INFORMATION ASSET, LLC
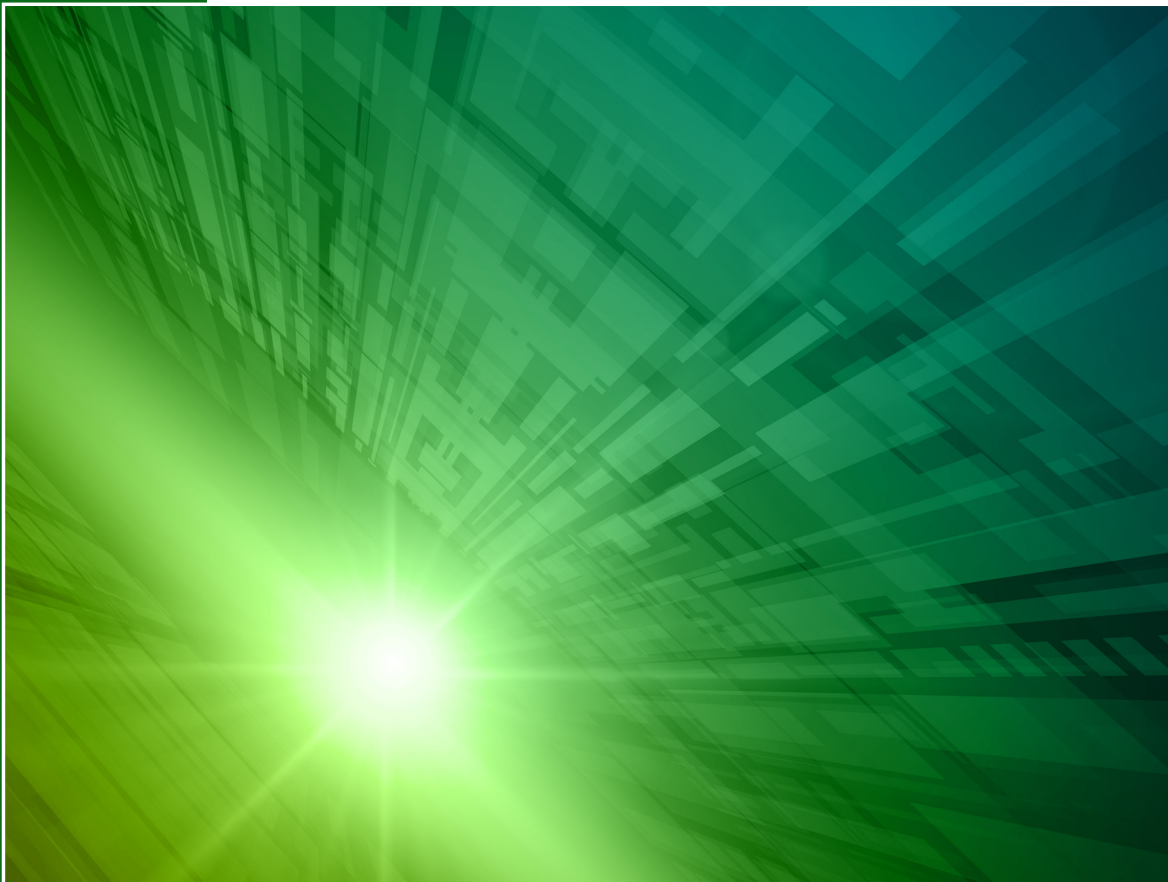
White Paper

An In-Depth Review of Data Governance Software Tools:

Reference Architecture, Evaluation Criteria, and Vendor Landscape

by Sunil Soares
January 27, 2014

Introduction

Data governance is the formulation of policy to optimize, secure, and leverage information as an enterprise asset by aligning the objectives of multiple functions. Data governance programs have traditionally been focused on people and process. Cost has historically been a key consideration because data governance programs have often started from scratch with little to no funding. As a result, Microsoft Excel and SharePoint are the tools of choice to document and share data governance artifacts. The marginal cost of these tools is zero, but they are often missing critical functionality, as we discuss later in this report. Notwithstanding the above, vendors have matured their offerings to the extent that organizations need to consider tools as a critical component of their data governance programs.

Reference Architecture for Enterprise Data Management

As shown in Figure 1, the reference architecture for Enterprise Data Management (EDM) includes 20 categories. At the very bottom, we have the data sources that need to be governed. Hadoop and streaming analytics are two categories focused on big data. Databases may be NoSQL, in-memory, relational, or legacy. We then have data modeling tools. Next, data integration tools may be for bulk data movement, data replication, or data virtualization.

The next four EDM categories are data discovery, data quality, business glossaries, and metadata. Then, we have text analytics, information policy management, master data management, and reference data management. The next two EDM categories are data warehouses and data marts, and analytics and reporting. Business process management, data security and privacy, and information lifecycle management are the next three EDM categories. The final EDM category is cloud.
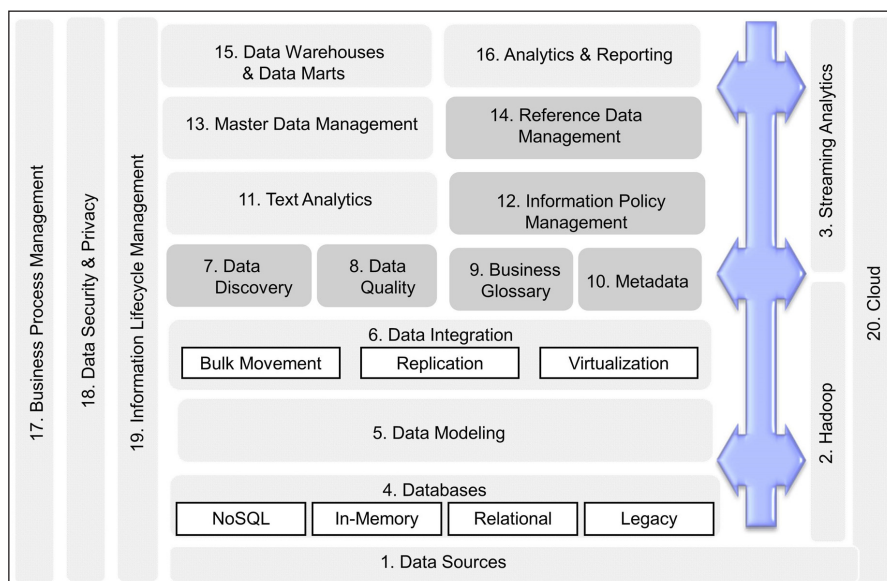
Figure 1: EDM reference architecture includes six highlighted categories for data governance.

