Data protection for big data environments

*Design real-time, business-driven data protection policies across the enterprise*
Introduction
Security fundamentals
Data protection challenges in the era of big data
Why not start now?
What are the benefits?
Why InfoSphere?
Introduction

Organizations are eager to harness the power of big data. But as new big data opportunities emerge, ensuring the veracity and security of source information becomes exponentially more difficult. If these challenges are not addressed directly, end users will lose confidence in the insights generated from big data. As a result, they will be unable to act on new opportunities and address important risks.

The tremendous volume, variety and velocity of big data mean that the old, manual methods of discovering, governing and correcting data might not apply. Organizations need to automate information integration and governance at the point of data creation and within big data applications. Effective information integration and governance can help protect information and improve the accuracy of big data insights.

Information integration and governance solutions must offer automated discovery and profiling, and they must facilitate an understanding of diverse data sets to provide complete context required to make informed decisions. They must be agile enough to accommodate a wide variety of data and seamlessly integrate with diverse technologies, from data marts to Apache Hadoop systems. Information integration and governance solutions also must monitor and protect sensitive information as part of big data applications.
Data security and privacy concerns in the new era of computing

Maintaining data security is critical. In this era of big data, the average cost of security-related events is estimated to cost American companies USD40 million per year.¹

Worldwide, cybercrime costs people and businesses USD388 billion in lost time and money—USD100 billion more than the cost of the world’s illegal drug market.² Identity theft is at a three-year high, affecting 12.6 million US consumers and costing more than USD21 billion each year.³

Data breaches not only cost money; they can also drive down stock prices and cause irreparable brand damage. Protecting brand reputation and preserving customer trust are two of the top three organizational goals that depend on good data protection.⁴

Data protection is also required by law. More than 50 international laws, such as Canada’s Privacy Act, Germany’s Federal Data Protection Act, Argentina’s Personal Data Protection Act and Korea’s Act on Personal Information Protection, mandate data protection.⁵

“I am going live in three months with our new Hadoop-based application. My biggest fear is that an auditor will shut us down because we do not have a secure audit trail telling us who is accessing sensitive data and when.”

— Large global bank
The rapidly increasing volume, variety and velocity of data are driving the immediate need for data protection. With 2.5 quintillion bytes of data created every day, organizations cannot wait to establish business-driven protection policies for keeping data safe. The rising volume of data and the growing number of analytics systems storing sensitive data exponentially increases the risk of breaches.

In many cases, organizations will need multiple data protection policies to accommodate a wide variety of data. Text, audio, log files and clickstreams have unique characteristics and challenges. At the same time, data protection policies need to support an increasing velocity of data, ensuring that no data is left unprotected.

IBM® InfoSphere® has evolved information integration and governance to help organizations create and enforce real-time data security and privacy policies to keep customer, business, personally identifiable information (PII) and other types of sensitive data safe.

“By focusing on the three S’s—service, speed and security, brands can improve customer lifetime value, strengthen security and increase brand loyalty. Consumer trust in technology is key. It should be used to reassure customers that their security concerns are being addressed while simultaneously improving the customer experience, speeding up the time and driving down the cost of each and every customer service interaction.”

— Simon Culmer, Managing Director, UK, Avaya
Security fundamentals: Integrate security for big data environments from the start

As organizations plan big data projects, business and IT leaders must recognize that data security cannot simply be added later—it must be built into big data implementations from the start. If there is a data security breach, the damage to customer relationships and the brand image cannot be undone. Organizations must replace reactive approaches with strategies that protect data, ensure data integrity and prevent breaches.

Big data implementations often involve sharing data, putting all types of data into a central repository. If security is not incorporated into this platform from the initial intake steps, the opportunity to protect data is lost. That means data protection policies should be business-driven, easily measured and applied to data as it moves across the enterprise through its lifecycle. Security policies must also be available on demand to respond to threats in real time, since users are accessing data through a growing number of devices, and they are sharing information across the enterprise at lightning speeds. Organizations must secure sensitive data at the source before that data proliferates and is exposed.

How do I know which data needs protection?
Sensitive data is everywhere. It is in production and nonproduction systems, reports and analytics systems, files, documents, applications and a wide range of other places. Data in these systems is often shared with third parties such as development partners, outsourced IT staff and third-party vendors. However, the source of this data is typically a transactional system or analytics platform.

