



Cloud Pak for Data
Version 3 or higher
Tutorial – Mortgage

Contents

1.	Prerequisites	5
2.	Setting up database and sample data.....	5
3.	Access Credentials	6
3.1.	Access credential for Db2 database.....	6
3.2.	Sign into Cloud Pak for Data web console as Administrator.....	6
4.	Create Connection	7
4.2.	Navigate to Connections	7
4.3.	Add connection	7
5.	Discover Assets	8
5.1.	Navigate to discover assets	8
6.	Add users	11
6.1.	Grant Catalog Permission	12
6.2.	Create Analytic Project	14
6.3.	Create Deployment Space	15
7.	Implement Business Glossary.....	17
7.1.	Download Business Glossaries.....	17
7.2.	Import Categories	18
7.3.	Import Terms	19
7.4.	Create a policy	20
7.5.	Create a rule	20
7.6.	Automated Discovery	21
7.7.	Add rule to metadata.....	21
8.	Access data as a Data Scientist	24
8.1.	Assets from Glossary	24
8.2.	Check Asset Details	25
9.	Data Virtualization	27
9.1.	Adding a new data source for Db2.....	27
9.2.	Adding Users to Data Virtualization.....	29
9.3.	Select tables for virtualization	30
9.4.	Creating Virtual Table	31
9.5.	Add virtual table to catalog	32
9.6.	Publish virtualized table.....	32

Cloud Pak for Data (v.3 or higher) – Tutorial

9.7. Access information for virtual table	33
10. Build Model	34
10.1. Navigate to analytics project	34
10.2. Create notebook	34
10.3. Review and run notebook.....	35
10.4. Test the model.....	36

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 a docker image. User will perform following tasks to predict if a prospective customer may default on their mortgage.

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

1. Prerequisites

- Access to an operational Cloud Pak for Data (v.3 or higher) Instance
- Install Git on the machine that you will use for the tutorial
- Docker or Podman available on the machine that you will use for the tutorial
- WKC, Data Virtualization and Watson Studio services enabled on Cloud Pak for Data

2. Setting up database and sample data

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

From your home directory, clone the tutorial sample files:

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

Change to the tutorials directory:

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

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

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

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

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

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

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

3. Access Credentials

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

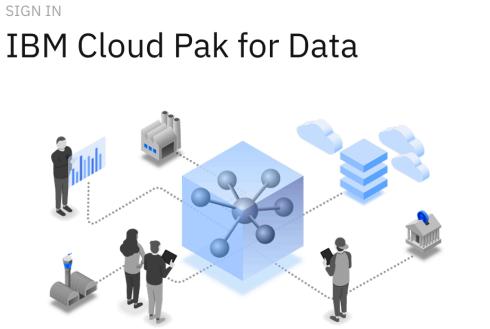
3.1. Access credential for Db2 database

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

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

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

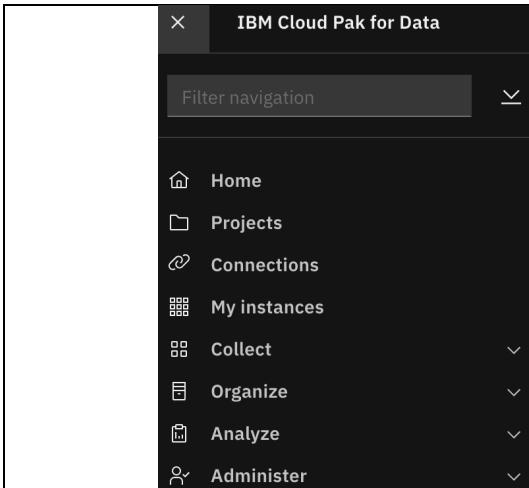
You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data web console. Starting from here all instruction need to execute on Cloud Pak for Data web console only. You need to login as admin who has administrator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

 <p>SIGN IN</p> <p>IBM Cloud Pak for Data</p> <p>Username admin</p> <p>Password</p> <p>Sign in →</p>	<p>Sigh into the Cloud Pak for Data web console as user ‘admin’ and password is ‘password’.</p>
---	---

