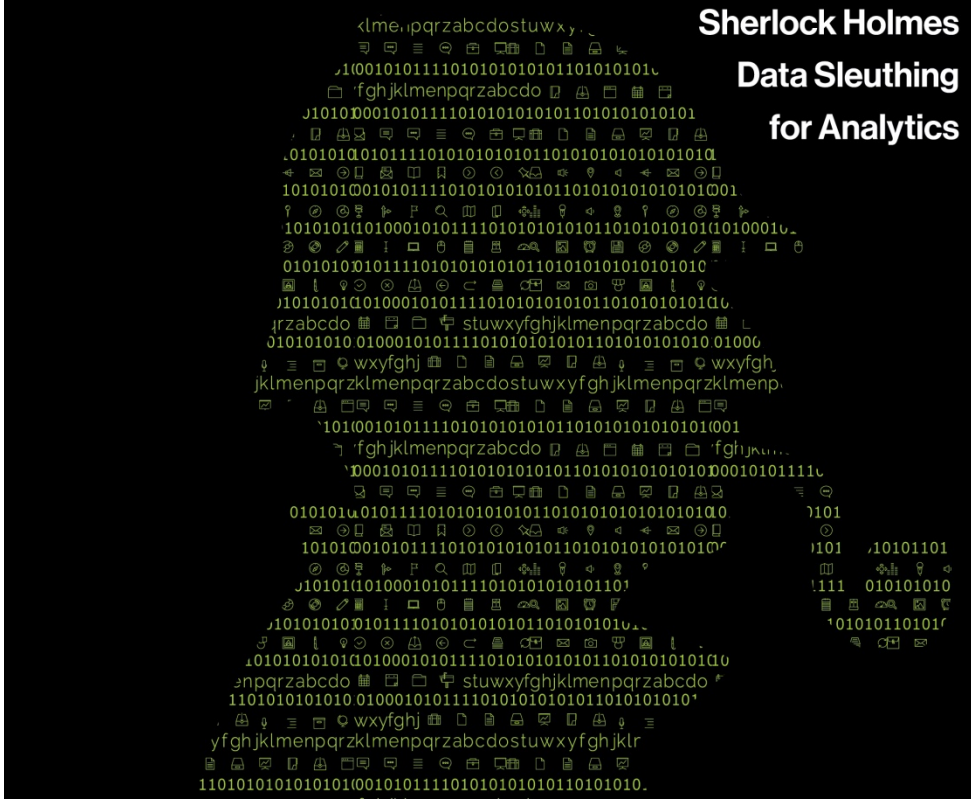
Bonnie K. O'Neil and Lowell Fryman

The Data Catalog

Sherlock Holmes Data Sleuthing for Analytics



The Data Catalog

Sherlock Holmes Data Sleuthing for Analytics

Bonnie K. O'Neil

Lowell Fryman

Technics Publications

Published by:

2 Lindsley Road, Basking Ridge, NJ 07920 USA
https://www.TechnicsPub.com

Edited by Jessica McCurdy-Crooks, cover design by Lorena Molinari

First Printing 2020

The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

Approved for Public Release; Distribution Unlimited. Public Release Case Numbers 19-3379, 19-3611, 19-3710, 19-3772, 19-3802, 19-3706, 19-3838, 19-3793, 19-3794, 19-3836, 19-3837, 19-3850, 19-3836, 19-3842, 19-3852, 19-3844

ISBN, print ed.	9781634627870
ISBN, Kindle ed.	9781634627887
ISBN, ePub ed.	9781634627894
ISBN, PDF ed.	9781634627900

Library of Congress Control Number: 2020931206

CHAPTER 7

One-Stop Shopping

The emphasis in this book is on the importance of a data catalog that is enterprise-wide, enabling searches that span silos. Such a catalog bridges these silos and makes data all over the enterprise discoverable, perhaps for the very first time. Some of the vendors and industry analysts call this the discovery of "dark data"—data that exists, but nobody knows it is there.

An enterprise data catalog, therefore, needs to:

- Support the scanning of large volumes of data
- Ingest a diversity of data types, formats (structured or unstructured) and platforms (on-premises, cloud, or hybrid cloud)
- Make this data understandable to both technical and non-technical users
- Facilitate searches across the diverse data types using ML
- Automate curation with ML

It seems like a natural expansion of functionality that an enterprise data catalog would offer more than just a data inventory but extend its reach to encompass enterprise data management functionality as well. The catalog's curation functionality blends well with data governance, as we saw in Chapter 5. Most of the major data management disciplines involve data governance, such as:

- Data quality, overseeing quality improvements and rules, ensuring that it increases over time

- Reference data and Master Data Management (MDM), putting standards and rules in place for master and reference consistency throughout the enterprise

- Data Integration, ensuring data is properly matched with similar data in other systems in the enterprise when it is brought together for summaries, ensuring that summaries are accurate

It could, therefore, be argued that a true enterprise data catalog is more than just a catalog. It expands and serves all the major data management disciplines in one unified platform.

Both IBM and Informatica initially offered specific data management products and tools, and then expanded their respective offerings by acquisition and in-house development. They both now offer complete data management platforms, which include a data catalog at their center. This chapter presents examples of both to illustrate expansive enterprise data catalog features, focusing on each one in turn then interspersed to spotlight different capabilities. We are calling this class of tools the One-Stop Shop.

Enterprise data catalogs

Most data catalog tools have evolved from other data-centric tools that were designed for a specific use case or solution. Although tools of this nature are often limited by their use case focus area and therefore difficult to add features, IBM and Informatica are exceptions.

An enterprise data catalog must be able to bridge silos and span across many different environments and sources. The "One-Stop Shop" catalog products do this. Both vendors have worked hard to ingest disparate sources that are not native to their respective platforms. Figure 7-1 from Informatica illustrates the various disparate sources and platforms that need to be supported by an enterprise catalog in order to be effective. The due diligence you perform when

researching catalog products for purchase must include a checklist for infrastructure and source types within your environment.

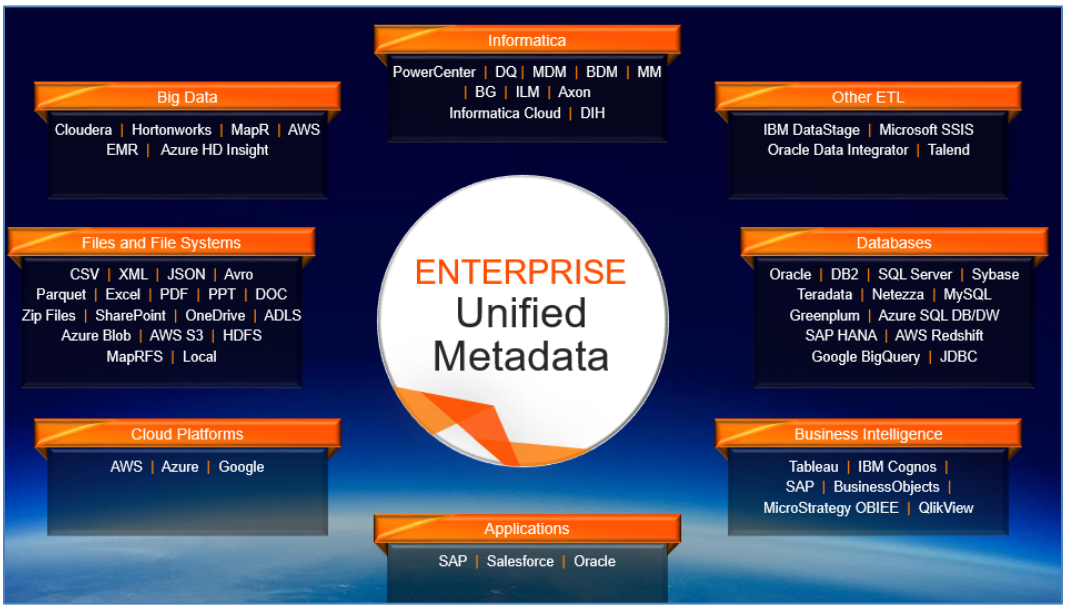


Figure 7-1. Types of sources and platforms in Informatica

There are two considerations for infrastructure: what the catalog runs on and what the ingest sources run on. An enterprise data catalog must be able to access a wide variety of data sources and also scale. See Figure 7-2 from Informatica, illustrating both diversity of ingest platforms and the CLAIRE platform on a high-performance parallel architecture (Hadoop cluster).

