Cloud Pak for Data

Version 2.5 or higher

Tutorial – Mortgage

Contents

1. Prerequisites 5
2. Setting up database and sample data 5
3. Access Credentials 6
3.1. Access credential for Db2 database 6
3.2. Sign into Cloud Pak for Data web console as Administrator 6
4. Create Connection 7
4.2. Navigate to Connections 7
4.3. Add connection 7
5. Discover Assets 8
5.1. Navigate to discover assets 8
6. Add users 11
6.1. Grant Catalog Permission 12
7. Implement Business Glossary 14
7.1. Download Business Glossaries 14
7.2. Import Categories 15
7.2. Import Terms 16
7.3. Create a policy 17
7.4. Create a rule 17
7.5. Automated Discovery 18
7.6. Add rule to metadata 19
8. Access data as a Data Scientist 21
8.1. Create analytic project 21
8.2. Assets from Glossary 22
8.3. Check Asset Details 22
9. Data Virtualization 24
9.1. Adding a new data source for Db2 24
9.4. Add virtual table to catalog 27
9.5. Publish virtualized table 27
9.6. Access information for virtual table 28
10. Build Model 29
10.1. Navigate to analytics project 29
10.2. Create deployment space 29

10.3. Create notebook 30
10.4. Review and run notebook 30
10.5. Test the model 32

Cloud Pak for Data is a single end to end platform for data management, governance and data science analytics. It provides a one stop shop for data scientists, data engineer and data stewards to collaborate on the platform to acquire, govern and extract best insights from the data in the least amount of time.

In this demo, user will use a set of a fictious mortgage data that available in Db2 database on a docker image. User will perform following tasks to predict if a prospective customer may default on their mortgage.

- Create connection from Cloud Pak for Data to Db2 database on cloud
- Discover Db2 assets from Cloud Pak for Data
- Transform the Db2 data using Data Virtualization
- Build a simple machine learning model for prediction

1. Prerequisites

- Access to an operational Cloud Pak for Data (v.2.5 or higher) Instance
- Install Git on the machine that you will use for the tutorial

2. Setting up database and sample data

Log in to the cluster where Cloud Pak for Data is deployed or log in to a Linux-based system (RedHat or Ubuntu) that can access the cluster over your network.

From your home directory, clone the tutorial sample files:

```
git clone https://github.com/IBM-ICP4D/icp4d-tutorials.git
```

Change to the tutorials directory:

```
cd icp4d-tutorials/tutorials/
```

The sample data-loading utility, load_samples.sh, provides an easy way to host a Db2 server and load it with sample data.

Run the following command to view the list of sample data that is provided in the load_samples.sh utility:

```
./load_samples.sh -l
```

Run the following command to load the sample data into a Db2 database:

```
./load_samples.sh -t mortgage-002
```

After the loading process completes, an instance of Db2 is hosted on your cluster as a Docker container.

3. Access Credentials

To work through the tutorial, you need access a Db2 database.

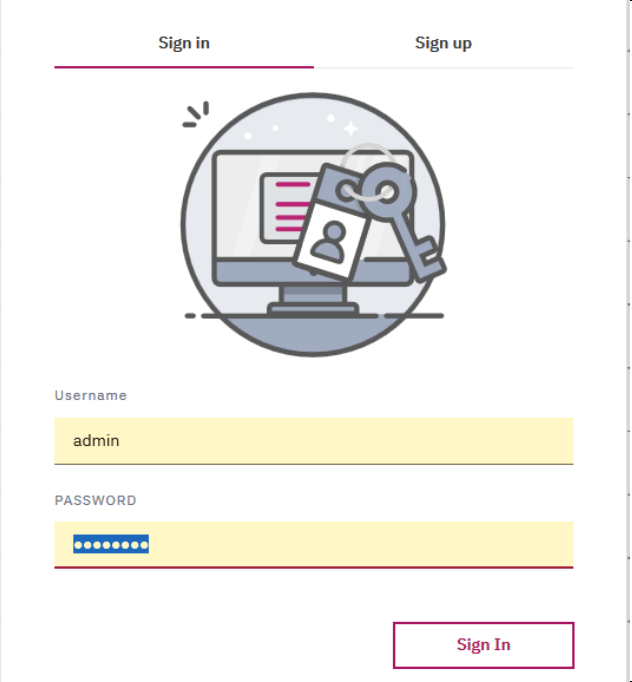
3.1. Access credential for Db2 database

For this tutorial you need JDBC connection to access to a Db2 database that hosted locally on Cloud Pak for Data. Following are JDBC connection credential for Db2:

JDBC Host name	<Same IP address as your web console>
Port number	50000
Database name	MORTGAGE
User ID	db2inst1
Password	password
Db2	Version 11.1
JDBC connection string	jdbc:db2://<same IP as Web Console>:50000/MORTGAGE

3.2. Sign into Cloud Pak for Data web console as Administrator

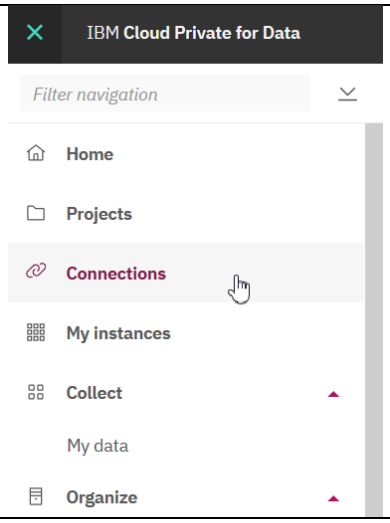
You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data web console. Starting from here all instruction need to execute on Cloud Pak for Data web console only. You need to login as admin who has administrator privileges.

	Sigh into the Cloud Pak for Data web console as user ‘admin’ and password is ‘password’.

4. Create Connection

Create a connection to the data source for Db2 database.

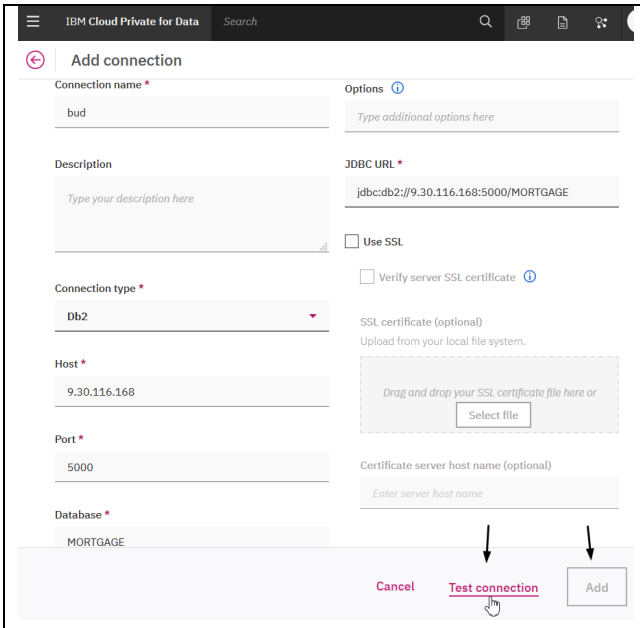
4.2. Navigate to Connections



On the left pane choose **Connections**.

Next, on the **Data Connections** window click on the Add connection icon.

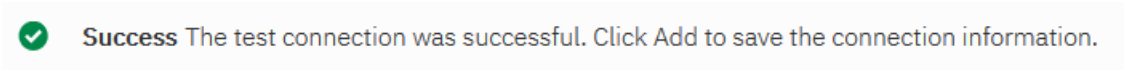
4.3. Add connection



Fill out the **Add Connection** information according to the information provided in step 3.1. Access credential for DB2. Credential used in following step is just an example.

1. For **Choose connection** use the drop-down menu and select 'Db2'.
2. Use 'Bud' as the **Connection name**
3. Use IP of the cluster node (where DB2 database) as **Host**
4. **Port** is '50000'
5. **Database is** `MORTGAGE`
6. **Username** is 'db2inst1' and **Password** is 'password'.

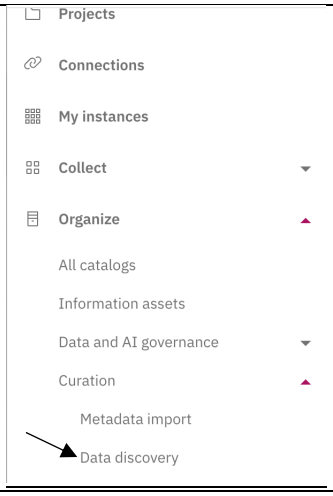
Next click on **Test Connection**, once it successful click on **Save Connection**.



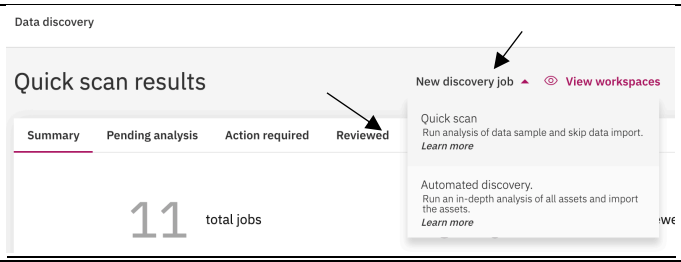
5. Discover Assets

Use the data source created above discover all data assets from Db2 database.

5.1. Navigate to discover assets

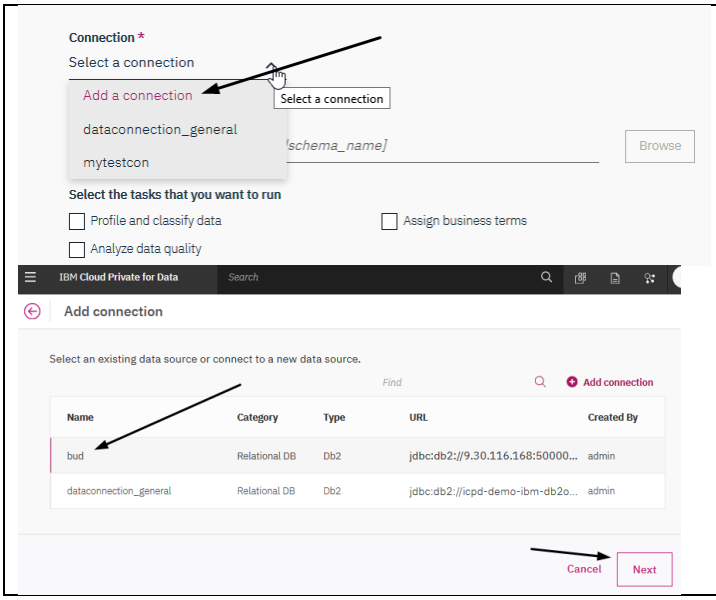


From **Organize** option on the left pane, choose **Curation** > **Data discovery**.



