



**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 in to 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 .....	10
7.	Implement Business Glossary.....	12
7.1.	Download Business Glossaries.....	12
7.2.	Import Categories .....	13
7.2.	Import Terms .....	14
7.3.	Create a policy .....	15
7.4.	Create a rule .....	15
7.5.	Add rule to metadata.....	16
8.	Access data as a Data Scientist .....	18
8.1.	Create analytic project.....	18
8.2.	Assets from Glossary .....	18
8.3.	Check Asset Details .....	20
8.4.	Enterprise search .....	20
10.	Navigate to data catalog.....	24
11.	Data Virtualization .....	26
11.1.	Adding a new data source for Db2 .....	26
11.5.	Add virtual table to catalog .....	28
11.6.	Publish virtualized table.....	28
11.7.	Access information for virtual table .....	29
11.8.	Deliver Dataset .....	29
12.	Build Model .....	32
12.1.	Navigate to analytics project .....	32

## Cloud Pak for Data (v.2.5 or higher) – Tutorial

12.2. Create deployment space .....	32
12.3. Create notebook .....	33
12.4. Review and run notebook.....	34
12.5. Test the model.....	35

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 fictitious 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

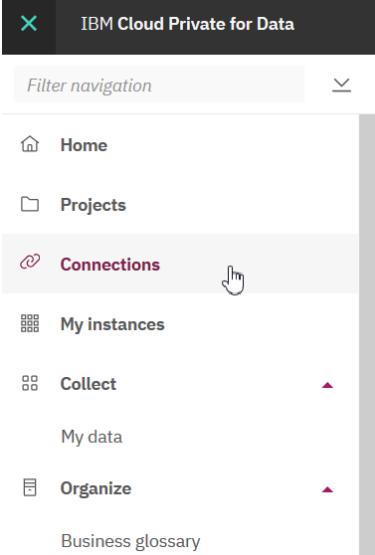
You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data web console. Starting from here all instruction need to execute on Cloud Pak for Data web console only. You need to login as admin who has administrator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

 Sign in      Sign up	<p>Sigh in to the Cloud Pak for Data web console as user ‘admin’ and password is ‘password’.</p>
Username <input type="text" value="admin"/> PASSWORD <input type="password" value="*****"/>	
<input type="button" value="Sign In"/>	

## 4. Create Connection

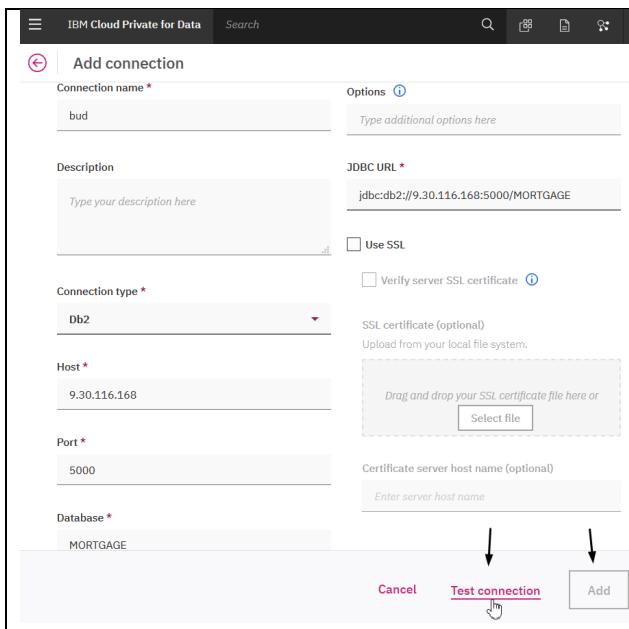
Create a connection to the data source for Db2 database.

### 4.2. Navigate to Connections



On the left pane choose **Connections**. Next, on the **Data Connections** window click on the  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/MORTGAG E'`
4. **Username** is ‘`db2inst1`’ and **Password** is ‘`password`’.

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



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

## 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

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

Quick scan job

**Connection \***  
bud

**Discovery root ⓘ**  
schema[MORTGAGE|DB2INST1] [Browse](#)

**Discovery options**

- Analyze columns
- Analyze data quality
- Assign terms
  - Use machine learning to assign terms
- Use data sampling

The maximum number of records included in the data set sample:  
1000

**Workspace \* ⓘ**  
Mortgage

[Cancel](#) [Discover](#)

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.

**Quick scan results**

New discovery job ▾ [View workspaces](#) [View automated discovery results](#)

Summary	Pending analysis	Action required	Reviewed		
<b>Status</b> <ul style="list-style-type: none"> <li><input checked="" type="radio"/> All jobs pending analysis</li> <li><input type="radio"/> Analyzing</li> <li><input type="radio"/> In queue for analysis</li> </ul>	<a href="#">Pause</a> <a href="#">View results</a>	1 item selected (select up to 15) <a href="#">Cancel</a>			
	<input checked="" type="checkbox"/> Job ID <input type="checkbox"/> Data assets <input type="checkbox"/> Connection <input type="checkbox"/> Started by <input type="checkbox"/> Processing time <input type="checkbox"/> Status <input type="checkbox"/> Status update				
	<input checked="" type="checkbox"/> qs_1571071613091 -    bud    admin    2 minutes 15 seconds    Analyzing -				

## 6. Add users

Create users with different roles.

<ul style="list-style-type: none"> <li>Home</li> <li>Projects</li> <li>Connections</li> <li>My instances</li> <li>Collect</li> <li>Organize</li> <li>Analyze</li> <li>Administer           <ul style="list-style-type: none"> <li>Manage platform</li> <li>Configure platform</li> <li>Gather diagnostics</li> <li>Manage users</li> </ul> </li> </ul>	<p>From <b>Administer</b> option on the left pane, choose <b>Manage users</b>.</p>
--	--

<p>IBM Cloud Private for Data</p> <p>Manage users</p> <p>Users Roles</p> <p>Add user Connect to an LDAP server</p>	<p>Switch tab to 'Users' and click on 'Add user'</p>
--	--

