Cloud Pak for Data

Version 2.5 or higher

Tutorial – Mortgage

Contents

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 IBM Bluemix Cloud. 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 on Cloud Pak for Data
- Use analytics dashboard to build visualizations
- Build a simple machine learning model from 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 git@github.com:IBM-ICP4D/ICP4DTutorial.git
```

Change to the tutorials directory:

```
cd ICP4XTutorial/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.

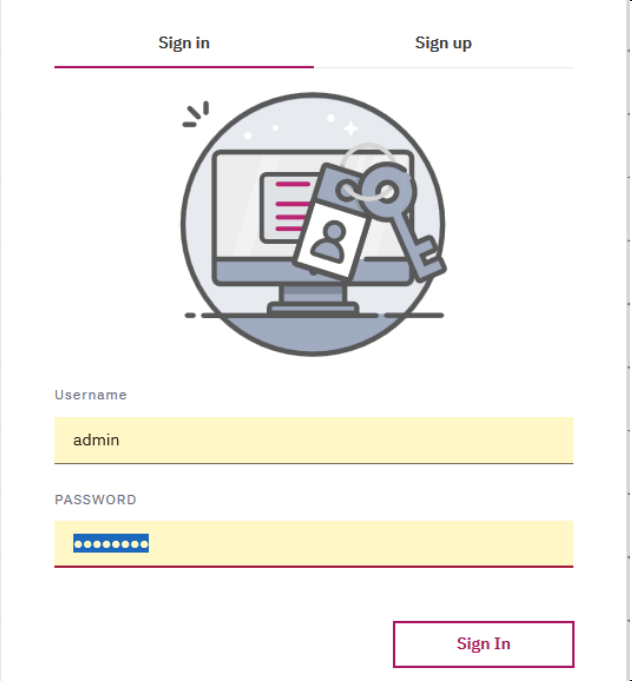
3.1. Access credential for Db2 database

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 in to Cloud Pak for Data web console as Administrator

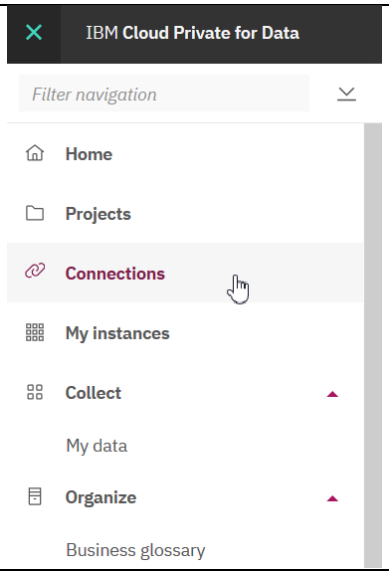
You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data web console. Starting from here all instruction need to execute on Cloud Pak for Data web console only. You need to login as admin who has administrator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

	Sigh in to 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.

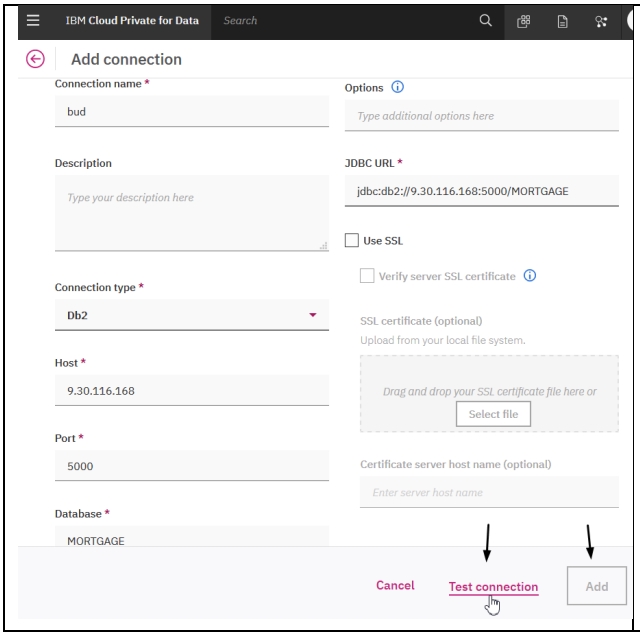
4.2. Navigate to Connections



On the left pane choose **Connections**.

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

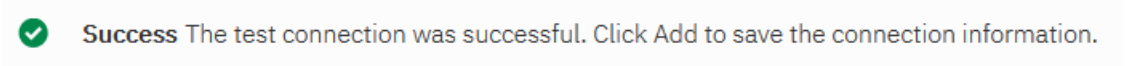
4.3. Add connection



Fill out the **Add Connection** information according to the information provided in step '2.1. 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 **Name**
3. **JDBC URL** is 'jdbc:db2://172.16.171.29:50000/MORTGAGE'
4. **Username** is 'db2inst1' and **Password** is 'password'.

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

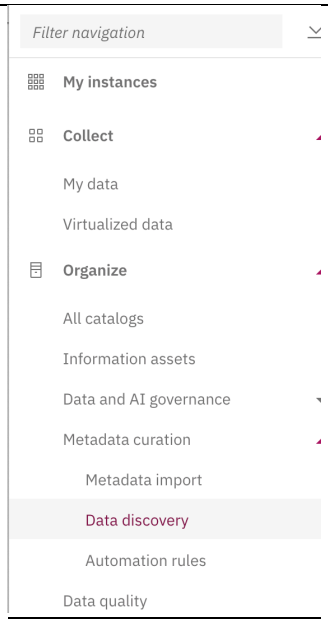


5. Discover Assets

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

5.1. Navigate to discover assets

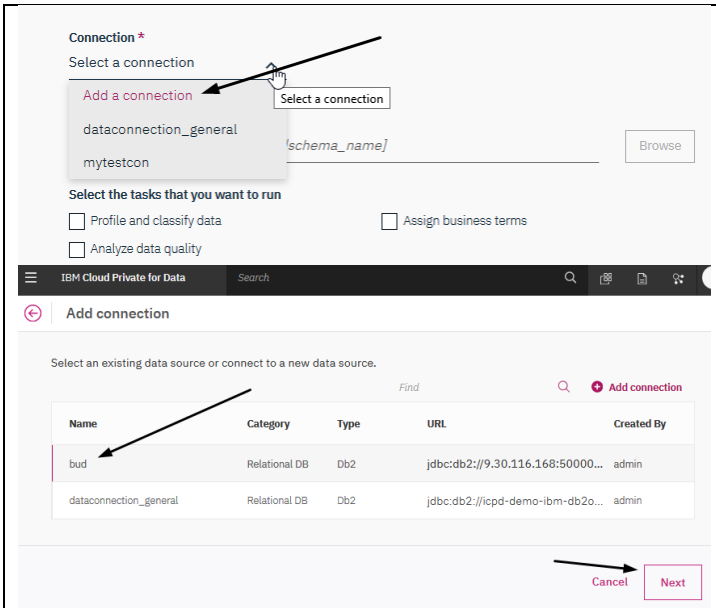
6.



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

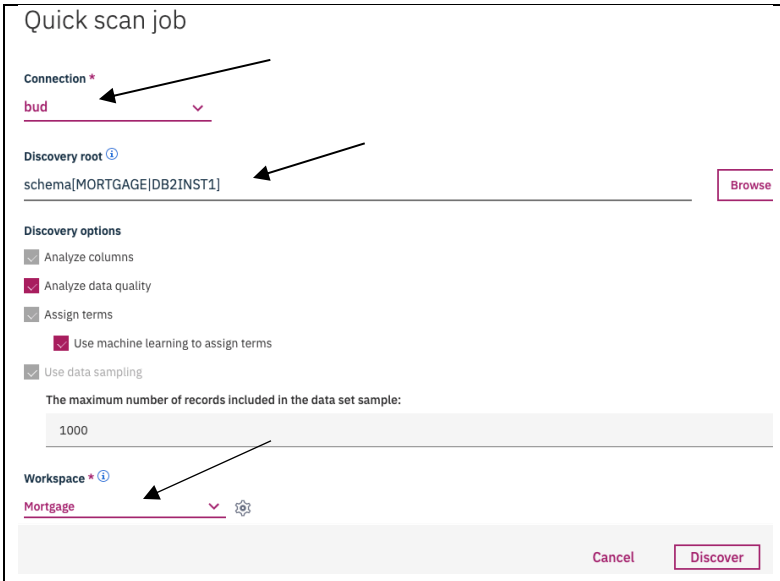
To select discover job

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



To discover assets

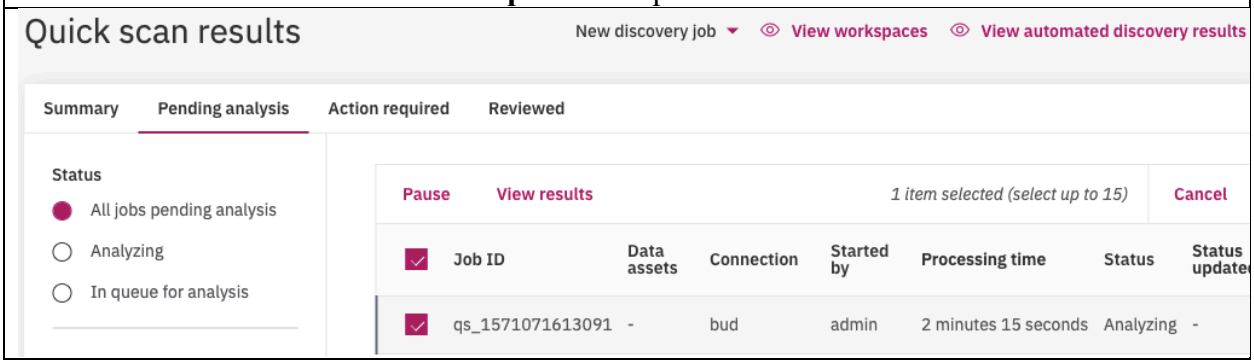
1. Click on **Add a connection**
2. Choose the connection named **bud** that you created previously, click **Next**



