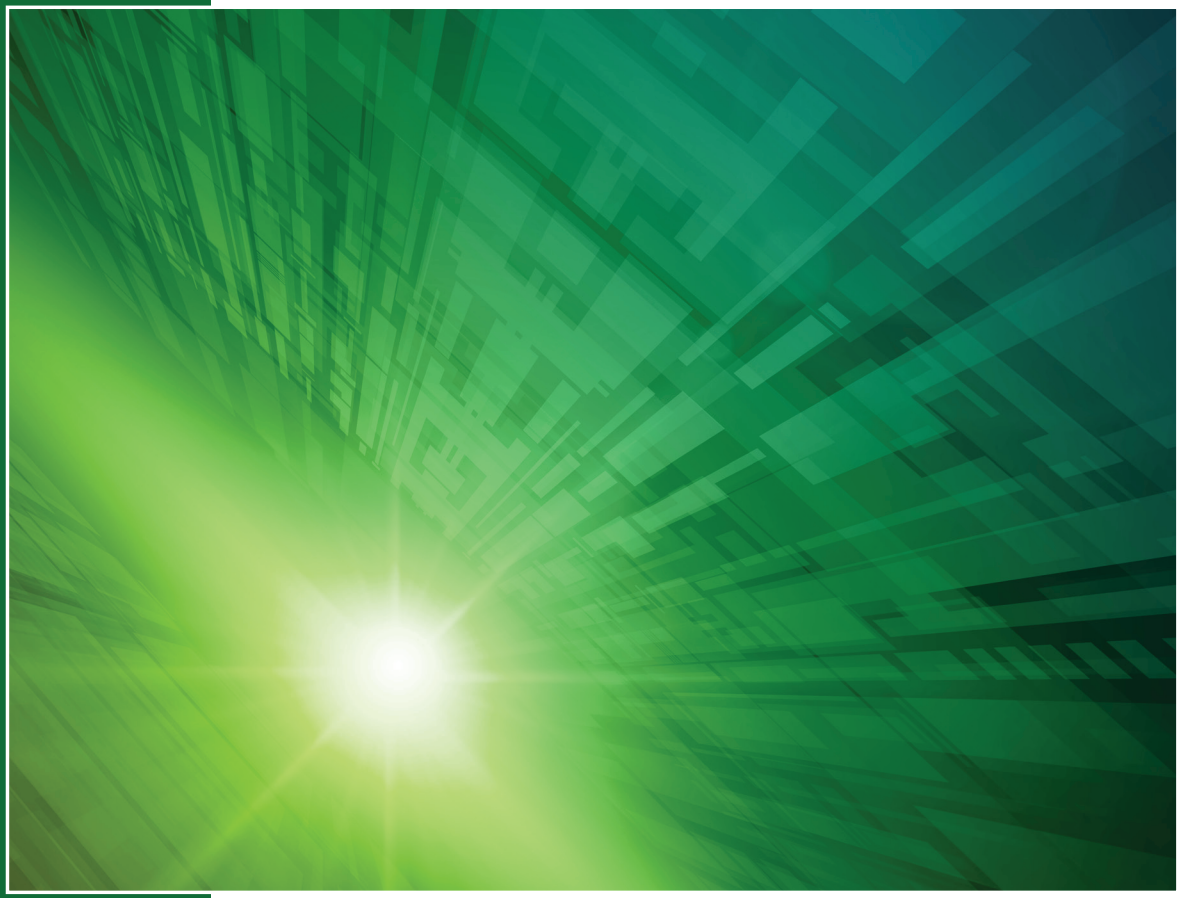
INFORMATION ASSET, LLC

Big Data Governance Tools:

Insights from Information Asset's Enterprise Data Management Lab

by Sunil Soares
June 6, 2014



Introduction

Big Data Governance is part of a broader information governance program that formulates policy relating to Big Data. Big Data is often referred to in the context of the "three Vs:"

- **Volume (data at rest)**—Big Data is generally large.
- **Velocity (data in motion)**—Often time-sensitive, streaming data must be analyzed with millisecond response times to bolster real-time decisions.
- **Variety (data in many formats)**—Big Data includes structured, semi-structured, and unstructured data such as email, audio, video, clickstreams, log files, and biometrics.

Organizations need to establish policies to ensure the maximum success of their Big Data programs. In this research report, we will examine critical activities to kick-start a Big Data Governance program with a specific focus on software tools for Hadoop.

1. Define the Business Problem

Leading organizations turn to technologies such as Hadoop, NoSQL, and stream computing to solve business problems that cannot be addressed with traditional tools. The first step is to define a high-level business problem that can be addressed by Big Data. We use a three-dimensional framework for Big Data Governance as shown in Figure 1:

- *Big Data types*—Big Data governance needs a heightened focus on the data itself. We have classified Big Data into five distinct types: web and social media, machine-to-machine, big transaction data, biometrics, and human generated.
- *Information governance disciplines*—The traditional disciplines of information governance also apply to Big Data. These disciplines are organization, metadata, privacy, data quality, business process integration, master data integration, and information lifecycle management. In this research report, we examine how tools can support the execution of these disciplines for Big Data.
- *Industries and functions*—Big Data analytics are driven by uses cases that are specific to a given industry or function.