<p>Add user</p> <p>Name: dst1</p> <p>Username: dst1</p> <p>Email: dst1@abc.com</p> <p>User roles:</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Administrator</li> <li><input type="checkbox"/> Business Analyst</li> <li><input type="checkbox"/> Data Engineer</li> <li><input checked="" type="checkbox"/> Data Scientist</li> <li><input type="checkbox"/> Data Steward</li> </ul> <p>Cancel Add</p>	<p>Fill out Add User information for a data scientist</p> <ol style="list-style-type: none"> <li>1. Name as <b>dst1</b></li> <li>2. Username is <b>dst1</b></li> <li>3. Use a valid email address</li> <li>4. Set Password as <b>dst1</b></li> <li>5. Choose the user roles as Data Scientist</li> </ol> <p>Click on <b>Add</b> to confirm the add user</p>
--	---

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

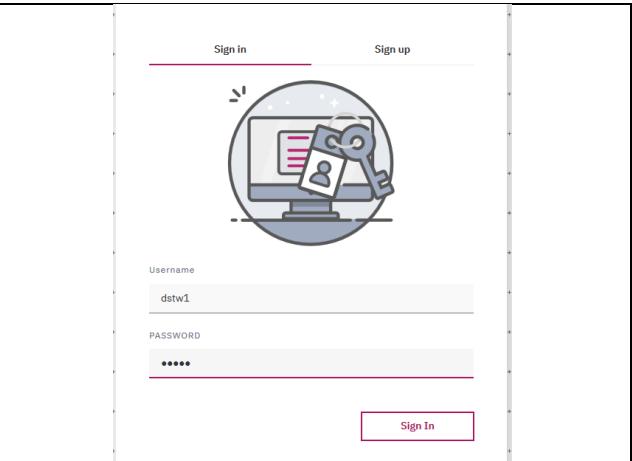
The screenshot shows the 'Manage users' page in the IBM Cloud Private for Data interface. A success message at the top left states 'Successfully updated user "dstw1"'. On the right side, a sidebar menu for the user 'admin' is open, showing options like 'Profile and settings', 'Getting Started', and 'About'. At the bottom of this sidebar is a 'Logout' button with a right-pointing arrow, which is highlighted with a mouse cursor. The main content area displays a table of users:

NAME ^	STATUS	USERNAME	DATE ADDED	USER ID	ROLES
admin	Approved	admin	--	999	Administrator + 4 ...
deng1	Approved	deng1	03/26/2019, 2:24 ...	1003	Data Engineer

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.



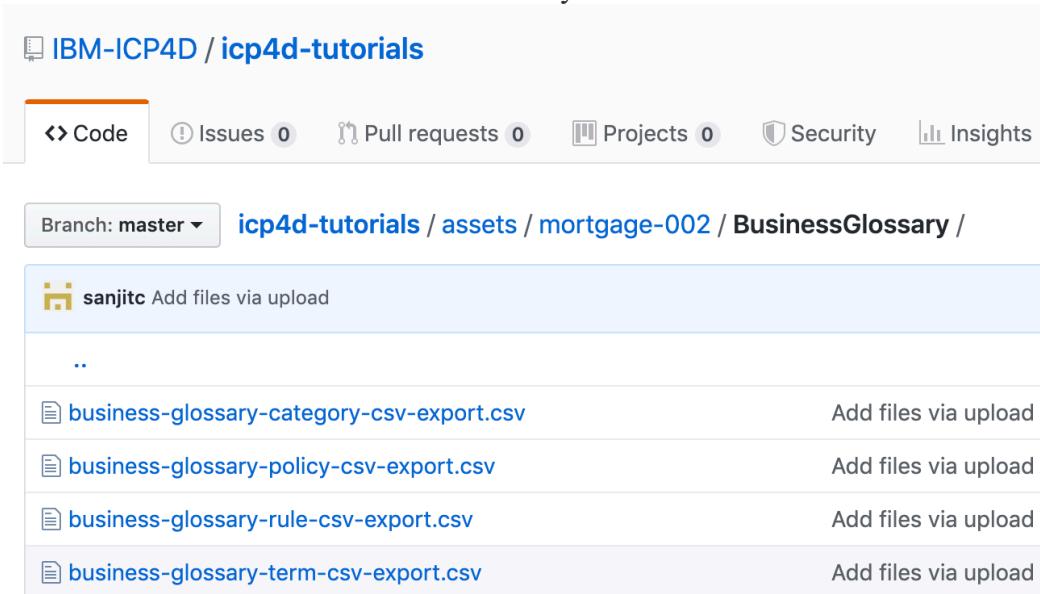
Sigh in to the Cloud Pak for Data web console as user ‘dstw1’ and password is ‘dstw1’ that you created earlier.

### 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.



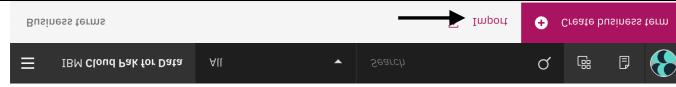
## 7.2. Import Categories

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

<p>Choose <b>Organize &gt; Data and AI governance &gt; Categories</b> from the left pane.</p> 	
	<p>Click on <b>Import</b> to import the CSV file contains category information that you downloaded from Git.</p>
 <p><b>Choose file</b></p> <p>Must be a CSV file.</p> <p><b>business-glossary-category-csv-export.csv</b></p> <p>The CSV file must conform to the template for importing governance artifacts. <a href="#">Learn more</a></p>	<p>Choose the CSV file location</p> <p><b>Click Next</b></p>
 <p><b>Next</b></p> <p><b>Choose file</b></p> <p><b>Set merging</b></p>	<p>Select merge option as <b>Replace all values</b></p> <p><b>Click Import</b></p>
 <p><b>Back</b></p> <p><b>Import</b></p>	

