**Data Fabric for Governance and Privacy**

**Tutorial 1: Trust Your Data**

# Background

Data assets can be enriched with information that helps users find data faster, to decide whether the data is appropriate for the task at hand, whether they can trust the data, and how to work with the data. Such information includes, for example, terms that define the meaning of the data, rules that document ownership or determine quality standards, or reviews.

Data stewards create asset profiles to understand the meaning of data and to assess its quality. Also, they add business context to data by assigning terms. Metadata enrichment automates this process thus increasing the data steward's productivity.

Data is useful only if its context, content, and quality are trusted. To keep it that way, data must continuously be evaluated and appropriate remediation be taken if required. Data stewards can configure recurring jobs to continuously track changes to the content and structure of date and then analyze only data that changed.

The information that is added to assets through metadata enrichment also helps to protect data because it can be used in data protection policies to mask data or to restrict access.

# Pre-requirements

## Demo environment

To demo MDE you need one of the below environments:

* IBM Cloud SaaS account with WKC Professional or Enterprise plan (for now until timed trial is in place).
* IBM Cloud SaaS Dev account
* CPD Software (starting from v 4.5 CPD Software will support Metadata Enrichment)

You need to have access to the relevant Category in glossary and be able to create projects or at least have access to the project.

## Relevant content in glossary

Relevant content in glossary is required to demo automatic term assignment.

For demo purposes, you can follow the next steps:

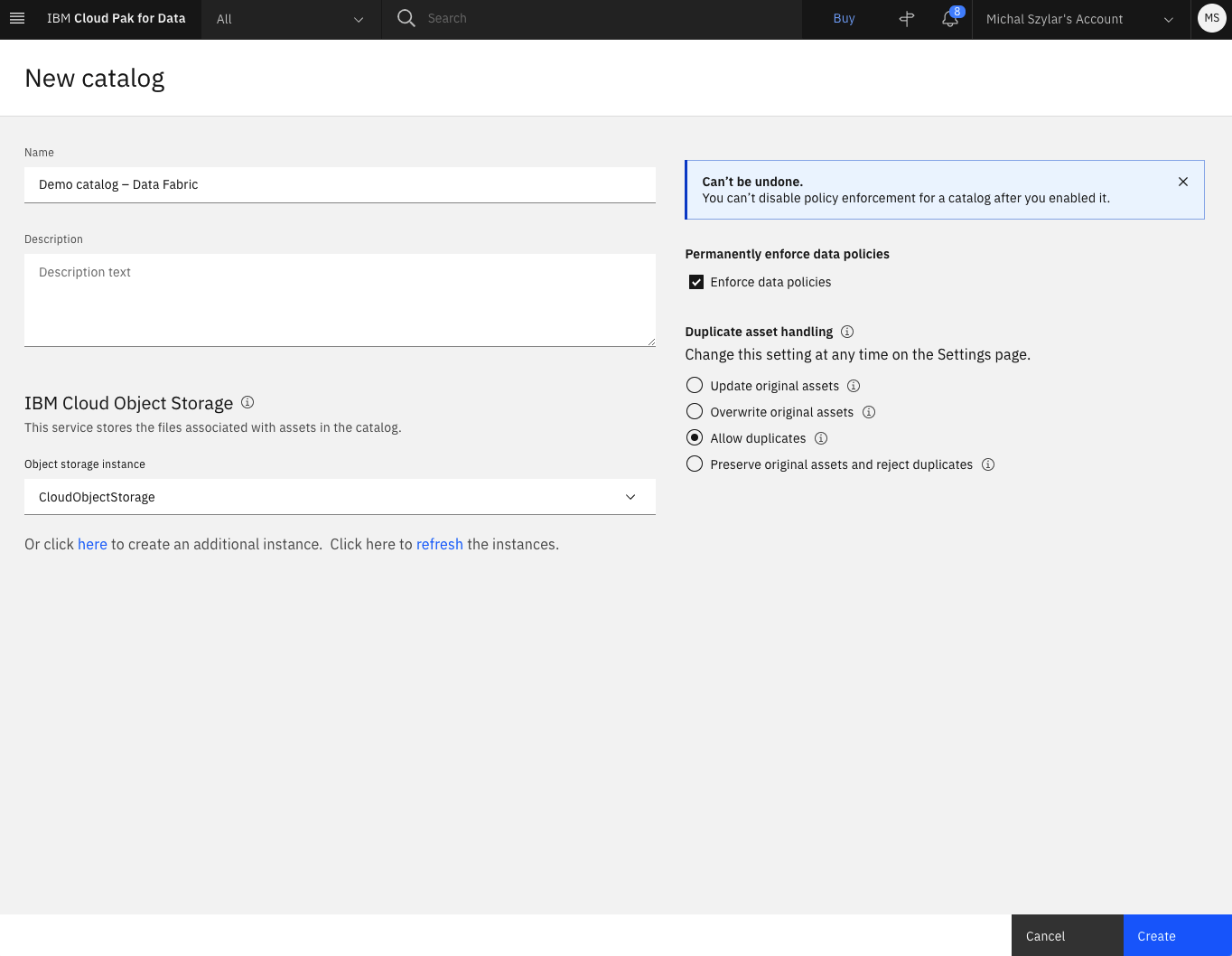
* Create a category called “**Banking”** and give yourself appropriate access permissions to it. Require “Manage governance categories” platform permission.
* Download the file below and import these terms into your environment (see IBM documentation if you do not know how to import terms): https://ibm.box.com/s/1hhhqiemvbogfraksfc31o6i9s6x9omd **“banking.csv”**. Require “Access governance artifacts” permission.
* You can also use Excel demo script: https://ibm.box.com/s/1hhhqiemvbogfraksfc31o6i9s6x9omd **“Import export business terms script.xlsx”** if you want to create your own set of business terms. This need to be then exported to csv and semicolon needs to be replaced by comas in text editor.