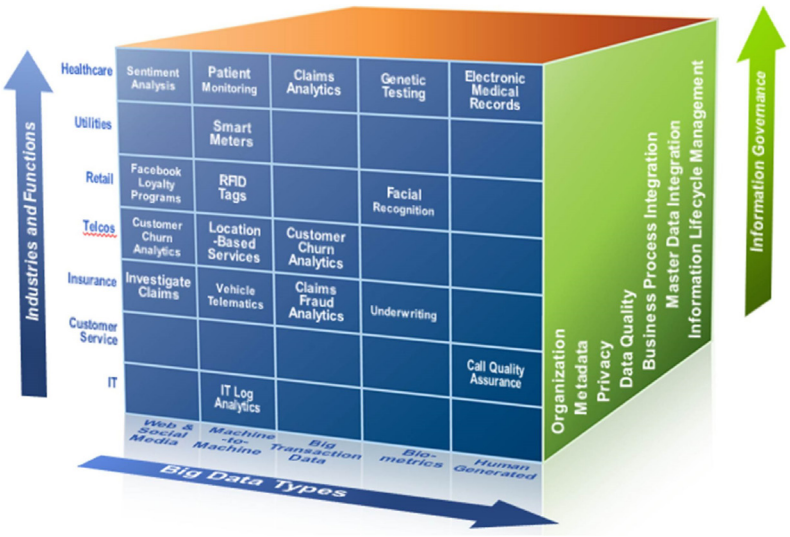


Figure 1: Big Data Governance framework.

Here are a couple of examples that leverage this framework:

Example 1

Solution: Smart meters in utilities
Big Data type: Machine-to-machine
Disciplines: Privacy, information lifecycle management

Several utilities are rolling out smart meters to measure the consumption of water, gas, and electricity at regular intervals of an hour or less. These smart meters generate copious amounts of "interval" data that needs to be governed appropriately. Utilities need to safeguard the privacy of this interval data because it can potentially point to a subscriber's household activities, as well as the comings and goings from his or her home. In addition, utilities need to establish policies for the archival and deletion of interval data to reduce storage costs.

Example 2

Solution: Sentiment analysis
Big Data type: Web and social media
Disciplines: Master data integration, data quality, privacy

The marketing department might want to use Twitter feeds for "sentiment analysis," to determine what users are saying about the company and its products or services. The analytics team needs to determine if, for example, references to "@Acme" and "Acme" both refer to "Acme Corporation." Integrating sentiment analysis with a customer's profile can also be challenging. In addition to privacy issues, the Twitter handle might reveal the user name in only 50 to 60 percent of cases. Finally, marketing might need to answer the following question: "Do we really believe that Twitter sentiment analysis is representative if users are younger and more affluent than our typical customers?"

2. Conduct an Inventory of Data in Hadoop

Many Big Data teams start by creating a "data lake" in Hadoop to act as a landing zone for data across the organization and for Big Data analytics. Left unchecked, these data lakes can quickly become unmanageable for data scientists who might have to deal with multiple copies of the same data set. Because you cannot govern what you do not know exists, the first step to governing Big Data is to build an inventory.

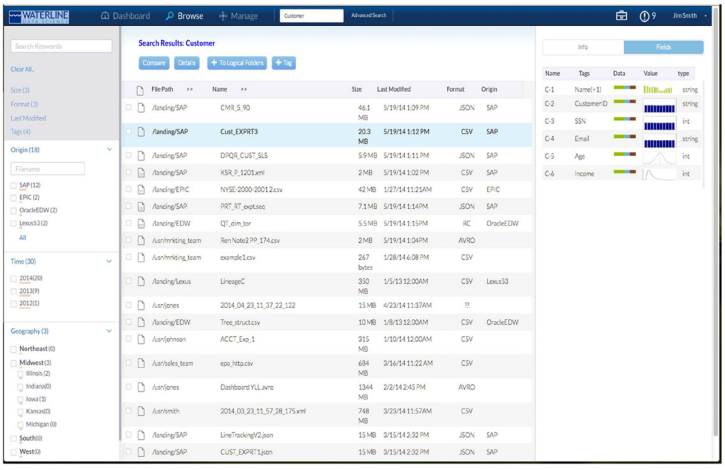


Figure 2: Hadoop Data Inventory file browser in Waterline Data Science.

In Figure 2, we show the Waterline Data Inventory for Hadoop by Waterline Data Science. The data scientist views the search results for Customer. In the left panel, the data scientist can also view the faceted search with results classified by categories. For example, based on the lineage discovered by Waterline, there are 12 files originating from SAP and three files that contain Midwestern states. On the right, the user can see profiling results and the data distribution for the fields in the selected file. In the example, the file is in CSV format without a header row, and Waterline Data Inventory used a combination of user tagging and automated data discovery to identify the meanings of the fields and to assign appropriate tags.

3. Assign Ownership for Data in Hadoop

Data governance leads need to assign the ownership of data whether or not it resides in Hadoop.

As shown in Figure 3, a retailer uses Data Advantage Group MetaCenter to manage data ownership for all types of data. For example, the Finance Reporting application and Data Mart 1 are owned by Finance and Merchandising, respectively. On the other hand, Big Data in the form of Twitter and Facebook are primarily managed in Hadoop and are under the ownership of Marketing. Finally, Supply Chain owns radio-frequency identification (RFID) data. MetaCenter also contains the names of the Data Executive, Managing Data Steward, and Data Steward for each application or data type. The Marketing department is accountable to establish and enforce policies relating to the acceptable use of Twitter and Facebook data.