## 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.

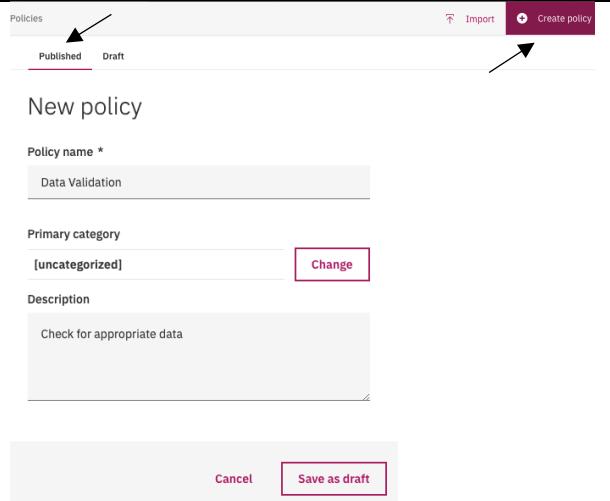
<p>The screenshot shows the first step of the import wizard: 'Choose file'. It has three steps: 'Choose file' (selected), 'Set merging', and 'Import'. A red box highlights the 'business-glossary-term-csv-export.csv' file in the file list. Below it, a note says: 'The CSV file must conform to the template for importing governance artifacts.' and a 'Learn more' link.</p>	<p>Choose the CSV file location</p> <p><b>Click Next</b></p>
<p>The screenshot shows the second step of the import wizard: 'Select merge option'. It has three options: 'Replace all values' (selected), 'Replace with defined values', and 'Replace empty values'. Below each option is a brief description. A red box highlights the 'Import' button at the bottom.</p>	<p>Select merge option as <b>Replace all values</b></p> <p><b>Click Import</b></p>

### 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**

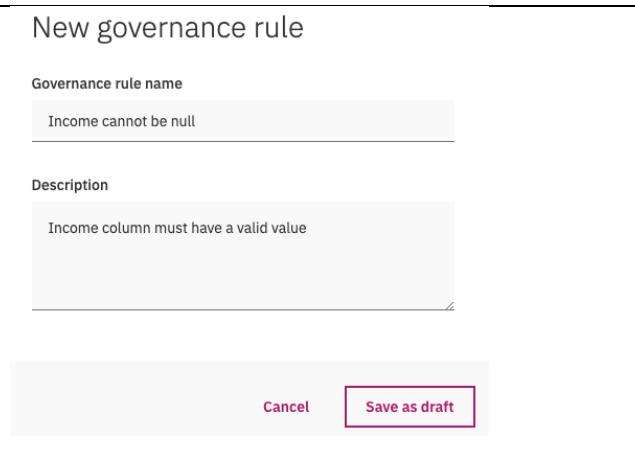
	<p>On the <b>New policy</b> window create a policy with following information and click on <b>Save as draft</b>:</p> <p><b>Name:</b> Data Validation  <b>Description:</b> Check for appropriate data</p> <p>It will take few minutes to appear under list of available policies.</p>
---	--

### 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**

	<p>On the <b>New governance rule</b> window create a rule with following information and click on <b>Save as draft</b>:</p> <p><b>Name:</b> Income cannot be null  <b>Referencing policies:</b> Data Validation  <b>Short Description:</b> Income column must have a valid value</p> <p>It will take few minutes to appear under list of available rules.</p>
---	---

	<p>Click on <b>Add policy</b> under <b>Parent policies</b> to assign the rule to it.</p>
---	--

## 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

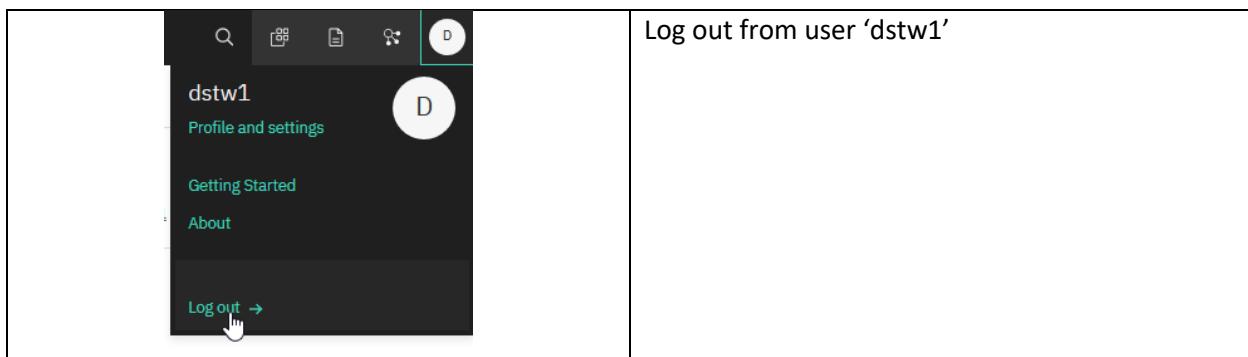
Select INCOME column

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

Scroll down to **Implement Rules** section

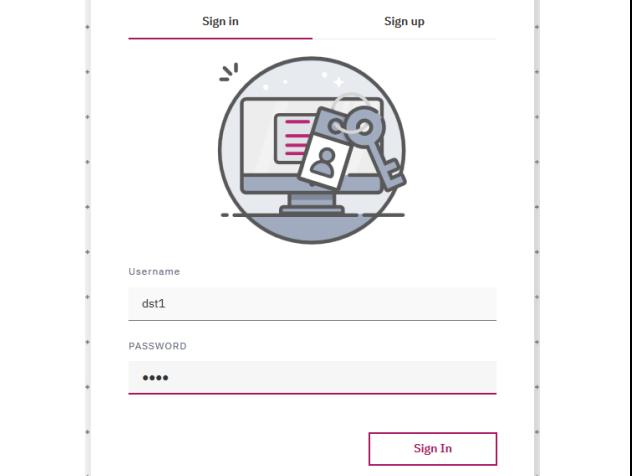
Search and select the rule **Income cannot be null** that you created earlier.

