make their infrastructures vulnerable to attack. It presents recommended best practices organizations can adopt to help make their infrastructures and operations more secure. And it discusses how adding advanced security software solutions from IBM to their big-data environment can fill gaps that big-data platforms by themselves do not address. It describes how IBM® Security Guardium®, an end-to-end solution for regulatory compliance and comprehensive data security, supports entitlement reporting; user-access and activity monitoring; advanced risk analytics and real-time threat detection analytics; alerting, blocking, encryption and other data protection capabilities, as well as automated compliance workflows and reporting capabilities, to stop threats.

The rise of big-data platforms

Many traditional database management tools and data processing applications cannot keep up with the huge quantity and rich complexity of data that's being created today. Organizations wishing to extract greater value from their data resources have recognized this, and are turning to big-data technologies such as the highly scalable Hadoop data and application framework and NoSQL databases.

This switch enables organizations to efficiently and effectively run data analysis on large sets of related data, and avoid the limitations that come with separate, smaller sets, including those containing the same total data. From a business standpoint, the change is a boon to supporting data-driven decision making; developing a 360-degree view of customer behavior and desires; and gaining business insights from historic reporting, real-time analysis or predictive modeling—even from analyzing and extracting insight from data that was previously considered unusable.

In practice, however, the arrival of big data, along with its relatively new and less-mature technology environments, can pose a security risk.

Traditional relational databases, which for many years have been the data software of choice, utilize structures (or "schemas") for organizing information in a way that helps protect data integrity. The longevity of the relational database platform also has given developers time to consider and take steps to provide defenses such as privileged identity management, authentication and access controls, encryption and masking. Newer NoSQL databases, on the other hand, are "schema-less," containing not only structured data but also diverse, complex and less predictable semi-structured, unstructured and polymorphic data. Perhaps most important when it comes to security, however, many NoSQL databases are designed to appeal to users interested in rapid development or to businesses pursuing low costs. In some cases it was assumed that these databases would be used only in safe, trusted environments. Security was not always a design concern.

Why one global bank moved to a NoSQL database

Challenge	Why NoSQL	Results
• Experienced delays of up to 36 hours in distributing data by batch • Was being charged multiple times globally for storing the same data • Incurred regulatory penalties from missing service level agreements • Had to manage 20 distributed systems containing the same data	• Dynamic schema: Easy to load initially and over time • Auto-replication: Data distributed in real time and read locally • Both cache and database: With a cache that is always up-to-date • Simple data modeling and analysis: Easy to change and understand	• Expects to save about 40USD million in costs and penalties over five years • Is charged only once for data • Supports data that is in sync globally and read locally • Has the capacity to move to one global shared data service

The security challenge of big data

The challenge arises when IT teams or even line-of-business managers seeking quick results or needing to make rapid business decisions go straight to big-data resources and initiate data analytics or other projects on their own. A project running on a relational database requires the intervention of a database administrator and often the security team—as well as processes for changing schemas and controls. All of this can be slow and time consuming. But in a big-data environment, it is often possible to bypass the security team entirely, leaving them unaware that someone has been running a project that may involve sensitive business, personnel or customer data on a less secure platform. And frequently IT teams don't stop to think that big data is also subject to regulatory compliance requirements (such as those for Payment Card Industry [PCI], Sarbanes-Oxley [SOX], etc.).

In today's hyper-competitive business environments, where time is usually of the essence, such a shortcut may seem understandable. A chief marketing officer facing a deadline for the rollout of a new product or the launch of a service in a new geography may not have the time or patience to wait for slower, traditional processes. IT or business users may be caught up in the excitement of a new technology, or the expediency, rapid results or lower costs associated with big-data platforms. They may simply dismiss security concerns, treating their endeavor as a pilot project conducted entirely in-house and safely behind the firewall, with no lasting consequences on the data and IT environment.