Data Governance Software

Of these 20 categories of tools, we believe that six are most closely associated with data governance programs. We refer to these categories in aggregate as data governance software. The categories are listed below:

1. *Data Discovery*
   Offerings include IBM InfoSphere Information Analyzer, Informatica Data Explorer, Oracle Enterprise Data Quality Profile and Audit, SAP Information Steward, SAS Data Management, and Trillium Software TS Discovery. These tools support a variety of data sources, including Hadoop.

2. *Data Quality*
   Offerings include IBM InfoSphere Quality Stage, Informatica Data Quality, Oracle Enterprise Data Quality, SAP Information Steward, SAS Data Management, and Trillium Software TS Quality.

3. *Business Glossary*
   Offerings include Adaptive Business Glossary Manager, ASG-metaGlossary, Collibra Business Semantics Glossary, Embarcadero CONNECT, IBM InfoSphere Business Glossary, Informatica Business Glossary, and SAS Data Management.

4. *Metadata*
   Offerings include Adaptive Metadata Manager, ASG-Rochade, Data Advantage Group Meta-Center, IBM InfoSphere Metadata Workbench, and Informatica Metadata Manager.

5. *Information Policy Management*
   Most business glossary vendors also support the management of information policies. In addition, Governance, Risk, and Compliance (GRC) platforms such as EMC RSA Archer GRC and IBM OpenPages GRC also offer capabilities to manage broader policies, including those relating to information.

6. *Reference Data Management*
   Offerings include Collibra Data Governance Center, IBM InfoSphere Master Data Management Reference Data Management Hub, Oracle Data Relationship Ma nagement, and Orchestra Networks EBX.

Integration Between Data Governance Tools and the Broader EDM Stack

An exhaustive examination of all the components of the EDM stack is beyond the scope of this research paper. However, we will discuss key components of the EDM stack as they relate to data governance tools:

1. *Data Integration*
   Offerings include IBM InfoSphere DataStage, Informatica PowerCenter, Pentaho, and Talend. Data integration tools have a few key points of integration with data governance:

   - Propagate reference data—Once a reference code value has been updated in the data governance tool, extract, transform, and load (ETL) jobs can propagate those code values to back-end applications.

   - Discover candidates for reference data—A health plan used ETL to discover new candidates for reference data. When loading rows, if the ETL job found a code value that was not part of the approved code list, it would "soft insert" that row into the data warehouse. The ETL tool would then flag the code value as a candidate for review by the data stewards.

   - Handle exceptions—Data exceptions, such as failed load jobs, may be handled by the data issue resolution tool.
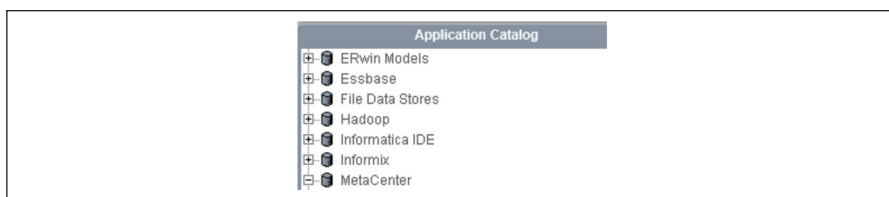
2. *Hadoop*
   As companies put more data into Hadoop stores, there will be an increasing demand for governance over that data. Most vendors either have or are in the process of adding data profiling, data quality, and metadata capabilities for Hadoop (see Figure 2).

   Figure 3 shows an example of a Hadoop data store within a metadata repository.
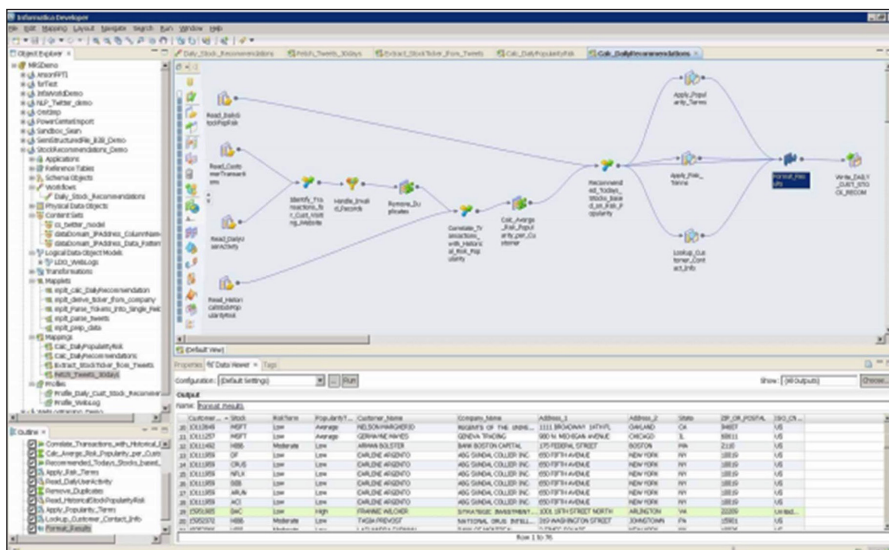
Figure 2: Informatica PowerCenter Big Data Edition supports data profiling, ETL, and complex data parsing on Hadoop.

