Enterprise Information Protection - The Impact of Big Data

By Mike Ferguson
Intelligent Business Strategies
March 2013
# Table of Contents

Introduction .................................................................................................................................................. 4

What is Enterprise Information Protection? ............................................................................................. 5

Where Are We in Data Audit, Access Security and Protection Today? ................................................. 8

Requirements for Enterprise Information Protection? ................................................................................ 10
  Information Classification Requirements ............................................................................................... 10
  Data Landscape Requirements .............................................................................................................. 11
  User Requirements .................................................................................................................................. 11
  Device Requirements ............................................................................................................................... 11
  Change Management Requirements ....................................................................................................... 12

An Introduction To Big Data .......................................................................................................................... 13
  Types of Big Data ...................................................................................................................................... 13
  What Is Big Data? ...................................................................................................................................... 14
  New Big Data Analytical Workloads ........................................................................................................ 14
  Technology Options for End-to-End Big Data Analytics ....................................................................... 15
  The New Enterprise Analytical Ecosystem ............................................................................................. 15

The Impact of Big Data on information Protection .................................................................................... 18
  Using Big Data Analytics for Security Analysis ...................................................................................... 19

New Security Requirements To Protect Information In Big Data Environments ...................................... 20

Implementing Information Protection in a Big Data Enterprise Using IBM Technologies ... ........... 23
  The IBM Big Data Platform ..................................................................................................................... 23
  Hadoop - IBM InfoSphere BigInsights ...................................................................................................... 24
  IBM InfoSphere BigInsights on IBM System zEnterprise ....................................................................... 24
  IBM PureData System for Analytics (powered by Netezza technology) ............................................. 24
  IBM InfoSphere Warehouse, Smart Analytics System and IBM PureData System for Operational Analytics ......................................................................................................................... 24
  IBM DB2 Analytic Accelerator (IDAA) ...................................................................................................... 25
  IBM Big Data Platform Accelerators ....................................................................................................... 25
  IBM Information Management for the Big Data Enterprise .................................................................... 25

IBM Products for Enterprise Information Protection ................................................................................ 25

IBM System z In A IBM Big Data Environment ....................................................................................... 26

Delivering Security For Big Data In a System z Environment ................................................................. 27
  Securing Information on IBM InfoSphere Streams .................................................................................... 28
  Securing Information on IBM System z InfoSphere Warehouse and IBM DB2 Analytics Accelerator ............................................................................................................................... 28
  Securing and Redacting Information on IBM BigInsights and IBM System z .................................. 29
  Securing and Redacting Information During Smart Consolidation .................................................. 30

Delivering Security From Big Data In a System z Environment .............................................................. 30
Conclusion ................................................................................................................................................. 31

Appendices - Key Requirements for Protecting Data and Preventing Security Breaches .. 32
  Data Landscape Requirements .............................................................................................................. 32
  Software Access Requirements ......................................................................................................... 32
  Environment Requirements ................................................................................................................. 33
  Usage Requirements .......................................................................................................................... 33
  Vulnerability Assessment Requirements ............................................................................................. 34
  Prevention Requirements ................................................................................................................... 35
  Enforcement Requirements ............................................................................................................... 38
  Performance Requirements ............................................................................................................... 39
INTRODUCTION

The shift from centralised computing in late ‘80s and early ‘90s increased complexity and made systems harder to manage.

In the late 1980’s and early 1990’s we saw a fundamental shift in computing when companies started to move away from centralised systems towards client server computing and distributed systems. In truth however, what happened was not so much distributed systems but more like ‘stand-alone’ autonomous systems. Each system was designed to support a specific business function and was comprised of an application deployed on its own servers with its own data. The arrival of multiple servers in the enterprise spawned a new problem – how to manage this new more complex environment. This resulted in agent-based distributed systems management software emerging to help systems administrators use the power of the network to manage multiple servers across the organisation.

While this helped people manage and monitor multiple systems, one ‘side-effect’ of the rise of distributed computing that was not well addressed was that the problem of distributed data. The impact of distributed data was significant in that data became difficult to share, keep consistent and synchronised especially as users of the different applications started to maintain data in different application specific databases. The result was that data became much more complex to manage. It needed to flow between systems to execute business processes. File transfers rocketed and demand grew for data to be integrated to support cross-functional management reporting and analysis. Batch jobs grew at a very rapid rate moving data between systems and the era of the data-warehouse was born. Today, we have moved on at pace, to web and mobile client enabled multi-tier computing, business process management and service oriented architecture. But still the problem of distributed data remains. If anything it is getting worse. Cloud computing is now upon us, meaning that business application systems now exist outside the enterprise as well as inside it. Multiple data warehouses have also emerged as line of business has taken hold of business intelligence and analytics. Data warehouse appliances have emerged; creating yet more data stores and for many, master and reference data is still not under control.

Today data is more distributed than ever and often still not under control.

Securing and protecting ‘at risk’ sensitive data in a distributed environment is a major challenge.

We are at the point now where data is heavily fractured making it harder and harder to manage. This is particularly the case for security where ‘at risk’ data needs to be protected. The challenge is significant. And yet despite this challenge, another wave of computing has arrived. That wave is Big Data, which is sweeping the industry by storm. Given the complexity already upon us in managing data and managing information security in a distributed computing environment, what then is the impact of big data on the enterprise? What is big data and how much of a challenge does it pose to the already stretched need to enforce enterprise information protection?

Big Data adds more complexity and introduces new requirements to keep data protected.

This paper looks at this problem by first defining what enterprise information security and privacy involves, then looking at requirements that need to be met before introducing big data and identifying what impact this has on those requirements. It then takes a look of it at what one vendor, IBM, is offering in the area of big data and what it is trying to do to help customers bring big data technologies on board while continuing to protect and secure their information.
WHAT IS ENTERPRISE INFORMATION PROTECTION?

In an information governance white paper I authored in November 2011\(^1\), I outlined the information protection challenge. A variation of Figure 1 in that paper is shown in Figure 1 below. This has been expanded in the areas of risk prevention and change management to show the overall complexity that most enterprises are facing when it comes to tackling enterprise information protection. It is without doubt a daunting problem and well worthy of a dedicated team to deal with it.

Looking at Figure 1 in a clockwise direction from the box entitled ‘Information’, tackling the challenge of enterprise information protection involves:

- Classifying information to define what is sensitive and what is not
- Understanding the existing data landscape to determine where that sensitive information is located in order to protect it
- Understanding the existing application portfolio and the information these applications access and maintain
- Ensuring enterprise information protection covers multiple environments
- Accounting for different types of users and the devices they chose to access applications and information with
- Upholding compliance regulations and legislation
- Assessing vulnerability to information security/privacy breaches

\(^1\) Information Governance: Audit and Protection on the IBM System z Platform
Defining policies, processes, roles, tests and actions needed to prevent breaches in security

Applying those policies to enforce information security

Managing change so that information security is not compromised by changes made

Maintaining information availability

In addition to this, enforcing enterprise information security and privacy also involves integration with other infrastructure software. A good example of this is corporate LDAP directories, which are needed to help enforce role-based access to information.

In order for enterprise information protection to be implemented, means defining what information has to be protected and then locating that data to be able to secure it. For most organisations, that data is stored in databases and files scattered across a highly distributed landscape of multiple DBMS and file systems that run on a range of operating systems on servers in multiple locations. This includes both on-premise and cloud based platforms. The fact that sensitive data could be widely distributed across this landscape increases the risks of security being compromised. Finding, controlling access to and protecting sensitive data content in this kind of environment is a real challenge without the right tools to help you.

In terms of users, identity management, user risk classification, authentication, authorisation and multi-device access security are all very much part of an enterprise information protection initiative to control user access. Both desktop and mobile devices need to be considered with the added complication that mobile devices can be easily lost or stolen. Types of user are also important when it comes to information access and authorisation. This would include external users such as customers, partners and suppliers, internal business users and IT professionals. Some IT professionals are ‘privileged’ in that they have administrative power that allows them access to potentially any data including sensitive customer, employee and financial data. For this reason, privileged IT professionals themselves need to be monitored and duties separated to control what they can and cannot do without authorisation.