But a project run in a "shadow IT" environment where rogue IT teams access data without regard to whether it should be kept private or secured can, in fact, have significant consequences in the potential exposure to risk or inability to comply with regulatory mandates. Additionally, big-data environments are frequently very open: Hadoop developers, data analysts, IT administrators and regular users have the same unfettered access to sensitive data that privileged users have. The data that's loaded into the big-data platform, and the insights that can be gleaned from it using analytics, is frequently sensitive (customer data, patient data, manufacturing data, business partner data, etc.)—and is a lucrative target for cybercriminals. Nonetheless, in many cases, there are no significant security or audit controls over sensitive data or user access.

The responsibility to protect big data

Security certainly should be of the utmost importance in the big-data world, and risk and governance issues—such as security, privacy and data quality—rank high among the challenges for big-data system designers and administrators. But many organizations have set security concerns to the side as they work to realize value from their big-data investments.

Meanwhile, the cost to business of a data breach continues to rise. A report by the Ponemon Institute found recently that the average cost of data theft had risen to 154USD for each lost or stolen record—with the average total cost of a single data breach rising to 3.8USD million. This represents a 23 percent increase just since 2013.[2] As big-data projects become more and more common, and as sensitive data proliferates and moves throughout those projects, the cost to the business will likely continue to grow: A breach of sensitive data in a big-data environment could be quite expensive.

Clearly, the time is now to make sure security teams and security best practices are included in the management and governance of all data—including big data—and in the deployment and administration of solutions to keep sensitive data secure.

The idea of security as an issue for projects such as in-house data analytics may be a new one for the non-IT personnel who are under pressure to use this data to speed business decisions. But making the security team a part of the process before the organization fails an audit, before a breach occurs or before "shadow IT" practices become the entrenched norm is critical.

That's because aside from the scope of big data—its volume, velocity and variety—the risk and threats, the IT processes, and the business urgency for protecting big data are fundamentally the same as on any other platform. And regardless of where in the Hadoop or NoSQL and big-data maturity lifecycles the organization stands—whether as a recent adopter or a company with an established big-data environment—the responsibility to protect data and comply with mandates is also the same.

The urgency of protecting big data

The volume, variety and velocity of big data are key elements in enabling analytics to drive business processes and help increase revenue. But with so much information arriving so fast, it can be difficult to ensure the integrity of big data, to control who has access to it, and to scale management controls to support regulatory compliance.

No matter what environment you are working in, it's entirely possible, in fact, for technical challenges—or inattention and lack of security control—to leave sensitive big data or the sensitive results of big-data analytics—dangerously exposed. A recent scan by one security researcher, for example, found some 30,000 instances—nearly 600 terabytes—of data managed by a popular database application accessible without authorization over the Internet due to configuration errors.[3]

Such exposures underline the critical need for security teams to get involved in analytics and other big-data projects from the work's inception—to always know what projects are planned and underway in their big-data environments, to control what "shadow IT" teams are doing, and to manage security for the use of big data and the results from big-data projects whenever possible.

