

# Lab Guide

## Data Governance and Privacy Watson Knowledge Catalogue

Cloud Pak for Data

IBM CSM ASEAN  
December 2022

## Table of Contents

<b>OVERVIEW.....</b>	<b>3</b>
<b>GETTING STARTED .....</b>	<b>4</b>
<b>PART 1: TRUST YOUR DATA .....</b>	<b>7</b>
<b>PART 2: PROTECT YOUR DATA .....</b>	<b>14</b>
<b>PART 3: KNOW YOUR DATA .....</b>	<b>17</b>

## Overview

Data governance is the overall management of data availability, relevancy, usability, integrity, and security in an enterprise. It helps organization manage their information, knowledge, and answer questions such as:

- What data do we have?
- What do we know about our information?
- Does this data adhere to company policies and rules?
- What is the quality of our data?

Take this lab to learn how to prepare trusted data with the Data Governance and Privacy use case of the data fabric. Your goal is to create trusted data assets by enriching your data and running data quality analysis.

The story for the lab is that Golden Bank has several departments that need access to high-quality customer mortgage data. As a Data Steward in the data governance team, you must sort and organize the company's data to provide high-quality and protected data assets that data consumers can easily find in a self-service catalog. In this lab, you will complete these tasks:

1. Create a data catalog.
2. Create a category.
3. Add business terms.
4. Import data into the project.
5. Enrich the data.
6. View the results of the metadata enrichment.
7. Publish assets to a catalog.

## Getting Started

### Sign up or log into Cloud Pak for Data

You must sign up for Cloud Pak for Data as a Service and provision the necessary services for the Data Governance and Privacy use case.

For this lab, you can use the following Cloud Pak for Data account or use a Cloud Pak for Data as a Service account.

Log in into Cloud Pak for Data with the following account:

CP4D URL: <this information will be provided during the practicum>

Username: <this information will be provided during the practicum>

Password: <this information will be provided during the practicum>

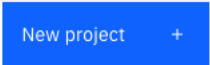
### Provision the necessary services

*Note: You may not need to provision necessary services when you are using the accounts prepared for the labs.*

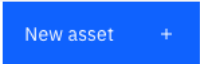
After you log in, follow these steps to verify or provision the necessary services.

1. From the Cloud Pak for Data navigation menu, choose **Services > Instances**.
2. Click **New Instance**, check if there is an existing **Watson Knowledge Catalog** service instance enabled.
3. If you need to create a Watson Knowledge Catalog service instance, click **New instance**.
  0. Select Watson Knowledge Catalog.
    1. Select the plan
    2. Click **Create**.
4. Repeat these steps to verify or provision other necessary services.