## Create a catalog

You can create a new catalog or use an existing catalog that you have access to.

To create a new catalog, open left hand side menu and click **View all catalogs** under **Catalogs** menu group.

Then click **Create Catalog**



For demo purposes, use “**Demo catalog – Data Fabric**“ as a name. Description is optional.

Set **Enforce data policies** option. Click **Create**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

We will return to the catalog in one of the next steps.

## Create project

You can create a new project or use an existing project that you have access to.

To create a new project, open left hand side menu and click **View all projects** under **Projects** menu group.

Then click **New project**

Graphical user interface, application, Teams

Description automatically generated

Click **Create an empty project**

Graphical user interface, website

Description automatically generated

For demo purposes, use “**Metadata Enrichment demo project**“ as a name. Description is optional. Click **Create**.

Graphical user interface, application, website

Description automatically generated

## Create a connection

Go to **Assets** tab in Project view. Click **New Asset** button. Select **Connection.**

Graphical user interface, website

Description automatically generated

Select connection type

Graphical user interface, application, website

Description automatically generated

Graphical user interface, application

Description automatically generated

For demo purposes, use “**Data Fabric Trial - Db2 Warehouse**“ as a connection name. Description is optional.

Graphical user interface, application

Description automatically generated

For demo purposes, we are using following credentials:

* + Name: Data Fabric Trial - Db2 Warehouse
  + Database: BLUDB
  + Hostname or IP address: db2w-ruggyab.us-south.db2w.cloud.ibm.com
  + Port: 50001
  + Username: CPDEMO
  + Password: DataFabric@2022IBM
  + Check the checkbox for “Port is SSL-Enabled”

Click the “Test Connection” button to make sure your entries are valid, and the connection is working. If the test connection is successful, click ‘Create’.