A prospective buyer needs to keep this in mind and determine not just the sources that will be ingested but also the infrastructure upon which the catalog resides. Catalog performance is dependent upon its underlying architecture, facilitating its ability to scan to profile data sources and perform ML algorithms in real-time.

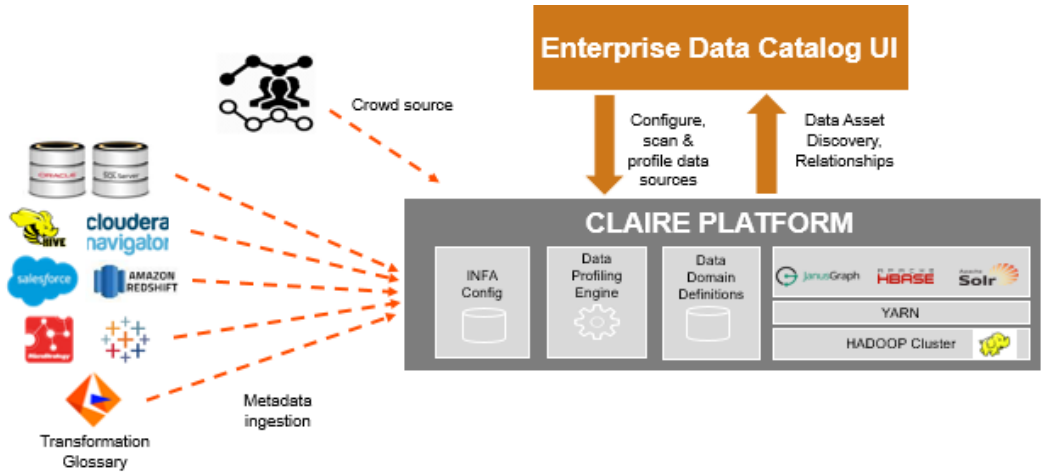


Figure 7-2. Enterprise data catalog architecture

Example: IBM

IBM created one of the first relational databases, DB2, and before that, IBM was known for mainframes and mainframe databases such as Information Management System (IMS). C.J. Date was one of the founders of relational theory, and he hailed from IBM. Then IBM acquired and built an entire suite of database management products under the InfoSphere umbrella, including the acquisition of Ascential bringing DataStage, ProfileStage, QualityStage and others; Master Data Management, including an in-house tool and the acquisition of Initiate; and others, including Information Governance, workflow, and Business Glossary. IBM introduced the Information Governance Catalog (IGC), a data catalog product with its main focus being data governance.

The integration of these capabilities has been challenging for IBM, but they have finally succeeded with many of them shown in Figure 7-3.

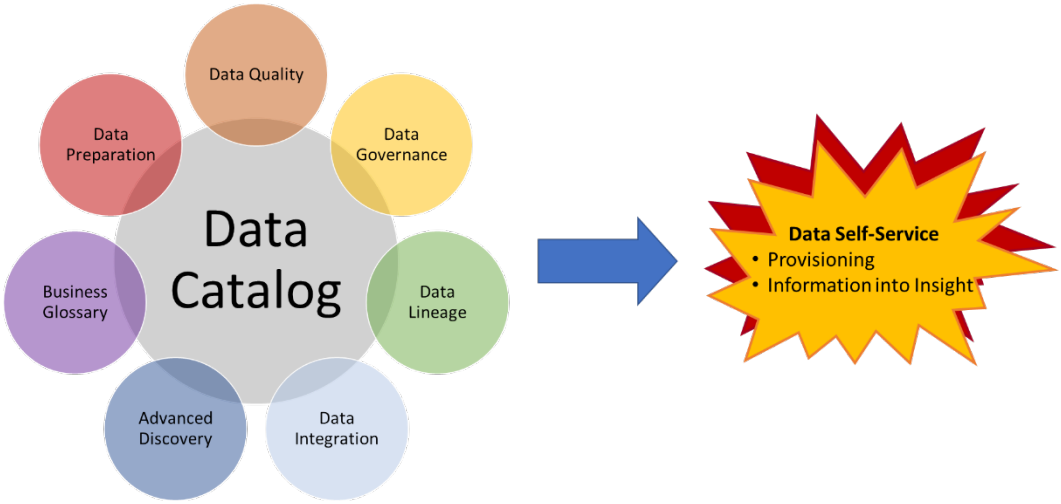


Figure 7-3. Integrated data catalog

IBM added its powerful AI engine Watson to this platform. You'll recall that Watson was the famous AI question-answering computer that won the *Jeopardy* TV game show's grand prize of $1 Million in 2011.[25] The Watson AI engine has been incorporated into many different products and services. IBM offered Watson Knowledge Catalog (WKC), an ML-empowered data catalog that was initially not integrated with the rest of the InfoSphere data management products. Bringing WKC together with IGC and the InfoSphere suite resulted in an enterprise data catalog and one-stop shop for data management. This is a fully integrated data platform called Cloud Pak for Data. Watson's technology is employed not just in the data catalog but also Watson's data prep (Watson Applications), Watson Studio, Watson Machine Learning, and Watson OpenScale. The Cloud Pak for Data platform encompasses all the data management capabilities embedded in the InfoSphere product line coupled with the Watson products, enabling data self-service to drastically reduce time to insight. See Figure 7-4.

[25] https://tek.io/2GzWiJk.

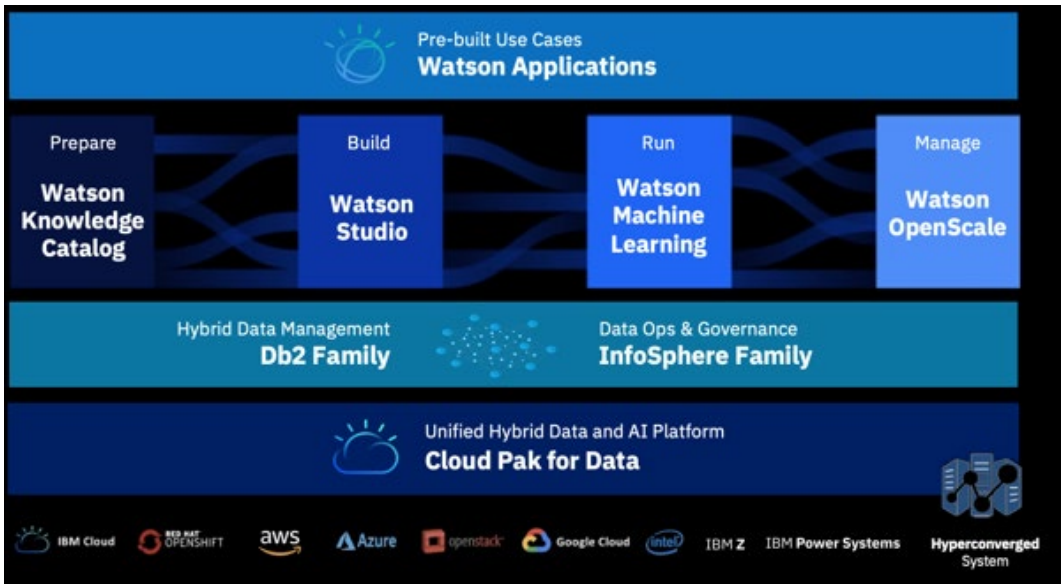


Figure 7-4. IBM Cloud Pak for Data

Figure 7-5 shows a deeper dive into IBM's many capabilities and components, showing five core operations driving a business-ready data pipeline:

- Metadata curation services: Watson Knowledge Catalog
- Metadata Management: Infosphere augmented with Watson ML
- Self-services interaction: catalog functions enabling collaboration and data citizen curation: Watson Knowledge Catalog
- Core governance and master data management: Information Governance Catalog plus IBM Master Data Management
- Machine learning and automation: Watson ML

The InfoSphere suite brings the data management functions of data quality, MDM, data integration, and data governance, together with Watson and IGC all into one platform. This complete platform delivers the Business Ready Supply Chain. See Figure 7-6.