Figure 3: ActiveLinx for Hadoop within Data Advantage Group MetaCenter.

3. *Master Data Management (MDM)*
   We were conflicted about whether to include MDM tools within data governance. In the end, we decided to keep them separate. Although MDM and data governance are closely related, we determined that the use cases were sufficiently distinct to warrant keeping these two separate. MDM vendors include IBM, Informatica, Oracle, Orchestra Networks, Riversand, SAP, SAS, Semarchy, Stibo Systems, and Talend. Notwithstanding the symbiotic relationship between MDM and data governance, the software tools are often not well integrated. MDM tools have embedded rules for column names, data enrichment, data validation, entity relationships, record matching, hierarchy management, de-duplication, survivorship, confidence thresholds, and record consolidation.

   Figure 4 shows a sample MDM data validation rule that requires source address to contain at least one address line and either a postal code or a city. MDM rules are often developed by data analysts and are not exposed to the business. MDM rules should be documented in the business glossary, integrated with the MDM hub, and exposed to data stewards.

Figure 4: Data validation rule in Semarchy Convergence for MDM.

4. *Analytics and Reporting*
   Offerings include IBM Cognos and SPSS, Microsoft SQL Server Reporting Services and SQL Server Analysis Services, Oracle Business Intelligence, Pentaho, QlikView, SAP BusinessObjects, SAS, and Tableau. There are three key points of integration between analytics/reporting and data governance tools:
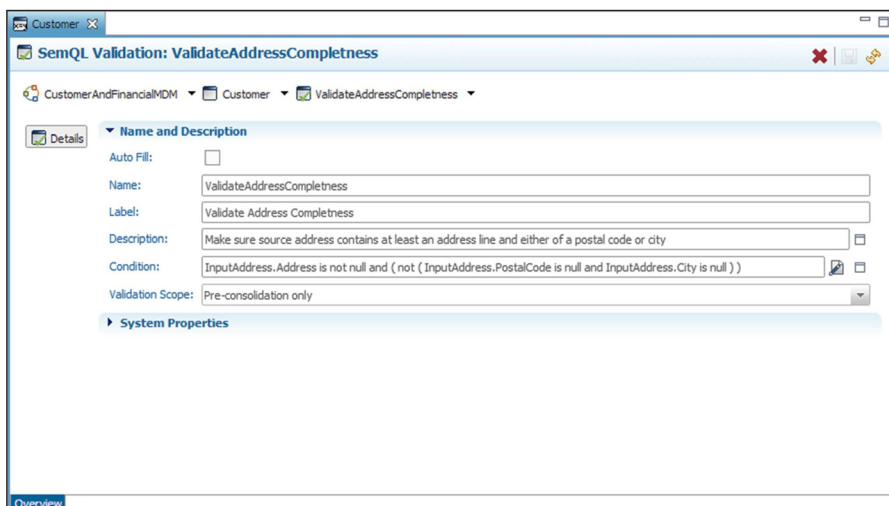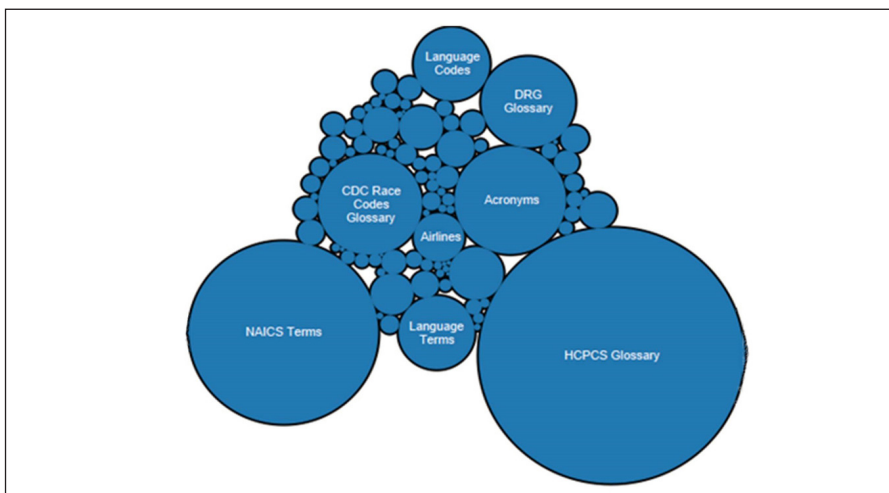

**Figure 5:** Visualization of data domains in Tableau.

- Data profiling—Users can export data profiling results to a reporting tool such as Tableau for further visual analysis (see Figure 19 later in this report).

- Data governance metrics—Users can export data artifacts into a reporting tool for visualization of data governance metrics. Figure 5 shows a visualization of data domains in Tableau.

- Integration with business glossary—Many business glossary tools ship with a client widget. Depending on the configuration, users can highlight a term in a reporting tool, click **Shift+F5**, and pull up the definition in the glossary.

5. *Business Process Management (BPM)*
   Offerings include IBM Business Process Manager and Pega Business Process Management. In addition, the Eclipse open-source framework includes a plug-in for Business Process Model and Notation (BPMN). There are two key integration points between BPM and data governance tools:

   - Data governance workflows—Data governance tools should leverage BPM capabilities to enforce data governance workflows. As part of these workflows, specific roles may be involved in changes to data governance artifacts, including business terms, data policies, business rules, and code values. Later in this report, Figure 11 shows an example of a data governance workflow.

