Governed MLOps Workshop

Data Privacy, Quality and Metadata Enrichment with Watson Knowledge Catalog

Document version: June 2023

Table of Contents

Data Quality and Data Privacy – Introduction

A key component of a Governed MLOps methodology is delivering well governed data with high quality and enforced privacy. High quality data produces high quality AI models. Data with enforced privacy guarantees compliance with the governance and regulatory requirements for the enterprise. To maintain data privacy and deliver high quality data, it is important to perform the following tasks after creating connections to, and possibly virtualizing, relevant data sources across a hybrid cloud environment:

1- Apply data protection rules to enforce masking and protection of sensitive data such as PII, Personally Identifiable Information, data.

2- Review and improve the quality of the data stored in these data sources.

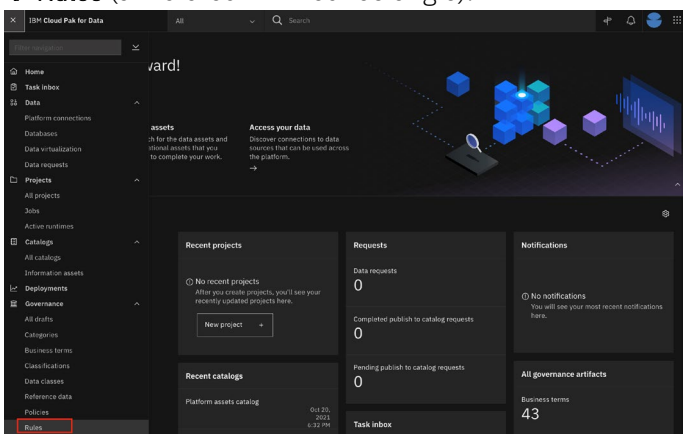
Data Privacy via Data Protection Rules

In Cloud Pak for Data, Watson Knowledge Catalog (WKC) enables data masking by applying data protection rules to discovered data assets based on a variety of conditions such as detected data classes and business terms. For example, a data protection rule can state that a data column should be redacted if it is classified as a Credit Card Number.

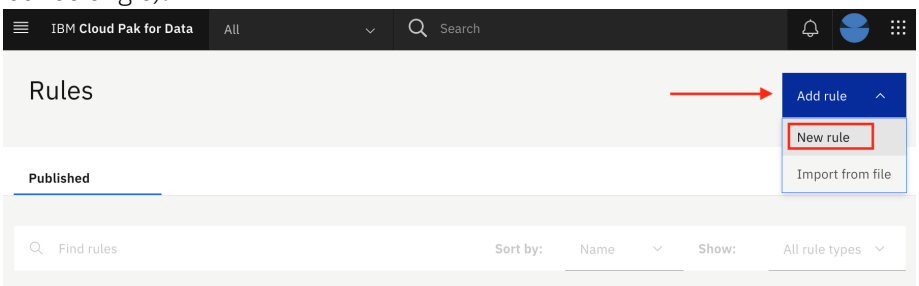
In this lab, you will login as a datasteward user (Data Steward role) and define some sample data protection rules which will be applied to the discovered data assets.

1- Login into Cloud Pak for Data as *datasteward* user.

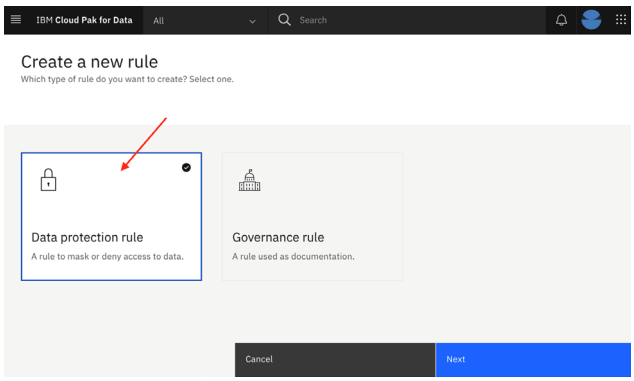
2- Navigate to Rules by clicking the Navigation menu (top left hamburger icon) and selecting **Governance → Rules** (annotated with red rectangle).



3- On the Rules page, click on **Add rule** (annotated with red arrow) and select **New rule** (annotated with red rectangle).

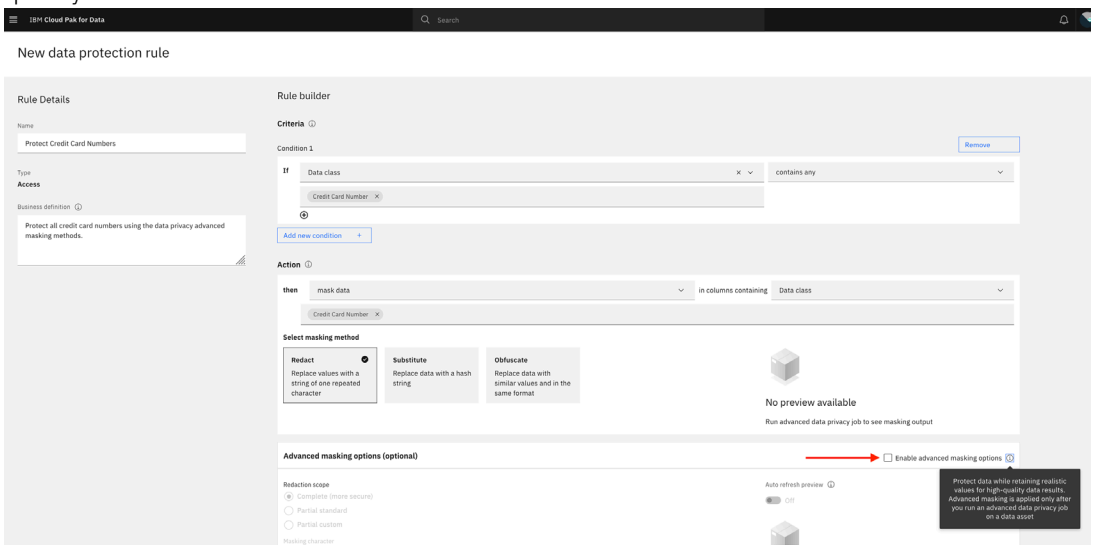


4- On the **Create a new rule page**, select the **Data protection rule** (annotated with red arrow) and click **Next**.



5- On the **New data protection rule** page, fill out the information as shown in the figure below indicating that data in columns with **Credit Card Number** data class should be masked using the Redact method.

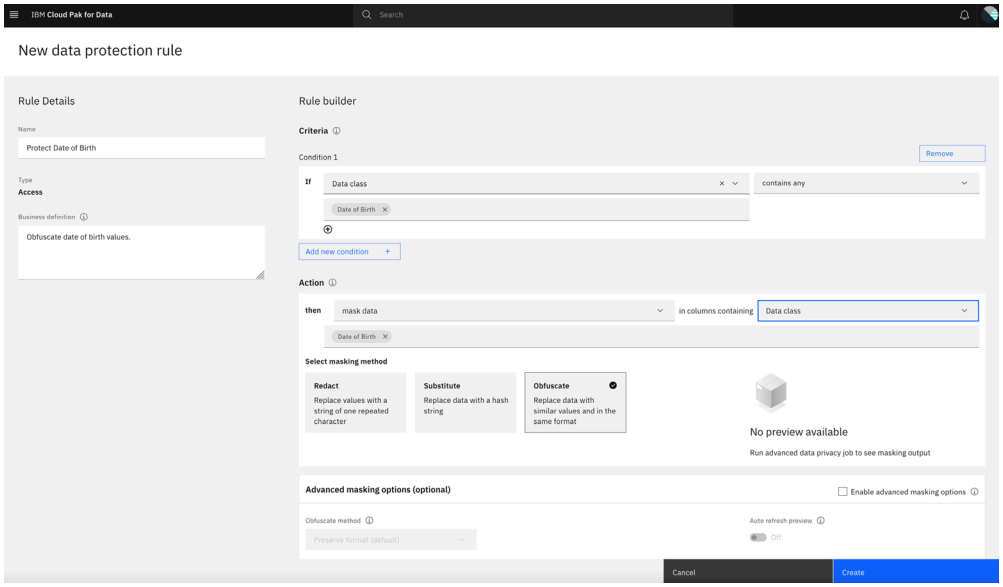
Note the different options for masking where the original data can be redacted, substituted, or obfuscated. Additionally, note that you can select the checkbox (annotated with red arrow) to **Enable advanced masking options** which allows you to protect data while retaining realistic values for high quality data results. Click **Create**.