## Create Metadata Import asset - Import data to the project

In **Project -> Assets** click **Add to project.** Select ‘Metadata Import’. For demo purposes, use “**Metadata import demo – Data** **Fabric**” as the name. Optional: enter a description. Optional: use existing tags or create new ones. Select connection created in the previous step.

Graphical user interface, application

Description automatically generated

Chose “This Project” for the import target. Click ‘Next’.

Select data scope. For the demo, from AI\_MORTGAGE schema, select:

* COMMERCIAL\_CLIENT
* CREDIT\_SCORE
* HOUSE\_PRICE
* MORTGAGE\_APPLICANTS
* MORTGAGE\_APPLICATION

Graphical user interface, application

Description automatically generated

Click ‘Select’.

Graphical user interface, application, website

Description automatically generated

Click “Next”

Graphical user interface, application

Description automatically generated

Click ‘Next’ on the Schedule optional step. Click ‘Create’.

Graphical user interface, application, Teams

Description automatically generated

Wait until it finishes - Proper UI notification is displayed.

Graphical user interface, application, website

Description automatically generated

# Metadata Enrichment

## Create Metadata Enrichment asset

Metadata enrichment automatically providing context, content, and quality information in a unified framework.

In Project -> Assets click ‘Add to project’. Select ‘Metadata Enrichment’.