Databases and warehouses such as IBM DB2® for Linux, UNIX and Windows (LUW), DB2 for z/OS®, Oracle Database, Teradata and IBM PureData™ excel at storing and optimizing online application processing (OLAP) and online transaction processing (OLTP) workloads containing structured and unstructured data. Most organizations depend on these platforms to support critical business operations and reporting.
At the same time, many organizations are turning to Hadoop for fast, reliable analysis of big data in real time. Many organizations deploy Hadoop alongside their existing IT systems to combine traditional data and new data sets in powerful ways. Solutions that augment Hadoop capabilities, such as IBM InfoSphere BigInsights™ and Cloudera, as well as NoSQL databases such as Greenplum and MongoDB, are growing in popularity.

All enterprise data stores and the data inside these systems need to be protected. In this era of big data, much of the information that organizations collect is sensitive. Organizations need to implement business-driven, centralized approaches that secure this rising volume of sensitive data and maintain privacy.

Organizations must secure data wherever it resides, including test, production and cloud environments.
Data protection challenges in the era of big data

The emergence of new security threats, such as spear phishing and advanced persistent threats, and changes in the way people access data make it difficult to protect big data. Users are accessing data around the clock, from a wider variety of devices than ever before—and fewer access points are corporately controlled.

Additional challenges presented by big data environments include:

- 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.
- The aggregation of data from across the enterprise means that sensitive data is in a single repository.
- Big data repositories present another data source to secure, and most existing data security and compliance approaches will not scale.

Adding to the complexity, organizations face new, more rigorous data security and privacy regulations from governments and industry organizations. Much of the data inside big data environments is financial, personal, intellectual property or other types of sensitive data. Consequently, that data is subject to regulations stemming from the Sarbanes-Oxley Act (SOX), the Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI-DSS), the Federal Information Security Management Act (FISMA), the EU Data Privacy Directive and other government mandates. At the same time, that sensitive data is often a primary target for hackers.
How should organizations approach big data security? Here are five best practices to follow:

• **Protect data as it moves into and out of big data environments:** As data is being aggregated, shared and turned into actionable insight, how can you be sure it won’t fall into the wrong hands or be accessed by someone without a valid business purpose?

• **Integrate security controls into big data platforms:** Data security and privacy policies shouldn’t be process- or technology-heavy. They should align with business priorities and seamlessly integrate with enterprise systems.

• **Leverage existing technologies to control and protect big data:** Don’t throw away all your best practices! Work to extend data masking, data encryption, and data monitoring and auditing across the enterprise to emerging platforms.

• **Build consensus about what constitutes sensitive data:** No data security and privacy strategy can be successful without agreement from a cross-functional team about what constitutes sensitive data. For example, what information constitutes a corporate secret? Agreement across lines of business, legal staff, security and IT is important to establish the right security and privacy strategy.

• **Automatically control access to big data resources and monitor user behavior:** New data sources are added, dropped and expanded every day. Your IT environment is always changing and your security and privacy strategies and strategies must keep up.
As part of their big data projects, organizations must implement data security strategies that continuously prevent, detect and address vulnerabilities. Data security strategies must not only offer strong defenses, but also provide active monitoring to identify suspicious activity in real time. Organizations must be able to quickly discover and remediate any security issues as they evolve and before they cause damage.
Why not start now?

Don’t be overwhelmed when it comes to securing big data environments. While the challenges may be different, you can build on the security fundamentals for traditional systems—like access controls, monitoring, auditing and data masking—to secure big data systems.

By building security into big data environments, organizations can satisfy a wide variety of mandates and integrate security with other IT management solutions. Organizations can implement security for leading authentication protocols, security information and event management (SIEM) solutions, ticketing systems, application servers, archival systems and other systems. The end goal is to improve security decision making based on prioritized, actionable insight, and be able to identify when an advanced targeted attack has bypassed traditional security controls and penetrated the organization.

But remember: the longer you wait, the more data you aggregate and the greater the risk of a breach. The time to get started is now.
The InfoSphere advantage
The InfoSphere portfolio, which includes InfoSphere Guardium® and InfoSphere Optim™, provides data-access and change-control capabilities, real-time data monitoring and auditing, data protection and loss prevention, vulnerability management and sensitive data discovery, and classification capabilities to support compliance requirements and prevent breaches.

InfoSphere data security and privacy capabilities are available on all leading big data platforms, including Cloudera, IBM PureData Systems, InfoSphere BigInsights and Teradata, as well as emerging NoSQL databases such as MongoDB, Greenplum and more.

