Big Data Governance

Big Data Governance Use Cases with Collibra

By **Sunil Soares**

February 10, 2015

Collibra White Paper

Collibra

http://www.collibra.com

sales@collibra.com

Summary

The initial focus of Big Data efforts has been on analytical use cases. However, Big Data Governance is becoming front and center as organizations grapple with certain challenges:

- How can we make our data scientists more productive by providing them with easy access to definitions, business rules, data standards, and reference data?
- How can we make our legal and compliance teams comfortable that we are using data in an acceptable manner?
- Do we have an agreed upon process to manage changes to our analytic models? Are these models only using approved data?
- How can we keep our data scientists and development teams in sync as we make changes to our Big Data platforms?
- Do we have assigned data owners to answer any questions relating to Big Data?
- Do we have an agreed upon process to manage data issues?

This white paper reviews a number of uses cases for Big Data Governance with Collibra Data Governance Center.

Table of Contents

Introduction

Big Data Governance is part of a Data Governance program that formulates, monitors, and enforces policies relating to Big Data. The initial focus of Big Data efforts has been on analytical use cases to support the needs of data scientists, with Data Governance being an after-thought. This point is evidenced by the fact that the landing page for Apache Hue does not have any modules that pertain directly to Data Governance (see Figure 1).

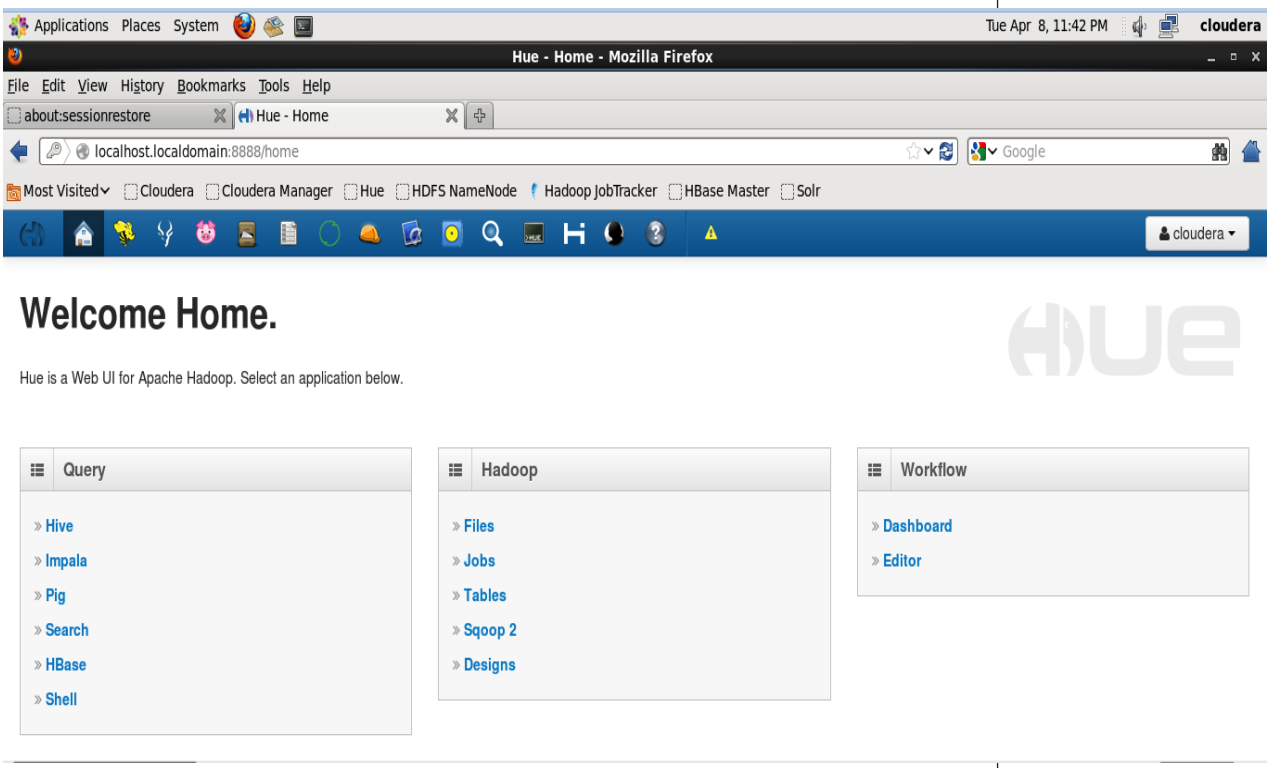


Figure 1: Landing page for Apache Hue has limited Data Governance functionality.

However, there will be an increasing focus on Big Data Governance as Big Data efforts become mainstream. Collibra Data Governance Center is market-leading software for Data Governance. This white paper will examine key use cases for Big Data Governance with Collibra Data Governance Center.

Enteprise Data Management Policies

The first step in governing Big Data is to adopt an overall policy as shown in Figure 2. Collibra supports workflows to gather policy approvals, and to log any changes to policies over time.