   - Mapping data policies and standards to business processes—BPM tools support business processes, but they are not well integrated with data governance processes. As shown in Figure 6, IBM Business Process Manager contains a simple workflow to on-board corporate banking customers. The data governance team should use the business glossary to document key data policies and standards for various activities in the business process. For example, a data policy for activity 5 might state that when a company announces the acquisition of another company, the legal hierarchy must be updated at the time of the announcement as opposed to the effective date of the acquisition.
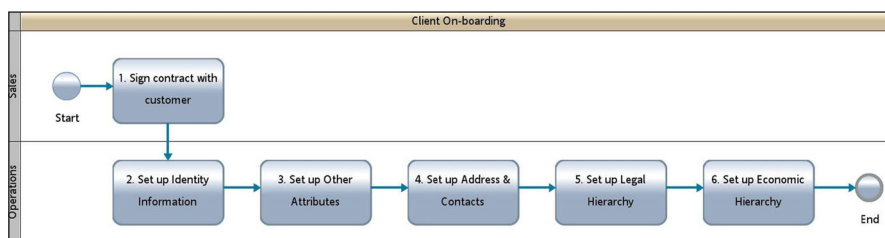
Figure 6: EDM reference architecture includes six highlighted categories for data governance.

6. *Data Security and Privacy*
   This category includes a number of subcategories, including data masking, data tokenization, database encryption, and database monitoring. Offerings include IBM InfoSphere Guardium, IBM InfoSphere Optim Data Privacy, Imperva, Informatica Data Masking, and Protegrity. A number of integration points exist between these products and data governance tools:

- Document security and privacy policies—Data stewards should use data governance tools to document data policies relating to data security and privacy. These policies may cover data ownership, user access, user revalidation, and the definition of sensitive data by jurisdiction.

- Discover hidden sensitive data—Once sensitive data has been defined, data discovery tools can discover hidden sensitive data. For example, Social Security Numbers may be hidden in a field called **EMP_NUM**. The data governance team should then flag this data in the metadata repository.

- Enforce security and privacy policies—Security and privacy tools should enforce the policies that have been documented within the data governance tools.

7. *Information Lifecycle Management*
Information lifecycle management is a process and methodology for managing information through its lifecycle, from creation through disposal, including compliance with legal, regulatory, and privacy requirements. Information lifecycle management encompasses multiple subcategories: information archiving, records and retention management, eDiscovery, and test data management. This category is often rolled up into a class called "Information Governance" tools. Although there is a lot of discussion, there are limited examples of integration between information lifecycle management and data governance tools. However, the emergence of big data with a focus on semi-structured and unstructured data will start to converge these two categories.

8. *Cloud*
Certain vendors, including Collibra, Informatica, Orchestra Networks, SAS, and Trillium, also offer cloud versions of their data governance tools. This reduces the upfront cost and effort to derive value from a data governance program.

Evaluation Criteria for Data Governance Software Tools

Here is a comprehensive set of evaluation criteria for data governance software tools:

1. *Cost*
Organizations should consider the following cost buckets when evaluating data governance tools:

- Software license cost—Organizations should also consider vendors with cloud-based offerings to reduce upfront capital expenditures.

- Implementation cost.

- Ongoing cost to maintain the tool.

- Ongoing cost to use the tool, including time allocation from data stewards and business analysts.

Figure 7 (below) provides an example of financial impact analysis for data governance.

2. *Ease of Use*

Data governance tools are meant to be used by business users. It goes without saying that usage will suffer if the tool is hard to use. There are several ease-of-use features. For example, some vendors have adopted a "Facebook-style" user interface and social collaboration features (see Figure 8).

Another feature is particularly interesting. Consider that you have two Dodd-Frank–related terms in banking:

- ''Depository Institution Holding Company'' means a bank holding company or a savings-and-loan holding company that is organized in the United States, including any bank or savings-and-loan holding company that is owned or controlled by a foreign organization, but does not include the foreign organization.

- "Savings and Loan Holding Company" means any company that directly or indirectly controls a savings association or that controls any other company that is a savings-and-loan holding company.
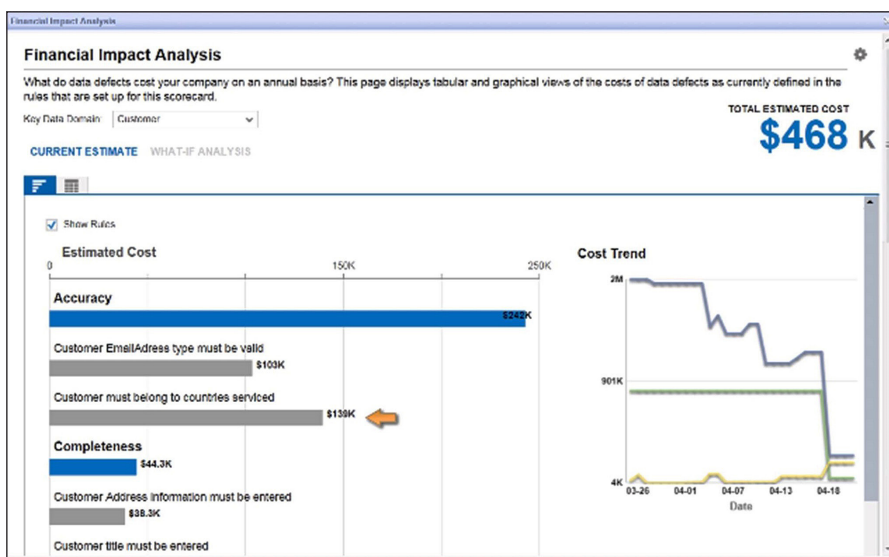
Figure 7: SAP Information Steward provides a financial impact analysis for email address.
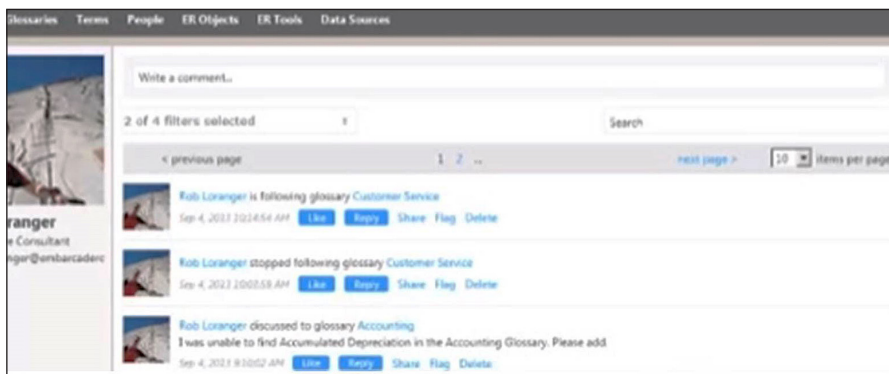
Figure 8: Embarcadero CONNECT is a business glossary with a Facebook-style user interface

