



Governed MLOps Workshop
**Data Privacy, Quality and
Metadata Enrichment
with Watson Knowledge Catalog**

Document version: June 2023

DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift is a trademark of Red Hat, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

© 2023 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.

Table of Contents

Data Quality and Data Privacy – Introduction..... 4

Data Privacy via Data Protection Rules..... 4

Data Quality and Data Discovery via Metadata Import and Enrichment 7

Summary 23

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:

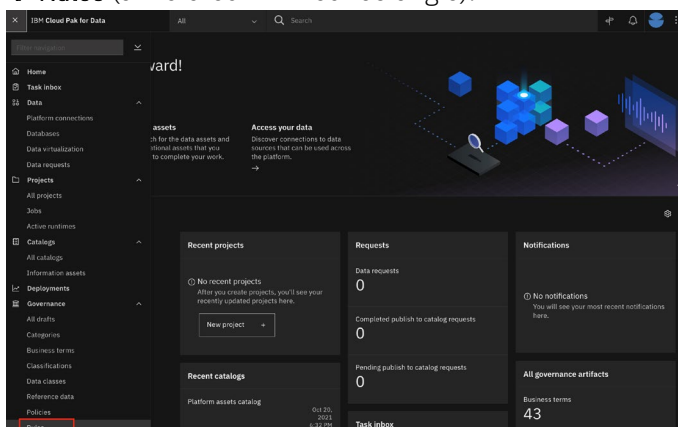
- 1- Apply data protection rules to enforce masking and protection of sensitive data such as PII, Personally Identifiable Information, data.
- 2- Review and improve the quality of the data stored in these data sources.

Data Privacy via Data Protection Rules

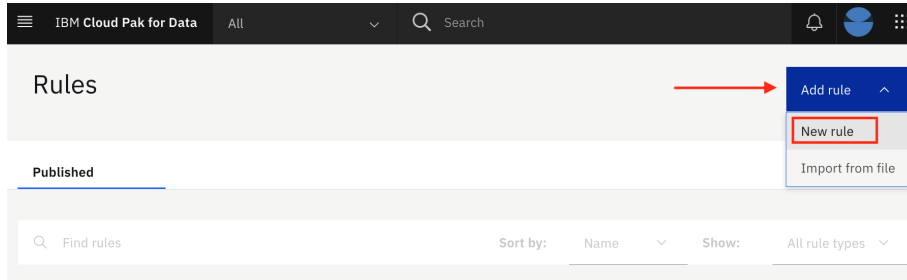
In Cloud Pak for Data, Watson Knowledge Catalog (WKC) enables data masking by applying data protection rules to discovered data assets based on a variety of conditions such as detected data classes and business terms. For example, a data protection rule can state that a data column should be redacted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

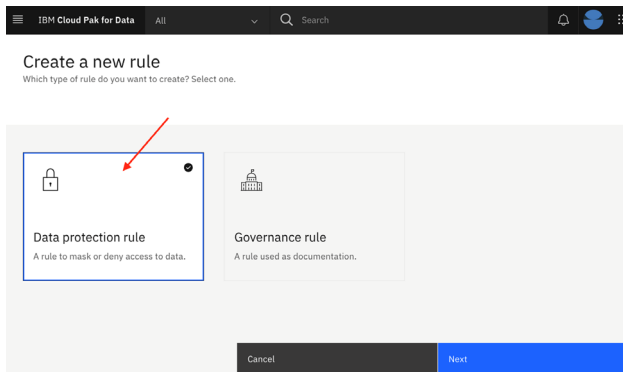
- 1- Login into Cloud Pak for Data as *datasteward* user.
- 2- Navigate to Rules by clicking the Navigation menu (top left hamburger icon) and selecting **Governance** → **Rules** (annotated with red rectangle).



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

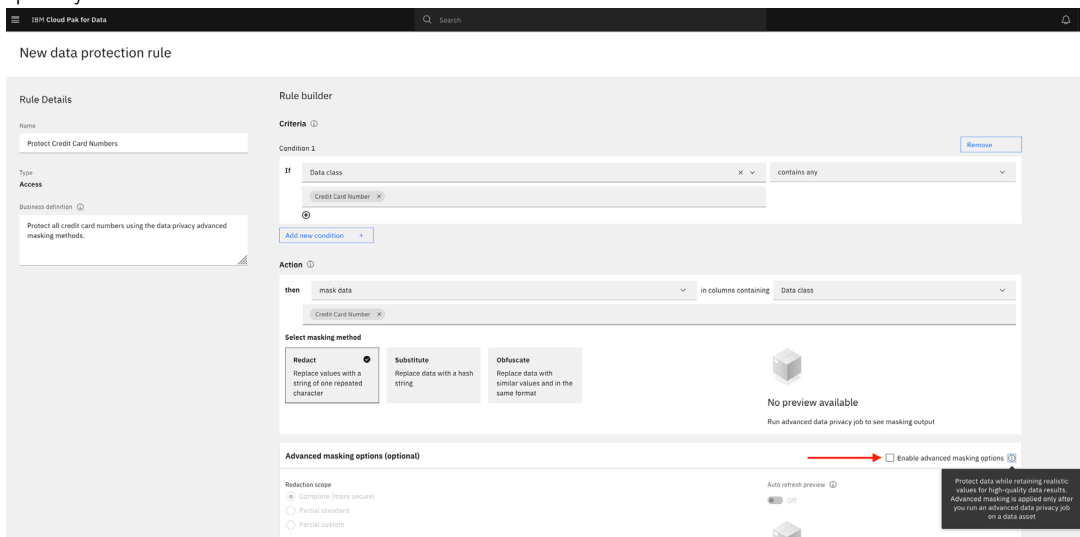


- On the **Create a new rule page**, select the **Data protection rule** (annotated with red arrow) and click **Next**.