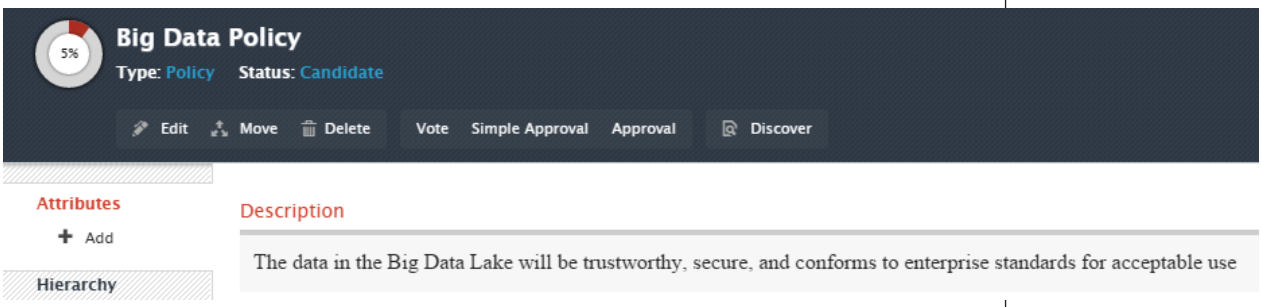


Figure 2: Enterprise Big Data Policy in Collibra

Figure 3 shows the underlying Enterprise Big Data standards in Collibra. These standards cover overall principles such as data inventory, data ownership, Critical Data Elements (CDEs), Critical Big Datasets (CBDs), data quality, information security, data lineage, and data retention. Collibra maintains a parent-child relationship between the Big Data policy and standards.

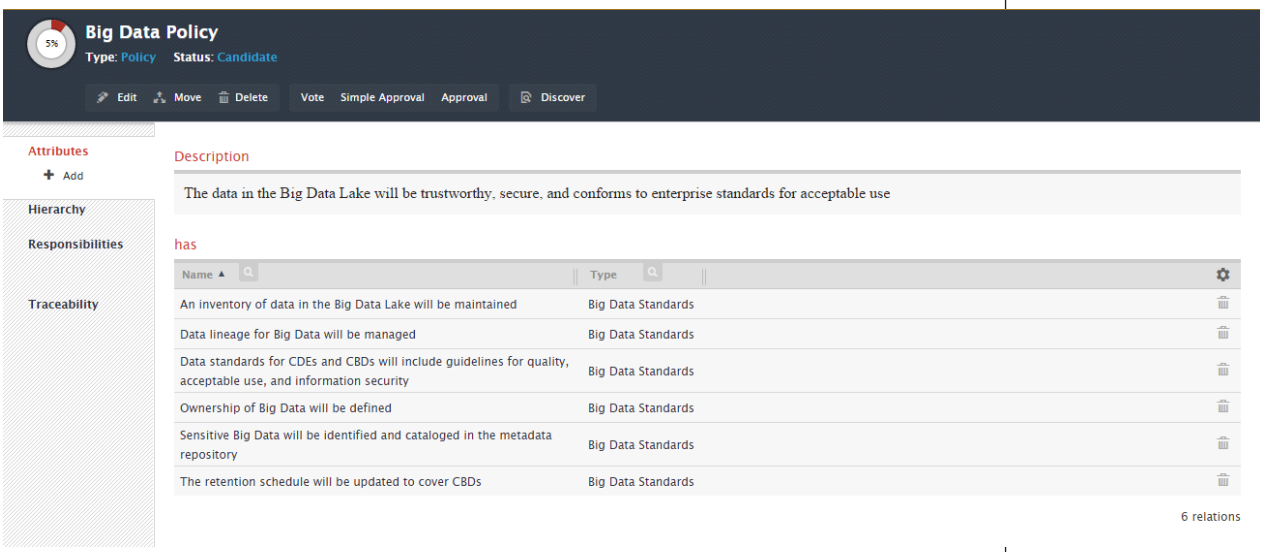
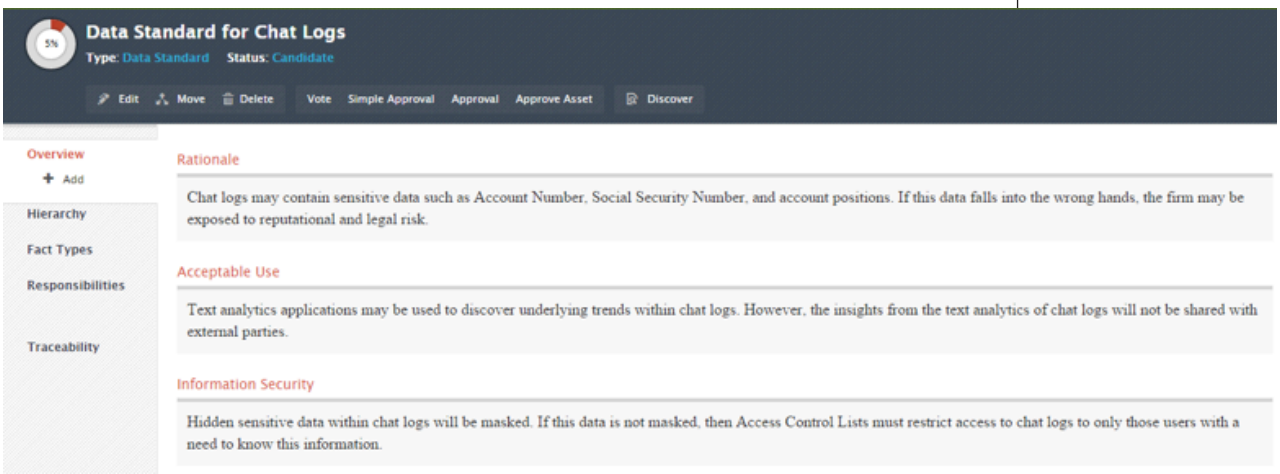