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**

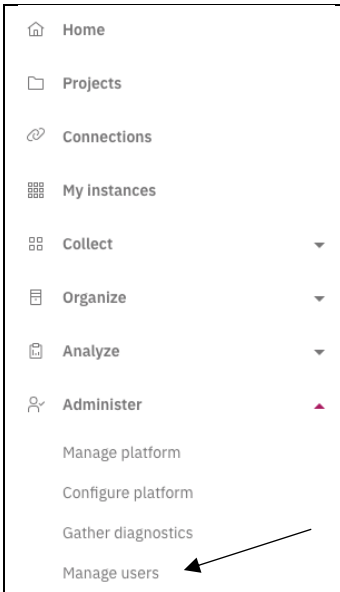
It may take few minutes to complete.

Click on **View results** or **View workspaces** to explore the discover assets.

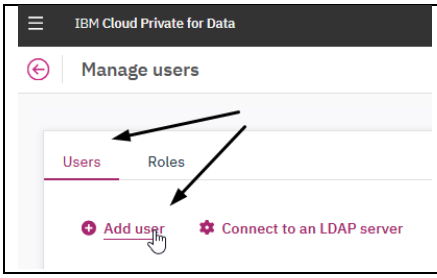


6. Add users

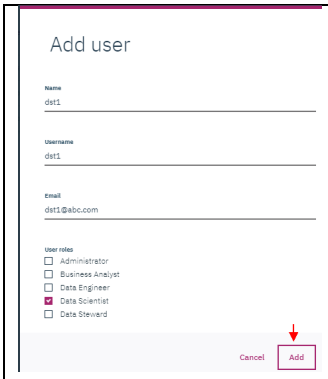
Create users with different roles.



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

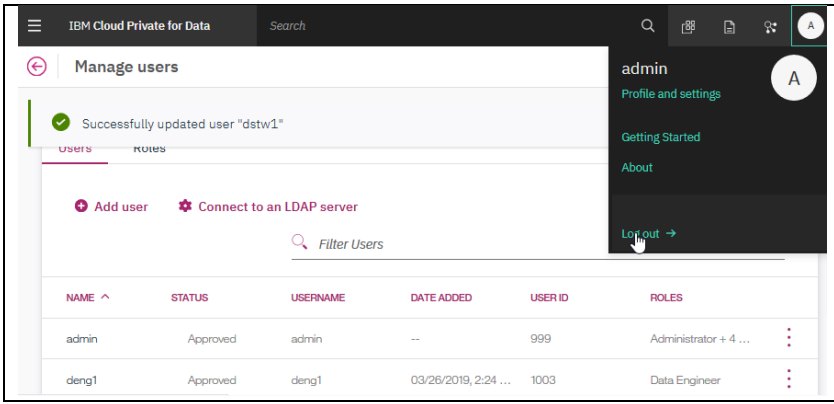
1. Name as **dst1**
2. Username is **dst1**
3. Use a valid email address
4. Set Password as **dst1**
5. Chose the user roles as Data Scientist

Click on **Add** 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**

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.

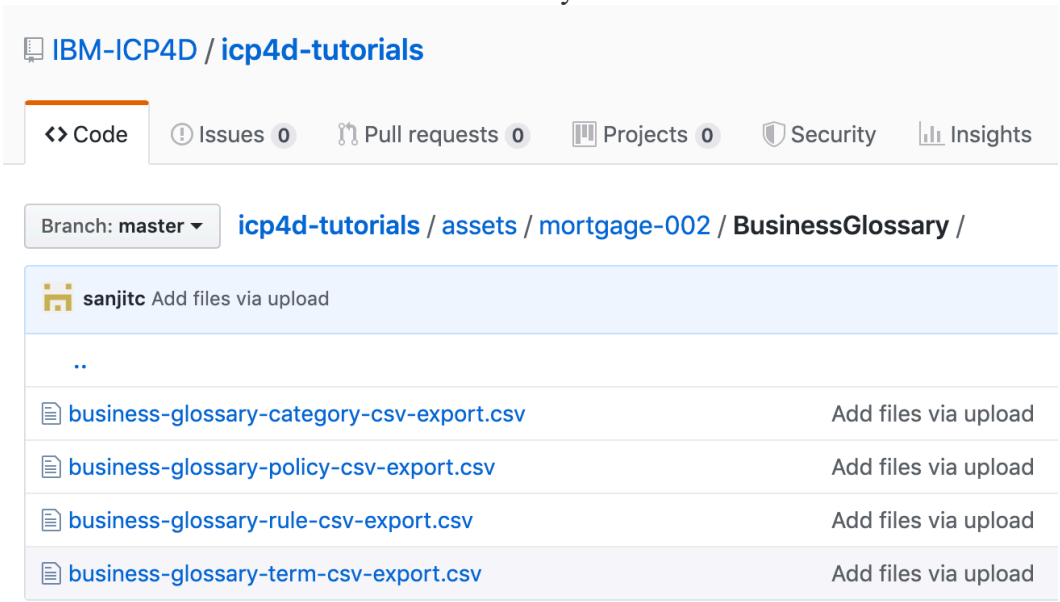
Sigh in to the Cloud Pak for Data web console as user ‘dstw1’ and password is ‘dstw1’ that you created earlier.

7.1. Download Business Glossaries

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

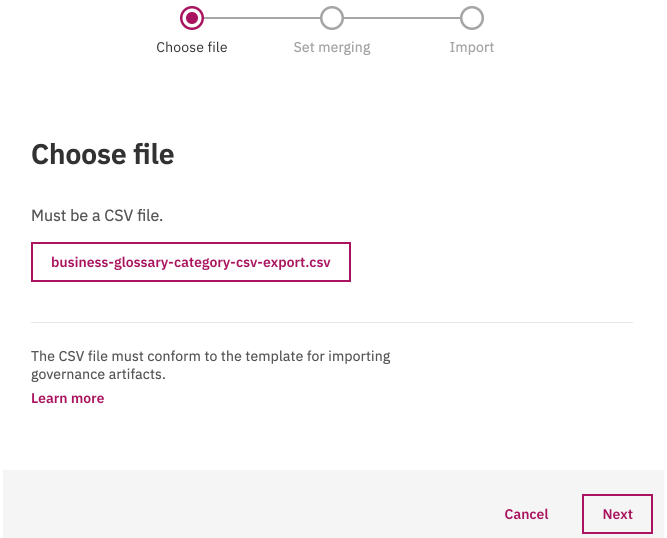
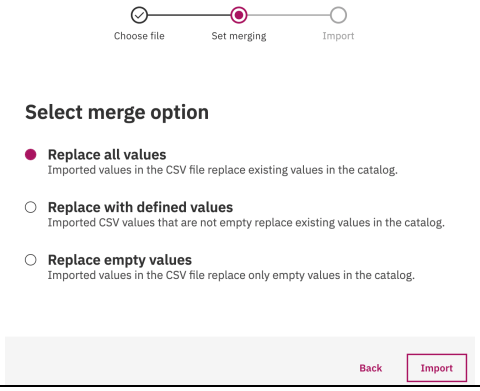
Go to: https://github.com/IBM-ICP4D/icp4d-tutorials/tree/master/assets/mortgage-002/BusinessGlossary

Download all four CSV files and save them locally.



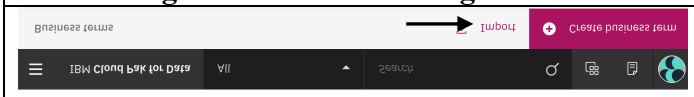
7.2. Import Categories

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

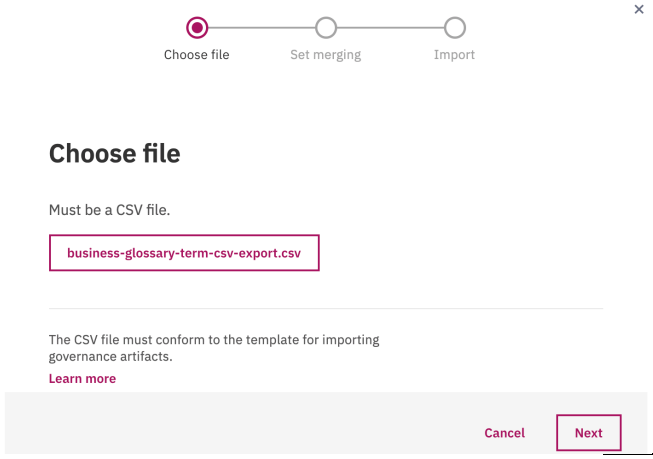
Choose **Organize** > **Data and AI governance** > **Categories** from the left pane.	
	Click on **Import** to import the CSV file contains category information that you downloaded from Git.
	Choose the CSV file location Click **Next**
	Select merge option as **Replace all values** Click **Import**

7.2. Import Terms

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

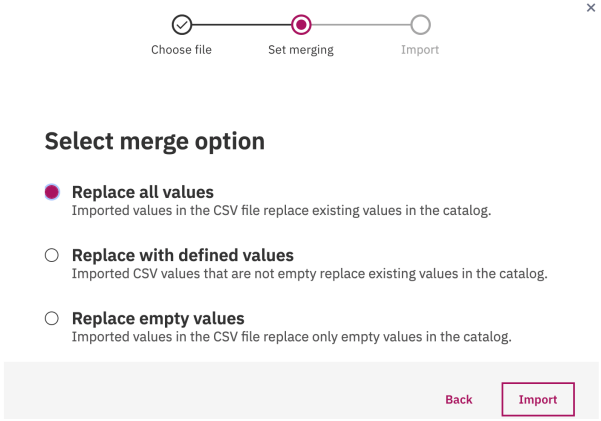


Click on **Import** to import the CSV file contains term information that you downloaded from Git.