IT developers who build systems and IT operations staff who manage and run those systems also need to be taken into account. IT developers often work with production data during development and testing. Therefore information has to be secured in development, testing and production environments.

Also information produced from accessing sensitive information may itself contain subsets of that sensitive information. An example of this is reporting. Therefore sensitive information in shared reports needs to be removed or redacted. If analytics work on sensitive data then again the information should remain protected while still allowing reports to be produced and retain context.

Enterprise information protection has to be holistic, covering all bases to avoid information risks that might breach legislation, cause non-compliance with regulations or adversely impact the organization's ability to meet its own business objectives. It involves being able to classify and locate
sensitive data, assess the vulnerability of the organisation to potential breaches in security, implement prevention measures to avoid putting data at risk, monitor events that may signal a problem and respond in a timely manner to minimise the impact of these events when they occur. Access control, sensitive data masking and network encryption are central to it.

In order to do this, companies need an enterprise information protection strategy to protect information as it flows throughout the enterprise. This enterprise information protection strategy needs to include:

- A set of objectives
- Metrics that show if you are on track to meeting the objectives
- Targets that need to be reached to protect specific data
- People accountable for reaching information protection targets
- Information security initiatives to ensure targets are reached
- People, processes and technologies to be used to protect data
- Security reports that need to be created

Companies then need to identify what the information risks are and what controls are in place to secure and protect information to reduce these risks. These controls may be in the form of access approval processes, data masking and encryption policies, auditing, backup policies, retention policies, and other checks and balances. If a violation occurs, then a damage limitation process is needed to manage losses and manage changes to procedures to avoid the same thing happening again. Also there need to be process in place to re-test security when changes are made so that risk exposure is not increased as a result of the changes.
WHERE ARE WE IN DATA AUDIT, ACCESS SECURITY AND PROTECTION TODAY?

Given the complexity of the information security challenge that most companies are facing, how far along are we in implementing enterprise security? A recent survey of 9600 responses in 138 countries asked the question “What process information security safeguards does your organisation have in place?”. The survey revealed the following:

<table>
<thead>
<tr>
<th>Overall information security strategy in place</th>
<th>62%</th>
</tr>
</thead>
<tbody>
<tr>
<td>An established security baseline for partners, customers and suppliers</td>
<td>39.9%</td>
</tr>
<tr>
<td>Centralized security information management process in place</td>
<td>43.3%</td>
</tr>
<tr>
<td>Standards/ procedures for infrastructure deployment established</td>
<td>43.4%</td>
</tr>
<tr>
<td>Identity management strategy</td>
<td>41%</td>
</tr>
<tr>
<td>Business continuity/ disaster recovery plans</td>
<td>39.1%</td>
</tr>
<tr>
<td>Standards and procedures in place for portable device security</td>
<td>40.5%</td>
</tr>
<tr>
<td>Authentication based on user risk classification</td>
<td>34.4%</td>
</tr>
<tr>
<td>Wireless security standards in place</td>
<td>42.8%</td>
</tr>
<tr>
<td>Employee security awareness training running?</td>
<td>43%</td>
</tr>
<tr>
<td>Cloud security strategy</td>
<td>25.6%</td>
</tr>
<tr>
<td>Mobile device security strategy in place</td>
<td>36.9%</td>
</tr>
<tr>
<td>Security strategy for employee use of personal devices</td>
<td>42.9%</td>
</tr>
</tbody>
</table>

What this survey clearly shows is that there is much more work to be done with 38% of the respondents having no overall information protection strategy in place at all. Also, with the exception of the 62% figure indicating that respondents do have an overall information security strategy in place, all the other percentage figures are below 50%. The weakest areas are cloud security, which is an area where application usage is growing in many organisations. The survey results also show that more thought is needed around user security to authenticate internal and external users based on risk. More also needs to be done around mobile device security, identity management and on establishing a security baseline for partners, customers and suppliers.

One area that stands out in this study is that less than half of the companies surveyed have a centralized security information management process in place. This indicates that the approach taken to implementing information security is distributed and may not be integrated in many organisations even though 62% of the 9600 respondents surveyed stated that they have an information security strategy in place. An integrated approach to enterprise information protection taking into account all the areas highlighted in Figure 1 is needed for comprehensive protection of sensitive information.

When it comes to implementation, the vast majority of information classified as sensitive in most organisations resides in ‘core platform’ databases. This includes sensitive data in:

- Core operational transaction processing databases and files
- Enterprise data warehouses and data marts including data warehouse appliances

Transaction processing systems are a classic place where sensitive data (e.g. customer financial information) resides. Many of these systems run on mainframes making them an important platform to include within the scope of an enterprise information protection program. But what is the link to big data? The answer lies in the relationship that mainframes and transaction processing have with the web.

A major strength of mainframe servers is their track record on very high availability and transaction processing. This is a key reason why they are often chosen as a platform for e-commerce, and customer self-service web computing. Mainframes also scale well in support of high transaction rates. For most organisations, the attraction of the web is twofold. First, the web offers global reach and secondly it reduces cost of operating by facilitating the introduction of self-service transaction processing. This has given rise to the need for 24x365 system availability. Also, the growth in mobile device usage and the switch to on-line commerce as a preferred way of transacting business has seen transaction volumes growing at a phenomenal rate.

The result is that increasing amounts of sensitive data are being captured in transaction system databases both on-premise (often on mainframes) and in the cloud. In addition a side effect of skyrocketing on-line activity is the growth in web log data, which has been nothing short of spectacular. Web logs are a popular type of big data. They are often analysed to derive additional customer insight that can be loaded into existing data warehouses to enrich what we already know. Given that web sites generating these web logs are often on mainframes, the mainframe has become a major source of both ‘big transaction’ structured data and multi-structured (web logs) big data. Both of these need to be secured.
Having defined what enterprise information protection is and looked at where organisations are in terms of tackling this challenge, the next question to ask is “What are the requirements for enterprise information protection?” These requirements need to cover the complete range of considerations shown Figure 1 to prevent security breaches. Much of this list of requirements has already been defined in an information governance white paper focussed on information audit and protection. These are shown in the Appendices at the back of this paper for convenience.

Rather than repeat those requirements here, it is worth adding additional requirements to that list to cover everything shown in Figure 1 in this paper that is not covered in the aforementioned paper. The reader is asked to combine the original list of requirements shown in the Appendices with the addition requirements for a complete set. These additional requirements are listed under the box headings shown in Figure 1 for convenience.

**INFORMATION CLASSIFICATION REQUIREMENTS**

- Structured information
  - It should be possible to define common data definitions for all master data (e.g. customer, product, asset, site, supplier etc.), reference data (code sets), transaction types, hierarchies and metrics in a business glossary and then classify what data is sensitive
  - It should be possible to define and attach policies to individual data item definitions and/or complete data entities to control data privacy and access security for master data, reference data, transaction data, relationship data (hierarchies) and metrics

- Unstructured data
  - It should be possible to define standard document and content types for the organization to describe what a document/image/rich media file is e.g. for documents the document types could include a supplier contract, a marketing brochure, a customer contract, a equipment maintenance manual etc.
  - It should be possible to define an enterprise taxonomy for the organization to describe what a document/image/rich media file is about e.g. a brochure is about an insurance

---

3 “Information Governance: Audit and Protection on the IBM System z Platform”, Ferguson, October 2011
product, a maintenance manual is about a specific make and model of asset (equipment)

- It should be possible to define and attach policies to individual document and content items to control privacy and access security associated with specific document/content types about a specific topic e.g. to secure access to contracts associated with a specific customer, financial reports associated with a business unit, or annual review documents associated with all employees.

### Data Landscape Requirements

*Some data streams may be considered sensitive*

- It should be possible to secure and protect data in motion even though it has not yet been stored

### User Requirements

*Need to manage users to govern what information they can access*

- It should be possible to centrally manage the identity of individual users or federate identity management so that a single view of all users accessing applications, information and software tools inside and outside the organization can be seen to prevent creation of duplicate users and so that authentication and authorization can be managed centrally.

