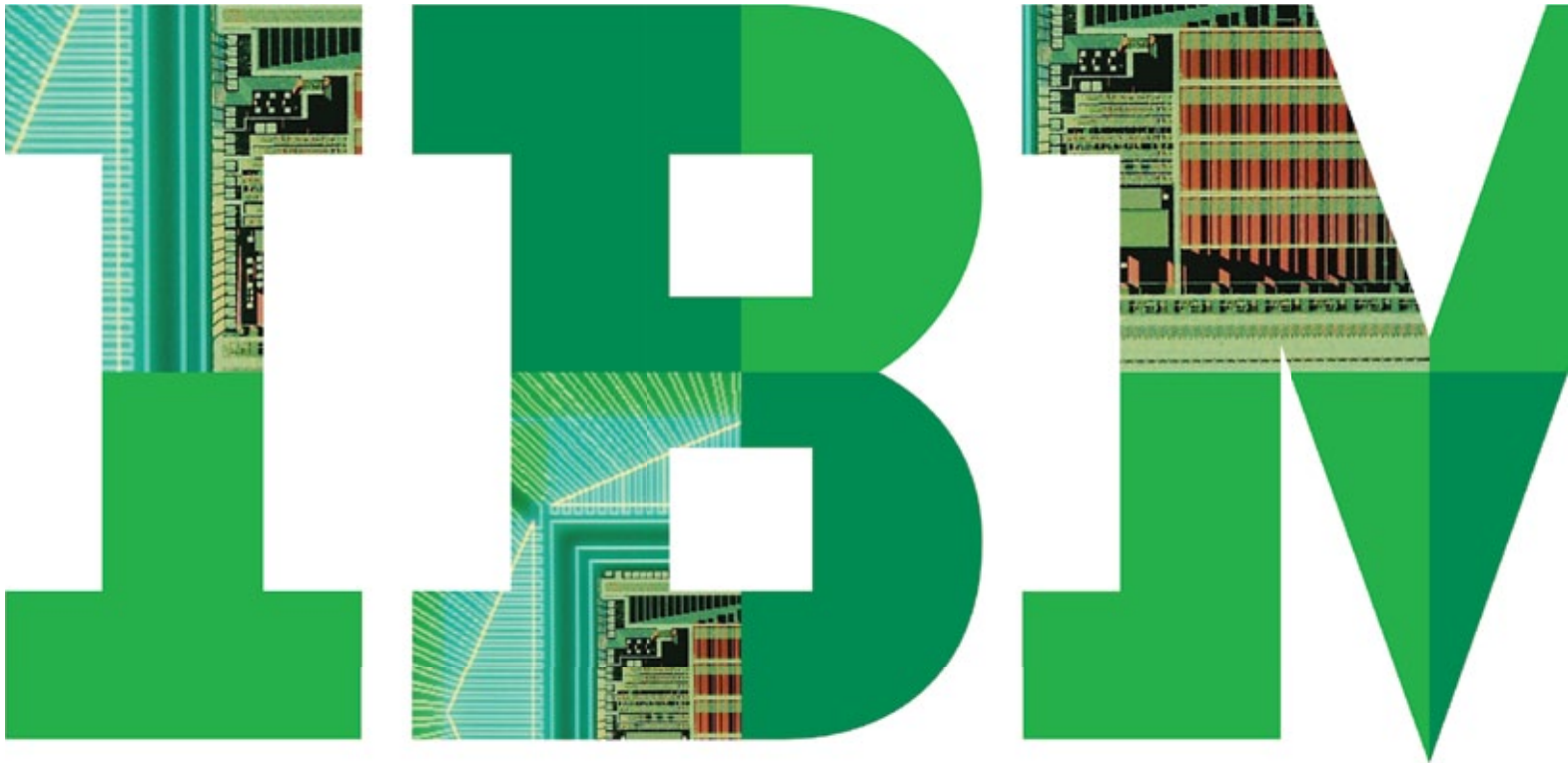
Top tips for securing big data environments

Why big data doesn't have to mean big security challenges



IBM

IBM

1 **What is big data?**

Handle and process data on an extreme scale to derive maximum value.

2 **Unique challenges**

Big data environments create significant opportunities, as well as security challenges.

3 **Raising security awareness**

Build security into big data environments to reduce costs, risks, and deployment pain.

4 **Security fundamentals**

3 steps to control and secure the extreme volumes of data.

5 **IBM InfoSphere Guardium solutions**

Improve security decision-making based on prioritized, actionable insight derived from monitoring big data environments.

6 **Conclusion**

The same security fundamentals for securing databases, data warehouses and file share systems can be applied to securing big data implementations.