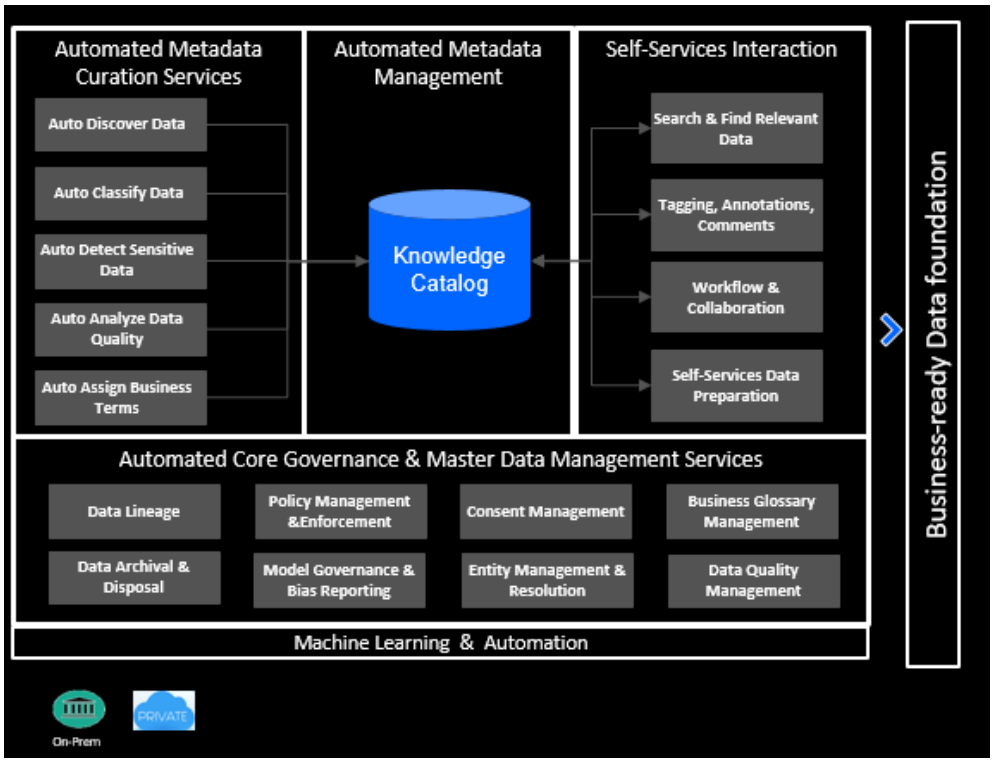


Figure 7-5. IBM capabilities and components

Enterprise Data Integration	Enterprise Data Quality	Enterprise Data Governance	Enterprise Data Consumption
IBM provides the most comprehensive, most integrated, and most scalable data integration platform provides the core data integration backbone running in the largest banks, telcos, retailers, insurance companies, etc.	IBM provides the most complete, most scalable, and most integrated data quality platform supporting data profiling; creation, execution, and monitoring of data validation rules and data matching and consolidation.	IBM provides an enterprise data governance platform that supports business and technical users on data governance teams. It eliminates the cost and complexity of integrating a stand-alone data governance platforms with data integration, quality, and consumption tooling.	IBM provides an enterprise data catalog that delivers self-service capabilities for data citizens (business analysts, data scientists) to search and explore information, to preview and refine information to act upon information in a secure environment driving new data science or analytics.

Figure 7-6. IBM's Business Ready Supply Chain

Example: Informatica

Informatica has been on a very similar adventure. Known in the early days of data warehousing for its extremely powerful ETL suite of tools (PowerCenter), Informatica was the first vendor to introduce an ETL product that had a server, scheduler, workflow, and lineage support. It was a natural extension for Informatica to produce graphical lineage from their ETL jobs and packages, and they released various iterations of a metadata tool. Today, like IBM, they offer a comprehensive data management platform but as a smorgasbord of coordinated plug-and-play components, bringing a powerful integrated solution. See Figure 7-7.

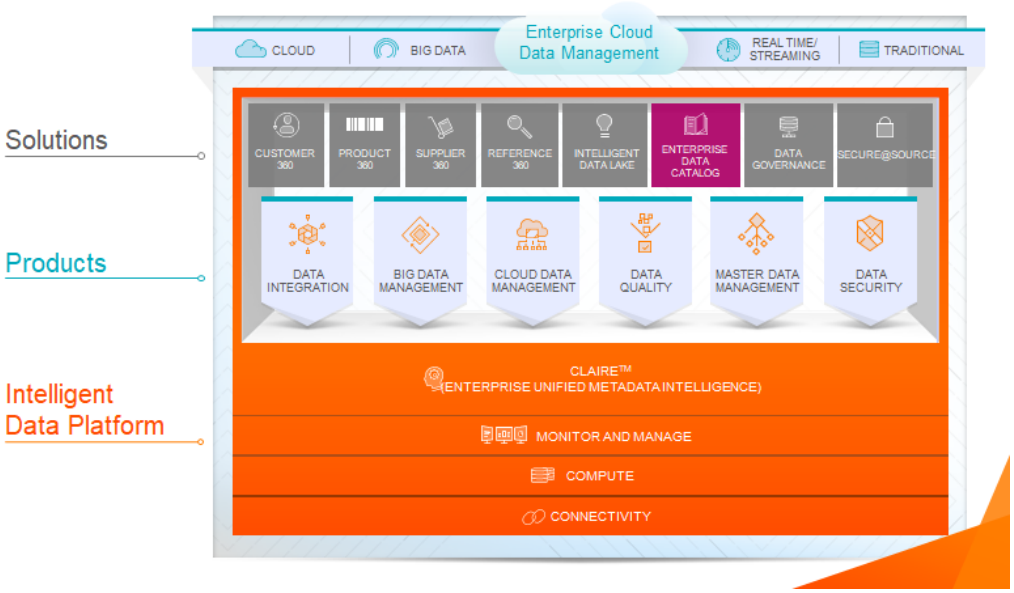


Figure 7-7. Informatica's suite of products

Informatica has CLAIRE, mentioned above, with AI in the middle of its name: The AI engine empowering ML to assist both curation and search.

Informatica's integrated suite of products includes:

- Business Glossary
- Data Quality
- Data Profiling
- Data Integration
- Big Data support
- Master/Reference Data Management
- Data governance, including workflow
- Security

Figure 7-8 shows a more detailed look into how the various components are integrated into the platform to deliver a complete data management solution.

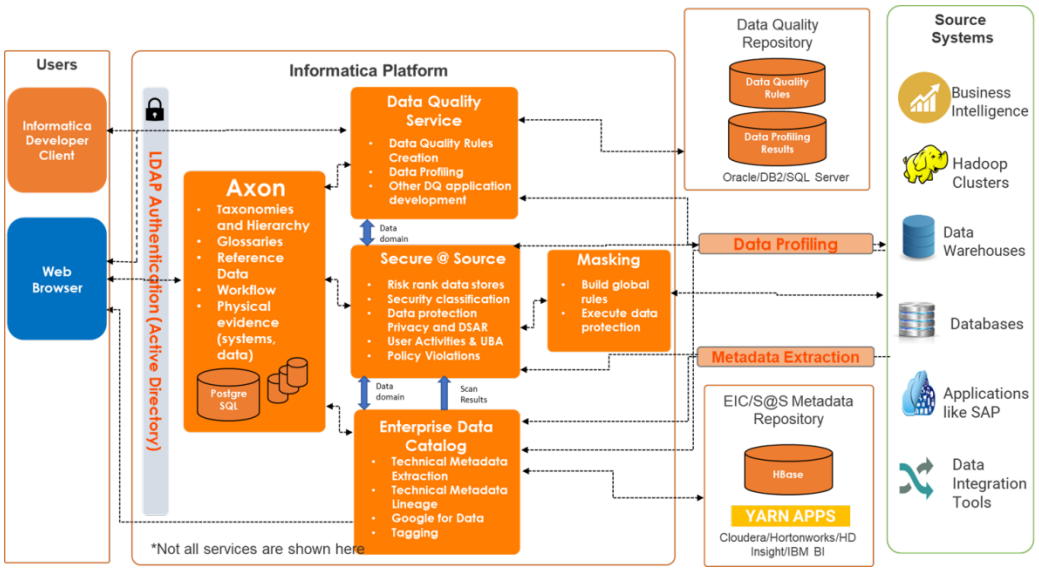


Figure 7-8. Informatica components: deep dive

The way Informatica offers their products differs from IBM. IBM sells them as a bundle, Cloud Pak for Data, and Informatica offers them as separate components. The latter's philosophy is that customers should be able to choose the components they want.