### Create a Project and add a Connection

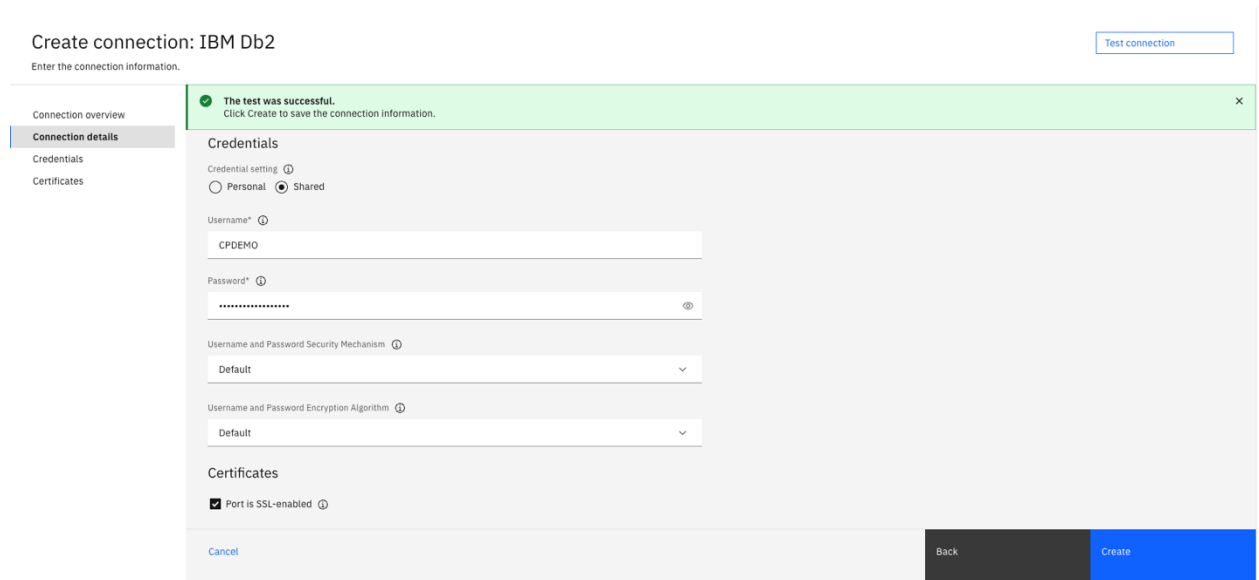
1. Navigate to **All Projects** in the menu **Projects** of the Cloud Pak for Data Navigation
2. Click **New project**. A blue rectangular button with the text "New project" in white, followed by a white plus sign.
3. Choose **Create an empty project**. Fill in the information details of the project
4. Click **Create**.

5. Navigate through different tabs in the project to verify the project. Note: If this is your first time accessing a project, you may see a guided tour asking if you want to tour of projects. For now, click **Maybe later**.

6. Click the **Assets** tab. Click **New asset**  -> **Data access tools > Connection**. Select **IBM Db2**. Fill in the database connection using the following connection information.

- 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
- Certificates: Select Port is **SSL-Enabled**

Click Test connection. If the test is successful. Then Click Create.



Create connection: IBM Db2 Test connection

Enter the connection information.

Connection overview  
**Connection details**  
Credentials  
Certificates

**The test was successful.**  
Click Create to save the connection information.

**Credentials**

Credential setting ⓘ  
☐ Personal ☒ Shared

Username\* ⓘ  
CPDEMO

Password\* ⓘ  
\*\*\*\*\*

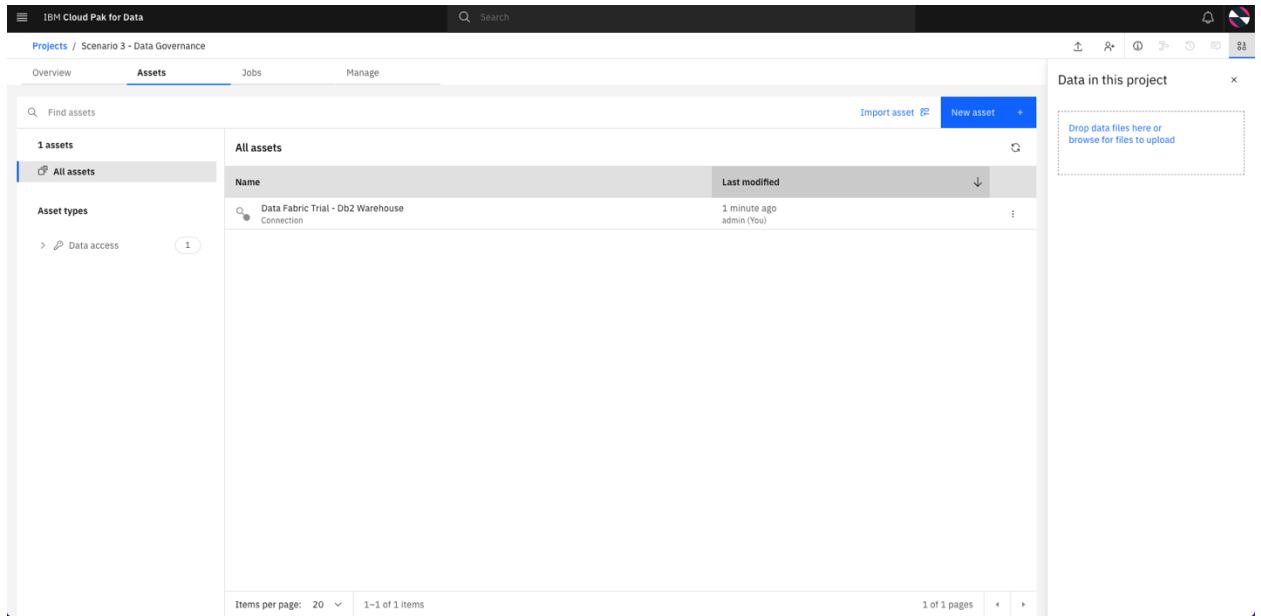
Username and Password Security Mechanism ⓘ  
Default

