IBM DataOps
White Paper

Cognitive Data Governance

Powered by Machine Learning to find and use governed data

Jo Ramos
Distinguished Engineer & Director – IBM Analytics

Rakesh Ranjan
Program Director & Data Scientist – IBM Analytics

Contents.

IBM Corporation
Route 100
Somers, NY 10589

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Introduction

The purpose of this white paper is to discuss how machine learning and deep learning techniques can impact the way companies implement metadata discovery and business term assignments using the IBM DataOps platform.

First, a set of definitions to ensure consistent understanding of the topic:

Definitions

Master Data: the consistent and uniform set of identifiers and extended attributes that describe the core entities of an enterprise, such as existing or prospective customers, products, services, employees, vendors, suppliers, hierarchies and the charts of accounts
Machine Learning: an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed
Data Governance: the overall management of data availability, relevancy, usability, integrity and security in an enterprise
Regulatory Compliance: an organization's adherence to laws, regulations, guidelines and specifications relevant to its business

Most organizations spend a great deal of time and energy wrestling dirty or poorly-integrated data. Their people either cannot find the right data or cannot trust the data they find. On top of that, they must deal with multiple regulations in their industry that are barriers to self-service and data democratization. As a result, organizations try to fix their data through a variety of labor-intensive tasks, from writing custom programs to global replace functions. As a result, the organization's data analysts and data scientists can find their productivity diminished.

This is particularly true within large organizations, where many years of mergers and acquisitions have resulted in an extremely complex data environment of diverse systems and databases. While organizations are busy maintaining these legacy data environments, they are constantly creating new data at unseen speed. Some try to solve this problem using master data management tools by unifying disparate data sources to achieve a single view of their critical business entities.

Several vendor tools approached this problem with a rule-based engine that unites a variety of data sources in their offerings. Rules are easy to implement and understood by many. However, rule-based engines do not scale very well. In the context of large enterprises, where organizations must deal with large amounts of data and a variety of disparate systems, machine learning technologies are now replacing rules engines.

Machine learning has proven remarkably powerful in accomplishing a wide variety of analytics objectives, such as predicting customer churn or detecting fraud in online credit card transactions. While identifying data similarities or unifying data may not be the most exciting application of machine learning, it is one of the most beneficial and financially valuable applications to IBM clients.

The Benefits of Managing Master Data

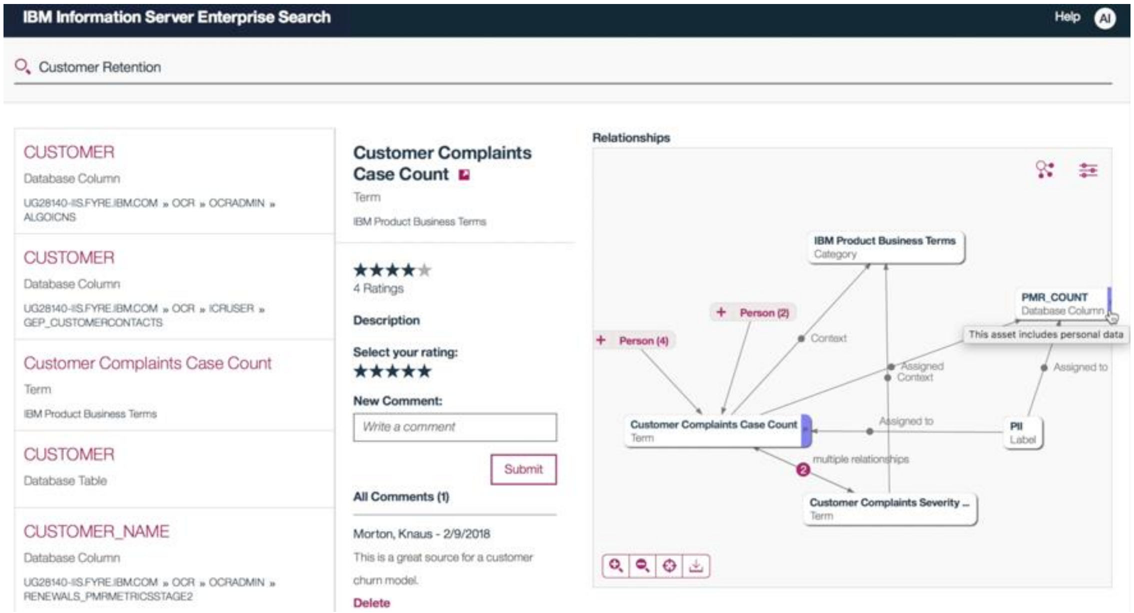
Building a data catalog can be very labor-intensive and time-consuming, which is why so many organizations give up on creating and updating a well-organized data catalog. They also face additional challenges, such as:

- Standardizing business definitions and creating a business glossary
- Cataloging all data sources and updating with clear business descriptions
- Linking business terms to data fields across all data sources

Time is not the only thing needed to build a robust data catalog. It can also be extremely expensive to hire domain experts who can perform this on an ongoing basis. This is where artificial intelligence and machine learning technology can help. IBM DataOps uses machine learning and neural networks to identify probabilistic matches of multiple data records that are likely to be the same entity, even if they look different. This enables analysis of master data for quality and business term relationships, a major pain point for IBM clients. Projects that used to take months can now be done in a few weeks.

While machine learning enables automation of tasks, there is always a need for human intervention in the process, like any other artificial intelligence or machine learning application. Through feedback learning, if the confidence score of match is below a certain threshold level, the system will refer the candidate data records to a human expert using the workflow. It is far more productive for those experts to deal with a small subset of weak matches that an entire dataset.

The benefits of this activity are huge—both for data curators and data scientists in any organization. Consider a new data scientist who is given a task to develop a machine learning model to detect customer churn for a specific product or service. While the data scientist has an idea on what needs to be accomplish, he or she has no idea what data sets he/she can use to start the task. With IBM data governance technology enabled with machine learning, the data scientists can easily search for business terms such as “customer retention” to get a graph view of all connected entities. Then they can drill down and get information about the quality and authenticity of the data.


IBM Information Server Enterprise Search

Customer Retention

CUSTOMER
Database Column
UG28140-IIS.FYRE.IBM.COM » OCR » OCRADMIN » ALGOICNS

CUSTOMER
Database Column
UG28140-IIS.FYRE.IBM.COM » OCR » ICRUSER » GEP_CUSTOMERCONTACTS

Customer Complaints Case Count
Term
IBM Product Business Terms

CUSTOMER
Database Table

CUSTOMER_NAME
Database Column
UG28140-IIS.FYRE.IBM.COM » OCR » OCRADMIN » RENEWALS_PMRMETRICSSTAGE2

Customer Complaints Case Count
Term
IBM Product Business Terms

★★★★☆
4 Ratings

Description

Select your rating:
★★★★★

New Comment:
Write a comment

Submit

All Comments (1)

Morton, Knaus - 2/9/2018
This is a great source for a customer churn model.
Delete

Figure 1: *IBM Information Server Enterprise Search showing Business terms and Technical metadata relationships*

What you see in the figure above is a result of automated data classifications and term assignments using a machine learning model. A classification or taxonomy is a way of understanding the world by grouping and categorizing. Many organizations use Social Security Numbers (SSN) to track a customer with various investment products, but they may appear in various forms such as a Tax Identification Number or Employee Identification Number. Using a traditional, rule-based engine, it is difficult to conclude that these three different terminologies refer to the same entity. By contrast, one term may also have different meanings in the same organization. Machine learning models offer a new way to train the system to describe a domain from the data that helps identify these relationships.

Auto Data Classification

The use of machine learning to perform data classification is a three-step process involving clustering, sorting and classifying. The goal of clustering is to find similar entities in data. The different characteristics used to compare data such as shape and size of an object are called *features*. The more features that are the same or close to being the same among two data values are considered similar data values. Several statistical techniques are available that can be employed to perform clustering. Some techniques ask the user to specify features in advance, whereas others develop the features through comparison of different data values.

The figure below shows the IBM DataOps software stack with core services that it offers. Some of these services are either in beta mode or will be available later this year.