- 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.

IBM Cloud Pak for Data

New data protection rule

Rule Details

Name: Protect Date of Birth

Type: Access

Business definition: Obfuscate date of birth values.

Rule builder

Criteria

Condition 1

If Data class contains any Date of Birth

Action

then mask data in columns containing Data class

Date of Birth

Select masking method

Redact
Replace values with a string of one repeated character

Substitute
Replace data with a hash string

Obfuscate
Replace data with similar values and in the same format

No preview available
Run advanced data privacy job to see masking output

Advanced masking options (optional)

Obfuscate method: Preserve format (default)

Auto refresh preview: Off

Cancel Create

Data Quality and Data Discovery via Metadata Import and Enrichment

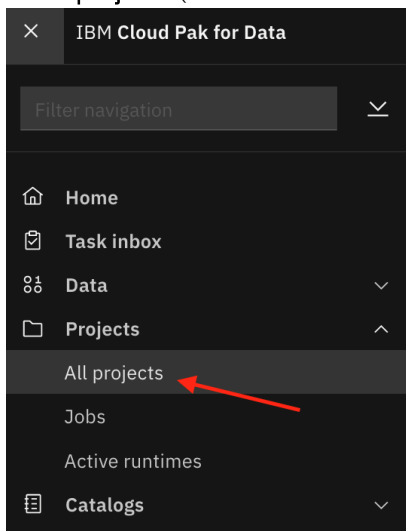
With IBM Cloud Pak for Data, you can analyze and curate data to ensure and improve data quality. During analysis, you can identify or assign data classes and business terms, or process automatically assigned data classes and business terms.

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.

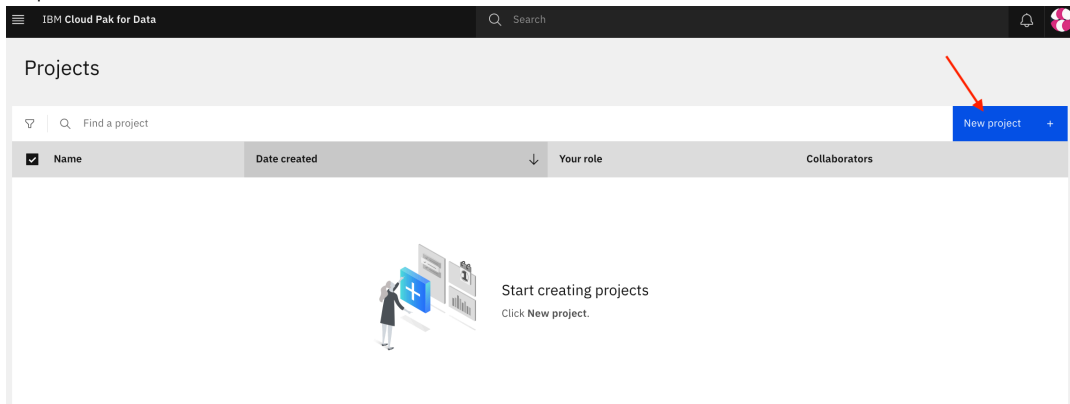
The pre-defined Data Quality Analyst role has the required permission to run and execute metadata import and enrichment jobs in order to evaluate the quality of the data discovered from the various data sources.

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.

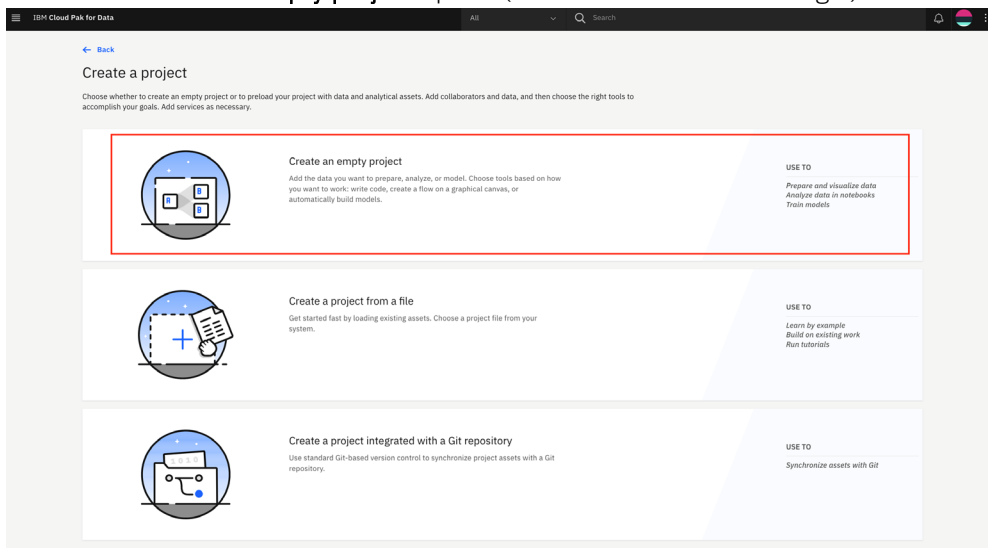
- 1- Login into Cloud Pak for Data as *dqanalyst* user.
- 2- Select All projects by clicking on the Navigation menu (top left hamburger icon) and selecting **Projects** → **All projects** (annotated with red arrow).



- 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.