<table>
<thead>
<tr>
<th>InfoSphere capability</th>
<th>Functions enabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery and classification of sensitive assets in structured and unstructured data sources</td>
<td>Automatically discover and classify data to enforce security policies.</td>
</tr>
<tr>
<td>Assessment of configurations and vulnerabilities and hardening of data stores such as Oracle Database, DB2, DB2 for z/OS and emerging NoSQL databases</td>
<td>Assess data vulnerabilities through a variety of techniques; for example, checking administrative-level access privileges and verifying that the application accessing data does not contain known vulnerabilities.</td>
</tr>
<tr>
<td>Change management across heterogeneous big data environments</td>
<td>Monitor data stores to ensure that changes do not compromise security. If such changes do occur, the right people will be alerted to take the appropriate corrective action.</td>
</tr>
<tr>
<td>Discovery and classification of sensitive assets in structured and unstructured data sources</td>
<td>Assess data vulnerabilities through a variety of techniques; for example, checking administrative-level access privileges and verifying that the application accessing data does not contain known vulnerabilities.</td>
</tr>
<tr>
<td>Discovery and classification of sensitive assets in structured and unstructured data sources</td>
<td>Assess data vulnerabilities through a variety of techniques; for example, checking administrative-level access privileges and verifying that the application accessing data does not contain known vulnerabilities.</td>
</tr>
<tr>
<td>Change management across heterogeneous big data environments</td>
<td>Monitor data stores to ensure that changes do not compromise security. If such changes do occur, the right people will be alerted to take the appropriate corrective action.</td>
</tr>
<tr>
<td>Change management across heterogeneous big data environments</td>
<td>Monitor data stores to ensure that changes do not compromise security. If such changes do occur, the right people will be alerted to take the appropriate corrective action.</td>
</tr>
<tr>
<td>Real-time activity monitoring for Hadoop-based platforms</td>
<td>Vigilantly monitor user interaction with the data, detecting any unusual pattern of access among privileged users and external-access users.</td>
</tr>
<tr>
<td>Access and entitlement management from a central location for all big data platforms</td>
<td>Set up and enforce policies to continually manage access to data stores.</td>
</tr>
<tr>
<td>Data encryption for data at rest inside big data repositories</td>
<td>Transform data through encryption to make it useless to hackers.</td>
</tr>
<tr>
<td>Data redaction for unstructured documents and forms</td>
<td>Automatically recognize and remove sensitive content from unstructured data sources such as scanned documents as well as PDF, TIFF, XML and Microsoft Word files. Safeguard sensitive data while supporting information sharing for business use.</td>
</tr>
<tr>
<td>Data masking as data moves into and out of Hadoop-based platforms</td>
<td>Mask data on demand, wherever and whenever across the enterprise. Apply masking to data at rest, data in flight, relational data, flat files, data sets such as IBM IMS™ or VSAM data sets, data in reports and in documents, data moving into or out of Hadoop, and more.</td>
</tr>
</tbody>
</table>
Semantic masking: Making secure, useful data a reality

IBM has pioneered a new approach to data masking known as *semantic masking* to keep pace with data privacy requirements in the new era of computing. The technology was developed in the IBM Zurich Research Lab to mask data in context based on rules to ensure accurate and consistent results for analytics.

Semantic masking facilities analytics and ensures sensitive data cannot be traced to an individual entity. The value of semantic masking is to retain the utility (usefulness) of the data while also adhering to compliance/regulation requirements.

Let's explore an example scenario. An international healthcare provider would like to conduct analysis on sensitive data from various sources and geographies across lifestyles, geographies and age groups. The provider wants to know if financial status (standard of living) has anything to do with the propensity for a family member to get diabetes. They need a significant volume of patient data to be able to reach a conclusion across all subsidiaries worldwide.

Without semantic masking, international medical and insurance laws prohibit this type of sensitive data aggregation because it can usually be tied to an individual entity. But with a semantic masking solution, the international healthcare provider is able to reach an outcome and provide analysis quickly because the algorithms ensure data is masked appropriately. For example, the provider needs to determine if there are enough Pancreatic Cancers (Symptom Code 157) to summarize all 157.xx into 157. Semantically masked data will have the same symptoms and gender but the age, family income and ethnicity are intelligently masked to the proper range and to a valid set of data points. Therefore, it is impossible to identify a person or tie a person to income bracket, exact age, gender or ethnic background, and researchers achieve valid results while protecting privacy.

