BIG DATA ANALYTICS—WHAT IT MEANS TO THE AUDIT COMMUNITY – PART 2

Markus Hardy, PMP, CISSP, CISA
Agenda

• **Big Data**
  - What is it / Types / Stats / New forms and growth
  - Value it might bring
  - The four “V’s”
  - Concerns

• **Analytics, Business Intelligence and Big Data Analytics**
  - What are they and how are they different
  - Stacks and the different types
  - Hadoop Ecosystem

• **Auditing Big Data**
  - Planning, Scope, Considerations
  - Operational Considerations
Auditing Big Data – Scope Planning

Scope of the Big Data related Audit –

- **Big Data Business Case.** What did the business present and get approved for as the business case, proven ROI, projected ROI, defined payback period, etc?

- **Big Data Initiative Metrics.** How does the company intend to measure success of the Big Data Initiative? (Quantitative, Qualitative, Business Performance, IT Performance, ?

- **Big Data PMO and SDLC Methods.** What methods are being used to manage Big Data initiatives – traditional SDLC? Agile? Same PMO standards?

- **Big Data Retention.** What is being done with the results of Big Data Analysis post exploration
Auditing Big Data – Scope Planning

Scope of the Big Data related Audit –

• **Big Data Ecosystem.** What Big Data technology framework is being used and what decisions are being made by using this Ecosystem

  - Computational – Hadoop and Map Reduce
  - Data Flow/Movement – Apache Pig
  - Data Access – Apache Hive
  - Stream Filtering and Pattern Recognition
  - Platform – Distributed Job Management – Hadoop Job Tracker
  - Distributed File Storage (HDFS)
  - Stream Capture (Apache S4 or IBM InfoSphere Streams)
  - Horizontally Scalable No SQL DB’s (Apache Hbase)
  - In Memory DB’s (eXtremeDB, Oracle Times Ten)

• **Big Data Techniques.** Is the company using basic or advanced Hadoop/Big Data techniques – Friend of Friends (FoF), Page Ranking, Bloom Filters, Crunch and Cascading
The Hadoop Ecosystem

June 21, 2012

Who has the most partners? Who is connected?

brought to you by Datameer

“Datameer Corp Infographic”
**Auditing – Classic Scope of moving data to functions**

- Pre-Data Warehouse
- Data Cleansing
- Data Repositories
- Front-End Analytics

**Data Flow**

- OLTP Server
- Meta-Data Repository
- ETL
- Data Warehouse
- Data Mart
- ODS

**Analytics**

- OLAP
- Data Mining
- Data Visualization
- Reporting

CONFIDENTIAL - NOT FOR DISTRIBUTION
Auditing Big Data – Scope Planning

Scope of the Big Data related Audit –

- **Big Data GRC** What Big Data security and compliance tools are being used and what do you plan to accomplish?
  
  Source Data Ingestion – How, how much, when, why
  Encryption – At rest, in motion, at end point
  Open Source Security Knowledge – Auditor’s base?
  Documented Policies- Do you have any

- **Big Data Skills** What are the critical skills sets and what risk do they pose to operations?
  
  Specialized skills – Parallel computing; statistical analysis, Predictive Analysis (R Hadoop, Rhipe, R, Mahout), Data Scientists
  Open Source / Pure Java / Java “Like”, JVM, SSL/TLS / Java 1.6 update 8
  Web HDFS, XML, JSON, Hadoop 0.23 REST API, GitHub, Sqoop, Flume
  High Level Languages – Crunch, Cascading, Pig, Hive
  Organizational Inventory – what do you have already that can be converted
Auditing – Present day scope of moving functions to data
Auditing – Classic Scope of moving data to functions
Auditing Big Data – Operational Planning

**Client Requests.**
Who is running specific big data requests, when, how and what functions are they looking to perform?

**Jobs and Functions.**
Are they running map-reduce jobs or functions, why, and what is being done with the results?

**Data Classification.**
Are they trying to download all of the sensitive data, or is this a normal query to gain insight into your customers?

