Cloud Pak for Data

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 Instance
- Install Git on the machine that you will use for the tutorial.

2. Setting up database and sample data

2.1. 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.

2.2. From your home directory, clone the tutorial sample files:

```
git clone https://github.com/sanjitc/ICP4XTutorial.git
```

2.3. 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.

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

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

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

```
./load_samples.sh -t mortgage-001
```

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.

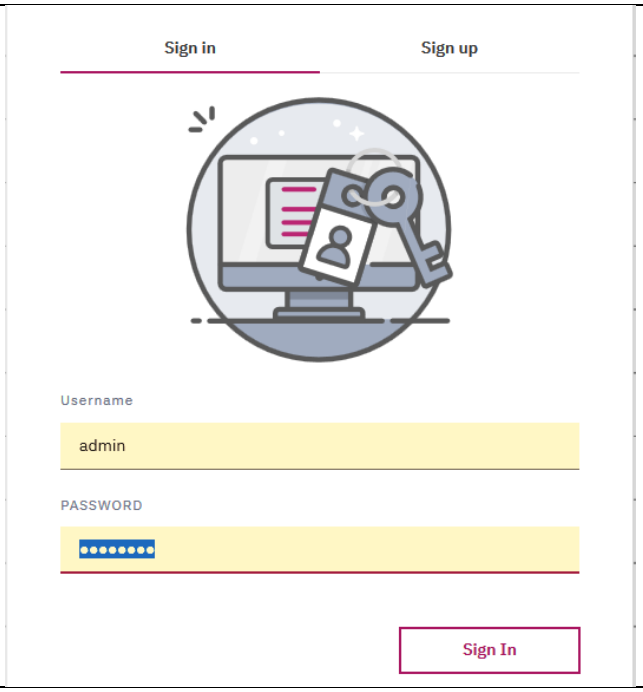
3.1. Access credential for Db2 database

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

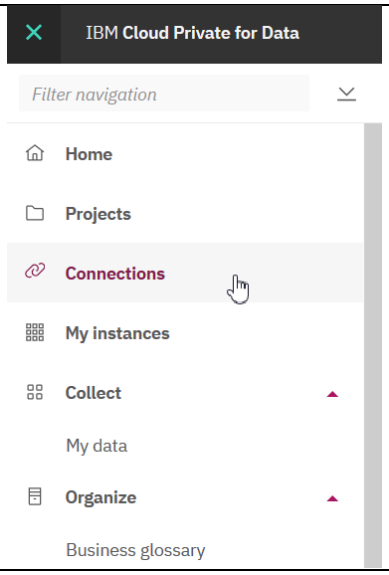
You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data web console. Starting from here all instruction need to execute on Cloud Pak for Data web console only. You need to login as admin who has administrator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

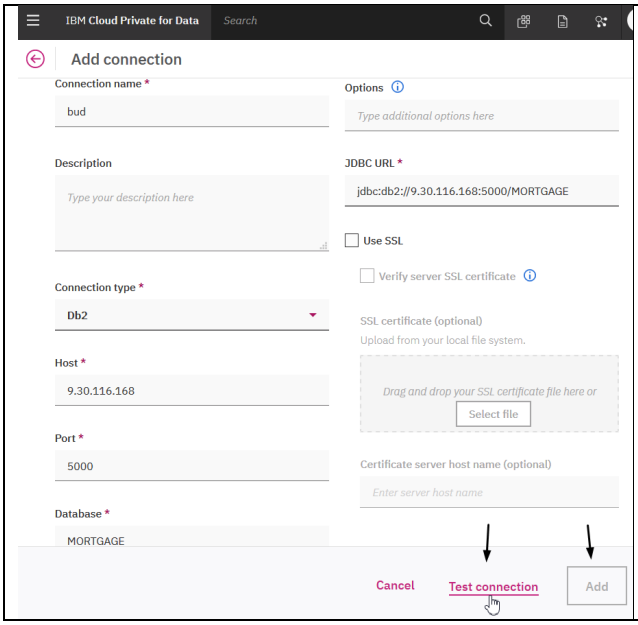
4.1. Navigate to Connections



On the left pane choose **Connections**.

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

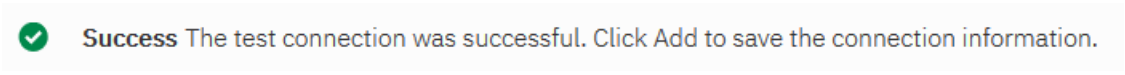
4.2. Add connection



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'.

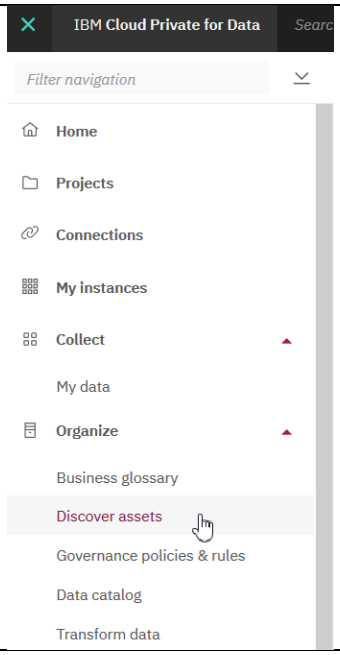
Next click on **Test Connection**, once it successful click on **Save Connection**.



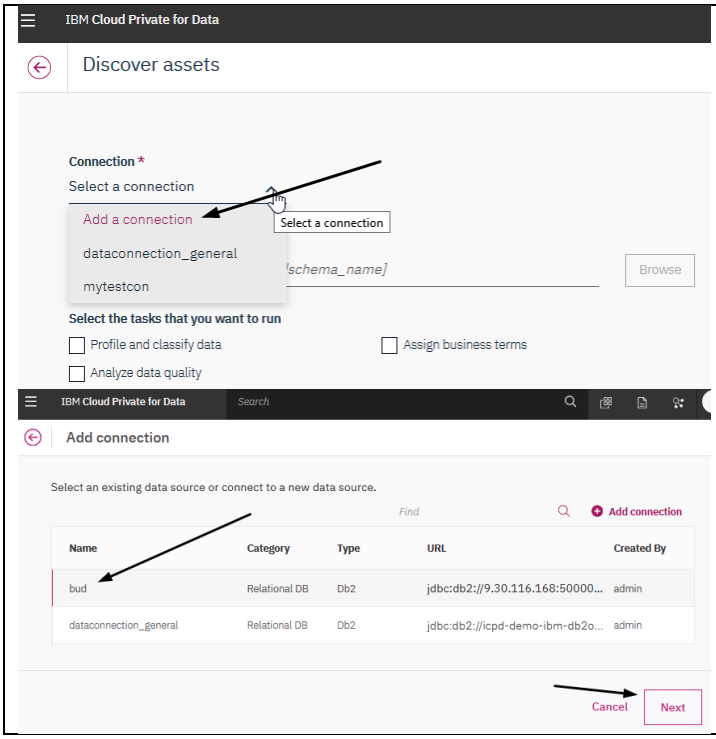
5. Discover Assets

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

5.1. Navigate to discover assets

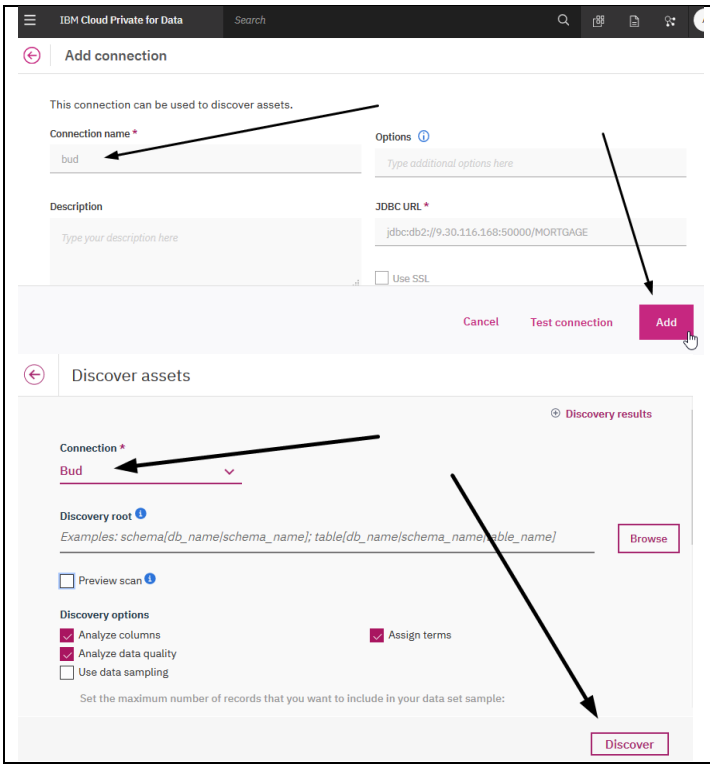


From **Organize** option on the left pane, choose **Discover assets**.



To discover assets

1. Add a connection
2. Choose the connection named ‘bud’ that you created previously, click Next



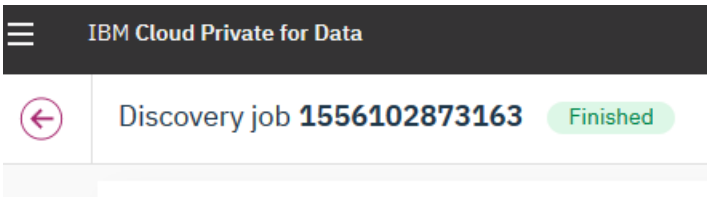
3. Add the bud connection

4. Choose the connection named ‘bud’ that you created previously.

5. Check ‘Analyze data quality’ under the ‘Select the tasks that you want to run’.

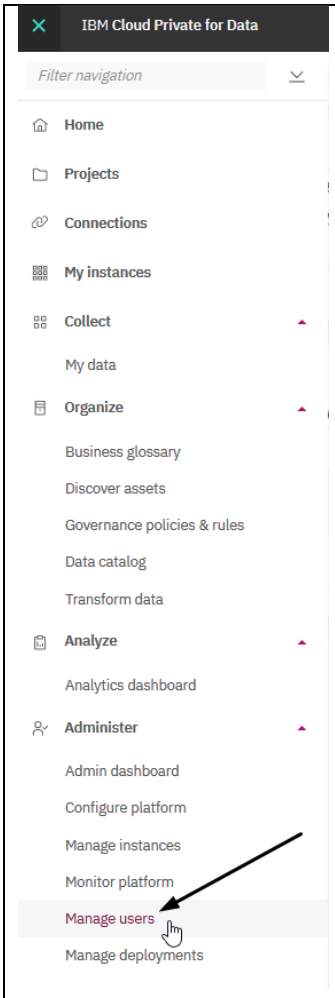
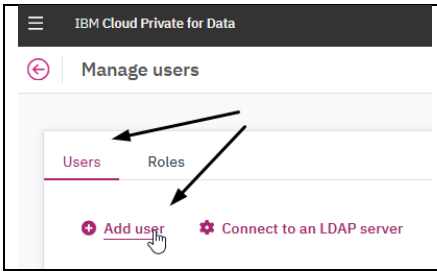
6. Click on **Discover**

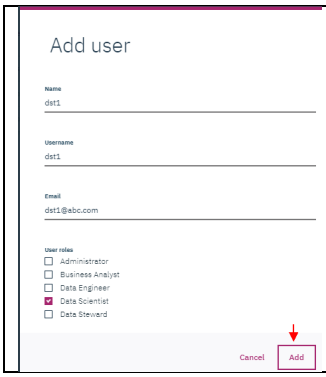
It may take few minutes to complete.



6. Add users

Create users with different roles.