Please consult the [Advanced data masking documentation] for more details.

6- Navigate back to Rules and create another data protection rule to obfuscate date of birth values by executing steps 3-5.

Fill out information for this rule as shown below.



Data Quality and Data Discovery via Metadata Import and Enrichment

With IBM Cloud Pak for Data, you can analyze and curate data to ensure and improve data quality. During analysis, you can identify or assign data classes and business terms, or process automatically assigned data classes and business terms.

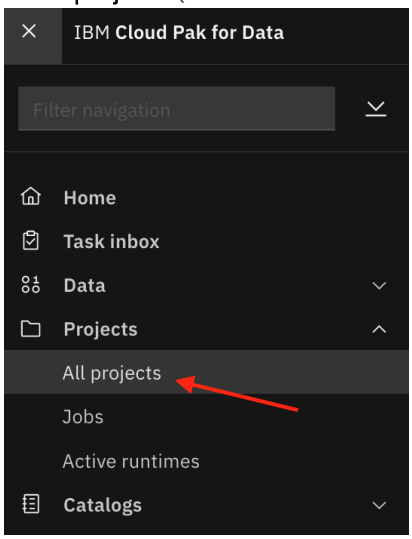
To discover assets and get insight about the quality and business content of tables and files analyzed from various data connections, you can leverage metadata import and metadata enrichment for purposes of scanning data assets, assigning business terms and data classes, and evaluating overall data quality.

The pre-defined Data Quality Analyst role has the required permission to run and execute metadata import and enrichment jobs in order to evaluate the quality of the data discovered from the various data sources.

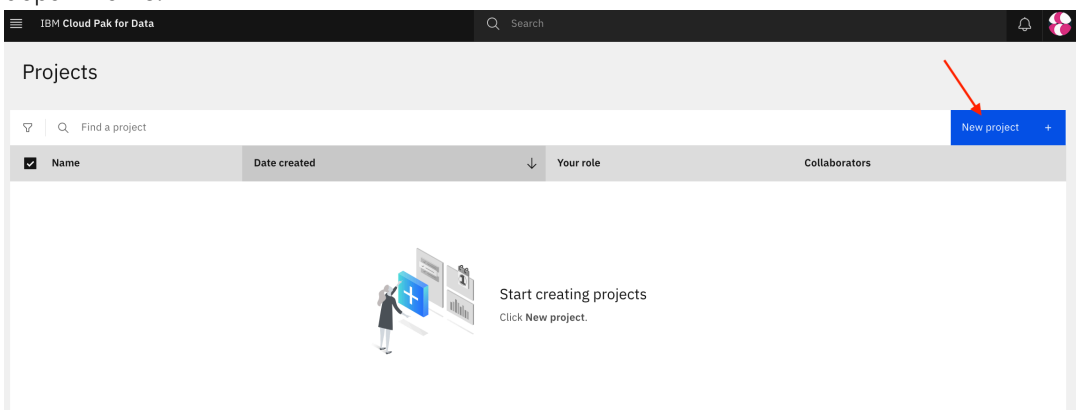
In this lab, as the dqanalyst user, you will run metadata import, metadata enrichment and data quality analysis to enrich data assets from the different data sources and assess their data quality.

1- Login into Cloud Pak for Data as *dqanalyst* user.

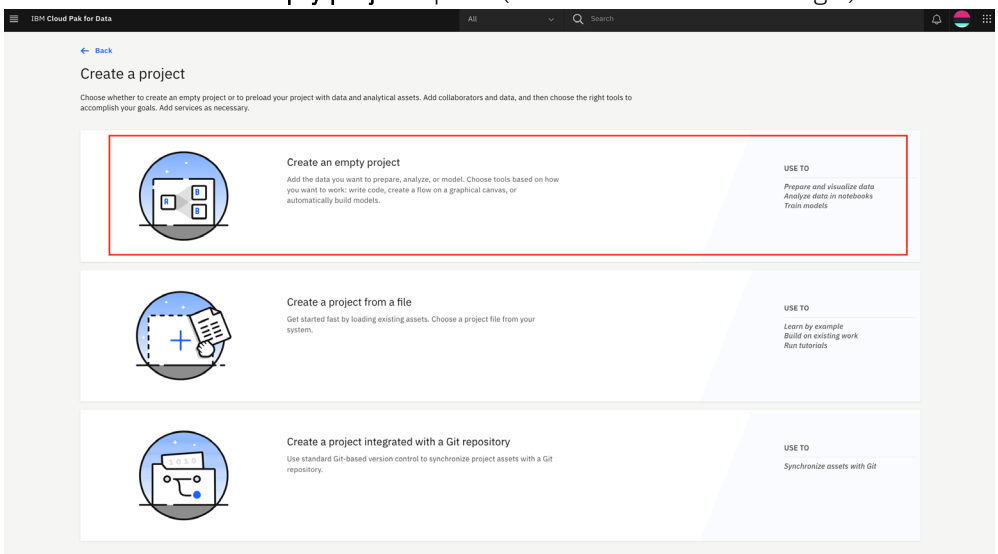
2- Select All projects by clicking on the Navigation menu (top left hamburger icon) and selecting **Projects → All projects** (annotated with red arrow).



3- Click on **New project** (annotated with red arrow) to create a new project. In IBM Cloud Pak for data, a project is how you organize your resources to achieve a particular goal. A project allows for high-level isolation, enabling users to package their project assets independently for different use cases or departments.



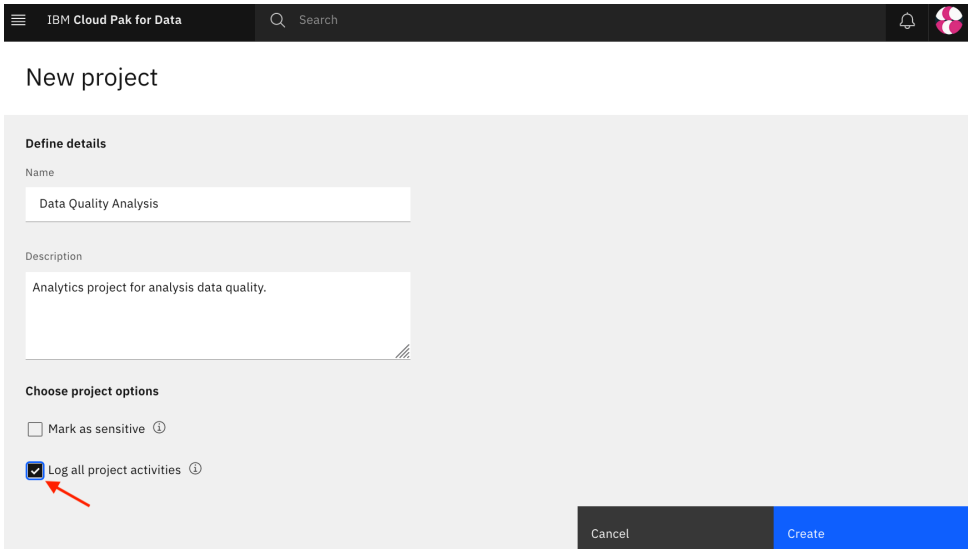
4- Select the **Create an empty project** option (annotated with red rectangle).