Choose the CSV file location

Click **Next**



Select merge option as **Replace all values**

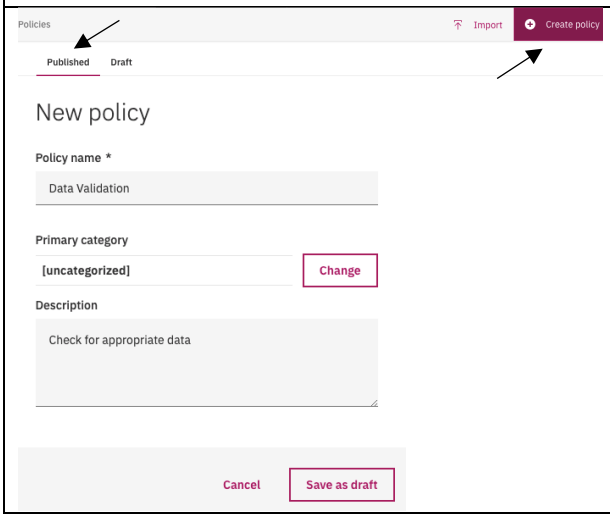
Click **Import**

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**



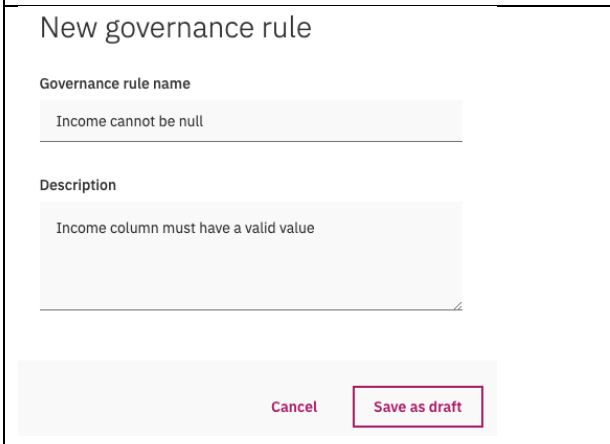
On the **New policy** window create a policy with following information and click on **Save as draft**:

Name: Data Validation
Description: Check for appropriate data

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

7.4. Create a rule

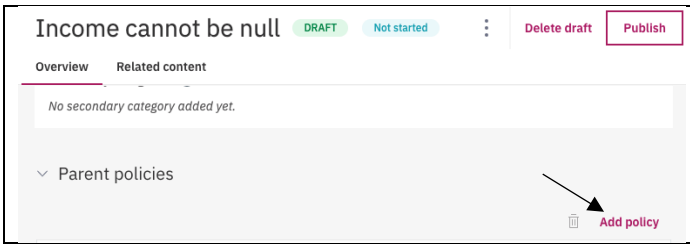
Choose **Organize** > **Data and AI governance** > **Rule** from the left pane
Select **Published** tab and click on **Create Rule**
Choose **Governance rule**



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

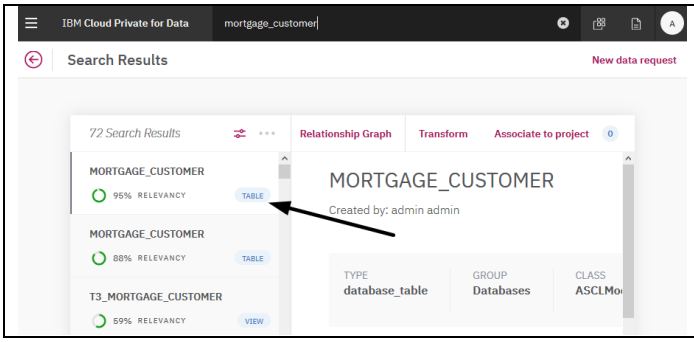
Name: Income cannot be null
Referencing policies: Data Validation
Short Description: Income column must have a valid value

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



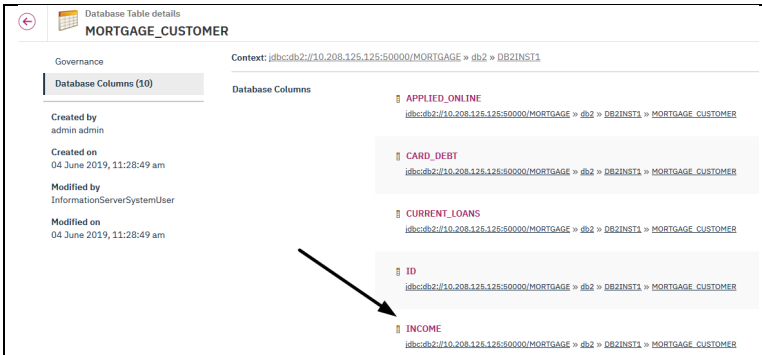
Click on **Add policy** under **Parent policies** to assign the rule to it.

7.5. Add rule to metadata



Click on the enterprise search, Search for 'mortgage_customer' and hit enter
From the search results select table 'mortgage_customer'

Click on **Details** tab at the top



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

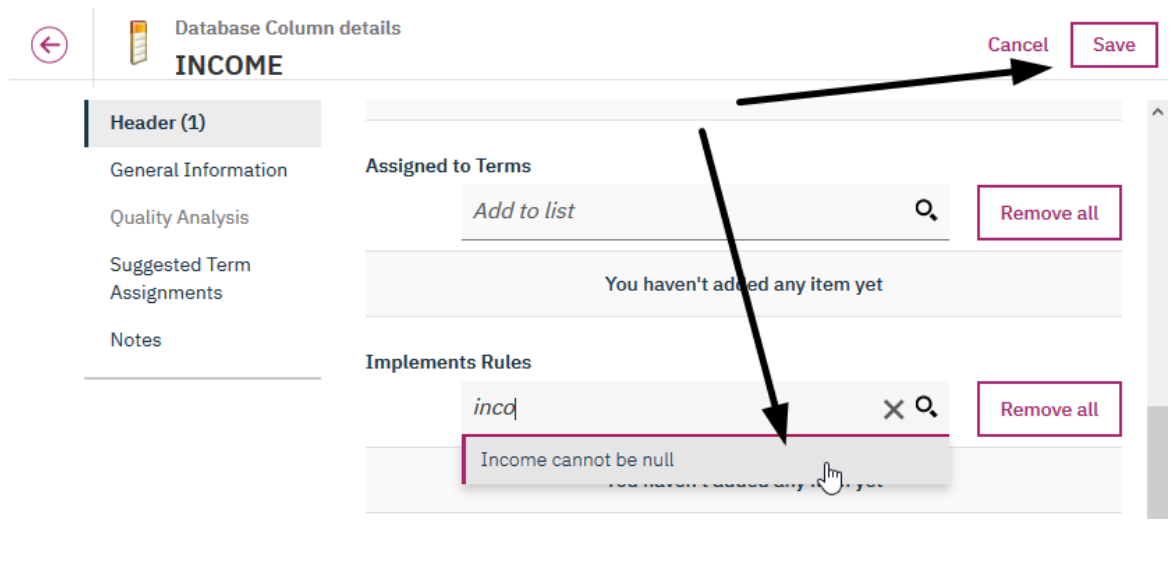
Select INCOME column

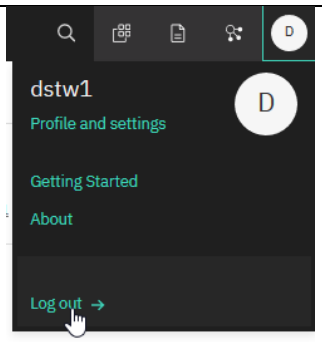
Next click on ⋮ icon (right top corner) and choose **Edit**

Scroll down to **Implement Rules** section

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

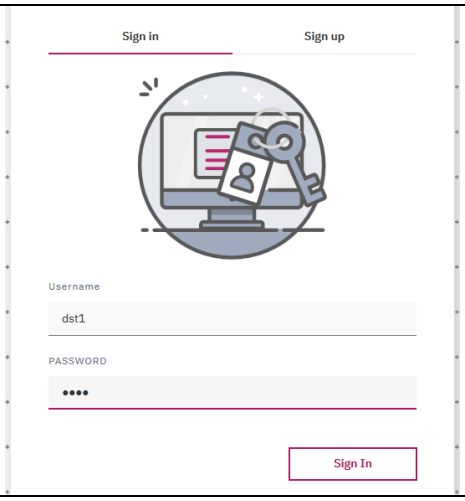
Click on **Save**



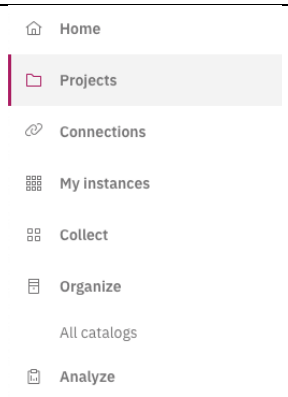
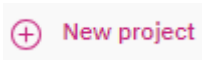
	Log out from user ‘dstw1’