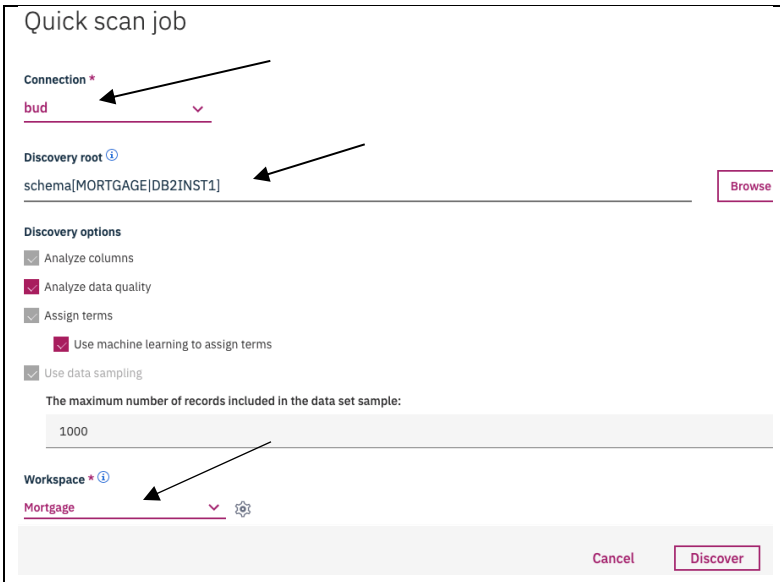
To select discover job

Navigate to **New discover job** > **Quick scan**



To discover assets

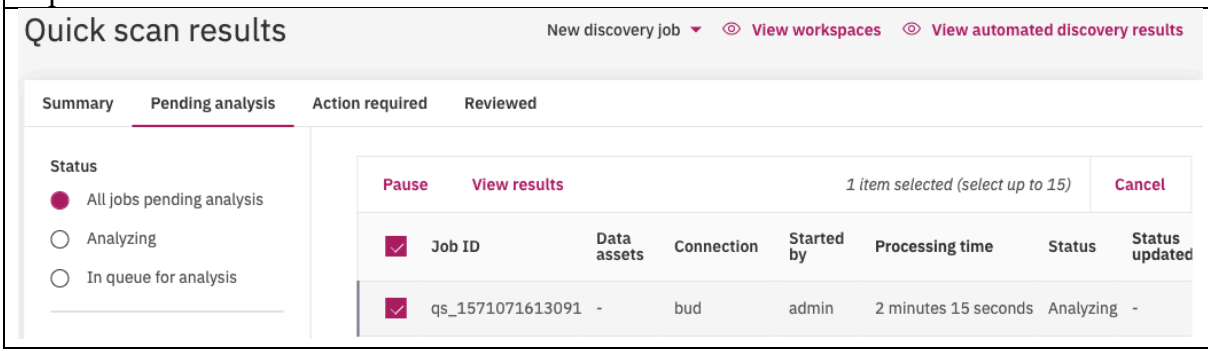
1. Click on **Add a connection**
2. Choose the connection named **bud** that you created previously, click **Next**

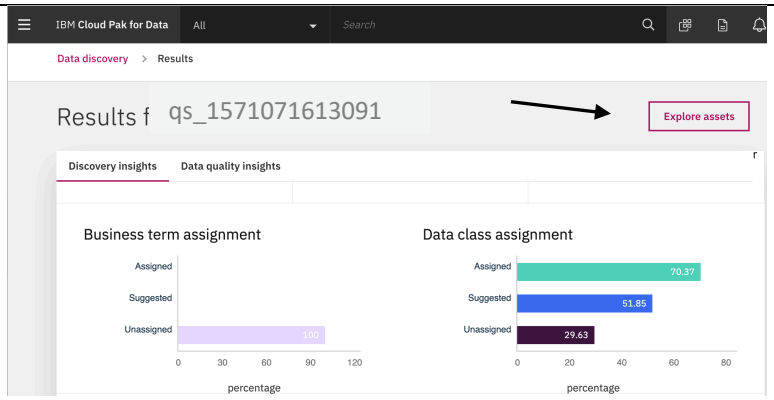


3. Choose the connection named **bud** that you created previously.
4. Select **Discover root** as **MORTGAGE** > **DB2INST1**
5. Check necessary **Discover options**
6. Click on **Add a workspace** under Workspace and named it as **Mortgage**. Click **Create**.
7. Click on **Discover**

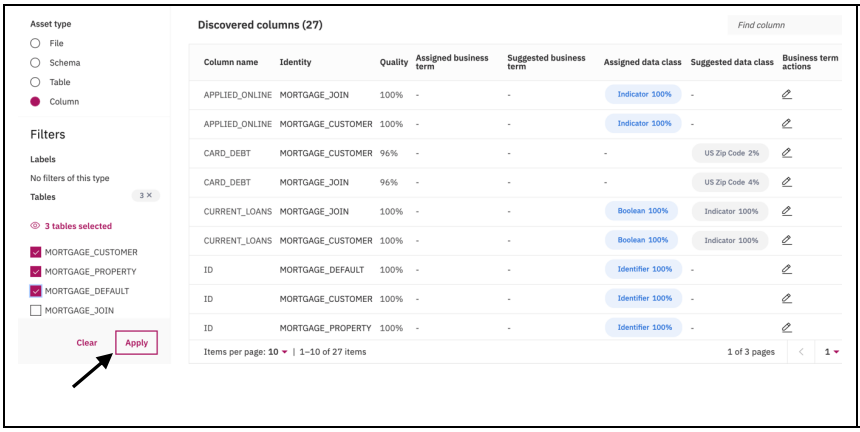
It may take few minutes to complete.

Click on **Quick scan results** > **Action required** > **View results** or **View workspaces** to explore the discover assets.





Review the discovery results using **Explore assets** tab

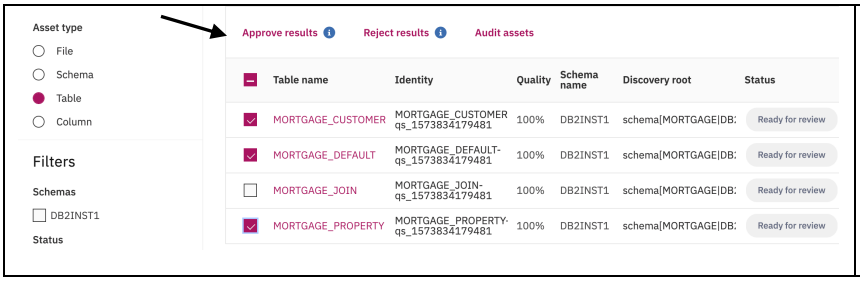


Review assets for proper business data class assignment, if needed you can adjust them.

Select **Asset type** as "**Column**"

Filters necessary tables using checkbox

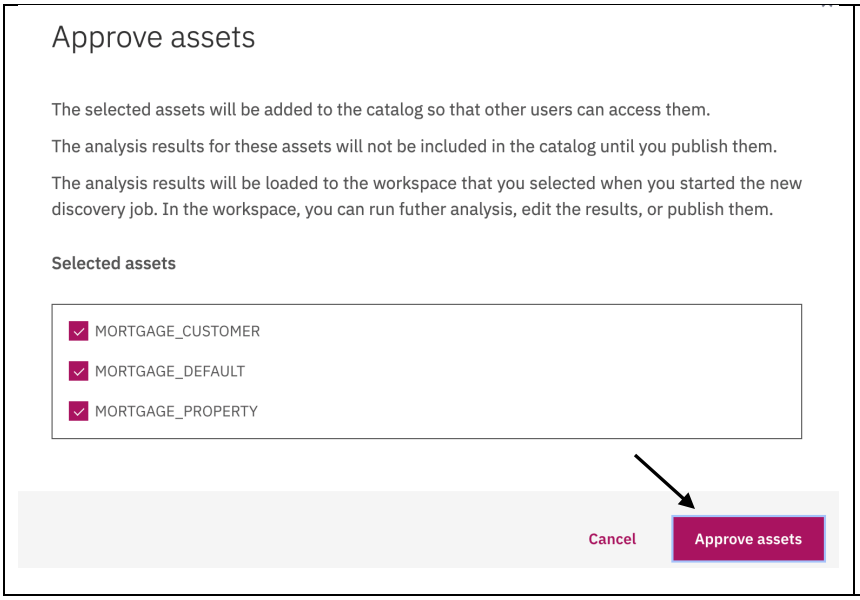
Click on **Apply**



Change **Asset type** as "**Table**"

Select all Mortgage related tables

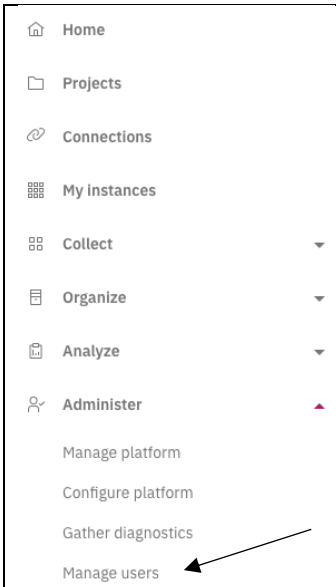
Click on **Approve results**



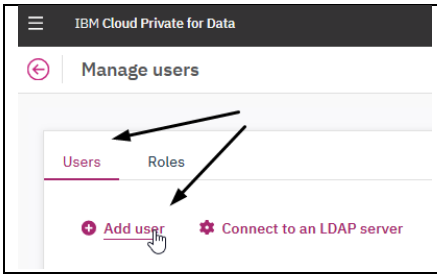
Click on **Approve assets**

6. Add users

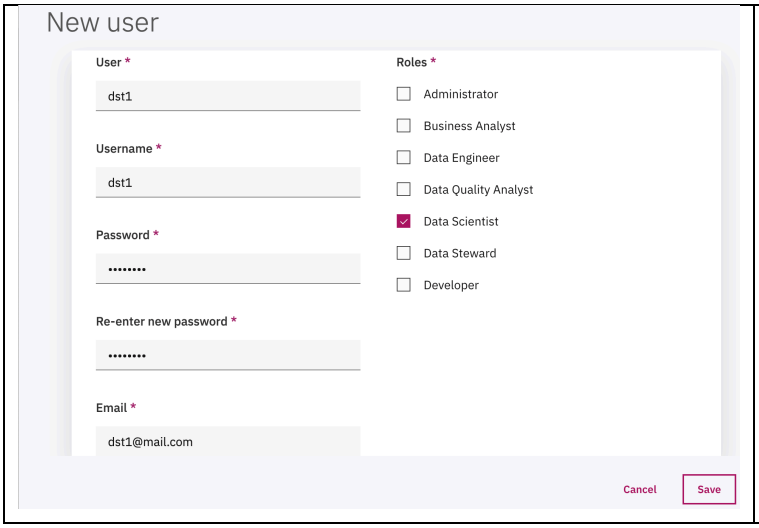
Create users with different roles.



From **Administer** option on the left pane, choose **Manage users**.