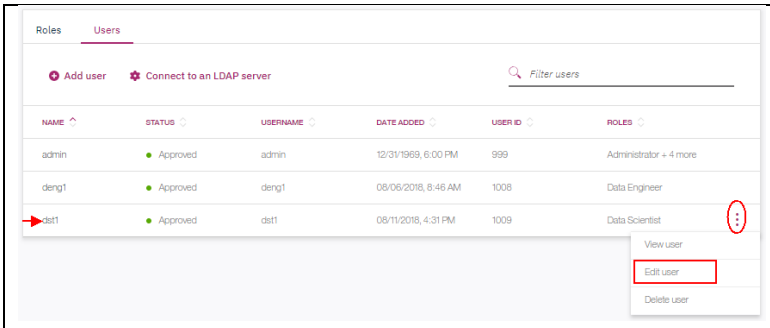
	From **Administer** option on the left pane, choose **Manage users**.
	Switch tab to ‘Users’ and click on ‘Add user’



Fill out Add User information for a data scientist

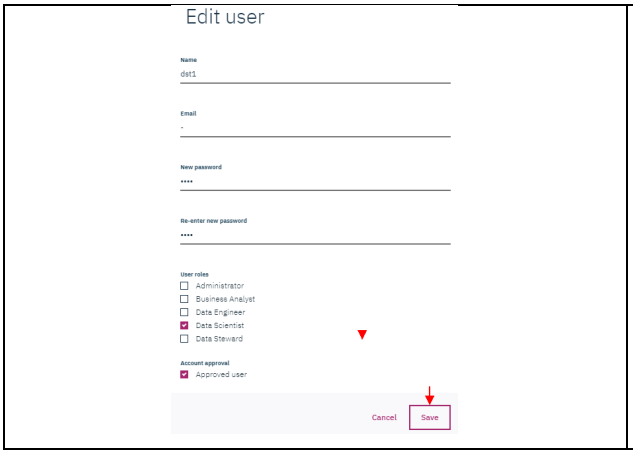
1. 'Name' as **dst1**
2. Username is **dst1**
3. Use a valid email address
4. Chose the user roles as Data Scientist

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



Before hand over user, change the password.

1. Access dst1 user setting by click on ⋮ icon
2. Choose 'Edit user'

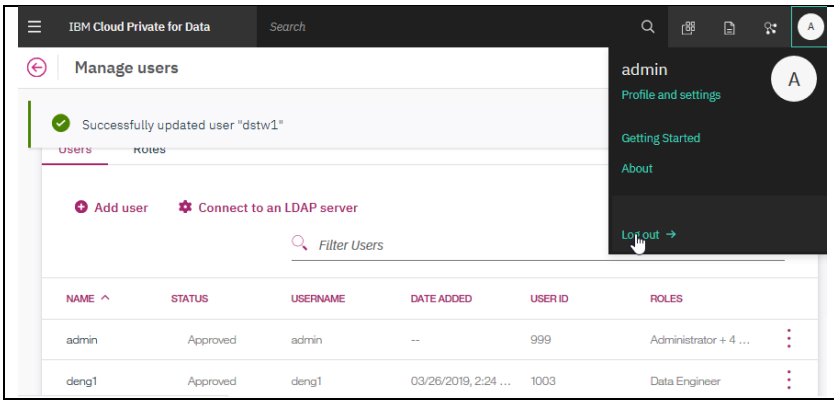


1. Type password as **dst1** in 'New password' and 're-enter new password' fields.

2. Click on Save

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

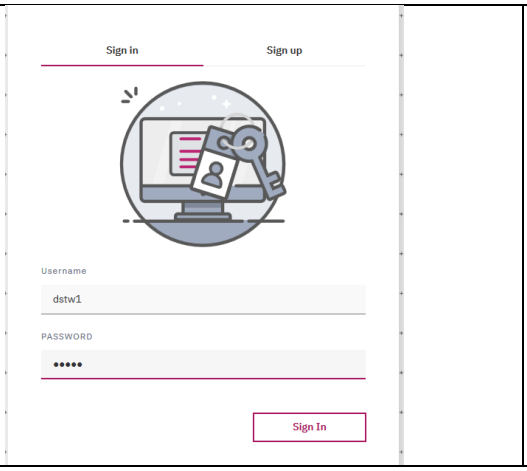
	User	Role	Password
•	deng1	Data Engineer	deng1
•	dstw1	Data Stewards	dstw1



Log out from user **admin**

7. Implement Policies and Rules

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

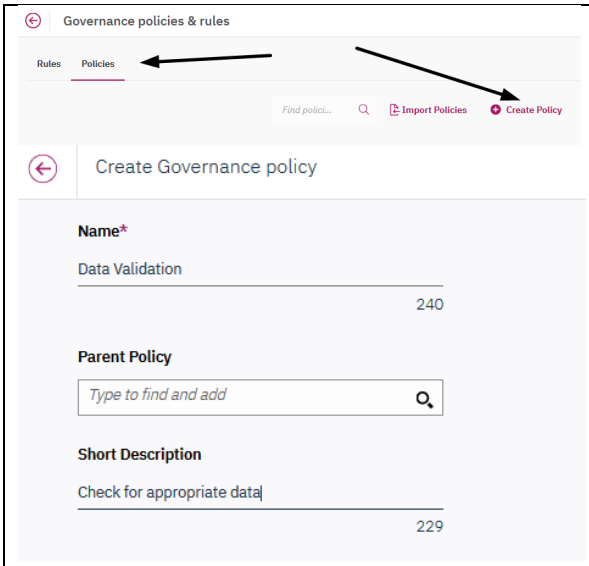


Sigh in to the Cloud Pak for Data web console as user ‘dstw1’ and password is ‘dstw1’ that you created earlier.

7.1. Create a policy

Choose **Organize** from the left pane, then select **Governance policies and rules**.

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



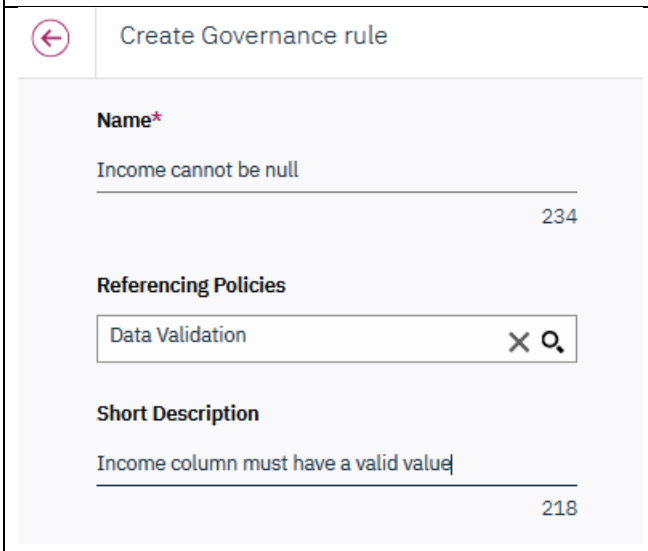
On the Create Information Governance Policy window create a polity with following information and click on **Save**:

Name: Data Validation
Short Description: Check for appropriate data

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

7.2. Create a rule

Select **Rules** tab and click on **Create Rule**

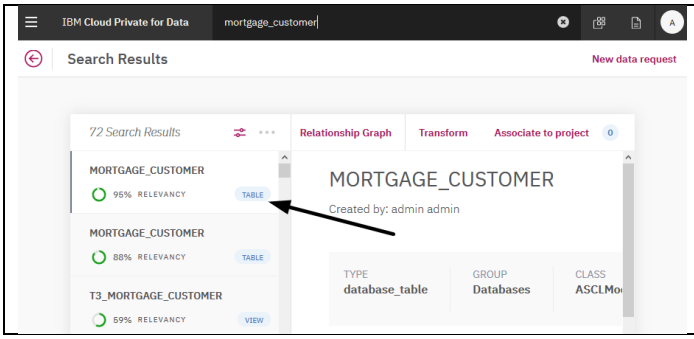


On the Create Information Governance Policy window create a rule with following information and click on **Save**:

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

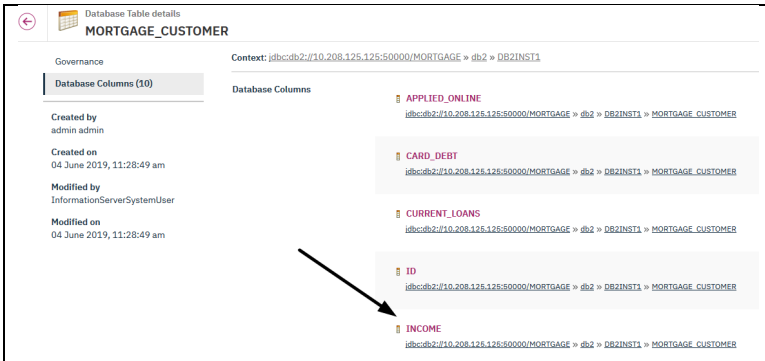
It will take few minutes to appear under list of available rules.

7.3. 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

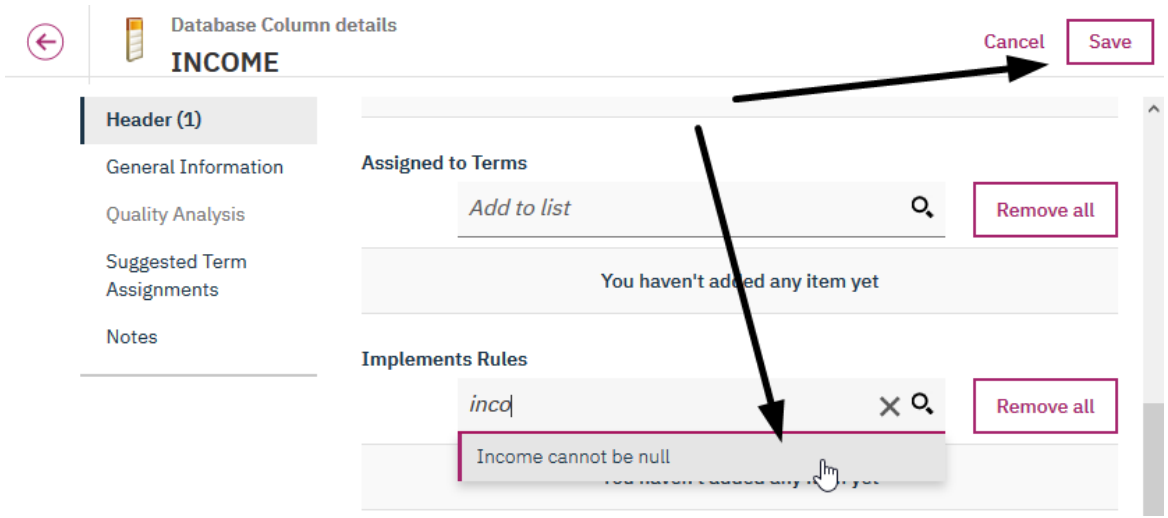
Select INCOME column

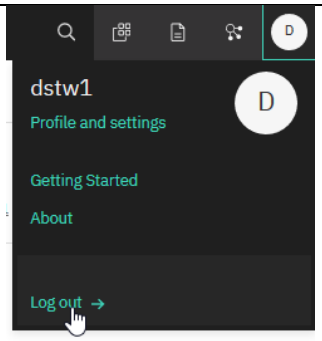
Next click on ⋮ icon (right top corner) and choose **Edit**