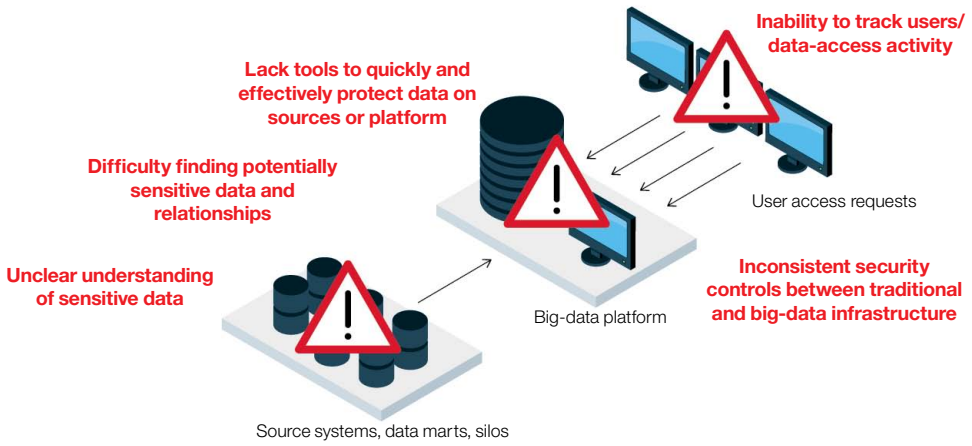
The scope of protecting big data

Being involved with big-data security from the beginning means understanding and controlling the entire big-data environment. You need to know where big data comes from, who is accessing it, the results of analytics or other projects, and where that information goes—all are key elements not only to using big data to move the business forward and increase revenue, but to protecting the organization from security breaches, data corruption and cyber theft.

In a nutshell, huge volumes of big data arrive from traditional environments, transactions, sensors, mobile devices or social media, and are stored in big-data repositories as structured, unstructured, semi-structured and polymorphic data. The sensitive pieces require protection when they are in motion and at rest, so data can be reliably accessed and utilized by authorized users.

A critical function for security is to manage all this data across the environment and at each stage of the big-data lifecycle, from the moment data arrives until it is analyzed to produce actionable business insights. Along the way, the data can pass through a maze of processes, as, for example, business users and data analysts search the data and investigate its possible uses, data scientists analyze the data to find business value, and application developers create ways to use that data to advance business operations.

Big-data technology security and privacy issues



The security challenge lies in managing and protecting huge masses of valuable, sensitive data that can be as irresistible to employees who want to use it for business projects as it is to cybercriminals who want to steal and sell it for gain. In some cases, sensitive big data also may face threats from malicious internal users who might also steal or leak the data, or damage or delete it. Big-data environments provide a big attack surface, giving employees and criminals many opportunities to gain access and wreak havoc.

What's more, big-data environments such as those built on Hadoop and NoSQL technologies don't have mature, built-in capabilities for protection. At varying levels, these big-data platforms are making progress on native security controls, such as authentication, authorization, auditing, and encryption. But these efforts are not enough to make big data an integral part of the enterprise data protection strategy. Information still has to be classified in a way that lets all parts of the organization reference it consistently, and the security team needs to be able to monitor data activity and access, mask sensitive data, and receive real-time alerts when appropriate. Ultimately, the security team needs broad capabilities for avoiding disclosure or leakage of sensitive data, preventing unauthorized changes to data and data structures, and reducing the cost and complexity of compliance.

Key capabilities missing from big-data platforms

Missing	Why it's important
Auditing with minimal impact	Helps improve performance and reduce costs
Real-time alerts	Enables the security team to take action before it's too late, helping stop data loss in real time
Segregation of duties	Avoids conflict of interest and the risk that can occur when one person has too many privileges and too much access
Advanced role behavior and data usage analytics	Reduces risk by enabling the security team to monitor everyone—including those doing the monitoring
Data security scalability, performance and centralization across data platforms	Helps increase visibility and control, close loopholes, reduce costs and increase efficiencies by eliminating data security silos
Assessment of data repositories for vulnerabilities and exposures	Addresses weaknesses; hardens and secures repositories

The good news for the security team, however, is that despite the challenges of scale, difficulty in knowing who is using sensitive big data, and multiple ways for users to access this valuable information, the core security concerns for protecting this data remain much the same as in a more traditional data environment—ensuring compliance, identifying and mitigating data risk, guarding against data breaches (including insider threats), protecting against segregation-of-duties violations, protecting the brand reputation, managing users and their access entitlements, and deploying data protection tools that have the right capabilities and level of scalability.

Best practices for securing big-data environments

In providing security for big-data environments, the place to begin is by establishing a series of effective best practices. These can help the security team better understand issues such as who is accessing big data, whether the users are authorized to access or change or delete data, whether privileged users are showing suspicious behavior, and whether the users are copying or removing sensitive data.