**Access Management.**
Is there an exceptional number of file permission exceptions, perhaps caused by a hacker algorithmically trying to get access to sensitive data?
Auditing – Hybrid scope

1. Data Warehouse
2. Virtual Data Mart
3. Dept Data Mart
4. Analytic Appliance
5. Hadoop

Query:
- Data Warehouse
- Virtual Data Mart
- Dept Data Mart
- Analytic Appliance
- Hadoop
Auditing Big Data – Operational Planning

**Job Authority.**
Are these jobs part of an authorized program to list accessing the data?

**Application Authority.**
Has some new application been developed that you were previously unaware existed?

What application packages are being used in conjunction with accessing Big Data and why? What rights and permissions are being used?

- **R** for Statistical Analysis
- **Lucerne** for text search
- **Avro** for data serialization
- **Zookeeper** for process coordination
- **Oozie** for workflow and job orchestration)
Auditing Big Iron – Fabric, Spines and Leafs

10GE Hadoop Cluster – 2560 Nodes

- 1RU Nodes 2 x 10GE (C1100)
- 40 Nodes per rack
- 64 racks, **2560 nodes**
- 2.5:1 oversubscribed Leaf

- Scale to 16 x Z9000 Spine switches (10G)
- Cost effective QSFP+ SR optics
- QSFP optical breakout cables (150m)
- Each Z9000 configured as 128 x 10G
Big Data Infrastructure - Auditing Big “Iron”

Traditional Three Tier Network Fabric Architecture

Two Tier Leaf – Spine Network

Blocked Links

Core

Aggregation

Access

Spine

Leaf

Network Fabric
Big Data Infrastructure - Auditing Big “Iron”

Network Fabric

Multi-Domain – Silo’d

Single Domain – Any App, Any where!

Silo 1

Silo 2

Silo 3

Web Servers

App Servers

New Apps
Auditing Big Data – Operational Planning

Data Replication and Recovery. What do you know about replication and recovery of data across Clusters and Nodes?

Cluster Management tools. How are they being used for recovery?

Named Nodes. How are they structured or recovered?

File System Management. Is it HDFS or other file systems (IBM’s GPFS)

Data Block Recovery Processes. How big are the blocks and how are they being recovered?

In Memory and/or No SQL Databases. How are they being used? What is being stored in Cache Memory and why? What decisions are being made using this approach? (Orace Times Ten, extremeDB, etc)
Auditing Big Data – Operational Planning

Disks, Stores and Schemes. What do you know about how disks, stores and scheme’s are performing or MTBF rates and how stores are being organized and recovered across clusters and nodes?

Disk Management. How are disks or nodes being recovered?

Meta Data Stores. How are they organized, backed up and being recovered? THIS IS CRITICAL

Key Value Stores. How are they organized and being recovered?

NameNode, Primary Master Nodes, Secondary Master Nodes, Worker Nodes. How are they being backed up and restored? NameNode and MasterNodes are CRITICAL

Compression Scheme’s. What are the Codec’s being used and what are their impacts (Deflate, qzip, bzip2, LZO, LZOP, Snappy)
Auditing Big Data – Operational Planning

Big Data Ecosystem Performance.
What do you know about performance related problems, their root causes and what alternative actions can be implemented to provide solutions?

Large Input Records. Evaluate who is trying to ingest what that’s causing bottlenecks?

Resource Contention. Evaluate who is running what jobs causing resources to be overwhelmed?

Race Conditions. Who keeps writing/ingesting the same data over and over again at multiple locations?

No process for handling bad data. What is being done to validate data quality (CRC’s?)

Hadoop versions. Do you have different versions of Hadoop running on different Clusters?
Auditing Big Data – Operational Planning

Security Frameworks and Management

**Open Ports on Nodes in Cluster.** Do you know which ports are being opened and why?

**Surface Area vs Open Ports.** What is being done to minimize surface area?

**Alternative Authentication Protocols.** What is being used for authentication and Account Management (LDAP, Active Directory, Kerberos, MIT Kerberos?)