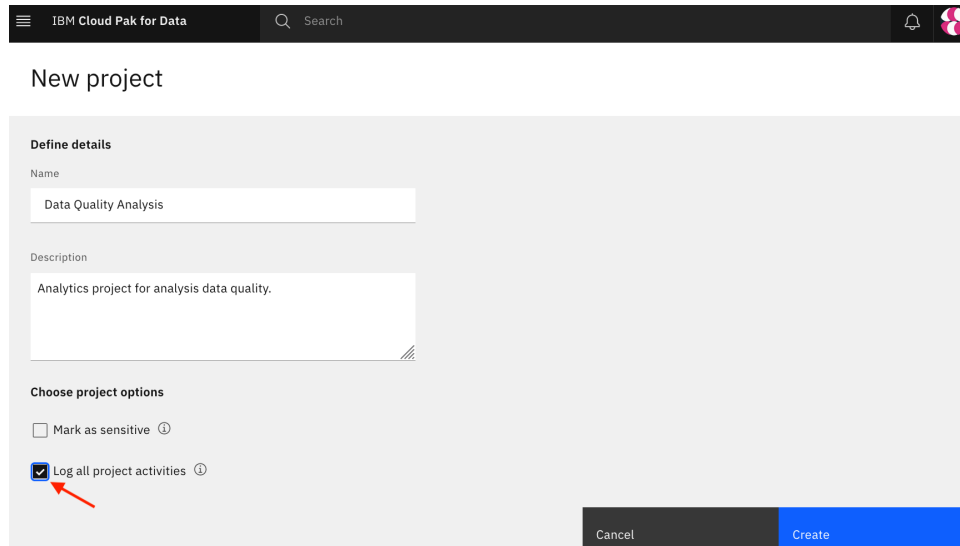
- 4- Select the **Create an empty project** option (annotated with red rectangle).



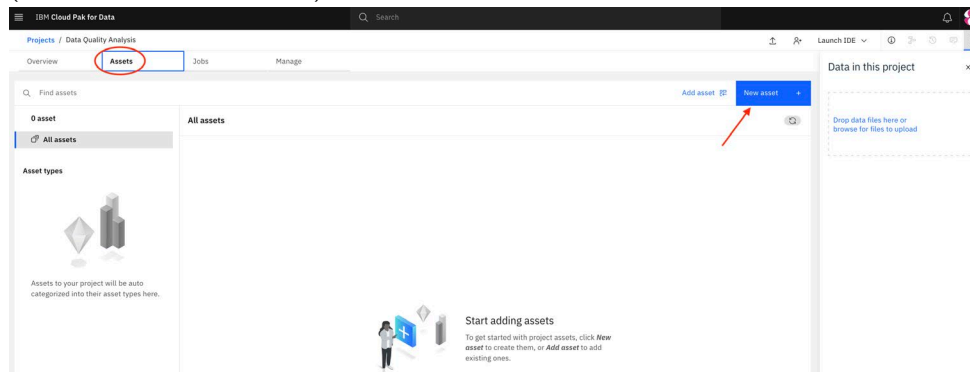
- 5- On the **New project** page, provide a **Name** and **Description** (optional) for the project.

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

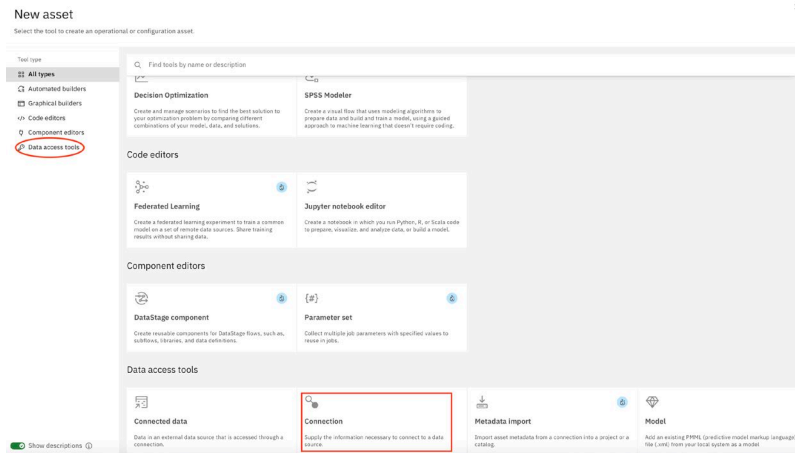
Click **Create**.



- 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.



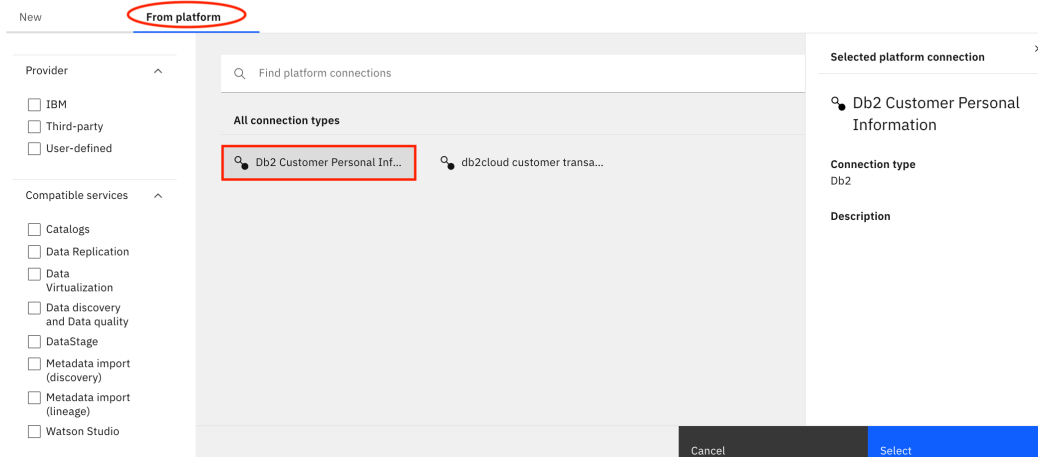
- 8- On the New connection page, select the **From platform** tab (annotated with red oval) to select from the data source connections already created at the platform level.