Figure 3: Enterprise Big Data Standards in Collibra

Formulate Standards for Critical Datasets

Data standards should also be defined for Critical Big Datasets (CBDs) such as chat logs, Facebook data, Twitter feeds, and sensor data. Figure 4 shows a data standard for chat logs in Collibra. Chat logs may contain sensitive data such as Account Number, Social Security Number, and account positions. If this data falls into the wrong hands, the firm may be exposed to reputational and legal risk. The data standard states that text analytics applications may be used to discover underlying trends within chat logs. However, the insights from the text analytics of chat logs will not be shared with external parties. The data standard also states that hidden sensitive data within chat logs will be masked. If this data is not masked, then access control lists must restrict access to chat logs to those users with a need to know this information. Once again, Collibra manages the approval workflows as well as a history of any changes to these data standards.

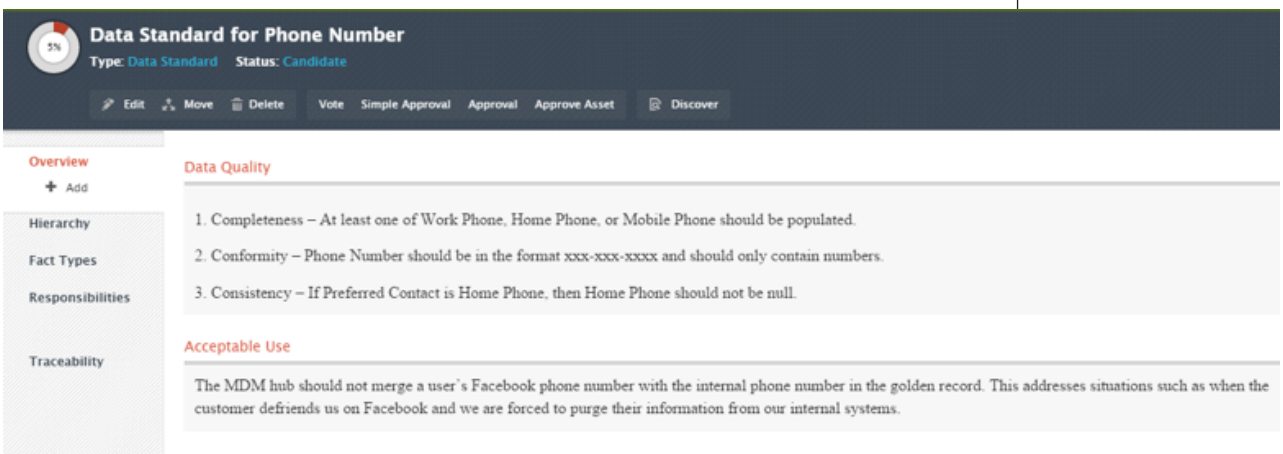
Figure 4: Data Standards for chat logs in Collibra



Policies for Critical Data Elements

Data standards should also be defined for Critical Data Elements (CDEs) such as phone number. As shown in Figure 5, the data standard includes data quality rules relating to the completeness, conformity, and consistency of phone numbers. The data standard also states that a user's Facebook phone number should not be merged with the golden record in the master data hub. This data standard addresses situations such as when a user defriends a company on Facebook. In this case, the company has to purge user-provided Facebook data from its internal systems, which would be hard to do if Facebook phone numbers were merged into the master data hub.

Figure 5: Data standard for phone numbers



Assign Ownership of Big Data

The Data Governance team also needs to assign ownership of enterprise data assets, including Big Data. These data owners are accountable for data definitions, data standards, acceptable use policies, data quality, and data access. As shown in Figure 6, these enterprise data assets may be the following:

- Traditional data domains such as customer, product, vendor, and chart of accounts.
- Datasets such as Twitter, Facebook, chat logs, and RFID data.
- Critical Data Elements such as phone number and product category.
- Data platforms such as Hadoop, Cassandra, the enterprise data warehouse, and Oracle databases.