### Device Requirements

- **Device security**
  - It should be possible to centrally manage mobile device security by creating configuration profiles containing device security policies, VPN configuration information, Wi-Fi settings, APN settings, email account settings, mail settings, and certificates that permit mobile smart phones and tablets to work with your enterprise systems.
  - It should be possible to enforce device authentication to secure access to a mobile device.
  - It should be possible to configure memory limits on mobile devices to limit the amount of information they are allowed to hold on the device.

- **Mobile application security**
  - It should be possible to enforce mobile application authentication via user ID and password so that users have to log in to applications that provide access to sensitive data.
  - It should be possible to enforce role-based access to application functionality from any device so that the user is only authorized to use specific application functionality.
  - It should be possible to allow access to specific applications from a mobile device for a set period after which access to those applications automatically expires.
• Data transmission security
  o It should be possible to encrypt sensitive data flowing over a public or private network to a desktop or mobile device or encrypt the data in the data store and decrypt before sending only if a user is authorized to see it

• Device Information security
  o It should be possible to clear mobile device caches of application specific information immediately an application closes or after a defined period
  o It should be possible for a user to control what subset(s) of information other users or user groups are allowed to see when sharing that information from a mobile device
  o It should be possible to enforce role-based access to application and information services from any device so that the user is only authorized to see specific information
  o It should be possible to allow access to specific information from a mobile device for a set period after which access to that information automatically expires

• Security monitoring and audit
  o It should be possible to log all mobile security violations and warnings for auditing and reporting purposes

CHANGE MANAGEMENT REQUIREMENTS

• It should be possible to define processes that require approval when changes are made to:
  o Information classifications
  o Security and privacy policies
  o Schema
  o Data replication and synchronization
  o Access and manipulation privileges
  o Application functionality

• It should be possible to run vulnerability testing before and after changes are made to ensure that security risk exposure has not been increased as a result of the changes made. It should also be possible to run vulnerability testing to look for database vulnerabilities when changes have not been made.

• It should be possible to co-ordinate controlled reversal of changes to information classification, schema, privileges and policies to previous versions of privileges if changes result in security breaches
AN INTRODUCTION TO BIG DATA

Now that we have looked at information security requirements, the next step is to look at big data and the impact this has on information security.

For many years, companies have been building data warehouses to analyse business activity and produce insights for decision makers to act on to improve business performance. These traditional analytical systems capture, clean, transform and integrate data from multiple operational systems before loading it into a data warehouse. However, even though this traditional environment continues to evolve, many new more complex types (varieties) of data have now emerged that businesses want to analyse to enrich what they already know. In addition, the rate (velocity) at which much of this new data is being created and/or generated and the volumes of data being analysed is far beyond what we have ever seen before.

TYPES OF BIG DATA

The most popular new types of data that organisations want to analyse include:

- **Web data** - e.g. web logs, e-commerce logs and social network interaction data
- **Industry specific big transaction data** - e.g., Telco call data records (CDRs), geo-location data and retail transaction data
- **Machine generated/sensor data** - to monitor everything from movement, temperature, light, vibration, location, airflow, liquid flow and pressure. RFIDs are another example.
- **Text** - e.g. from archived documents, external content sources or customer interaction data (including emails for sentiment analysis)

Customers and prospects are creating huge amounts of new web data in the form of social network interactions. A good example is Twitter data. In addition, on-line news items, weather data, competitor web site content, and even data marketplaces are now available as candidate data sources for business consumption.

Within the enterprise, web logs are growing as customers switch to on-line channels as their preferred way of transacting business and interacting with companies. That means that transaction data is on the increase. Large volumes of structured data should therefore also be considered a type of big data. Archived data warehouse data is also being resurrected for analysis and increasing amounts of sensor networks and machines that generate data are being deployed to instrument and optimise business operations. The result is an abundance of new data sources, rapidly increasing data volumes and a flurry of new data streams that all need to be analysed.
**WHAT IS BIG DATA?**

Big data can be broken into two areas:

- **Big Data Transaction Processing** (a.k.a. Big transactions)
- **Big Data Analytics**

Big Data transaction processing is about extreme volumes of transactions that may update data in relational DBMSs, NoSQL DBMSs or file systems. Typically, relational DBMSs are used as it is often the case that so-called ACID properties are found missing in many NoSQL DBMSs. This is only a problem if it is unacceptable to lose a transaction e.g. a Banking deposit.

Big Data Analytics is about advanced analytics on traditional structured and multi-structured data\(^4\). It is a term associated with the new types of workloads and underlying technologies needed to solve business problems that we could not previously support due to technology limitations, prohibitive cost or both.

Big data analytics is therefore *not* just about data volumes. It may be the case that data volumes are moderate but that data complexity (variety of data type) and analytical complexity are significant. Big Data analytics is about analytical workloads that are associated with some combination of data volume, data velocity (the rate at which data is generated) and data variety that may include complex analytics and complex data types. It can also be associated with *both* structured and multi-structured data.

**NEW BIG DATA ANALYTICAL WORKLOADS**

The emergence of new data sources and the need to analyse everything from live data streams in real time to huge amounts of unstructured content together with traditional structured content, has made many businesses realise that they are now in an era where the spectrum of analytical workloads is so broad that it cannot all be dealt with using a single enterprise data warehouse.

New big data analytical workloads have emerged that have taken us beyond the traditional data warehouse. These are:

1. Analysis of data in motion
2. Exploratory analysis of un-modelled multi-structured data
3. Accelerating ETL and analytical processing of un-modelled data to enrich data in a data warehouse or analytical appliance
4. Analysis of relationship in a graph
5. Complex analysis of structured data

---

\(^4\) Multi-structured data can be semi-structured like email or XML or unstructured data like text and video.
6. The storage and re-processing of archived data

TECHNOLOGY OPTIONS FOR END-TO-END BIG DATA ANALYTICS

To support these new analytical workloads, additional technologies have emerged beyond the traditional data warehouse RDBMS. These include:

- Stream processing software
- Analytical RDBMSs
- Hadoop solutions (could be on-premise or in the cloud)
- NoSQL DBMSs e.g. graph DBMSs

Stream processing software, is used to support the automatic analysis of data-in-motion in real-time or near real-time. Its purpose is to identify meaningful patterns in one or more event streams and trigger action to respond to them as quickly as possible. This software therefore provides the ability to build real-time analytic applications whose job it is to continuously keep different parts of a business operation optimized.

All of the other technologies mentioned support big data workloads that analyse data at rest where data is stored prior to analysis taking place.

Analytical requirements and data characteristics will dictate the technology deployed in a big data environment. However, the following table tries to match the each workload analysing big data at rest to the appropriate data storage platform.

<table>
<thead>
<tr>
<th>Big Data Analytical Workload</th>
<th>Big Data Storage Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploratory analysis of un-modelled multi-structured data e.g. web logs, unstructured content, filtered sensor data, email</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Complex analysis of structured data or for data warehouses that have ‘light’ mixed workloads</td>
<td>Analytic RDBMS Appliance</td>
</tr>
<tr>
<td>Storage and re-processing of archived data</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Accelerating ETL processing of structured and un-modelled data</td>
<td>Hybrid: Hadoop and Analytical DBMS</td>
</tr>
<tr>
<td>Social Graph Link analysis</td>
<td>NoSQL Graph DBMS</td>
</tr>
</tbody>
</table>

THE NEW ENTERPRISE ANALYTICAL ECOSYSTEM

Looking at these technologies, it is not difficult to conclude that dealing with big data involves the use of multiple underlying technology platforms each of which is optimised for specific big data analytical workloads. These platforms are in addition to the data warehouse. Big Data analytics can however include the traditional data warehouse environment because some

5 For more information on these technologies please refer to the paper “Architecting a Big Data Platform for Analytics”, Ferguson, September 2012
Enterpise Information Protection – The Impact of Big Data

analytical workloads may need both traditional and workload optimised platforms to solve a business problem.