Username and Password Encryption Algorithm ⓘ  
Default

**Certificates**  
☒ Port is SSL-enabled ⓘ

Cancel Back Create

Once the database is connected, you will see it in Assets.



7. Download the dataset banking.csv in [this Github folder](#) (or in [this box folder](#) in case you cannot download from Github) and save it to your computer. You'll use that file in a later step.

The lab contains three parts:

- Part 1: Trust Your Data
- Part 2: Protect Your Data
- Part 3: Know Your Data

Now, you are ready to begin the lab.

## Part 1: Trust Your Data

### Step 1: Create a catalog

Before you start working with data, create a catalog where you will publish data to share it with your organization. If you already have a catalog, you can skip this step. Otherwise, follow these steps to create a catalog.

**Note:** If this is your first time accessing a catalog, you may see a guided tour asking if you want to tour of catalogs. For now, click **Maybe later**.

1. From the Cloud Pak for Data navigation menu, choose **Catalogs > View all catalogs**.
2. Click **Create Catalog**.
3. For the *Name*, type *Mortgage Approval Catalog*. Type the catalog name, exactly as shown with no leading or trailing spaces.
4. Select **Enforce data protection rules**, confirm the selection, and accept the defaults for the other fields.
5. Click **Create**.

Your catalog is now ready for you to share assets with your organization.

## Step 2: Create a category

You need a category to contain the business terms that you'll import in the next step. Categories act like folders to organize your governance artifacts and the people who can author and manage those artifacts. Follow these steps to create a category.

1. From the Cloud Pak for Data navigation menu, choose **Governance > Categories**.
2. Click **Add category > New category**.
3. For the name, type *Banking*.
4. Click **Create**.

The Banking category is ready for you to import business terms



### Step 3: Add business terms

Now import business terms into the new category. You'll use them to enrich your data assets in a later step. Business terms are standardized definitions of business concepts so that your data is described in a uniform and easily understood way across your enterprise. Follow these steps to import the business terms from a file.

1. From the Cloud Pak for Data navigation menu, choose **Governance > Business terms**.
2. Click **Add business term > Import from file**.
3. Click **Add file**. Select the **banking.csv** file that you downloaded earlier.
4. Click **Next**.
5. Select **Replace all values**, and click **Import**.
6. Click **Go to task** to see the draft business terms. If you miss the notification, then from the Cloud Pak for Data as a Service navigation menu, choose **Task inbox**.
7. Select the **Publish business terms** checkbox, and then click **Publish**. Click **Publish** to confirm.

You are now ready to import the data to a project which you will then enrich with the imported business terms.

## Step 4: Import data to a project

The project includes a connection to a Db2 Warehouse instance which contains the mortgage assets. You can import technical metadata associated with the data assets into a project or a catalog to inventory, evaluate, and catalog these assets. Technical metadata describes the structure of data objects. Follow these steps to import the data assets.