As you will note, the second term is embedded in the definition of the first term. As a result, the two terms need to be linked. This can be done manually, but it can become a real challenge when dealing with thousands of business terms in the glossary. Any tool that can auto-link these terms will offer tremendous advantages in terms of productivity of the data stewards.

3. *Custom Attributes*
Data governance tools ship with out-of-the-box (OOTB) attributes such as "Definition," "Short Description," "Long Description," and "Example." However, tools should also allow administrators to create custom attributes, such as "Security Classification" and "Critical Data Element" (see Figure 9).

4. *Custom Relationships*
Most data governance tools ship with OOTB relationships such as "Synonym" and "Acronym." Other OOTB relationships could be "Replaces/Replaced By," indicating that "Vendor" replaces "Supplier" and "Supplier" is replaced by "Vendor." Additional OOTB relationships may include "Assigned Assets" or "Implemented By" to associate a business term such as "Vendor Number" with the database column **SUPP_NUM**. Finally, "Allowed Values" can link a business term or column name to the underlying reference data.

Figure 9: IBM InfoSphere Business Glossary allows administrators to create custom attributes.

Figure 10: Data stewardship interface in Orchestra Networks EBX.

Data governance tools should support custom relationships. For example, the term "Exposure" may have a "Regulated By" relationship to the "Dodd-Frank" regulation. On the other hand, the Dodd-Frank regulation should have a "Regulates" relationship to the term "Exposure." This allows data stewards to view all the data artifacts that are regulated by a given regulation.

5. *Data Stewardship*
   Data governance tools should support stewardship over data artifacts. This means that administrators should be able to add and delete data stewards. Data stewards should also be able to manage categories, business terms, code lists, code values, data policies, data standards, data quality rules, data quality metrics, data issues, master data rules, and other data artifacts (see Figure 10 on previous page).

6. *Custom Roles*
   Data governance tools typically ship with OOTB roles such as "Steward." However, every organization has its own data governance vernacular. Accordingly, data governance tools should also support custom roles, such as "Data Owner," Data Executive," "Data Sponsor," "Stakeholder," "Subject Matter Expert," "Responsible," "Accountable," "Consulted," and "Informed."

7. *Approval Workflows*
   Data governance tools should support OOTB and custom workflows for different processes, such as adding a business term, adding a code value, resolving a data issue, and adding a business rule. For example, a large bank wanted to add **KO** (Kosovo) to its list of countries. However, Kosovo was not on the list of ISO 3166-1 alpha-2 country codes. The Enterprise Data Management team developed the following workflow:

   - Regional Finance Steward (Europe) will propose the change.

   - Global Finance Steward will approve the change.

   - Enterprise Data Management will add the new entry to the code table for Country.

   Figure 11 shows a BPMN workflow that can be imported into a data governance tool.
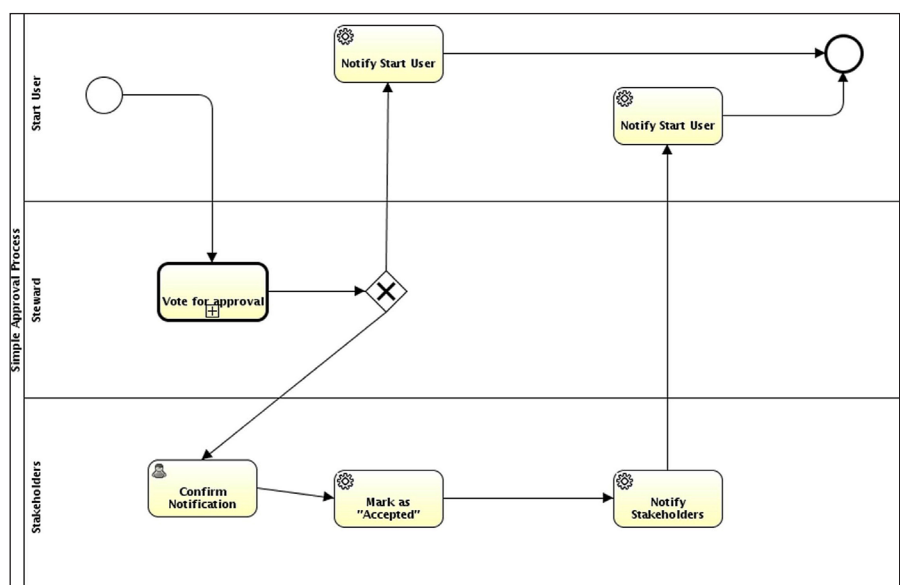
Figure 11: Simple approval workflow in Collibra Data Governance Center.

8. *Data Policies, Standards, and Processes*
    Data governance organizations need to maintain a hierarchy of data policies, standards, and processes. These artifacts need to be reviewed and approved by the data governance council. These data policies, standards, and processes may be maintained in Microsoft Word or PDF format and stored on Microsoft SharePoint. These artifacts may also be maintained within GRC tools, such as EMC RSA Archer GRC or IBM OpenPages GRC. However, it is preferable to manage the artifacts within data governance tools.