Figure 2 shows the extended end-to-end analytical environment needed to support the big data analytical workloads discussed as well as traditional ad hoc query processing, analysis and reporting. Big Data does not replace a data warehouse. On the contrary, the data warehouse is an integral part of the extended analytical environment. Some refer to this new environment as the ‘enterprise analytical ecosystem’ or ‘logical data warehouse’.

It can be seen from this architecture that event stream processing of data-in-motion can be done on sensor data, or indeed any other event data source like financial markets for example. When variations in event data occur, event-processing software analyses the business impact and can take action if required. Filtered events can then be picked up by information management software and loaded into Hadoop for subsequent historical analysis. If any further insight is produced using batch map/reduce analytical processing, that insight may then be fed into a data warehouse. For un-modelled multi-structured data, this data can be loaded directly into Hadoop using information management software where data scientists can conduct exploratory batch analysis on this data in sandboxes. Alternatively search-based BI tools can be used to analyse the data using indexes built in Hadoop with map/reduce utilities. If data scientists produce any valuable insight, it can also be loaded into the data warehouse to enrich the structured data already there and so make this insight available to traditional BI tool users.

Complex analysis of structured data is undertaken on analytical DBMS appliances using in-database analytics. Again, if any insight is produced or any new predictive/statistical models created, then this can be moved into the data warehouse for use by information consumers in reports, dashboards and scorecards. Storage and re-processing of archived data can be managed in Hadoop with batch map/reduce applications or the aforementioned front-end tools used to analyse this data. In-Hadoop analytics (custom or pre-built) can be used as needed. Finally with respect to accelerating ETL processing on structured and un-modelled data, information management tools can be used to exploit Hadoop analytics and/or in-database analytics in analytical DBMS appliances (or both) for this purpose. Traditional data warehouse workloads continue as normal.
Given that this new extended analytical environment has a mix of traditional data warehouse and big data workload optimised systems, there is also a need to simplify access to these data stores for users with traditional front-end tools. This is achieved through data virtualisation software, which is designed to present data as if it is available in a single data store. Behind the scenes it uses pushdown optimization and other advanced optimization techniques to answer business queries. Also master data is available to feed consistent dimension data to all analytical environments.
THE IMPACT OF BIG DATA ON INFORMATION PROTECTION

Now that we have introduced big data and understood what it is, the next question is “What is the impact of big data on information protection?”

The impact of big data in the enterprise, is that introduces:

- New sources of information
- Data in motion as well as additional data at rest
- Multiple analytical data stores in a more complex analytical environment (with some of these data stores possibly being in the cloud)
- Big Data Platform specific storage e.g. Hadoop Distributed File System (HDFS), Analytical RDBMS Columnar Data Store, or a NoSQL Graph database
- New analytical workloads
- Sandboxes for data scientists to conduct exploratory analytics
- New tools and applications to access and analyse big data
  - More complex information management in a big data environment to
    - Supply data to multiple analytical data stores
    - Move data between big data analytical systems and Data Warehouses

All this impacts on data security. The data landscape is now more complex particularly because each big data analytical platform may have a different way to store data often with no standards in sight in this fast moving area of technology.

New sources data need to be secured and protected across heterogeneous big data platforms. This includes big transaction data and e-commerce logs in particular. Rich sets of structured and multi-structured data brought into a big data store for analysis could easily attract cyber criminals. An example might be data sources like customer data, location sensor data from smart phones, customer interaction data, on-line transaction data and web logs all being brought into Hadoop for analysis. Security around big data is therefore an issue.

Big Data adds data-in-motion and new file based analytical data stores to the data landscape thereby making it more complex to manage security. Figure 2 shows that master data can be supplied to Big Data platforms to analyse big data by various dimensions. This data may be supplied in various formats depending on the big data platform it is being loaded into. For example, data loaded into Hadoop, is likely to be supplied in files to distribute across a Hadoop cluster. Sensitive data may also be taken from traditional data warehouses into these environments to help with big data batch analytics. New analytical workloads mean new applications running on platforms like Hadoop and NoSQL databases as well as analysis of data-in-motion. Also new tools are accessing this data in one or more analytical data stores. Access control environments therefore needs to be
extended to manage access to big data analytical tools and applications as well as data held in new Big Data platforms.

In addition a new type of user has emerged - the data scientist. Data scientists are highly skilled power users who need a secure environment where they can explore un-modelled multi-structured data and/or conduct complex analyses on large amounts of structured data. Creating a project environment where small teams of data scientists can work and collaborate in ‘sandboxes’ on Hadoop and/or analytical RDBMS appliances is very much a part of big data analytics.

However sandbox creation and access needs to be controlled, as does data going into and coming out of these sandboxes.

**USING BIG DATA ANALYTICS FOR SECURITY ANALYSIS**

Finally, there is another side to this. Big Data analytics may well be able to help fight the security problem by being used to detect cyber crime. Analysing data in motion to identify fraud is one example of this. Also analysis of access activity to see what users access what data and what have they done to that data.
NEW SECURITY REQUIREMENTS TO PROTECT INFORMATION IN BIG DATA ENVIRONMENTS

Based on what have learned, the introduction big data into the enterprise demands that the requirements defined earlier for enterprise information protection, be extended. This is to move beyond structured data in existing transaction processing systems and data warehouses to also protect data in big data environments. Note that it is possible to define requirements focussed on information protection for big data environments and information protection from big data environments. The former is associated with protecting information in big data environments while the latter is about using real-time big data analytics to produce insights to help protect information in big data and traditional environments. The following requirements are associated with protecting information in big data environments and should be added to those already documented:

- It should be possible to classify data streams containing sensitive data associated with analyzing Big data in motion as ‘at risk’ and to know where in a data stream that sensitive data resides
- It should be possible to define or classify which files being loaded into a Big Data platform (e.g. Hadoop HDFS) contain sensitive data and to know where on a Big Data platform that data resides
- It should be possible to define and apply policies that encrypt sensitive structured data and multi-structured Big data in motion and sensitive structured data and multi-structured big data at rest
- It should be possible to define and apply policies that redact sensitive structured data and multi-structured big data in motion and sensitive structured data and multi-structured big data at rest in any big data analytical data store
- It should be possible to encrypt and redact structured sensitive data and multi-structured data when moving this data between big data and traditional data stores during analytical processing
- It should be possible to control access to all sensitive data files stored in file based big data analytical data stores
- It should be possible to monitor and log all administrative activity associated with sensitive data streams and sensitive data files in big data environments
- It should be possible to control who is allowed to create big data analytical sandboxes on top of big data analytical platforms e.g. Hadoop HDFS and/or analytical DBMSs
- It should be possible to control who is allowed to access data in big data analytical sandboxes built on top of Hadoop HDFS and/or analytical DBMSs
It should be possible to control what data sources can be loaded into big data analytical data stores and sandboxes for analysis and to log what sensitive data files are loaded and when.

It should be possible to encrypt and/or redact sensitive data brought into big data analytical data stores and sandboxes.

It should be possible to control which analytic applications have access to sensitive data streams and to report on what streams were accessed by what applications and when.

It should be possible to control which Map Reduce analytic applications have access to sensitive data files in Hadoop and other NoSQL data stores and to report on what files were accessed by what applications and when.

It should be possible to control which software tools have access to sensitive data in Hadoop and other NoSQL data stores and to report on what data was accessed by what tools and when.

It should be possible to control access to sensitive data in Hadoop whether that data is in HDFS, HBase or Hive structures.

It should be possible to define and classify structured data derived from unstructured data using text analytics in Hadoop as sensitive.

It should be possible to encrypt and redact structured sensitive data produced from text analytics in Hadoop to secure and protect it.

It should be possible for structured sensitive data produced from text analytics in Hadoop to be kept secure and protected if included in search indexes built on big data.

Search based tools accessing sensitive big data should not show this data to unauthorized users.