1. From the Cloud Pak for Data navigation menu, choose **Projects > View all projects**.
2. Click the project you created.
3. Click the **Assets** tab.
4. Click **New asset**.
5. Scroll down to select **Metadata Import**.
6. For the name, type *Mortgage data - metadata import*.
7. Click **Next** to continue.
8. For Select target, select **This project**, and click **Next** to continue.
9. For Select scope, click **Select connection**.
  0. Select the **Data Fabric Trial - Db2 Warehouse** connection.
    1. Click the **WKC\_MORTGAGE** schema.
    2. Select the following tables:
      - COMMERCIAL\_CLIENT
      - CREDIT\_SCORE
      - HOUSE\_PRICE
      - MORTGAGE\_APPLICANTS
      - MORTGAGE\_APPLICATION
    3. Review the list of tables in the side panel, and then click **Select**.
10. Click **Next** to continue to the schedule.
11. Click **Next** to continue to the review.
12. Review the summary of the import, and click **Create**. The import job will start.
13. Click the **refresh** icon to watch the status change from Queued to In progress to Imported. When the job run is complete, you will see the five assets listed.

IBM Cloud Pak for Data

Projects / Scenario 3 - Data Governance / Mortgage data - metadata import

Metadata import complete, 5 assets were imported successfully.

Mortgage data - metadata import

Metadata import

Imported assets

5 assets

Name	Type	Context	Last imported	Status
MORTGAGE_APPLICANTS	Relational table	WKC_MORTGAGE/MORTGAGE_APPLICANTS	Dec 23, 2022, 07:11 AM	Imported
HOUSE_PRICE	Relational table	WKC_MORTGAGE/HOUSE_PRICE	Dec 23, 2022, 07:11 AM	Imported
CREDIT_SCORE	Relational table	WKC_MORTGAGE/CREDIT_SCORE	Dec 23, 2022, 07:11 AM	Imported
MORTGAGE_APPLICATION	Relational table	WKC_MORTGAGE/MORTGAGE_APPLICATION	Dec 23, 2022, 07:11 AM	Imported
COMMERCIAL_CLIENT	Relational table	WKC_MORTGAGE/COMMERCIAL_CLIENT	Dec 23, 2022, 07:11 AM	Imported

Items per page: 20 1-5 of 5 items 1 of 1 pages

About this metadata import

Description

This is a test

Import details

Goal

Discover

Connection

Data Fabric Trial - Db2 Warehouse

Scope

Assets: 5

Import target

Scenario 3 - Data Governance

Job details

Job name: Mortgage data - metadata impo...

Last run: Run 1

Dec 23, 2022, 07:11 AM

Schedule

No schedule configured

Related assets

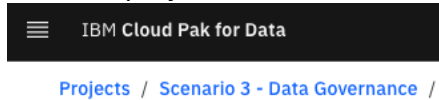
Mortgage data - metadata import job

Your next step is to enrich the imported data assets with the imported business terms.

## Step 5: Enrich the imported data

You can enrich data assets with information that helps users to 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. Follow these steps to enrich the imported data.

1. Click the project name in the navigation trail.



2. On the Assets tab, click **New asset**.
3. Select **Metadata enrichment**.
4. For the name, type *Mortgage data - metadata enrichment*.
5. Click **Next** to continue.
6. Click **Select data from project**.
  0. Select **Metadata Import**.
    1. Select **Mortgage data - metadata import** which includes the following assets:
      - COMMERCIAL\_CLIENT
      - CREDIT\_SCORE
      - HOUSE\_PRICE
      - MORTGAGE\_APPLICANTS
      - MORTGAGE\_APPLICATION
    2. Click **Select**.
  7. Click **Next** to continue to the enrichment objective.
  8. Select all enrichment objectives:
    0. Profile data
      1. Analyze quality
      2. Assign terms
  9. Click **Select categories**.
    0. Select **[uncategorized]** and **Banking** (or the category you created earlier).
      1. Click **Select**.
  10. For the Sampling, select **Basic**.
  11. Click **Next** to continue to the schedule.
  12. Click **Next** to continue to the review.
  13. Click **Create**.
  14. The metadata enrichment asset will display. The job may take several minutes to complete. Click the **refresh** icon to watch the status change from Queued to In progress to Finished. When the job run is complete, you will see the five assets listed.

## Step 6: View the results of the metadata enrichment

After Metadata enrichment run is completed, follow these steps to view the enriched data.