Switch tab to 'Users' and click on 'Add user'



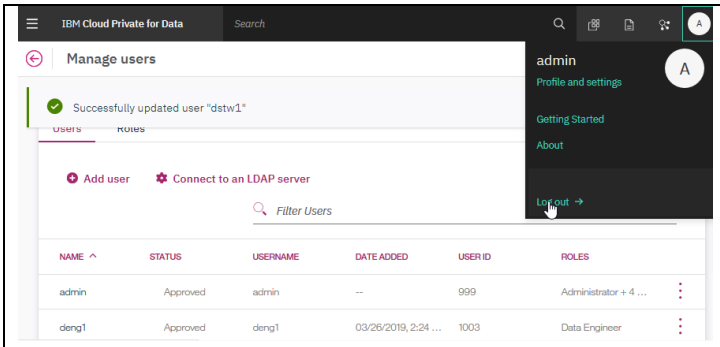
Fill out Add User information for a data scientist

1. User as **dst1**
2. Username is **dst1**
3. Use a valid email address
4. Set Password as **dst1**
5. Chose the user roles as Data Scientist

Click on **Save** to confirm the add user

Follow same steps in Add User section (above) and two more account. Create **deng1** for Data Engineer and **dstw1** a data steward.

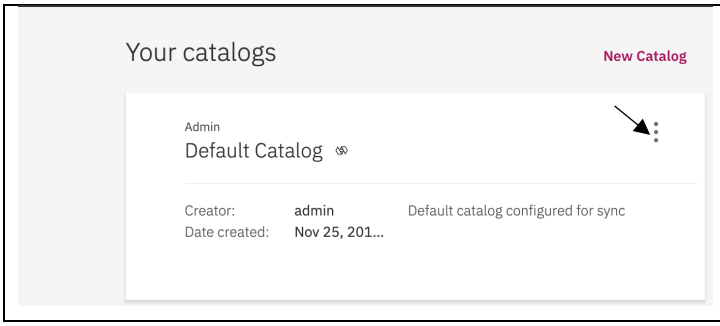
	User	Role	Password
•	deng1	Data Engineer	deng1
•	dstw1	Data Stewards	dstw1



Log out from user **admin**

6.1. Grant Catalog Permission

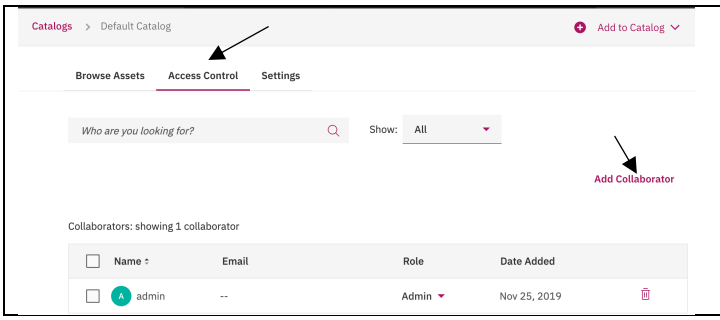
With Watson Knowledge Catalog, you use catalogs to easily find and share your data and other assets. A catalog is like a private community for your organization. It's a way to organize resources for many data science projects: data assets, analytical assets, and the users who need to use the assets. You can manage access to the catalog by adding collaborators with specific roles that determine their permissions to perform actions.



Go to **Organize** > **All catalogs**

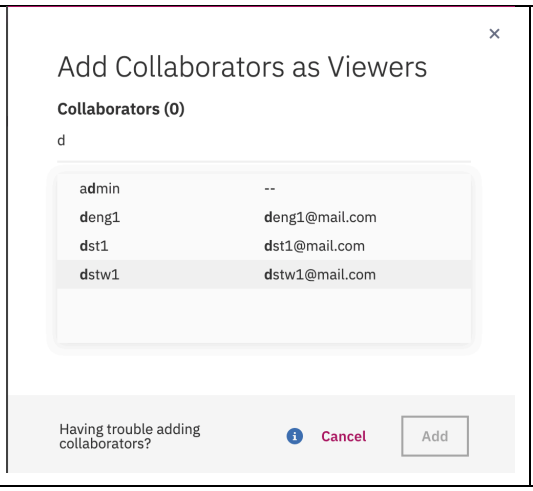
Select **Default Catalog** and click on action icon

Choose **View**



Go to **Access Control** tab

Click on **Add Collaborator**



Add ‘deng1’ user as collaborator with editor role

Add ‘dstw1’ user as collaborator with editor role

Add ‘dst1’ user as collaborator with viewer role

7. Implement Business Glossary

Cloud Pak for Data enables you to structure your enterprise information in a logical way, discover relationships between assets, and keep your data always up-to-date. You can import existing glossary with categories, terms, information governance policies and rules.

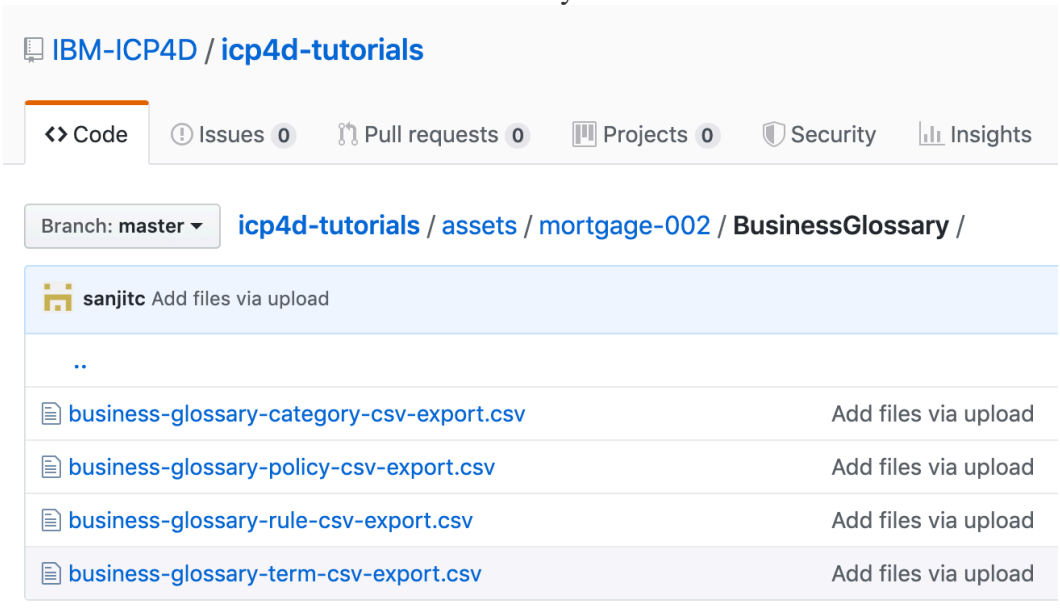
	Sigh into the Cloud Pak for Data web console as user 'dstw1' and password is 'dstw1' that you created earlier.

7.1. Download Business Glossaries

First download business glossaries from the GIT to your local machine.

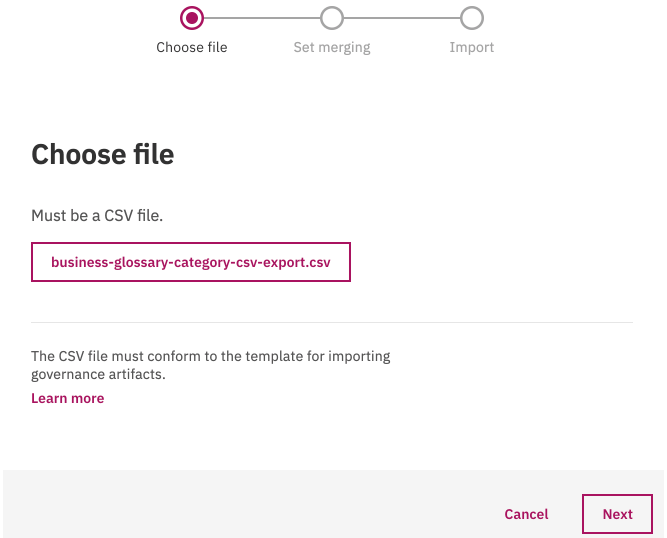
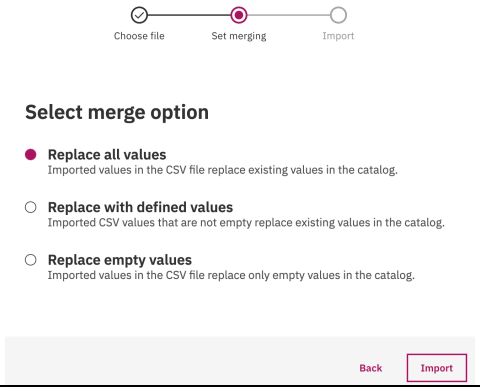
Go to: https://github.com/IBM-ICP4D/icp4d-tutorials/tree/master/assets/mortgage-002/BusinessGlossary

Download all four CSV files and save them locally.



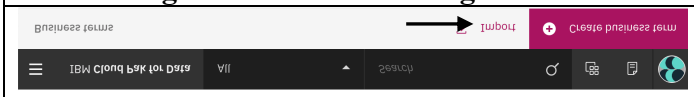
7.2. Import Categories

Sequence is important when importing business glossaries. Make sure import categories before do the terms.

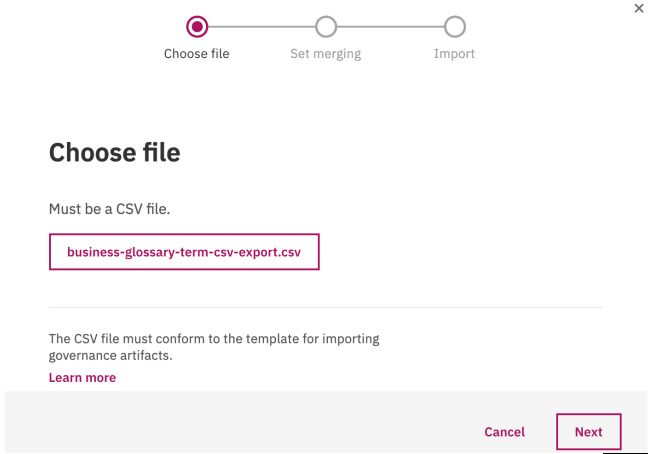
Choose **Organize** > **Data and AI governance** > **Categories** from the left pane.	
	Click on **Import** to import the CSV file contains category information that you downloaded from Git.
	Choose the CSV file location Click **Next**
	Select merge option as **Replace all values** Click **Import**

7.2. Import Terms

Choose **Organize** > **Data and AI governance** > **Business terms** from the left pane.