It should be possible to integrate software tools and big data analytic applications with Active Directory and/or LDAP based directories.

It should be possible to apply policies to sensitive data in Hadoop and have those policies enforced whether the data is stored in HDFS, HBase or Hive structures.

It should be possible to control access to external table functions in RDBMSs that access and manipulate big data in other big data data stores outside of the RDBMS itself.

It should be possible to report on which users and which applications accessed and/or manipulated sensitive data in a Big Data platform via external table functions on an RDBMS, and when.

It should be possible to encrypt and/or redact sensitive big data prior to providing that data to an application, RDBMS (via external table function), software tool or user not authorized to see it if that data is not already protected.
• It should be possible to control access to sensitive data in Data Warehouses and other data stores in an extended analytical ecosystem from MapReduce applications running on Hadoop.

• It should be possible to monitor sensitive data file counts and file sizes in file-based Big Data platforms.

• It should be possible to detect and block unauthorized access to sensitive data streams and sensitive data files in big data analytical environments.

• It should be possible to detect and block unauthorized access to sensitive data in Data Warehouses, MDM systems and other data stores from applications running on Hadoop and/or other Big Data platforms.

• It should be possible to secure, protect and audit all activity on sensitive big data irrespective of whether that data resides on-premise or in the cloud.

• It should be possible to secure, protect and audit all activity on sensitive big data irrespective of whether it is in test, development or production environments.

With respect to the use of analytics to help protect information the following requirements should be added:

• It should be possible to leverage Big Data platforms and analytics to collect, monitor and analyse security information to help solve advanced security and risk use cases.
IMPLEMENTING INFORMATION PROTECTION IN A BIG DATA ENTERPRISE USING IBM TECHNOLOGIES

Having defined what enterprise information protection is, looked at the requirements for this and also looked at big data and its impact, this section of the paper looks at how technologies from one vendor, IBM, can be used to help manage information security in a big data environment.

IBM provides both an enterprise platform for big data and also software to secure and protect big data platforms and data.

THE IBM BIG DATA PLATFORM

IBM provides a number of integrated technology components for end-to-end analytics on data in motion and data at rest. These components include:

- An stream processing engine for real-time analysis of data in motion
- A data warehouse platform supporting traditional analysis and reporting on structured data at rest
- An arrangement of analytical appliances optimised for specific advanced analytical workloads on big data
- An appliance for accelerating operational analytic query processing
- An integrated suite of self-service BI tools for ad hoc analysis and reporting including support for mobile BI
- Search based technology for building analytic applications offering free form exploratory analysis of multi-structured and structured data
- Predictive analytics for model development and decision management
- Applications and tools for content analytics
- Pre-built templates to quick start analytical processing of popular big data sources
- A suite of integrated information management tools to govern and manage data in this new extended analytical environment

Together, this set of technologies constitutes the IBM Big Data Platform as shown below. This platform includes three analytical engines to support the broad spectrum of traditional and big data analytical workloads. These are:

---

6 For more information on IBM Big Data Platform please refer to the paper “Architecting a Big Data Platform for Analytics”, Ferguson, September 2012
• IBM InfoSphere Streams for continuous analytics of data-in-motion
• IBM BigInsights - A Hadoop System
• Data Warehouse (could be one or more data stores)

In addition to IBM InfoSphere Streams, the IBM Big Data Platform also includes:

**Hadoop - IBM InfoSphere BigInsights**

IBM InfoSphere BigInsights is IBM’s commercial distribution of the Apache Hadoop system. It has been designed for exploratory analysis of large volumes of multi-structured data. IBM InfoSphere BigInsights ships with standard Apache Hadoop software. However, IBM has strengthened this by adding a number of features to make it more robust including a Posix compliant file system and storage security.

**IBM InfoSphere BigInsights on IBM System zEnterprise**

With respect to IBM System z, IBM has announced a version of InfoSphere BigInsights Enterprise Edition that will run within the zEnterprise on the zEnterprise BladeCenter Extension(zBX) frame. This means that Hadoop can run on virtualized HX5 Linux blades using virtual disk.

**IBM PureData System for Analytics (powered by Netezza technology)**

IBM PureData System for Analytics powered by Netezza technology is the next generation Netezza Appliance optimised for advanced analytical workloads on structured data.

**IBM InfoSphere Warehouse, Smart Analytics System and IBM PureData System for Operational Analytics**

The IBM PureData System for Operational Analytics is based on IBM Power System. The IBM Smart Analytics System is a modular, pre-
integrated real-time Enterprise Data Warehouse optimized for operational analytic data workloads available on IBM System x, IBM Power System or IBM System z servers. Both the IBM PureData System for Operational Analytics and the IBM Smart Analytics System family include IBM InfoSphere Warehouse 10 software running on DB2 Enterprise Server Edition 10. DB2 10 includes a new NoSQL Graph store. Also automated optimized data placement leveraging Solid State Disk (SSD) is included inside the new PureData System solution.

**IBM DB2 Analytic Accelerator (IDAA)**

IBM DB2 Analytics Accelerator is a IBM Netezza 1000™ Appliance specifically designed to offload complex analytical queries from operational transaction processing systems running DB2 mixed workloads on IBM System z. It can also be used with IBM DB2 for z/OS based data warehouses to accelerate complex query processing.

**IBM Big Data Platform Accelerators**

IBM has built over 100 sample applications, user defined toolkits, standard toolkits, industry accelerators and analytic accelerators to expedite and simplify development on the IBM Big Data Platform.

**IBM Information Management for the Big Data Enterprise**

InfoSphere Information Server, InfoSphere Foundation Tools, InfoSphere Optim and InfoSphere Guardium provide an integrated suite of tools for governing and managing data. It includes tools for defining, modelling, profiling, cleaning, integrating, virtualising, protecting and moving data across all traditional and big data analytical platforms. It supports connectors to IBM BigInsights, The IBM PureData System for Operational Analytics, IBM PureData System for Analytics powered by Netezza technology, IBM DB2 Analytics Accelerator as well as IBM InfoSphere Master Data Management. IBM InfoSphere Information Server also integrates with IBM InfoSphere Streams to pump filtered event data into IBM InfoSphere BigInsights for further analysis.

**IBM PRODUCTS FOR ENTERPRISE INFORMATION PROTECTION**

The following table shows IBM’s products for Enterprise Information Protection

<table>
<thead>
<tr>
<th>IBM Information Management Products on System z Required to Implement Information Protection</th>
<th>Usage in Protecting Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM RACF</td>
<td>Managing role based access to data and services</td>
</tr>
<tr>
<td>IBM DB2 10 DBMS for Z/OS</td>
<td>Data encryption for DB2 data</td>
</tr>
<tr>
<td>IBM InfoSphere Guardium Data Encryption for IMS and DB2 Databases</td>
<td>Data encryption for DB2 and IMS data using System z hardware</td>
</tr>
<tr>
<td>IBM InfoSphere Business</td>
<td>Manage common data definitions for</td>
</tr>
<tr>
<td>Glossary for Linux on System z</td>
<td>master data and transaction data and allow business users to highlight information to be protected</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>IBM InfoSphere Discovery</td>
<td>Discovery of data and data relationships to identify the location of the data to be protected</td>
</tr>
<tr>
<td>IBM InfoSphere Optim Data Privacy solution for z/OS</td>
<td>Managing Data Privacy, information retention and archive processing</td>
</tr>
<tr>
<td>IBM InfoSphere Guardium Data Redaction</td>
<td>Protects sensitive unstructured data contained in documents &amp; forms from unintentional disclosure</td>
</tr>
<tr>
<td>IBM InfoSphere Guardium Data Activity Monitor and Vulnerability Assessment for System z</td>
<td>Real-time database activity monitoring Monitor privileged users e.g. DBAs Monitor enterprise application users for fraud Enforce database change control Prevent database leaks</td>
</tr>
<tr>
<td>IBM Security zSecure Audit and Admin (and their integration with Qradar)</td>
<td>Enables you to detect and report security events and exposures on z/OS, DB2, CICS, IMS, UNIX, and Linux on System z on systems with RACF, CA ACF2, or Top Secret.</td>
</tr>
<tr>
<td>IBM Qradar</td>
<td>A suite of products for collecting, storing, analyzing and querying log, threat, vulnerability and risk related data</td>
</tr>
<tr>
<td>IBM Security Key Lifecycle Manager</td>
<td>Centralizes and automates the encryption key management process</td>
</tr>
</tbody>
</table>