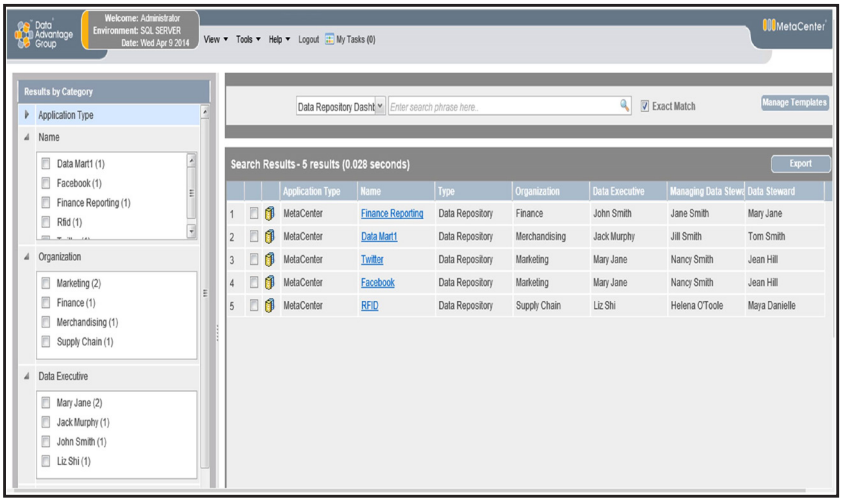


Figure 3: Data ownership in Data Advantage Group MetaCenter.

4. Provision a Semantic Layer for Analytics in Hadoop

Although Hadoop is a powerful analytics platform, it is meaningless without the appropriate semantic layer to provide meaning to key business terms. In the next few screenshots, we explore a simple MapReduce word count program that executes within Cloudera. The business definitions for the key terms are governed in Collibra Data Governance Center.

	A	B
1	Name	definition
2	affiliate	Section 2 of the Bank Holding Company ActFor purposes of this Act the term "affiliate" means any company that controls, is controlled by, or is under common control with another company (see definition
3	agency	Section 1 of the International Banking Act"Agency" means any office or any place of business of a foreign bank located in any State of the United States at which credit balances are maintained incidental to
4	bank	Section 3 of the Federal Deposit Insurance Act.The term "bank"-- (A) means any national bank and State bank, and any Federal branch and insured branch; and (B) includes any former savings association.
5	bank hold	Section 2 of the Bank Holding Company Act(a)(1) Except as provided in paragraph (5) of this subsection, "bank holding company" means any company which has control over any bank or over any company
6	branch	Section 1 of the International Banking Act "branch" means any office or any place of business of a foreign bank located in any State of the United States at which deposits are received.
7	commerci	Section 602 of Dodd FrankA company is a "commercial firm" if the annual gross revenues derived by the company and all of its affiliates from activities that are financial in nature (as defined in section 4(k)
8	exposure	Section 23 of the Federal Reserve ActAn insured depository institution's "exposure" to another depository institution means--A.all extensions of credit to the other depository institution, regardless of nat

Figure 4: Banking-related business terms relating to the Dodd-Frank legislation in the United States.

Figure 4 shows a list of banking terms related to the Dodd-Frank legislation in the United States.

In Figure 5, we execute the MapReduce code on the .csv file that contains the Dodd-Frank terms.

In Figure 6, we view the results of the MapReduce word count program in Cloudera. The words affiliate and bank appear once each.

Finally, we can view the business definitions of these terms in Collibra Data Governance Center as shown in Figure 7. In this way, Collibra provides the semantic layer for Hadoop analytics in Cloudera.

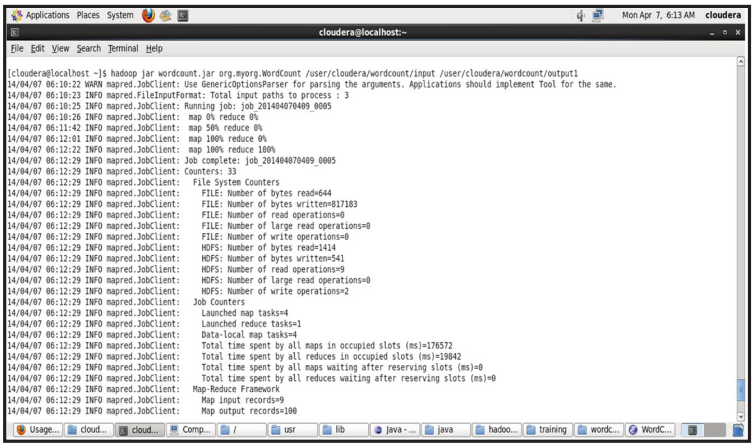


Figure 5: Executing MapReduce code on .csv file with Dodd-Frank business terms.

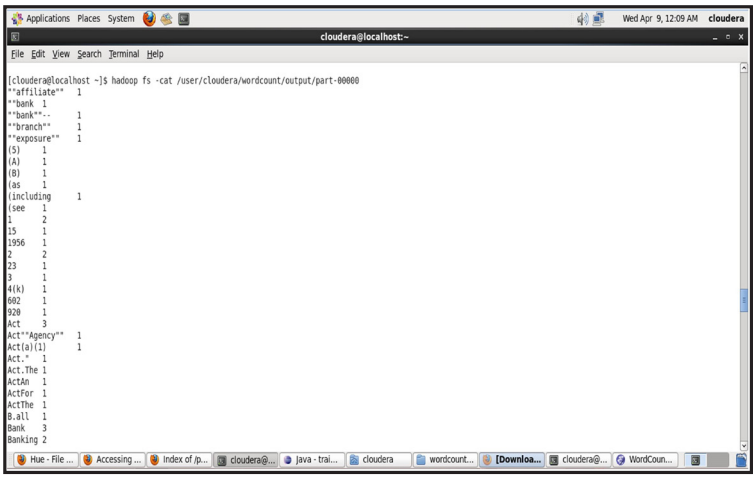


Figure 6: Viewing the results of the MapReduce word count program in Cloudera.