Click on **Save**

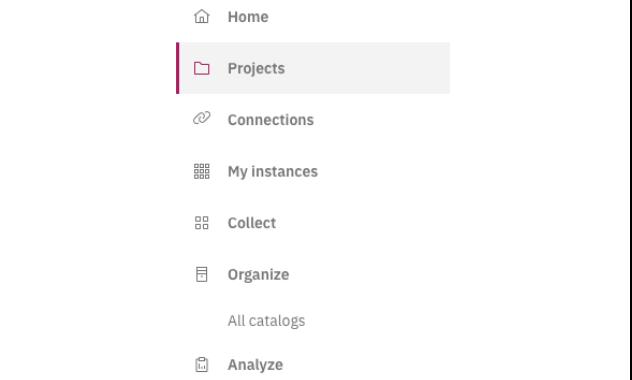


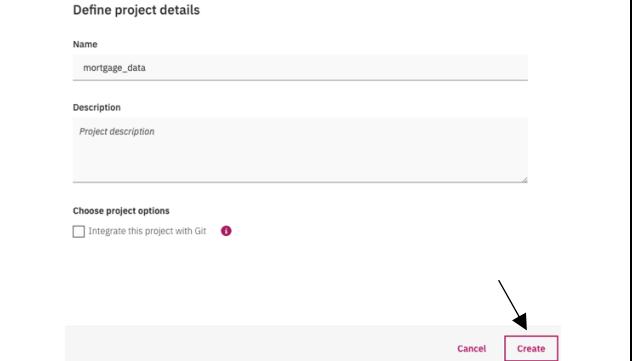
## 8. Access data as a Data Scientist

Explore the data require for build a model

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

### 8.1. Create analytic project

	<p>Create a new analytical project by ‘Projects’ from right pane.</p> <p>Click on the  icon</p> <p>Select <b>Create an empty project</b></p>
--	--

	<p>Provide a project name and click <b>Create</b></p>
---	---

### 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**.

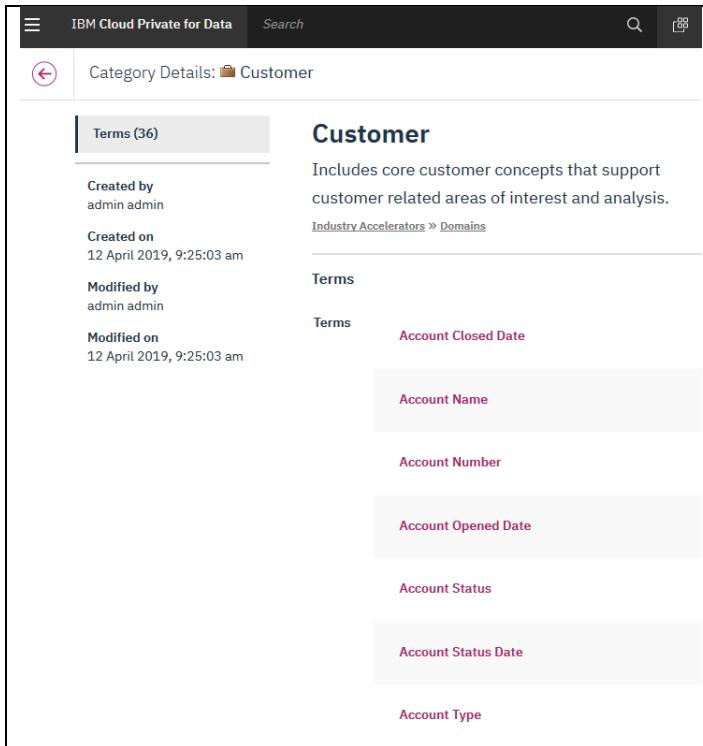
You should have all mortgage related information as follows. Click on each **ASSET NAME, TERMS** for additional information. The **TERM DESCRIPTION** provides a basic information about each term.

Category	ASSET NAME	CATEGORY DESCRIPTION	TERMS	TERM DESCRIPTION
	Address Information	Location related glossary for a JK insurance customer	<a href="#">Customer Zipcode</a> <a href="#">Continuity Of Address Segment</a> <a href="#">Address part 1</a> <a href="#">Customer City</a> <a href="#">Address part 2</a> <a href="#">Customer Street Suffix</a> <a href="#">Customer Street Name</a> <a href="#">Customer State</a> <a href="#">Customer House Label</a> <a href="#">Country Of Residence</a>	Current zip code for customer's address Customer City Current suffix for street for customer address Current street name for customer's address Current state of residence for a customer House number with optional suffix
	Crown Jewels	All data that is sensitive customer info per regulatory obligations	<a href="#">Sensitive Personal Data</a>	Any data deemed to be sensitive personal info for a customer
	Insurance Customer Details	Category for individual insurance customers	<a href="#">Gender</a> <a href="#">Market Segment</a> <a href="#">Summary</a>	Customer's gender, if known Customer Market Segment Summary information about a JKLV insured customer

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 screenshot shows the 'Customer' asset details page. At the top left is the navigation bar with 'IBM Cloud Private for Data' and a search bar. Below the navigation is a breadcrumb trail: 'Category Details: Customer'. A sidebar on the left lists 'Terms (36)' and various metadata: Created by (admin admin), Created on (12 April 2019, 9:25:03 am), Modified by (admin admin), and Modified on (12 April 2019, 9:25:03 am). The main content area is titled 'Customer' with a description: 'Includes core customer concepts that support customer related areas of interest and analysis.' It includes a link to 'Industry Accelerators > Domains'. Below this, there are several 'Terms' listed: Account Closed Date, Account Name, Account Number, Account Opened Date, Account Status, Account Status Date, and Account Type. Each term is preceded by a small red icon.

The asset **Customer** shows different terms associated with it.

Check each **Terms** for additional information.

### 8.4. Enterprise search



The screenshot shows the enterprise search interface. At the top left is the navigation bar with 'IBM Cloud Private for Data'. The search bar contains the word 'mortgage' with a magnifying glass icon highlighted by a red box. To the right of the search bar are three small icons: a magnifying glass, a gear, and a document.

Click on the enterprise search

Search for 'mortgage' and hit enter

**MORTGAGE\_PROPERTY**

**MORTGAGE\_PROPERTY**

Database Table  
idbc:db2://9.30.116.168:50000 /MORTGAGE  
db2 >> DB2INST1

★★★★★ 0 Ratings None Quality score

Description

Select your rating:

New Comment:  
Write a comment

All Comments (0)

Submit

**Relationships**

```

graph TD
    LOCATION[LOCATION Database Column] -- Context --> MORTGAGE_PROPERTY[MORTGAGE_PROPERTY Database Table]
    MORTGAGE_PROPERTY -- Context --> SALE_PRICE[SALE PRICE Database Column]
    MORTGAGE_PROPERTY -- Context --> ID[ID Database Column]
    MORTGAGE_PROPERTY -- Context --> Context
    MORTGAGE_PROPERTY -- Context --> Context
    MORTGAGE_PROPERTY -- Context --> Context
  
```

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.