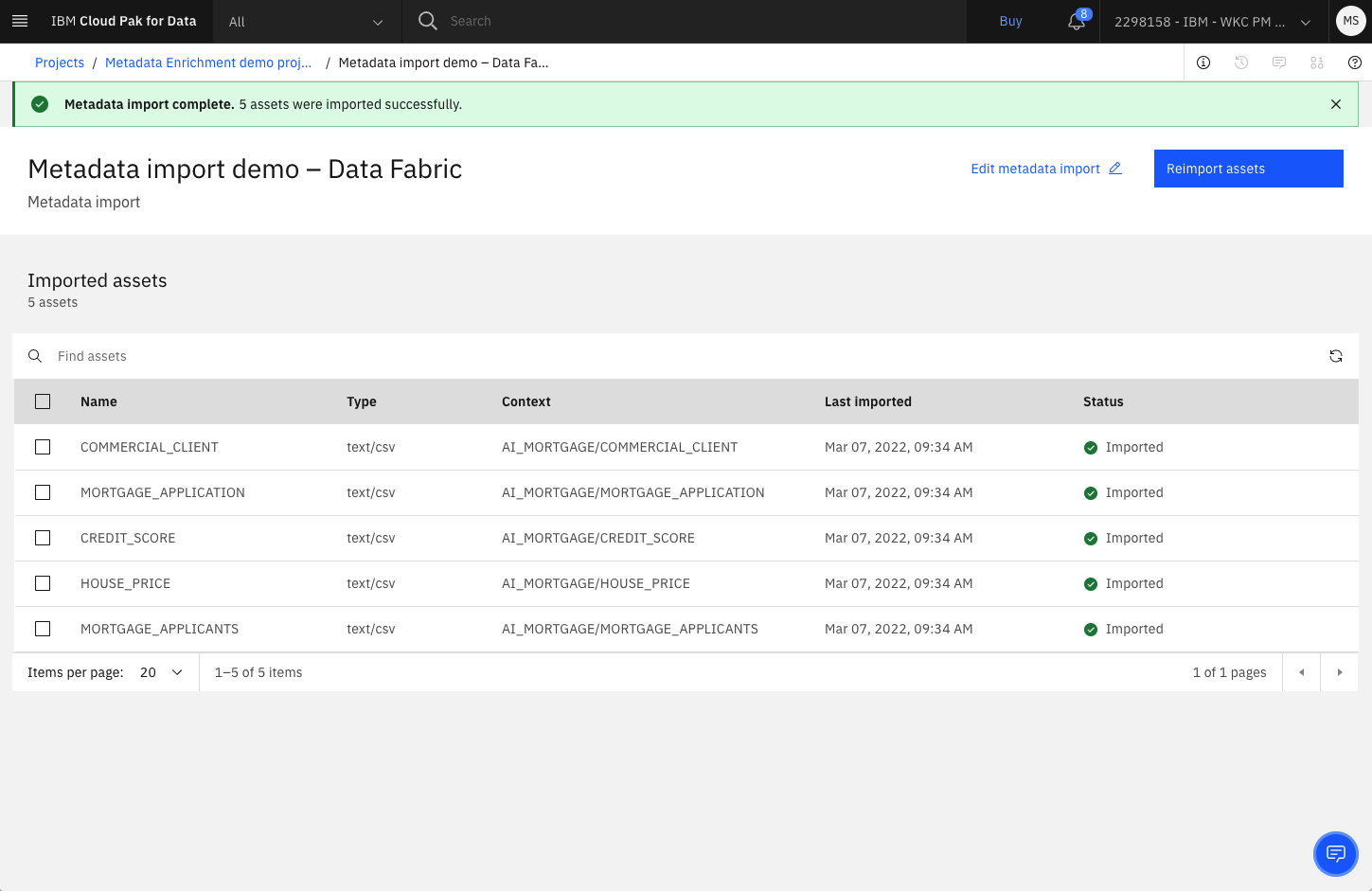
Graphical user interface, website

Description automatically generated

For demo purposes use **Metadata enrichment demo - Data Fabric** as the name or whatever you would like to use.

Click ‘Next’.

Select data from project. Click on data assets and select data assets you want run MDE. I select all six available tables.



Click ‘Select’.

Graphical user interface, application, Teams

Description automatically generated

Click ‘Next’. Select all available options for Enrichment objective.

Graphical user interface, text, application

Description automatically generated

Click ‘Select categories’ - !!! **Important**!!!. Selection in this step will impact the outcome of the metadata enrichment. For demo purposes, select uncategorized and Banking categories. The Uncategorized category contains data classes required by profiling. As well, in the Banking category there are some Terms created specifically for the BANKDEMO assets in this demo.

Graphical user interface, text, application

Description automatically generated

Click ‘Select’.

Graphical user interface, application, website

Description automatically generated

You can leave sampling as it is (Basic). Click ‘Next’.

Graphical user interface, text, application, Teams

Description automatically generated

Click ‘Next’.

Graphical user interface, application, Teams

Description automatically generated

Click ‘Create’. After the last step MDE process starts and relevant job is created. The process can take ~2-3 minutes.

Graphical user interface, application

Description automatically generated

## Jobs / runs

You can use jobs for troubleshooting and to get fine graded insight into all the runs.

You can see job status when you go to hamburger menu->Jobs.

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

To get details on specific run, click on the relevant run.

Graphical user interface, text, application

Description automatically generated

## Result of MDE – profiling

After Metadata enrichment run is completed, you can navigate back to project “**Metadata Enrichment demo project” -> Assets tab and from the left hand side menu select assets** “**Metadata enrichment demo - Data Fabric**”.

Graphical user interface, text, application

Description automatically generated

Next click on the **Columns** tab, and on the **City** column. In the right hand side panel, you can find profiling information like: Format, Frequency distribution, Statistics.

Graphical user interface, application

Description automatically generated

This analysis (also known as column analysis) literally captures the profile of the data. Structured profiling provides sample values, formats (aka patterns), detection of data properties (lengths, data types, min/max, etc.) and frequency distributions.

For columns with unclear names like “CSRidUpdate21”, sample values that look like email address help users understand what is contained in that column. For a column called “ZIP”, a frequency distribution with 80% 5-digit U.S. zip codes and 20% 6 digit-Canadian zip codes helps a user know that this data is from the worldwide database. Profiling gives consistent information to help users understand data fast.