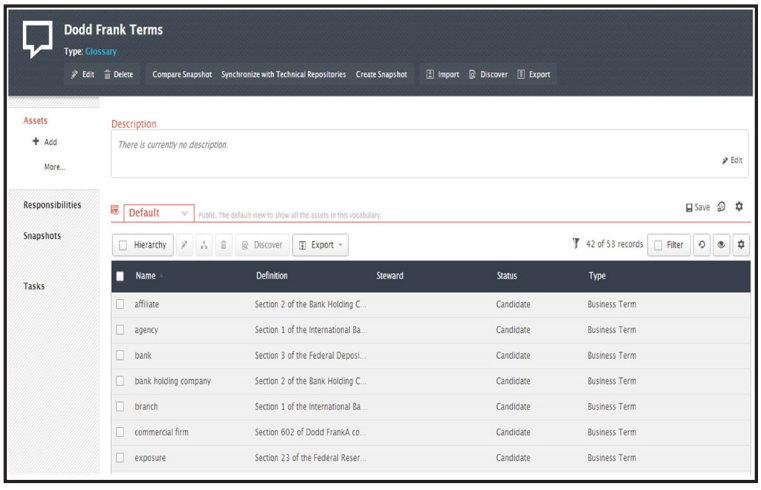


Figure 7: Viewing definitions of business terms in Collibra Data Governance Center.

5. View the Lineage of Data In and Out of Hadoop

As Hadoop becomes mainstream, organizations will use the platform for mission-critical applications. This means that data governance teams will need to include Hadoop within their data lineage views.

In Figure 8, ASG-Rochade shows the detailed forward lineage of the **EMP_EXPENSE_FACT** table. The lineage report shows the **emp_expense_fact** Hive table as well as two sets of Sqoop ETL jobs.

Figure 9 shows how a user can use Data Advantage Group MetaCenter to browse the employee table in Hive.

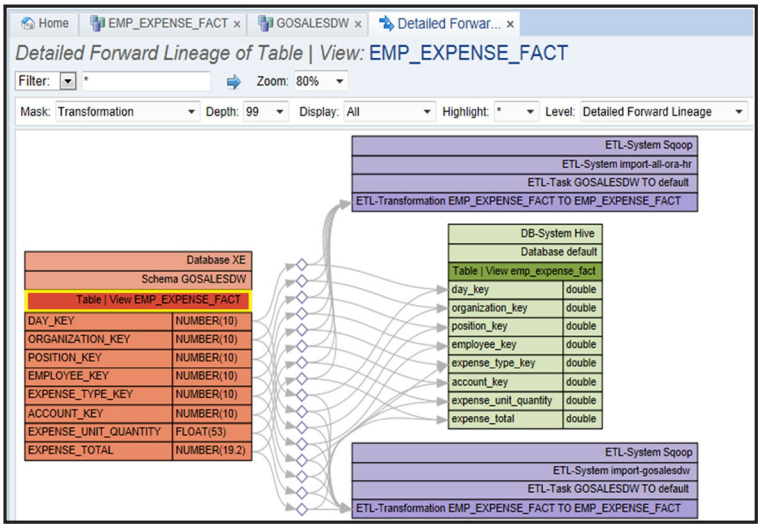


Figure 8: Detailed Forward Lineage of EMP_EXPENSE_FACT table in ASG-Rochade.

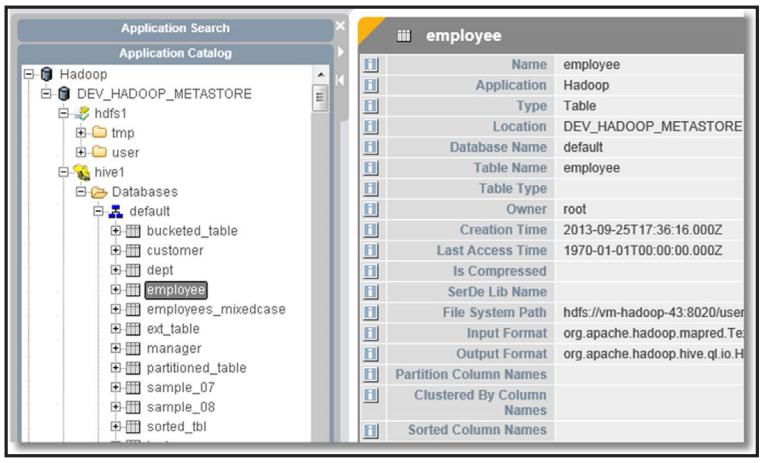


Figure 9: Browsing a Hive table in Data Advantage Group MetaCenter.

6. Manage Reference Data for Hadoop

Hadoop implementations also need high-quality reference data, as shown in the remainder of this section.

In Figure 10, the **GDP_Test.csv** file contains the Gross Domestic Product (GDP) in millions of U.S. dollars by country. This data is in unsorted format and will be used as the input file for a Pig script.

In Figure 11, we have created a Pig script to sort the countries in descending order of GDP. The script runs in Cloudera and takes the **GDP_Test.csv** file as an input file. The script places the result in a file called **output**.

Figure 12 shows the results of the Pig script in sorted format.

In Figure 13, we can see the reference data for country codes in Data Advantage Group MetaCenter. In the left panel, we can view a list of country codes. In the right panel, we can see that **FR** represents **France**.