**39 Search Results**

**Filter**

Search Result	Relevancy	Type
mortgage_property	67% RELEVANCY	TABLE
mortgage_default	67% RELEVANCY	TABLE
mortgage_customer	67% RELEVANCY	TABLE
mortgage customer	67% RELEVANCY	TERM
mortgage_join	65% RELEVANCY	TABLE
mortgage_default	64% RELEVANCY	COLUMN

Go back to the enterprise **Search Result**

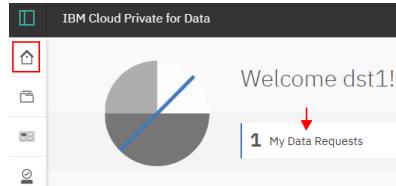
The enterprise search will return all objects that mentioned word mortgage but as a data scientist you don't have access to any of those objects.

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.

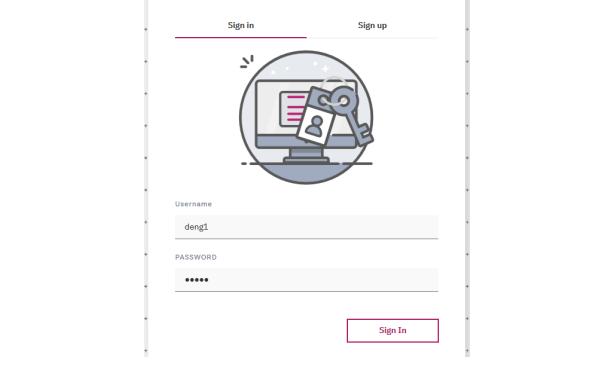
At this point you need to wait for data engineer to address the data request.

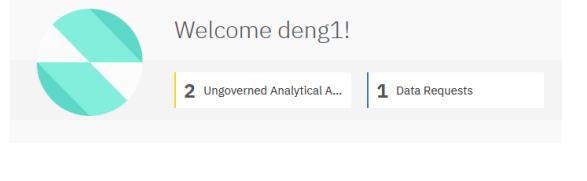
You can go to the home page by clicking on icon from left pane and check the status of the data request.



Sign out from user **dst1**

## 9. Review data request

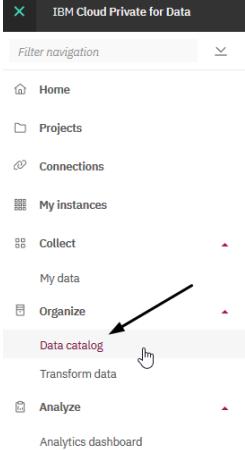
	<p>Sigh in to the Cloud Pak for Data web console as user ‘deng1’ and password is ‘deng1’ that you created earlier.</p>
---	--

	<p>After sing in Click on <a href="#" style="border: 1px solid red; padding: 2px;">Go to your home page</a></p> <p>Check the <b>Data Request</b> tab on the home page.</p>
---	--

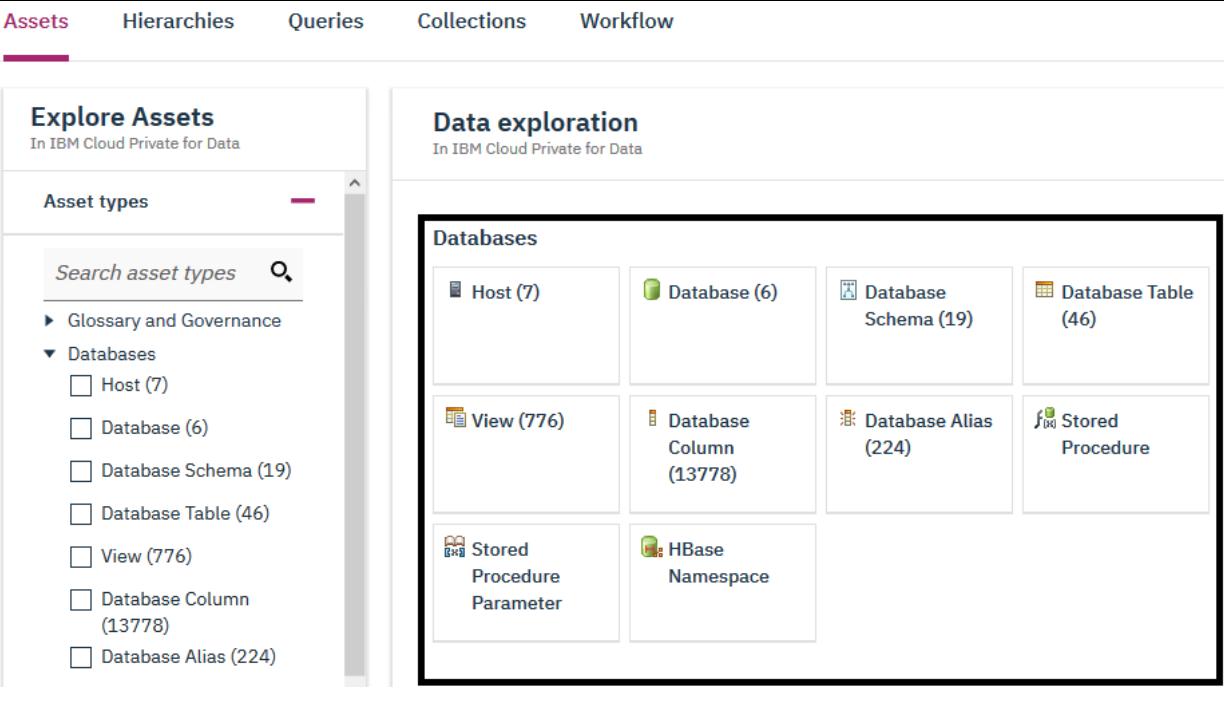
<p>Click on the new data request that submitted by data scientist earlier for review. After reviewing the request click on Action in top right corner and select assign to me.</p> <div style="border: 1px solid #ccc; padding: 10px; margin-top: 10px;"> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>ID</th> <th>Name</th> <th>Status</th> <th>Requested by</th> <th>Assigned to</th> <th>Priority</th> <th>Last updated</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Mortgage_Data_Access</td> <td>Claimed</td> <td>dst1</td> <td>deng1</td> <td>Medium</td> <td>3 Jun 2019, 8:15 PM</td> </tr> <tr> <td>2</td> <td>Mortgage_Data_Access_Request</td> <td>Claimed</td> <td>dst1</td> <td>deng1</td> <td>Medium</td> <td>3 Jun 2019, 8:41 PM</td> </tr> <tr> <td>3</td> <td>Mortgage_Data_request1</td> <td>Claimed</td> <td>dst1</td> <td>deng1</td> <td>High</td> <td>3 Jun 2019, 8:39 PM</td> </tr> <tr> <td>4</td> <td>CustData</td> <td>New</td> <td>admin</td> <td>Unassigned</td> <td>High</td> <td>4 Jun 2019, 9:03 AM</td> </tr> </tbody> </table> <div style="text-align: center; margin-top: 10px;"> <span style="border: 1px solid #ccc; padding: 2px;">Search</span> <span style="border: 1px solid #ccc; padding: 2px;">New</span> <span style="border: 1px solid #ccc; padding: 2px;">Import</span> <span style="border: 1px solid #ccc; padding: 2px;">Export</span> </div> <div style="margin-top: 10px;"> <p>Action ▾</p> <div style="background-color: #f0f0f0; padding: 5px; border-radius: 5px; width: fit-content;"> <a href="#">Transform data</a>  <a href="#" style="color: #0070C0; font-weight: bold;">Assign to me</a> </div> </div> </div>							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
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.