It is obvious that both products would offer strong integration with their own tools. Data lineage and visibility into transformations within the Informatica PowerCenter has always been strong. See Figure 7-9, which shows a portion of a data flow. The expression used in the output field is displayed by selecting the icon shown in Figure 7-9. See Chapter 9 for a detailed discussion of data lineage. Data integration is best displayed as lineage diagrams.

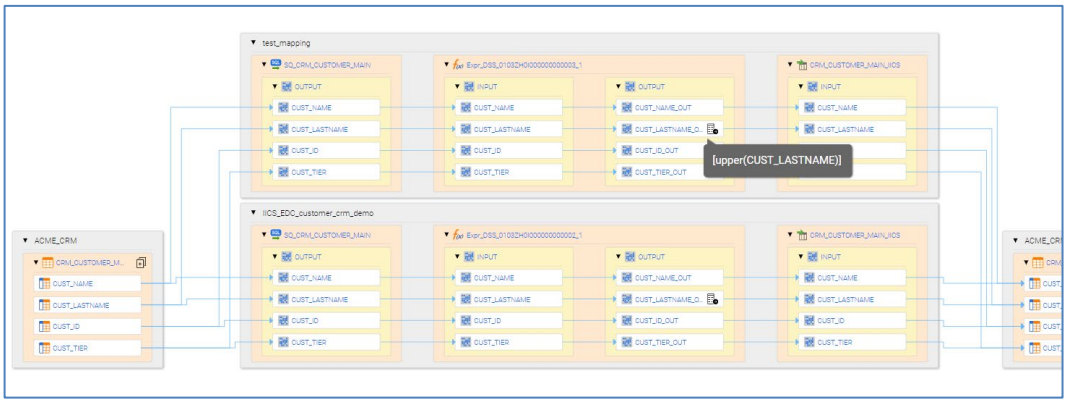


Figure 7-9. Lineage depicted as a data flow

Reference data support

Reference data is used to organize or categorize other data. Often, reference data includes code sets that serve as shorthand for longer named elements. They are extremely helpful, especially for relational databases, because they avoid redundancy and typographical errors. A reference data table contains the codes, their values, and descriptions. An application that requires the data specified in the code set would only use the code, and a join in the database would bring the description to the requesting application. The beauty of this system is that codes are maintained in one central place.

Figure 7-10 from Informatica shows the types of reference data that can exist in an organization.

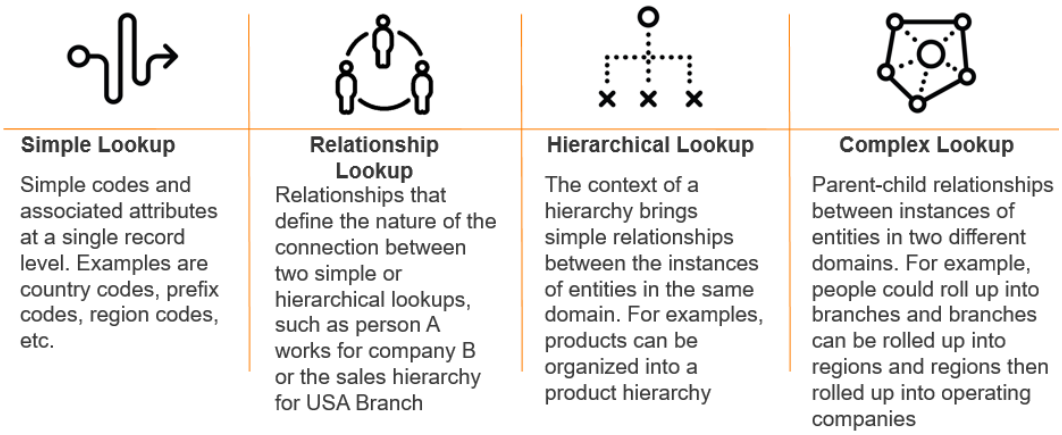


Figure 7-10. Types of reference data

Issues occur, however, when different applications use different sets of codes for the same data. For example, suppose an enterprise tracks projects, and each department has a different set of codes that represent the status of a project. Here's the Sales Department's code set:

Table 7-1. Sales Department Project Codes

Code	Value	Description
100	Proposed	The project has been scoped and proposed to management.
200	Pending	The project has been submitted and is waiting for approval.
300	Approved	The project has been approved.
400	Active	The project is underway.
500	Canceled	The project has been canceled.
600	Completed	The project has been completed.

Here's the IT Department's code set:

Table 7-2. IT Department Project Codes

Code	Value	Description
I	Initiated	The project is in the proposal phase.
U	Under Review	The proposal is under review.
S	Submitted for Approval	The project has been submitted for approval.
P	Pending	The project is on hold.
A	Approved	The project has been approved.
D	Denied	The project has been rejected and approval is not granted.
O	Active	The project is active and ongoing.
X	Canceled	The project has been canceled.
C	Completed	The project has been completed.

Notice that these code lists are different from each other and may not map easily. The organization would need to determine how they want to report on projects enterprise-wide.

Figure 7-11 shows a graphic from Informatica illustrating the mapping problem from three different systems, and three different code sets. The tool will maintain these mappings to the enterprise codes. Fortunately, standards bodies exist to create standardized lists of codes that many organizations use. For example, the International Standards Organization (ISO) maintains several lists of geographic codes such as countries in the world. They maintain 3166 Country Codes[26] and several varieties based on length and whether subdivisions (such as US States) are included. However, even when using these lists, you must ensure you specify the issuing standards organization.

[26] https://www.iso.org/iso-3166-country-codes.html.

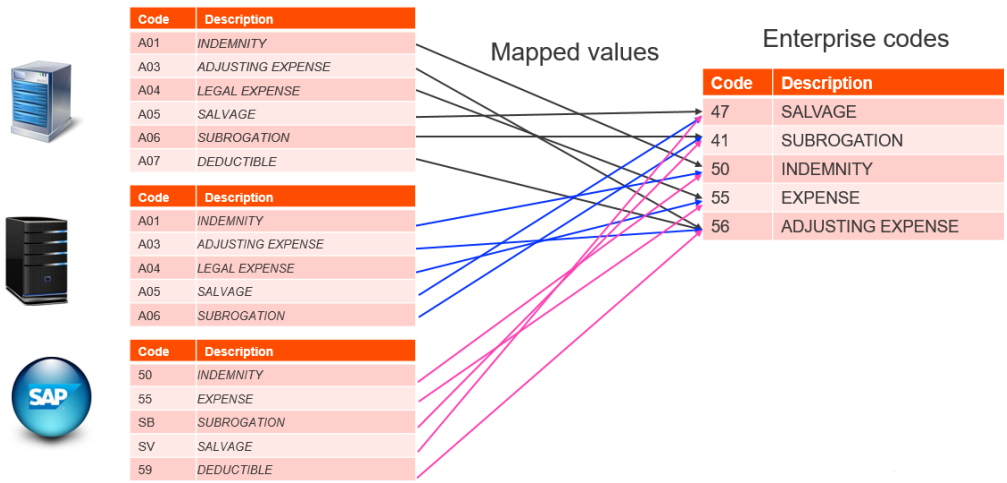


Figure 7-11. Reference data mapping

It can be an interesting problem when different standards bodies maintain the same type of reference data. At least three standards bodies maintain "standard" lists of airport codes:

- International Air Transport Association (IATA)
- International Civil Aviation Organization (ICAO)
- Federal Aviation Administration (FAA)

The FAA maintains only airport codes for airports in the United States, so it contains no international codes. Sometimes the three code sets will agree, and sometimes they differ. Harmonization of reference data is important for enterprise data management. Data catalog tools help maintain and provide visibility into reference data mappings.

One-Stop Shop tools provide a window into reference data management and governance. Reference data in IBM WKC can be either imported from a Comma Separated Values (CSV) file or entered manually. WKC provides a simple drag and drop interface, very familiar to Excel users, see Figure 7-12. You drag and

drop the fields into the three categories of metadata: Code, Value and Description, just as you would in an Excel file import.