**Big Data/Hadoop/HDFS Hardening.** What is being done to lock down the Big Data ecosystem?

**Mechanisms to protect data.** What is being done to protect Big Data at rest and in motion?
## Auditing Big Data – Tools and Techniques

Table 1. Hadoop service ports to monitor  (SEE SLIDE 49)

<table>
<thead>
<tr>
<th>Service</th>
<th>Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS Name Node</td>
<td>8020, 50470, and 50070</td>
</tr>
<tr>
<td>HDFS Thrift plugin for Hue (NameNode)</td>
<td>10090</td>
</tr>
<tr>
<td>MapReduce Job Tracker</td>
<td>8021, 9290, and 50030</td>
</tr>
<tr>
<td>HBase Master</td>
<td>600000 and 60010</td>
</tr>
<tr>
<td>HBase Region</td>
<td>60020</td>
</tr>
<tr>
<td>HBase Thrift plugin</td>
<td>9090</td>
</tr>
<tr>
<td>Hive Server</td>
<td>10000</td>
</tr>
<tr>
<td>Beedwax Server</td>
<td>8002</td>
</tr>
<tr>
<td>Cloudera Manager Agent</td>
<td>9001</td>
</tr>
</tbody>
</table>
Auditing Big Data – IBM’s Systematic Approach
Auditing Big Data – IBM’s Approach
Two tools that are Native to Hadoop Ecosystem

(1) **MRUnit** is a Apache MRUnit ™ is a Java library that helps developers unit test Apache Hadoop map reduce jobs

(2) **JobTracker UI**
SCAP is a method for using specific standards to enable automated vulnerability management, measurement, and policy compliance evaluation (e.g., FISMA compliance). The National Vulnerability Database (NVD) is the U.S. government content repository for SCAP.

SCAP, pronounced “ess-cap”, combines a number of open standards that are used to enumerate software flaws and configuration issues related to security.

They measure systems to find vulnerabilities and offer methods to score those findings in order to evaluate the possible impact. It is a method for using those open standards for automated vulnerability management, measurement, and policy compliance evaluation.

SCAP defines how various standards (referred to as SCAP 'Components') are combined.
SCAP Components

• Common Vulnerabilities and Exposures (CVE)
• Common Configuration Enumeration (CCE)
• Common Platform Enumeration (CPE)
• Common Vulnerability Scoring System (CVSS)
• Extensible Configuration Checklist Description Format (XCCDF)
• Open Vulnerability and Assessment Language (OVAL)

Starting with SCAP version 1.1

• Open Checklist Interactive Language (OCIL) Version 2.0

Starting with SCAP version 1.2

• Asset Identification
• Asset Reporting Format (ARF)
• Common Configuration Scoring System (CCSS)
• Trust Model for Security Automation Data (TMSAD)
SCAP Checklists

- They standardize and enable automation of the linkage between computer security configurations and the NIST Special Publication 800-53 (SP 800-53) controls framework.

- Is meant to perform initial measurement and continuous monitoring of security settings and corresponding SP 800-53 controls.

- Future versions will likely standardize and enable automation for implementing and changing security settings of corresponding SP 800-53 controls.
SCAP Checklists

• In this way, SCAP contributes to the implementation, assessment, and monitoring steps of the NIST Risk Management Framework.

• Accordingly, SCAP is an integral part of the NIST FISMA implementation project.
Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing remote procedure calls.

Cassandra is a distributed and Open Source database. Designed to handle large amounts of distributed data across commodity servers while providing a highly available service. It is a NoSQL solution that was initially developed by Facebook. It is structured in the form of key-value.

Chukwa is a Hadoop subproject devoted to large-scale log collection and analysis. Chukwa is built on top of the Hadoop distributed filesystem (HDFS) and MapReduce framework and inherits Hadoop’s scalability and robustness.

Chukwa also includes a flexible and powerful toolkit for displaying monitoring and analyzing results, in order to make the best use of this collected data.

Crunch
Cascading  Provides a higher level of abstraction for Hadoop, allowing developers to create complex jobs quickly, easily, and in several different languages that run in the JVM, including Ruby, Scala, and more. In effect, this has shattered the skills barrier, enabling Twitter to use Hadoop more broadly.

Comparators  Two ways you may compare your keys is by implementing the `org.apache.hadoop.io.WritableComparable` interface or by implementing the `RawComparator` interface. In the former approach, you will compare (deserialized) objects, but in the latter approach, you will compare the keys using their corresponding raw bytes.

Flume  is a framework for populating Hadoop with data. Agents are populated throughout ones IT infrastructure – inside web servers, application servers and mobile devices, for example – to collect data and integrate it into Hadoop.
**Hama** is a distributed computing framework based on Bulk Synchronous Parallel computing techniques for massive scientific computations e.g., matrix, graph and network algorithms. It's a Top Level Project under the Apache Software Foundation.