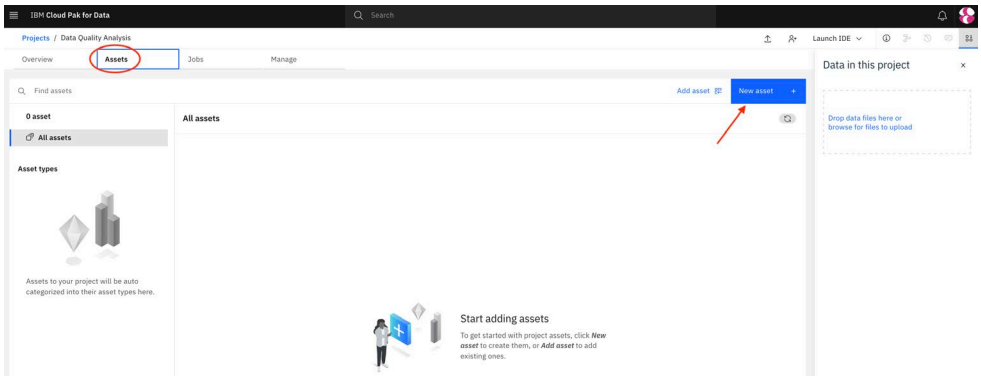
5- On the **New project** page, provide a **Name** and **Description** (optional) for the project.

Select the check box next to **Log all project activities** (annotated with red arrow; this is optional but helps track project activities).

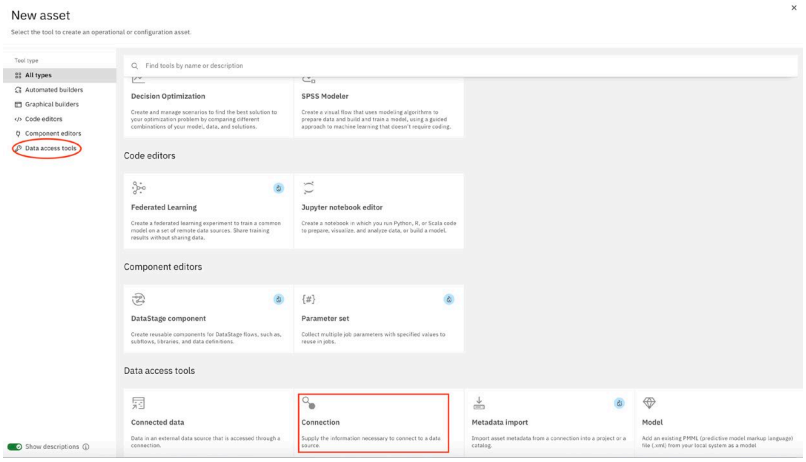
Click **Create**.



6- On the project's page, click the **Assets** tab (annotated with red oval) and then click **New asset** (annotated with red arrow).

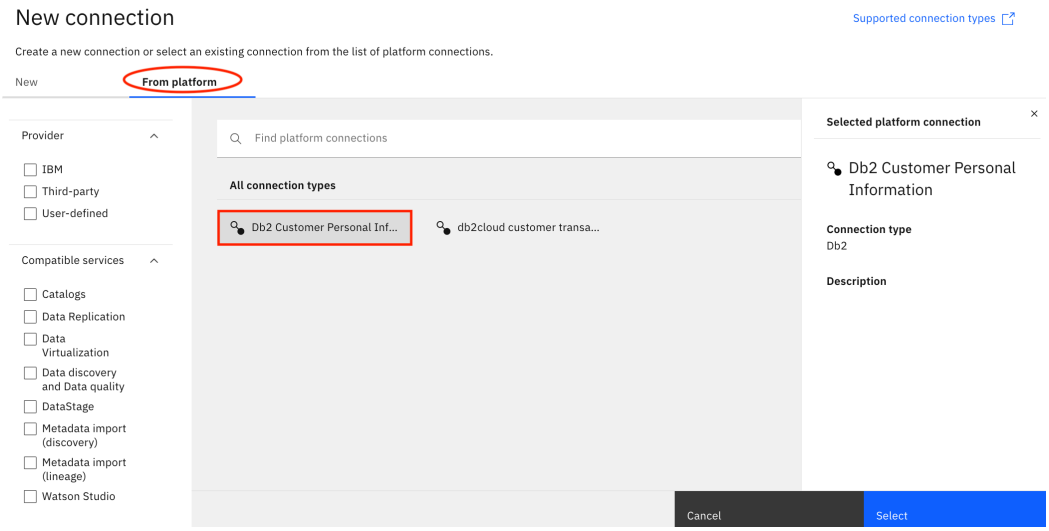


7- Select **Connection** (annotated with red rectangle) as the asset type to add to the project. You can also select the **Data access tools** (annotated with red oval) under Tool type to filter asset types available.



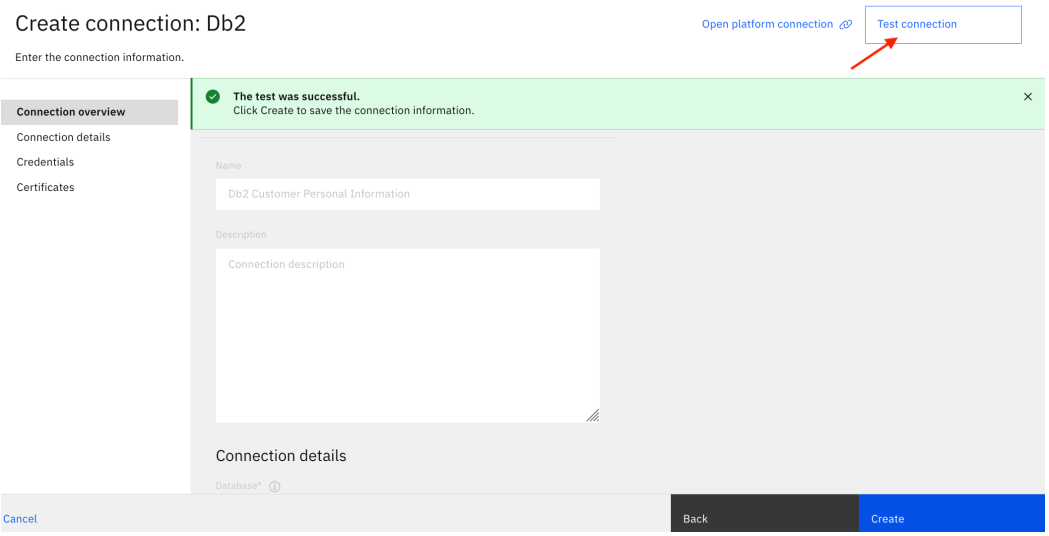
8- On the New connection page, select the **From platform** tab (annotated with red oval) to select from the data source connections already created at the platform level.

Then select the **Db2 Customer Personal Info** connection (annotated with red rectangle) and click **Select**. As a reminder, the Db2 Customer Personal Info connector is for the on-prem Db2 database running on the same Cloud Pak for Data cluster.



9- Review the connection details and click **Test connection** (annotated with red arrow) to confirm the connection information is correct.

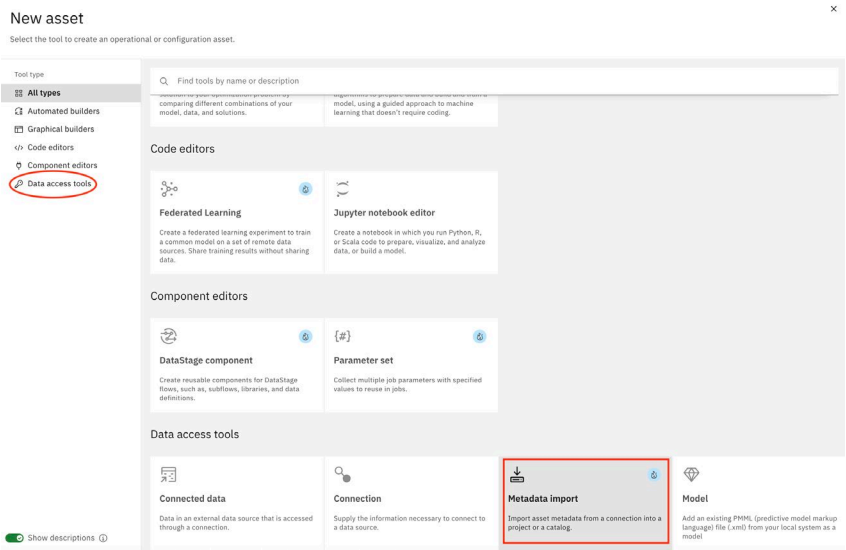
Once you get the green confirmation bar indicating successful connection, click **Create**.