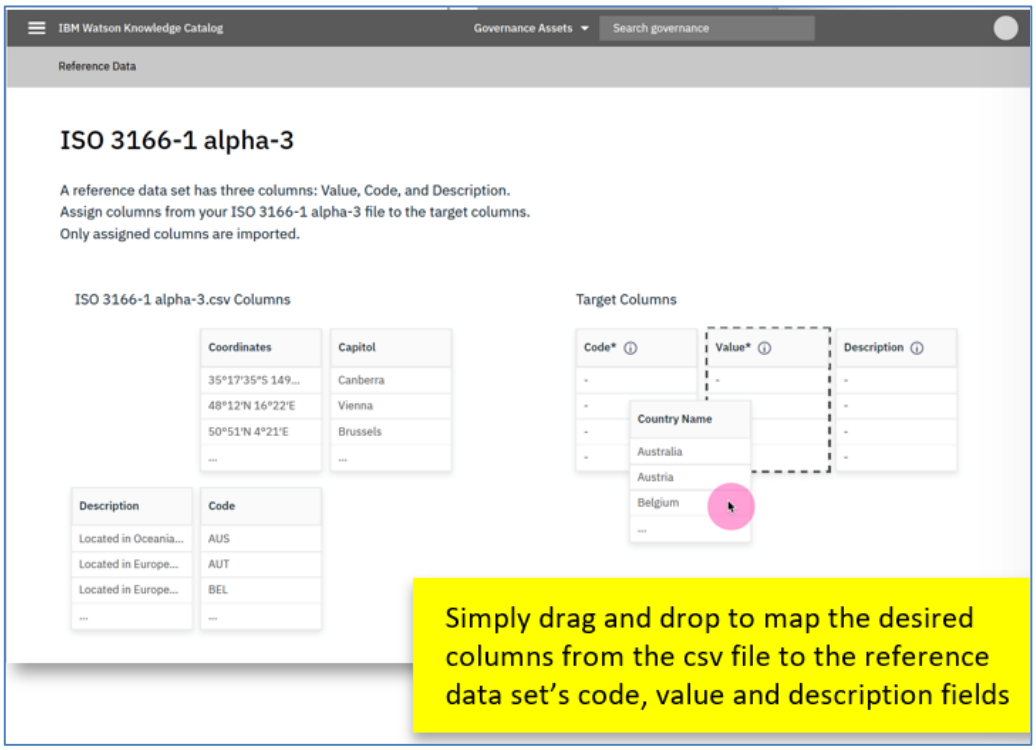


Figure 7-12. Drag and drop reference data fields

You can then see the results of the import in Figure 7-13.

Reference data can get very complicated. Some reference data is in a hierarchical form, such as the medical diagnostic code set International Statistical Classification of Diseases (ICD-10). Other examples are product hierarchies and organizational charts. Informatica provides support for hierarchical code sets with versioning and collaboration, see Figure 7-14.

Code	Value	Description
AUS	Australia	Located in Oceania.
AUT	Austria	Located in Europe
BEL	Belgium	Located in Europe
CAN	Canada	Located in North America
CHN	China	Located in Asia
DNK	Denmark	Located in Europe
FIN	Finland	Located in Europe
FRA	France	Located in Europe
GHA	Ghana	Located in Africa
MAR	Morocco	Located in Africa

Items per page: 10 | 1 - 10 of 40 items

Figure 7-13. Reference data imported

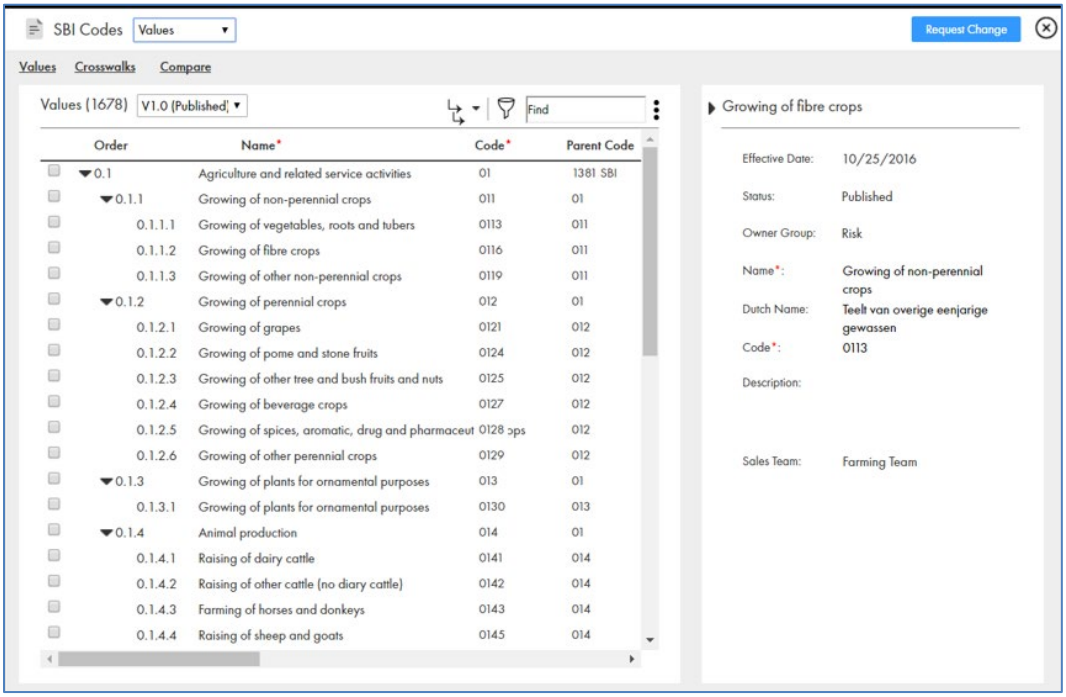


Figure 7-14. Hierarchical code set

Business glossary

IBM has a well-developed Business Glossary component, with both searching and editing term capabilities. The workflow follows the author's Governance Lite™[27] methodology which is a lightweight governance framework specially designed for glossary terms. It is reactive governance, allowing users to enter terms or modify descriptions that are then sent to the data steward for approval. A search for terms will show the status of the term, see Figure 7-15. This search revealed a term called Alternate Service Identifier, which shows a status of Inactive, highlighted in pink.

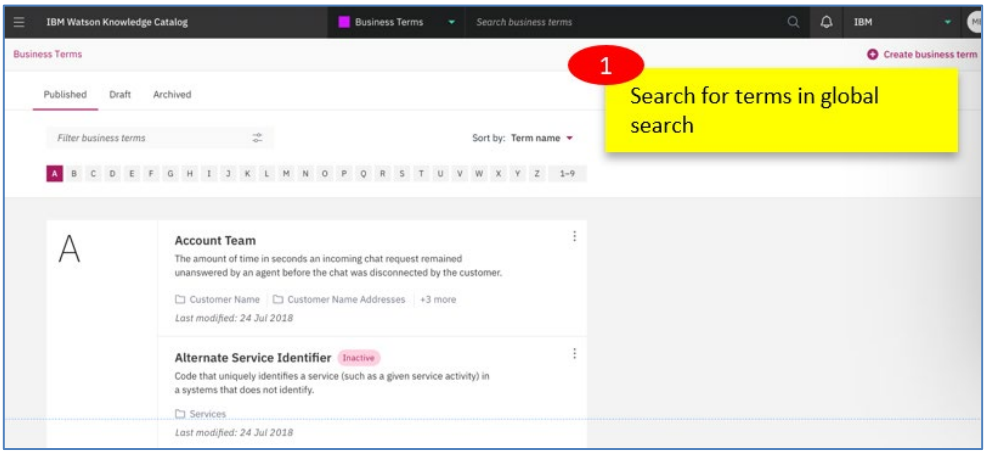


Figure 7-15. Glossary search

The user can edit the description or create a new term (see the "Create business term" at the top right corner) in Figure 7-15. The status of the term changes to "Draft" shown in blue right after the term name when an edit is made, see Figure 7-16. The user can then submit it for approval (upper right box) or delete their changes. The submittal activates the term governance workflow by notifying the appropriate data steward.

---

[27] Governance Lite™ was introduced in this online newsletter: https://bit.ly/2Gu7bMy.