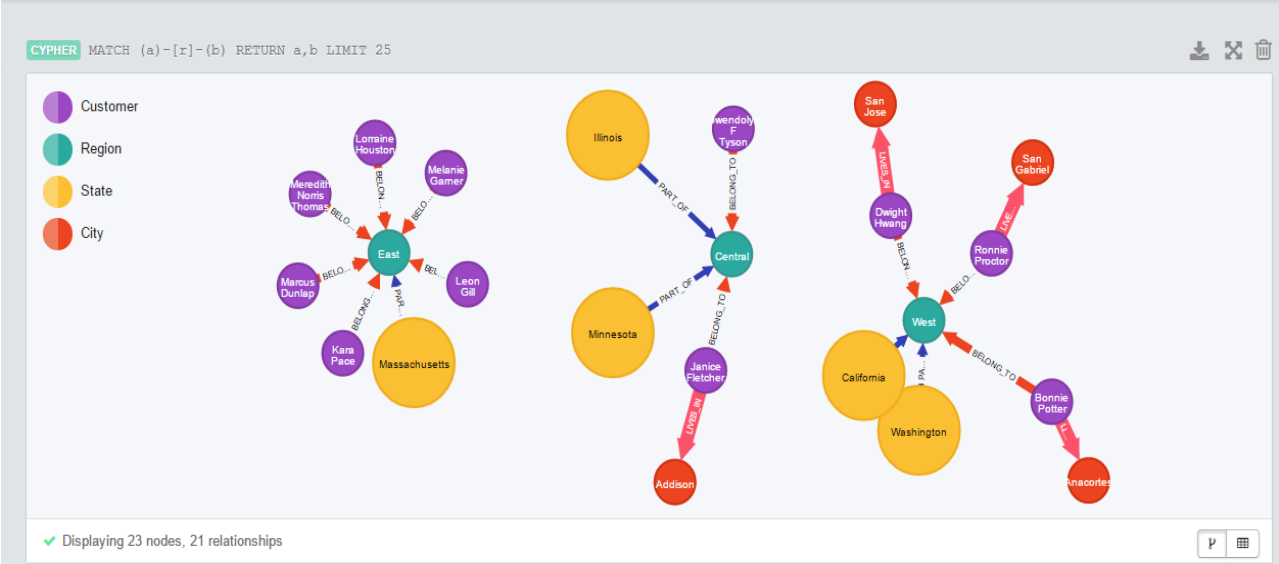
Figure 6: Data Ownership in Collibra

Name ▲	Organization	Type	Data Executive	Data Steward	Managing Data Steward
Cassandra	Finance	Data Platform	John Smith	Mary Jane	Jane Smith
Facebook	Marketing	Dataset	Mary Jane	Jean Hill	Nancy Smith
Hadoop	Merchandising	Data Platform	Jack Murphy	Tom Smith	Jill Smith
Phone Number	Marketing	Critical Data Element	Mary Jane	Jean Hill	Nancy Smith
RFID	Supply Chain	Dataset	Liz Shi	Maya Danielle	Helena O'Toole
Twitter	Marketing	Dataset	Mary Jane	Jean Hill	Nancy Smith

Semantic Layer for Big Data analytics

Big Data applications need high quality business terms to support analytic use cases. Figure 7 shows customer data in the Neo4j NoSQL graph database. By way of background, NoSQL ("Not Only SQL") databases are a category of database management systems that do not use SQL as their primary query language. NoSQL databases include Apache HBase, Apache Cassandra, MongoDB, and the Neo4j graph database.

Figure 7: Visual depiction of customer details in Neo4j NoSQL graph database



Neo4j's strong graphical capabilities can be complemented by Data Governance functionality in Collibra. Figure 8 shows the definition of the term 'Customer' in Collibra.

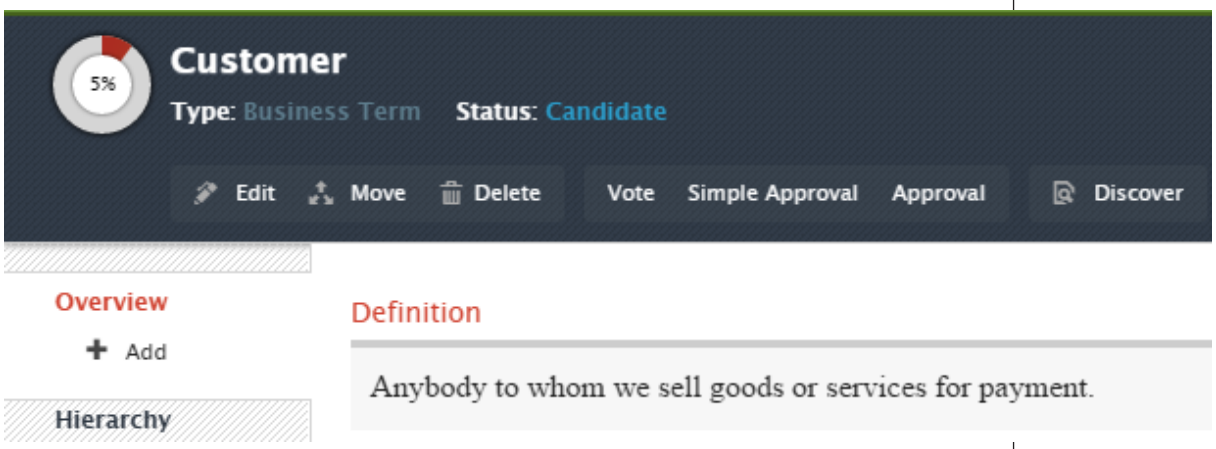


Figure 8: Definition of 'Customer' in Collibra

Govern Analytic Models for Big Data

The Data Governance team needs to manage metadata for analytic models in Collibra. As shown in Figure 9, this metadata includes the model name, model description, the business data steward, the data scientist owner, and the technical data steward. Additional attributes such as the model drivers, model methodology, and model create date can also be managed in Collibra.

Name	Description	Business Data Steward	Data Scientist Owner	Technical Data Steward	Status	Type
High Risk Options Propensity Model	High Risk Options Propensity M...	John	Jane	Mary Jose	Candidate	Data Model

Figure 9: Model metadata in Collibra

Figure 10 shows a traceability diagram in Collibra. The High Risk Options Propensity Model is dependent on the Net Income, Date of Birth, and Net Worth CDEs. The Net Worth CDE also has a documented data standard. In addition, the Net Worth CDE is represented by the nw_500 column in the cust_500 Hive table.