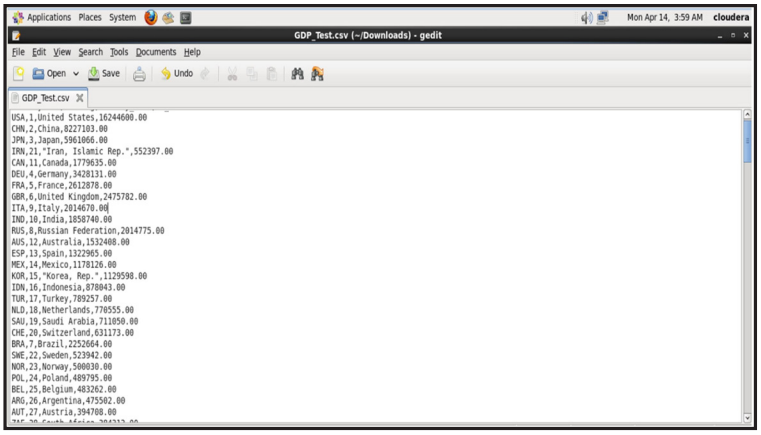


Figure 10: GDP_Test.csv file contains Gross Domestic Product by country.

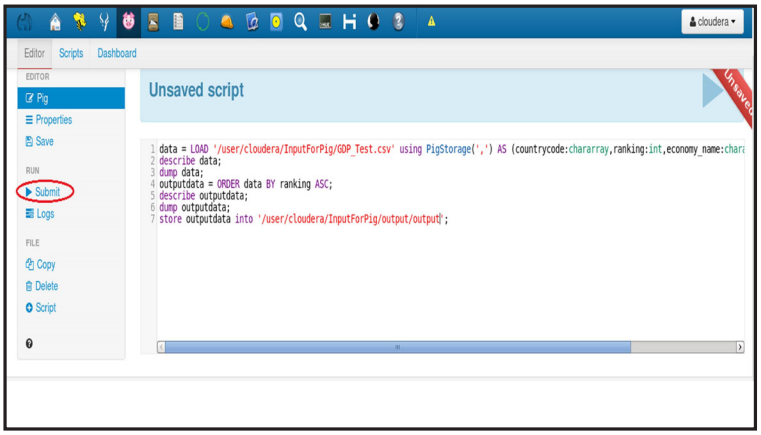


Figure 11: Pig script to sort countries in descending order of GDP.

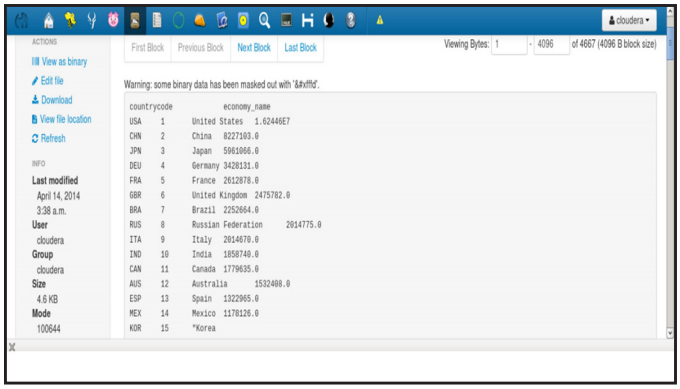


Figure 12: Output of Pig script shows a sorted list of countries in descending order of GDP.

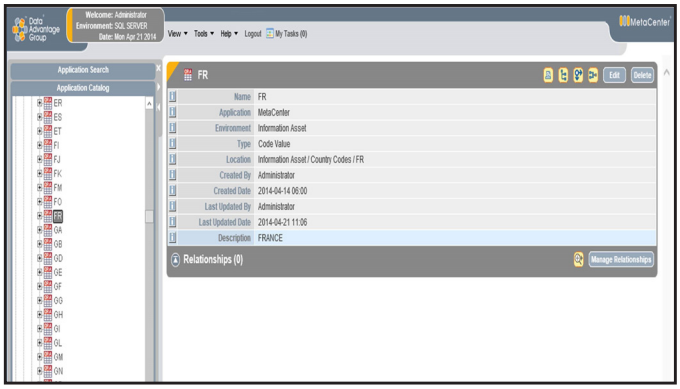


Figure 13: Data Advantage Group MetaCenter contains country codes relating to GDP.

7. Profile Data Natively in Hadoop

Organizations also need to profile data natively in Hadoop without moving large volumes into structured data stores.

Figure 14 shows the Global IDs HDFS File Profiler. The screenshot shows a column analysis for the product data in semi-structured format in the **Inventory.txt** file. The profiler has uncovered 49 columns, including **FAMILY, PRODUCTGROUP, COMMODITYGROUPID, PRODUCT_TYPE, and SELL_SWAP_INDICATOR.**

In Figure 15, Talend Data Quality profiles data natively in Hive. By leveraging the Hadoop MapReduce distributed architecture, Talend Data Quality can cleanse, match, and de-duplicate large volumes of data without having to move it prior to processing.

Informatica also supports data profiling natively on Hadoop through Informatica Developer and the browser-based Informatica Analyst to better understand the data, identify data quality issues, collaborate on data flow specifications, and validate mapping, transformation, and rules logic.

In Figure 16, the data analyst can view unique values, null values, percent nulls, inferred data types, and other descriptive statistics to identify outliers and anomalies.