Best practices can help identify whether existing security measures are sufficient—or if they need to do more. They can create a foundation for three key steps to ensuring big-data security: planning ahead, thinking strategically, and securing and protecting the data. Security strategies and best practices for big-data environments include:

Discover and classify sensitive data before it moves to the big-data environment

Most firms don't comprehensively classify all their data, but that may be short-sighted: Classification prior to moving data into Hadoop or a NoSQL database enables the right security policies to be established downstream.

- Determine which data is sensitive and where it is stored—and automate this process so it can be run on-demand or on a regular basis
- Prioritize "crown jewels" of enterprise information above all else
- Work with application owners and compliance officers to help identify and classify data
- Work with business managers to determine the level of data sensitivity and the impact a data breach would have on the business
- Share the data glossary, privacy policies and project blueprints with stakeholders

Put access controls into place

Access policies are critical to ensuring that users have the right access to the right data without creating undue risk and exposure. Be sure unauthorized users, IP addresses and programs do not have access to sensitive data, inappropriate communications between applications, or the ability to bypass security controls.

- Turn on programs for user authentication
- Determine appropriate roles for each user
- Establish a range of hours and days of the week for authorized access
- Put into place allowable file system operations for access to the data, including read, write, delete, rename, application or process

Establish real-time data activity monitoring and auditing

To ensure data security and support regulatory compliance, it is critical to understand the who, what, where, when, and how of big-data access. Rigorously monitor and carefully manage access to sensitive big-data resources.

- Evaluate current access privileges and entitlements and monitor new entitlements on an ongoing basis
- Create a secure, detailed, verifiable audit trail of data access activities
- Intensively monitor activity involving sensitive data; be especially mindful of privileged user activity
- Use real-time alerting to immediately investigate suspicious activity
- Use analytics to uncover suspicious patterns of behavior or fraud
- Establish an audit trail for data access and usage to help prevent the loss of sensitive data
- Centralize and automate the reporting of audit data
- Integrate data-access monitoring insight from big data into the access analysis for the rest of the data environment.

Protect data with masking, encryption or redaction

To keep sensitive information secure, it must be protected from misuse, fraud and loss—but to maintain its usefulness for building business value, protection measures also must preserve the ability to perform big-data analytics.

- Mask data to protect actual information while having a functional substitute for occasions when the real data is not required; de-identify sensitive data at its source or within the big data platform
- Encrypt data at rest and in motion to create versions that are unreadable by cybercriminals or unauthorized internal users who might gain access to them through the infrastructure
- Redact information to protect unstructured data in textual, graphical and form-based documents
- Employ a combination of measures to achieve better information governance and regulatory compliance

Assess and address vulnerabilities

To reduce the chance of a data breach or cyber attack, be sure to understand any weaknesses in the sensitive data stores. Once the vulnerabilities have been found, put plans in place to mitigate those weaknesses.

- Check for security risks such as missed or incorrectly installed patching
- Develop a review process to identify unauthorized connections
- Update server images frequently to ensure that systems are not running vulnerable, outdated instances of data sources and other software
- Ensure that configurations are correct, including for firewalls and ports
- Correlate real-time log data with other activity to provide context and prevent attacks

Enable compliance management:

Compliance mandates apply to big-data environments just as they do to other platforms. The rush for big-data benefits does not excuse organizations from the need to support compliance rules and mandates.