What is big data?

Information technology drives innovation and has revolutionized the way businesses, governments and individuals work and interact. The ability to harness big data has opened the door for world-wide collaboration in real time and is no doubt a game changer. Big data has generated an enormous amount of discussion and debate in the press, on blog sites, amongst analysts and top technology firms. However, with all the noise it's hard to discern the capabilities, practical uses, and challenges of big data technologies.

Big data spans three dimensions: volume, velocity and variety.

Volume	Velocity	Variety
Every day 2.5 quintillion bytes of data are generated from new and traditional sources including climate sensors, social media sites, digital pictures and videos, purchase transaction records, cellphone GPS signals, and more.	Sometimes 2 minutes is too late. For time-sensitive processes, such as detecting fraud, a real time response is required.	Big data is any type of data—structured and unstructured—such as text, sensor data, audio, video, clickstreams, log files and more.

Big data environments help organizations process, analyze and derive maximum value from these new data formats, as well as traditional structured formats, in real time or for future use to make more informed decisions cost effectively. Forrester Research defines big data as "a set of skills, techniques, and technologies for handling data on an extreme scale with agility and affordability."

Some examples of big data projects include:

- Turning 12 terabytes of Tweets into improved product sentiment analysis
- Scrutinizing 5 million trade events created each day to identify potential fraud
- Monitoring 100's of live video feeds from surveillance cameras to identify security threats

As big data environments ingest more data, organizations will face significant risks and threats to the repositories containing this data. Failure to balance data security and quality reduces confidence in decision making. In fact, research shows business leaders who feel uncertain about analytical outputs will find reasons to reject them unless they develop high levels of trust in the data and know the data is secure.

A paradox exists. Organizations are generating more data now as compared to any other point in history, and yet they don't understand its relevance, context or how to protect it.

Unique challenges of securing big data

Big data environments create significant opportunities. However, organizations must come to terms with the security challenges they introduce, for example:

- Schema-less distributed environments, where data from multiple sources can be joined and aggregated in arbitrary ways, make it challenging to establish access controls
- The nature of big data—high volume, variety and velocity—makes it difficult to ensure data integrity
- Aggregation of data from across the enterprise means sensitive data is in a repository
- Big data repositories present another data source to secure and most existing data security and compliance approaches will not scale

Big data environments allow organizations to aggregate more and more data—much of which is financial, personal, intellectual property or other types of sensitive data. Most of the data is subject to compliance regulations such as Sarbanes-Oxley Act (SOX), Health Insurance Portability and Accountability Act (HIPAA), Payment Card Industry Data Security Standard (PCI-DSS), Federal Information Security Management Act (FISMA) and the EU Data Privacy Directive. Sensitive data is also a primary target for hackers.

Data security professionals need to take an active role early. The reality is that pressure to make quick decisions can result in data security professionals being left out of key decisions or be seen as inhibitors of business growth. However, the risk of lax data security is well known and documented.

Corporations and their officers may face fines from $5,000 USD to $1,000,000 USD per day, and possible jail time if data is misused. According to the 2011 Cost of Data Breach Study conducted by the Ponemon Institute (published March 2012) the average organizational cost of a data breach is $5.5M USD. Data breaches can cost their companies an average of $194 USD per compromised record.

Hard penalties are only one example of how organizations can be harmed; other negative impacts resulting from a data breach include share price erosion and negative publicity resulting in irreparable brand damage.

At this time, only the most forward thinking and highly innovative firms have deployed big data environments. Thus, a window of opportunity is open to establish a set of best security practices.

Protecting data security is a detailed, continuous responsibility which should be part of every best practice. Protecting data requires a holistic approach to protect organizations from a complex threat landscape across diverse systems. Data security compliments other security measures such as endpoint security, network security, application security, physical site security and more to create an in-depth defense approach.

> *"Auditors are looking more closely at how access to the huge data stores in these systems is controlled, and enterprises are being pressured to adopt more aggressive and expansive data controls".*
>
> – Gartner: "Database Activity Monitoring Is Evolving Into Database Audit and Protection," February 2012

Raising security awareness in big data environments

Data security can be addressed in an efficient and effective manner to satisfy all parties. Start big data security planning immediately. Building security into big data environments will reduce costs, risks, and deployment pain.

So what needs to be protected? Most organizations are turning to Hadoop-based systems for fast, reliable analysis of big data. Many organizations deploy Hadoop along side their existing database systems, allowing them to combine traditional structured data and new unstructured data sets in powerful ways. Hadoop consists of reliable data storage using the Hadoop Distributed File System (HDFS), a column-oriented database management system that runs on top of HDFS called HBase and a high-performance parallel data processing technique called MapReduce.