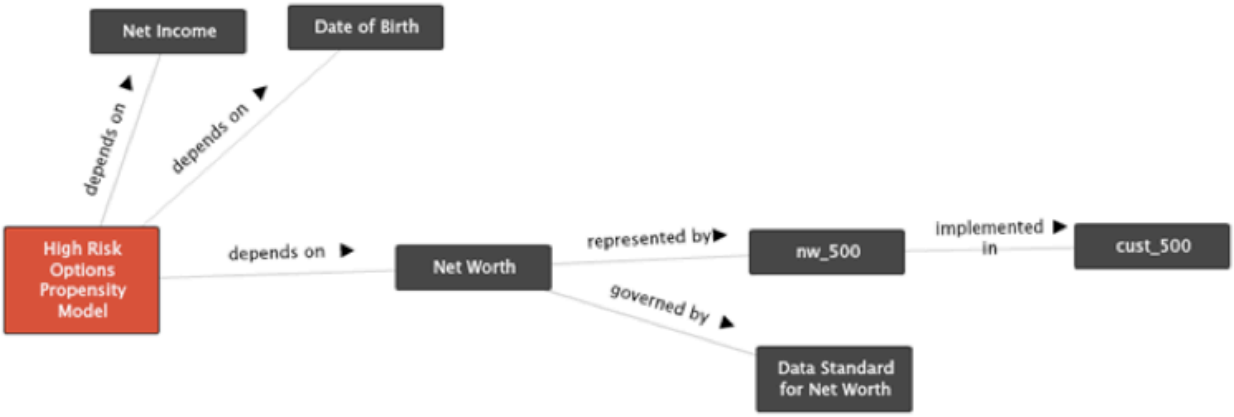
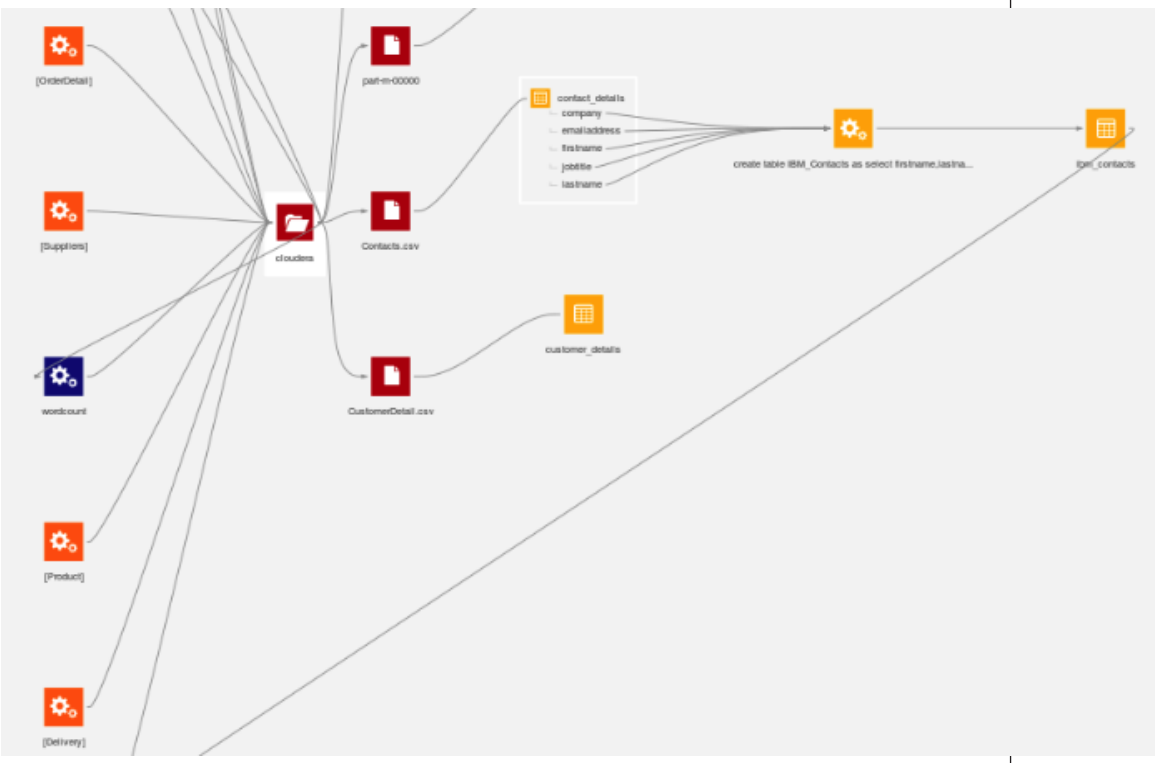


Figure 10: Traceability diagram showing analytic models, CDEs, data standards, Hive column, and Hive tabe in Collibra