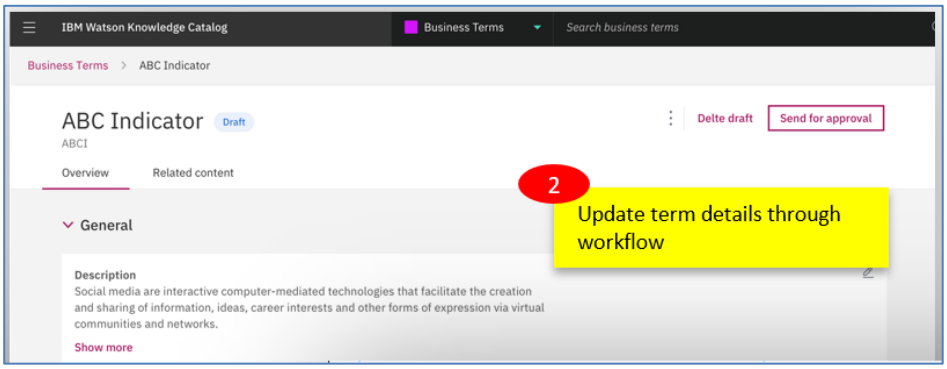


Figure 7-16. Update a term

Term relationships

Figure 7-17 shows the user's ability to specify relationships between terms, including the ability to have type hierarchies. This is very powerful and can greatly assist searches by retrieving related assets. It not only shows that a term is related but also how they are related. Elaborate hierarchies can be established.

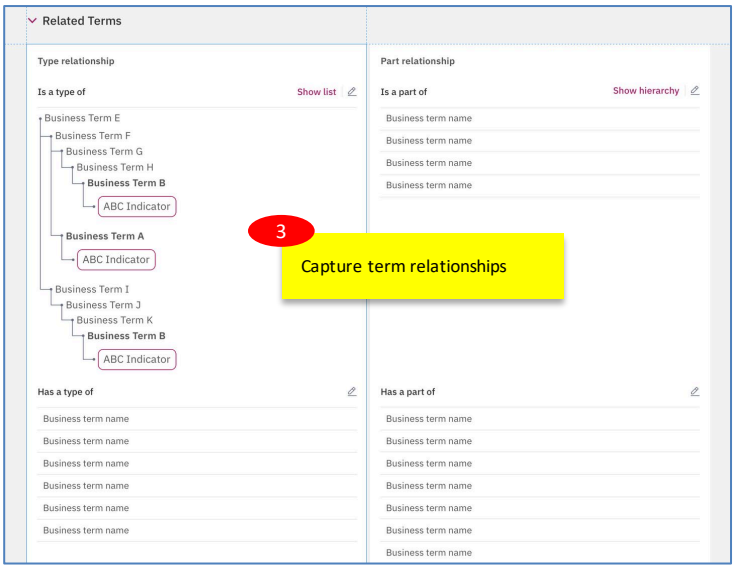


Figure 7-17. Term relationships

Data quality support

Both vendors offer integration with their data quality products.

Figure 7-18 shows the start of a data quality investigation in Informatica. The user surveys the CRM_CUSTOMER_MAIN table associated with the business term "Customer." Notice the green "seal of approval" next to the table name, indicating that it is a certified data set.

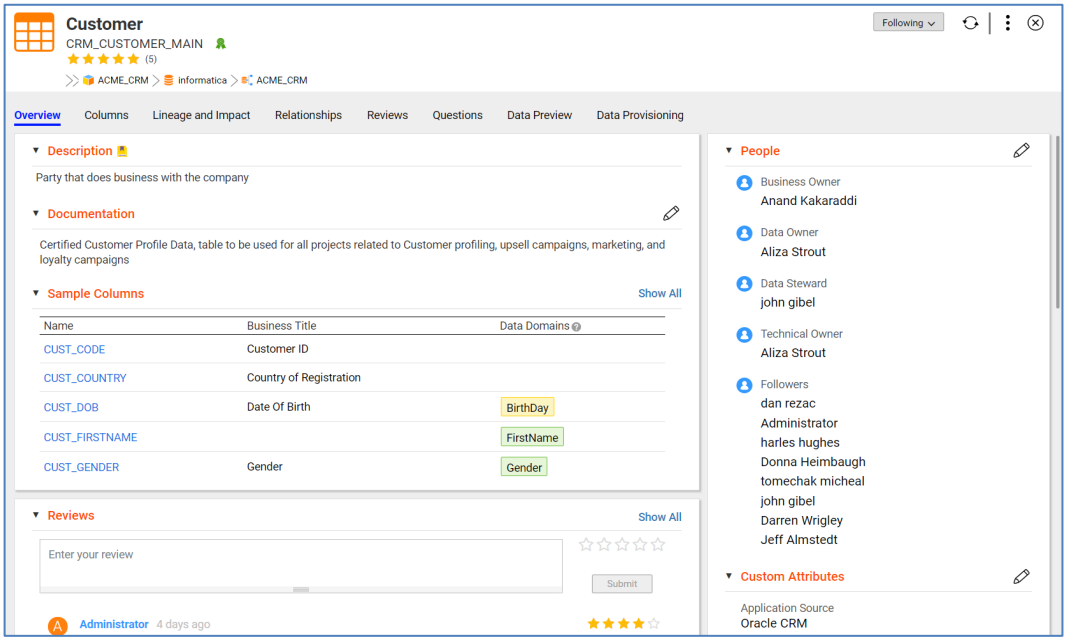


Figure 7-18. Customer table overview screen

The user clicks on the CUST_COUNTRY column name, which brings up the next screen showing the overview for this column, see Figure 7-19. The panel on the right shows information about the table to which this column belongs, including the Application Source, the business description ("Country of Origin"), the business unit ("Sales and Marketing"), and even the load frequency (Weekly).

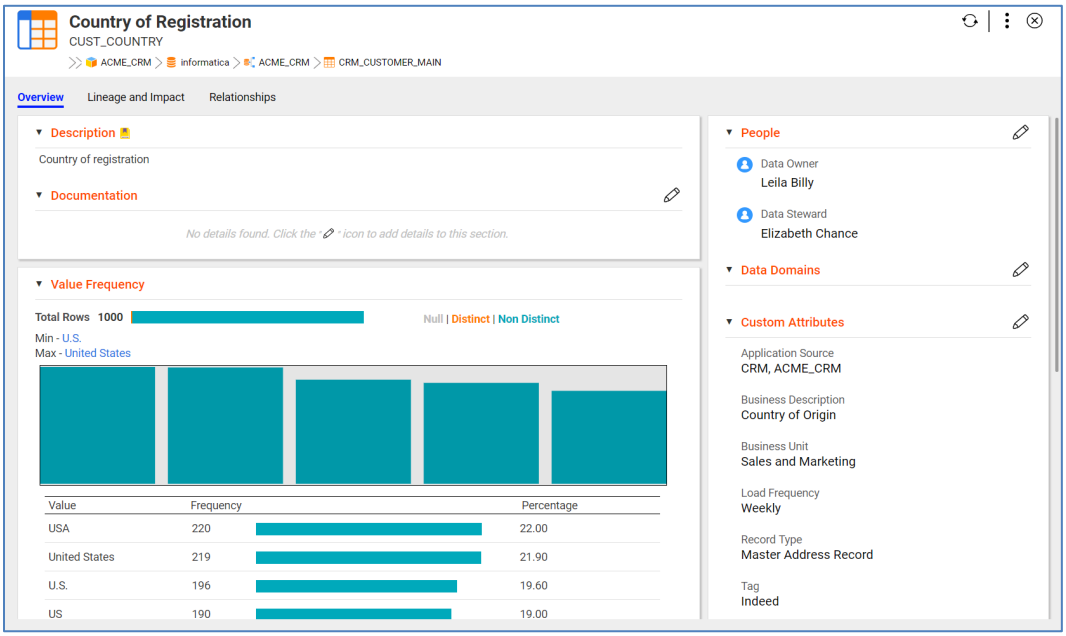


Figure 7-19. Country column overview

The screen in Figure 7-20 provides some profiling details to give a clue as to its data quality (or lack thereof). There are 1000 rows in the table, and the histogram shows that many values are recurring, which is normal. What is not normal are the various ways that the United States has been represented. Based on the minimum and maximum values, it appears that all the records in the table are for Customers in the United States. The highest count value is the string "USA," occurring 220 times. The user scrolls down, revealing more detailed information, shown in Figure 7-20.