	<p>Next go back to <b>Organize</b> option on the left pane and choose <b>Data catalog</b>.</p>
---	--

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\_DEFAULT’ 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

<input type="checkbox"/>	<a href="#"><b>/MORTGAGE</b></a> ↳ db2 » SYSTOOLS	Modified by InformationServerSystemUser on Jun 3, 2019, 6:52 PM
<input type="checkbox"/>	<a href="#"><b>HMON_COLLECTION</b></a>	Modified by InformationServerSystemUser on Jun 4, 2019, 11:30 AM
<input type="checkbox"/>	<a href="#"><b>MONGO_MORTGAGE_DEFAULT</b></a>	Created by admin on Jun 3, 2019, 6:34 PM
<input type="checkbox"/>	<a href="#"><b>MONGO_MORTGAGE_PROPERTY</b></a>	Modified by InformationServerSystemUser on Jun 3, 2019, 6:34 PM
<input type="checkbox"/>	<a href="#"><b>MORTGAGE_CUSTOMER</b></a>	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.

<ul style="list-style-type: none"> <li>• Click <b>Collect &gt; Virtualized data &gt; Menu &gt; Virtualize</b></li>   <li>• Select tables <b>MORTGAGE_CUSTOMER</b>, <b>MORTGAGE_PROPERTY</b> and <b>MORTGAGE_DEFAULT</b> from <b>MORTGAGE</b> database, then click <b>Add to cart</b></li>   <li>• Click <b>View cart</b></li>   <li>• Click <b>Next</b></li> </ul>	<p>Menu   Virtualize</p> <p>Browse for: Tables Files View cart (0)</p> <p>Filters Available tables 4 tables</p> <p>Databases Find tables by name... <input type="button" value="Add to cart"/></p> <table border="1"> <thead> <tr> <th>Table</th> <th>Schemas</th> <th>Database</th> </tr> </thead> <tbody> <tr> <td>MORTGAGE_JOIN</td> <td>DB2INST1</td> <td>MORTGAGE</td> </tr> <tr> <td>MORTGAGE_CUST...</td> <td>DB2INST1</td> <td>MORTGAGE</td> </tr> <tr> <td>MORTGAGE_DEFALT</td> <td>DB2INST1</td> <td>MORTGAGE</td> </tr> <tr> <td>MORTGAGE_PROP...</td> <td>DB2INST1</td> <td>MORTGAGE</td> </tr> </tbody> </table>	Table	Schemas	Database	MORTGAGE_JOIN	DB2INST1	MORTGAGE	MORTGAGE_CUST...	DB2INST1	MORTGAGE	MORTGAGE_DEFALT	DB2INST1	MORTGAGE	MORTGAGE_PROP...	DB2INST1	MORTGAGE
Table	Schemas	Database														
MORTGAGE_JOIN	DB2INST1	MORTGAGE														
MORTGAGE_CUST...	DB2INST1	MORTGAGE														
MORTGAGE_DEFALT	DB2INST1	MORTGAGE														
MORTGAGE_PROP...	DB2INST1	MORTGAGE														

<ul style="list-style-type: none"> <li>• Select <b>Neither</b></li> <li>• Uncheck the box for <b>Submit to catalog</b></li> <li>• Click <b>Virtualize</b> to complete the process</li> </ul>	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Table</th> <th>Schema</th> <th>Source schema</th> <th>Host/Database</th> <th>Grouped tables</th> </tr> </thead> <tbody> <tr> <td>MORTGAGE_CUSTOMER</td> <td>USER999</td> <td>X ▾</td> <td>DB2INST1</td> <td>169.46.33.180:MORTGAGE</td> </tr> <tr> <td>MORTGAGE_DEFAULT</td> <td>USER999</td> <td>X ▾</td> <td>DB2INST1</td> <td>169.46.33.180:MORTGAGE</td> </tr> <tr> <td>MORTGAGE_PROPERTY</td> <td>USER999</td> <td>X ▾</td> <td>DB2INST1</td> <td>169.46.33.180:MORTGAGE</td> </tr> </tbody> </table>	Table	Schema	Source schema	Host/Database	Grouped tables	MORTGAGE_CUSTOMER	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE	MORTGAGE_DEFAULT	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE	MORTGAGE_PROPERTY	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE
Table	Schema	Source schema	Host/Database	Grouped tables																	
MORTGAGE_CUSTOMER	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE																	
MORTGAGE_DEFAULT	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE																	
MORTGAGE_PROPERTY	USER999	X ▾	DB2INST1	169.46.33.180:MORTGAGE																	

#### 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.

- |  |
|--|
| <ul style="list-style-type: none"> <li>• Click <b>Collect &gt; Virtualized data &gt; Menu &gt; SQL editor</b> to access the editor.</li> <li>• Copy the following SQL statement and paste it on the editor</li> <li>• Click on <b>Run all</b></li> </ul> |
|--|

```
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;
```