    Some organizations may elect to maintain their data policies, standards, and processes in custom repositories but provide a hyperlink from within the data governance tool. Figure 12 shows a sample hierarchy. A data policy for data ownership relates to one or more data standards, including one for data roles. The data roles standard relates to one or more data processes, including one for data stewardship meetings. The data governance tool should support this hierarchy of data policies, standards, and processes.

9. *Data Lineage and Impact Analysis*
    Business users often have to answer questions such as "Where did this data come from?" and "Where is it going?" and "What happens to it along the way?" and "What is the impact if we drop this column?" Organizations need a strong metadata foundation to be able to answer these questions. Metadata tools need to ingest technical metadata from a variety of data sources, including data modeling tools, relational databases, COBOL copybooks, ETL tools, reporting tools, and Hadoop. Figure 13 shows an example of a data lineage report.
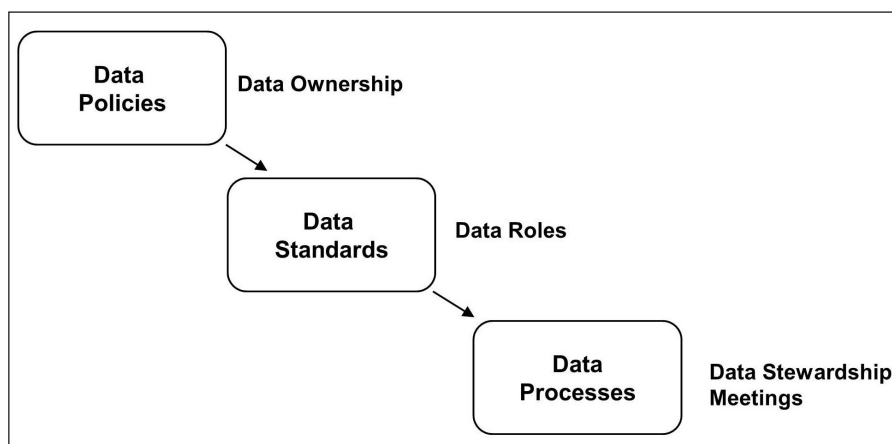

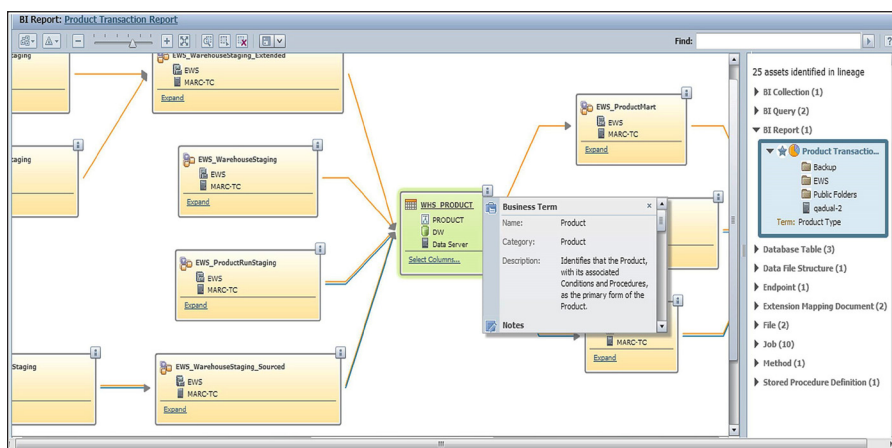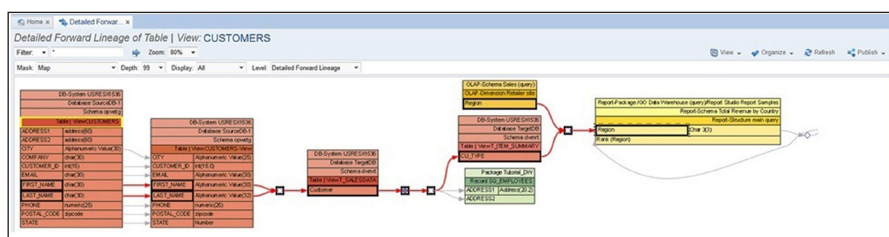**Figure 12:** Hierarchy of data policies, standards, and processes.

Figure 13: Data lineage report in IBM InfoSphere Metadata Workbench.

Figure 14 (below) shows a detailed forward lineage report for the **CUSTOMERS** table.

10. *Hierarchy of Data Artifacts*
    Data governance tools need to support a hierarchy of data artifacts. In the example in Figure 15, an insurance carrier has a policy that says that IT cannot change rating attributes without approval from the underwriting department. A rating attribute is a data attribute that can affect insurance premiums. Examples of rating attributes may include date of birth, gender, smoking status, and city of residence. A child policy deals specifically with data quality for National Change of Address. An associated data rule validates the state against a list of states based on the ISO 3166-2:US standard. The data rule relates to the business term "State." Finally, the business term "State" links to the **ADDRESS** table and to the "State" code table. The data governance tool should support the management of this hierarchy of data artifacts.

11. *Data Issue Resolution Process*
    Data governance tools should provide capabilities to document, assign, reassign, track, close out, and report on data issues. Data governance tools should also support OOTB and custom workflows to deal with data issues. As shown in Figure 16, these tools should also support email notifications so that users become aware when they need to perform a task as part of the data issue resolution workflow. Data discovery and data quality tools should also be able to handle data exceptions in a systematic manner.