1. From the Mortgage data - metadata enrichment screen, click the **Columns** tab.
2. In the list of Columns, locate the **EMAIL\_ADDRESS** column for the MORTGAGE\_APPLICANTS asset.
  0. Click the **Three dots** menu at the end of the EMAIL\_ADDRESS for MORTGAGE\_APPLICANTS row, and choose **View column details**.
    1. In the side panel on the Details tab, you will see profiling information such as: Format, Frequency distribution, Statistics.
    2. In the side panel, click the **Governance** tab. This tab includes the data classes and business terms that were auto-assigned during the metadata enrichment. Note that you may also see suggested business terms and data classes, and manually assign them.
  3. To review the suggested terms and manually assign them:
    0. Click **Suggested business terms**.
      1. For Address, click **Assign**.
3. At the end of the EMAIL\_ADDRESS column for the MORTGAGE\_APPLICANTS asset row, click the **Three dots** menu, and choose **View data quality**. Watson Knowledge Catalog automatically generates a data quality score for each column and data asset by analyzing every value in every record according to pre-built dimensions.
4. Click the **X** to close the Data quality window.
5. For the CITY column for the CREDIT\_SCORE asset, click the **Three dots** menu, and choose **Mark as reviewed**.
6. Click the **Assets** tab.
7. In the list of Assets, for the MORTGAGE\_APPLICANTS asset, click the **Three dots** menu, and choose **View asset details**.
  0. In the side panel, click the **Governance** tab to see business term auto assignment.
    1. To manually assign business terms, click the **Edit** icon.
    2. Search for *social*. If there are no results, then make sure the dropdown list is set to **All terms** instead of Suggested terms.
    3. Select **Social Security Number**.
    4. Click **Assign**.

The next step is to publish the enriched data to a catalog to share with your organization.

## Step 7: Publish data to a catalog

Now that you have enriched data, you want to publish those data assets to a catalog so data scientists and data analysts can use the enriched data assets. Follow these steps to store the enriched data assets in a catalog for others to have access to the trusted data.

1. Click the project name in the navigation trail.
2. Click the **Assets** tab.
3. Navigate to **Data > Data asset**.
4. Select the **COMMERCIAL\_CLIENT**, **HOUSE\_PRICE**, **MORTGAGE\_APPLICANTS**, and **MORTGAGE\_APPLICATION** data assets from the list, and click Publish to catalog.
  0. For the Target catalog, select **Mortgage Approval Catalog**.
    1. For the MORTGAGE\_APPLICANTS asset, click the **pencil** icon, and change the name to MORTGAGE\_APPLICANTS\_TRUST.
    2. For the Tag, type trusted, and click + (plus sign).
    3. Notice that the data asset and the connection asset will be added to the catalog. Click **Publish**.
5. Deselect all checked assets, then select the checkbox next to the **CREDIT\_SCORE** asset from the list, and click **Publish to catalog**.
  0. For the Target catalog, select **Mortgage Approval Catalog**.
    1. For the Tag, type confidential, and click + (plus sign).
    2. For the Tag, type trusted, and click + (plus sign).
    3. Click **Publish**.
6. From the Cloud Pak for Data navigation menu, choose **Catalogs > View all catalogs**.
7. Click **Mortgage Approval Catalog**.
8. In the Filter by > Any tag drop down list, select **trusted**. Verify that the five data assets were added to the catalog.

As a Data Steward on the governance team, you have learned how to sort and organize the company's data in order to provide high-quality and protected data assets that data consumers can easily find in a self-service catalog.

## Part 2: Protect Your Data

Take this lab to protect your data with the Data Governance and Privacy use case of the data fabric trial. Your goal is to control access to data across services in the data fabric.

The story for the tutorial is that Golden Bank has several departments that need access to high-quality customer mortgage data. As a Data Steward on the governance team, you will create data protection rules in order to protect confidential mortgage data.