Notice all the different ways that the "United States" has been entered into this column, and all of them have fairly high counts, meaning there is no format standardization. This is also revealed in the pattern counts, shown to the right. There are 175 rows with the pattern "X.X.X.," which matches the value "U.S.A." in the histogram pane. The importance to the analyst is that they would have to perform standardization and cleansing on this field to have the values appear in

the same format in the resultant data set. The analyst also knows that there is no reason to create a filter for selecting only customers in the United States because there are no international customers in this data set. However, if there were, the analyst would have to cleanse the data and unify it to a single format in order to perform the filter.

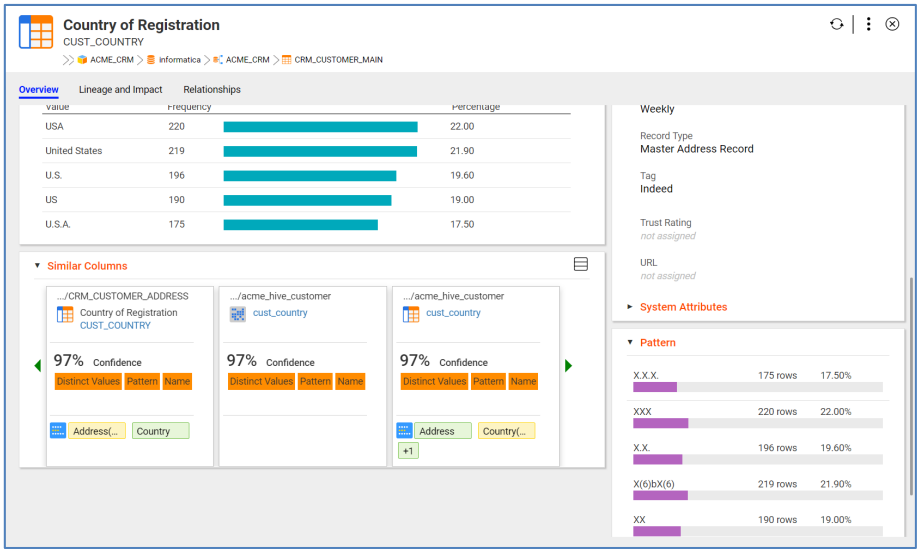


Figure 7-20. Column details for data quality inspection

This screen also contains a very helpful pane called "Similar Columns." This helps in Reference Data Management discussed above, to facilitate discovery of reference data candidates to standardize a value like Country across the entire enterprise. It also helps the analyst locate potential data sets to join with their data that might have descriptive information to enrich their final study.

Data quality visualization

IBM provides visualizations without having to leave the catalog environment. Figure 7-21 shows a different data set with many diverse countries. The pie chart helps to visualize which countries have the most entries.

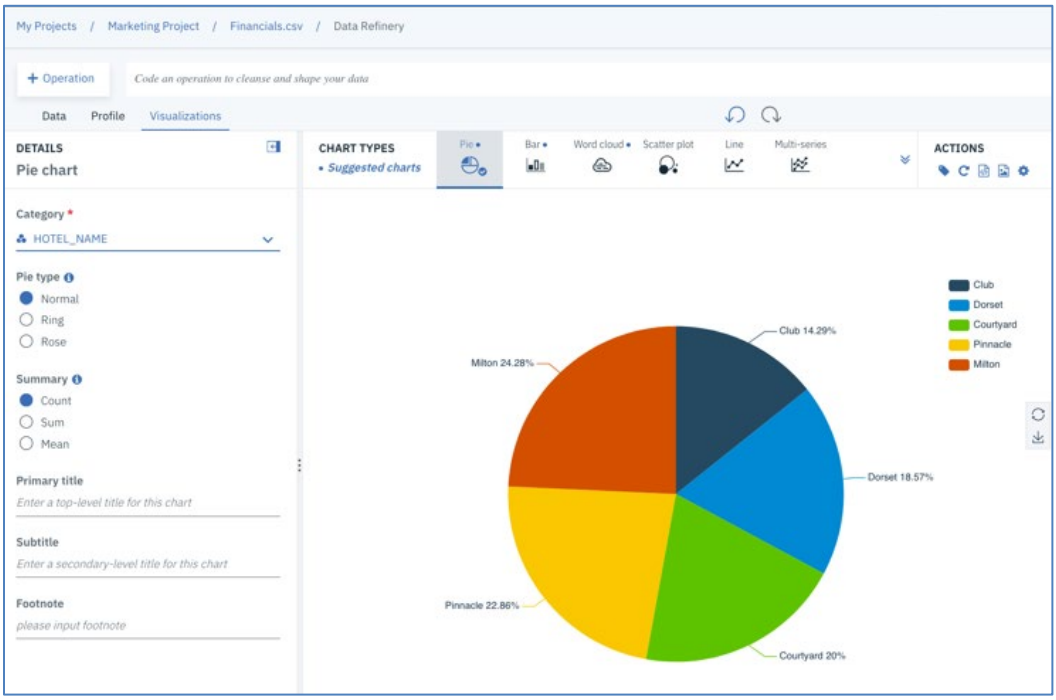


Figure 7-21. Visualization inside the catalog

Data quality rule enforcement

You can perform data quality enforcement with rules and display them and their statuses in the data catalog. Figure 7-22 shows the status of automated rules in Informatica. The red, yellow, and green help to pinpoint potential problems.

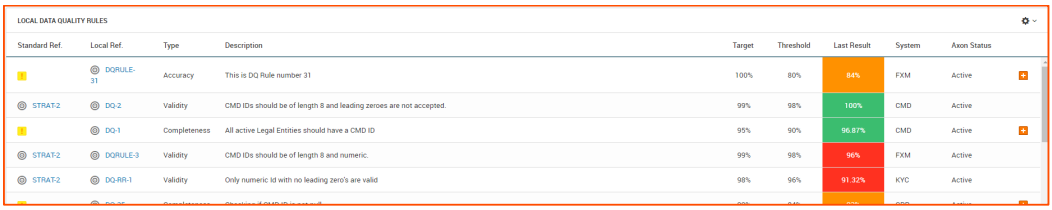


Figure 7-22. Rule status in Informatica

Informatica can also display rule enforcement and data quality levels in systems and business processes. See Figure 7-23. Notice the dimensions of data quality that are measured: Validity, Completeness, Accuracy, and Timely.

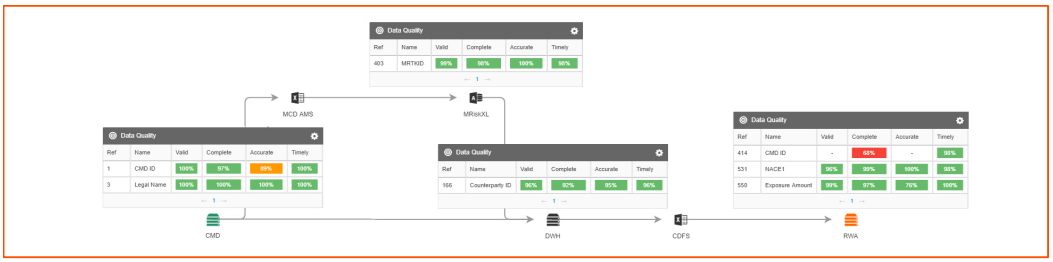


Figure 7-23. Rule enforcement and quality levels

Data quality can also be measured for data assets over time, see Figure 7-24 for an example.

Figure 7-24. Data quality over time

Policies and rules

Policies are laws, regulations, and corporate mandates dictating the proper handling of data. IBM provides a hierarchical structure of Policies, meaning that Policies can have Sub-Policies. Rules can be associated with Policies, adding the actual enforcement of the Policy. An example of a Policy could be that customer addresses must be valid and must be in the same country as stated in the field titled "Country of Residence." There could be two rules that enforce this policy: One that calls address-validation software, and a specialized rule that matches

the Customer Country field with the Country of Residence in another data set. Figure 7-25 shows the Policy hierarchy in the left pane and the descriptive information. Note that categories can be assigned to Policies, although this one doesn't have any categories. This policy is a draft and is clearly "under construction." Note the Draft status, shown in blue text name to the name. The user can either delete the draft or sent it for approval. The two choices are shown in red at the top right. The hierarchy shows that this Policy will be a larger policy with sub-policies underneath it.