<pre>* Untitled - 1</pre> <pre>1 2 3 CREATE VIEW MORTGAGE_JOIN_VIEW 4 AS 5 SELECT A.ID, INCOME, APPLIED_ONLINE, RESIDENCE, YRS_CURRENT_ 6       YRS_CURRENT_EMP, NO_OF_CARDS, CARD_DEBT, CURRENT_LOAN 7       LOAN_AMOUNT, SALE_PRICE, LOCATION, MORTGAGE_DEFAULT 8 FROM   MORTGAGE_CUSTOMER A, 9        MORTGAGE_PROPERTY B, 10       MORTGAGE_DEFAULT C 11 WHERE  A.ID = B.ID 12 AND    A.ID = C.ID; 13</pre>
---

- |  |
|--|
| <ul style="list-style-type: none"> <li>• Click <b>Collect &gt; Virtualized data &gt; Menu &gt; My virtualized data</b> to access the virtual table <b>MORTGAGE_JOIN_VIEW</b></li> <li>• Check the box associated with <b>MORTGAGE_JOIN_VIEW</b></li> <li>• Click on the table actions menu </li> </ul> |
|--|

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

Grant access to

All data virtualization users [?](#)    Specific users [?](#)

[Users](#)   [Roles](#)

Search [?](#) [Revoke](#) [?](#) [Grant access](#)

<input type="checkbox"/>	Name	Username	Role	User ID	Access level

### 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
- Choose **Submit to catalog** from table action
- Click on **Confirm**

Menu [?](#) | My virtualized data

Find [?](#)

Total tables: 23 [?](#) Access to some tables is restricted by policies. [?](#)

[Assign](#) [Join view](#) [Add tab](#)

<input type="checkbox"/>	Table	Schema	Created on
<input checked="" type="checkbox"/>	MORTGAGE_JOIN_VIEW	USER999	18 Oct 2019 20:17:28
<input type="checkbox"/>	V2	USER999	18 Oct 2019 20:09:27
<input type="checkbox"/>	V1	USER999	18 Oct 2019 20:08:40
<input type="checkbox"/>	MORTGAGE_PROPERTY	USER999	18 Oct 2019 20:08:40

### 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.

( Pending Publish to Catalog Requests

Search [?](#)

Name	Type	Project	Owner	Date Updated	Status
> USER999.MORTGAGE_JOIN_VIEW	view	-	admin	21 October 2019, 2:49PM	Pending
> USER999.Currency USER999.Country	table	-	admin	17 October 2019, 8:40AM	Pending

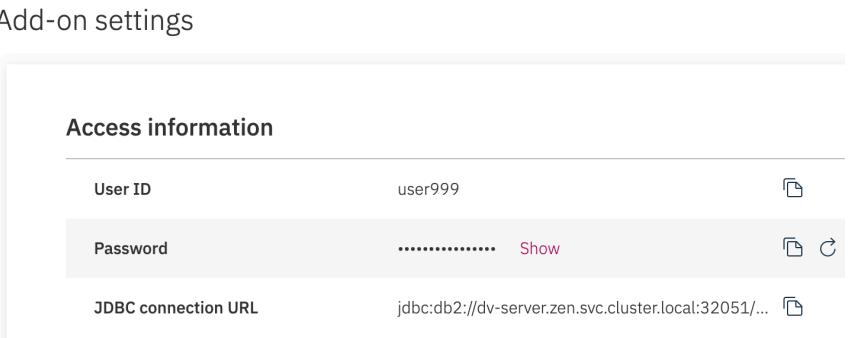
**Pending Publish to Catalog Requests**

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

	<ul style="list-style-type: none"> <li>• Click on  icon on left for virtual table <b>MORTGAGE_JOIN_VIEW</b> that you created</li> <li>• Click on <b>Approve</b></li> </ul>
--	---

### 11.7. Access information for virtual table

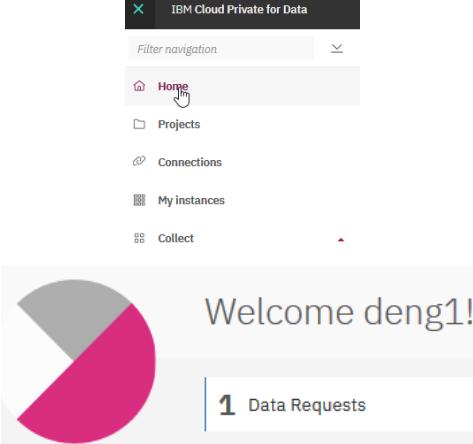
To access virtual table from external application, you need the JDBC connection information. Click on **Collect > Virtualized data > Menu > Add-on settings** to find out access information. You will use this information later in the building model section.



The screenshot shows the 'Add-on settings' page under the 'Menu' dropdown. It displays the following access information:

Access information	
User ID	user999
Password	..... <a href="#">Show</a>
JDBC connection URL	jdbc:db2://dv-server.zen.svc.cluster.local:32051/...

### 11.8. Deliver Dataset



The screenshot shows the home page of IBM Cloud Private for Data. The left sidebar includes navigation links for Home, Projects, Connections, My instances, and Collect. The main area displays a welcome message 'Welcome deng1!' and a 'Data Requests' section indicating 1 request.

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.

## Data requests + Add new data request

	Name	ID	Status	Last Updated
1	Mortgage_Data_Access	2	New	27 Mar 2019, 11:15 AM

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.

New data request

Overview    Columns    **Source**

### Source

Data source name  
mortgage\_join

DB2

Username  
db2inst1

Password  
.....

JDBC URL  
169.45.83.218

+ Add new dataset

	Remote data set name	Description	Schema	Table
1	mortgage_join		db2inst1	mortage_join

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

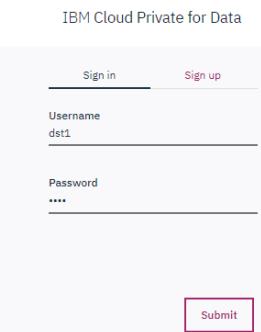
NAME	ID	STATUS	REQUESTED BY	ACCEPTED BY	LAST UPDATED	ACTIONS
mortgagedata1	7	Delivered	dst1	deng1	6 Aug 2018, 12:55 PM	
Mortgage_Data_Access	9	Accepted	dst1	deng1	15 Aug 2018, 1:17 AM	

The screenshot shows the IBM Cloud Private for Data dashboard. At the top right, there is a user profile dropdown menu with the text "Signed in as: deng1". Below this, there are links for "Getting Started", "Settings", and "Sign Out". A red box highlights the "Sign Out" link. To the left of the dashboard, there is a sidebar with icons for Home, Data Requests, and Settings. The main area displays a "Welcome deng1" message and a circular icon with a teal X.