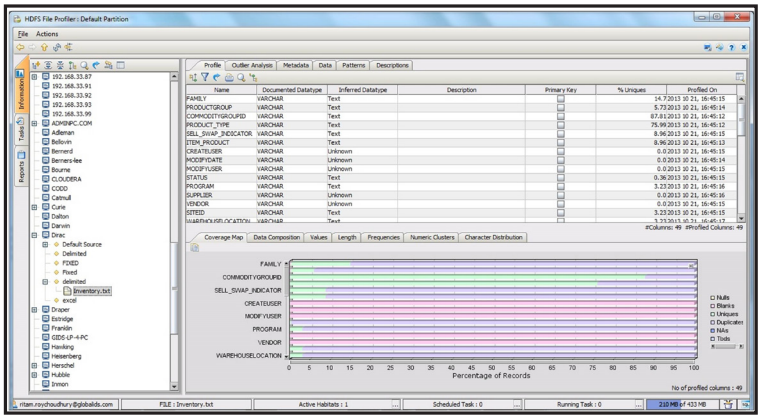


Figure 14: Global IDs HDFS File Profiler.

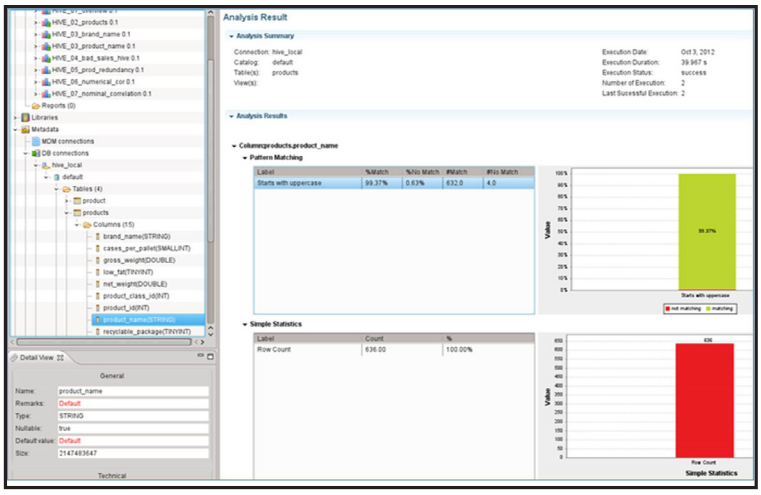


Figure 15: Talend Data Quality profiles data natively in Hive.

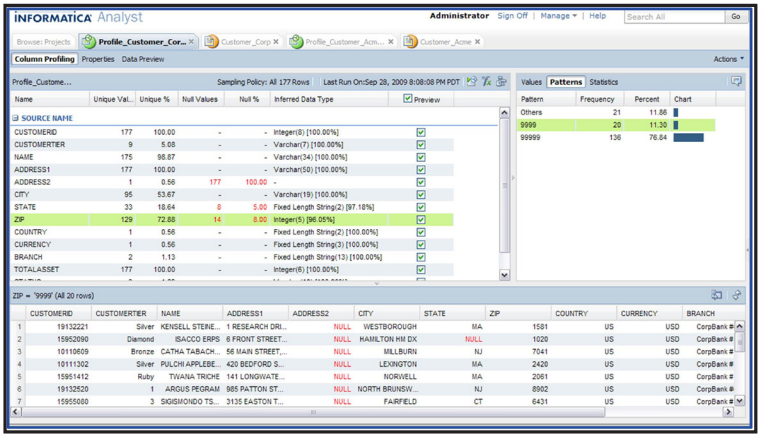


Figure 16: Informatica Data Quality Big Data Edition profiles data natively in Hadoop.

8. Discover Data Natively in Hadoop

Data profiling tools should also natively discover hidden data in Hadoop.

Figure 17 shows how Informatica Data Quality Big Data Edition can be set up to discover **IPAddress** and **StockSymbol** data within a **WebLogs** Logical Data Object in Hadoop.

As shown in Figure 18, Informatica Data Quality Big Data Edition shows the results of the data domain discovery for the **WebLogs** Logical Data Object. We can see that 100 percent of the values in the **IP_Address** column conform to the **IPAddress** pattern. The same holds true for the **Stock_Symbol** column relative to the **StockSymbol** pattern.

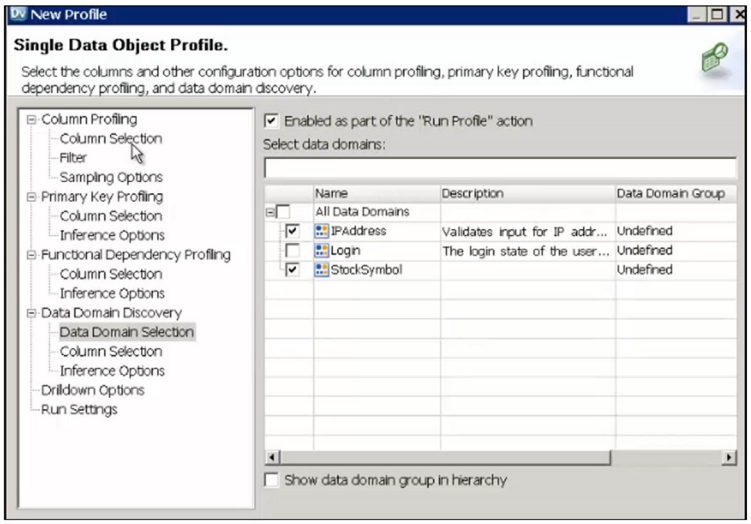


Figure 17: Informatica Data Quality Big Data Edition supports data domain discovery within Hadoop.