**HBase** is a non-relational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily.

**HCatalog** is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored.

**HDFS** (Hadoop Distributed File System) the storage layer of Hadoop, is a distributed, scalable, Java-based file system adept at storing large volumes of unstructured
**Impala** (By Cloudera) provides fast, interactive SQL queries directly on your Apache Hadoop data stored in HDFS or HBase using the same metadata, SQL syntax (Hive SQL), ODBC driver and user interface (Hue Beeswax) as Apache Hive. This provides a familiar and unified platform for batch-oriented or real-time queries.

**MapReduce** is a software framework that serves as the compute layer of Hadoop. MapReduce jobs are divided into two (obviously named) parts. The “Map” function divides a query into multiple parts and processes data at the node level. The “Reduce” function aggregates the results of the “Map” function to determine the “answer” to the query.

**Hive** is a Hadoop-based data warehousing-like framework originally developed by Facebook. It allows users to write queries in a SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, etc.
Hue (Hadoop User Experience) is an open source web-based interface for making it easier to use Apache Hadoop. It features a file browser for HDFS, an Oozie Application for creating workflows and coordinators, a job designer/browser for MapReduce, a Hive and Impala UI, a Shell, a collection of Hadoop API and more.

Kafka (developed by LinkedIn) is a distributed publish-subscribe messaging system that offers a solution capable of handling all data flow activity and processing these data on a consumer website. This type of data (page views, searches, and other user actions) are a key ingredient in the current social web.

Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing and statistical modeling and implements them using the Map Reduce model.

MongoDB is a NoSQL database oriented to documents, developed under the open source concept. It saves data structures in JSON documents with a dynamic scheme (called MongoDB BSON format), making the integration of the data in certain applications more easily and quickly.
Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages – such as Map Reduce, Pig and Hive -- then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

Pentaho offers a suite of open source Business Intelligence (BI) products called Pentaho Business Analytics providing data integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities.

Pig Latin is a Hadoop-based language developed by Yahoo. It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
**Sqoop** is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop. It allows users to specify the target location inside of Hadoop and instruct Sqoop to move data from Oracle, Teradata or other relational databases to the target.

**Storm** is a system of real-time distributed computing, open source and free, born into Twitter. Storm makes it easy to reliably process unstructured data flows in the field of real-time processing, which made Hadoop for batch processing.

**Thrift** "Thrift is a software framework for scalable cross-language services development. It combines a software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml."

**WebHDFS** Apache Hadoop provides native libraries for accessing HDFS. However, users prefer to use HDFS remotely over the heavy client side native libraries. For example, some applications need to load data in and out of the cluster, or to externally interact with the HDFS data. WebHDFS addresses these issues by providing a fully functional HTTP REST API to access HDFS.
**Key Value Stores** – Key value stores allow the application to store **its data in a schema-less way**. The data could be stored in a datatype of a programming language or an object. Because of this, there is **no need for a fixed data model**.

**NoSQL (commonly interpreted as "not only SQL")** is a broad class of database management systems identified by non-adherence to the widely used relational database management system model. NoSQL databases **are not built primarily on tables**, and generally **do not use SQL for data manipulation**.

**Semi-Structured Data.** is a form of structured data that **does not conform with the formal structure** of data models associated with relational databases or other forms of data tables (XML, EDI)

**Structured Data.** Data derived from current day SQL oriented Relational Data Base Management Systems (RDBMS) such as Oracle, Microsoft, Sybase, etc
ZooKeeper is a software project of the Apache Software Foundation, a service that provides centralized configuration and open code name registration for large distributed systems. ZooKeeper is a subproject of Hadoop.
Unstructured Data. MetaData, Texts (Email’s), Analog, Health Records, Web Logs, Photo’s, Tweets, Video’s, FaceBook Messages (Java Server Oriented Notation (JSON))
“As data becomes a more valuable asset for most organizations, data architectures will need to change between old and new databases and systems to unlock the value.”

“It’s not just the rising volume of data that will challenge organizations, but rather developing new data architectures for effectively handling both structured and unstructured information.”

http://www.youtube.com/watch?v=qd1aZb4ORmw