Hadoop environments need to be protected using the same rigorous security strategies applied to traditional database systems, such as databases and data warehouses, to support compliance requirements and prevent breaches.

Security strategies which should be implemented for Hadoop environments include:

- Sensitive data discovery and classification: Discover and understand sensitive data and relationships before the data is moved to Hadoop so that the right security policies can be established downstream.
- Data access and change controls: Establish policies regarding which users and applications can access or change data in Hadoop.
- Real-time data activity monitoring and auditing: Understand the who, what, when, how and where of Hadoop access and report on it for compliance purposes.
- Data protection: Transform data in Hadoop through masking or encryption.
- Data loss prevention: Establish an audit trail for data access and usage to ensure data is not lost.
- Vulnerability management: Understand weaknesses and put policies in place to remediate.
- Compliance management: Build a compliance reporting framework into Hadoop to manage report generation, distribution and sign off.

Organizations need to be able to answer questions like:

1. Who are running specific big data requests?
2. Are users authorized to make requests?
3. What map-reduce jobs are users running?
4. Are users trying to download sensitive data or is the request part of a job requirement, for example, a marketing query?

Compliance mandates are enforced the same across big data environments and more traditional data management architectures. In other words, the rush for big data benefits is not an excuse for overlooking security. Organizations should be prepared to follow audit requirements (see figure 1).

The Compliance Mandate

Audit Requirements	COBIT (SOX)	PCI-DSS	ISO 27002	Data Privacy & Protection Laws	NIST SP 800-53 (FISMA)
1. Access to Sensitive Date (Successful/failed SELECTS)		✓	✓	✓	✓
2. Schema Changes (DDL) (Create/Drop/Alter Tables, etc.)	✓	✓	✓	✓	✓
3. Data Changes (DML) (Insert, Update, Delete)	✓		✓		
4. Security Exceptions (Failed logins, SQL errors, etc.)	✓	✓	✓	✓	✓
5. Accounts, Roles & Permissions (DCL) (Grant, Revoke)	✓	✓	✓	✓	✓

DDL - Data Definition Language (aka schema changes)
DML - Data Manipulation Lauguage (data value changes)
DCL - Data Control Lauguage

Figure 1. Compliance and audit requirements.

Security fundamentals: Three tips to improve security in big data environments

According to The Future Of Data Security And Privacy: Controlling Big Data, Forrester Research, Inc., January 26, 2012, organizations can control and secure the extreme volumes of data in big data environments such as Hadoop by following a three step framework (see figure 2).

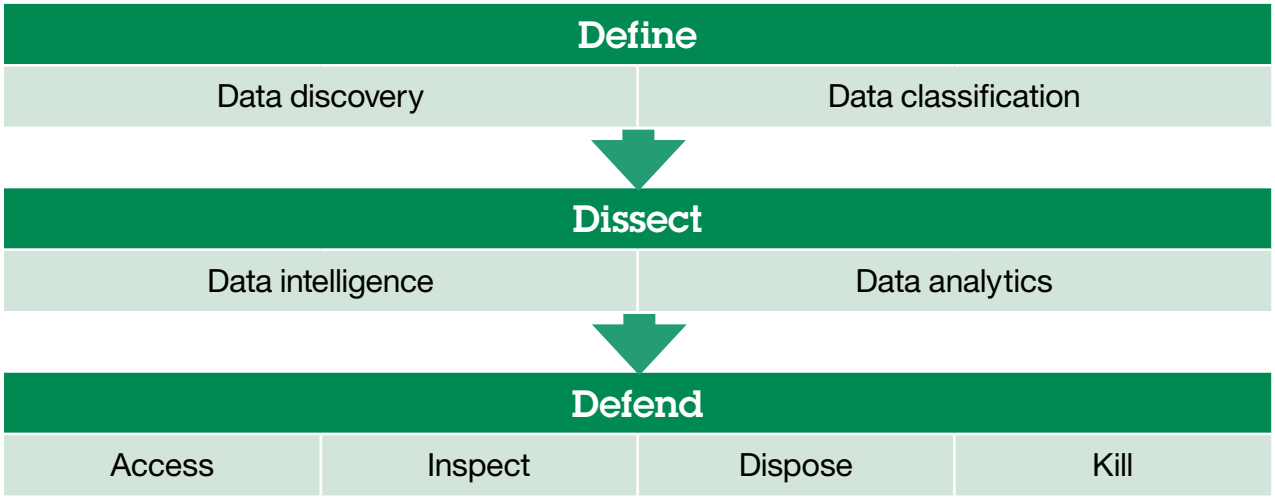


Figure 2: The Future Of Data Security And Privacy: Controlling Big Data, Forrester Research, Inc., January 26, 2012 Big Data Security and Control Framework

Define

Most organizations are just starting down the path of implementing a big data environment, so they don't know which types of data (structured or unstructured) they want to include in a big data repository like Hadoop.