In this part of the lab, you will complete these tasks:

1. Create a data protection rule to deny access.
2. Create a data protection rule to mask data.

### **Prerequisites**

Complete the Trust your data lab to import and enrich data assets and publish them to a catalog.

## Step 1: Create a data protection rule to deny access

A data protection rule controls access to a data asset. You can either mask data in the data asset or deny access to the data asset. Follow these steps to create a data protection rule to deny access to confidential information in some of the mortgage data assets.

1. From the Cloud Pak for Data navigation menu, choose **Catalogs > View all catalogs**.
2. Open the **Mortgage Approval Catalog**.
3. Click the **CREDIT\_SCORE** data asset. Notice that it contains the confidential tag. You will create a rule to deny access to this data asset.
4. From the Cloud Pak for Data navigation menu, choose **Governance > Rules**.
5. Click **Add rule > New rule**.
6. Select **Data protection rule**.
7. Click **Next**.
8. For the Name, type *Confidential Information*.
9. For the Business definition, type Rule to prevent unauthorized users from accessing data assets that have been tagged as confidential.
10. For Condition 1, select the following options.
  - Select **Tag**.
  - Select **contains any**.
  - Type *confidential*.
11. For the Action, **select deny access to data**.
12. Click **Create**. This rule will now deny access to data for anyone trying to access data assets that tagged as “Confidential”. This rule applies in the Catalog Preview, Catalog Download, Data Refinery, and Project Asset preview. Note that the rule doesn’t apply to the person who created the rule or added an asset to a catalog
13. Now if you log in with another account, you will what others see when trying to access the CREDIT\_SCORE data asset.

## Step 2: Create a data protection rule to mask data

Some of the mortgage data assets include personal identifiable information which you need to protect, but the rest of the columns contains valuable information that should be available to a broader audience. That's where data masking comes in handy. Follow these steps to create a data protection rule that will mask data assets containing columns with a US Social Security Number.

1. From the Cloud Pak for Data navigation menu, choose **Catalogs > View all catalogs**.
2. Click **Mortgage Approval Catalog**.
3. In the catalog, click the **MORTGAGE\_APPLICANTS\_TRUST** data asset.
4. Click the **Asset** tab to preview the data. Notice that one of the columns contains Social Security Numbers.
5. Click the eye icon for the Social Security Number column. Notice that this column was auto-assigned the Social Security Number business term. You will create a rule to mask this column.
6. From the Cloud Pak for Data navigation menu, choose **Governance > Rules**.
7. Click **Add rule > New rule**.
8. Select **Data projection rule**.
9. Click **Next**.
10. For the Name, type *Redact Social Security Number*.
11. For the Business definition, type *Rule to redact Social Security Number*.
12. For Condition 1, select the following options:
  - Select **Business term**.
  - Select **contains any**.
  - Start typing social, and then select **Social Security Number**.
13. For the Action, select **mask data**. Business term and *Social Security Number* will be filled in for you.
14. For the masking options, select **Redact**. This will replace the data with Xs. You can hover over each masking option to see example data masked using the selected option.
15. Click **Create**. This rule redacts columns with US Social Security Numbers in data assets.
16. Log in with other account to see what other users will see accessing the MORTGAGE\_APPLICANTS data asset.

As a Data Steward on the governance team, you have learned how to create data protection rules in order to protect confidential mortgage data.

### Next steps

You are now ready to know your data by evaluating, sharing, shaping, and analyzing data in the data fabric. Check the Know your data lab.



## Part 3: Know Your Data

Take this part to work with your trusted and protected data with the Data Governance and Privacy use case of the data fabric trial. Your goal is to evaluate, share, shape, and analyze data in the data fabric.

The story for the tutorial is that Golden Bank has several departments that need access to high-quality customer mortgage data. As a Data Analyst, you will need to search for and find the right data, understand and trust its content, and then prepare it for other data analysts and data scientists to use.

In this part, you will complete these tasks:

1. Understand data assets.
2. Enrich assets and create relationships.
3. Add enriched data to a project.
4. Visualize the data.
5. Prepare the data for analytics and data science project.