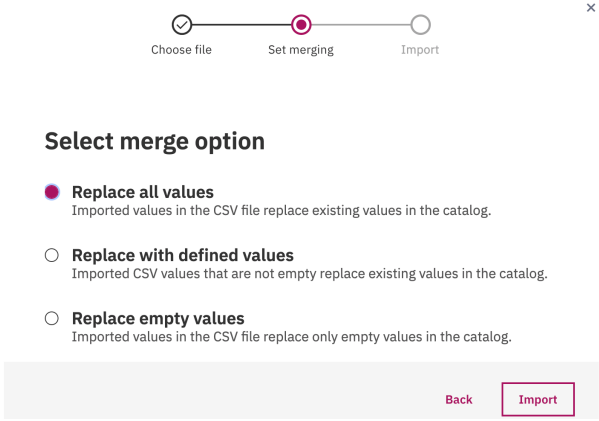


Click on **Import** to import the CSV file contains term information that you downloaded from Git.



Choose the CSV file location

Click **Next**



Select merge option as **Replace all values**

Click **Import**

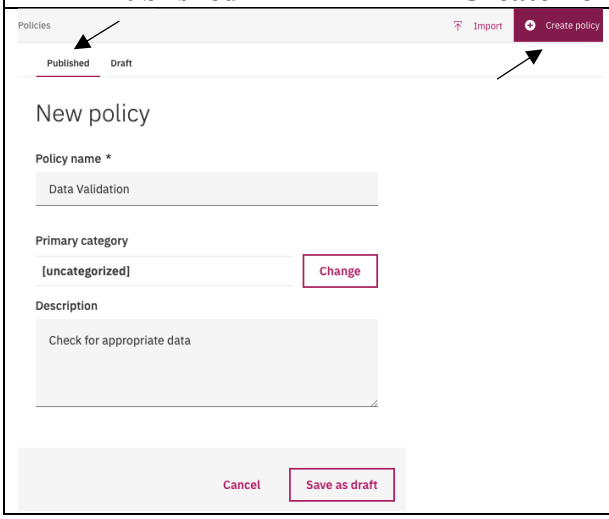
Review each imported business terms and then publish

7.3. Create a policy

Create governance policies and rules for the entire organization to ensure clarity and compatibility among departments, projects, or products.

Choose **Organize** > **Data and AI governance** > **Policy** from the left pane

Select **Published** tab and click on **Create Policy**



On the **New policy** window create a policy with following information and click on **Save as draft**:

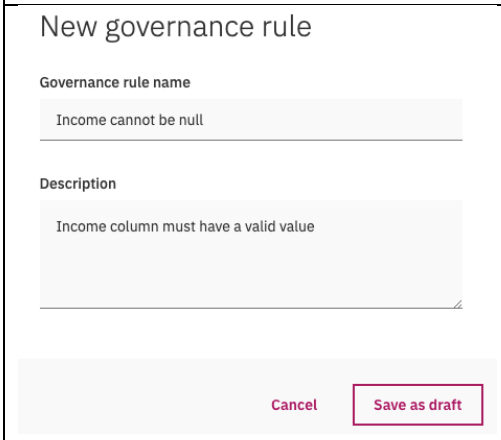
Name: Data Validation
Description: Check for appropriate data

It will take few minutes to appear under list of available policies.

Once new policy available let's publish it.

7.4. Create a rule

Choose **Organize** > **Data and AI governance** > **Rule** from the left pane
Select **Published** tab and click on **Create Rule**
Choose **Governance rule**



On the **New governance rule** window create a rule with following information and click on **Save as draft**:

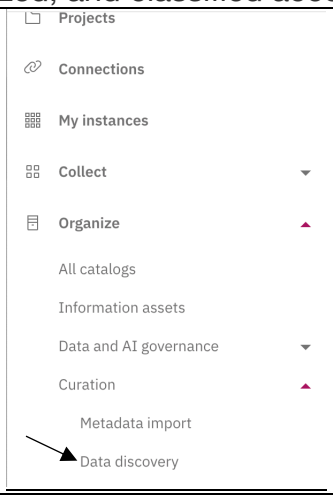
Name: Income cannot be null
Description: Income column must have a valid value

It will take few minutes to appear under list of available rules. Once the new rule is available, publish it.

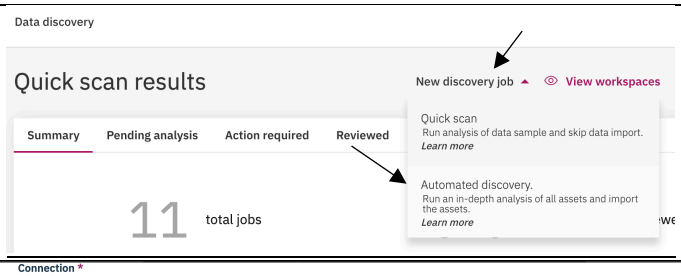
Click on **Add policy** under **Parent policies** to assign the **Data Validation** rule to it.

7.5. Automated Discovery

Re-run discover assets to add data to the catalog. During the discovery the data is imported, analyzed, and classified according the glossary you imported/created earlier.

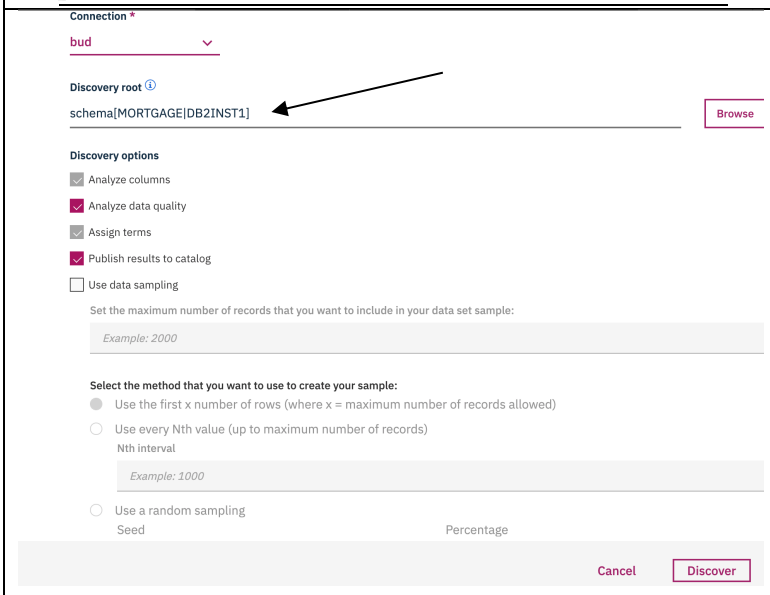


From **Organize** option on the left pane, choose **Curation** > **Data discovery**.



To automated discover job

Navigate to **New discover job** > **Automated discovery**



To discover assets

Choose the connection named **bud** that you created previously.

Select **Discover root** as **MORTGAGE** > **DB2INST1**

Check necessary **Discover options**

Select Workspace as **Mortgage**.

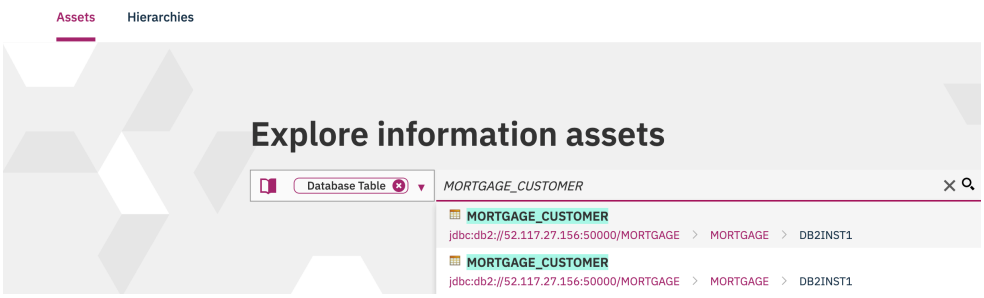
Click on **Discover**

Wait till import and analyze phase complete.

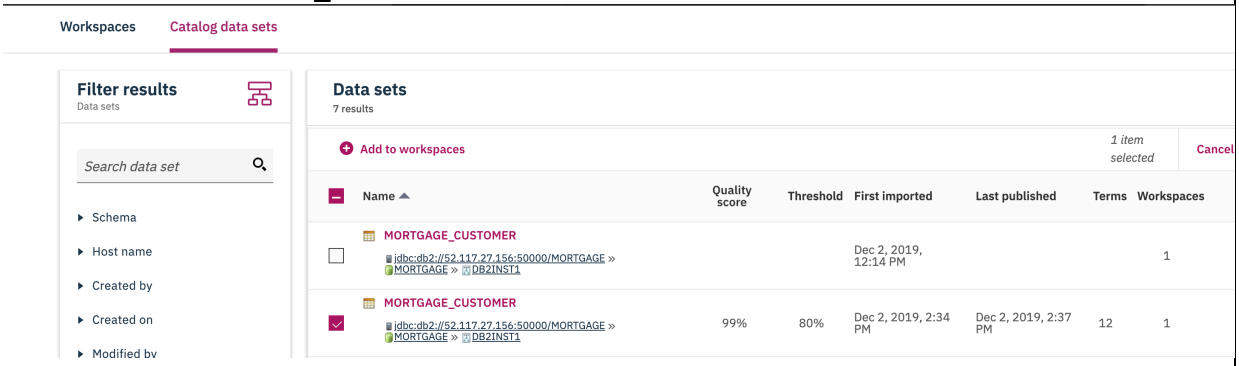
7.6. Add rule to metadata

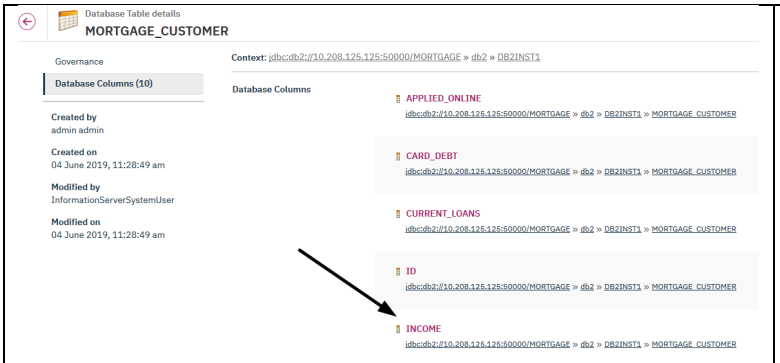
Go to **Organize** > **Information assets**