In addition to these software solutions the System z10 EC mainframe provides cryptographic hardware features available on consist of the following:

- CP (central processor) assist for Cryptographic Functions (CPACF)
- Configurable Crypto Express2 (CEX2)
- Trusted Key Entry (TKE) workstation
- Smart Card Reader

**IBM SYSTEM z IN A IBM BIG DATA ENVIRONMENT**

Many companies have built data warehouses on IBM System z close to the transaction data sources also stored there. Given the growth in transaction data on IBM System z both in DB2 z/OS and IMS, many organisations have chosen to build data warehouses on IBM System z close to that transaction data in order to make it possible to integrate business intelligence and analytics into operational business processes. In context right-time operational BI is a growing trend. In addition many organisations run the IBM InfoSphere Warehouse on IBM System z or IBM’s Smart Analytics System on zEnterprise, both of which can run coupled with an IBM DB2 Analytics Accelerator for z/OS. It is also the case that IBM System z transaction data stores are popular sources feeding into other RDBMS and big data analytical platforms. That may
include other data warehouse platforms or IBM BigInsights where transaction data may be archived and analysed.

Big Data platforms are also a place where analysis takes place on multi-structured data. Data Scientists produce insights from that data and can then move those insights into data warehouses to enrich what an organisation already knows. For those organisations running data warehouses on IBM System z, DB2 can integrate with IBM’s InfoSphere BigInsights Hadoop Platform. The first phase of this integration is the availability of connectors between IBM DB2 on z/OS and IBM InfoSphere BigInsights. JAQL server running on InfoSphere BigInsights can accept JAQL query processing requests from IBM DB2 z/OS via new JAQLSubmit and HDFSRead user defined functions. This means that it is possible to embed a JAQL query (JAQL script) in DB2 for z/OS application SQL statements that execute map reduce applications on Hadoop to analyse data held in Hadoop HDFS.

It is also a direction of IBM to build DB2 Map Reduce functions that can be invoked in DB2 z/OS SQL and offloaded onto IBM DB2 Analytic Accelerator. This would allow a new breed of DB2 z/OS analytic applications to be built to apply advanced analytics to information stored in its native form in IBM InfoSphere BigInsights.

Given the increased level of integration between Big Data platforms and data warehouses, it is also possible to move data from IBM InfoSphere BigInsights Hadoop into data warehouses. With respect to data warehouses developed on DB2 z/OS, map reduce applications running in Hadoop can move data into DB2 z/OS partitioned tables. In order to do this, the map reduce applications running on Hadoop would need access to the DB2 Catalog information to understand DB2 data format and also to match parallel reduce threads to partitions in DB2 for z/OS.

**DELIVERING SECURITY FOR BIG DATA IN A SYSTEM z ENVIRONMENT**

IBM System z transaction data is an obvious target for people and/or applications trying to gain unauthorised access to sensitive data in core transaction databases and so IBM’s InfoSphere Guardium offers full audit and protection for data in DB2 z/OS databases, IMS databases and VSAM files. Also when moving data from/to IBM System z to/from other platforms for analysis (including data warehouses on System z), it is possible to mask and redact structured sensitive data. This could happen as part of an ETL job or even a MapReduce application for example.

Compliance mandates that information security and protection policies should be equally enforced across big data environments as well as traditional data management architectures. To extend the reach of information security to Big Data platforms means at least being able to answer the following questions:

- Who is issuing specific big data requests?
- What map-reduce jobs are they running?
- Are they trying to download sensitive data?
- Are there any files in Big Data platforms containing sensitive data
• Are there an unusually large number of file permission exceptions, perhaps caused by a hacker algorithmically trying to get access to sensitive data?

• Are these jobs part of an authorized program list accessing the data? Or has some new application been developed that you were previously unaware existed?

To deal with these requirements, IBM has extend edit security technologies to support big data and new big data analytical environments. In particular, security is supported within IBM InfoSphere Streams, IBM BigInsights, IBM InfoSphere Warehouse and IBM Smart Analytics System on System zEnterprise. Several products including IBM InfoSphere Guardium have been extended and IBM has combined QRadar with BigInsights to analyze security event data to produce security intelligence. Protecting sensitive data as it moves between System z and IBM Big Data analytical platforms is also supported. These are covered in more detail below.

Securing Information on IBM InfoSphere Streams

IBM InfoSphere Streams maintains its own security configuration for each Streams instance. It also provides security templates that you can specify with each Streams instance. Therefore, by dedicating specific instances and their associated security template to different data streams, it becomes possible to isolate access to data streams containing sensitive data. With respect to user authentication, IBM InfoSphere Streams can integrate with one of two supported authentication services. These are Pluggable Authentication Module (PAM the default), or an external Lightweight Directory Access Protocol (LDAP) server. InfoSphere Streams uses access control lists (ACLs) to enforce security and so when a Streams instance is created, the authorization policy template specifies the initial ACLs for that instance. Administrators can then adjust this going forward. The ACLs control access to InfoSphere Streams commands, as several require user authentication to execute them. Finally, the security events for any instance of InfoSphere Streams can be optional captured in an audit log file to record activity on that instance.

Securing Information on IBM System z InfoSphere Warehouse and IBM DB2 Analytics Accelerator

Many data warehouses running on IBM System z are built using the IBM InfoSphere Warehouse, which is based on IBM DB2 z/OS. Several System z security components integrate to manage security in this environment including:

• Databases security on DB2 10 for z/OS
• Databases security on IMS data sources
• Access security via RACF
• InfoSphere Guardium Database Activity Monitoring and Vulnerability Assessment
• InfoSphere Guardium Data Encryption for DB2 and IMS databases

InfoSphere Guardium uses a variety of methods to collect events in order to gather the necessary information to protect IBM System z DB2, IMS and
VSAM data stores. Each InfoSphere Guardium S-TAP collects event information on database activity to build up a comprehensive picture on access to and manipulation of sensitive data either in IBM System z transactional data sources or in IBM InfoSphere Warehouse. In addition, it is possible to mask sensitive data using InfoSphere Optim Data Masking.

**Securing and Redacting Information on IBM BigInsights and IBM System z**

With respect to Hadoop activity in IBM BigInsights, IBM InfoSphere Guardium has now been extended to audit Hadoop activity to monitor events and exceptions.

The events that can be monitored include the following:

- Session and user information
- HDFS operations commands (e.g., cat, tail, chmod, chown, expunge etc.)
- MapReduce jobs - including job, operations and permissions
- Exceptions – e.g., authorization failures
- Hive/HBase queries

To make event monitoring possible, IBM InfoSphere BigInsights Hadoop includes a ‘Guardium Proxy’ which is installed InfoSphere BigInsights Hadoop NameNode. The proxy allows BigInsights to send Apache log data as messages from Hadoop logs to the InfoSphere Guardian collector for analysis and reporting. However because Hadoop does not log exceptions, there is no way for the proxy to send exceptions to InfoSphere Guardian. To overcome this limitation and make it possible for exceptions to also be reported, IBM InfoSphere Guardium also provides an S-TAP for Hadoop. S-TAPs can be installed on the HBase master, the Hadoop JobTracker, the Hadoop NameNode and Hive Server.