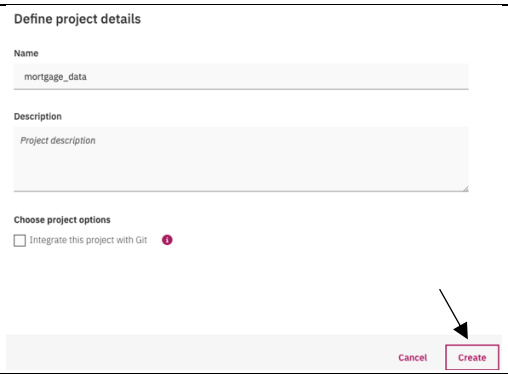
8. Access data as a Data Scientist

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**

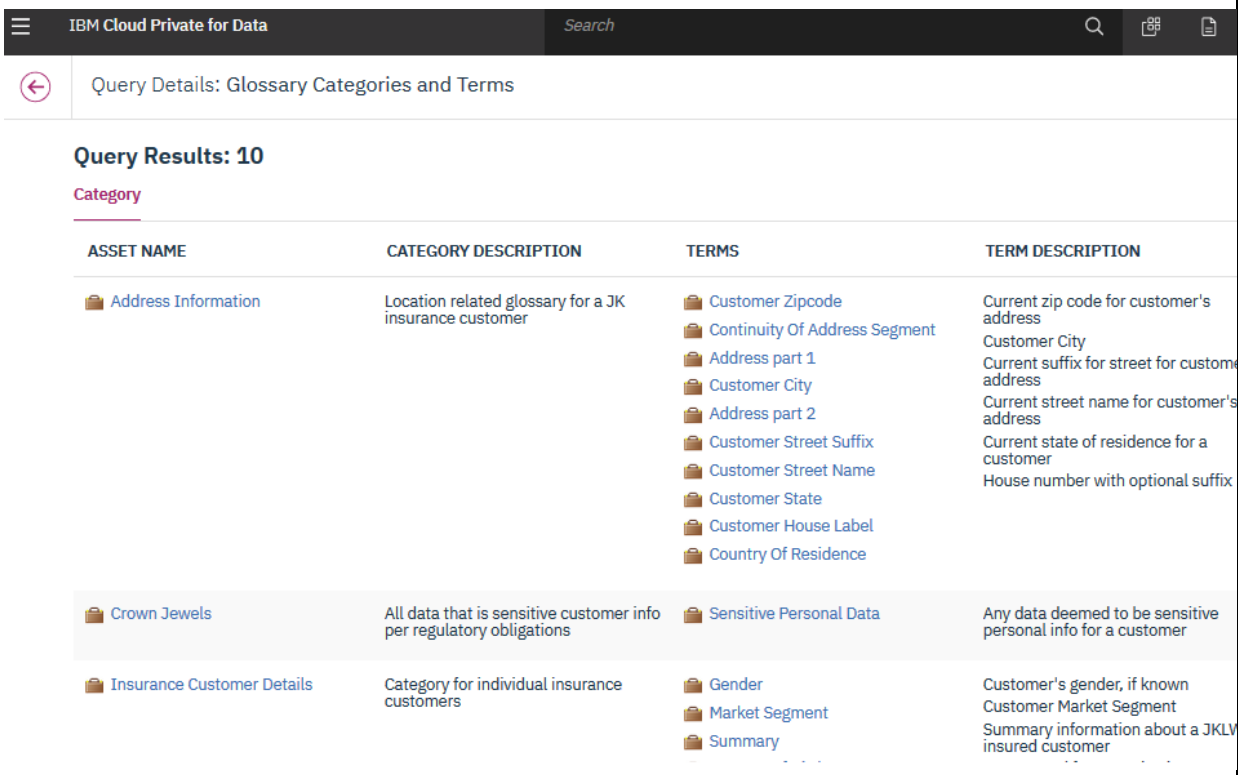
	Provide a project name and click **Create**

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.

Choose **Organize** from the left pane, select **Data Catalog** -> **Queries -> Glossary Categories and Terms**.

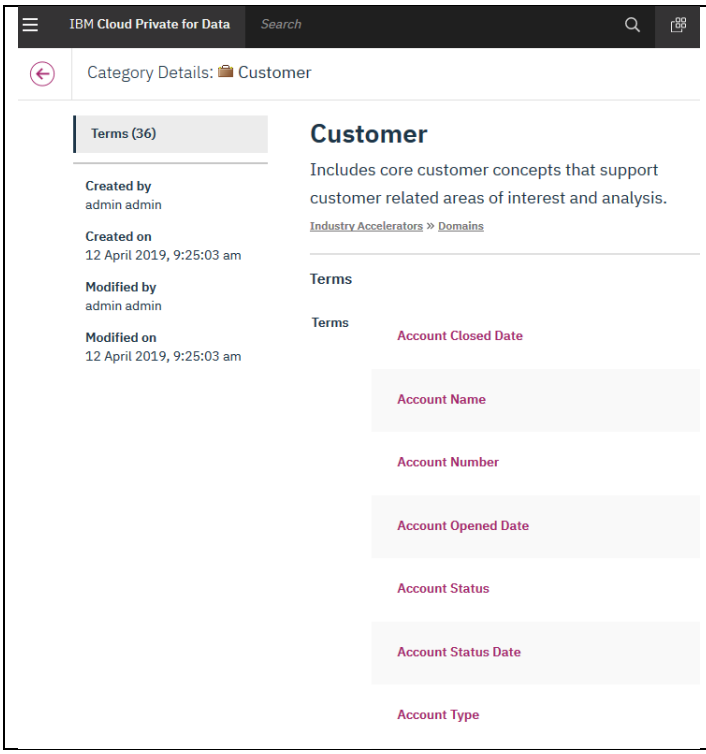
You should have all mortgage related information as follows. Click on each **ASSET NAME**, **TERMS** for additional information. The TERM DESCRIPTION provides a basic information about each term.



For example, click on ASSET NAME **Customer**

8.3. Check Asset Details

Go through each item related to mortgage in glossary to have better idea about data you need for your project.



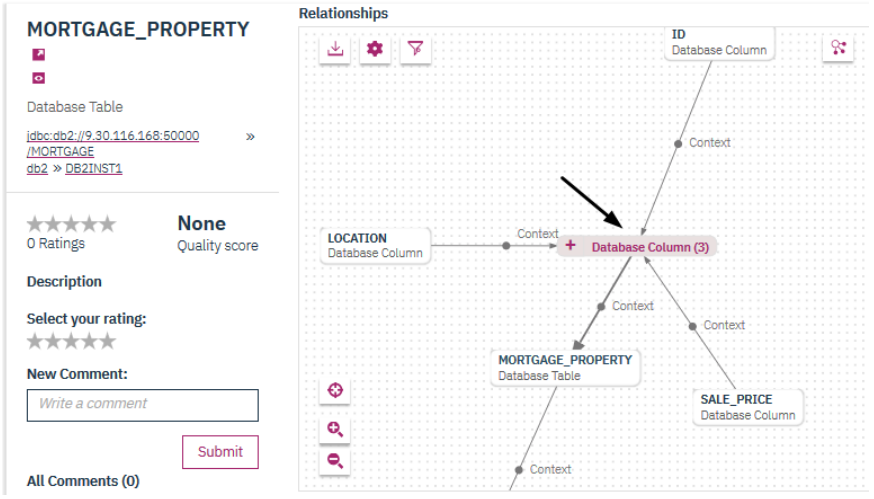
The asset C**ustomer** shows different terms associated with it.

Check each **Terms** for additional information.

8.4. Enterprise search

Click on the enterprise search

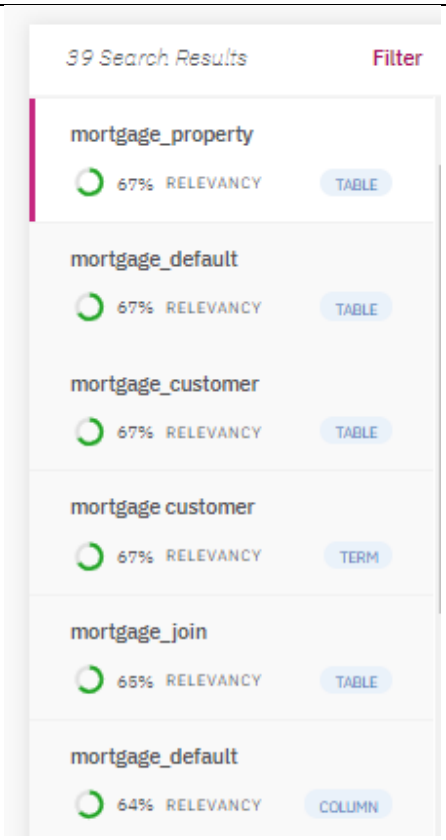
Search for ‘mortgage’ and hit enter



Choose the **mortgage_property** table and click on **Relationship Graph** to see details about the table.