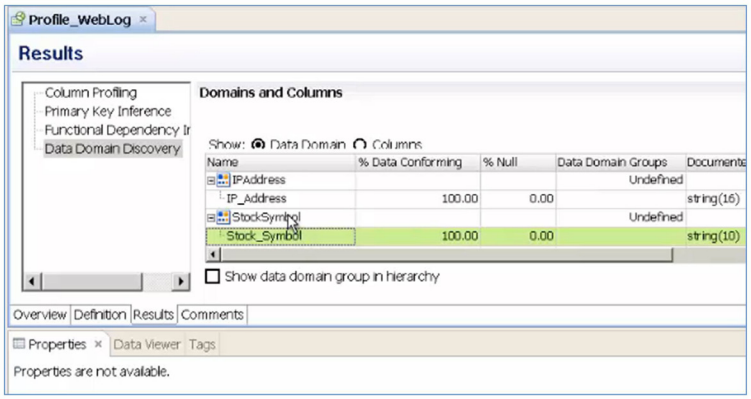


Figure 18: Informatica Data Quality Big Data Edition shows the results of the data domain discovery exercise.

9. Execute Data Quality Rules Natively in Hadoop

Data quality tools should also execute rules natively in Hadoop.

In Figure 19, we show a sample web log file called **DailyWebLogDump.txt**.

In Figure 20, we can see the mappings in Informatica Data Quality Big Data Edition. The **DailyWebLogDump.txt** is read as a binary source file in the **Read** mapping on the left. The file is then parsed in the **POC_ Session** transformation in the middle and the results passed to the Joiner mapping on the right.

Figure 21 provides further detail on how the binary file is parsed in Informatica Data Quality Big Data Edition. The parser conducts a text search for VISITOR_id. It then populates any associated text found in parentheses in the VISITOR_ID element in the VISITOR object in a predefined schema.

Figure 19: Sample web log file in semi-structured format.

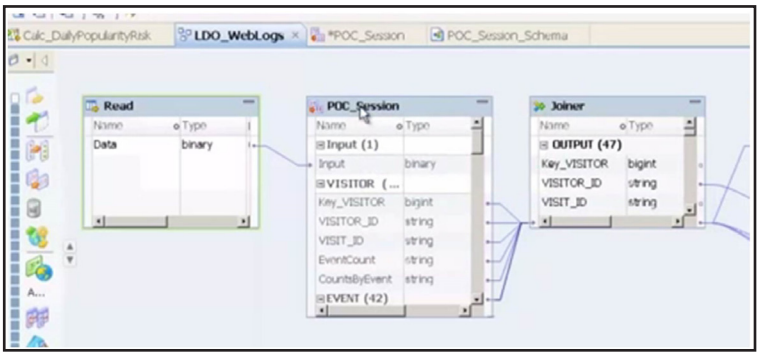


Figure 20: Mappings in Informatica Data Quality Big Data Edition.

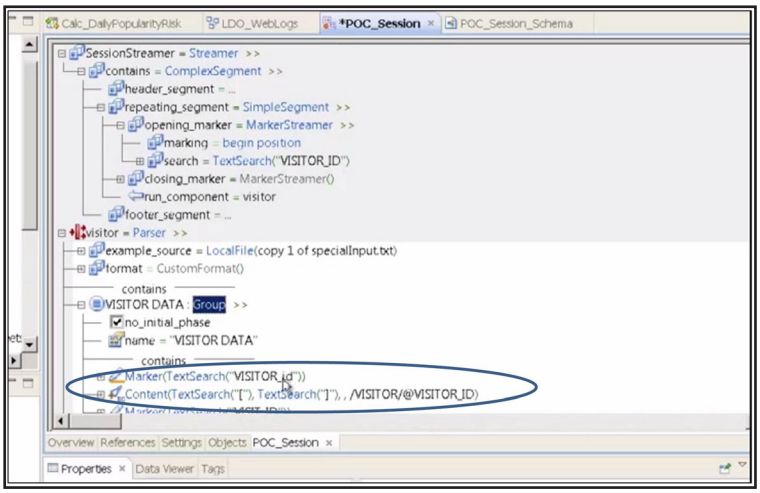


Figure 21: Complex data parsing with Informatica Data Quality Big Data Edition.

10. Integrate Master Data Management with Hadoop for Improved Performance

Organizations can also leverage the power of Hadoop to turbo charge their MDM hubs. For example, marketing organizations often need to match lists of prospects against internal records to remove any customers who have made do-not-call elections. These large data sets push the limits of existing computational resources when IT needs to match 200 million prospects against a database of 100 million customers and turn around the results to marketing in 24 hours. MDM vendors such as IBM and Informatica have integrated their MDM hubs with Hadoop.

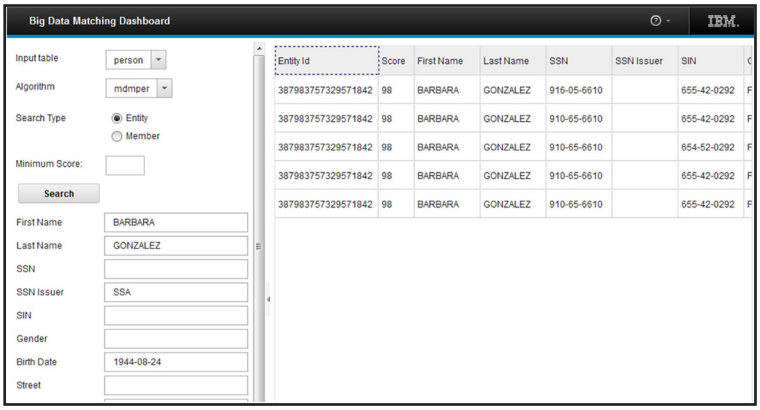


Figure 22: Probabilistic matching engine within IBM InfoSphere BigInsights ("Big Match").