4. Create Connection

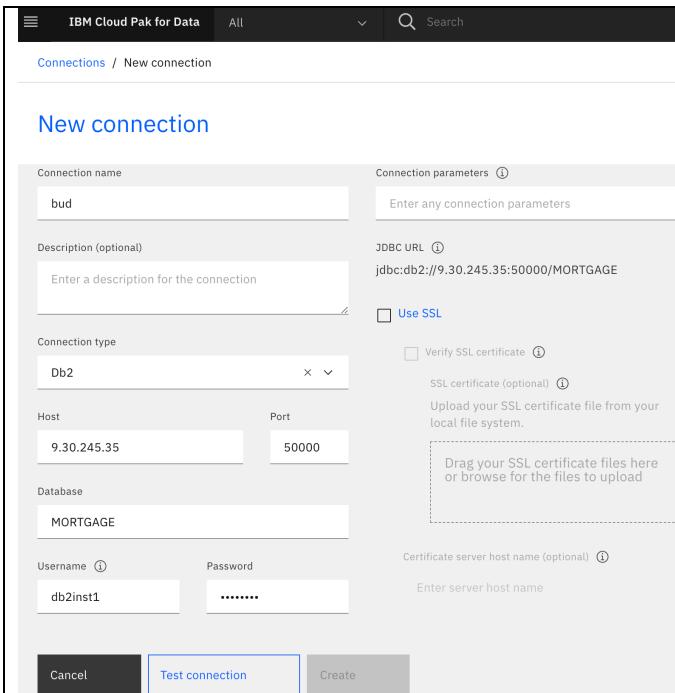
Create a connection to the data source for Db2 database.

4.2. Navigate to Connections



On the left pane choose **Connections**. Next, on the **Data Connections** window click on the **New connection** + icon.

4.3. Add connection



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

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

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

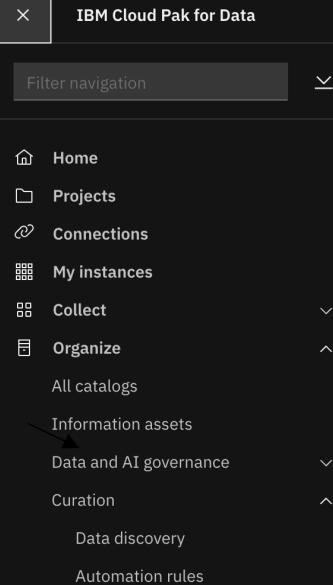


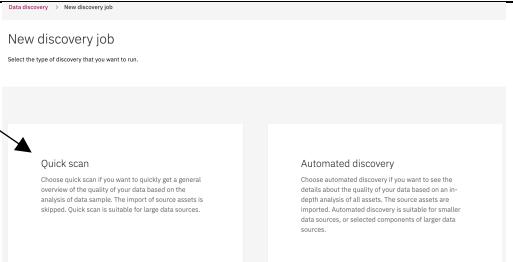
Success The test connection was successful. Click Add to save the connection information.

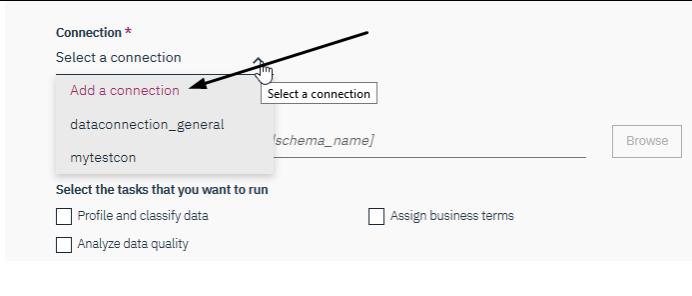
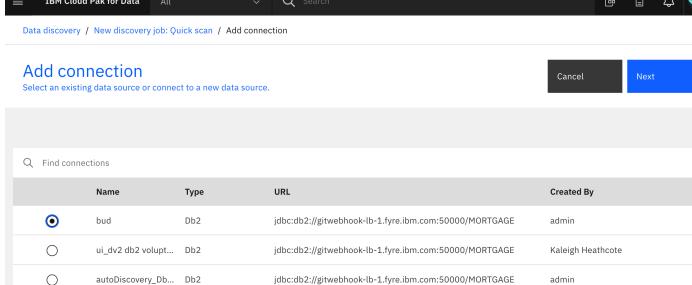
5. Discover Assets

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

5.1. Navigate to discover assets

	<p>From Organize option on the left pane, choose Curation > Data discovery.</p>
---	--

	<p>Click on Quick scan</p>
--	-----------------------------------

	<p>To discover assets</p> <ol style="list-style-type: none"> 1. Click on Add a connection 2. Choose the connection named bud that you created previously, click Next
	

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

It may take few minutes to complete.