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.

New connection

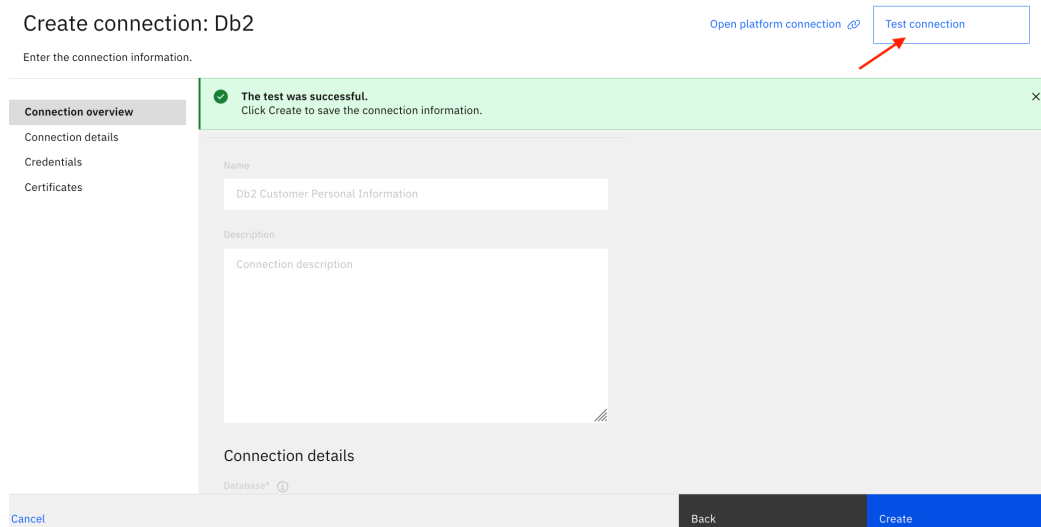
Create a new connection or select an existing connection from the list of platform connections.

[Supported connection types](#)



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

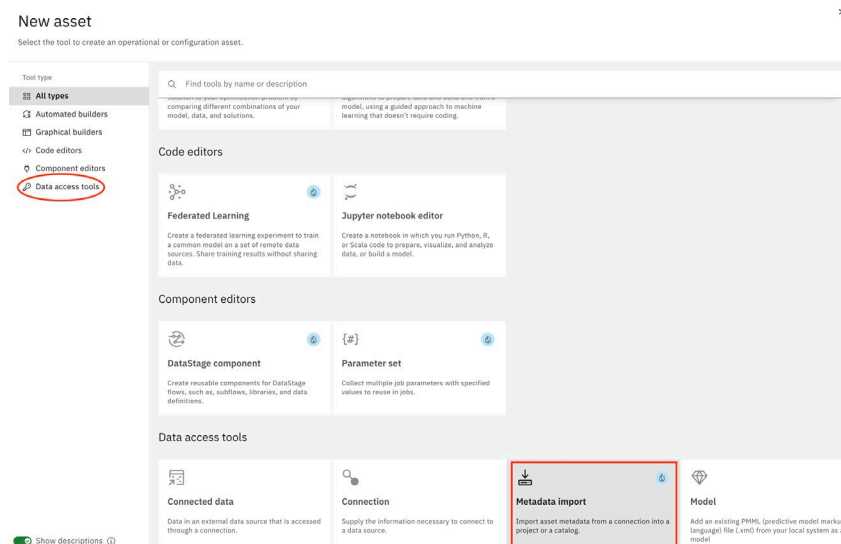
Once you get the green confirmation bar indicating successful connection, click **Create**.



- 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**.

New asset

Create a metadata import

- Define details
- Select target
- Select scope
- Set schedule
- Optional
- Review import

Define details

Specify some basic information about your metadata import to make it easy to identify.

Name

Description (optional)

Tags (optional)

Add tags to make assets easier to find.

[Cancel](#)
[Back](#)
[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.

New asset

Create a metadata import

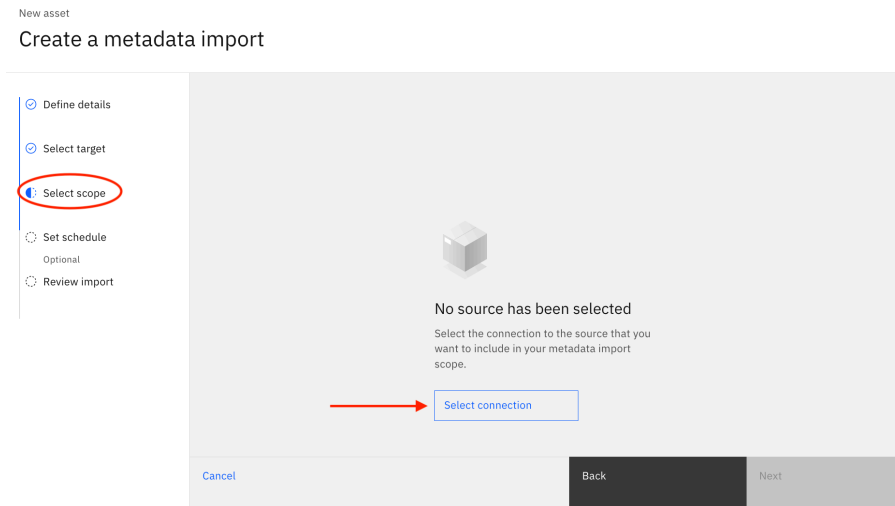
- Define details
- Select target
- Select scope
- Set schedule
- Optional
- Review import