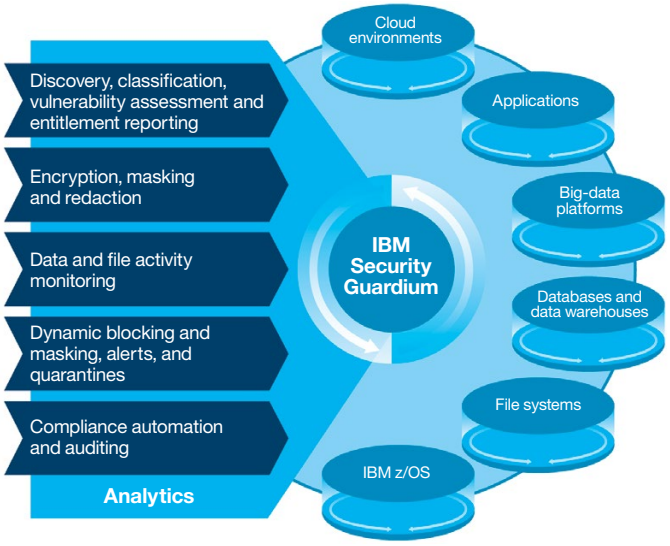
- Build a compliance reporting framework into Hadoop or NoSQL to manage report generation, distribution and sign-off
- Integrate security activities with business processes to support audit compliance; send reports to the right people at the right time for review and sign-off
- Help reduce the cost of compliance by automating and centralizing controls and by simplifying the audit review processes

IBM Security Guardium: Data protection for environments of all types

The need to monitor and control big-data environments grows from a lack of visibility into the data within the environment, who is accessing it and how it's being used. Is there a new request in the system? Who is accessing data? Are developers or end users accessing and using sensitive data? Are certain data access patterns abnormal, and do they indicate an attack? Are the security team's management processes affecting the performance of the platform? And ultimately: Are these processes keeping the sensitive big data safe?

Organizations that rely on Hadoop and NoSQL platforms alone may not be able to answer all of these questions. That's because the big-data platforms have limited auditing and alerting capabilities. Their ability to control data access, match it to users' roles, and establish and maintain segregation of duties is still maturing. And they also lack capabilities for data protection—such as data masking and encryption. In short, they provide a big-data analytics environment, not a data security and compliance solution. Even when their native security controls mature, for true sensitive big-data protection, they should be supplemented with a complete and scalable data-protection solution that enhances visibility and protects data across the board.

Guardium is a comprehensive data protection solution that can support the most complex of environments—whether clients are leveraging Hadoop or NoSQL platforms, file systems, mainframe environments, or all of the above. Guardium provides automated and centralized controls that help secure the entire sensitive data environment with zero to near-zero impact on the performance of data stores, and it is highly flexible and scalable.



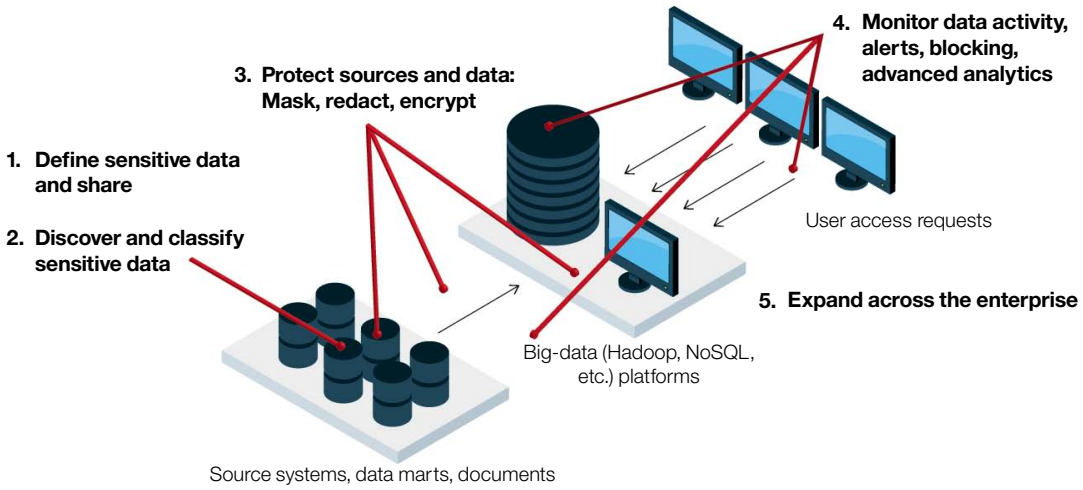
Designed to support Hadoop and NoSQL big-data environments as well as more traditional technologies, Guardium uses automation and cognitive intelligence to analyze sensitive data risk and safeguard sensitive data. For databases, data warehouses and file systems, Guardium supports automated discovery and classification of sensitive data, protects sensitive data with the right capabilities, and provides automated compliance workflows and pre-built compliance reports. Guardium is able to adapt to changes in the IT environment—whether those changes include adding new users, adding new technologies, or adjusting to the changing volumes and types of data flowing throughout the enterprise.