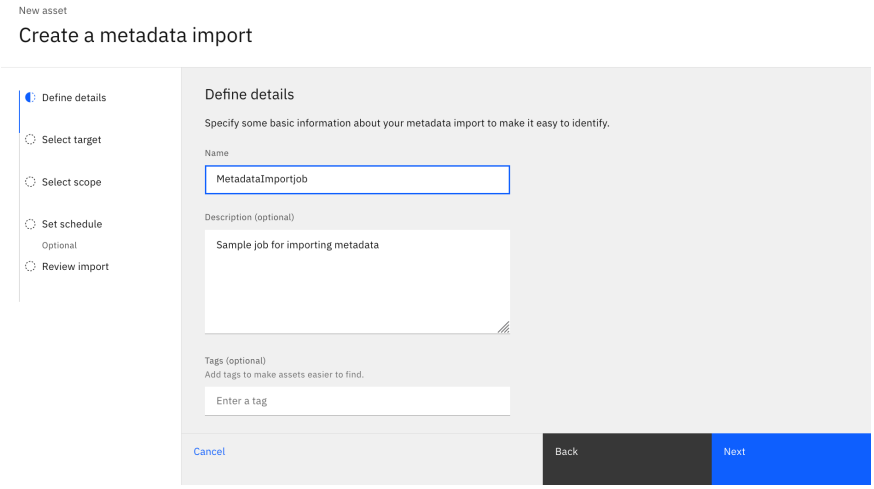
10- Next you will run a Metadata import against the connected data source. On the project's page, click **New asset** again.

Select the **Metadata import** (annotated with red rectangle) as the asset type to add to the project.

You can also select the **Data access tools** (annotated with red oval) under Tool type to filter asset types available for faster access to the Metadata import function.

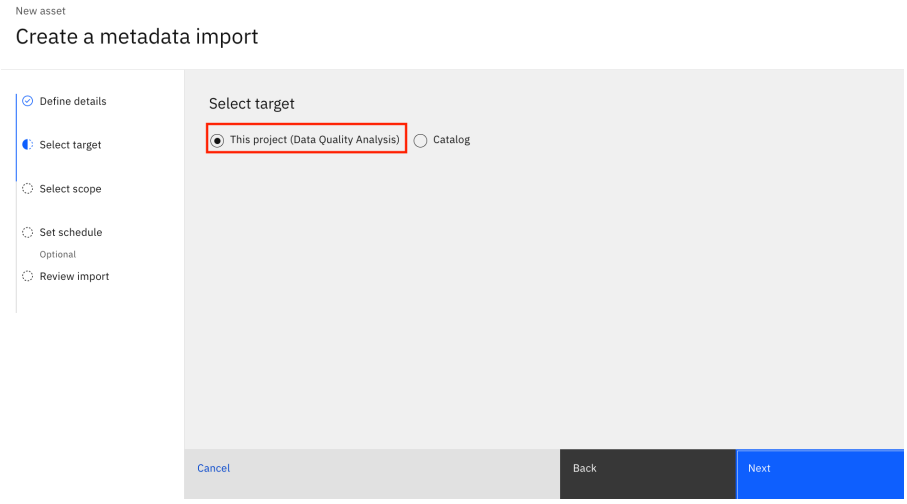


11- Provide a **Name** and **description** (optional) for the metadata import job, then click **Next**.

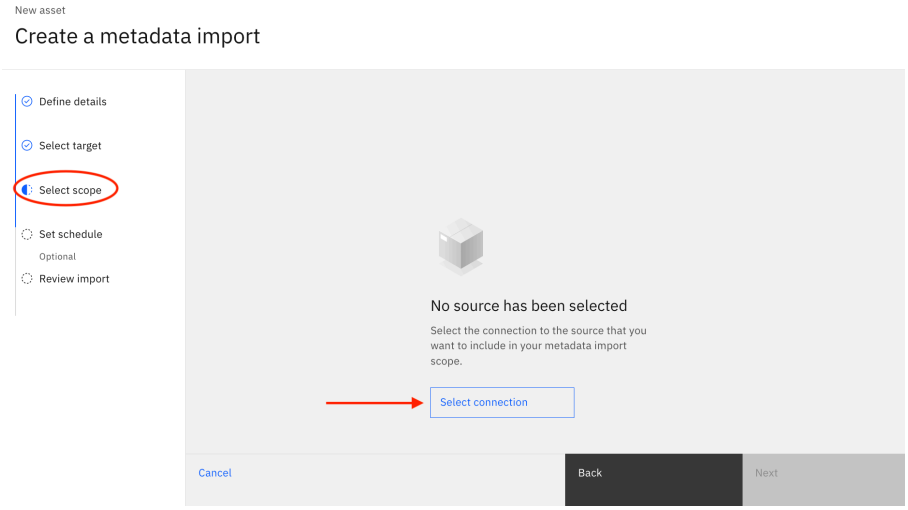


12- Select the target as the current project, **Data Quality Analysis**, (annotated with red rectangle) and click **Next**.

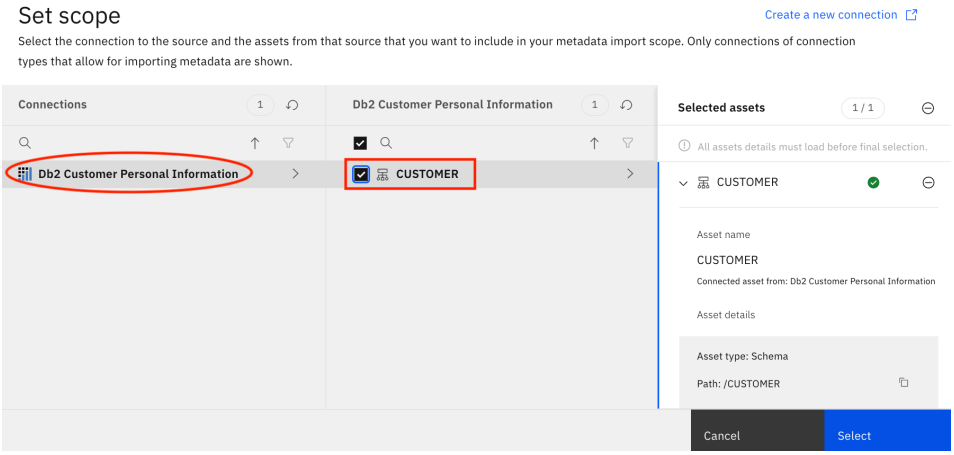
Note that you can also import metadata into a catalog but for this module, we import metadata into the project as we'll also run metadata enrichment as the next step.