IBM InfoSphere Guardian supports monitoring of

- IBM InfoSphere BigInsights 1.4 or later
- Cloudera CDH3 and CDH4
- Hive (requires the use of MySQL as the Beeswax database)

It also ships with several out-of-the-box reports for Hadoop, including reports on:

- MapReduce activity for both BigInsights and Cloudera
- Unauthorized MapReduce jobs
- Hue/Beeswax activity (assumes use of the Thrift message format and the MySQL database)
- HDFS, HBase, and Hive activity
- Exception report

Some of the reports are parameterized to allow you to define run time parameters to control overtime periods between which you want to look at activity. Rules can also be adjusted in the Hadoop security policy to control precisely what InfoSphere Guardian monitors. In addition, authorised users and sensitive objects in Hadoop can both be defined to InfoSphere Guardian in separate groups so that InfoSphere Guardian knows what is sensitive and who is authorized to access it. Any unauthorized users accessing sensitive data can therefore be tracked.
With respect to HBase, IBM InfoSphere Guardium will show the actual commands that flow to HBase e.g., Alter, count, create, drop, get, put, list, etc. to detail specific HBase activity.

Securing and Redacting Information During Smart Consolidation

As we have seen from Figure 2, the impact of big data is that new analytical data stores have emerged over and above traditional data warehouses to handle new more complex analytical workloads. This includes IBM’s DB2 for z/OS, which has been enhanced to support graph analytics with the inclusion of a new graph data store. This more complex analytical environment has put more pressure on data management technologies, which now have a much greater role to play. Data has to be supplied into and moved between data stores during analytical processing at an ever-increasing rate. This includes moving data between IBM System z and other platforms for analysis. That makes it more challenging to protect sensitive data. To overcome this IBM has added APIs to IBM InfoSphere Optim to make it possible to mask and redact structured sensitive data on-demand as it is loaded into or moved between systems. On-demand masking and redaction could therefore be invoked as part of an InfoSphere DataStage ETL job or in a MapReduce application providing data into DB2 z/OS. This helps to keep data protected in a big data environment.

DELIVERING SECURITY FROM BIG DATA IN A SYSTEM z ENVIRONMENT

In a world where data is becoming more distributed, it is getting harder to secure and protect sensitive data. Now that big data has arrived, there is even more to manage. To help overcome this IBM have sought to exploit big data to manage security. IBM Security Intelligence with Big Data combines the real-time security visibility of the IBM QRadar Security Intelligence Platform with custom analytics on the IBM Big Data Platform.

IBM QRadar performs real-time correlation, anomaly detection and reporting for immediate threat detection, and also sends enriched security data to IBM InfoSphere BigInsights for analysis along with vast amounts of data from unstructured and semi-structured sources. In a world where sophisticated cyber criminals are now at work, many organizations now need to accommodate combining event data with the variety and volume of data needed to solve advanced security and risk use cases. IBM Security Intelligence with Big Data is IBMs answer to this requirement. Integration between IBM QRadar and IBM BigInsights is two-way in that security intelligence produced from big data analytics can be fed back to QRadar, providing a facility for closed-loop, continuous learning.

The result is a solution that collects, monitors, analyzes, explores and reports on security and enterprise data in ways previously not possible. Key capabilities of this solution include:

- Real-time correlation and anomaly detection
- High-speed querying of security intelligence data
- Flexible big data analytics across structured and unstructured data – including security data; email, document and social media content; full packet capture data; business process data; and other data
- Graphical front-end tool for visualizing and exploring big data
- Forensics for deep visibility
CONCLUSION

Big Data has had a major impact on enterprise information protection. The arrival of big data into the enterprise has had a major impact enterprise information protection. New data stores have emerged increasing the distribution of data and the complexity of securing and protecting that data along with it. What we have seen is that big data has forced companies to have to define new requirements when it comes to both data management and enterprise information protection. It is now harder to protect sensitive data as it may move around between different transactional and analytical data stores as companies create new analytical workloads.

Figure 1 shows how much of a daunting challenge Enterprise information protection is. Yet despite the complexity, IBM covers most of this with the products that it offers. It has recognised that transaction data is where a significant amount of sensitive data resides and as a result has not only protected key databases on IBM System z but has also introduced technologies that can monitor, audit and protect that data as it moves across data bases and data stores in a distributed heterogeneous environment. The use of technologies like QRadar, InfoSphere Guardium and InfoSphere Optim along with database security, RACF and even hardware encryption on IBM System z go a long way towards closing off all the avenues potentially open to those attempting to gain unauthorised access to sensitive data. The extension of InfoSphere Guardium to support Hadoop, securing access to streaming data, on-demand data masking and redaction during data movement and the use of IBM InfoSphere BigInsights to produce security intelligence are all weapons in the technology arsenal needed to guarantee enterprise information protection. While there is more to do to wrestle big data to the ground, for those introducing big data into a distributed computing environment where information security is high on the agenda, IBM is certain to be shortlisted as a key supplier of software to tackle enterprise information protection.

IBM security software has been extended to embrace and protect information in a big data environment

More work has still to be done to get complete control of big data
APPENDICES - KEY REQUIREMENTS FOR PROTECTING DATA AND PREVENTING SECURITY BREACHES

DATA LANDSCAPE REQUIREMENTS

• It should be possible for enterprise information audit and protection software to simultaneously and continually protect and secure data held in multiple heterogeneous relational and non-relational DBMSs throughout the enterprise.

• It should be possible for enterprise information audit and protection software to simultaneously and continually monitor and audit database activity across multiple heterogeneous relational and non-relational DBMSs throughout the enterprise.

• It should be possible for enterprise information audit and protection software to simultaneously and continually monitor activity for data held in multiple file systems throughout the enterprise.

• It should be possible for enterprise information audit and protection software to simultaneously and continually monitor activity for data held in specific locations throughout the enterprise.

• It should be possible for enterprise information audit and protection software to protect and secure access to sensitive data down to the individual record and field levels in multiple heterogeneous relational and non-relational DBMSs throughout the enterprise.

SOFTWARE ACCESS REQUIREMENTS

• It should be possible to register instances of software applications and tools on different systems and in different locations throughout the enterprise as authorized or non-authorized so as to be capable of detecting access to sensitive data from non-authorised application and tool instances.

• It should be possible for enterprise information audit and protection software to be able to identify, monitor and audit all activity from all instances of applications and tools used to...
access and manipulate sensitive data across multiple heterogeneous relational and non-relational DBMSs and file systems throughout the enterprise

**ENVIRONMENT REQUIREMENTS**

- It should be possible for enterprise information audit and protection software to protect and secure data in all:
  - Development,
  - Test
  - Production

  environments across multiple heterogeneous relational and non-relational DBMSs and file systems throughout the enterprise. Since many customers still don’t have information audit and protection controls in production, the production environment is the priority.

- It should be possible for enterprise information audit and protection software to monitor and audit database and file based activity in all
  - Development
  - Test
  - Production

  environments across multiple heterogeneous relational and non-relational DBMSs and file systems throughout the enterprise

**USAGE REQUIREMENTS**

- It should be possible for enterprise information audit and protection software to be able to distinguish between:
  - External users
  - Internal business end users
  - Privileged users e.g. database administrators
  - IT developers
  - IT operations personnel

- It should be possible to monitor and audit privileged user behavior across heterogeneous relational and non-relational DBMSs and file systems throughout the enterprise

- It should be possible for enterprise information audit and protection software to integrate with corporate user directories such as LDAP directories or Microsoft Active Directory to automatically discover active users within the enterprise and to automatically monitor new user creation
• It should be possible for enterprise information audit and protection software to integrate with relational and non-relational DBMSs to automatically monitor new user creation.

• It should be possible for enterprise information audit and protection software to automatically discover privileged users declared in relational and non-relational DBMSs and operating systems throughout the enterprise.

• It should be possible for enterprise information audit and protection software to automatically discover user privileges and who granted these privileges and when.

**Vulnerability Assessment Requirements**

Vulnerability assessment requirements are associated with the ability to gauge the likelihood of data risk exposure such as unauthorised access to sensitive data and to specifically pinpoint risks that need to be addressed. The requirements are as follows:

• It should be possible to mark data items defined in a business glossary as ‘data at risk’ so as to create a ‘data at risk register’ within a business glossary that is visible to authorized business users to make people aware of sensitive data.