Scroll down to **Implement Rules** section

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

Click on **Save**





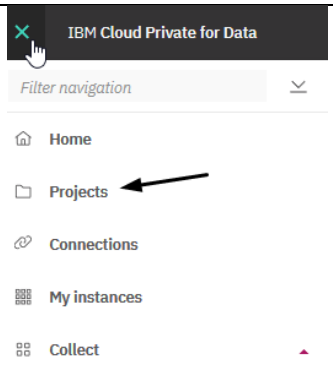
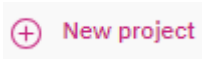
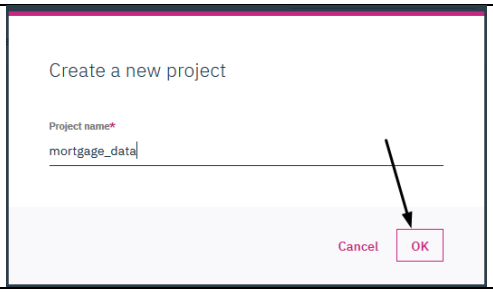
Log out from user ‘dstw1’

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

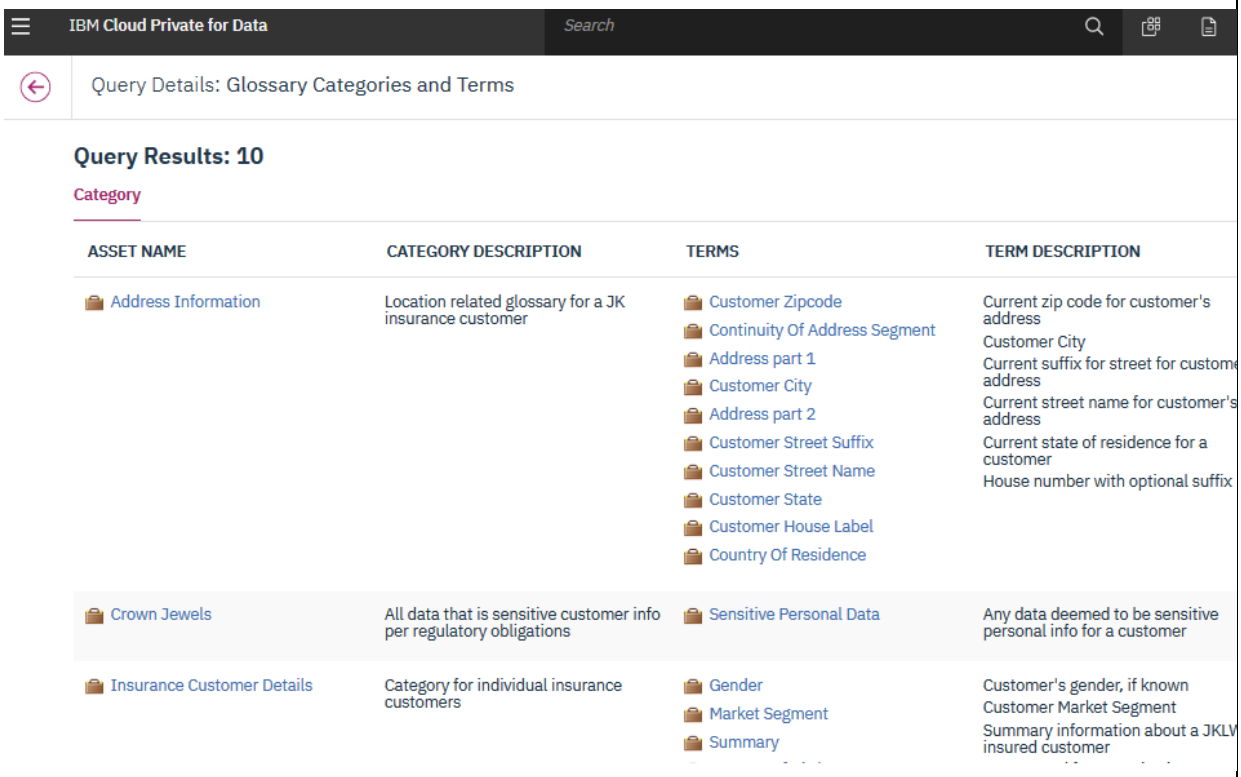
	Create a new analytical project by 'Projects' from right pane. Click on the  icon
	Provide a project name and click **OK** On the next 'Create project' window, click on **Create**

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, the select **Data Catalog** -> **Queries -> Glossary Categories and Terms**.

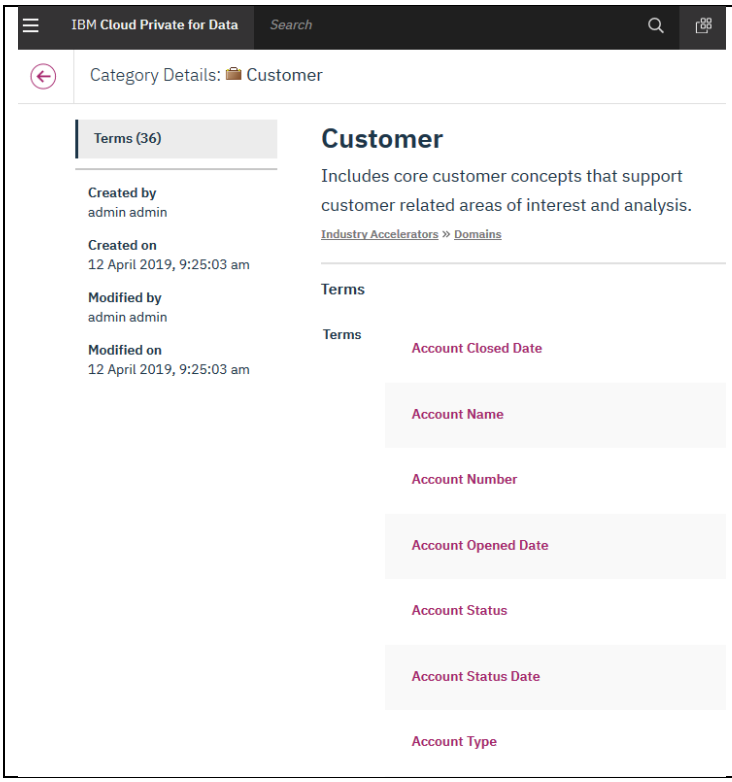
You should have all mortgage related information as follows. Click on each **ASSET NAME**, **TERMS** for additional information. The TERM DESCRIPTION provides a basic information about each term.



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.



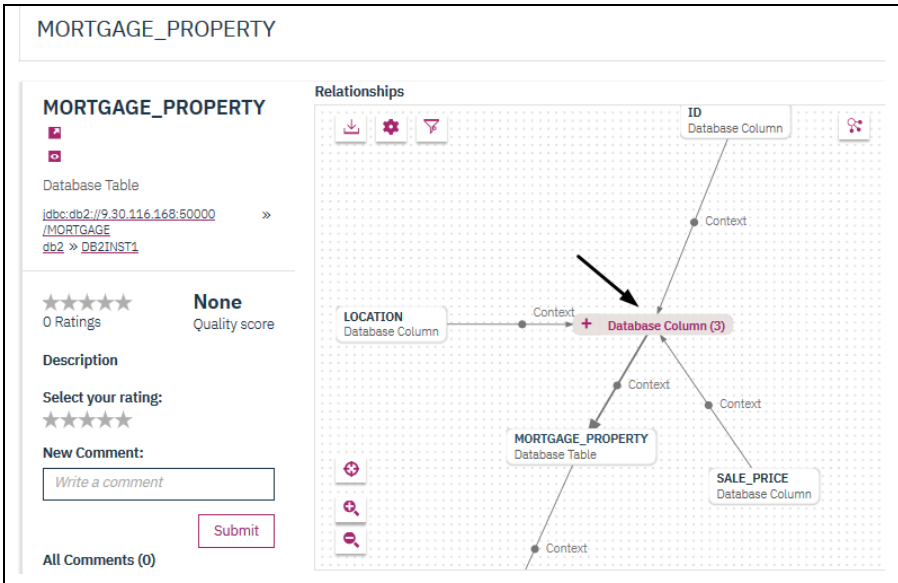
The asset **Customer** shows different terms associated with it.

Check each **Terms** for additional information.

8.4. Enterprise search

Click on the enterprise search

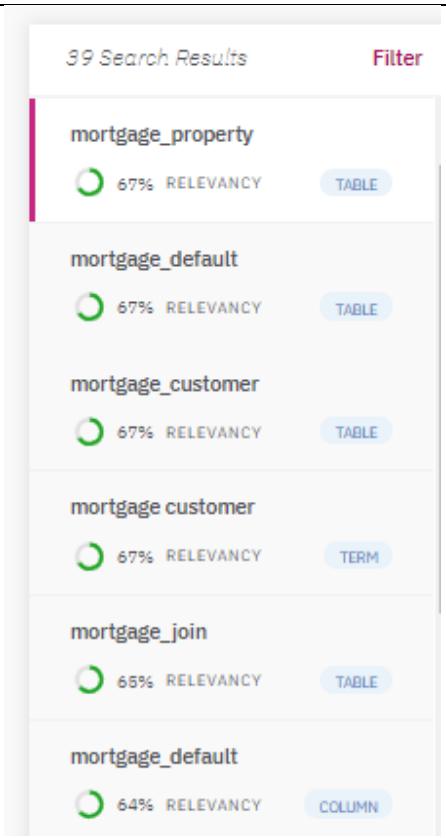
Search for 'mortgage' and hit enter



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.

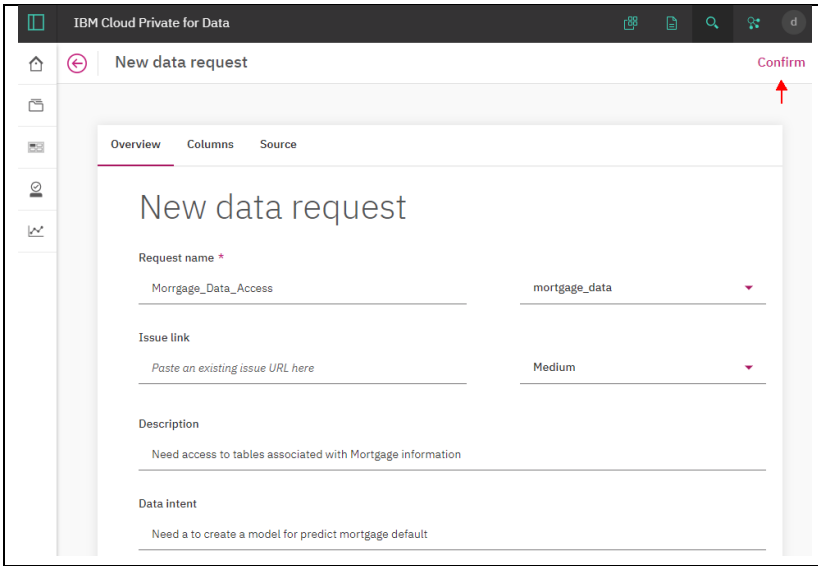
Same way you can view other mortgage related tables.



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.

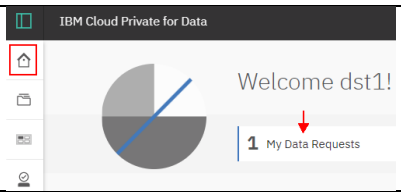
Click on the New Data Request on top right corner for request access to mortgage related datasets.

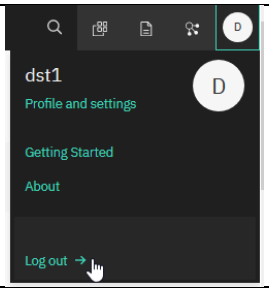


Fill up the **New Data Request** form with detail information as much possible, so a data engineer can provide accurate dataset. Click Confirm and then Submit request.

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.

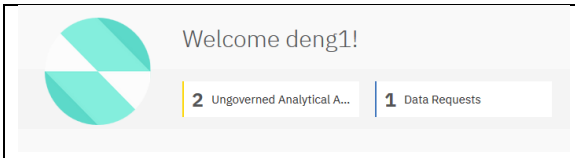




Sign out from user **dst1**

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.