Select target

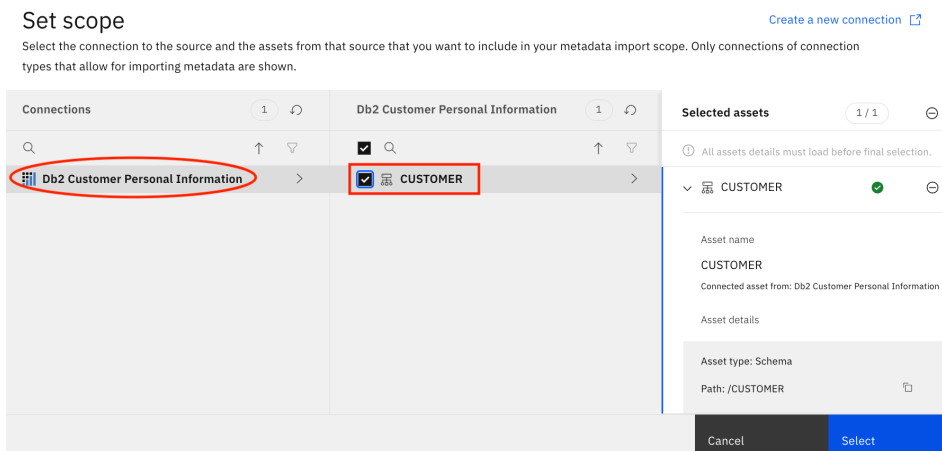
☒ This project (Data Quality Analysis)
 ☐ Catalog

[Cancel](#)
[Back](#)
[Next](#)

- 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**.

New asset

Create a metadata import

- Define details
- Select target
- Select scope
- Set schedule
- Optional
- Review import

Selected assets

1 assets

1 item selected Delete Cancel

<input checked="" type="checkbox"/>	Name	Type	Context
<input checked="" type="checkbox"/>	CUSTOMER	schema	/CUSTOMER

Items per page: 10 1-1 of 1 items 1 of 1 pages

Cancel Back 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**.

New asset

Create a metadata import

- Define details
- Select target
- Select scope
- Set schedule
- Optional
- Review import

Schedule (optional)

Job name

MetadataImportJob job

☐ Schedule off

Leave this turned off if you want to run the job on demand.

Cancel Back Next

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

New asset

Create a metadata import

- Define details
- Select target
- Select scope
- Set schedule
- Optional
- Set advanced options
- Optional
- Review import

Advanced options

Update on reimport

By default, all properties are updated when assets are reimported. If you want any of the following properties to remain unchanged, clear the respective check box.

- ☒ Asset name
- ☒ Asset description
- ☒ Column description

Cancel Back Next

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

New asset

Create a metadata import

Define details

Select target

Select scope

Set schedule

Optional

Review import

Details

Metadata import name
MetadataImportJob

Description
Sample job for importing metadata

Target

Project
Data Quality Analysis

Scope

Connection
DB2 Customer Personal Information

Selected assets
Data assets: 1

Schedule

Job name
MetadataImportJob job

No schedule created

Cancel

Back

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).

IBM Cloud Pak for Data

Projects / Data Quality Analysis / MetadataImportJob

MetadataImportJob

Metadata import

Imported assets

4 assets

Name	Type	Context	Last imported	Status
CUSTOMER_PERSONAL_INFO	Relational table	CUSTOMER/CUSTOMER_PERSONAL_INFO	Aug 08, 2022, 09:16 AM	Imported
CUSTOMER_ACTIVITY	Relational table	CUSTOMER/CUSTOMER_ACTIVITY	Aug 08, 2022, 09:16 AM	Imported
CUSTOMER_CHURN	Relational table	CUSTOMER/CUSTOMER_CHURN	Aug 08, 2022, 09:16 AM	Imported
CUSTOMER_TRAINING_DATA	Relational table	CUSTOMER/CUSTOMER_TRAINING_DATA	Aug 08, 2022, 09:16 AM	Imported

Items per page: 20 | 1-4 of 4 items | 1 of 1 pages

Summary message: Metadata import complete. 4 assets were imported successfully.

About this metadata import

Description
Sample job for importing metadata

Import details

Asset
Discoverer

Connection
DB2 Customer Personal Information

Scope
Assets: 1

Import target
Data Quality Analysis

Job details

Job name: MetadataImportJob job

Last run: Run 1
Aug 08, 2022, 09:16 AM

Schedule
No schedule configured

Related assets

@ MetadataImportJob job

Tags
No tags added yet.

Created by
dipanjyot, Aug 08, 2022, 09:16 AM

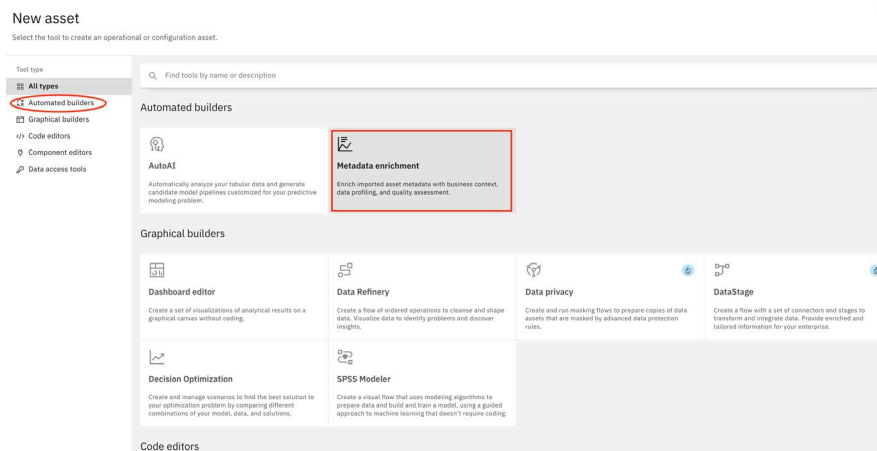
Modified by
Not applicable, Aug 08, 2022, 09:16 AM