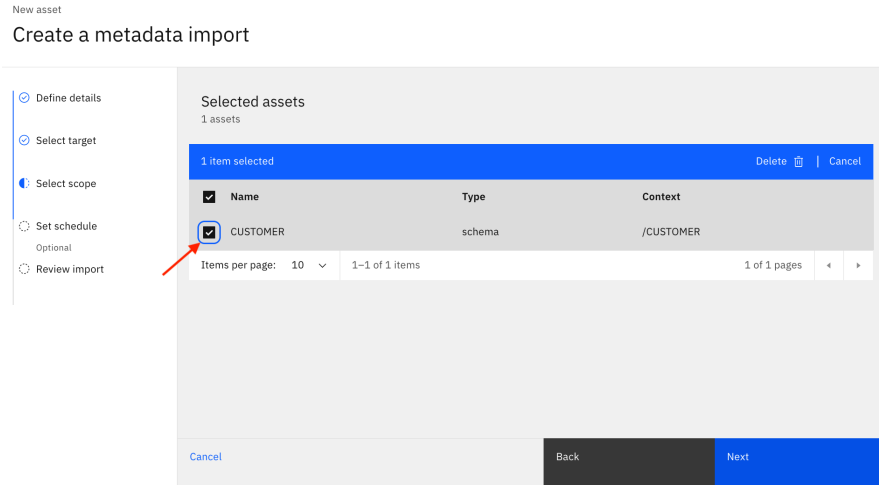
13- On the **Select scope** step (annotated with red oval), click **Select connection** (annotated with red arrow).



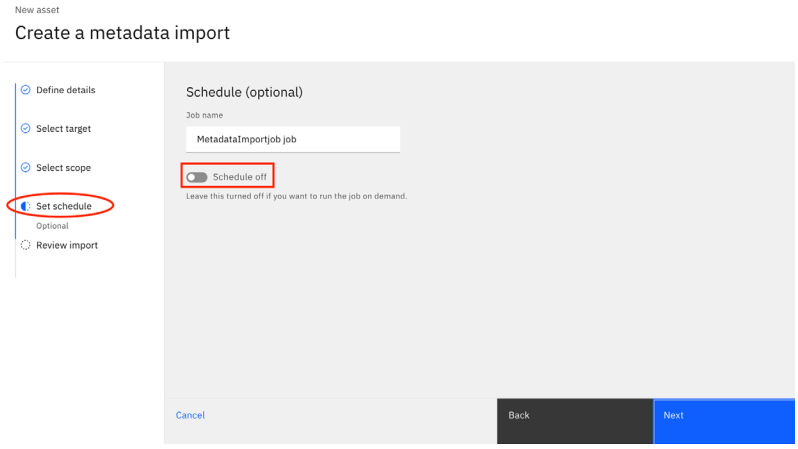
14- On the Set scope page, select the **Db2 Customer Personal Information** connection (annotated with red oval), select the **CUSTOMER** schema (annotated with red rectangle), and click **Select**.



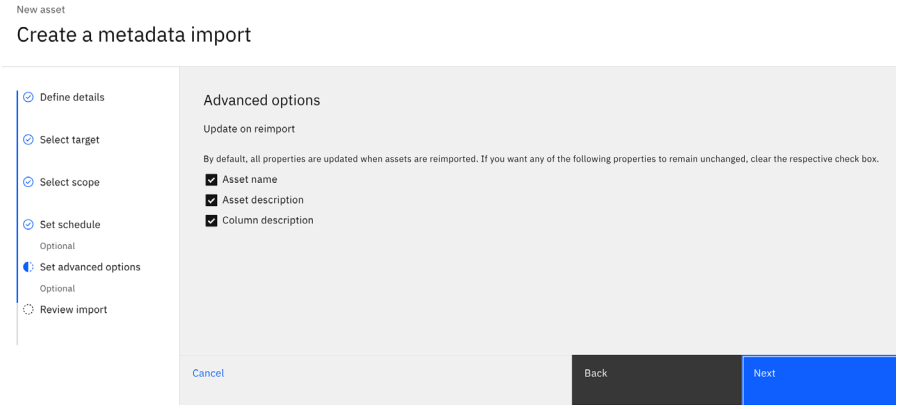
15- Select the **CUSTOMER schema** (annotated with red arrow) to import metadata for that schema and click **Next**.



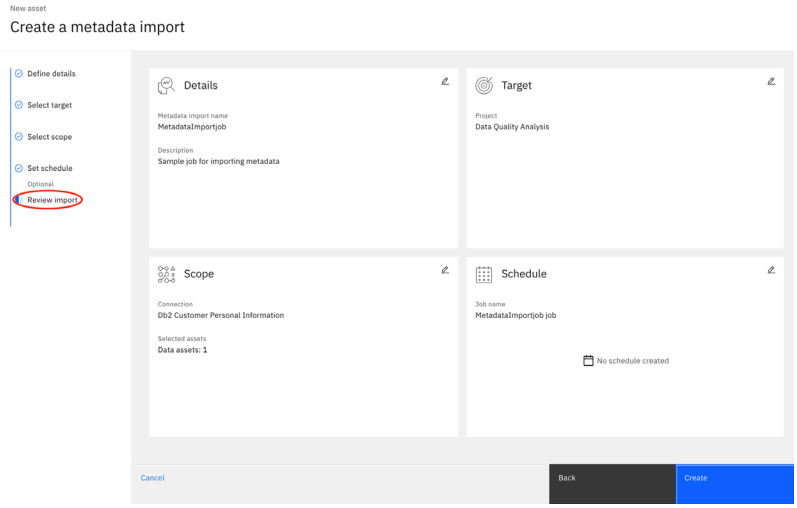
16- Next, on the Set schedule step (annotated with red oval), you can define a schedule for periodically running the metadata import. For this lab, keep the **Schedule off** setting (annotated with red rectangle) and click **Next**.



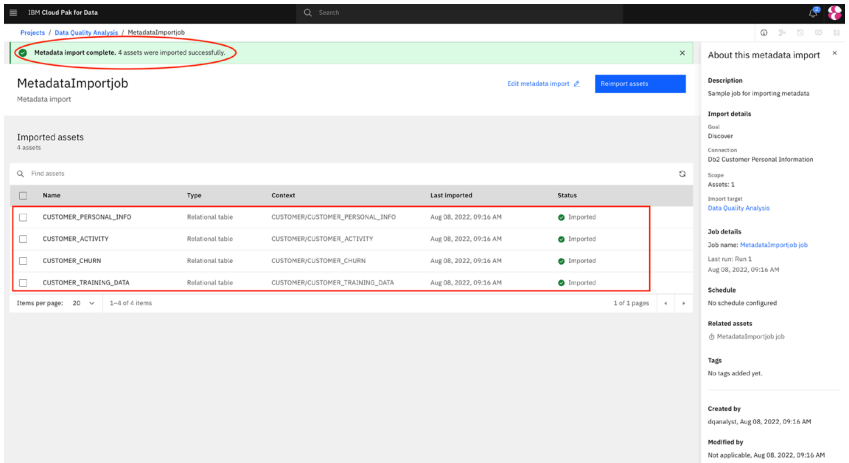
17- On the Set advanced option step, keep the default selections and click **Next**.



18- On the Review import step (annotated with red oval), review the various configurations for this metadata import job and click **Create**.



19- The Metadata import job runs and imports metadata from the CUSTOMER schema and returns the imported assets (annotated with red rectangle) with a summary message (annotated with red oval).

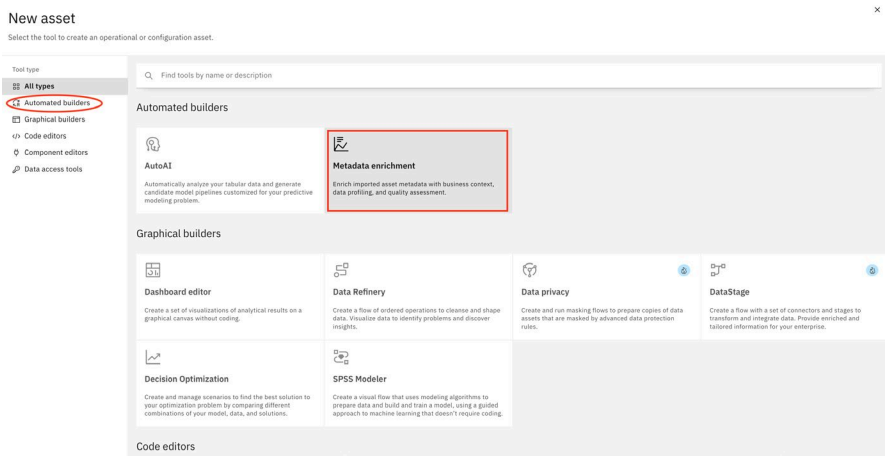


20- Next, you will run metadata enrichment to enrich the imported assets with the defined business terms and data classes.