## Result of MDE – class and term assignment

In the columns view you can also see data class and business term auto assignment. You can navigate to the right-hand side panel -> Governance to adjust auto assigned terms and data classes:

Graphical user interface, application

Description automatically generated

Locating, assessing, and managing your most critical data elements is a prerequisite for unlocking business value, complying with regulatory standards, and reducing risk. Watson Knowledge Catalog scales the productivity of data stewards by automating the data curation process with patented technology that mirrors how human beings recognize and curate data.

This automatic business term assignment help connect business meaning to the data — establishing signal from noise. This process of assigning meaning is often referred to using phrases like mapping or auto-tagging. The value of assigning business terms is that data with the right business term is easier to find using search, easier to protect using data protection rules, and easier to analyze using automation rules to trigger specific data quality rules.

Like automatic business term assignment, automatic data class assignment (aka data classification) allows for automatic detection of business meaning and value. Automatic data class assignment analyzes actual data values to assign the best data class, and this result is one of a few factors used in the automatic business term assignment framework.

Automatic data class assignment is run in conjunction with profiling, together known as column analysis. IBM provides over 200 out of the box data classes. These range from a core set (identifier, code, text, quantity, etc.) to specific domains (credit card, email address, etc.). In some instances, these data classes provide validation of the data, like verifying valid credit card number's structure.

## Result of MDE – data quality

To see data quality details, you can click on 3 dots icon -> View data quality:

Graphical user interface, application, Teams

Description automatically generated

Managing and trusting data at enterprise scale depends on extensible framework for analyzing data quality. Using pre-built or customized measures of quality, a user can easily understand the nature of their data from a single number and then take action on their most critical assets.

WKC generates a data quality score for each column and data asset right out of the box. The data quality score is calculated for every data asset and column by analyzing every value in every record according to pre-built dimensions. Think about this scoring like a nutrition label for a data asset — it’s a standardized set of metrics that captures the overall quality of a data asset. Using this data quality score, customers can quickly locate and value and risk.

DQ scores help users locate, diagnose, then prioritize data quality issues. Scores also help identify and monitor business value. For example, if my data quality score is high and few DQ dimension violations are found, I can be confident that a data asset is of quality.

Table

Description automatically generated

Under 3dots menu, you can also mark asset or column as reviewed.

Graphical user interface, text, application, Teams

Description automatically generated

## Publishing assets to catalog

After all assets are reviewed, you can publish one or all of them to a catalog.

Select relevant assets and click -> Publish to catalog

Graphical user interface, application

Description automatically generated

Select target catalog created in one of the previous steps.

Graphical user interface, application, Teams

Description automatically generated

Click **Publish. You can navigate to “Demo catalog – Data Fabric” to check results.**

Graphical user interface, application, Teams

Description automatically generated

# Conclusion

Data is like water—best served fresh, well-filtered, and with confidence in its origins and quality. A business’ data-driven decisions are only as good as the data that underlies them. The goal of metadata enrichment is to create value and meaning from data.

Metadata Enrichment enables customers to go from backroom to showroom, from data source to data catalog, faster. It enriches metadata with understanding. How? Automatically providing context, content, and quality information in a unified framework.

Automatic data class and business term assignment help automatethe data curation processand enable faster, more scalable data governance. Profiling gives users confidence in the meaning and structure of their data. Data quality scoring and dimensions allow for easy assessment of business value and detection of data quality issues.

With infinite time, people, and a lot of handwritten code, customers could replicate almost all of this functionality. This functionality is valuable because it’s fast, scalable, and pre-built to deliver ROI from enterprise data.

# Resources and documentation

* Metadata Enrichment documentation can be found here: <https://dataplatform.cloud.ibm.com/docs/content/wsj/governance/metadata-enrichment.html#supp-assets>
* You can learn more about Terms and Data Classes in Governance artifacts documentation: <https://dataplatform.cloud.ibm.com/docs/content/wsj/governance/governance.html?audience=wdp>
* Metadata Import documentation: https://dataplatform.cloud.ibm.com/docs/content/wsj/manage-data/metadata-import.html?audience=wdp