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

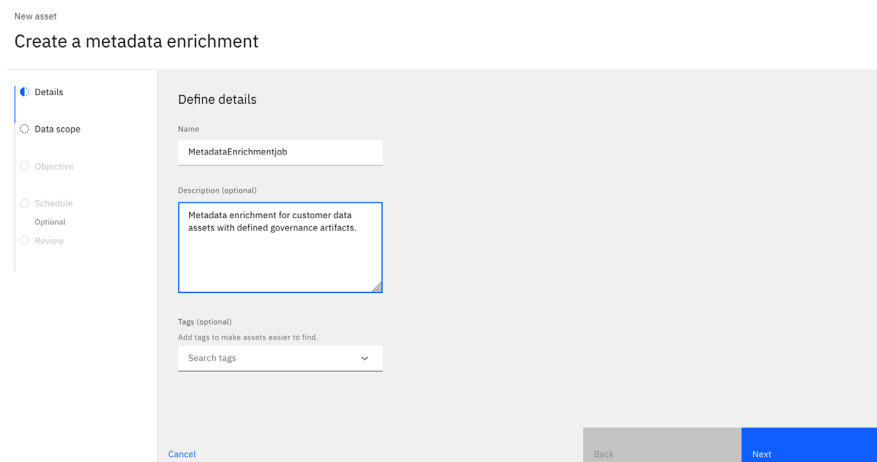
Click on Data Quality Analysis project breadcrumb and on the project's Assets page, click **New asset** again.

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

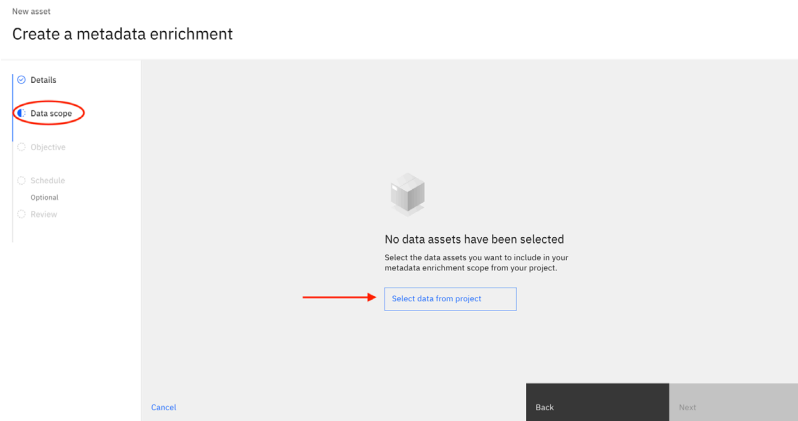
You can also select **Automated builders** (annotated with red oval) under Tool type to filter asset types available for faster access to the Metadata enrichment function.



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

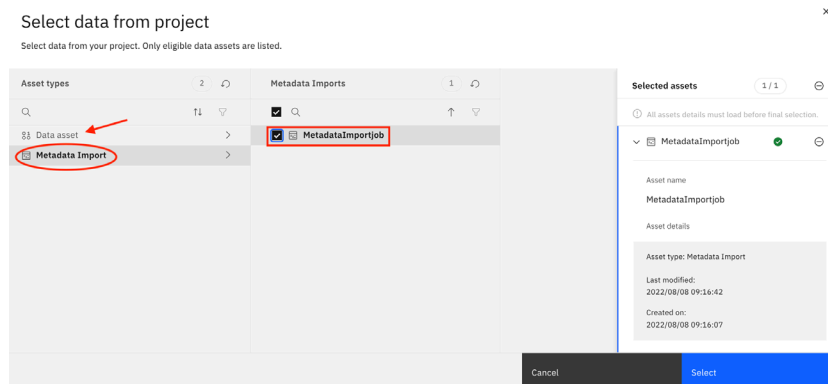


22- On the Data scope step (annotated with red oval), click **Select data from project** button (annotated with red arrow).

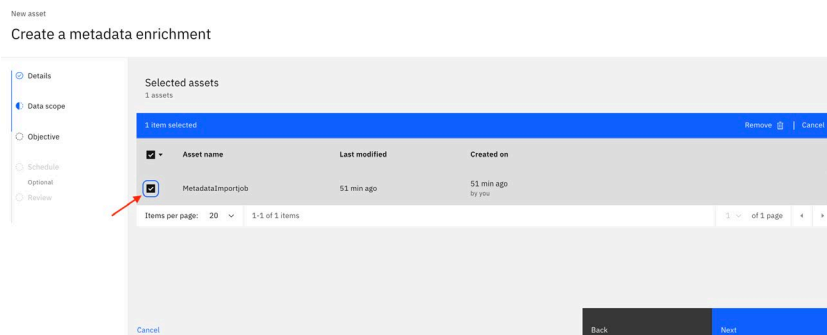


23- On the Select data from project page, you can either select specific data assets from the project (annotated with red arrow) or you can select the Metadata Import job (annotated with red oval) to apply enrichment to all the data assets imported via that Metadata import job.

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).

New asset

Create a metadata enrichment

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

New asset

Create a metadata enrichment

- Details
- Data scope
- Objective
- Schedule
- Optional
- Review**

Details

Metadata enrichment name
MetadataEnrichmentjob

Description
Metadata enrichment for customer data assets with defined governance artifacts.