Core DataOps services

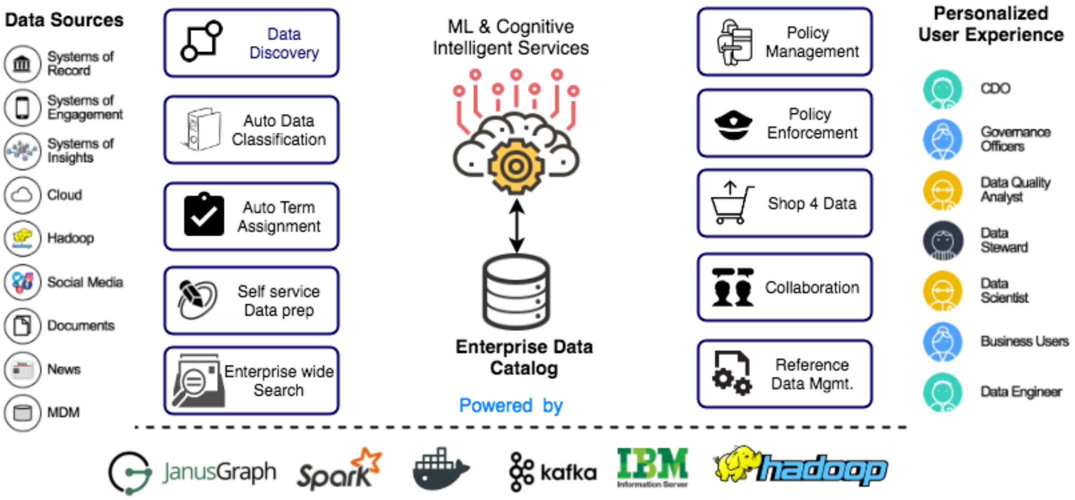


***Figure 2**: IBM Information Server 11.7 core services*

The IBM DataOps software stack is focused on providing personalized user experiences and targeted outcomes for a variety of personas. For example, business users often struggle to find the data they need due to data silos and lack of a central data catalog – oftentimes turning to IT for help. Even when business users find the data they are looking for, they struggle to understand it due to the following reasons:

- The data sources are unlabelled, leaving users to guess the context and meaning of each data field
- The data sources have schemas and data fields that use abbreviations or acronyms that are difficult to understand, specifically from legacy systems

- The data fields are repeated on multiple data sources with different labels, creating a need to standardize business terms

IBM DataOps is enabling business and technical users in the enterprise to identify data easily, and also discover quality data that is useful and endorsed by other users in the organization.

Importance of Feedback Learning in the Discovery Process

Many organizations do not have a well-defined relationship between their business terms and technical metadata. IBM DataOps has employed a technique to pre-train models with industry-oriented datasets and labels to provide candidates matches to users as soon as they load their business glossaries into the IBM Governance Catalog. Users receive confidence scores associated with every candidate term, helping them pick the right one. If a match meets the threshold of 80 percent, the term is automatically assigned to the metadata and if not, the system perceives this as a knowledge input to retrain the machine learning model. This dynamic feedback loop enables data curators to become more productive with enhanced results on subsequent discoveries of metadata.