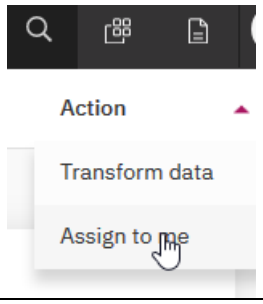


After sing in Click on Go to your home page

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

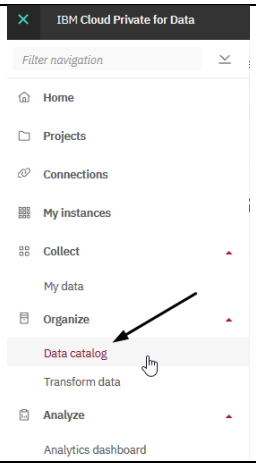
Click on the new data request that submitted by data scientist earlier for review.
After reviewing the request click on Action in top right corner and select assign to me.

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



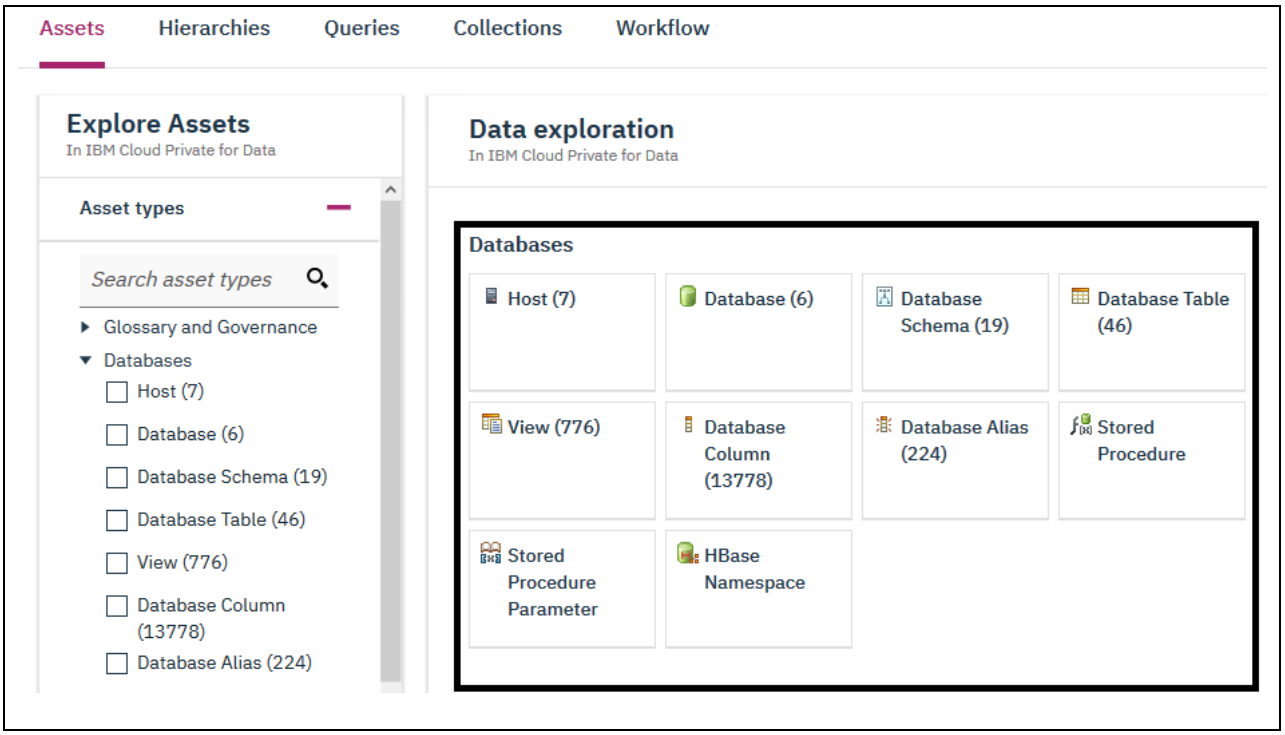
10. Navigate to data catalog

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

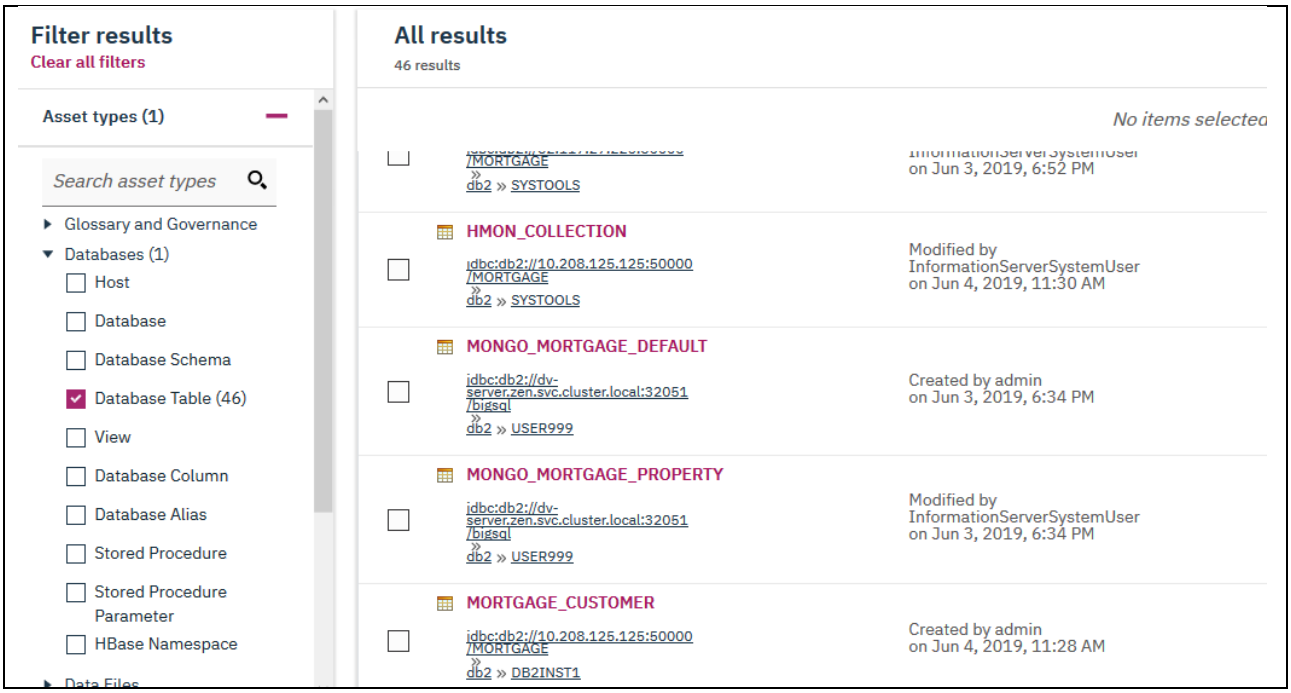


Next go back to **Organize** option on the left pane and choose **Data catalog**.

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**.



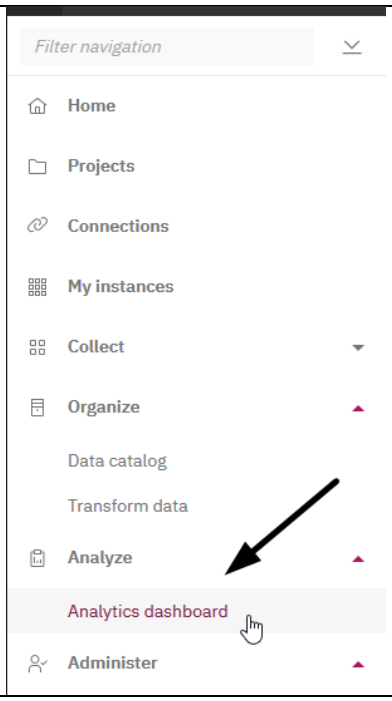
Under the **Database Tables** you can see ‘MORTGAGE_CUSTOMER’, ‘MORTGAGE_DEFAUT’ and ‘MORTGAGE_PROPERTY’ tables, cataloged from Db2 database.



11. Analyze Data

With the analytics dashboard, you can build sophisticated visualizations of your analytics results and communicate the insights you've discovered in your data on the dashboard. Then, share the dashboard with others. The analytics dashboard tool in Cloud Pak for Data provides a great way for a line-of-business user to begin investigating data for patterns and insights. The dashboard can then be handed off to a data scientist for deeper analysis and predictive modeling.

11.1. Navigate to Create Dashboard

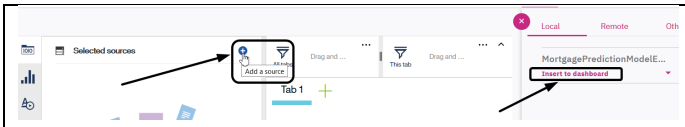


Let's create a new dashboard. From the menu, click **Analyze**, then click **Analytics dashboard**. From the Dashboards page, click Create dashboard.

Type a name for your dashboard (for example Mortgage), select a project (mortgage_data), and click Create.

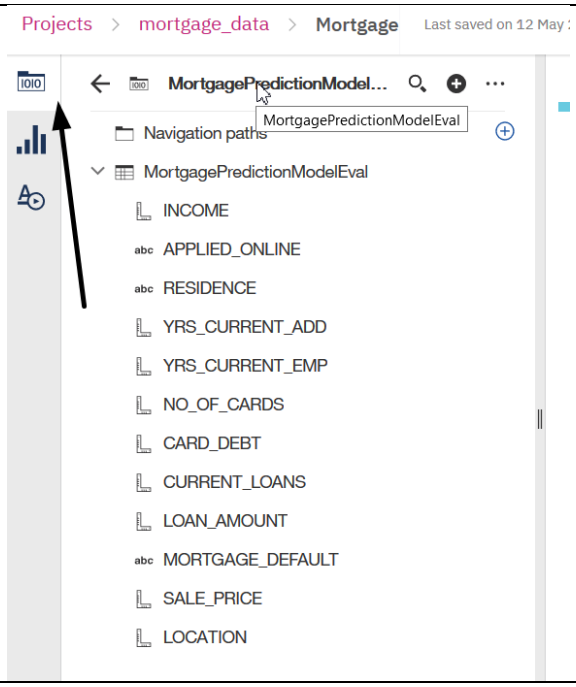
Select a template, for example the freeform tabbed template. A new empty dashboard opens.