Figure 22 shows Big Match from IBM. The company has embedded the IBM InfoSphere Master Data Management probabilistic matching engine within the IBM InfoSphere BigInsights Hadoop distribution.

11. Mask Sensitive Data in Hadoop

Big Data teams also need to mask sensitive data within their Hadoop data lakes. Many vendors, such as IBM and Informatica, support data masking functionality natively in Hadoop.

Figure 23 shows the batch interface in HDFS that uses IBM InfoSphere Optim Data Privacy Solution to mask sensitive data.

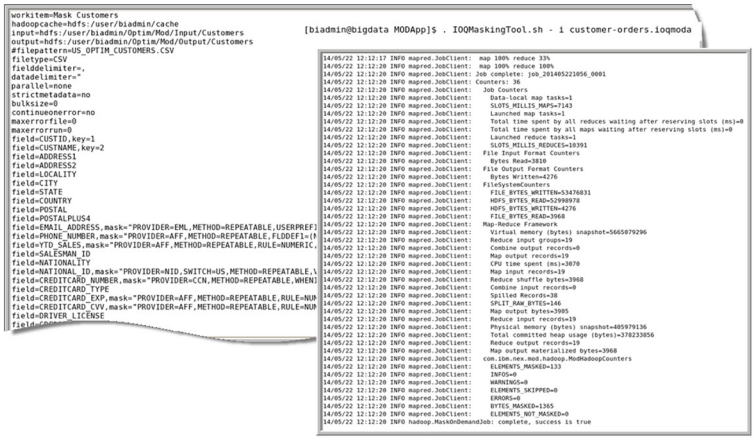


Figure 23: IBM InfoSphere Optim Data Privacy Solution masks sensitive data natively in Hadoop.

12. Manage Workflows Associated with The Approval Process for Key Data Artifacts

Data governance tools should also manage the approval process to add, modify, and delete key data artifacts in Hadoop. For example, the data governance team manages a list of key business terms in Hive.

As shown in Figure 24, the team uses the Simple Approval workflow in Collibra Data Governance Center to manage the process to add a new business term. Once the term has been approved, Collibra initiates a HiveQL query to update the glossary in Hive. This process can also be used to approve other data artifacts in Hadoop, such as product lists and reference data.

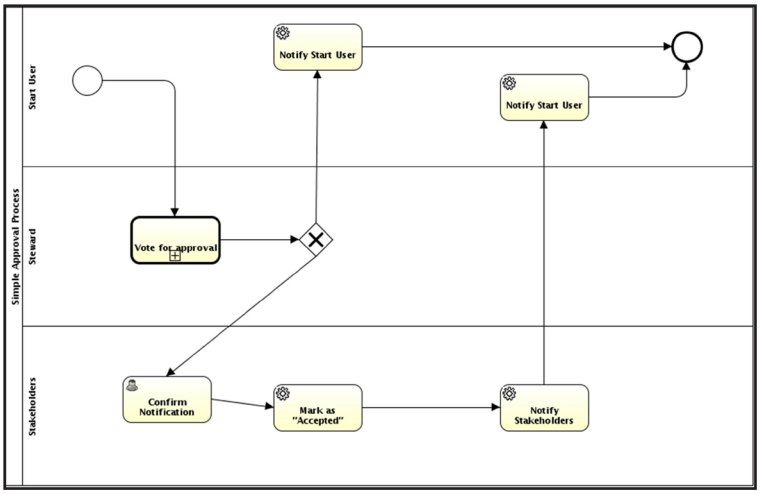


Figure 24: Simple approval workflow in Collibra Data Governance Center.

Glossary

Hadoop
Apache Hadoop is an open source software library that supports the distributed processing of large data sets across thousands of computers based on commodity hardware.

Hadoop Distributed File System (HDFS)
HDFS is a distributed file system that is designed to be highly fault tolerant and to run on low-cost hardware.

Hive
Apache Hive is a data warehousing infrastructure that sits on top of Hadoop. Hive provides a SQL-like interface called HiveQL to query large volumes of data in Hadoop. Because Hive insulates users from having to learn the intricacies of MapReduce programming in Java, it is a great transition for relational database programmers who are looking to work with Hadoop.

MapReduce
MapReduce applications can process vast amounts (multiple terabytes) of data in parallel on large clusters in a reliable, fault-tolerant manner.

NoSQL
A database management system that does not use SQL as its primary query language.

Pig
Apache Pig is a platform for analyzing large semi-structured data sets in Hadoop. It uses a procedural language called Pig Latin that insulates users from learning the intricacies of MapReduce programming in Java. Hive and Pig evolved as separate Apache projects for the analysis of large data sets. Hive is better suited to users who are familiar with SQL. Pig, on the other hand, is ideal for users who are familiar with procedural programs like Microsoft Visual Basic and Python.

Sqoop
Apache Sqoop is a tool that supports the movement of massive volumes of data between Apache Hadoop and structured data stores such as relational databases. Many ETL vendors also support similar functionality.

Stream computing
A class of applications that perform high-performance, low-latency processing of large volumes of data leveraging parallel processing capabilities without landing data to disk.

About the Author

Sunil Soares is the Founder and Managing Partner of Information Asset, a consulting firm focused on Data Governance and Enterprise Data Management. He is the author of several books, including *Selling Information Governance to the Business and Big Data Governance.* His book on *Data Governance Tools* will be published in Fall 2014. For more information, please visit www.information-asset.com.