• It should be possible for enterprise information audit and protection software to be capable of automatically discovering sensitive data in heterogeneous relational and non-relational DBMSs and files throughout the enterprise to determine where sensitive data is located and what policies to apply to protect and secure access to it. Automatically discovered sensitive data attributes should be mapped to common definitions in the business glossary so that it becomes possible to use the glossary to highlight sensitive data items in heterogeneous data stores right across the enterprise that qualify for vulnerability assessment testing.

• It should be possible to define protection policies and enforcement mechanisms for specific business glossary data items marked as ‘at risk’ and have these policies enforced enterprise wide.

• Enterprise information audit and protection software should provide a pre-built set of vulnerability tests available to test if sensitive data is exposed to unauthorized access or if it is not masked.

• Enterprise information audit and protection software should provide a pre-built set of vulnerability tests available to test exposures cause by ‘loose’ privileges allocation in DBMSs e.g. GRANT…WITH GRANT OPTION or GRANT…TO PUBLIC.
• Enterprise information audit and protection software should provide a pre-built set of vulnerability tests to test for non-compliance of specific line item requirements defined within regulations and/or legislation

• It should be possible to extend pre-built vulnerability tests and to create additional custom built vulnerability tests that can be run to assess vulnerability of data to other risks

PREVENTION REQUIREMENTS

These requirements define the capabilities enterprise information audit and protection software need to provide to minimise exposure to data risks across DBMSs and file systems in the enterprise. The requirements are as follows:

• Enterprise information audit and protection software should provide a set of pre-defined
  o Policies
  o Roles
  o Tests
  o Templates
to speed up implementation and enforcement of information protection across the enterprise and to comply with security and privacy regulations

• It should also be possible to define custom policies to protect and secure access to sensitive data residing in one or more heterogeneous DBMSs and file systems across the enterprise

• It should be possible to define custom policies to mask and encrypt sensitive data on one or more instances of a database, a file, or structure in one or more heterogeneous DBMSs and files systems across the enterprise

• It should be possible to group policies in any way to make it easier to administer the protection and security of at risk data. For example:
  o Common policies that can be enforced enterprise wide
  o System specific policies
  o Policies associated with a specific master data entity e.g. customer or employee data
  o Policies associated with a specific transaction and its data
  o Policies associated with specific business intelligence
  o Policies associated with a specific database or file and its data

Out-of-the-box pre-built templates help organisations get started quickly
Policy based access to sensitive data is paramount

- It should be possible to define policies that control access to sensitive data. These policies should be capable of being applied at multiple different levels including:
  - Access to specific data in all systems across the enterprise
  - Access to specific data across all instances of a database
  - Access to specific data in specific database or file instances
  - Access to specific data in a specific database or file structure
  - Access to specific data by specific users or user groups
  - Access to specific data by applications and software tools
  - Access to specific data at a specific location
  - Access to specific data only at specific times e.g. only within working hours
- It should be possible to define policies that govern the masking and encryption of data in one or more heterogeneous relational and non-relational DBMSs and files throughout the enterprise
- It should be possible to define policies that restrict the ability to change the schema of any database in any DBMS instance or instances
- It should be possible to define policies that prevent manipulation of data by unauthorized transactions, software tools and users
- It should be possible to restrict the ability to change information protection policies to only authorized users
- It should be possible to restrict the powers of privileged users across one or more databases, files, DBMSs, file systems and locations
- It should be possible to separate the duties of privileged users from approvers so that privileged user activity can be formally controlled across development, test and production environments
• It should be possible to lock down production databases to prevent privileged users from creating changes to information protection policies, and schemas and also to unlock a database and once unlocked to monitor it to make sure all changes by privileged users audited so that they can be reported

• It should be possible to define processes that prevent privileged users from escalating their own privileges or the privileges of others

• It should be possible to flag policies as needing to be enforced in real-time or on a scheduled basis

• It should be possible to define expiry dates for information protection policies and to implement policy version control

• It should be possible to define information audit and protection ‘agents’ that can be deployed to monitor specific data access behavior and database activities in real-time. This includes monitoring privileged user activity, end user access to sensitive data, access to sensitive data from application transactions and tools, file open and close, etc. These software information audit and protection agents should be capable of being deployed in specific heterogeneous relational and non-relational DBMSs and file systems across all defined locations and platforms in the enterprise

• It should be possible to define policy-based actions or action sequences that can be invoked by information audit and protection software agents in real-time to uphold protection policies and neutralize threats when unauthorized or suspicious behavior occurs

• It should be possible to monitor all units of work associated with sensitive data in a database

• It should be possible to monitor access to database image copies containing sensitive data

• It should be possible to monitor changes to database metadata

• It should be possible for information protection agents to generate and emit audit entries that can be sent back to a centralized auditing function for the enterprise that automatically records all audited activity in a tamper-proof repository

• It should be possible to define policies to prevent circumvention of a DBMS by monitoring and denying access to underlying files used by a DBMS to store sensitive data

**ENFORCEMENT REQUIREMENTS**

• It should be possible to audit all of the following in real-time

Approval workflows are needed to control privileged user behaviour

Policies need to be enforced in real-time on a continuous basis

A centralized auditing function is needed
- Schema changes including what the changes were, who made the changes, when they were made, who approved them and when
- Policy changes including what the changes were, who made the changes, when they were made, who approved them and when
- Access and changes to sensitive data by any user, application object, transaction unit of work, query plan or software tool
- Access to image copies
- All SQL statements accessing sensitive data
- Time of access to sensitive data (e.g. outside working hours)
- Outbound transactions from databases e.g. caused by triggers
- Privilege escalations
- Login failures
- New user IDs created
- Sharing of user IDs

- It should be possible to detect, audit and block the following in real-time
  - Access to sensitive data by an unauthorized end user
  - Access to sensitive data by an unauthorized application object, transaction or query plan
  - Unauthorized database activity e.g. privilege escalation
  - Unauthorized policy changes
  - Unauthorized schema changes
  - Unauthorized opens and closes of a file
  - Unauthorized outbound transactions coming from a database
  - Access to sensitive data by an unauthorized software tool e.g. TOAD
  - User account creation by an unauthorized user, application or tool

- It should be possible to notify nominated user(s) about security exceptions, login failures and privilege escalations in real-time via a user-defined alerting mechanism (e.g. email, SMS, dashboard alert, etc.) and escalate the alert if necessary if receipt of alert is not acknowledged

- It should be possible for enterprise information protection software to provide pre-built out-of-the-box reports and dashboards for monitoring and auditing purposes on unauthorized access, suspicious behaviour, privilege
escalations, login-failures, schema changes, policy changes etc.

- It should be possible for enterprise information audit and protection software to provide pre-built out-of-the-box reports on regulatory compliance violations
- It should be possible to create custom built reports from the audit repository to satisfy specific questions

**PERFORMANCE REQUIREMENTS**

> A solution needs to scale

- It should be possible for enterprise information audit and protection software to scale across all databases and file systems in the enterprise
- It should be possible for enterprise information audit and protection software to impose minimal overhead on application databases and the daily running of operational and analytical workloads by using agent software to emit data back to a centralised audit repository
About Intelligent Business Strategies

Intelligent Business Strategies is a research and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, analytical processing, data management and enterprise business integration. Together, these technologies help an organisation become an intelligent business.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an analyst and consultant he specialises in business intelligence and enterprise business integration. With over 31 years of IT experience, Mike has consulted for dozens of companies on business intelligence strategy, big data, data governance, master data management, enterprise architecture, and SOA. He has spoken at events all over the world and written numerous articles. He has written many articles, and blogs providing insights on the industry. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates, an independent analyst organisation. He teaches popular master classes in Big Data Analytics, New Technologies for Business Intelligence and Data Warehousing, Enterprise Data Governance, Master Data Management, and Enterprise Business Integration.

Water Lane, Wilmslow
Cheshire, SK9 5BG
England
Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-mail: info@intelligentbusiness.biz

Enterprise Information Protection – The Impact of Big Data
Copyright © 2013 by Intelligent Business Strategies
All rights reserved