The planning phase presents the perfect opportunity to start a dialog across data security, legal, business and IT teams about sensitive data understanding, discovery and classification. A cross-functional team should identify where data exists, decide on common definitions for sensitive data, and decide what types of data will move into Hadoop. Also, organizations should establish a life-cycle approach to continuously discover data across the enterprise.

Dissect

Big data environments are highly valuable to the business. However, data security professionals also benefit because big data repositories like Hadoop can store security information. Data security professionals can leverage big data environments to more efficiently prioritize security intelligence initiatives and more effectively place the proper security controls.

For example, understanding more specifics about potential attackers such as who or what is accessing data in Hadoop and how and when the data is being accessed, can help align data security strategies such as sending a real-time alert notifying the information security group to take action.

Defend

Aggregating data by nature increases the risk that a cybercriminal or insider (malicious or otherwise) can compromise sensitive information. Therefore, organizations should strictly limit the number of people who can access repositories like Hadoop.

Big data environments should include basic security and controls as a way to defend and protect data. First, access control ensures that the right user gets access to the right data at the right time. Second, continuously monitoring user and application access is highly important especially as individuals changes roles or leave the organization. Monitoring data access and usage patterns can alert security teams to potential abuse or security policies violations like an administrator altering log files. Typically internal attackers or cybercriminals will leave clues or artifacts about their breach attempts that can be detected through careful monitoring. Monitoring helps ensure security policies are enforced and effective.

Organizations can secure data using data abstraction techniques such as encryption or masking. Generally, cybercriminals cannot easily decrypt or recover data after it has been encrypted or masked. The unfortunate reality is that organizations need to adopt a zero trust policy to ensure complete protection.

IBM InfoSphere Guardium solutions support a data security and control framework

According to The Future Of Data Security And Privacy: Controlling Big Data, Forrester Research, Inc., January 26, 2012, there are several actions organizations can take today to better secure big data environments.

- Move your controls closer to the data itself
- Leverage existing technologies to control and protect big data
- Ask legal to define clear policies for data archiving and data disposal
- Diligently control access to big data resources and watch user behavior

IBM® InfoSphere® Guardium® secures Hadoop environments by:

- Vigilantly monitoring Hadoop activity from applications and users (both internal and external) in real time, and alerting on policy violations; tracking user interaction with the data to detect unusual patterns from both privileged users and external access and alerting SIEM dashboards for appropriate remediation such as alerting, blocking or connection termination.
- Auditing activity in Hadoop and reporting on activities; to fulfill compliance requirements and support forensic investigations, IT can gather Hadoop activity data into non-repudiable audit trails and appropriately formatted reports. Separation of duties is a key best practice since IT staff must not be able to tamper with reports about the systems they manage. Out-of-the-box preconfigured policies and reports are available as well as sign off management and entitlement reports.
- Enforcing change controls across the high volume, velocity and variety of big data.
- Implementing automated and centralized controls across database, data warehouses, file shares and Hadoop.
- Encrypting and masking data to make it unusable.

Conclusion: Build security into big data environments

Organizations don't have to feel overwhelmed when it comes to securing big data environments. The same security fundamentals for securing databases, data warehouses and file share systems can be applied to securing Hadoop implementations. InfoSphere Guardium solutions scale to protect both traditional data management architectures and big data environments and protect against a complex threat landscape including insider fraud, unauthorized changes and external attacks while remaining focused on business goals and automating compliance.

InfoSphere Guardium prevents leaks from databases, data warehouses and big data environments such as Hadoop, ensures the integrity of information and automates compliance controls across heterogeneous environments. It provides a scalable platform that enables continuous monitoring of structured and unstructured data traffic as well as enforcement of policies for sensitive data access enterprise-wide. A secure, centralized audit repository combined with an integrated workflow automation platform streamlines compliance validation activities across a wide variety of mandates. It leverages integration with IT management and other security management solutions, such as SIEMs like QRadar, to provide comprehensive data protection across the enterprise.

The end goal is to improve security decision-making based on prioritized, actionable insight derived from monitoring big data environments, like Hadoop, and identify when an advanced targeted attack has bypassed traditional security controls and penetrated the organization.

For more information: **ibm.com**/guardium

IBM Corporation
Software Group
Route 100
Somers, NY 10589 U.S.A.

Produced in the United States of America
October 2012

IMB14137-USEN-00