Figure 14: Detailed forward lineage report for the CUSTOMERS table in ASG-Rochade.
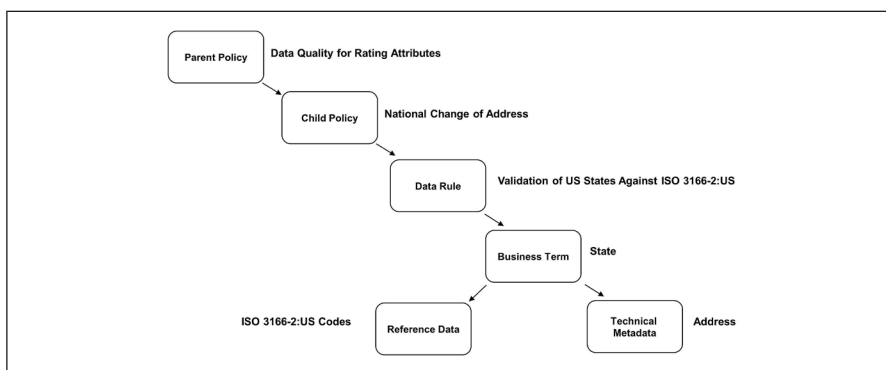
Figure 15: Hierarchy of data artifacts at an insurance carrier.
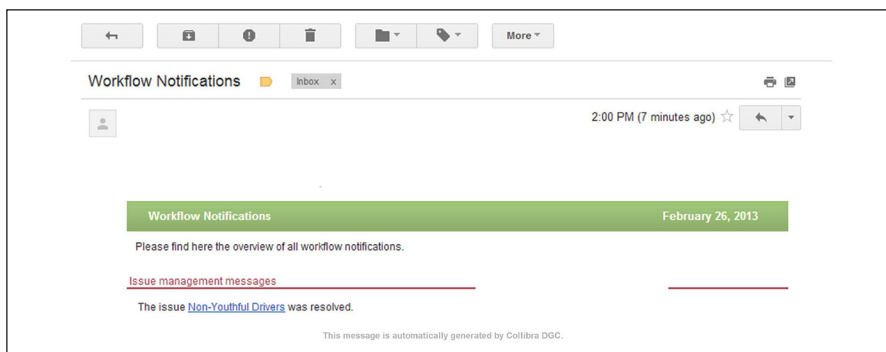
Figure 16: Automated email notification from Collibra Data Governance Center.

In addition, some organizations have repurposed issue management tools such as BMC Remedy and HP Quality Center to also handle data exceptions.

12. Support for Internal Audit

Many large institutions, especially those in the financial services industry, are subject to scrutiny from regulators regarding their data management practices. These institutions need to demonstrate that they have the appropriate practices to improve the trustworthiness of their data. A best practice is for these institutions to use Internal Audit as the "first line of defense" with the regulators. The data governance teams should document their policies, standards, and processes in the tool. They should then request Internal Audit to report on the adherence to these policies, standards, and processes. The Internal Audit review may focus on some of these questions:

- Data ownership— "Does every data repository have a data owner?" (Figure 17 shows an example of a data ownership matrix.)

- Data dictionary— "Does the business glossary contain a data dictionary for each repository?"

- Security—"Is the data owner accountable for determining who has access to each data repository?" "How often are users revalidated?"

- Business rules— "Does the busi- ness glossary contain business rules for each repository?" (Figure 18 provides an example of a business rules editor.)

Figure 17: Data ownership of enterprise applications in Data Advantage Group MetaCenter.

Figure 18: Rule editor for postal codes in SAP Information Steward.

13. Data Governance Metrics

Data governance teams need to establish metrics to monitor the ongoing performance of the program. Data governance tools should provide the following metrics on an overall basis and for different views, including by data steward, data owner, data repository, application, and data domain.

- Business Glossary—Number of candidate terms, number pending approval, and number approved.

- Reference Data—Number of candidate code values, number pending approval, and number approved.

- Data Issues—Number of outstanding data issues and number resolved in the previous period.

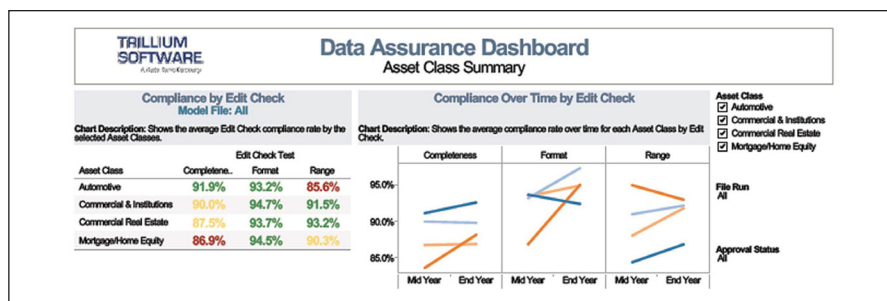- Data Quality Scorecard—Data Quality Index by application and by critical data element (see Figure 19).

Figure 19: Trillium Data Assurance Dashboard exposed in Tableau.

Vendor Landscape

In this section, we review a list of notable vendors for data governance software tools.

Adaptive

Adaptive offers the following data governance capabilities:

- Business glossary—Adaptive Business Glossary Manager offers a business-oriented user interface to manage data artifacts.

- Metadata—Adaptive Metadata Manager supports data lineage and impact analysis.

ASG

ASG has a very strong footprint in the metadata marketplace and offers the following data governance capabilities:

- Metadata—ASG-Rochade is the company's enterprise metadata repository with a vast array of scanners for third-party applications.

- Business glossary—ASG-metaGlossary supports governance over data semantics in a business-oriented manner.

Collibra

Collibra is a high-growth startup focused exclusively on data governance. Collibra Data Governance Center offers the following capabilities either directly or via integrations with third parties:

- Business glossary—Collibra Business Semantics Glossary has a slick user interface that is targeted at business users. The business glossary offers a highly customizable approach to semantic modeling for business terms, policies, rules, and other data artifacts.

- Reference data management—Collibra Reference Data Accelerator is integrated with the business glossary and offers the ability to do mappings between code sets and business terms.