Enrich Metadata for Big Data

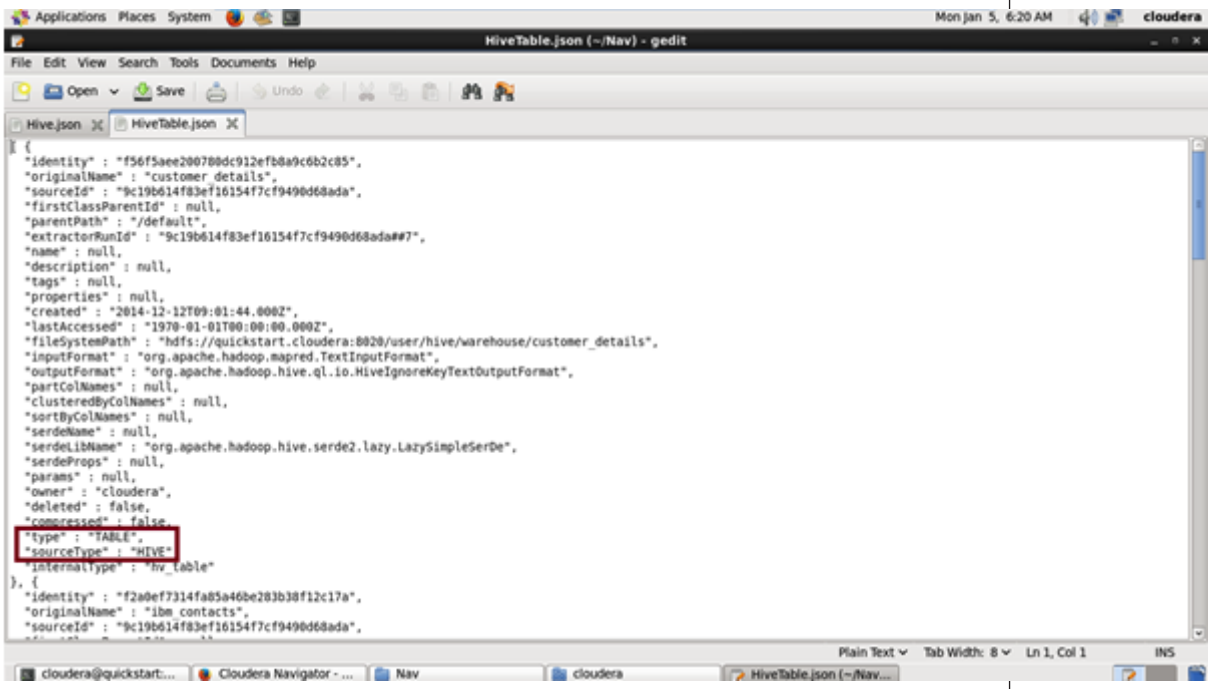
Collibra can also enrich the metadata for Big Data. Figure 11 shows Hadoop data lineage in Cloudera Navigator that includes Sqoop jobs, HDFS files, Hive tables, and Hive queries.

Figure 11: Hadoop data lineage in Cloudera Navigator



Collibra can supplement the metadata in Cloudera Navigator with additional content, such as business terms and data ownership. In Figure 12, the Cloudera Navigator metadata can be exported in JavaScript Object Notation (JSON) format.

Figure 12: Cloudera Navigator metadata in JSON format



The Cloudera Navigator metadata can then be imported into Collibra using a number of techniques including Python scripting. In Figure 13, Collibra shows the data steward for each Hive table that was imported from Cloudera Navigator. These Hive tables can also be associated with business terms in Collibra.

Name ▲	Status	Type	Data Steward
contact_details	Candidate	Hive Table	Jane
contacts	Candidate	Hive Table	Jane
customer_details	Candidate	Hive Table	Jane
customerdetails	Candidate	Hive Table	Jane
ibm_contacts	Candidate	Hive Table	Jane

Figure 13: Data stewardship for Hive tables in Collibra

Automate Manual Processes

Collibra's workflow functionality can formalize and automate various manual processes associated with the Big Data Lake. These out-of-the-box or custom workflows include the following:

Onboard a new data asset

Figure 14 shows a simple approval workflow that involves a steward and stakeholder. A more complex workflow may also include additional stakeholders such as legal and compliance.