Tags
—

Objective

Enrichment options
Profile data | Analyze quality | Assign terms

Categories
Telco Churn | Customer Data, 1 more

Sampling
From top: Basic

Data scope

Structured data
1 data asset

Schedule

Job name
MetadataEnrichmentjob

Starts
—

Repeats
—

Data scope of reruns
All data assets

Cancel

Back 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.

IBM Cloud Pak for Data

Projects / Data Quality Analysis /

MetadataEnrichmentjob

Assets (4) Columns

Find assets (search by asset or connection name, path, or business term)

Last refreshed: 1/28/23, 11:33 AM

Assets	Source	Business terms	Data quality	Review status	Enrichment status
CUSTOMER_ACTIVITY	Db2 Customer Perso... / CUS...	—	93%	Finished	Jan 28, 2023, 11:32 AM
CUSTOMER_CHURN	Db2 Customer Perso... / CUS...	1 suggested	100%	Finished	Jan 28, 2023, 11:32 AM
CUSTOMER_PERSONAL_INFO	Db2 Customer Perso... / CUS...	View more	97%	Finished	Jan 28, 2023, 11:32 AM
CUSTOMER_TRAINING_DATA	Db2 Customer Perso... / CUS...	—	96%	Finished	Jan 28, 2023, 11:32 AM

About this metadata enrichment

Description

Enrichment details

Enrichment options
Profile data | Analyze quality | Assign terms

Sampling method
BASIC

Publish details

Job details

Job name: MetadataEnrichmentjob

Last run: —

Schedule details

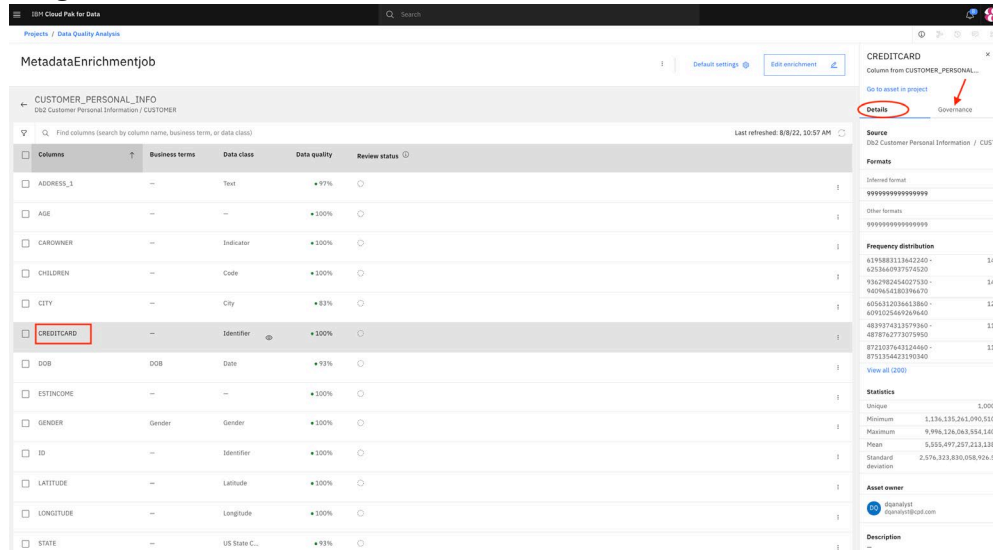
Next run
—

Repeats
—

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.

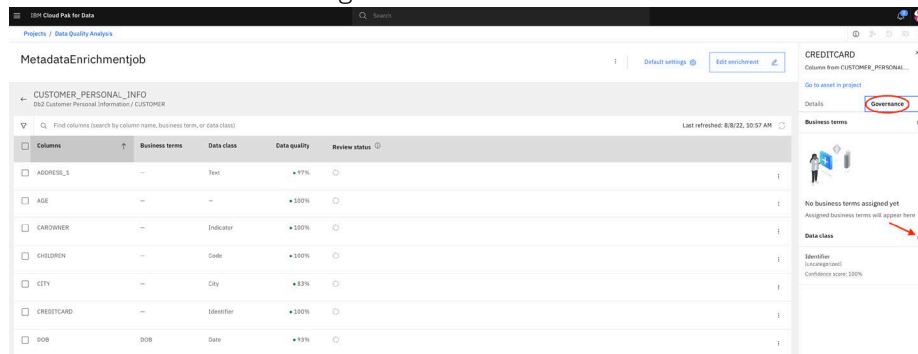
Next click the **Governance** tab (annotated with red arrow) for CREDITCARD column to review the assigned terms and data classes.



The screenshot displays the 'MetadataEnrichmentJob' results for the 'CUSTOMER_PERSONAL_INFO' table. The table lists various columns with their business terms, data classes, data quality, and review status. The 'CREDITCARD' column is highlighted with a red rectangle. The right sidebar shows the 'Details' tab for the 'CREDITCARD' column, which includes the following information:

- Source:** DPM Customer Personal Information / CUSTOMER_PERSONAL_INFO
- Inferred format:** 9999999999999999
- Frequency distribution:** A table showing the distribution of values for the CREDITCARD column.
- Statistics:** A table showing the statistical properties of the CREDITCARD column.
- Asset owner:** A user icon and name.
- Description:** A text field for describing the column.