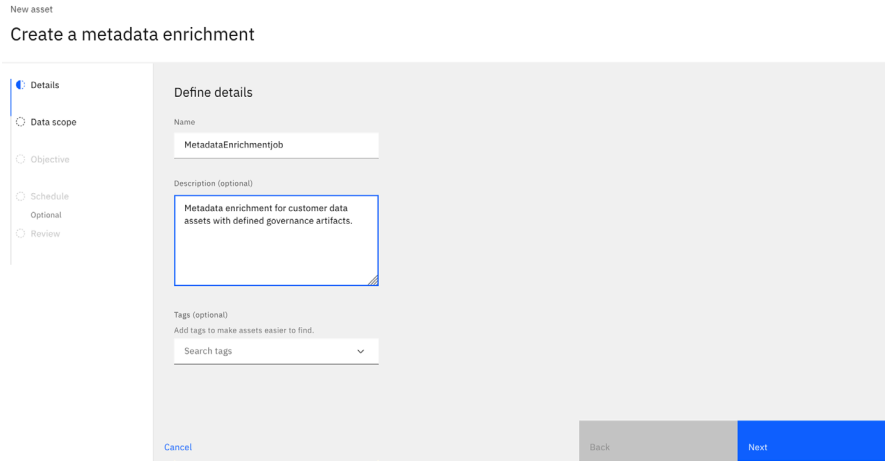
Click on Data Quality Analysis project breadcrumb and on the project's Assets page, click **New asset** again.

Select the **Metadata enrichment** tile (annotated with red rectangle) as the asset type to add to the project.

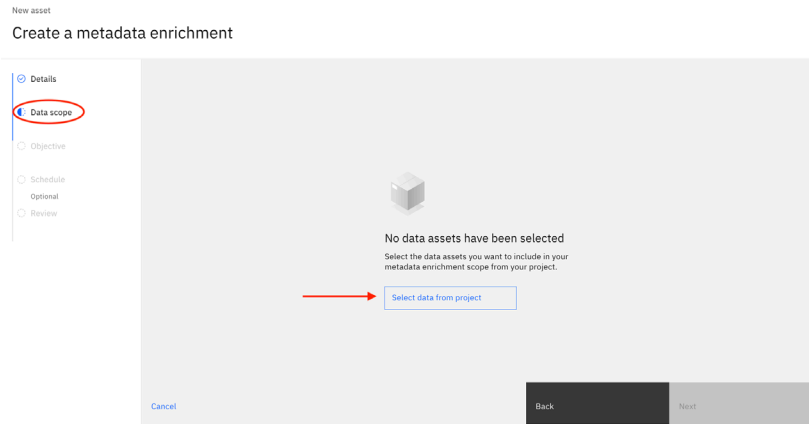
You can also select **Automated builders** (annotated with red oval) under Tool type to filter asset types available for faster access to the Metadata enrichment function.



21- Provide a **Name** and **description** (optional) for the metadata import job, then click **Next.**

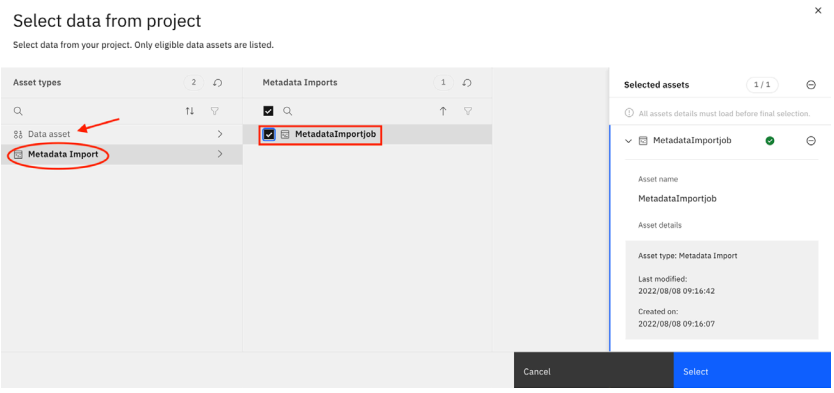


22- On the Data scope step (annotated with red oval), click **Select data from project** button (annotated with red arrow).

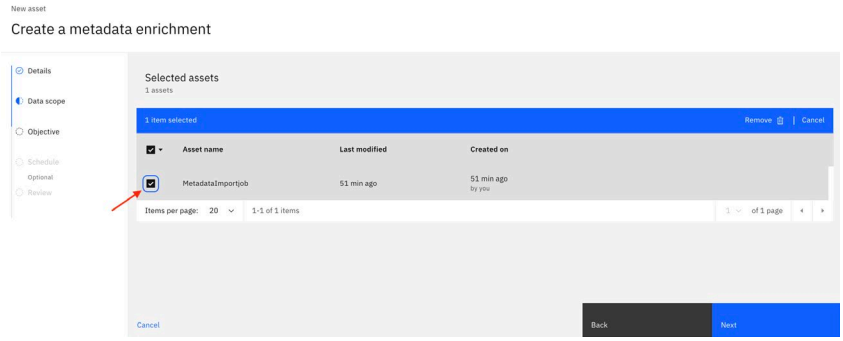


23- On the Select data from project page, you can either select specific data assets from the project (annotated with red arrow) or you can select the Metadata Import job (annotated with red oval) to apply enrichment to all the data assets imported via that Metadata import job.

For this lab, select the **Metadata Import** type (annotated with red oval) and then click the check box next to **MetadataImportjob** (annotated with red rectangle). Click **Select**.



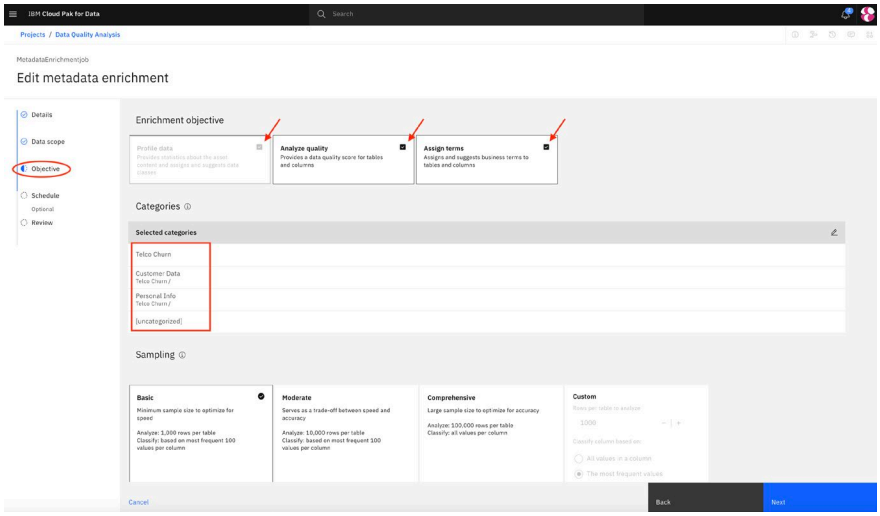
24- Next, select the **MetadataImportjob** (annotated with red arrow) and click **Next**.



25- On the Objective step (annotated with red oval), click the check boxes next to **Profile data**, **Analyze quality**, and **Assign terms** (annotated with red arrows) to tell the Metadata enrichment job to perform all these tasks.

Also, select the Categories to be the **Telco Churn** and [uncategorized] (out of the box with WKC) categories to limit the enrichment to the specified set of terms within that category.