<table>
<thead>
<tr>
<th>Original URL</th>
<th>NAME</th>
<th>HOME ADDRESS</th>
<th>SYMPTOM CODE</th>
<th>HOUSEHOLD INCOME</th>
<th>HOME PHONE</th>
<th>SEX</th>
<th>ETHNICITY</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>John Smith</td>
<td>5 Rue de la Paix Paris, France</td>
<td>157.0, 157.1, 157.2, 157.3, 185</td>
<td>75,000</td>
<td>01 58 71 12 34</td>
<td>M</td>
<td>Caucasian</td>
<td>43</td>
</tr>
<tr>
<td>Asia</td>
<td>Steve Jones</td>
<td>199 Huangpu Road Hongkou Shanghai, China</td>
<td>157.0, 185, 493.00, 493.02</td>
<td>44,000</td>
<td>021 6393 1234</td>
<td>M</td>
<td>Asian</td>
<td>21</td>
</tr>
<tr>
<td>North America</td>
<td>Angela Garcia</td>
<td>926 Central, Reno NV, 89521</td>
<td>157.1, 185, 493.02</td>
<td>83,000</td>
<td>775 587 1578</td>
<td>F</td>
<td>Latino</td>
<td>47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantically Masked URL</th>
<th>NAME</th>
<th>HOME ADDRESS</th>
<th>SYMPTOM CODE</th>
<th>HOUSEHOLD INCOME</th>
<th>HOME PHONE</th>
<th>SEX</th>
<th>ETHNICITY</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>Jerry Jones</td>
<td>5 Rue de la Paix Paris, France</td>
<td>157, 185</td>
<td>79,500</td>
<td>01 55 27 12 34</td>
<td>M</td>
<td>Caucasian</td>
<td>41</td>
</tr>
<tr>
<td>Asia</td>
<td>Steve Smith</td>
<td>199 Huangpu Road Hongkou Shanghai, China</td>
<td>157, 185, 493</td>
<td>29,000</td>
<td>021 6322 5767</td>
<td>M</td>
<td>Asian</td>
<td>21</td>
</tr>
<tr>
<td>North America</td>
<td>Angela Hernandez</td>
<td>1142 Main, Reno NV, 89521</td>
<td>157, 185, 493</td>
<td>68,500</td>
<td>775 773 1142</td>
<td>F</td>
<td>Latino</td>
<td>47</td>
</tr>
</tbody>
</table>
What are the benefits?

Focusing on data security and privacy within big data implementations can deliver significant value to organizations.

Organizations can:

- **Prevent data breaches:** Avoid disclosing or leaking sensitive data
- **Help ensure data integrity:** Prevent unauthorized changes to data, data structures, configuration files and logs
- **Reduce the cost of compliance:** Automate and centralize controls and simplify the audit review process

- **Protect privacy:** Prevent disclosure of sensitive information by masking or de-identifying data in databases, applications and reports, on demand, across the enterprise

Read on for a few examples of how big data security works in real-world scenarios.

**Protecting data for a global bank**

A major global bank needed to secure enterprise data and preserve data integrity across multiple business units, including the retail, corporate, investment and mortgage divisions. The bank needed to pass upcoming audits for multiple regulations, such as PCI-DSS and SOX.

The data requiring protection included financial, HR, enterprise resource planning (ERP), credit card, PII and intellectual property. The data existed across a diverse set of databases and warehouses, including Oracle, SQL Server, Sybase, DB2 for LUW, DB2 for z/OS and IBM i, IBM Informix®, MySQL and Teradata (over 2,000 database instances).
Using InfoSphere Guardium Data Activity Monitor, the bank saved USD1.5 million per year in storage costs, since it no longer relied on native audit trails. In addition, the bank reduced security costs by USD20 million. InfoSphere Guardium Data Activity Monitor is now a standard part of the bank’s infrastructure, protecting against both internal and external threats. Implementing this solution also has helped foster a culture shift by creating a greater awareness of the importance of data security across the enterprise.

**Ensuring regulatory compliance for a health benefits company**
A health benefits company that offers a wide variety of health and life insurance products is among the fastest-growing health and life insurance companies in the United States. The key to the company’s success has been a consumer-centric strategy that relies on analytics. However, this strategy is often inhibited because of the highly regulated nature of the US healthcare industry, specifically the US Affordable Care Act and HIPAA.