Search for Database Table name MORTGAGE_CUSTOMER



Click on **MORTGAGE_CUSTOMER** data set





On Database Table Details window choose **Database Columns** from left

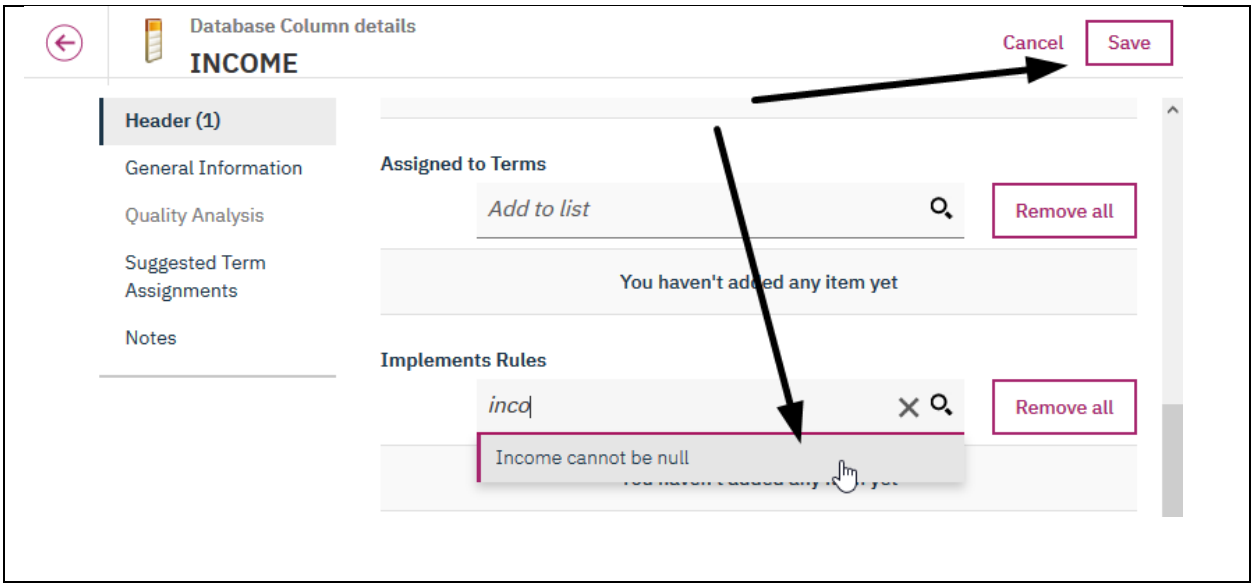
Select INCOME column

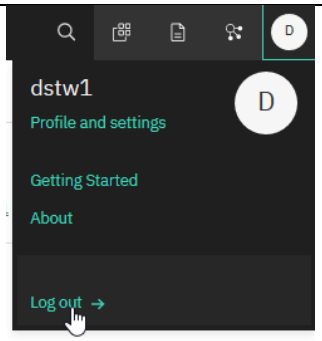
Next click on ⋮ icon (right top corner) and choose **Edit**

Scroll down to **Implement Rules** section

Search and select the rule **Income cannot be null** that you created earlier.

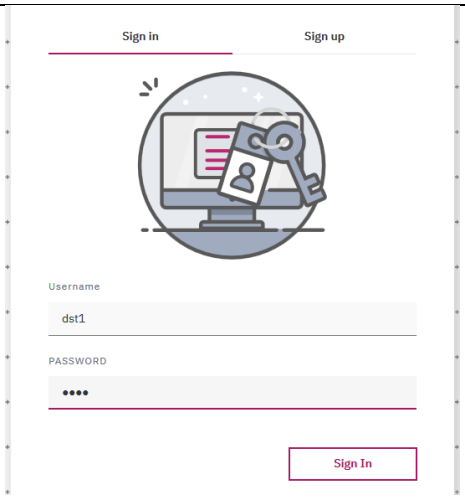
Click on **Save**



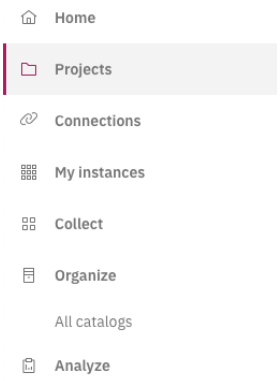
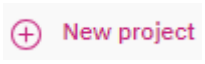
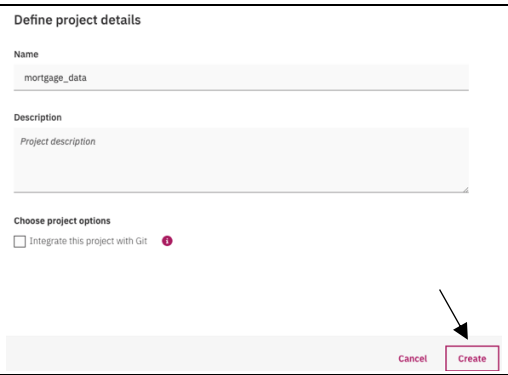
	Log out from user ‘dstw1’

8. Access data as a Data Scientist

Explore the data require for build a model

	Sigh in to the Cloud Pak for Data web console as user ‘dst1’ and password is ‘dst1’ that you created earlier.

8.1. Create analytic project

	Create a new analytical project by ‘Projects’ from right pane. Click on the  icon Select **Create an empty project**
	Provide project name ‘mortgage_data’ and click **Create**

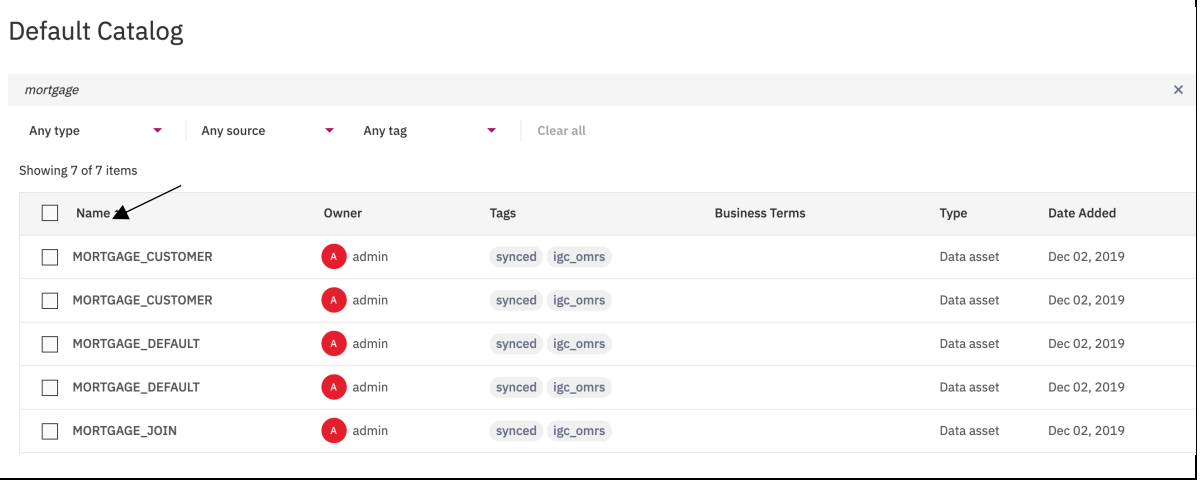
8.2. Assets from Glossary

Let's look for mortgage related terms in glossary to get an idea about different data assets available on the system.

Go to **Organize** > **All catalogs** and choose **Default Catalog**

Search for word **Mortgage** from **Browse Assets** to find all mortgage related assets.

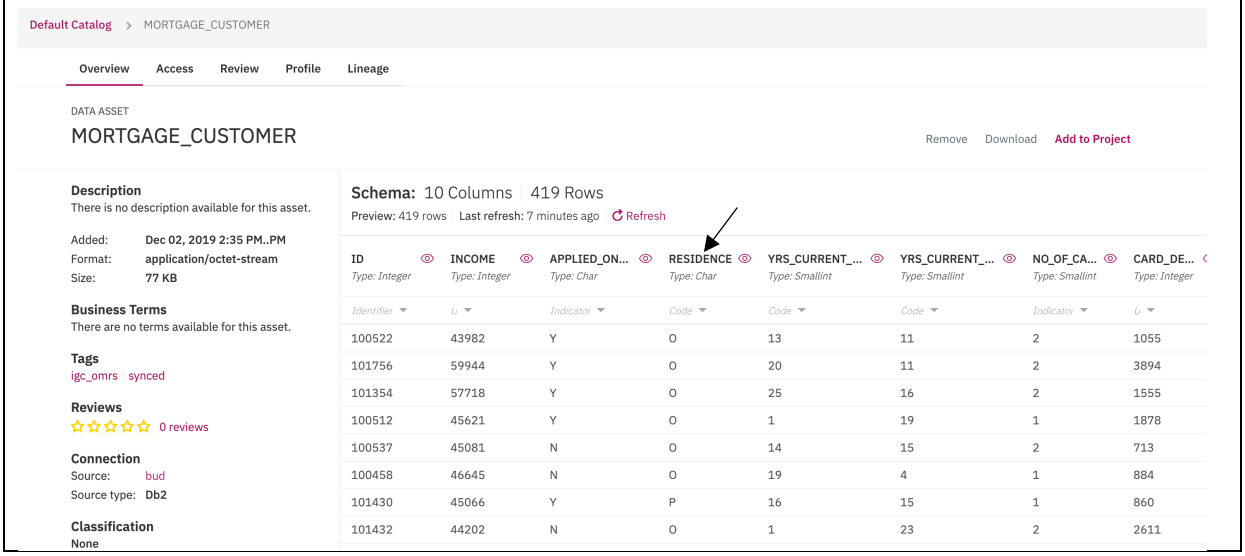
Click on each assets for additional information.



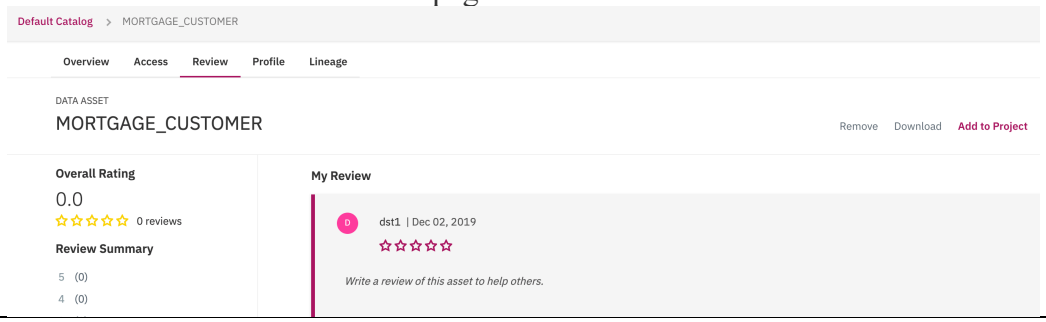
8.3. Check Asset Details

Go through each data assets related to mortgage in glossary to have better idea about data you need for your project. For example, check the MORTGAGE_CUSTOMER.