Click on the '+' next to **Database Column** to expand list of columns in the table.

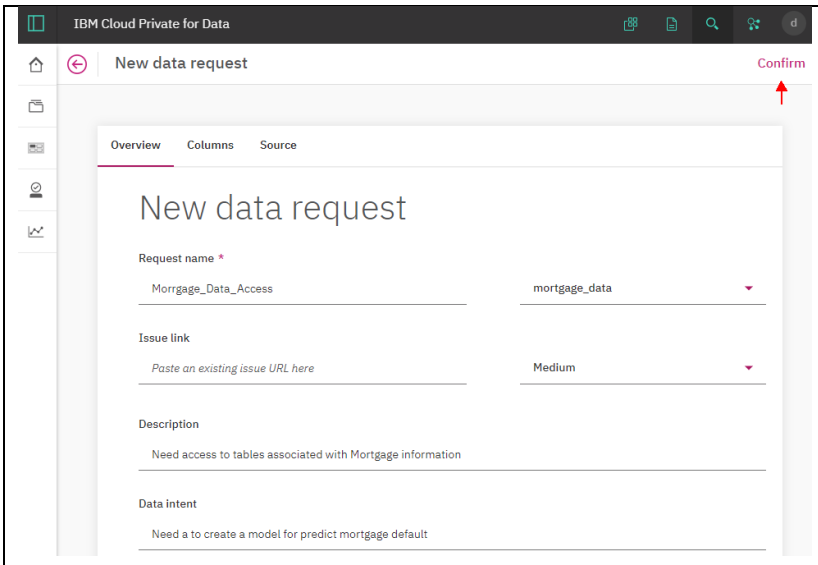
Same way you can view other mortgage related tables.



Go back to the enterprise **Search Result**

The enterprise search will return all objects that mentioned word mortgage but as a data scientist you don't have access to any of those objects.

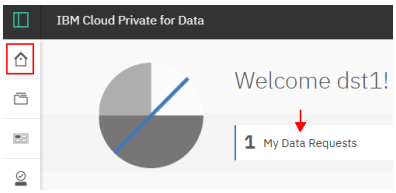
Click on the New Data Request on top right corner for request access to mortgage related datasets.

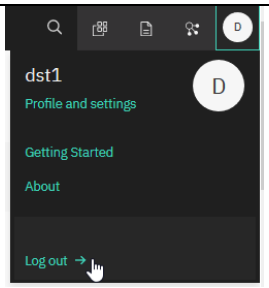


Fill up the **New Data Request** form with detail information as much possible, so a data engineer can provide accurate dataset. Click Confirm and then Submit request.

At this point you need to wait for data engineer to address the data request.

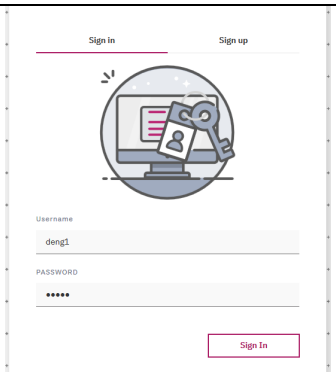
You can go to the home page by clicking on  icon from left pane and check the status of the data request.



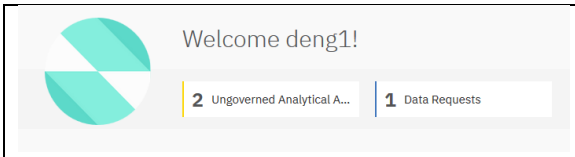


Sign out from user **dst1**

9. Review data request



Sigh in to the Cloud Pak for Data web console as user ‘deng1’ and password is ‘deng1’ that you created earlier.

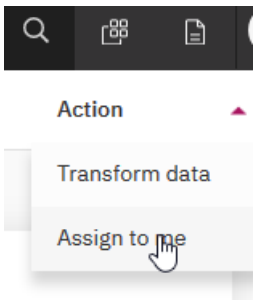


After sing in Click on Go to your home page

Check the **Data Request** tab on the home page.

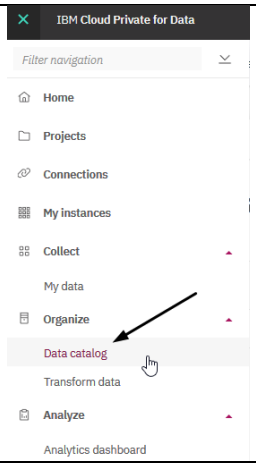
Click on the new data request that submitted by data scientist earlier for review.
After reviewing the request click on Action in top right corner and select assign to me.

ID	Name	Status	Requested by	Assigned to	Priority	Last updated
1	Mortgage_Data_Access	Claimed	dst1	deng1	Medium	3 Jun 2019, 8:15 PM
2	Mortgage_Data_Access_Request	Claimed	dst1	deng1	Medium	3 Jun 2019, 8:41 PM
3	Mortgage_Data_request1	Claimed	dst1	deng1	High	3 Jun 2019, 8:39 PM
4	CustData	New	admin	Unassigned	High	4 Jun 2019, 9:03 AM



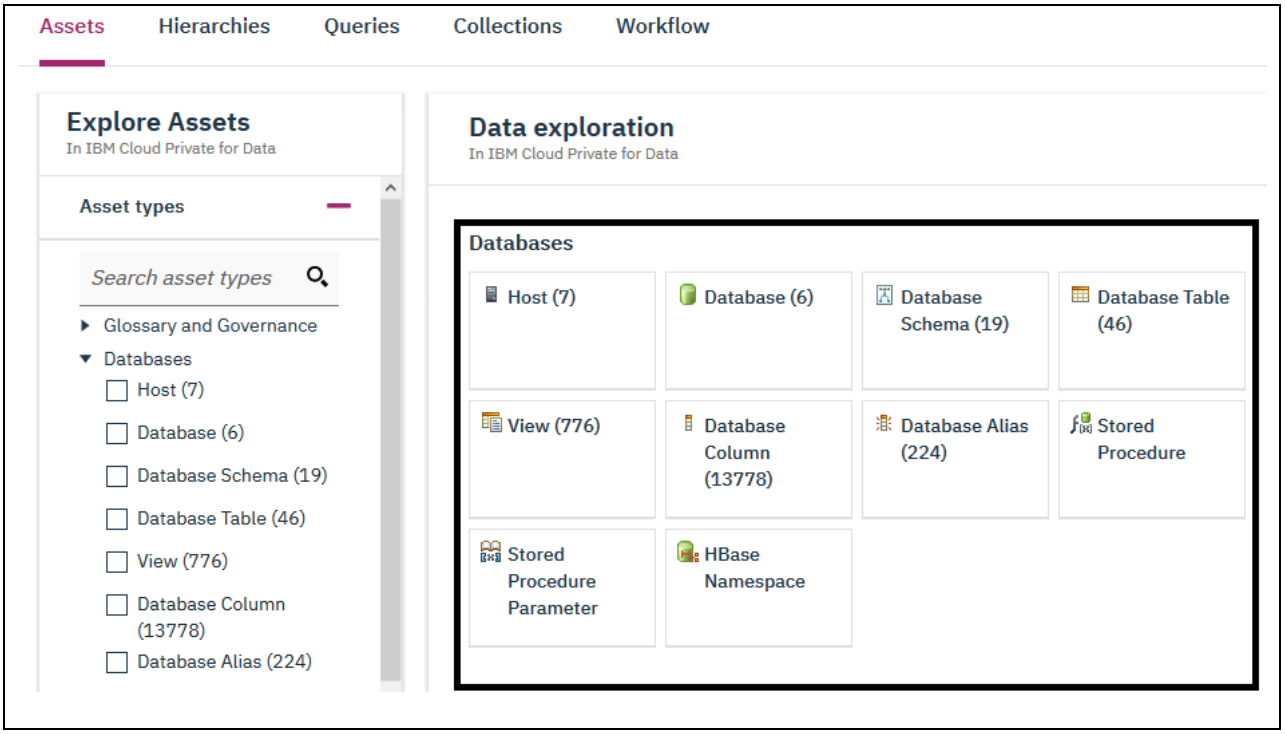
10. Navigate to data catalog

Once discover assets process completed. All database objects automatically cataloged in Cloud Pak for Data. You can review those database object in the catalog.