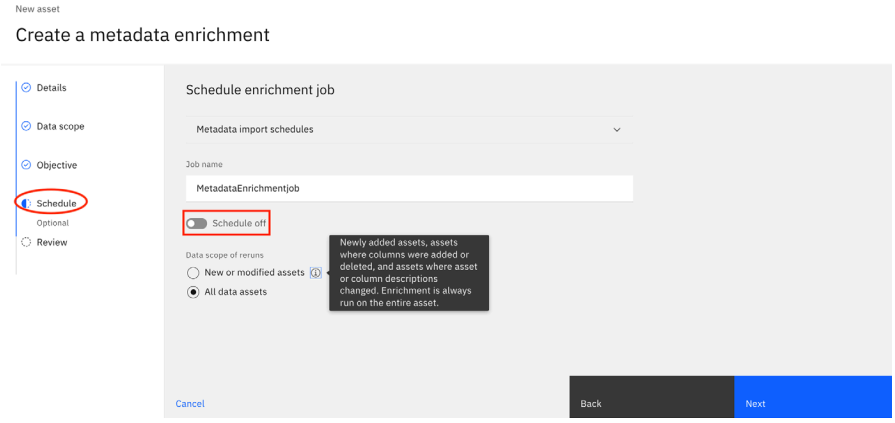
You can also select one of the available Sampling options but for this step, please keep the Basic sampling selection. Click **Next**.



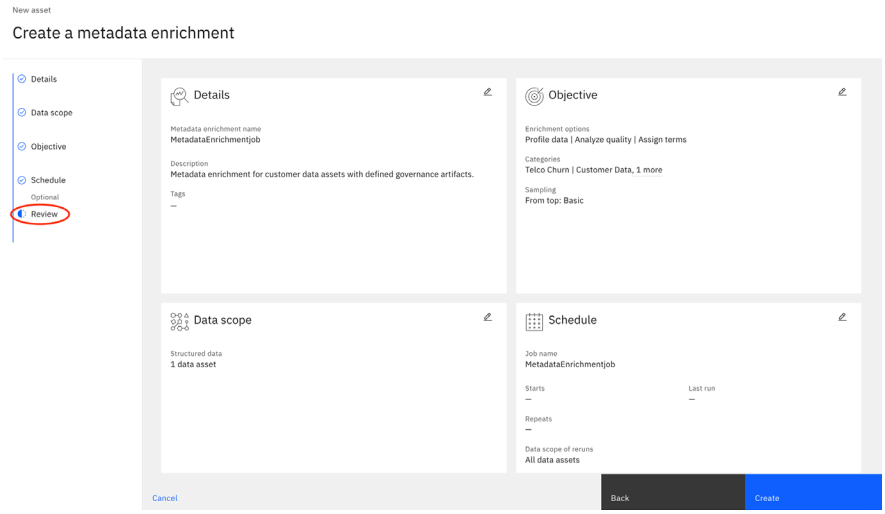
26- Next, on the Schedule step (annotated with red oval), you can define a schedule for periodically running the metadata enrichment.

For this lab, keep the **Schedule off** setting (annotated with red rectangle) and click **Next**.

Note that you can also specify whether the scheduled job should run on all data assets or only the New or modified assets (either newly added data assets or data assets where there are new columns).



27- On the Review step (annotated with red oval), review the various configurations for this metadata enrichment job and click **Create.**

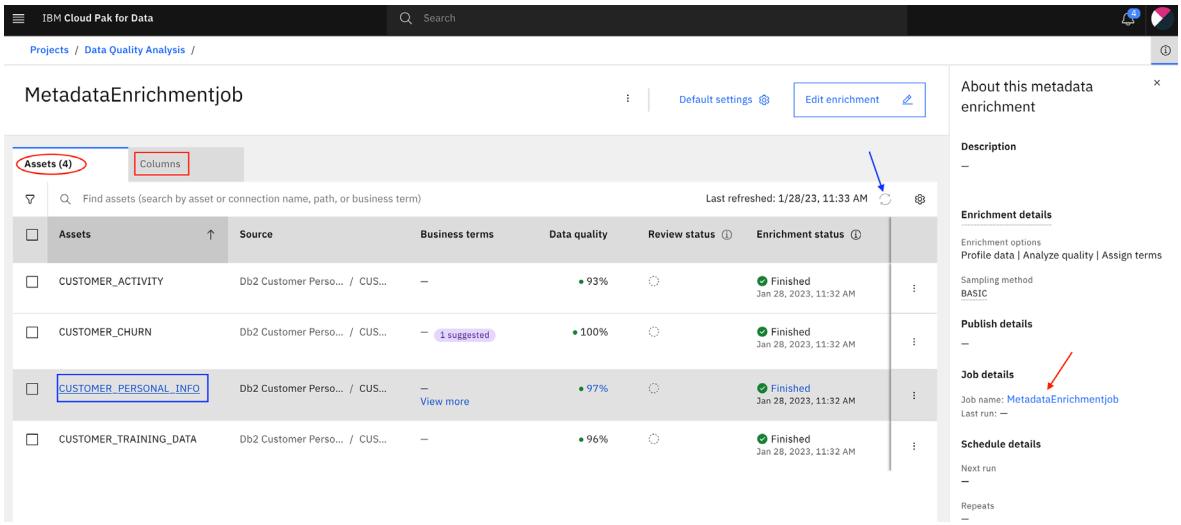


28- Click the Refresh icon (annotated with blue arrow) periodically to refresh status of the metadata enrichment job. The job should not take more than a couple of minutes to finish.
When the job completes, you will see the enrichment and data quality results for the discovered data assets (based on imported metadata).

Click back and forth between the **Assets** view (annotated with red oval) and the **Columns** view (annotated with red rectangle) to see the assigned Business terms and data classes as well as the Data quality results for the discovered assets (tables) or columns (table columns).

Also, please note the Job name under Job details (annotated with red arrow) which you can click to view the status of the running job.

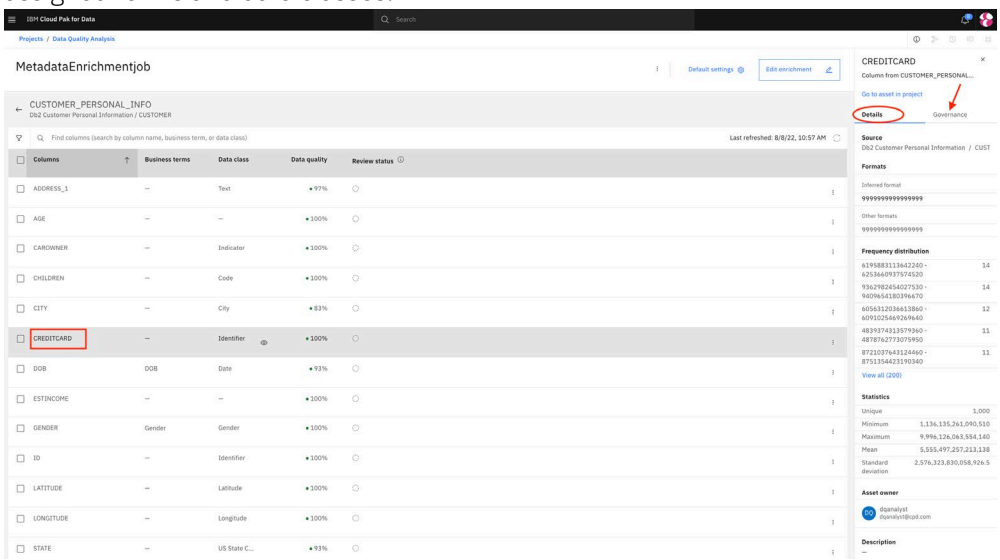
Click the **CUSTOMER_PERSONAL_INFO** asset (annotated with blue rectangle) to dive deeper into the analysis results for that table.