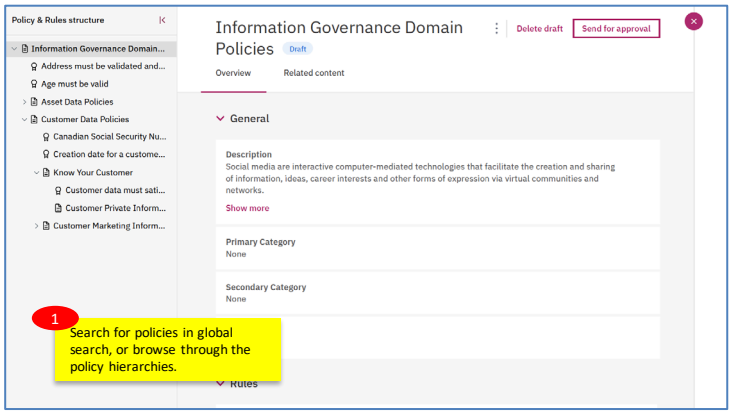


Figure 7-25. IBM policies

Figure 7-26 shows a rule called "Address In-Country," which probably validates the country field described above. It shows that it is performing "automated enforcement" and the 1,366 times when this was done in March 2019. It even shows that the number of enforcements is up 47.36% from last month. It shows that 100% of the data is anonymized and no access is granted. The number of policy enforcements is tracked over time in the chart for the last month.

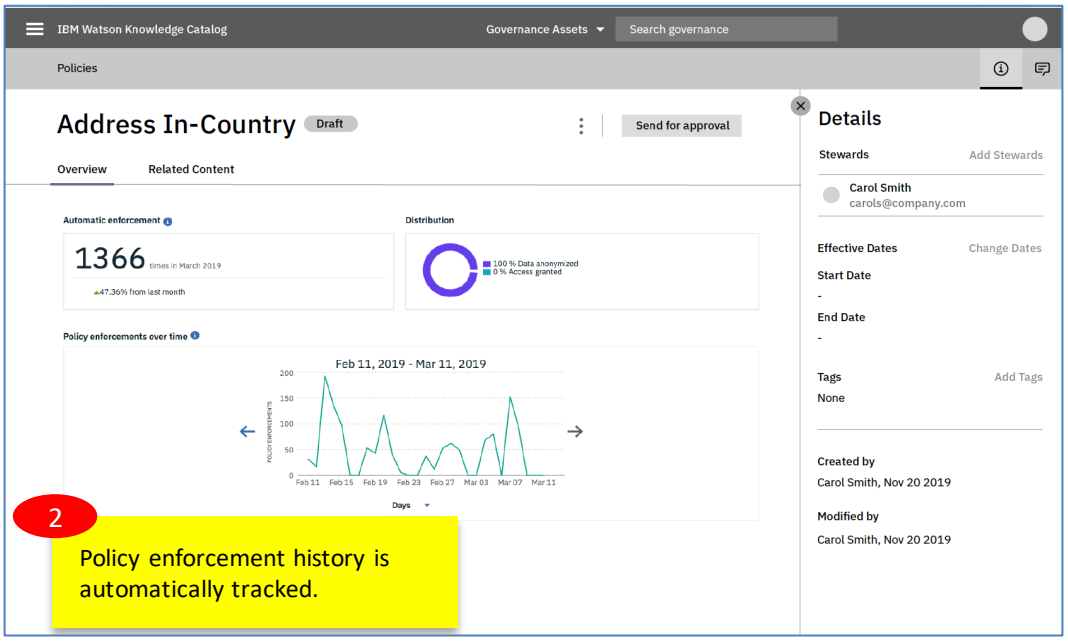


Figure 7-26. Address in-country rule

Workflow

Both vendors offer complete data governance suites in addition to data quality functionality. IBM uses the RACI Matrix to assign governance roles to assets and workflows. Figure 7-27 shows the request for approval (1), the tracking of progress (2), and the ability to add comments (3). This allows users to add questions, asking for more clarification or providing a rationale for their decisions. One of the roles of data governance is protecting sensitive data. This was discussed in Chapter 5, but for our purposes here, it is pertinent to point out that this is one of the functions offered by the One-Stop Shop integrated catalogs. They provide data profiling to display granular, descriptive technical metadata supplying clues about the contents of the data, but only when allowed and appropriate.

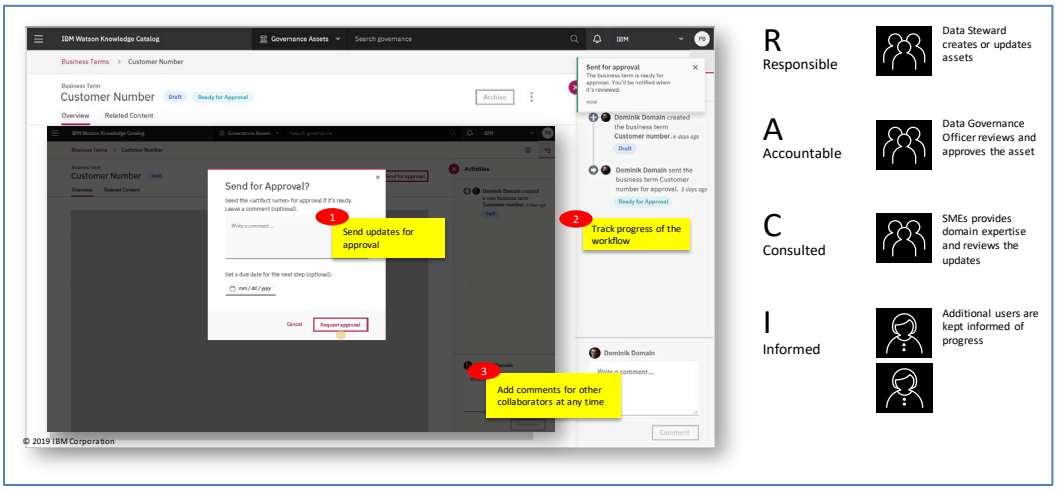


Figure 7-27. RACI governance workflow

Figure 7-28 illustrates a screen from IBM showing a data profile but masking sensitive columns of profiles that are not allowed to be shown. Notice the shields indicating that the profiles for the two columns in the figure are unavailable. The importance of data governance in data catalog implementations was discussed in Chapter 5.

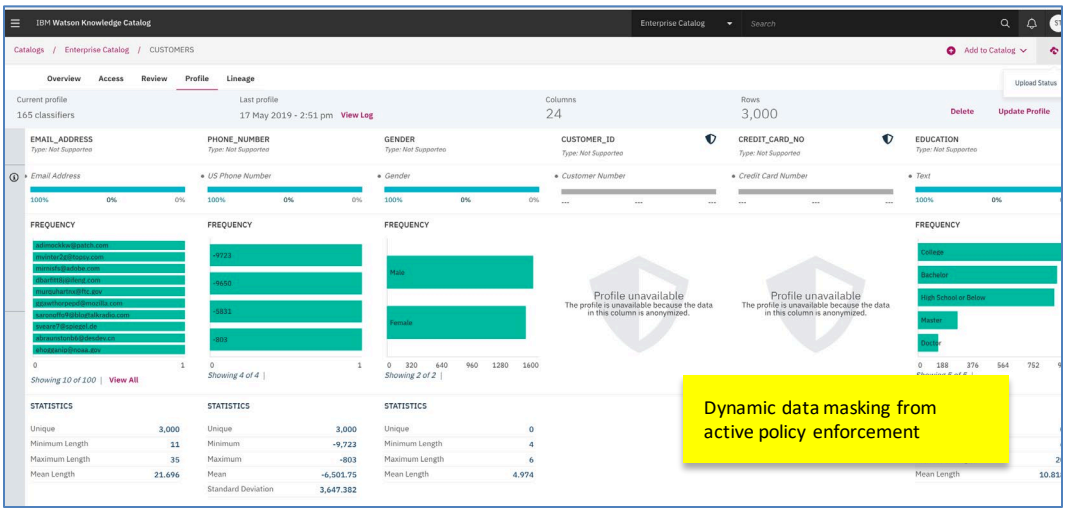


Figure 7-28. Masking sensitive columns

Key points

This chapter focused on the integrated enterprise data catalog, bringing together many different data management capabilities featuring a data catalog at the center. It is important that an enterprise data catalog has functionality in many, if not all the main data management disciplines to enable data management maturity, facilitate curation, and provide trusted data to analysts. Two examples of this kind of catalog are those provided by IBM and Informatica, each offering a vast array of data management components. This chapter highlighted each data management discipline, providing examples of each from these two vendors.