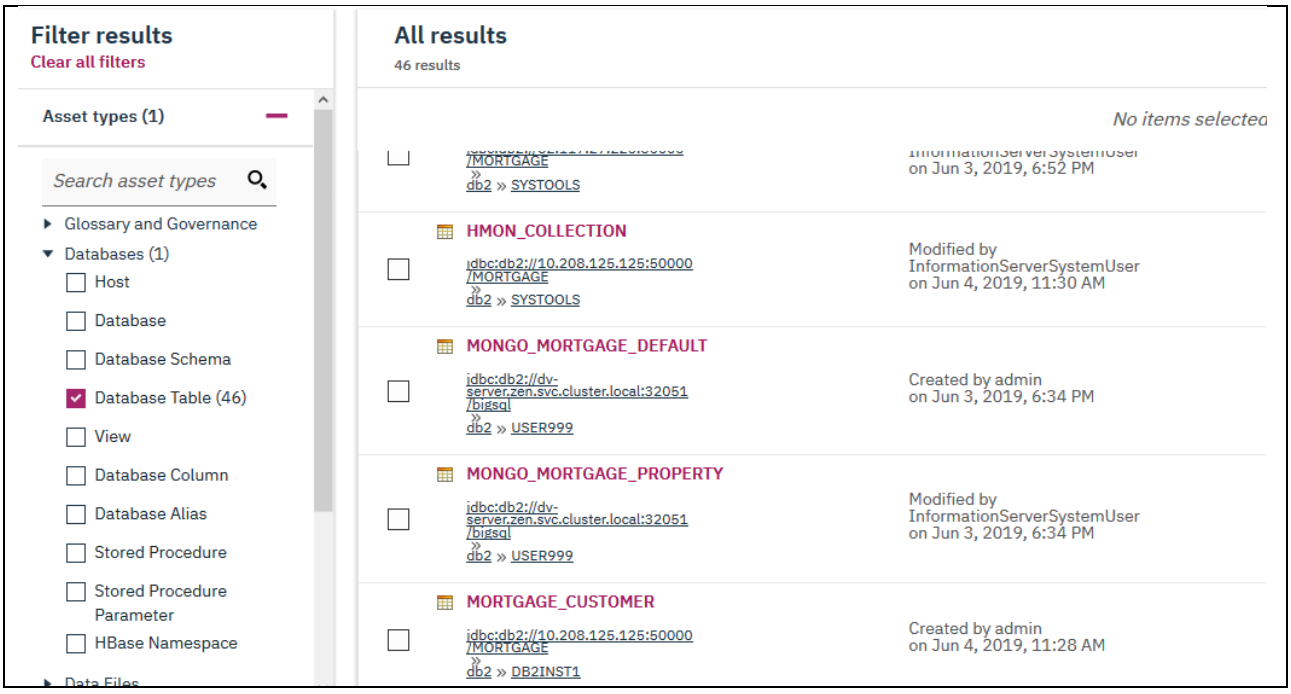


Next go back to **Organize** option on the left pane and choose **Data catalog**.

At this point Cloud Pak for Data should displays all the database objects. You can click each individual object under **Databases** to explore the catalog generated from discover asset previously. Click on the **Database Table** to check tables discovered from Db2. Take a look into the database named **mortgage**.



Under the **Database Tables** you can see 'MORTGAGE_CUSTOMER', 'MORTGAGE_DEFAUT' and 'MORTGAGE_PROPERTY' tables, cataloged from Db2 database.


Filter results
Clear all filters
Asset types (1)
Search asset types
Glossary and Governance
Databases (1)
Host
Database
Database Schema
Database Table (46)
View
Database Column
Database Alias
Stored Procedure
Stored Procedure Parameter
HBase Namespace
Data Files
All results
46 results
No items selected
/MORTGAGE
db2 » SYSTOOLS
on Jun 3, 2019, 6:52 PM
HMON_COLLECTION
jdbc:db2://10.208.125.125:50000/MORTGAGE
db2 » SYSTOOLS
Modified by InformationServerSystemUser on Jun 4, 2019, 11:30 AM
MONGO_MORTGAGE_DEFAULT
jdbc:db2://dv-server.zen.svc.cluster.local:32051/bigsql
db2 » USER999
Created by admin on Jun 3, 2019, 6:34 PM
MONGO_MORTGAGE_PROPERTY
jdbc:db2://dv-server.zen.svc.cluster.local:32051/bigsql
db2 » USER999
Modified by InformationServerSystemUser on Jun 3, 2019, 6:34 PM
MORTGAGE_CUSTOMER
jdbc:db2://10.208.125.125:50000/MORTGAGE
db2 » DB2INST1
Created by admin on Jun 4, 2019, 11:28 AM

11. Data Virtualization

Context: Data virtualization (DV) integrates 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.

11.1. Adding a new data source for Db2

Context: 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.

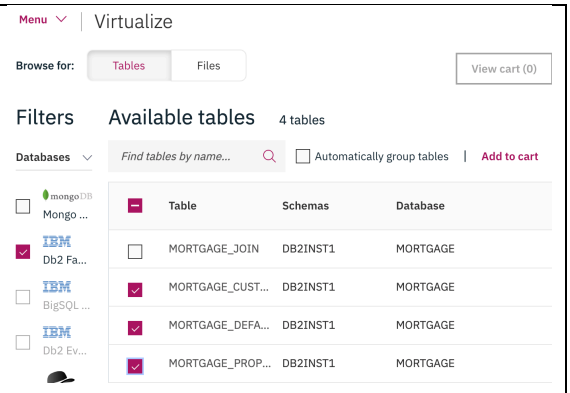
Define a data connection to Db2. Use your existing Db2 database connection for Db2 data source.

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

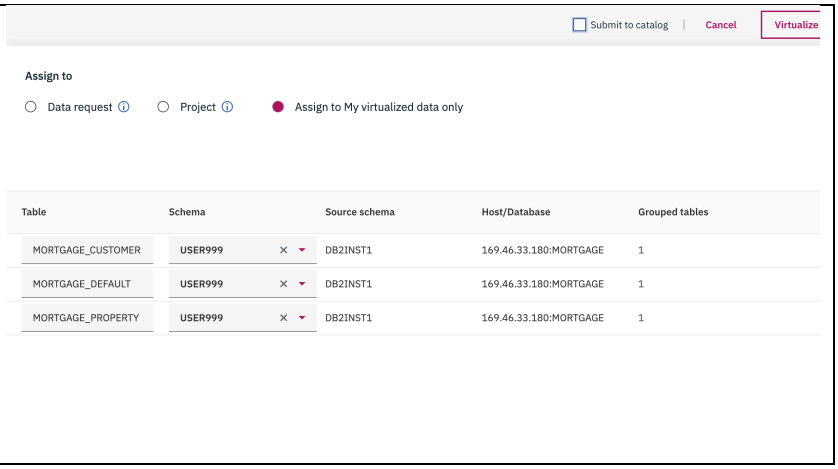
11.3. Select tables for virtualization

Context: 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** > **Virtualized data** > **Menu** > **Virtualize**
- Select tables **MORTGAGE_CUSTOMER, MORTGAGE_PROPERTY and MORTGAGE_DEFAULT** from **MORTGAGE** database, then click **Add to cart**
- Click **View cart**
- Click **Next**



- Select **Neither**
- Uncheck the box for **Submit to catalog**
- Click **Virtualize** to complete the process

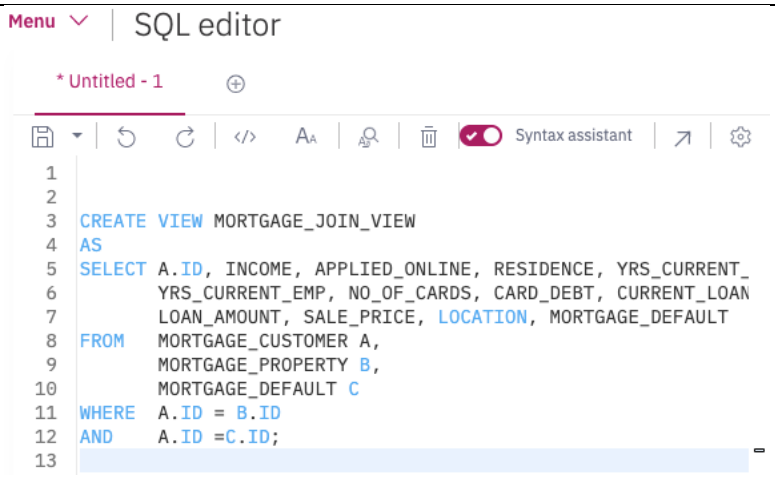


11.4. Creating virtual table

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.