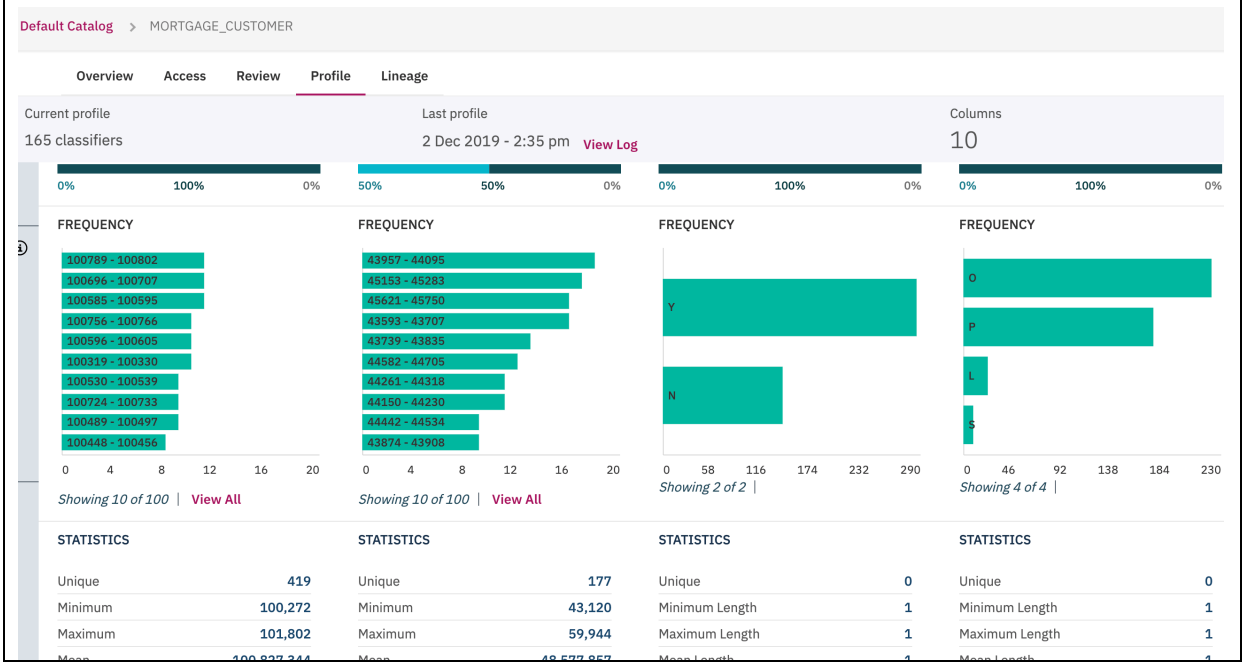
The asset **Overview** tab shows the asset properties, such as the description, tags, format, size, and date added. You'll see a **preview** of the contents of the asset if the asset type supports previews and you have the proper permissions. Check individual column header description.

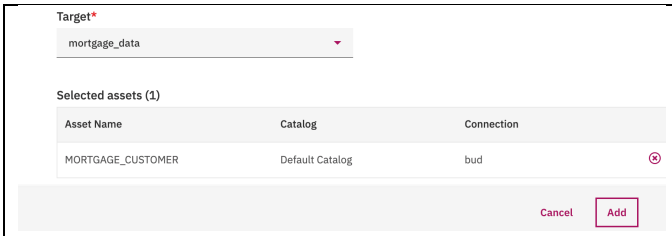


The **Review tab** shows the ratings and reviews of the asset by catalog collaborators. You can rate the asset and write a review on this page.



The **Profile tab** shows **profile information** about the contents of the asset. The profile of a data asset includes generated metadata and statistics about the textual content of the data. It contains relational or structured data shows information about each column in the data set, based on the first 5000 rows of data. The profile shows the frequency of the inferred data classes and statistics about the data for each column.





Once find right data asset use **Add to Project** tab to include it in your project. Select **Target** project as 'mortgage_data' and click on **Add**.

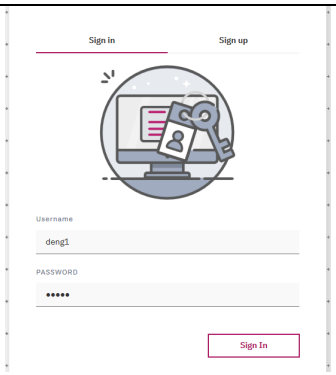
Similar way add MORTGAGE_DEFAULT and MORTGAGE_PROPERTY assets to your project.

9. Data Virtualization

Many time as a data engineer, you can receive requests for data from others. If you decide that a request requires data to be virtualized, You can use Data Virtualization (DV).

Assume you are a data engineer and need to deliver a data request that combined data sets of MORTGAGE_CUSTOMER, MORTGAGE_PROPERTY and MORTGAGE_DEFAULT.

DV allows integrate data sources across multiple types and locations and turns it into one logical data view. In this case, you have data across three different tables. Creating a virtual table you can quickly view data from different tables.

	Sigh into the Cloud Pak for Data web console as user 'deng1' and password is 'deng1' that you created earlier.

9.1. Adding a new data source for Db2

DV supports many relational and non-relational data sources (as well as files that reside on a local disk or network file system) that you can add to your data source ecosystem. After a data source has been added, any user that has virtualize permission can create virtual tables. DV agents connect to relational data sources using JDBC protocol. In this tutorial you will add a data source for Db2 database.

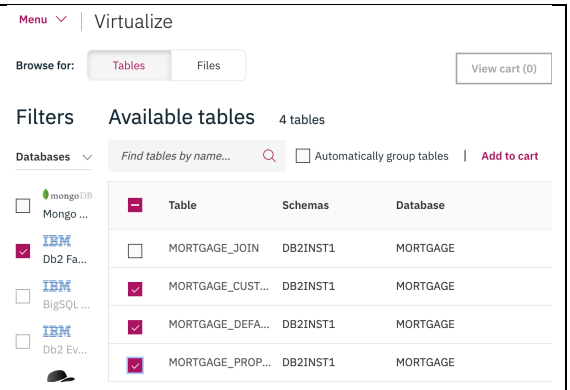
Define a data connection to Db2. Use your existing Db2 database connection for Db2 data source.

1. Go to **Collect** > **Data Virtualization** > **Menu** > **Data sources**
2. Click **Add** > **Add data source** > **Add connection**
3. Select **bud** that you created earlier and click **Next**

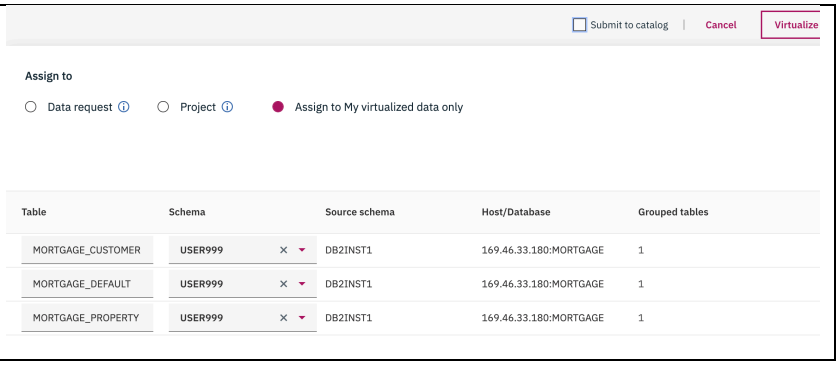
9.2. Select tables for virtualization

The most common mechanism for virtualizing data is to create a "view" or virtual table. Virtual tables can be full or segment of data from one or more tables. You can then run queries against the resulting virtual table.

- Click **Collect** > **Data virtualization** > **Menu** > **Virtualize**
- Select tables **MORTGAGE_CUSTOMER, MORTGAGE_PROPERTY and MORTGAGE_DEFAULT** from **MORTGAGE** database, then click **Add to cart**
- Click **View cart**
- Click **Next**



- Uncheck the box for **Submit to catalog**
- Click **Virtualize** to complete the process

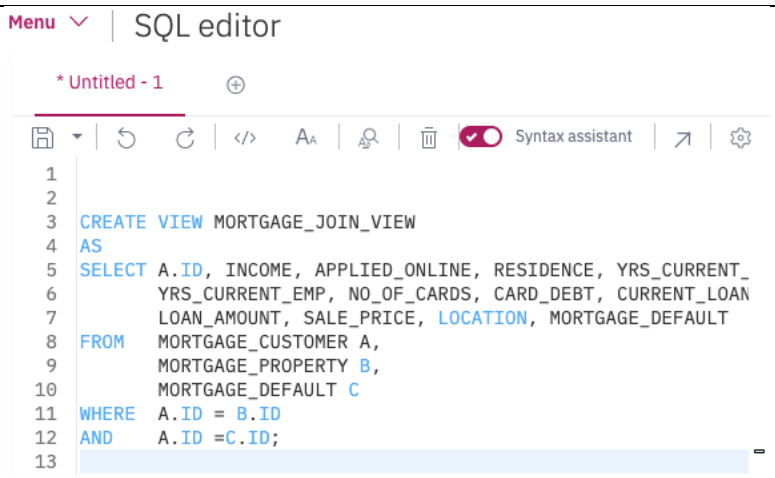


9.3. Creating virtual table

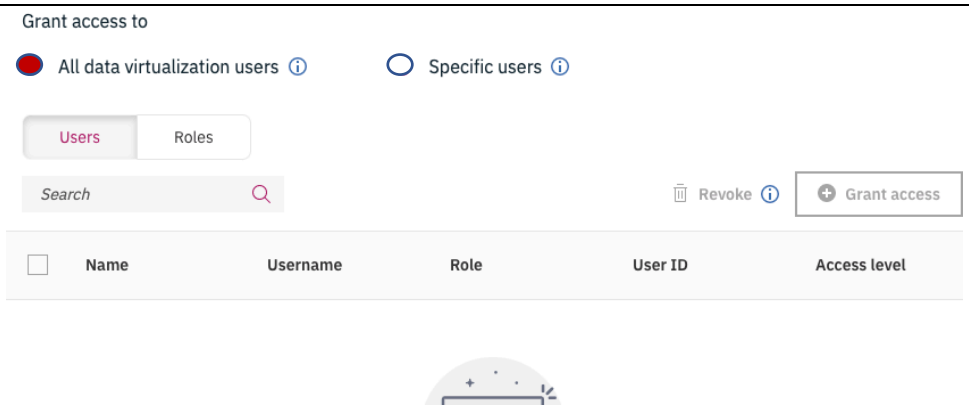
You can create a new virtual table based on existing tables under **My data** section. You can use "drag and drop" or write your own SQL to create the view.