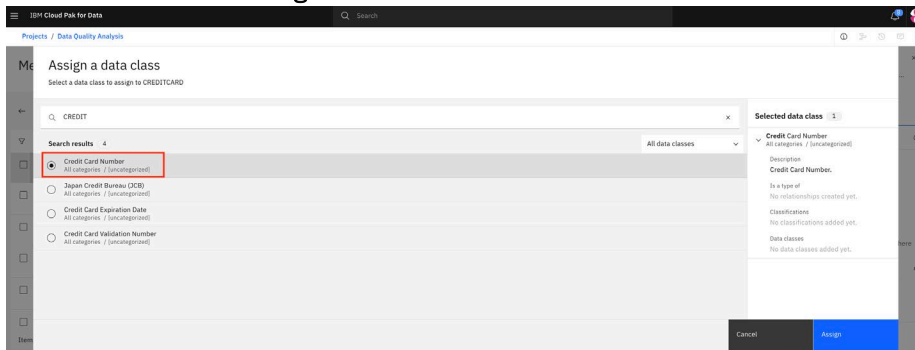
30- On the Governance tab (annotated with red oval), click the pencil icon (annotated with red arrow) next to Data class to edit the assigned data class.



The screenshot displays the 'MetadataEnrichmentJob' results for the 'CUSTOMER_PERSONAL_INFO' table. The table lists various columns with their business terms, data classes, data quality, and review status. The 'CREDITCARD' column is highlighted with a red rectangle. The right sidebar shows the 'Governance' tab for the 'CREDITCARD' column, which includes the following information:

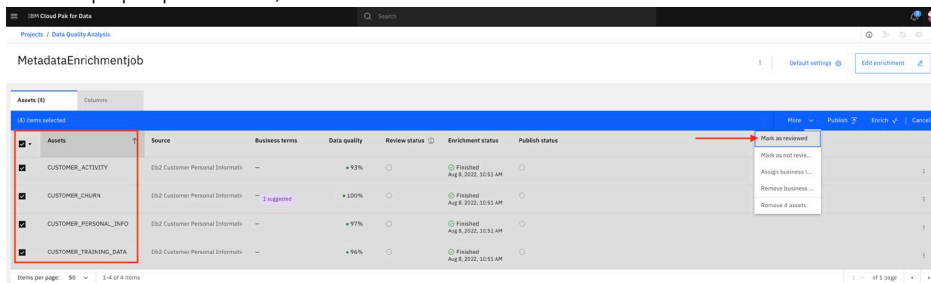
- Business terms:** A section for managing business terms associated with the column.
- Data class:** A section for managing the data class assigned to the column. A red oval highlights the 'Governance' tab, and a red arrow points to the pencil icon next to the 'Data class' label.

- 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**.

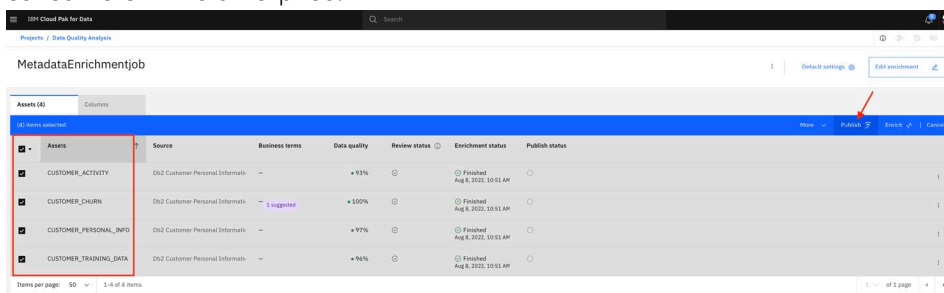


- 32- After reviewing the details for the data assets and columns in terms of statistics and governance enrichments, navigate back to the **Assets** view and select **all the discovered assets** (annotated with red rectangle) and then select **More → Mark as reviewed** (annotated with red arrow).

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

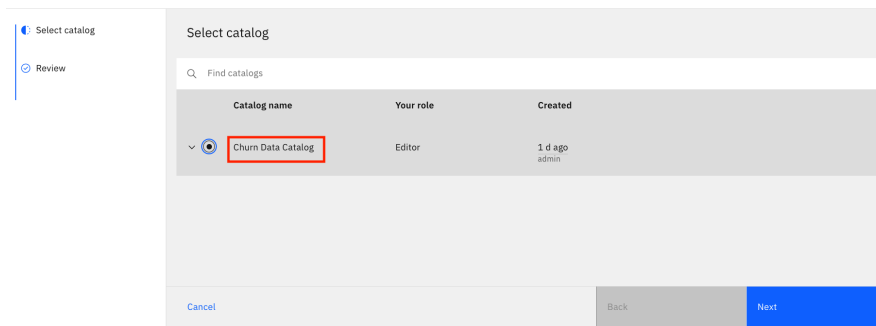


- 33- Now that you've reviewed, fixed/updated, and verified the data assets in terms of governance enrichments and quality results, **select all data assets** (annotated with red rectangle) and click **Publish** (annotated with red arrow) to publish assets to the catalog and make available for other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



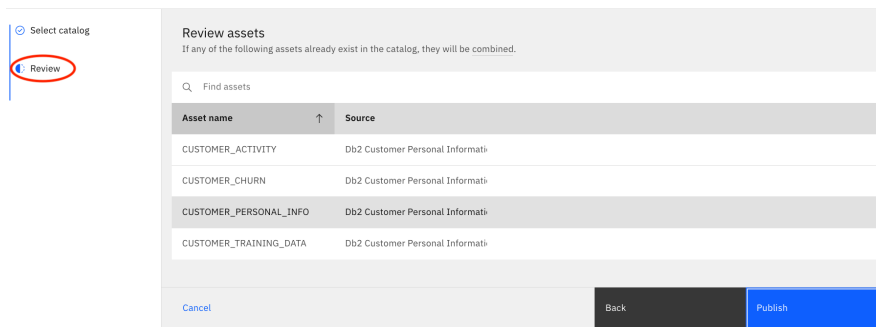
34- On the Publish to catalog page, select the **Churn Data Catalog** (annotated with red rectangle) and click **Next**.

Publish to catalog



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

Publish to catalog



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.