11.2. Add a data source



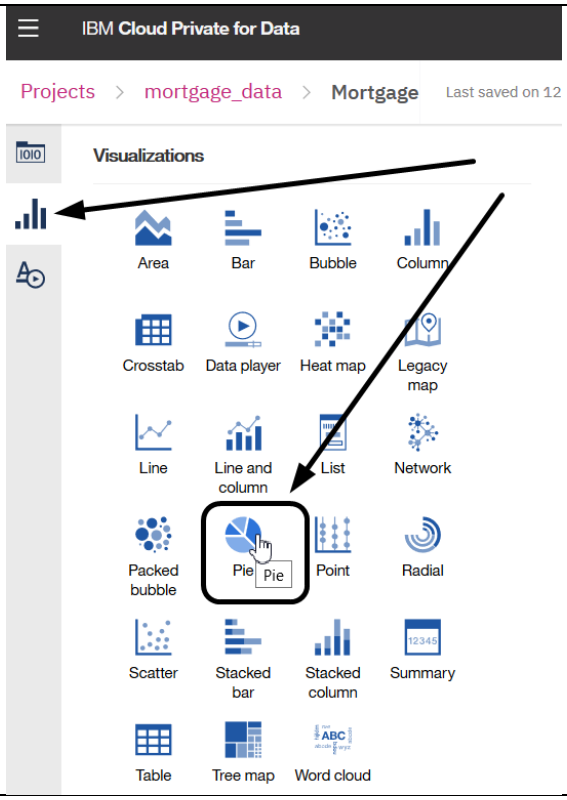
In the newly created dashboard click on **Add a source** and then, on the right, click on **Insert to dashboard** under MortgagePredictionModelEval

11.3. Explore Visualizations and Widgets



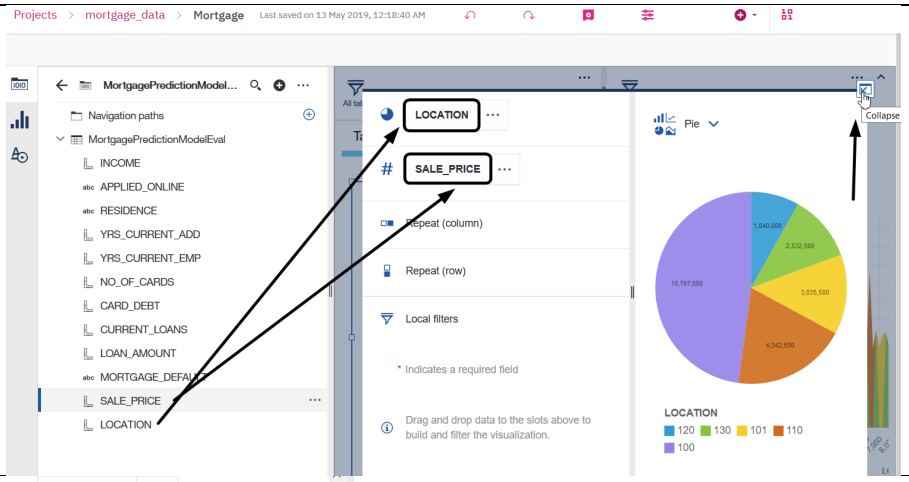
Once the data source has been added it can be viewed in the left pane.
Below Data Sources are Visualizations. There are many different types of graphs that you can use to visualize the data from the selected data source connection. Drill down into your source connection and select the data segments that you want to visualize.
Below Visualizations there are Widgets, such as text, media, web pages, images, and shapes to be added to the dashboard.

11.4. Add a graph



Select the pie graph.

11.5. Modify the graphs



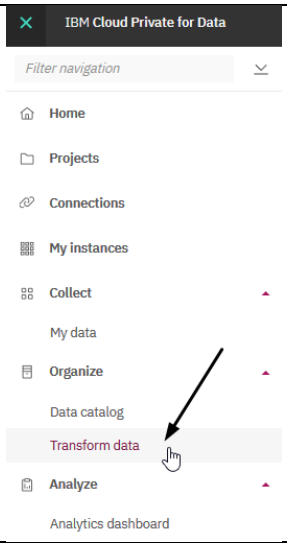
Drag and drop Location from the data source to Segments in the graph, and Sale Price for size. This shows total value of properties for each location.
Click Collapse in the top right corner of the graph.

Add a 2nd graph from the Visualization menu: Line and column. Fill in the data as on the screenshot. This graph shows total income of applicants and the value of mortgages granted for each location.

12. Transform Data

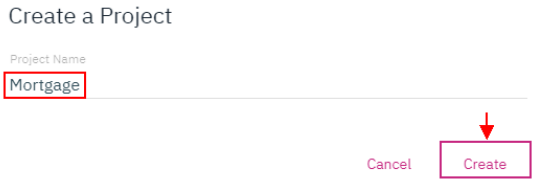
With Cloud Pak for Data, you can create, edit, load, and run transformation jobs. Cloud Pak for Data has features like built-in search, automatic metadata propagation, and simultaneous highlighting of all compilation errors, which make developers more productive.

12.1. Navigate to transform data



Let's transform the data now. Go back to **Organize** option on the left pane and choose **Transform data**.

12.2. Create a Project



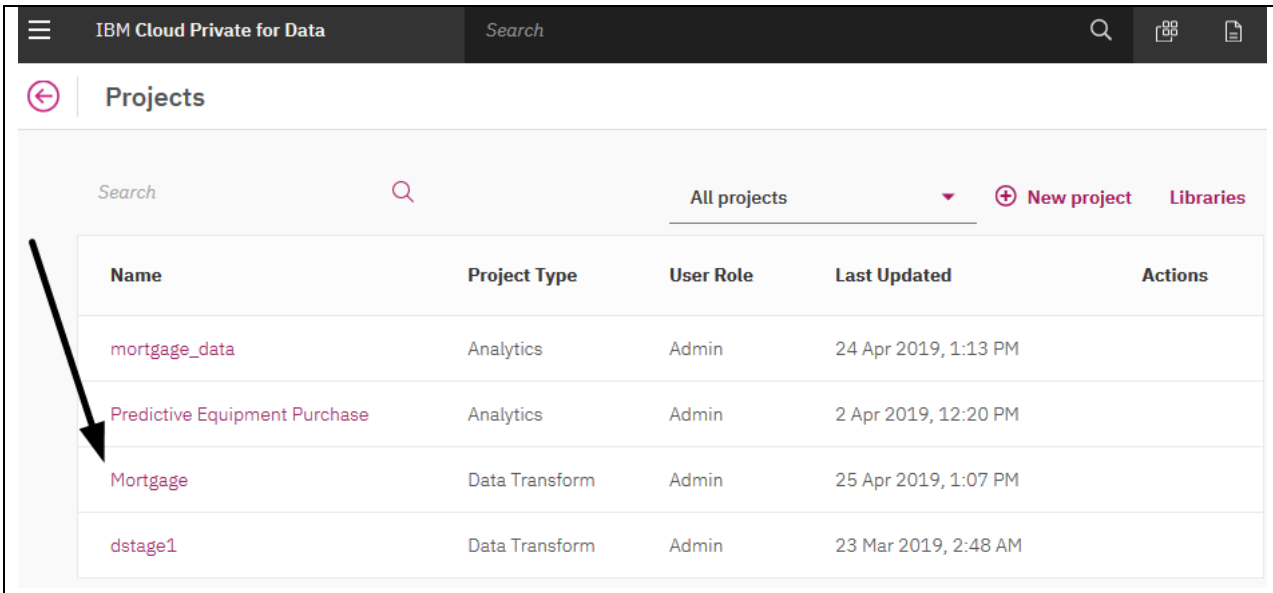
Next create a project by clicking on Create icon on top left corner.

On **Create a Project** window use the **Project Name** as 'Mortgage'.

Click on **Create**

It may take few minutes to complete.

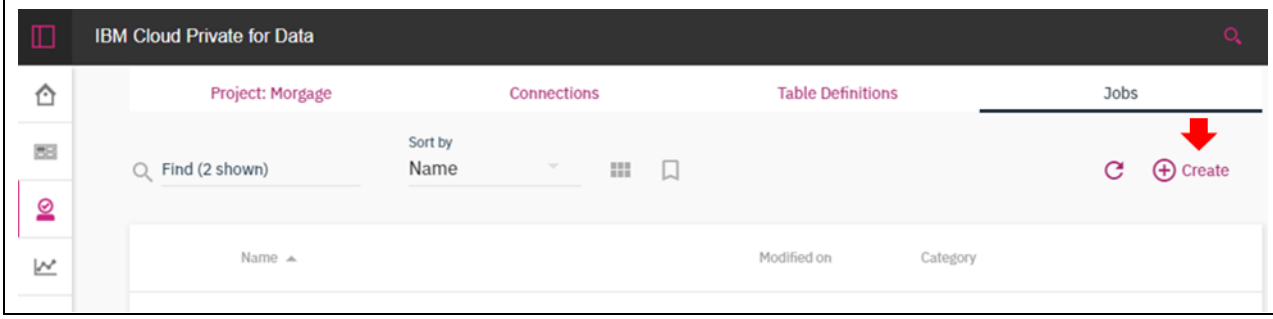
Once the project is created it will be listed under the **Projects.**



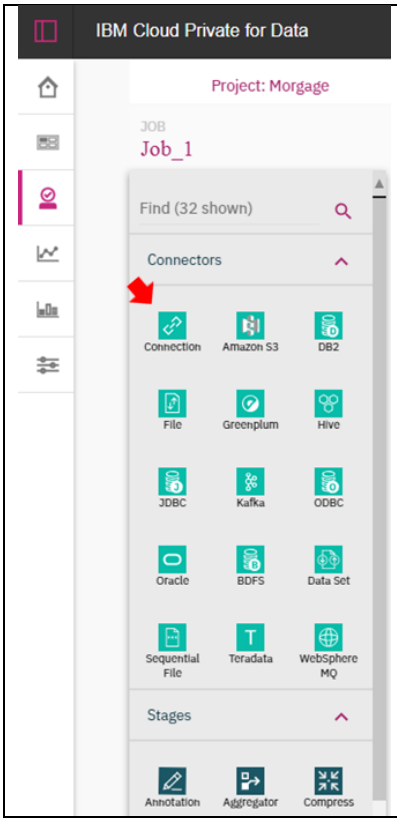
Click on the project **Name** 'Mortgage'

12.3. Create a job

Let's create a job by clicking on Create icon on top left corner.



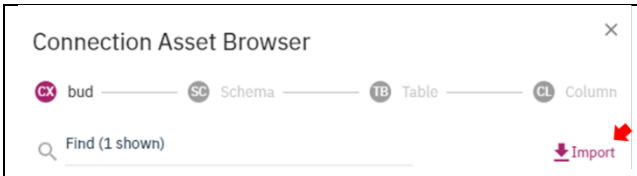
12.4. Add tables from asset browser



The create job operation will open a palette on the left.

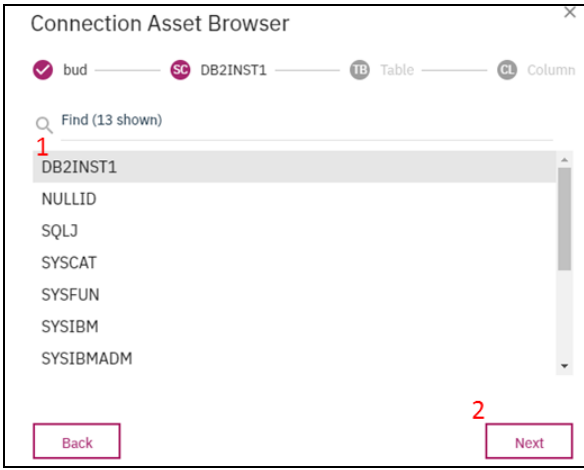
Click on the 

icon, drag it on the right pane and click once again. This will open the **Connection Asset Browser** window.

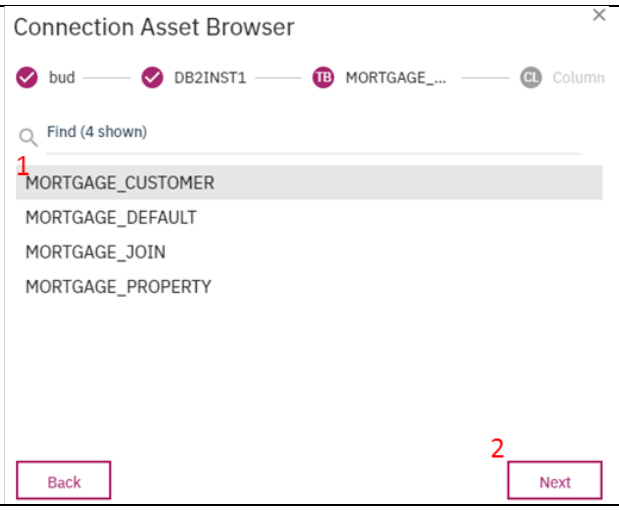


On the **Connection Asset Browser** window, Click on the **Import** to use the connection that you created earlier on step 4.2.