- Click **Collect** > **Data virtualization** > **Menu** > **SQL editor** to access the editor.
- Copy the following SQL statement and paste it on the editor
- Click on **Run all**

```
CREATE VIEW MORTGAGE_JOIN_VIEW
AS
SELECT A.ID, INCOME, APPLIED_ONLINE, RESIDENCE, YRS_CURRENT_ADD,
       YRS_CURRENT_EMP, NO_OF_CARDS, CARD_DEBT, CURRENT_LOANS,
       LOAN_AMOUNT, SALE_PRICE, LOCATION, MORTGAGE_DEFAULT
FROM   MORTGAGE_CUSTOMER A,
       MORTGAGE_PROPERTY B,
       MORTGAGE_DEFAULT C
WHERE  A.ID = B.ID
AND    A.ID =C.ID;
```



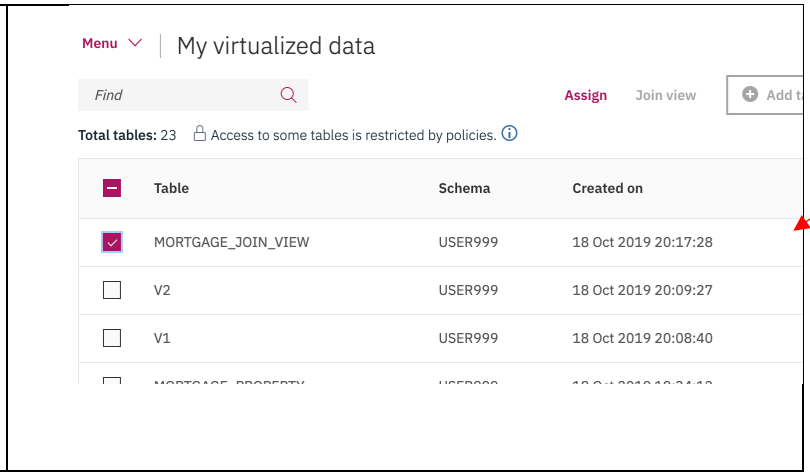
- Click **Collect** > **Data virtualization** > **Menu** > **My virtualized data** to access the virtual table MORTGAGE_JOIN_VIEW
- Check the box associated with MORTGAGE_JOIN_VIEW
- Click on the table actions menu ⋯
- Select **Manage access** option
- On grant access window select All data virtualization users
- Click **Add**



9.4. Add virtual table to catalog

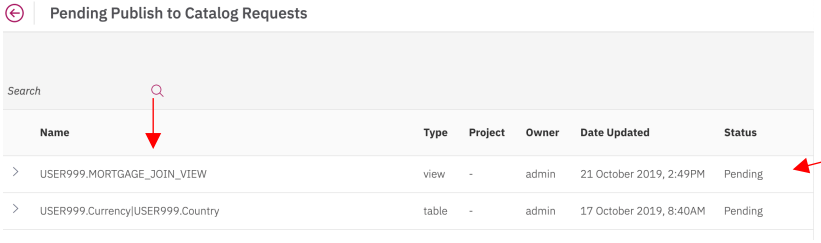
Once you create a virtual table, you can add it to the catalog, making it easily searchable.

- Click **Collect** > **Data virtualization** > **Menu** > **My virtualized data** to find the virtual table just created.
- Mark the checkbox associated with virtual table
- Chose **Submit to catalog** from table action
- Click on **Confirm**



9.5. Publish virtualized table

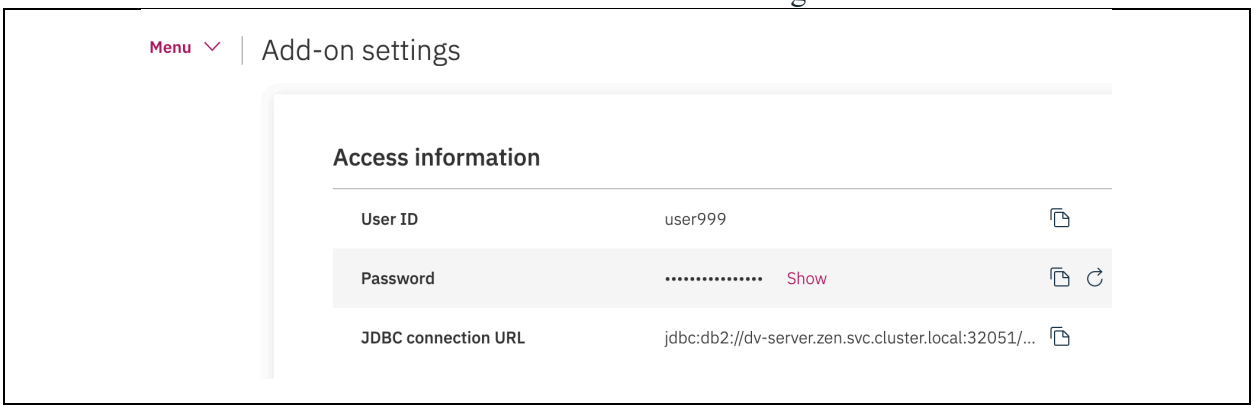
A data steward needs approve the published request before the asset is added to the global data catalog. You signed in as user 'admin', it should allow to publish the virtual table.

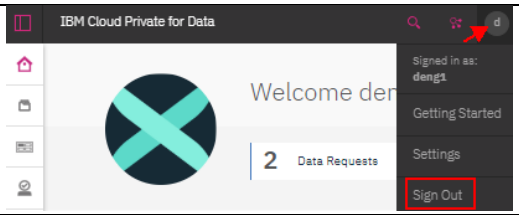


- Click on access the **Home** page
- Click on **Pending Publish to Catalog Requests**
- Click on icon on left for virtual table **MORTGAGE_JOIN_VIEW** that you created
- Click on **Approve**

9.6. Access information for virtual table

To access virtual table from external application, you need the JDBC connection information. Click on **Collect** > **Data Virtualization** > **Menu** > **Service settings** to find out access information. You will use this information later in the building model section.

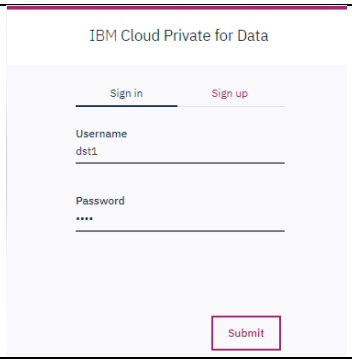




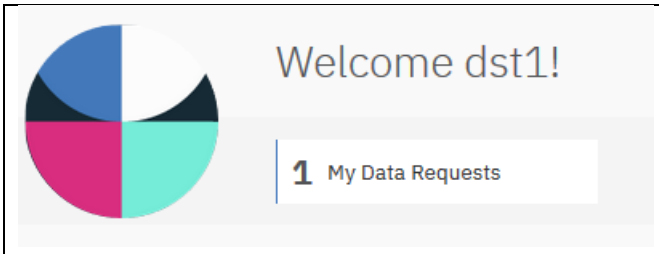
Sign out from user **deng1**

10. Build Model

With Cloud Pak for Data, you can collaborate with other team members on analytic projects to create visualizations and machine learning models with data from your enterprise. In this step you will build a simple model to predict the possibilities of mortgage default by customer. The object of this model is to show the functionality of Cloud Pak for Data, not the prediction accurecy. One can use lot more data and build a compmex algorithem to get better accurecy.



Sign: in to the Cloud Pak for Data web console as user ‘dst1’ and password is ‘dst1’ that you created earlier.



At this point data engineer deliver the data set for the data you requested.
You can go to the home page by clicking on icon from left pane

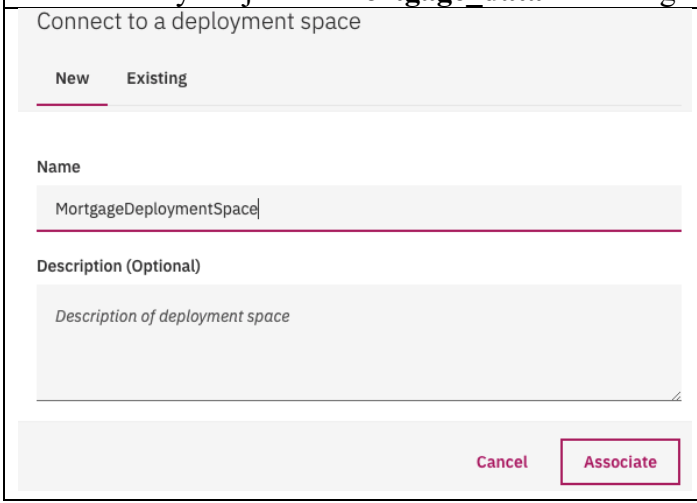
10.1. Navigate to analytics project

Select **Projects** option from the left pane and click on the analytics project ‘mortgage_data’ that you created earlier.

10.2. Create deployment space

Create a separate deployment space for your project ‘mortgage_data’.

Choose : My Projects > **mortgage_data** > Settings > Associate a deployment space > New



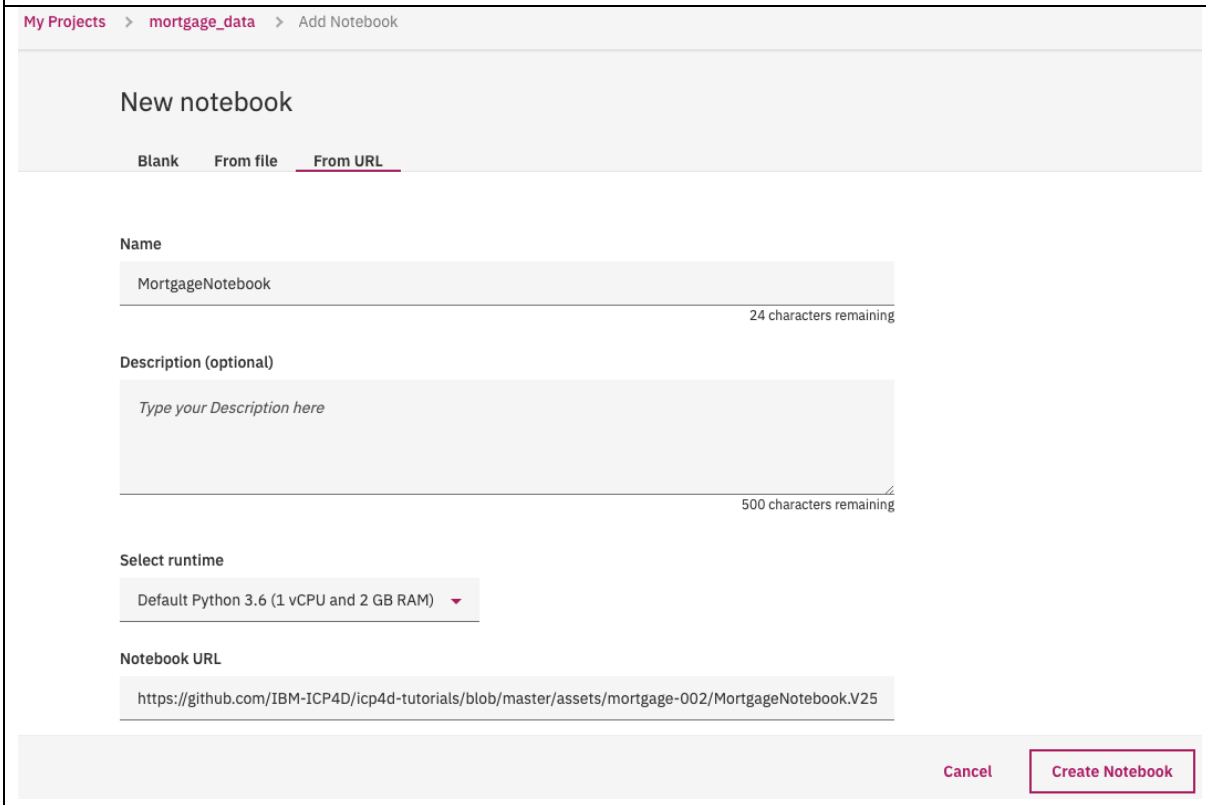
Name new deployment space as ‘MortgageDeploymntSpace’

Click on **Associate**

10.3. Create notebook

Create a notebook from a predefined Jupyter notebook that available on Github.