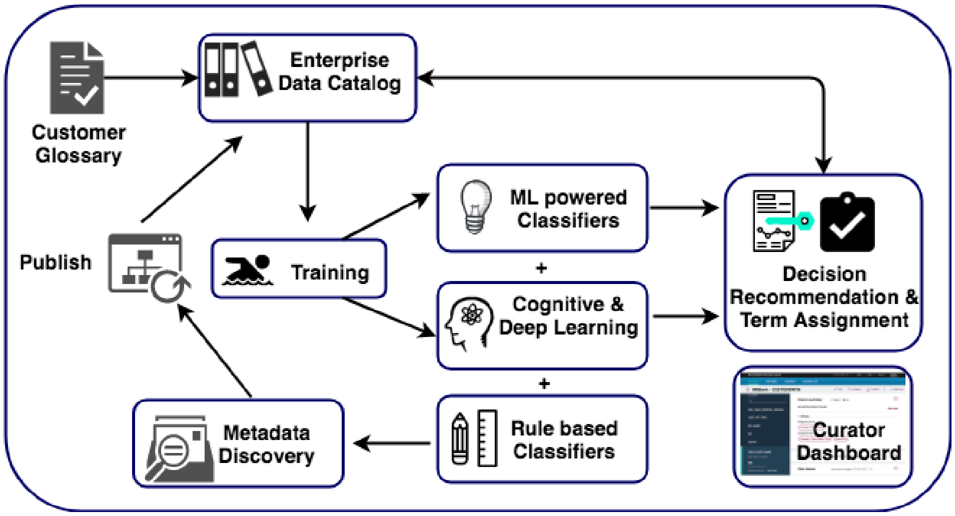
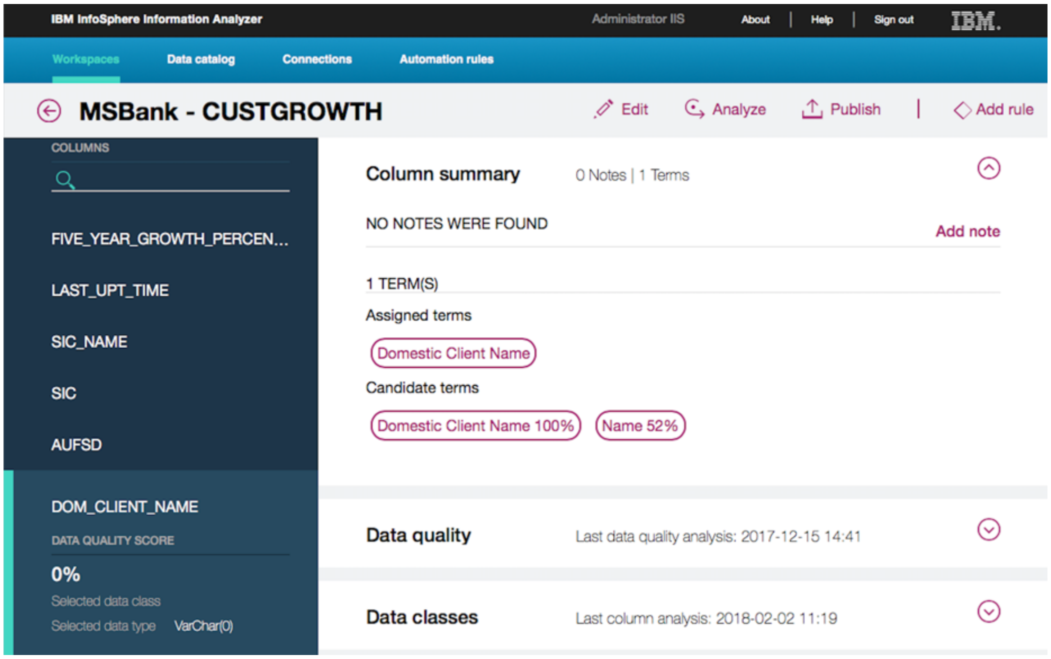


***Figure 3**: IBM Information Server 11.7 – ML powered metadata discovery process*

Traditional techniques of metadata matching and assignments are rule-based. While machine learning models to a better job with ambiguous data sets, they are not replacements for existent application rules, especially in cases where existing application rules and regular expression have proven to work well. IBM DataOps uses machine learning to complement existing rules and provide optimized results where systems can learn from a customer's data domain.



MSBank - CUSTGROWTH

COLUMNS

FIVE_YEAR_GROWTH_PERCEN...

LAST_UPT_TIME

SIC_NAME

SIC

AUFSD

DOM_CLIENT_NAME

DATA QUALITY SCORE

0%

Selected data class

Selected data type VarChar(0)

Column summary 0 Notes | 1 Terms

NO NOTES WERE FOUND

1 TERM(S)

Assigned terms

Domestic Client Name

Candidate terms

Domestic Client Name 100% Name 52%

Data quality Last data quality analysis: 2017-12-15 14:41

Data classes Last column analysis: 2018-02-02 11:19

***Figure 4**: Automated business term assignment using machine learning*

Applying Regulation to Business Data

When businesses try to find, understand, and use the information in their organization, they use several approaches to organize data. They physically move data from systems of engagement to systems of insight, such as data warehouses, data marts or connecting a graph database straight to the instance data. They build data lakes that create a landing zone of all the information that they might be interested in. Data scientists fish for what is useful. And at the edge of the lake they create sandboxes so others can sift through what is coming into the lake and start mining for gold. Ultimately, the most important aspect is knowing what data exists, what it means, where it comes from, how it has been manipulated (also known as data lineage) and what users can do with the data.

To identify what data exists, IBM has offerings for discovery and auto classification – such as StoredIQ and Information Analyzer. These product use machine learning models to read customer data, understand the pattern and learn from it. IBM DataOps addresses key issues of on-boarding regulations to a governed data lake. The IBM approach combines IBM Industry Models with neural networks model to on-ramp a variety of regulations into the IBM Governance Catalog and then map the regulatory terms to the business terms.