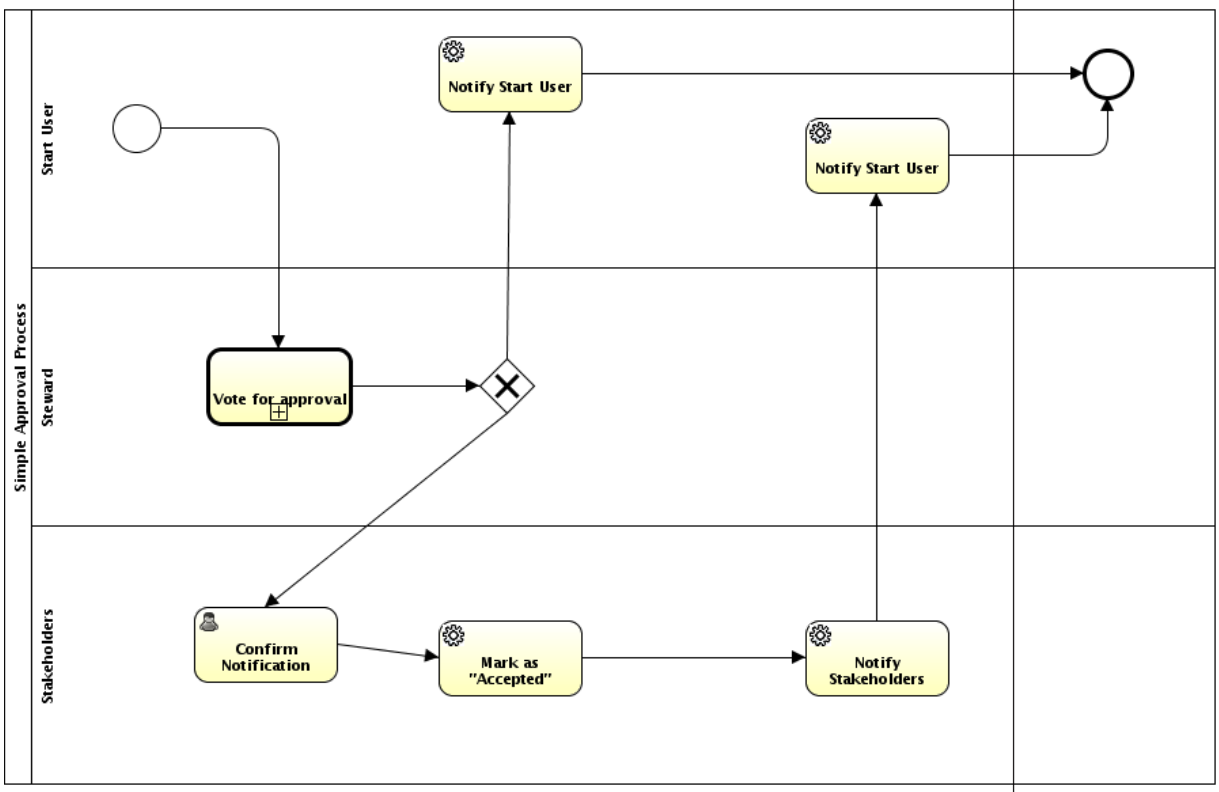


Figure 14: Simple approval workflow in Collibra

Approve the creation of a new Analytical Model

Figure 15 shows a custom form to propose a new analytic model. A business user requests a new analytic model to upsell life insurance to recently married couples. Collibra will then route this form to the following stakeholders for approval:

- The data management team, which will investigate the requirements in more detail.
- The data scientist team, which will determine if a similar model already exists.
- The legal team, which will sign-off on any compliance issues.

Once this form has been approved in Collibra, the data scientists will begin the process of creating the analytical model.

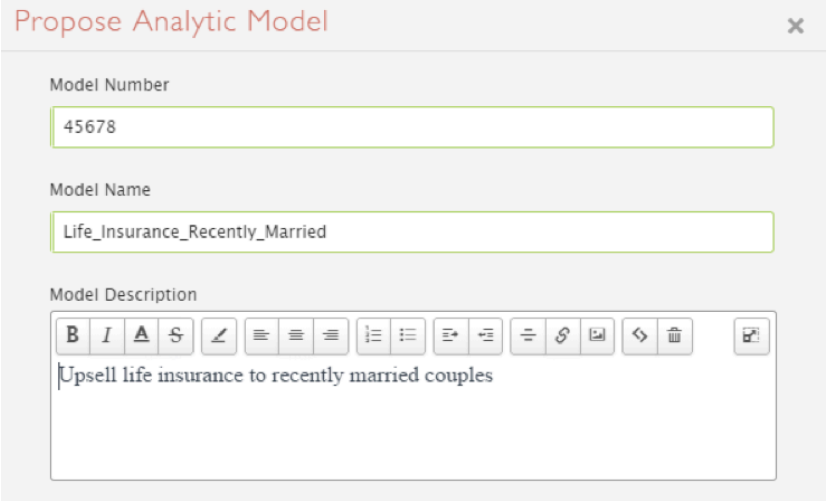


Figure 15: Custom Collibra form to onboard a new analytical model

Manage Data Issues related to Big Data

Figure 16 shows a data issue form that states that the *emp_num* column in the *cust* Hive column contains Social Security Numbers. The marketing department will investigate this issue. Collibra will manage this issue throughout its lifecycle from logging, data owner assignment, investigation, and resolution.