### **Prerequisites**

Complete the Trust your data and Protect your data parts as above:

- Trust your data part to import and enrich data assets and publish them to a catalog.
- Protect your data part to create data protection rules and masking flows to protect data.

## Step 1: Understand data assets

Data assets in catalogs are much more than pointers to data. They contain information about the format and meaning of the data and statistics about the data values. Follow these steps to understand the value of data assets.

1. From the Cloud Pak for Data navigation menu, choose **Catalogs > View all catalogs**.
2. Open the **Mortgage Approval Catalog**.
3. The featured assets section shows **Recently added assets**, assets that **Watson recommends** which are suggested assets from AI and machine learning based on your past usage and popularity, and **Highly rated** assets that catalog collaborators have rated and reviewed.
4. Click **Hide featured assets** to close that section.
5. Search for *mortgage*.
6. Click **MORTGAGE\_APPLICANTS\_TRUST** to view that catalog asset. The Overview tab provides basic information about the asset such as the description, a rating, tags, where the asset is located, business terms, data classes, and related assets.
7. Click the **Profile** tab. The profile information helps you understand the content, the quality, and usability of the data.
8. Scroll to the right to locate the **ZIP\_CODE** column.
9. The data class that was automatically assigned to the *ZIP\_CODE* column is *Commercial and Government Entity*, but the values are actually zip codes. You can easily reclassify this column. Click the drop down list to see other possible data classes and their confidence levels. Select **US Zip Code**.
10. Click the **Asset** tab to see a preview of the data.
11. To view column metadata, click the eye icon for the *EMPLOYMENT\_STATUS* column to see the assigned business terms. Click **Close** to close the column metadata window.

You explored the type of information that Watson Knowledge Catalog automatically adds to data assets during metadata enrichment. In the next step, you'll see how you can enrich data assets.

## Step 2: Enrich assets and create relationships

You can make assets more valuable by adding information to them. For example, you can add your opinion of the asset, update asset properties, and create relationships to link assets. Follow these steps to enrich assets and create relationships.

1. For the **MORTGAGE\_APPLICANTS\_TRUST** catalog asset, click the **Review** tab. Rate and comment on this asset so that others can find the asset easily.
  0. Select **5 stars** for the rating.
  1. For the review, type *This contains high quality customer data from the mortgage system.*
  2. Click **Submit**.
2. Click the **Overview** tab.
3. To edit the asset name, click the **pencil** icon next to the asset name.
  0. Change the name to **MORTGAGE\_APPLICANTS\_TRUST\_PROTECT**.
  1. Click **Apply**.
4. In the Description section, click the **+** icon.
  0. Type Mortgage applicants from the Mortgage System.
  1. Click **Add**.
5. Because this asset relates to mortgage loans, next to Business terms, click **+** (plus sign).
  0. Search for loan.
  1. Select **Loan**.
  2. Click **Add**.
6. Because this asset contains personal information, next to Classifications, click **+** (plus sign).
  0. Select **Personally Identifiable Information**.
  1. Click **Add**.
7. Because this asset is related to other mortgage assets, next to Related assets, click **Add asset**.
  0. Select **Is related to**, and click **Next**.
  1. Select the **CREDIT\_SCORE** and **MORTGAGE\_APPLICATION** assets, and click **Add**.
8. Click **MORTGAGE\_APPLICATION** to view that related asset.

You've made these assets more valuable by reviewing, updating properties, and adding relationships to the assets.

### Step 3: Add enriched data to a project

The data analysts team needs the mortgage applicants data in the mortgage analysis project to refine, visualize, analyze, and use as training data for models. Follow these steps to add the enriched data to a project.

1. Click **Mortgage Approval Catalog** in the navigation trail.
2. At the end of the *MORTGAGE\_APPLICANTS\_TRUST\_PROTECT* catalog asset row, click the **Three dots** menu, and choose **Add to project**.
  0. In the Target drop down list, select the project.
    1. Click **Add**.
3. When the notification displays, click **Go to project**. If you miss the notification, then:
  0. Click the Cloud Pak for Data navigation menu, choose Projects > View all projects.
    1. Click the project.
4. In the project, click the **Assets** tab to see the *MORTGAGE\_APPLICANTS\_TRUST\_PROTECT* data asset.