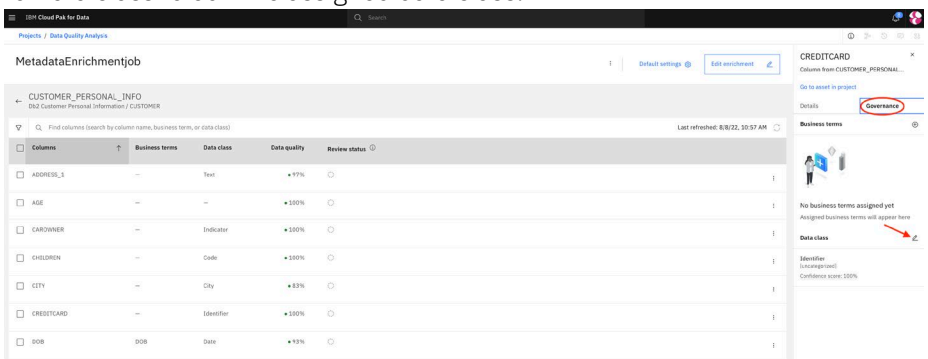
29- Review the enrichment results for the CUSTOMER_PERSONAL_INFO table.

Specifically, select the **CREDITCARD** column (annotated with red rectangle) and review the Details for that column in terms of inferred data format, frequency distribution, and statistics.

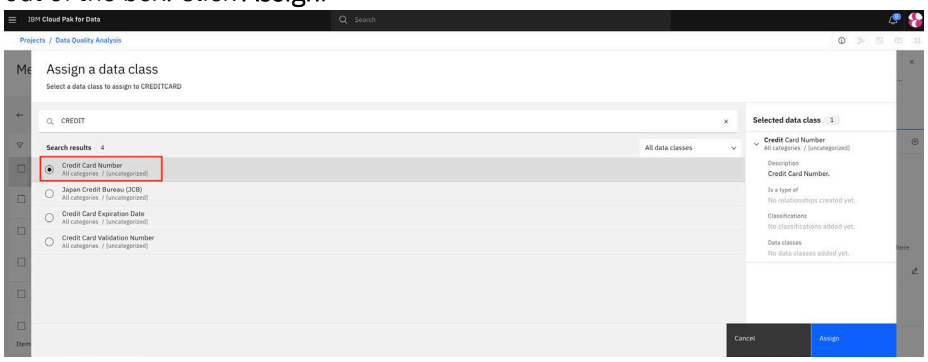
Next click the **Governance** tab (annotated with red arrow) for CREDITCARD column to review the assigned terms and data classes.



30- On the Governance tab (annotated with red oval), click the **pencil icon** (annotated with red arrow) next to Data class to edit the assigned data class.

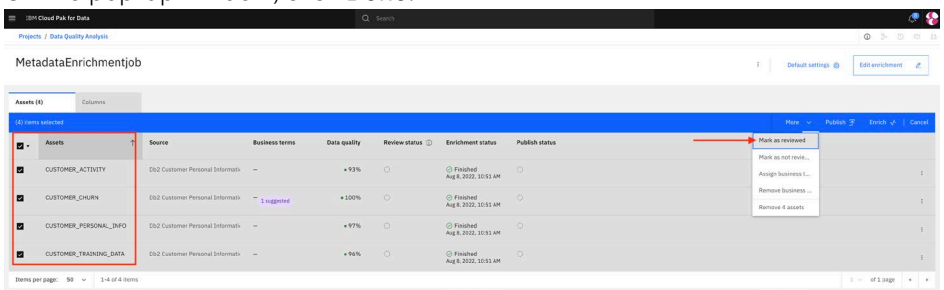


31- On the *Assign a data class* page, search for **CREDIT** term and then select the **Credit Card Number** data class (annotated with red rectangle) which is one of the pre-defined data classes available with WKC out of the box. Click **Assign**.

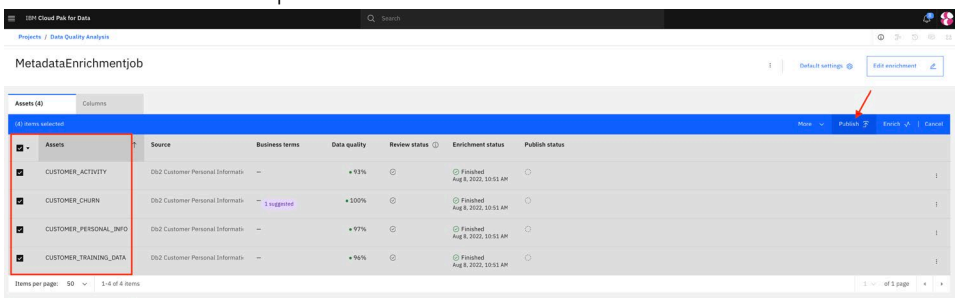


32- After reviewing the details for the data assets and columns in terms of statistics and governance enrichments, navigate back to the **Assets** view and select **all the discovered assets** (annotated with red rectangle) and then select **More ➔ Mark as reviewed** (annotated with red arrow).

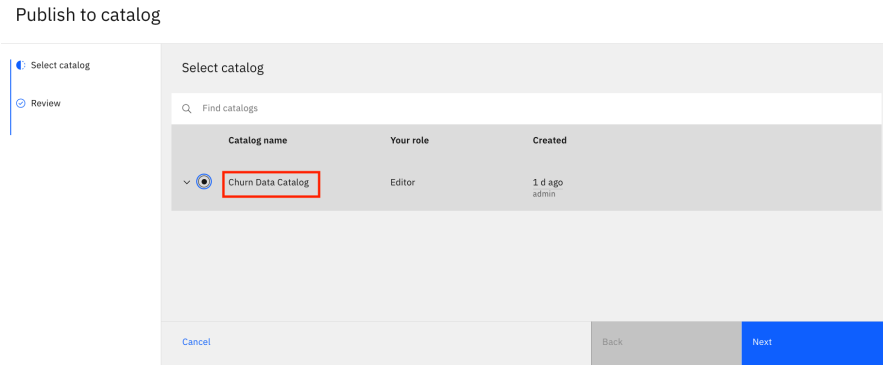
On the pop-up window, click **Done**.



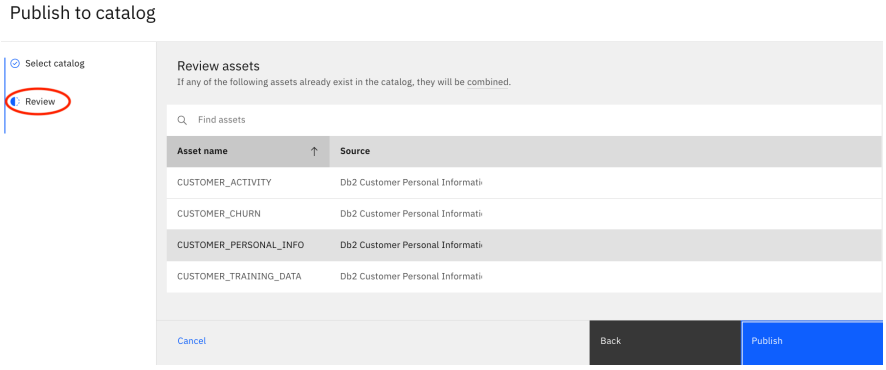
33- Now that you've reviewed, fixed/updated, and verified the data assets in terms of governance enrichments and quality results, **select all data assets** (annotated with red rectangle) and click **Publish** (annotated with red arrow) to publish assets to the catalog and make available for other data consumers in the enterprise.



34- On the Publish to catalog page, select the **Churn Data Catalog** (annotated with red rectangle) and click **Next**.



35- On the Review step (annotated with red oval), review the assets and click **Publish**.



36- Monitor the Publish job until it is completed.

Summary

To quickly re-cap, we've illustrated how to run metadata import and metadata enrichment to discover and enrich the data assets behind the connected data sources as well as assess the data quality at the asset level and the column level. Correct enrichment of governance artifacts helps make the data more discoverable and ready for self-service by data consumers across the enterprise. Guaranteeing high quality data is critical for a Governed MLOps methodology that delivers trust in AI models so business leaders can adopt such models confidently.