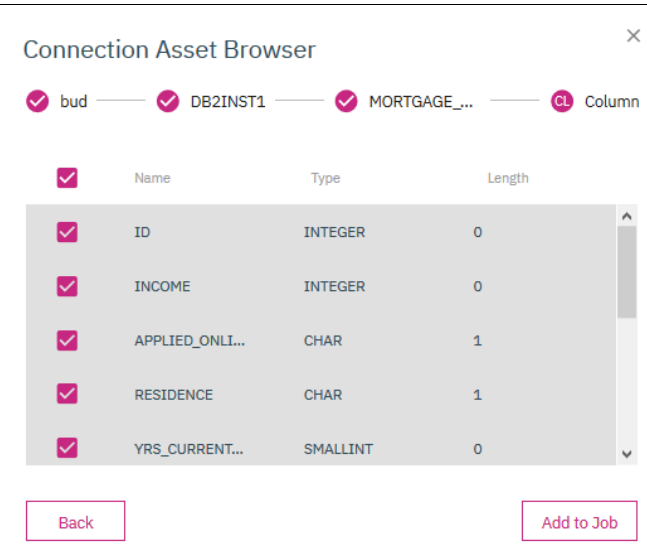
If connection name already exists, just select it and click **Next**.



1. Chose the schema named 'DB2INST1'
2. Click **Next**



1. Choose table named ‘MORTGAGE_CUSTOMER’
2. Click **Next**

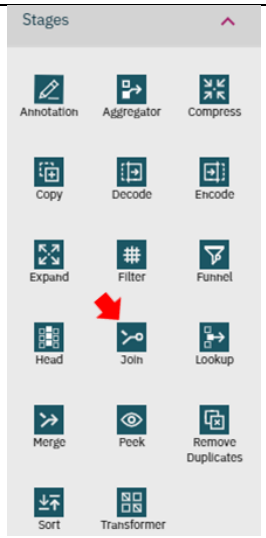


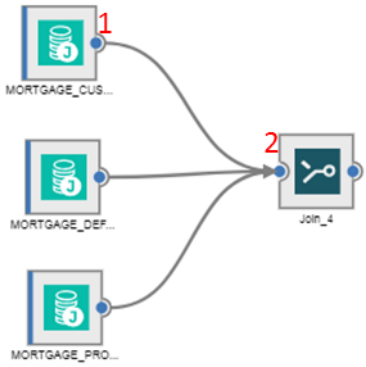
Review the column name and datatype from table ‘MORTGAGE_CUSTOMER’ and click **Add to Job**.

Repeat the above step in 5.4. for add tables ‘MORTGAGE_DEFAUT’ and ‘MORTGAGE_PROPERTY’ to the job.

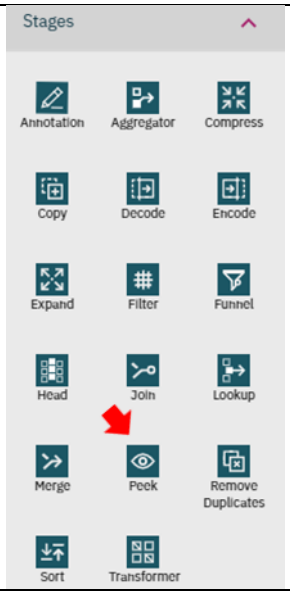
Once all three tables added to the job, you should have three tiles on right pane.

12.5. Join tables

	Now join the data. Click on the  icon from the palette on left, drag it on the right pane and click once again.

	Connect one table tile at a time to the join tile. To connect tiles, click on two blue dots one at a time.

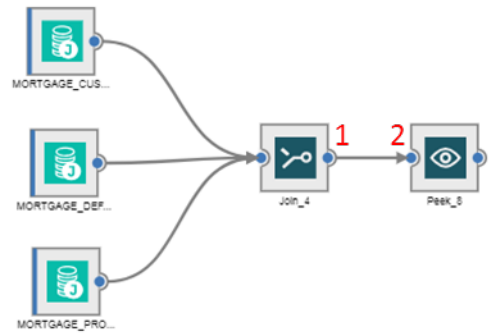
12.6. Preview output from join



Add peek to look into join output.

Click on the 

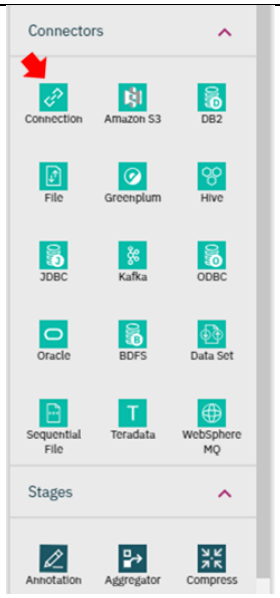
icon from the palette on left, drag it on the right pane and click once again.



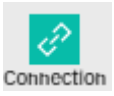
Connect join to the peek tile.

To connect tiles, click on two blue dots one at a time.

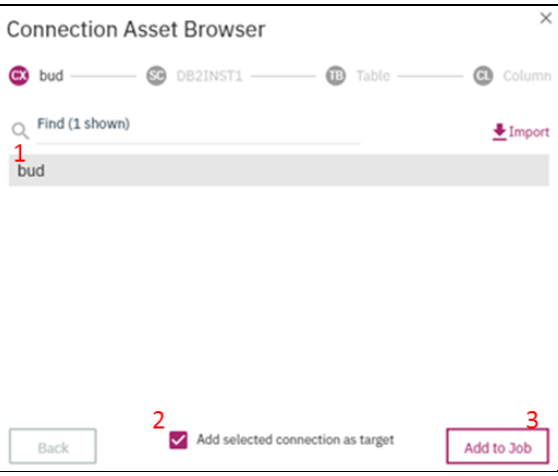
12.7. Store output from join



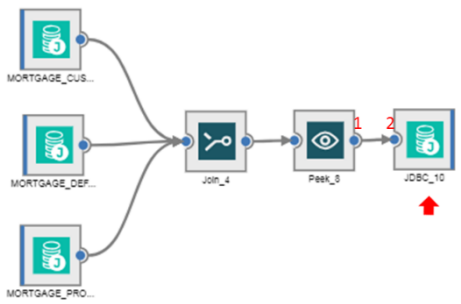
Next save the persistent data from join to the target at Db2 database.

Click on the 

icon, drag it on the right pane and click once again. This will open the **Connection Asset Browser** window.



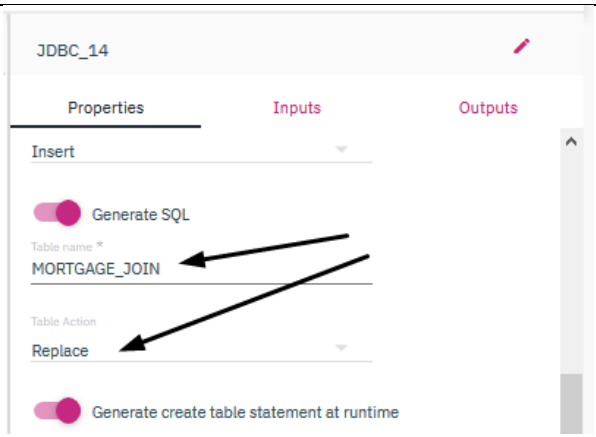
1. On the **Connection Asset Browser** window, click on connection that you created earlier in step 4.2.
2. Use check box **Add selected connection as target**
3. Click **Add to Job**



Join the target table tile with the peek

To connect tiles, click on two blue dots one at a time.

Once join completed, double-click on the new target table tile to make some adjustment.



On the Job properties pane

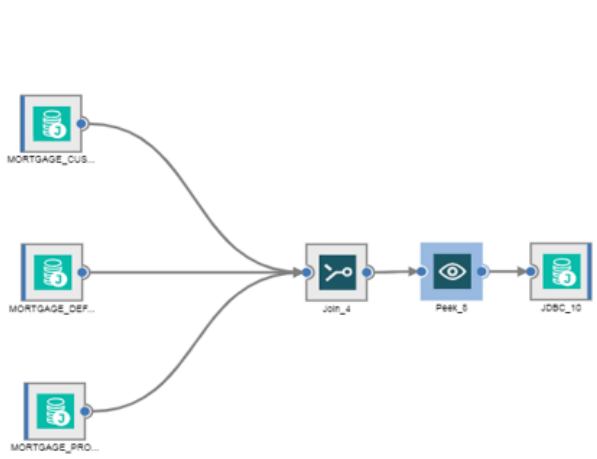
1. Use 'MORTGAGE_JOIN' as target **Table name**
2. Fro the **Table Action** dropdown menu chose 'Replace'
3. Click on **OK**

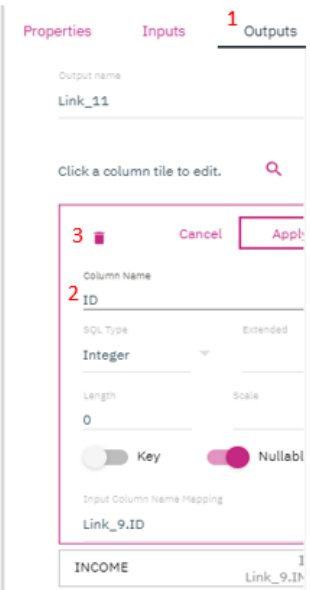
12.8. Transform output data

Let's go back to the Peek tile and double-click on it.

1. Choose the **Outputs** tab on the top right
2. Remove the column name **ID** by clicking on that column.
3. Click on the icon.
4. Click OK

For machine learning to predict mortgage default, it will use all columns, except the **ID**.

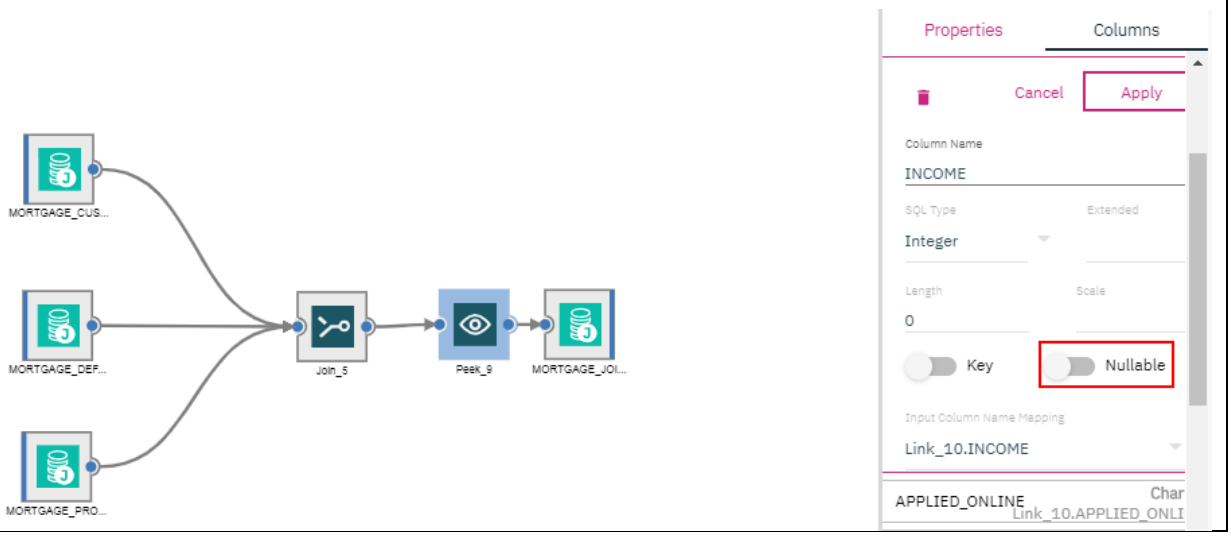




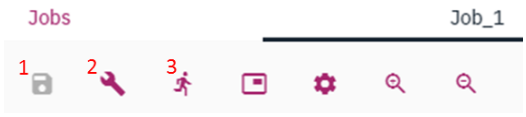
12.9. Apply governance rule

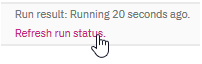
Go back to the **Peek** tile and double-click on it.