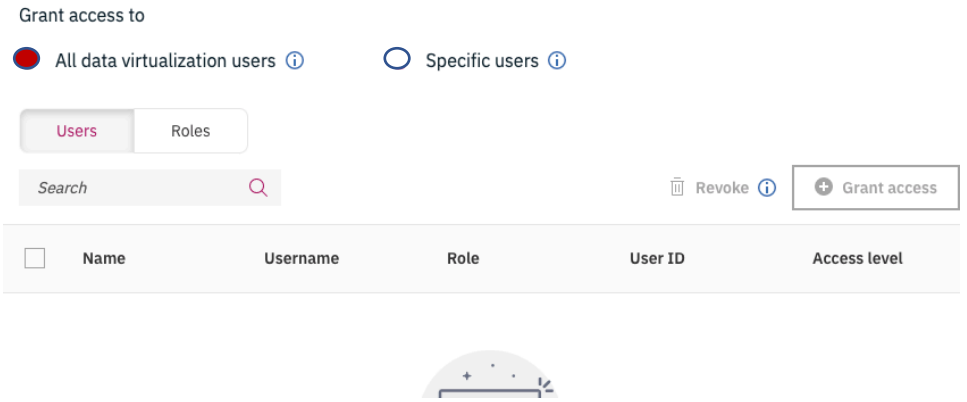
- Click **Collect** > **Virtualized data** > **Menu** > **SQL editor** to access the editor.
- Copy the following SQL statement and paste it on the editor
- Click on **Run all**

```
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** > **Virtualized data** > **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 ⋯

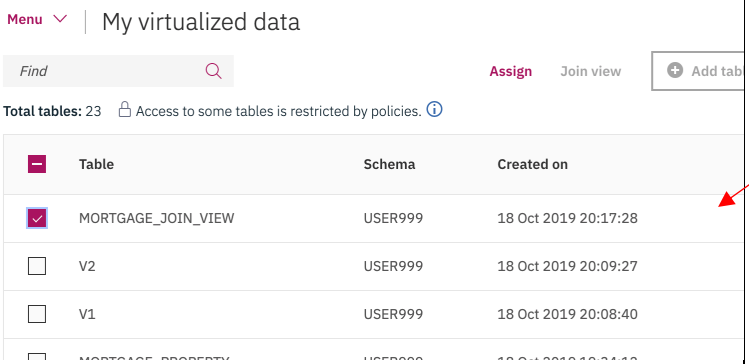
- Select **Manage access** option
- On grant access window select All data virtualization users
- Click **Continue**



11.5. Add virtual table to catalog

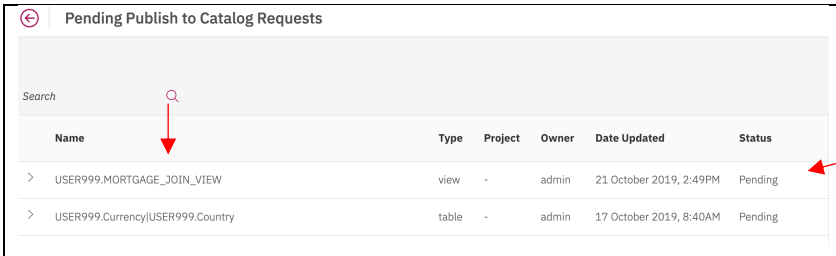
Once you create a virtual table, you can add it to the catalog, making it easily searchable.

- Click **Collect** > **Virtualized data** > **Menu** > **My 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**



11.6. Publish virtualized table

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

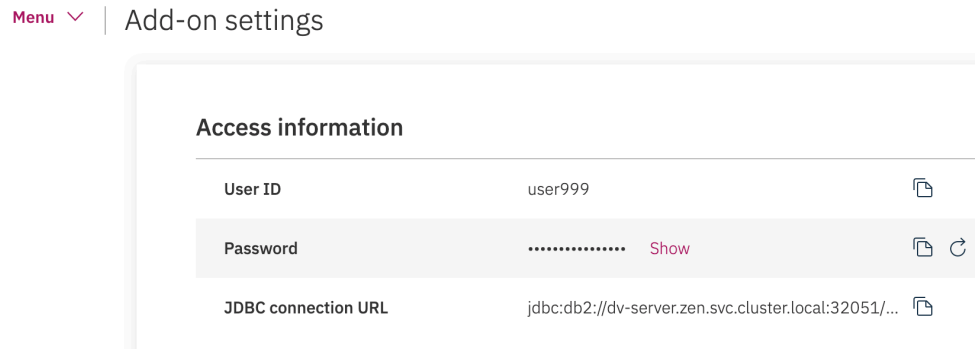


- Click on access the **Home** page
- Click on **Pending Publish to Catalog Requests**

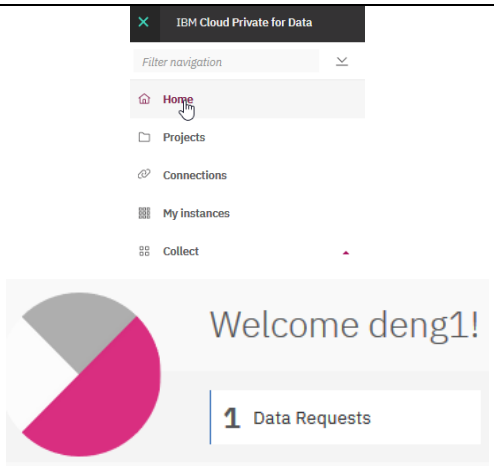
	• Click on ⋮ icon on left for virtual table **MORTGAGE_JOIN_VIEW** that you created • Click on **Approve**

11.7. Access information for virtual table

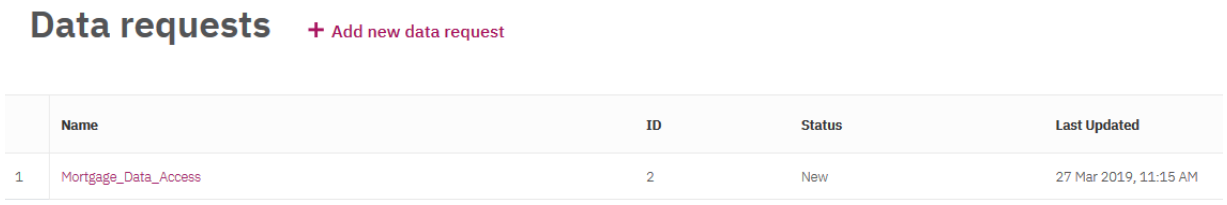
To access virtual table from external application, you need the JDBC connection information. Click on **Collect** > **Virtualized data** > **Menu** > **Add-on settings** to find out access information. You will use this information later in the building model section.



11.8. Deliver Dataset

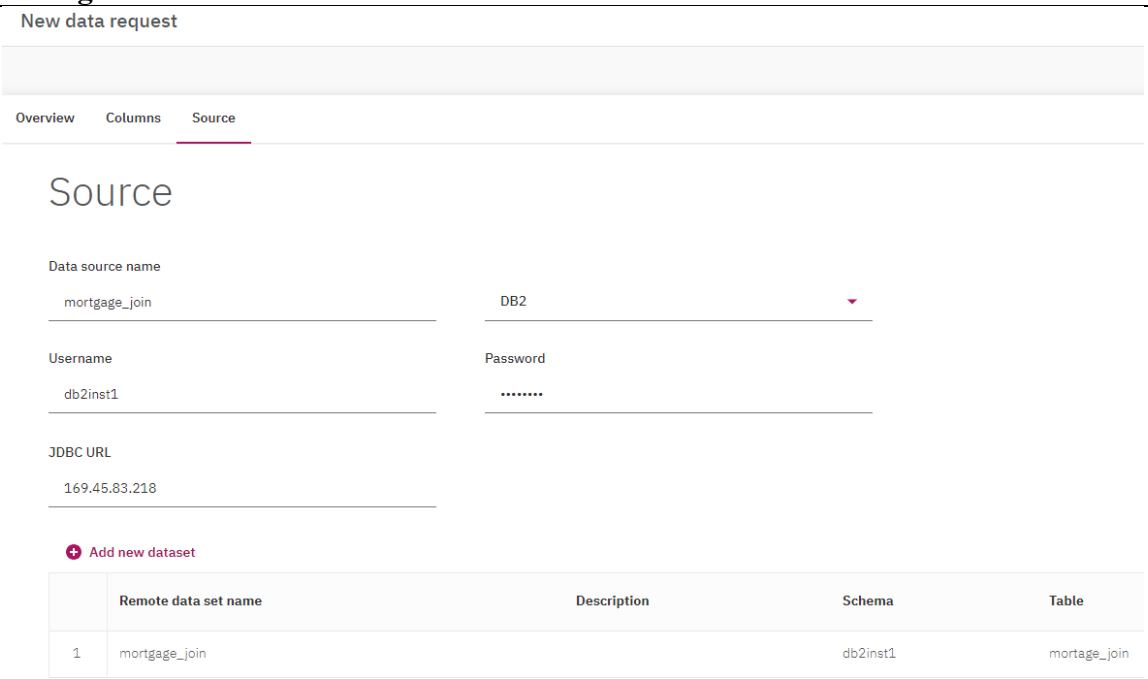
	Go to the home page by clicking on icon from left pane and check the data request tab.

Click on the data request for update that submitted by data scientist earlier.