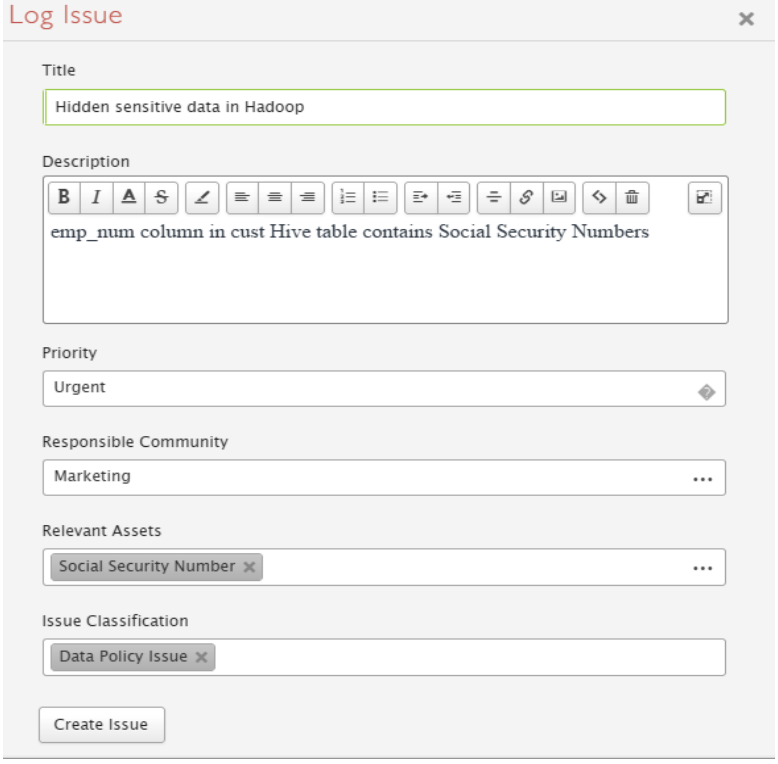


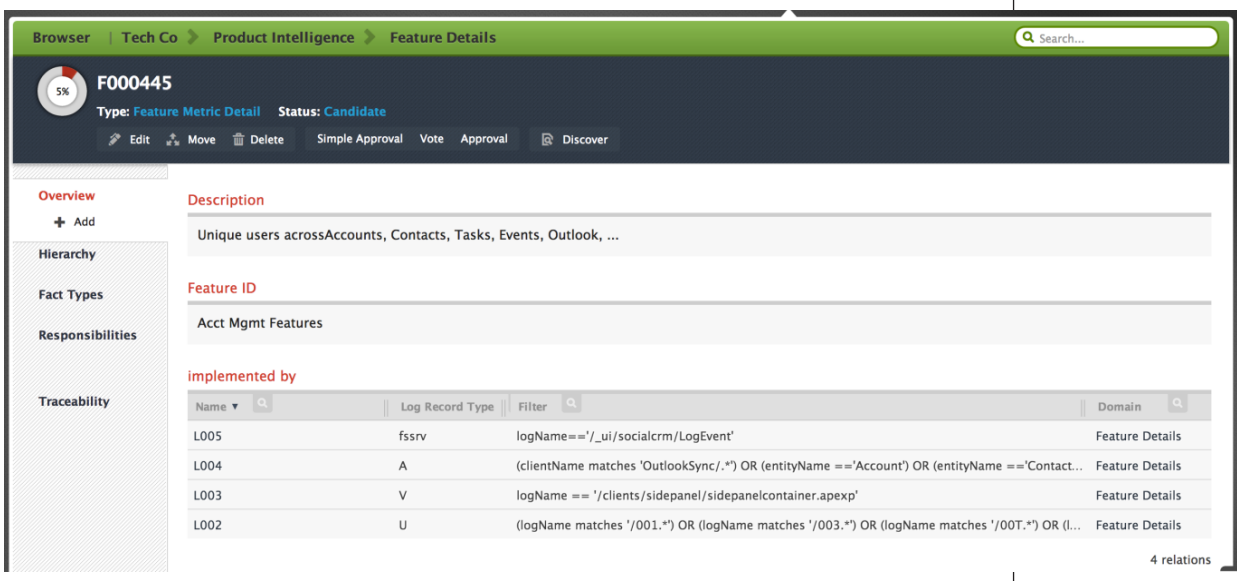
Figure 16: Collibra form to log a new data issue relating to Hadoop

Reference Data for Big Data Queries

Big Data repositories often store data in unstructured or semi-structured format. Because it is no longer possible to use a pre-defined data model, the data scientist must rely on actual data values or reference data to formulate queries. If the data scientist uses reference data values that are different from the actual values, the queries will return incorrect results.

The individuals that update reference data are often different from the people that query the data to derive insights. For example, technology companies often query terabytes of product logs to understand how often certain features are used. Because product logs are in semi-structured format, data scientists cannot rely on the product log data model to distinguish between different features. They need to know the reference data that is used for each product feature, to correctly aggregate the usage metrics per feature. When the development team changes how certain features are logged between releases, the data scientists need to be aware of these changes. If not, they will come to incorrect conclusions. Collibra workflows can be used to ensure that development teams and data scientists sign off on any changes to product logs. In Figure 17, Collibra manages how the 'F000445' product feature is implemented across multiple log files.

Figure 17: Reference data for the feature 'F000445' across multiple log files in Collibra



Govern changes to Attributes in NoSQL databases

As shown in Figure 18, NoSQL databases such as Apache Cassandra have a flexible data model. While this flexibility empowers the business, it loses the automatic governance of changes associated with the rigid data model of traditional relational databases. Collibra governance workflows ensure that all necessary documentation, quality, and impact analysis checks have been performed before changing the NoSQL data model. This improves the quality and consistency of content in the NoSQL database, and ensures that data can be correctly aggregated.

Figure 18: Apache Cassandra has a flexible data model

```
File  Edit  View  Search  Terminal  Help
Connected to Demo Cluster at localhost:9160.
[cqlsh 4.1.1 | Cassandra 2.0.8.39 | CQL spec 3.1.1 | Thrift protocol 19.39.0]
Use HELP for help.
cqlsh> USE customerdb;
cqlsh:customerdb> CREATE TABLE customer (cust_id uuid, user_name varchar, first_name varchar, last_name varchar, email list<varchar>, password varchar,create
d_on timestamp, PRIMARY KEY (cust_id,user_name));
```

About the Author

Sunil Soares is the Founder and Managing Partner of Information Asset, a consulting firm focused on Data Governance, Big Data Governance, and Enterprise Data Management. He is the author of several books, including Selling Information Governance to the Business, Big Data Governance, and Data Governance Tools.