#### Step 4: Visualize the data

You need to cleanse and refine the mortgage applicants data to get it ready for your analytical tools and models. A quick and easy way to determine how it needs to be shaped is to visualize the data in Data Refinery. Note that the visualization is based on the first 5,000 rows of the data. Follow these steps to visualize the data.

1. Click the **MORTGAGE\_APPLICANTS\_TRUST\_PROTECT** data asset to preview the data. (After the Click it is Prepare data; alternatively Refine option is available from the **Three dots** menu of **MORTGAGE\_APPLICANTS\_TRUST\_PROTECT**)
2. Click **Refine** to open the data asset in Data Refinery, and wait for the data to be read and processed.
3. In the Information panel, click the **X** to close the panel.
4. In the Steps panel, click the **X** to close the panel.
5. Click the **Visualizations** tab.
6. For the Column to visualize, select **EMPLOYMENT\_STATUS**.
7. Click **Visualize data**. The tool selects a pie chart as the best chart type for this column which shows the distribution of applicants by employment status. Notice that there are several suggested chart types indicated by a blue dot next to bar, word cloud, and sunburst.
8. For the *Chart type*, select the **Bubble** chart type. The Bubble chart is one easy way to quickly visualize the distribution of values in a particular dataset.
9. From the Chart type drop-down, select the **Relationship** chart type.
10. This chart type requires two columns. Select these columns:
  0. For the Column field, select **EMPLOYMENT\_STATUS**.
  1. Click **Add another column**.
  2. For the second Column, select **EDUCATION**.
11. With the *Relationship* chart, you can select endpoints to see the relationships. For example, you can see applicants employment status by level of education.

You are now ready to cleanse the data.

## Step 5: Prepare the data for analytics and Data Science

You can't process applicants without a social security number, so you need to review the data and remove any applicants without social security numbers.

To prepare the MORTGAGE\_APPLICANTS\_TRUST\_PROTECT data, you will:

- View the frequency of values in the Social\_Security\_Number column.
- Filter the applicants with missing values from the Social\_Security\_Number column.

Follow these steps to prepare the data.

1. In the Data Refinery, click the **Profile** tab.
2. Scroll to the right to locate the Social\_Security\_Number column. Notice that there are several missing values.
3. Click the **Data** tab to filter out these records. Notice at the bottom of the screen, Data Refinery indicates that the FULL DATA SET is 1101 rows.
4. If the Steps panel is not visible, click **Steps** to open the panel.
5. Click **New step**.
  0. In the Cleanse section, select **Filter**.
  1. In the Column field, select the **Social\_Security\_Number** column.
  2. In the Operator field, select **Is not empty**.
  3. Click Apply. Notice at the bottom of the screen, Data Refinery now indicates that the FULL DATA SET is 1000 rows because the rows with missing Social Security Numbers have been filtered out. Notice that a new step displays in the Steps panel showing the Filter operation.
6. Click the **Profile** tab.
7. Scroll to the right to locate the Social\_Security\_Number column. Notice that the missing values are gone.
8. From the toolbar, click the **Save** icon.
9. From the toolbar, click the Export icon, and choose **Export current data as CSV**.
  0. Save the **MORTGAGE\_APPLICANTS\_TRUST\_PROTECT\_shaped.csv** to a local folder.
  1. Navigate to that folder, and open the CSV file which contains 1000 rows and no applicants are missing the social security number.
10. Return to Cloud Pak for Data, and click the project in the navigation trail.
11. Click All assets, and locate the new Data Refinery flow asset with the name MORTGAGE\_APPLICANTS\_TRUST\_PROTECT\_flow.

As a Data Analyst for Golden Bank, you have learned how to search for and find the right data, understand and trust its content, and then prepare it for other data analysts and data scientists to use.