1. Choose the **Outputs** tab on the top right
2. Click on the INCOME column
3. Turn off the **Nullable** option, according to the “Income cannot be null” rule
4. Click **OK**

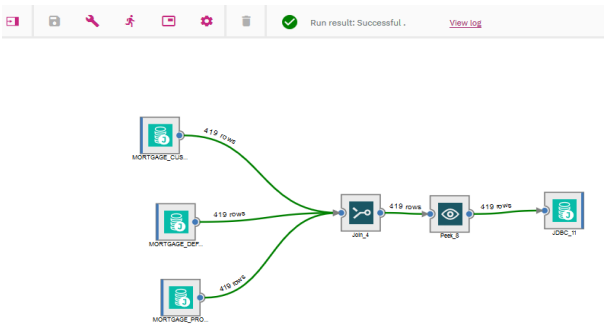


12.10. Execute job

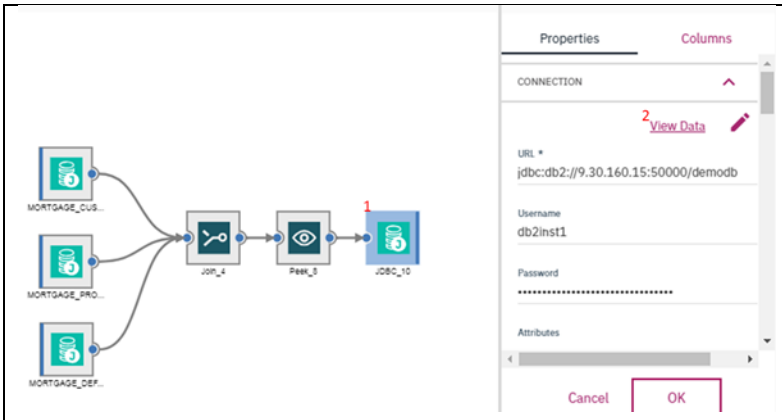


1. Click on the icon to save the job. You can use the default job name.
2. Use icon for compile the job
3. Next click on icon to run the job that will bring a **Job Run Options** window. Don’t change anything, just click on **Run**. Run may take few minutes to complete.
4. Click on the 

Refresh run status to refresh the display.



12.11. Preview output data

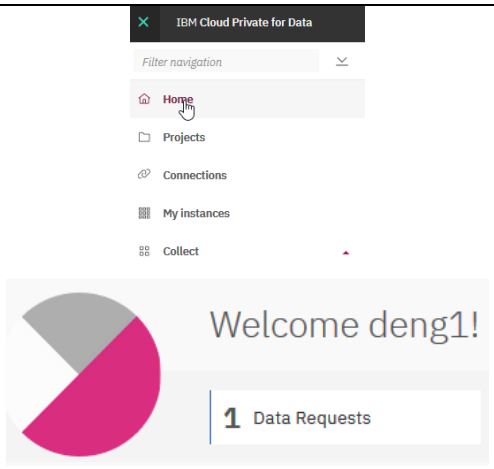


Let's take a quick look into the final data.

1. Click on the new target table tile
2. Click on **View Data**

The View Data will pop up a window with all the data. Once you done with review the data, close the window.

Deliver Dataset

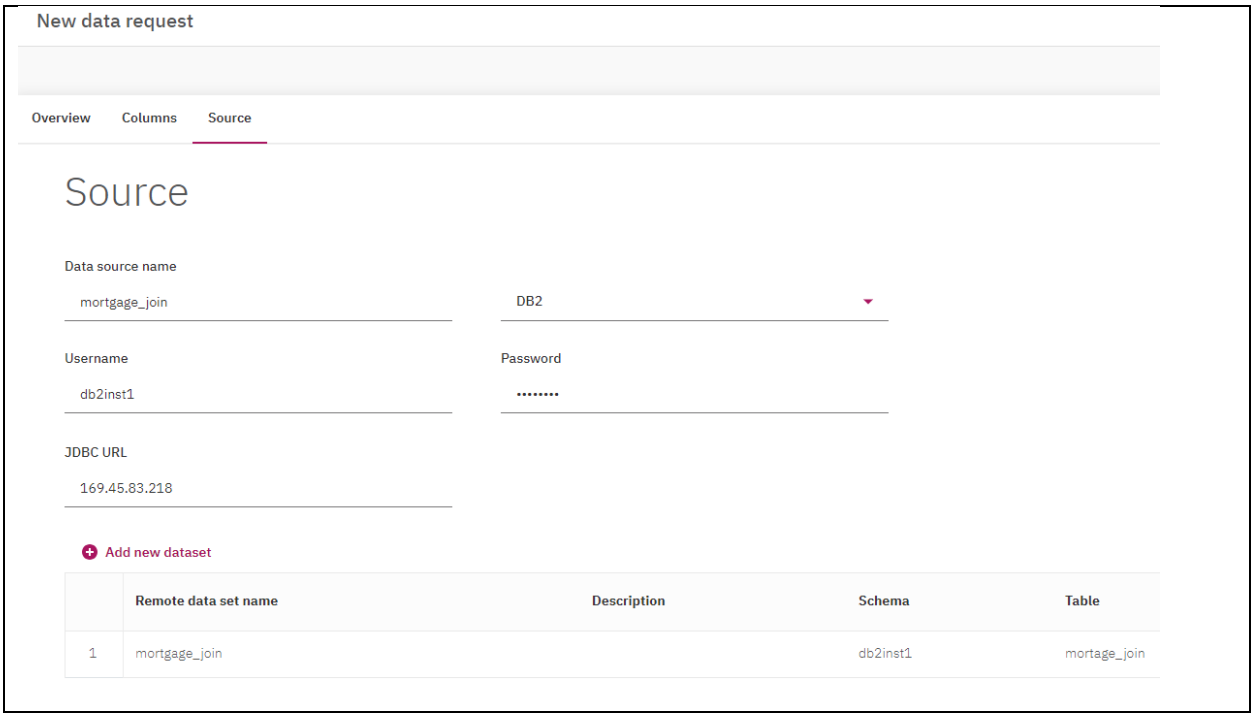


Go to the home page by clicking on icon from left pane and check the data request tab.

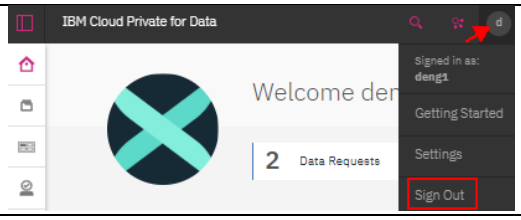
Click on the data request for update that submitted by data scientist earlier.

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. Use the IP address of master-1 node in case of JDBC URL.



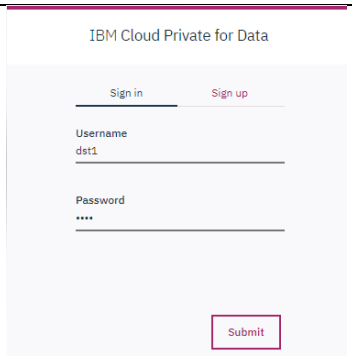
Click on the data request and change the status to **Deliver**.

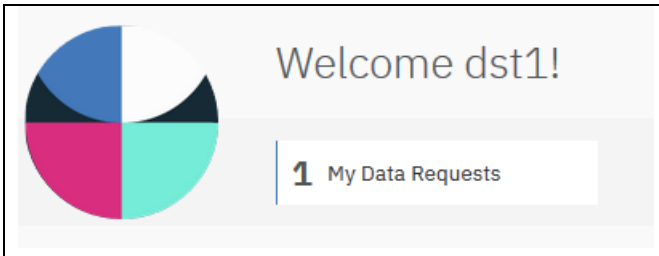


Sign out from user **deng1**

13. 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

13.1. Navigate to analytics project

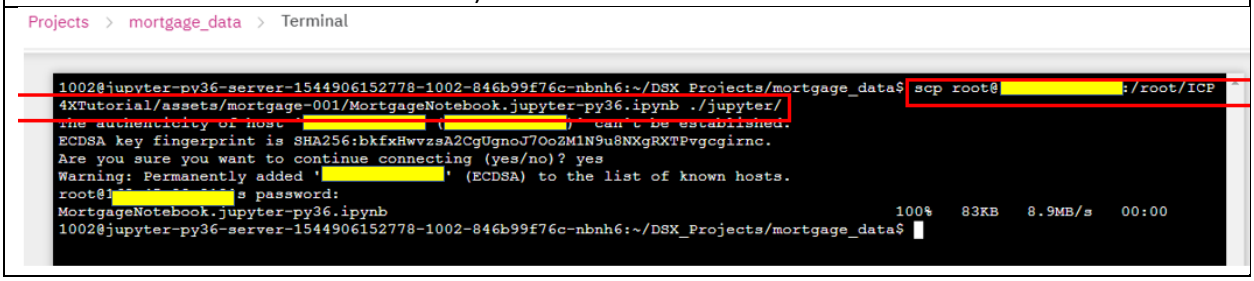
Select **Projects** option from the left pane and click on the analytics project 'mortgage_data' that you created earlier.

13.2. Create a model

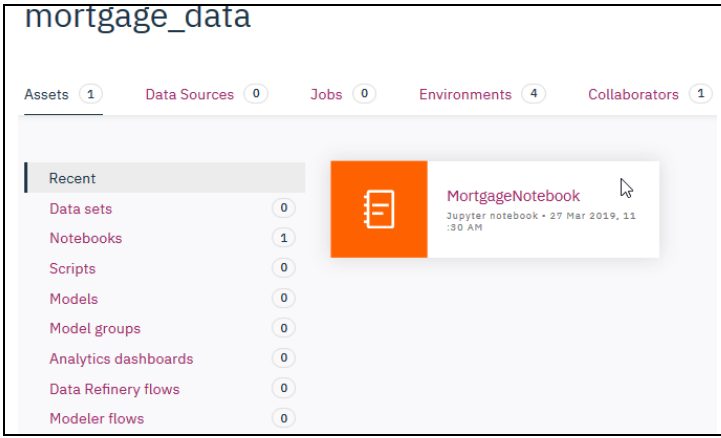
	Next, choose the **Launch Terminal with Python** from top right corner.

Copy (scp using root) a predefined Jupyter notebook from **~/ ICP4XTutorial/assets/mortgage-001/ MortgageNotebook.jupyter-py36.ipynb** on master-1 node; to **./jupyter/** directory under current project. Jupyter notebook was downloaded earlier from the Git repository.

(This step is needed just for this tutorial to create a model easily. In real life a data scientiest will not have access of root on master-1 node.)