- Go to : My Projects > **mortgage_data** > Add to project
- Chose asset typew as Notebook
- The new notebook needs to create from URL
- Name the notebok as **MortgageNotebok**
- Use notebook URL as **https://github.com/IBM-ICP4D/icp4d-tutorials/blob/master/assets/mortgage-002/MortgageNotebook.V25.jupyter-py36.ipynb**
- Click on **Create Notebook**



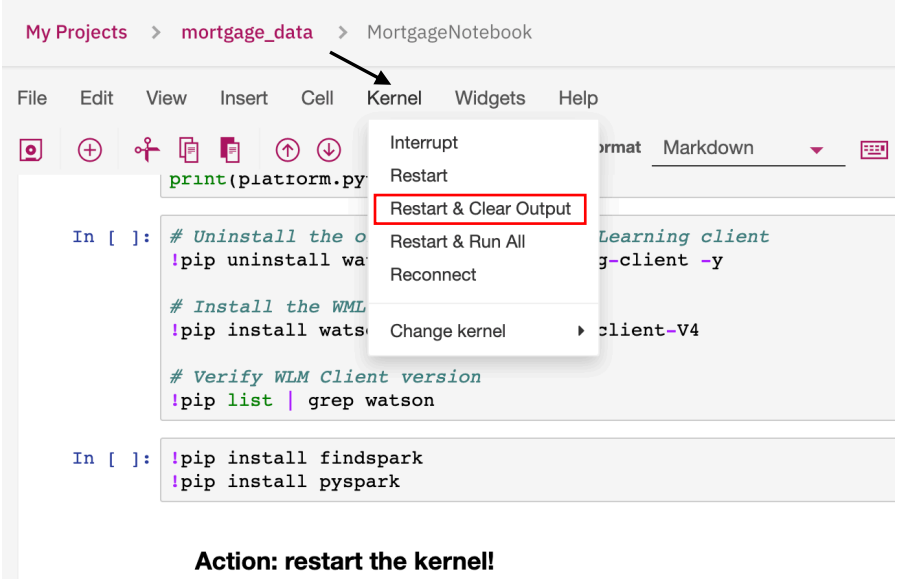
10.4. Review and run notebook

The majority of the code in the notebook is standard open source code that's used for various steps in the predictive analytics process.

Switch to edit mode by clicking on icon from top of the screen.

Do not run all cells at once. Follow the instruction below to run the notebook.

Run the **Step 1: Intall** section first. Once all package installed make sure resttart the Python kernel before move on next step.



Go the **Step 2: Authenticate** section and update the **url**, **username** and **password** fields with your CPD UI console details and access credential.

Step 2: Authenticate

```
[ ]: WML_CREDENTIALS = {
        "instance_id": "openshift",
        "url" : "https://zen-cpd-zen.apps.testcluster.demo.ibmcloud.com",
        "username":"admin",
        "password": "passw0rd",
        "version": "2.5.0"
     }
```

In the next notebook cell, update the **dsn_url, dsn_uid and dsn_pwd** values with the information available from **Collect > Virtualized data > Menu > Add-on settings**.

```
[ ]: #Enter the values for you database connection found under data virtualization
     dsn_url = "jdbc:db2://dv-server.zen.svc.cluster.local:32051/bigsql"  # e.g.
     dsn_uid = "user1022"                                                 # e.g.
     dsn_pwd = "sw?#@lT_674MfPI5"                                         # e.g.
```

Run all cells between step 2 and 6.
You may need to change the MORTGAGE_JOIN_VIEW schema name in step 3, according to your environment.

On **Step 7: Set default space** , run the first cell and find out the **GUID** for space name **MortgageDeploymentSpace**.

On the next cell replaced the GUID with one that you found above.

```
In [ ]: # Example: client.set.default_space('b49e13e8-ec68-408d-84a1-957e28c154b1')
        client.set.default_space('GUID')
```

Run through remaing cells, so that it generates and deployed the model.

Before exit, save the notebook .

10.5. Test the model

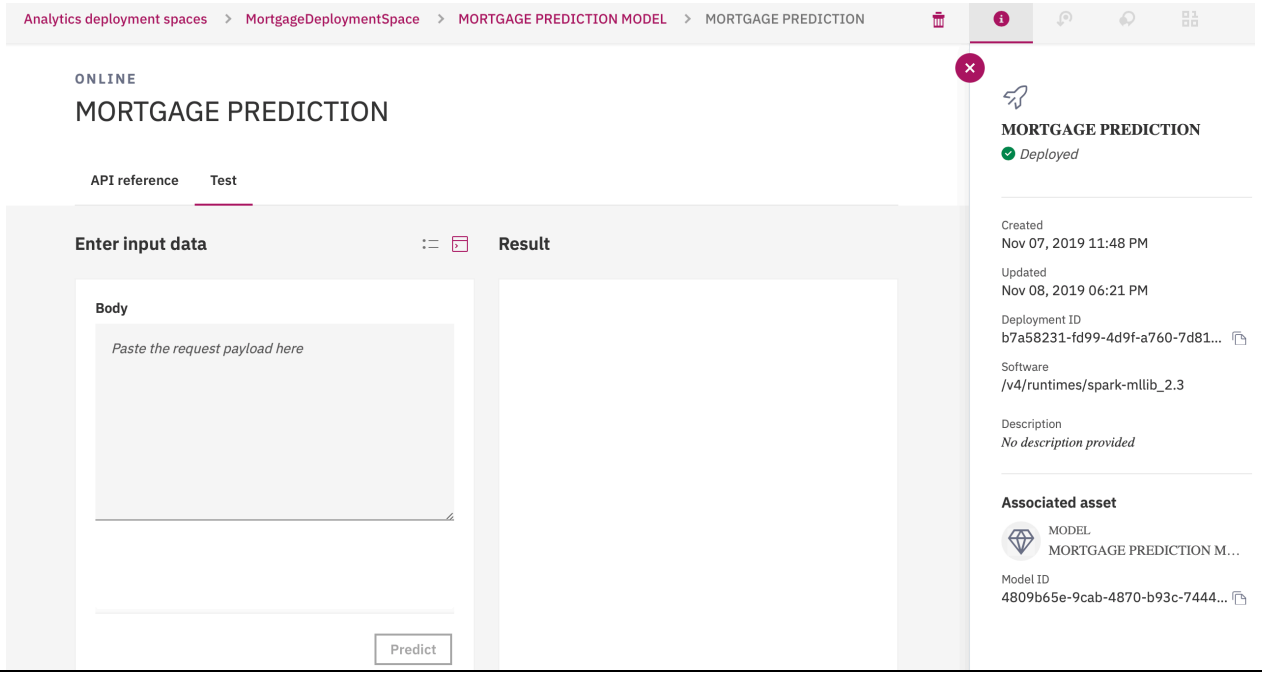
Go to: Analyze > Analytics deployment to access deployed model

Select the **MortgageDeplymentSpace** from the list of analytic deployment space

Click on the **MORTGAGE PREDICTION MODEL**

Choose the **MORTGAGE PREDICTION** model

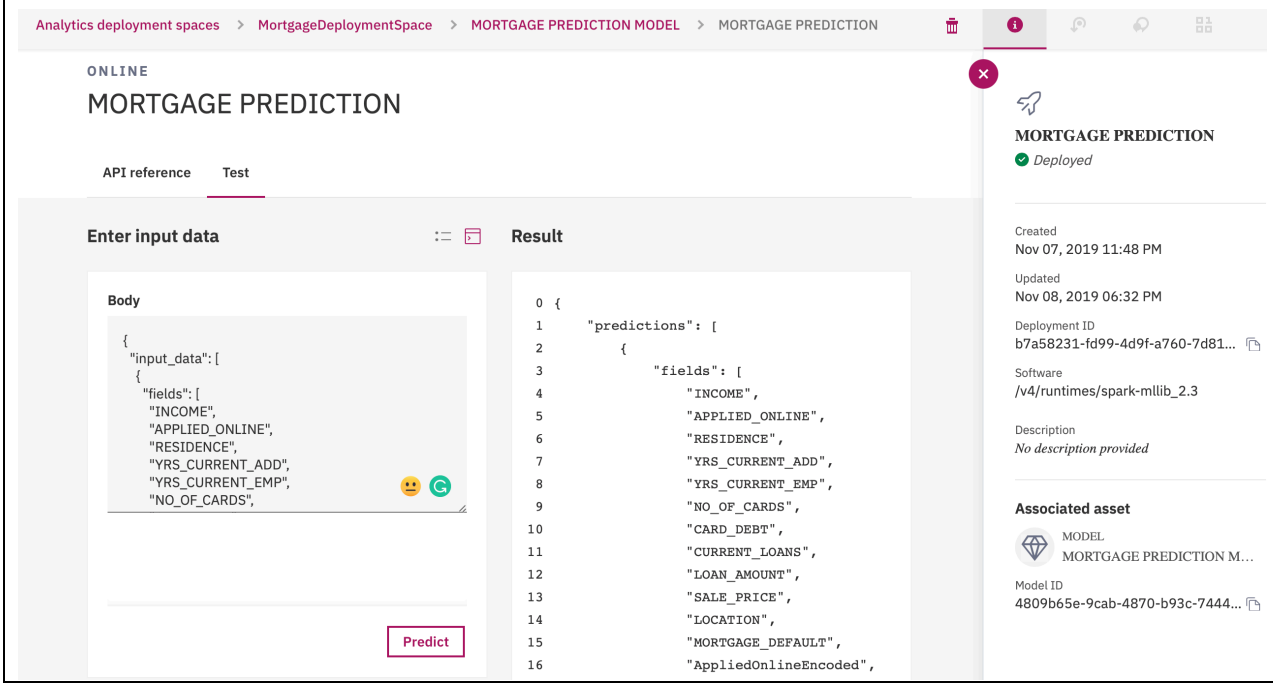
Click on **Test** tab



<table>
<tr>
<td>

```
{
  "input_data": [
    {
      "fields": [
        "INCOME",
        "APPLIED_ONLINE",
        "RESIDENCE",
        "YRS_CURRENT_ADD",
        "YRS_CURRENT_EMP",
        "NO_OF_CARDS",
        "CARD_DEBT",
        "CURRENT_LOANS",
        "LOAN_AMOUNT",
        "SALE_PRICE",
        "LOCATION"
      ],
      "values": [
        [
          43151,
          "N",
          "P",
          6,
          9,
          1,
          750,
          1,
          8600,
          320000,
          110
        ]
      ]
    }
  ]
}
```

</td>
<td>Copy this sample data and paste it on the Enter input data box.

Click on Predict</td>
</tr>
</table>

According on input values, model will predict and displays the result.