Guardium is able to support regulatory compliance needs, monitoring and sensitive data protection for Hadoop and NoSQL systems, supporting the full journey from compliance to data security, while using automation to help keep costs low.

IBM Security Guardium and big-data security

The IBM approach to big-data security begins with discovering and classifying sensitive data in source systems to provide visibility into which sensitive information is on the system before loading it into big-data environments. Guardium also supports redaction, encryption and masking to suppress, scramble or hide sensitive data—both in the original data source and for sensitive data in the big-data environment. Using data-activity monitoring and vulnerability assessment, Guardium can provide visibility into data-access patterns in Hadoop and NoSQL environments—including the who, what, when and where of user actions—as well as to exceptions in operations such as authorization and access control failures. It provides leading-edge analytic and cognitive capabilities to help security analysts uncover threat and forensic insights. To further protect data, Guardium can support and enforce segregation of duties, block or quarantine suspicious user IDs, and send real-time alerts to the security team to notify them of unusual activities.

The IBM approach to securing big-data environments



Guardium also integrates with the broader security ecosystem. It integrates with security products from many vendors, but of particular note are integrations with IBM Security Privileged Identity Manager, IBM Security Identity Governance and Intelligence, and IBM QRadar®. Guardium provides 360-degree integration with the data intelligence gathering and analytics capabilities of IBM QRadar Security Intelligence Platform, powered by IBM Sense Analytics™, which collects and understands security data from across the organization, transforms large amounts of raw security data into meaningful insights, and empowers the security team to speed investigations with integrated intelligence to limit the impact of a breach.

Conclusion

As the volume, variety and velocity of data and use of big-data analytics continues to increase, enterprises need to vigilantly safeguard the sensitive data in their Hadoop and NoSQL platforms—and across the rest of their environment—in real time. They need the ability to track data access and perform advanced analytics to detect unusual user behavior; protect sensitive data with capabilities including data redaction, encryption and masking, blocking, and quarantining; receive alerts when threats and unusual behavior are detected; and take steps to harden the environment.

Deployed with best practices for controlling and monitoring data activity, Guardium enables centralized, automated data protection across heterogeneous enterprise environments—including databases, applications, mainframe environments, cloud environments, and, of course, big-data environments. With the ability to scale to protect both traditional data sources and big-data platforms, Guardium is a comprehensive solution for securing the big-data environment—and for automating and streamlining compliance initiatives.

For more information

To learn more about IBM Security Guardium, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/guardium

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

IBM Security
Route 100
Somers, NY 10589

Produced in the United States of America
November 2016

Please Recycle

[1] Cindy Compert, "Football and a Crystal Ball: Data Privacy Predictions for 2016," *IBM Security Intelligence*, January 28, 2016. https://securityintelligence.com/football-and-a-crystal-ball-data-privacy-predictions-for-2016/

[2] "2015 Cost of a Data Breach Study: Global Analysis," *Ponemon Institute*, May 2015. http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&htmlfid=SEW03053WWEN&attachment=SEW03053WWEN.PDF

[3] Jaikumar Vijayan, "About 30,000 Instances of MongoDB Exposed on Web, Security Researcher Says," *IBM Security Intelligence*, July 23, 2015. https://securityintelligence.com/news/about-30000-instances-of-mongodb-exposed-on-web-security-researcher-says/

WGW03258-USEN-00