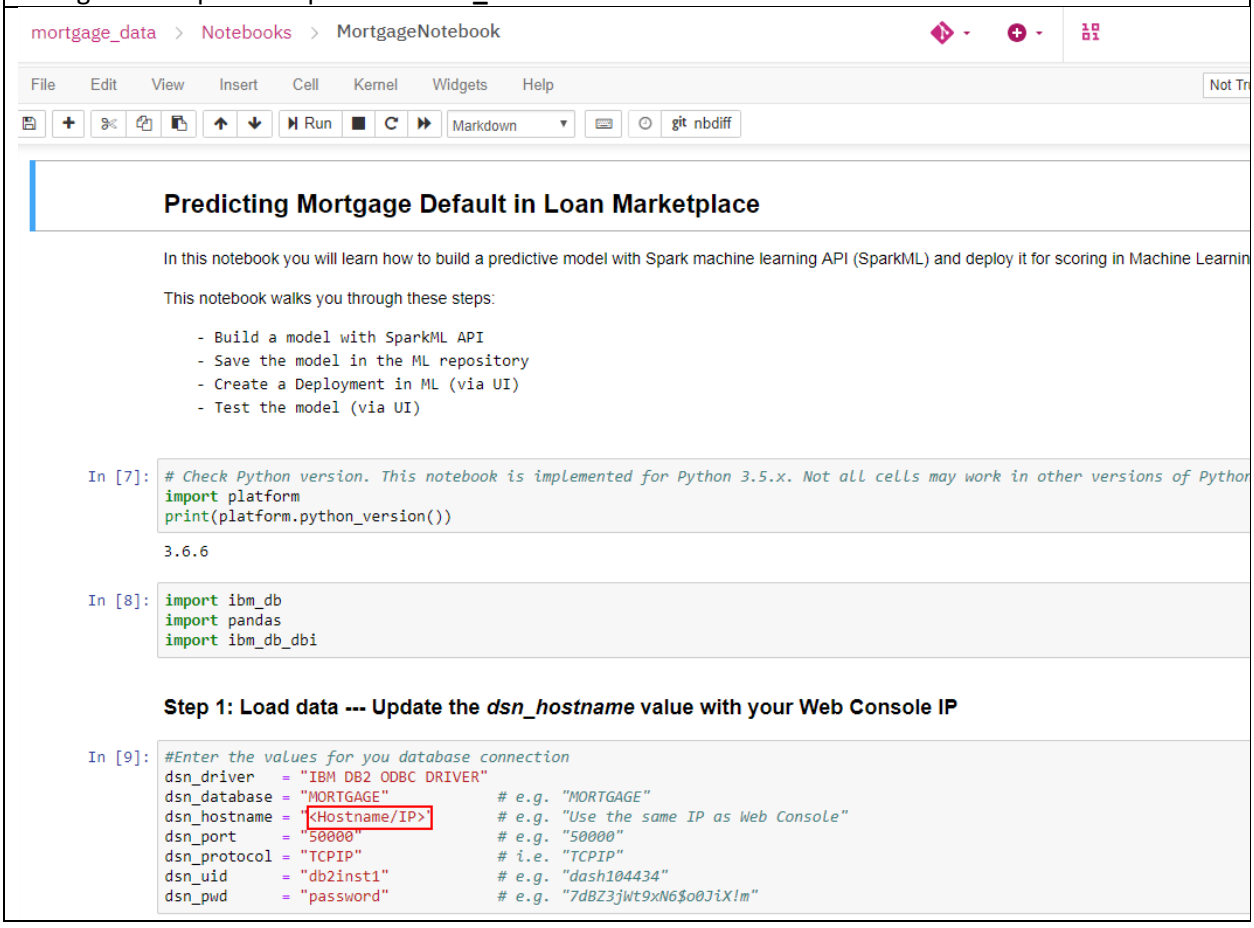
	Go back to project **mortgage_data** by clicking on the project name from top left.

	Open the predefined notebook called **MortgageNotebook**.

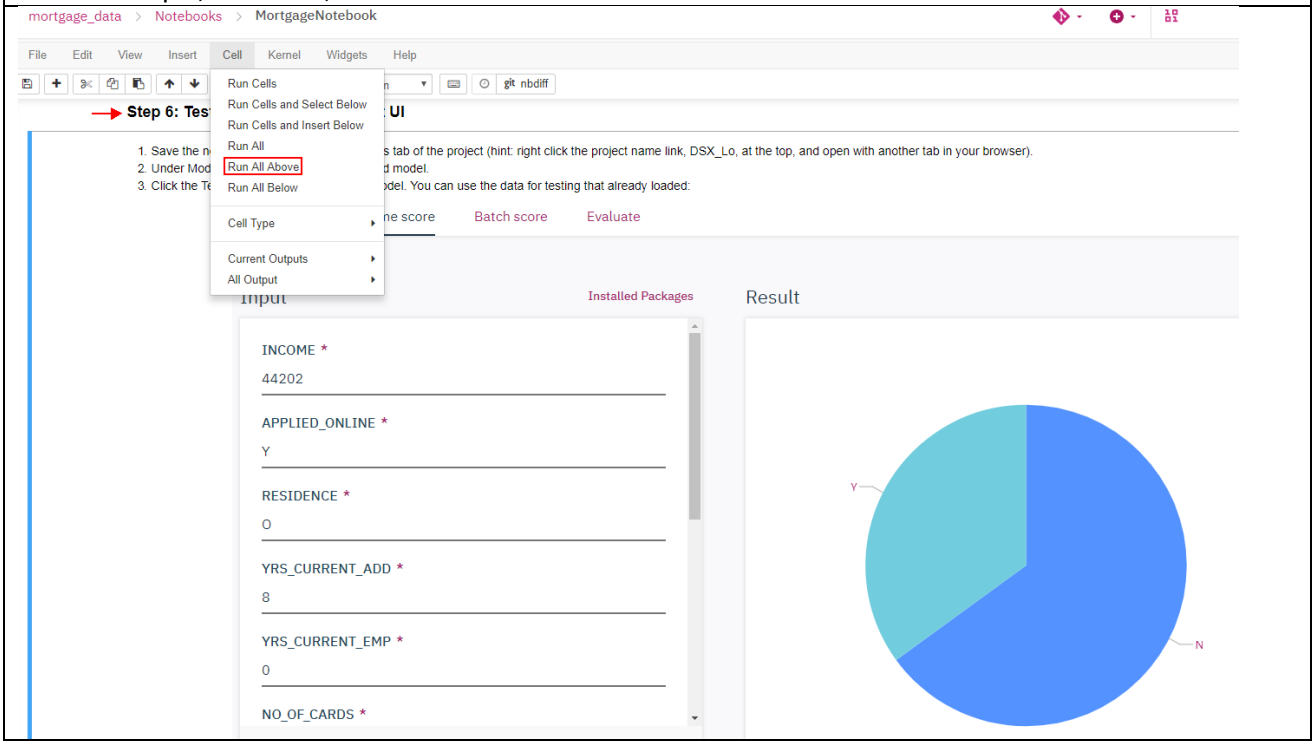
13.3. Review and Run notebook

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

First go the Step 1 and update the **dsn_hostname** value with the IP address of master node-1.

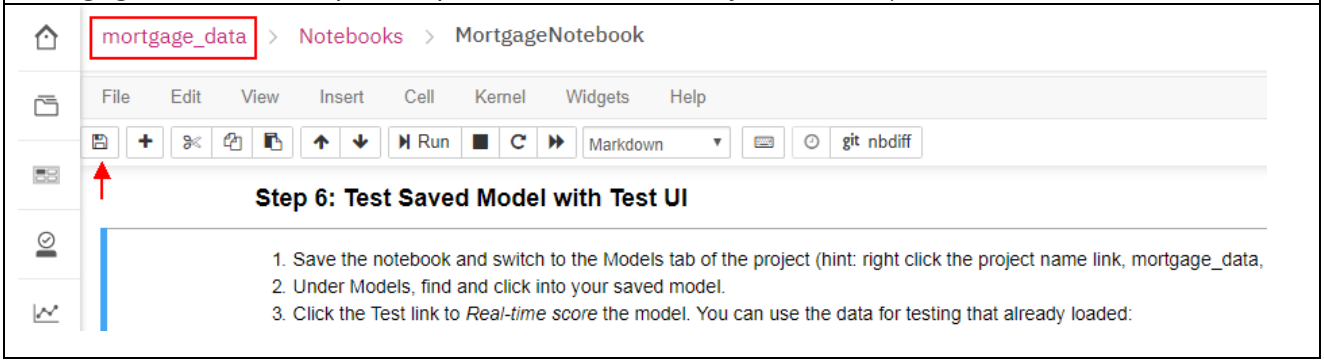


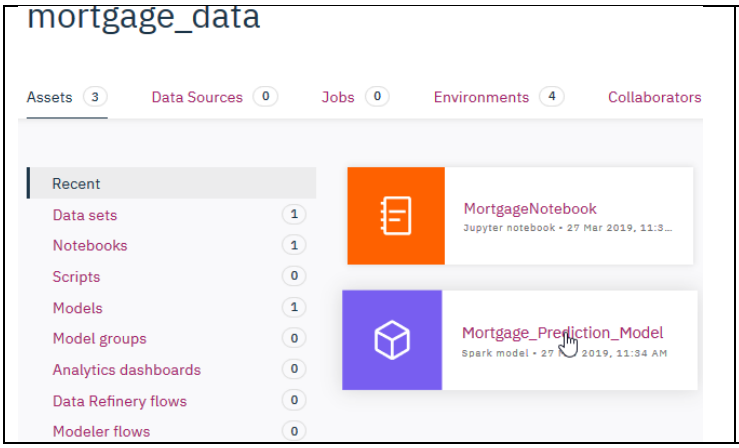
Run through it so that you generate a model. The easiest way to do this is to open the notebook, scroll down to Step 6, click on it, then in the menu select Cell -> Run all above.



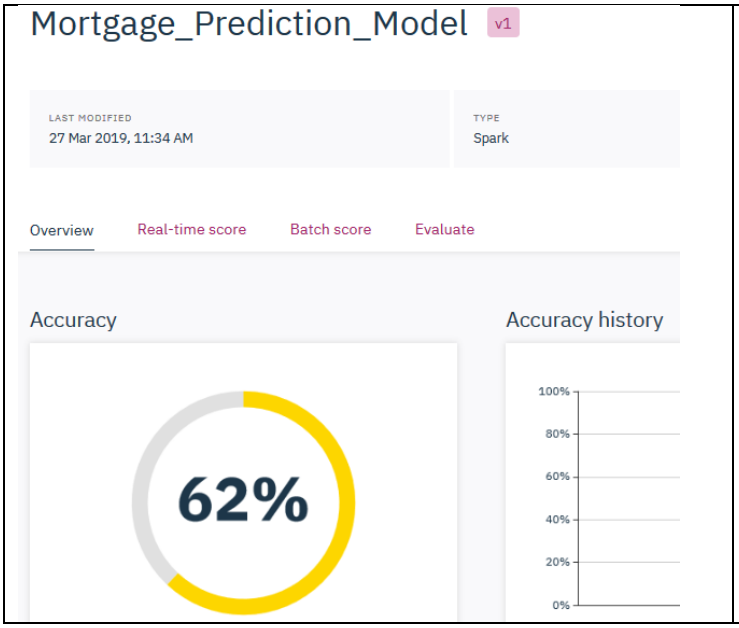
13.4. Test the model

Save the notebook and switch to the Models tab of the project (hint: right click the project name link, **mortgage_data**, at the top, and open with another tab in your browser).



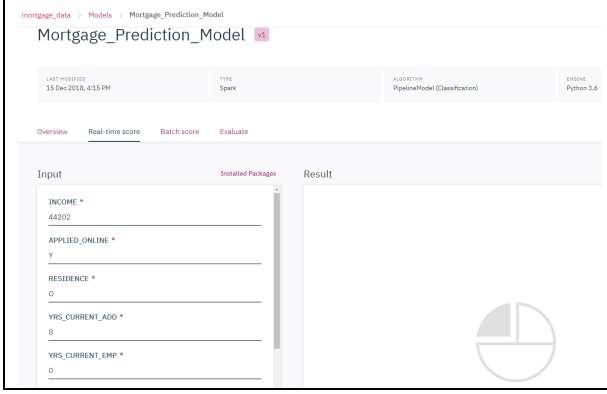


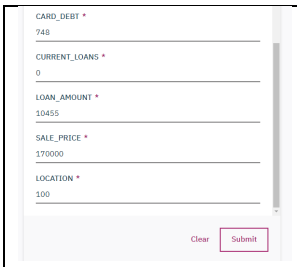
Chose the **Mortgage_Prediction_Model**



Click on the **Real-time score** to test the model.

Once your model is open, check the the mortgage default predeiction based on the sample data in the **Input** section.





If you want you can change some values in the Input section.
Then clien on **Submit**.

According on input values, model will predict the possibilities of mortgage default and produce a pie chart.