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 accuracy. One can use lot more data and build a complex algorithm to get better accuracy.

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



### 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

<p>Connect to a deployment space</p> <p>New Existing</p> <p>Name MortgageDeploymentSpace</p> <p>Description (Optional) <i>Description of deployment space</i></p> <p style="text-align: right;">Cancel Associate</p>	<p>Name new deployment space as 'MortgageDeploymentSpace'</p> <p><b>Click on Associate</b></p>
--	--

### 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 type as Notebook
- The new notebook needs to create from URL
- Name the notebook as **MortgageNotebook**
- 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**

<p>My Projects &gt; mortgage_data &gt; Add Notebook</p> <p>New notebook</p> <p>Blank From file <b>From URL</b></p> <p>Name MortgageNotebook <small>24 characters remaining</small></p> <p>Description (optional) <i>Type your Description here</i> <small>500 characters remaining</small></p> <p>Select runtime Default Python 3.6 (1 vCPU and 2 GB RAM)</p> <p>Notebook URL <a href="https://github.com/IBM-ICP4D/icp4d-tutorials/blob/master/assets/mortgage-002/MortgageNotebook.V25">https://github.com/IBM-ICP4D/icp4d-tutorials/blob/master/assets/mortgage-002/MortgageNotebook.V25</a></p> <p style="text-align: right;">Cancel <b>Create Notebook</b></p>
---

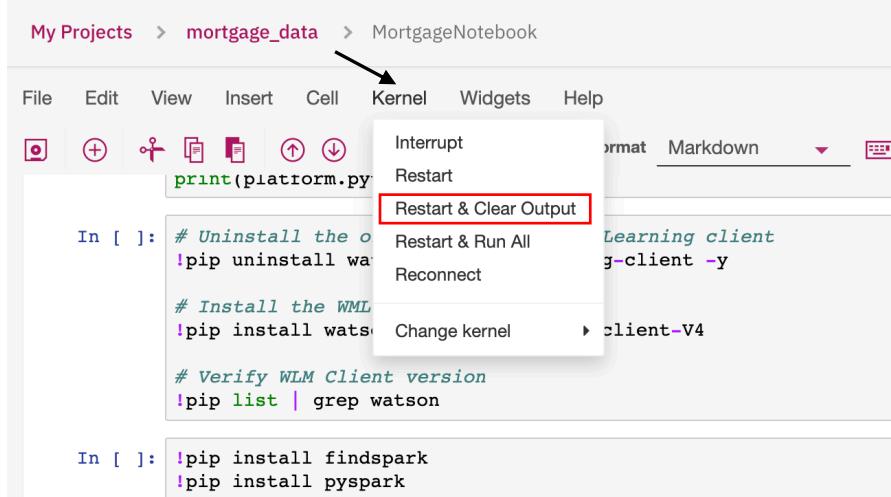
## 12.4. Review and run notebook

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 restart the Python kernel before move on next step.



My Projects > mortgage\_data > MortgageNotebook

Kernel

- Interrupt
- Restart & Clear Output
- Restart & Run All
- Reconnect
- Change kernel
- client-V4

```

In [ ]: # Uninstall the old Watson packages
!pip uninstall watson

# Install the WML Client
!pip install watsongears-client -y

# Verify WLM Client version
!pip list | grep watson

In [ ]: !pip install findspark
!pip install pyspark

```

**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 your database connection found under data virtualization
dsn_url = "jdbc:db2://dv-server.zen.svc.cluster.local:32051/biggsql" # 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 remaining cells, so that it generates and deploys the model.

Before exit, save the notebook .

## 12.5. Test the model

Go to: Analyze > Analytics deployment to access deployed model

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

Click on the **MORTGAGE PREDICTION MODEL**

Choose the **MORTGAGE PREDICTION** model

Click on **Test** tab

The screenshot shows the 'Analytics deployment spaces' interface. The path is: Analytics deployment spaces > MortgageDeploymentSpace > MORTGAGE PREDICTION MODEL > MORTGAGE PREDICTION. The 'Test' tab is selected. On the left, there's a 'Test' panel with tabs for 'API reference' and 'Test'. The 'Test' tab is active, showing a 'Enter input data' section with a 'Body' field containing placeholder text 'Paste the request payload here' and a 'Predict' button. To the right, there's a detailed view of the 'MORTGAGE PREDICTION' model. It shows the model is 'Deployed'. Below that, it lists 'Created' (Nov 07, 2019 11:48 PM), 'Updated' (Nov 08, 2019 06:21 PM), 'Deployment ID' (b7a58231-fd99-4d9f-a760-7d81...), 'Software' (/v4/runtimes/spark-mllib\_2.3), and 'Description' (No description provided). At the bottom, it shows an 'Associated asset' section with a 'MODEL' icon and the text 'MORTGAGE PREDICTION M...' and 'Model ID' (4809b65e-9cab-4870-b93c-7444...).

```
{
  "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.

The screenshot shows the Cloud Pak for Data interface with the following details:

- Header:** Analytics deployment spaces > MortgageDeploymentSpace > MORTGAGE PREDICTION MODEL > MORTGAGE PREDICTION
- Page Title:** ONLINE MORTGAGE PREDICTION
- Sub-Header:** API reference Test
- Left Panel (Enter input data):**
  - Body:** A code editor containing the JSON input data provided in the previous step.
  - Predict Button:** A pink button at the bottom of the code editor.
- Right Panel (Result):**
  - Result:** A code editor showing the predicted output JSON.
  - Associated asset:** Shows the model asset information.
- Model Asset Details:**
  - MORTGAGE PREDICTION** (Deployed)
  - Created:** Nov 07, 2019 11:48 PM
  - Updated:** Nov 08, 2019 06:32 PM
  - Deployment ID:** b7a58231-fd99-4d9f-a760-7d81...
  - Software:** /v4/runtimes/spark-mllib\_2.3
  - Description:** No description provided