- Data stewardship—Collibra Data Stewardship Manager offers the ability to assign and manage data issues.

- Workflows—Collibra Data Governance Center ships with OOTB workflows that can also be customized using BPMN.

- Metadata—Collibra does not ship with a metadata repository today. However, the company is focused on integrating with key metadata repositories. For example, a large financial institution allows users to author business terms in Collibra and then leverages the REST API to move the terms into a third-party metadata repository.

- Data discovery and data quality—Collibra does not ship with a data discovery and data quality tool. However, Collibra and Harte-Hanks Trillium announced an integration so that users can run data profiling jobs in Trillium TS Discovery and view the results in Collibra.

INFORMATION ASSET, LLC

Data Advantage Group

Data Advantage Group is a company that is exclusively focused on metadata. Data Advantage Group MetaCenter offers the following capabilities:

- Metadata—MetaCenter has strong scanner capabilities based on its ActiveLinx technologies for real-time metadata from third-party applications. ActiveLinx scanners exist for many data repositories, including data modeling tools, relational databases, COBOL, ETL, reporting tools, and Hadoop.

- Business glossary—MetaCenter ships with a business glossary.

- Reference data—MetaCenter supports the creation and maintenance of reference data.

- Workflows—The latest version of MetaCenter includes support for workflows to manage business terms and reference data.

Embarcadero Technologies

Embarcadero has a strong footprint within the data modeling community with ER/Studio. The company recently announced CONNECT, which provides a business glossary with a "Facebook-style" user interface, social enterprise collaboration, and integration with data models.

Harte-Hanks Trillium

Trillium has a strong heritage in data quality. The company offers the following capabilities either directly or through third-party integrations:

- Data discovery and profiling—TS Discovery supports manual and automated data profiling and discovery.

- Data quality—TS Quality offers data cleansing and standardization, address verification and correction, and de-duplication.

- Business glossary and reference data—Trillium offers these capabilities through integration with Collibra Data Governance Center.

IBM

IBM has a robust end-to-end platform for data governance with the following capabilities:

- Data discovery—IBM InfoSphere Information Analyzer and IBM InfoSphere Discovery offer manual and automated data profiling capabilities. Data rules generated in Information Analyzer can be stored in the Metadata Workbench and are accessible from the Business Glossary.

- Data quality—IBM InfoSphere QualityStage offers data validation and cleansing capabilities and is integrated with IBM InfoSphere DataStage from an ETL perspective.

- Business glossary—IBM InfoSphere Business Glossary supports the governance of business terms and policies.

- Metadata—IBM InfoSphere Metadata Workbench includes scanners for a number of IBM platforms, such as DB2, Cognos, DataStage, and Information Analyzer, as well as third-party applications. We anticipate that IBM InfoSphere Metadata Workbench will add support for Hadoop distributions, including IBM InfoSphere BigInsights.

- Information policy management—IBM InfoSphere Business Glossary supports policy management.

- Reference data management—IBM InfoSphere MDM Reference Data Management Hub supports reference data mapping.

- Information lifecycle management—IBM InfoSphere Discovery allows the discovery of a complete business object such as "customer" that can then be archived with IBM InfoSphere Optim Data Growth Solution. Market reports indicate that IBM is also integrating Business Glossary and Optim from a policy enforcement perspective.

- Security and privacy—IBM InfoSphere Discovery supports the discovery of hidden sensitive data. We also anticipate that IBM will integrate IBM InfoSphere Business Glossary and IBM InfoSphere Guardium to manage and enforce security policies.

Informatica

Informatica has a robust data governance platform that includes the following capabilities:

- Data discovery—Informatica Data Explorer is a data profiling tool to enable data stewards and business analysts to discover anomalies in all kinds of data, including Hadoop.

- Data quality—Informatica Data Quality (IDQ) supports matching and global address cleansing. We have observed growing market momentum around IDQ, especially relating to its exception-handling capabilities.

- Metadata—Informatica Metadata Manager is a key feature of Informatica PowerCenter Advanced Edition. The tool supports metadata management and data lineage.

- Business glossary—Informatica Business Glossary is a key feature of Informatica PowerCenter Advanced Edition. Market reports indicate that Informatica is getting ready for a major new release of its business glossary with a slick user interface and data stewardship capabilities.

- Big data—Informatica has made major investments to bridge the huge installed base of Power-Center skills to Hadoop. As a result, developers can develop in the PowerCenter environment and then deploy their data integration jobs to Hadoop without having to learn the intricacies of MapReduce, Pig, and Hive. Informatica also supports data profiling, masking, and cleansing on Hadoop.

Orchestra Networks

Orchestra Networks is a fast-growing company with European roots but a growing U.S. presence. Orchestra Networks EBX is a multi-domain MDM and reference data management platform. EBX has strong hierarchy management capabilities as well as embedded workflows for master data and reference data. EBX has established a strong beachhead of reference data customers. We anticipate that Orchestra Networks will also release a business glossary to govern business terms, business rules, and column definitions to support the broader EBX platform.

SAP

SAP offers a robust data governance solution for organizations with a strong investment in the SAP platform. SAP Information Steward coupled with SAP Data Services provides data stewards with capabilities for data profiling, rules management, de-duplication, data cleansing, data quality scorecards, financial impact analysis, business definitions, and metadata. SAP's large installed base automatically catapults these tools onto the shortlist for organizations looking to operationalize data governance in a business-focused manner.

SAS

SAS has now fully folded its DataFlux division back into the mother ship. Following this reorganization, SAS has also consolidated its enterprise data management tooling under the SAS Data Management platform. SAS Data Management provides consolidated capabilities for data integration, data profiling, data quality, data monitoring, business glossary, metadata, master data management, reference data management, and entity resolution.

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*. For more information, please visit www.information-asset.com.