This client was having trouble keeping up—too many regulations, too many applications to test and very rapid application updates motivated a focus on test data and privacy.

In the past, the company was making full copies of its production database for testing purposes, which caused an increased risk of exposing sensitive personal information. By implementing InfoSphere Optim, the company has accelerated deployment of new statistical analysis models and reduced development time by 10 to 20 percent. The new solution is also helping the company maintain compliance in test environments through data masking.
**Speeding time-to-security**

McAfee is a dedicated security technology company delivering solutions and services that help secure systems and networks around the world. McAfee products protect consumers and businesses from the latest malware and emerging online threats.

Recently, the company needed to rapidly implement new security and compliance measures to pass an upcoming PCI-DSS audit. McAfee needed to protect millions of PCI transactions while also maintaining strict service-level agreements (SLAs) with Internet service provider (ISP) customers such as Comcast and Cox. The goal was to implement automated security and compliance capabilities across a complex infrastructure, including multiple data centers and clustered databases, in less than 48 hours.

McAfee met its goals by implementing InfoSphere Guardium Data Activity Monitor. The company eliminated the manual process of correlating database log files and automated database activity monitoring into its larger security infrastructure.

“McAfee needed a solution with continuous real-time visibility into all sensitive cardholder data in order to quickly spot unauthorized activity and comply with PCI-DSS. But given our significant transaction volumes, performance and reliability considerations were crucial.”

— McAfee representative
Why InfoSphere?

**Build security and privacy into big data environments**
InfoSphere Guardium and InfoSphere Optim solutions help organizations implement the data security and privacy capabilities they need for big data environments. With these solutions, organizations can 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 and InfoSphere Optim solutions provide a complete data protection approach to secure diverse data types (structured, unstructured, online and offline) across different locations, including production and nonproduction (development, test and training) environments and big data platforms. At the same time, their data security and privacy capabilities help organizations streamline compliance by implementing unified controls and consistent enforcement of security policies across the entire enterprise.

InfoSphere Guardium and InfoSphere Optim solutions for data security and privacy are:

- **Business-driven:** Set policies based on business requirements
- **Simple to use:** Leverage out-of-the-box providers for protection policies
- **Ready for big data:** Support for leading databases, operating systems and applications across mainframe, distributed and big data environments
Leverage the power of integrated InfoSphere capabilities

As the foundation of the IBM big data platform, InfoSphere provides market-leading functionality across all the capabilities of information integration and governance. InfoSphere creates confidence in big data by helping to ensure that it is trustworthy and protected.

InfoSphere is designed to handle big data. It provides optimal scale and performance for massive data volumes, agile and rightsized integration and governance for the increasing velocity of data, and support for a wide variety of data types and big data systems. InfoSphere makes big data and analytics projects successful by delivering the confidence to act on insight.

InfoSphere capabilities include:

• **Metadata, business glossary and policy management:** Define metadata, business terminology and governance policies with IBM InfoSphere Business Information Exchange.

• **Data integration:** Handle all integration requirements, including batch data transformation and movement (InfoSphere Information Server), real-time replication (InfoSphere Data Replication) and data federation (InfoSphere Federation Server).

• **Data quality:** Parse, standardize, validate and match enterprise data with InfoSphere Information Server for Data Quality.

• **Master data management:** Act on a trusted view of your customers, products, suppliers, locations and accounts with InfoSphere Master Data Management (MDM).

• **Data lifecycle management:** Manage the data lifecycle from test data creation through retirement and archiving with InfoSphere Optim.

• **Data security and privacy:** Continuously monitor data access, protect repositories from data breaches and support compliance with InfoSphere Guardium. Ensure sensitive data is masked and protected with InfoSphere Optim.
For more information
To learn more about the IBM approach to information integration and governance for big data, as well as the InfoSphere platform, InfoSphere Optim and InfoSphere Guardium, check out these resources:

- [ibm.com/software/data/information-integration-governance](ibm.com/software/data/information-integration-governance)
- **E-book**: Planning a data security and auditing deployment for Hadoop
- **E-book**: Top 3 Myths about Big Data Security
- **E-book**: Top tips for securing big data environments
- **White paper**: Three guiding principles to improve data security and compliance
- **White paper**: Understanding holistic database security: 8 steps to successfully securing enterprise data sources
- **White paper**: Data masking everywhere