Take, for example, the Global Data Protection Regulation (GDPR). There are four processes before GDPR terminology can relate to business terminology to help leverage them for privacy regulations:

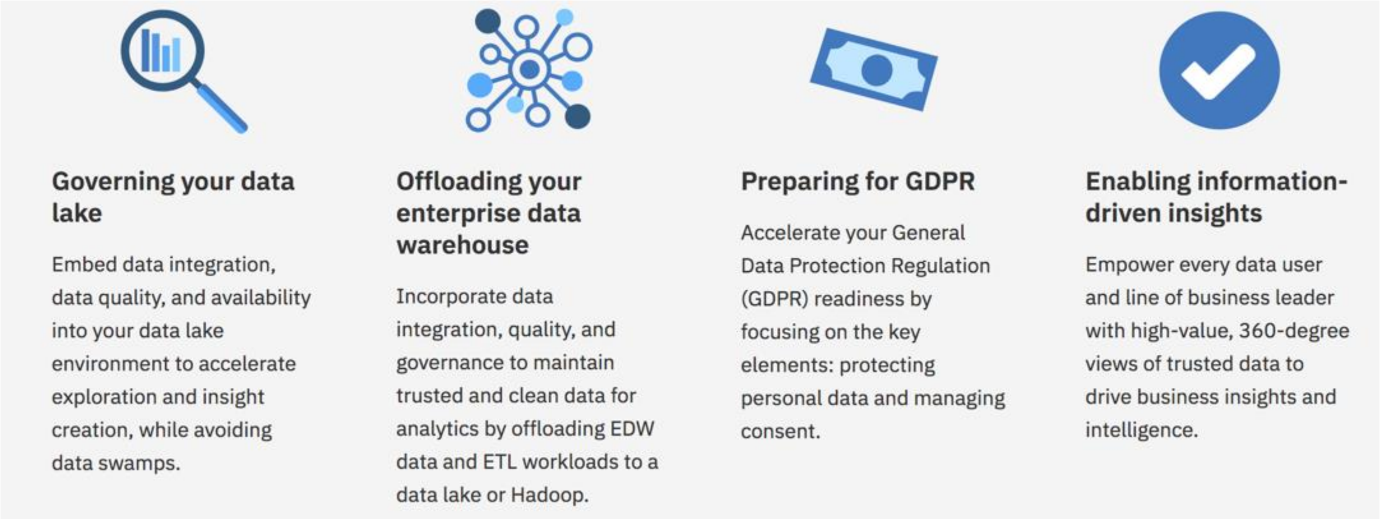
1. Support content terms must be manually extracted from GDPR documents (various sections and articles)
2. Hierarchies must be created for key categories
3. Supportive content terms must be matched manually with the business terms by domain experts
4. These supportive terms must be mapped to the business data model

IBM DataOps uses machine learning to create a neural network model that interprets regulations based on its experience with similar regulations. This not only extracts supportive content terms from raw documents but also create a well-formed taxonomy that can easily be ingested into the IBM Governance Catalog.

Conclusion

Companies in the current trend of global regulations and emergence of the use of hybrid cloud environments (public, private and hybrid) will have to work on managing their Master Data Entities more efficiently and use AI and machine learning to understand the impact of regulations and changes on their business. They need a solution that scales their business processes and helps them throughout their journey of data curation, data discovery, quality and governance. IBM DataOps solutions are designed to help clients achieve that with ease and efficiency.

Watch this webcast to learn how you can faster insights from your data using cognitive data governance in IBM Information Server.



Governing your data lake

Embed data integration, data quality, and availability into your data lake environment to accelerate exploration and insight creation, while avoiding data swamps.

Offloading your enterprise data warehouse

Incorporate data integration, quality, and governance to maintain trusted and clean data for analytics by offloading EDW data and ETL workloads to a data lake or Hadoop.

Preparing for GDPR

Accelerate your General Data Protection Regulation (GDPR) readiness by focusing on the key elements: protecting personal data and managing consent.

Enabling information-driven insights

Empower every data user and line of business leader with high-value, 360-degree views of trusted data to drive business insights and intelligence.