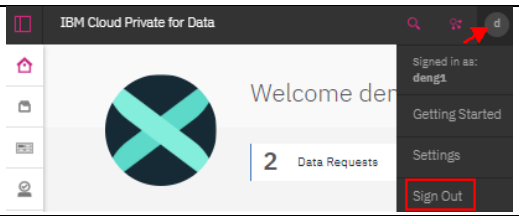


Click on the **Source** and fill out all the necessary information. This information will be picked up by the data scientist later.

Add the **remote data** set information that you created during data transformation. In this case remote data set is MORTGAGE_JOIN_VIEW. Use the **Access information** from the **Add-on settings** information from DV.



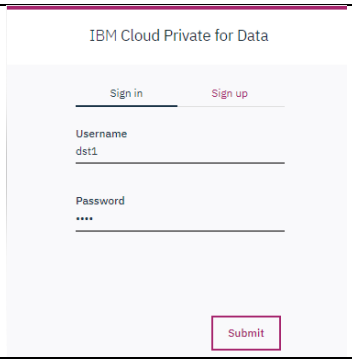
Click on the data request and change the status to **Deliver**.

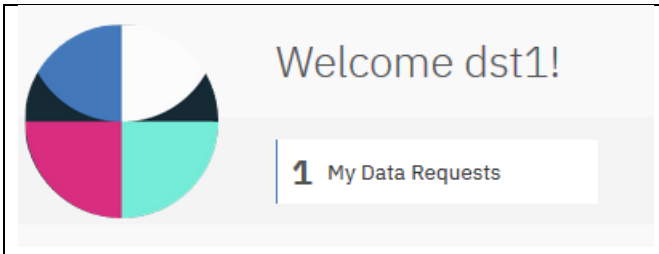


Sign out from user **deng1**

12. 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

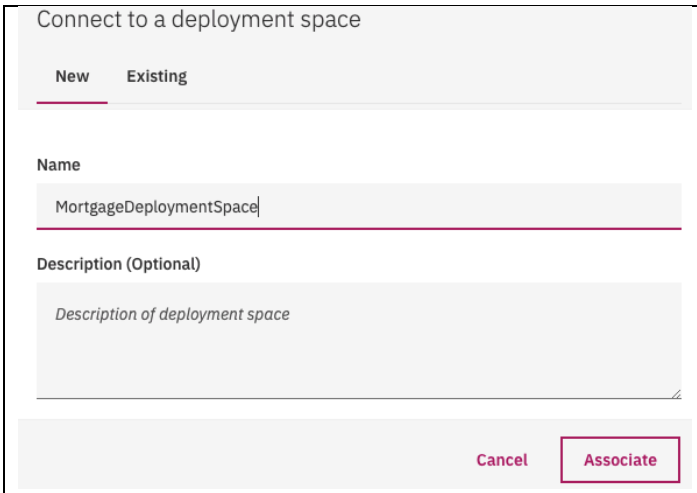
12.1. Navigate to analytics project

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

12.2. Create deployment space

Create a separate deployment space for your project 'mortgage_data'.

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



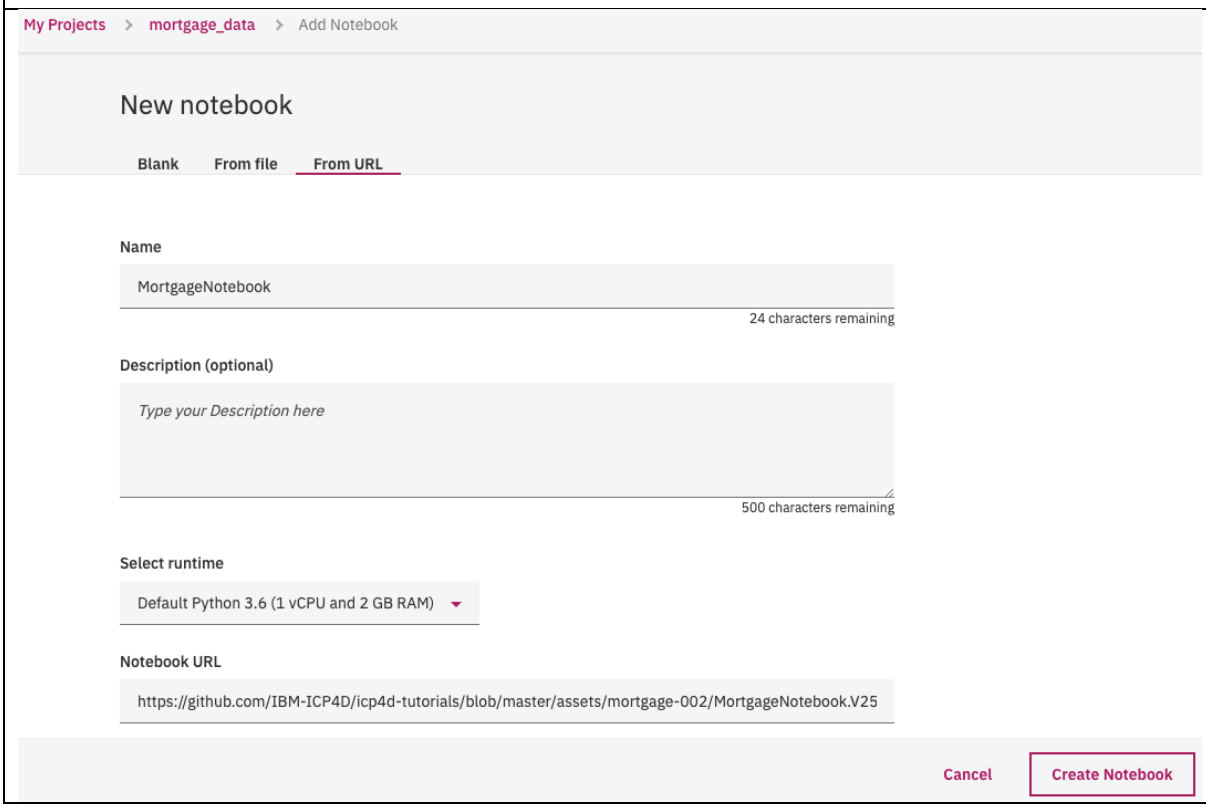
Name new deployment space as 'MortgageDeploymntSpace'

Click on **Associate**

12.3. 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**



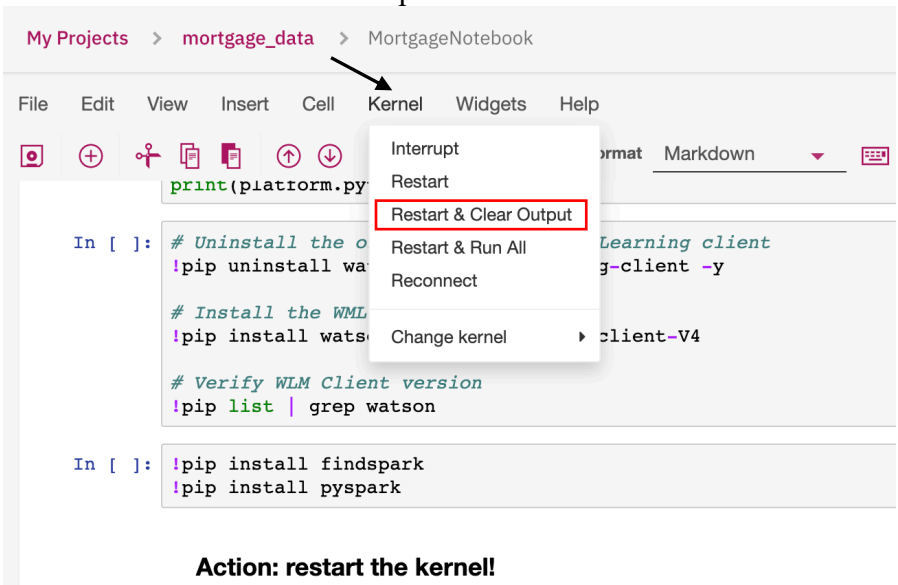
12.4. Review and run notebook

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.



Action: restart the kernel!

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.

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 .

12.5. Test the model

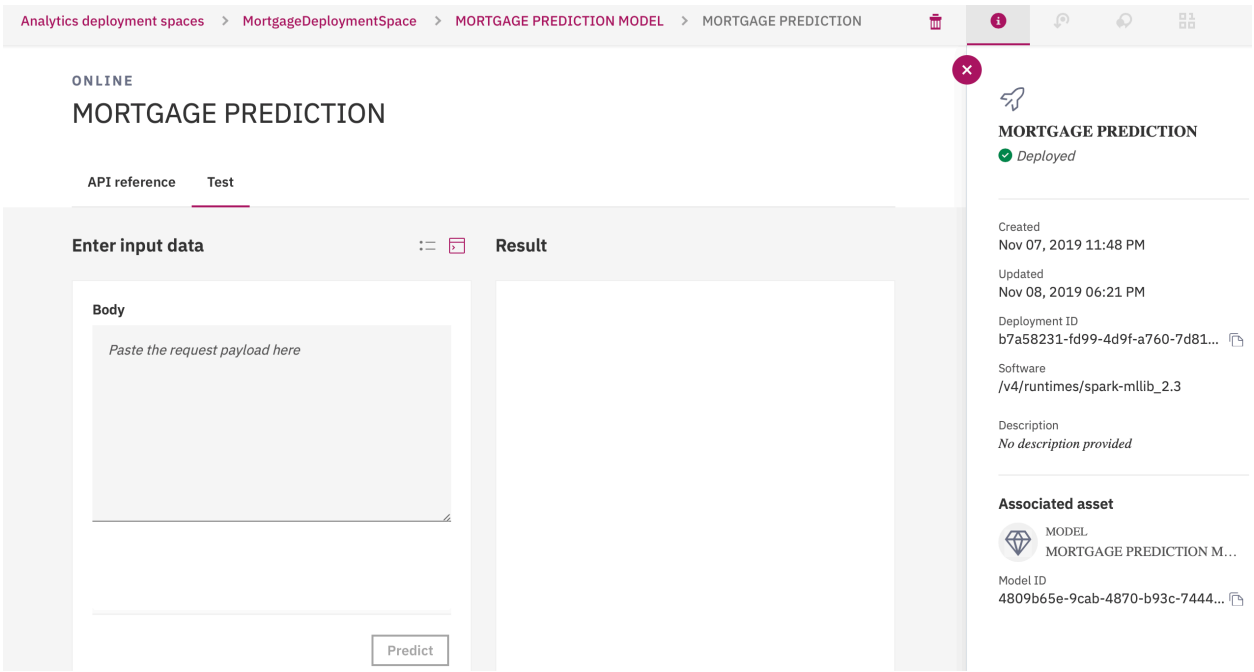
Go to: Analyze > Analytics deployment to access deployed model

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



```
{
  "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
        ]
      ]
    }
  ]
}
```

Copy this sample data and paste it on the **Enter input data** box.

Click on **Predict**

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