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

Job ID	Data assets	Connection	Started by	Processing time	Status	Status u
qs_1585614853064	4	bud	admin	42 seconds	Ready for review	March 3

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

Discovered columns (27)								Find column
Asset type	Column name	Identity	Quality	Assigned business term	Suggested business term	Assigned data class	Suggested data class	Business term actions
○ File	APPLIED_ONLINE	MORTGAGE_JOIN	100%	-	-	Indicator 100%	-	
○ Schema	APPLIED_ONLINE	MORTGAGE_CUSTOMER	100%	-	-	Indicator 100%	-	
○ Table	CARD_DEBT	MORTGAGE_CUSTOMER	96%	-	-	-	US Zip Code 2%	
● Column	CARD_DEBT	MORTGAGE_JOIN	96%	-	-	-	US Zip Code 4%	
Filters	CURRENT_LOANS	MORTGAGE_JOIN	100%	-	-	Boolean 100%	Indicator 100%	
Labels	CURRENT_LOANS	MORTGAGE_CUSTOMER	100%	-	-	Boolean 100%	Indicator 100%	
No filters of this type	ID	MORTGAGE_DEFAULT	100%	-	-	Identifier 100%	-	
Tables	ID	MORTGAGE_CUSTOMER	100%	-	-	Identifier 100%	-	
3 tables selected	ID	MORTGAGE_PROPERTY	100%	-	-	Identifier 100%	-	
MORTGAGE_CUSTOMER								
MORTGAGE_PROPERTY								
MORTGAGE_DEFAULT								
MORTGAGE_JOIN								
<input type="button" value="Clear"/>	<input type="button" value="Apply"/>							

Items per page: 10 | 1–10 of 27 items

1 of 3 pages < > 1 *

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

Select Asset type as “Column”

Filters necessary tables using checkbox

Click on Apply

Approve results					
Asset type	Table name	Identity	Quality	Schema name	Discovery root
○ File	MORTGAGE_CUSTOMER	MORTGAGE_CUSTOMER_qs_1573834179481	100%	DB2INST1	schema[MORTGAGE]DB: Ready for review
○ Schema	MORTGAGE_DEFAULT	MORTGAGE_DEFAULT_qs_1573834179481	100%	DB2INST1	schema[MORTGAGE]DB: Ready for review
● Table	MORTGAGE_JOIN	MORTGAGE_JOIN_qs_1573834179481	100%	DB2INST1	schema[MORTGAGE]DB: Ready for review
○ Column	MORTGAGE_PROPERTY	MORTGAGE_PROPERTY_qs_1573834179481	100%	DB2INST1	schema[MORTGAGE]DB: Ready for review

Change Asset type as “Table”

Select all Mortgage related tables

Click on Approve results

Approve assets

The selected assets will be added to the catalog so that other users can access them.

The analysis results for these assets will not be included in the catalog until you publish them.

The analysis results will be loaded to the project that you selected when you started the new discovery job. In the project, you can run further analysis, edit the results, or publish them.

Selected assets (3)

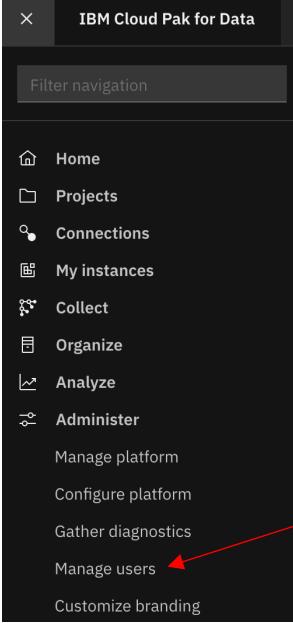
MORTGAGE_CUSTOMER
 MORTGAGE_DEFAULT
 MORTGAGE_PROPERTY

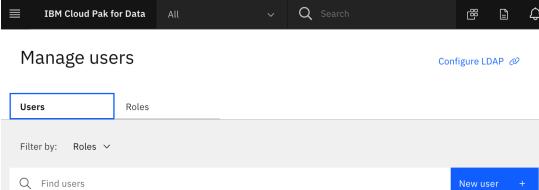
Cancel
Approve

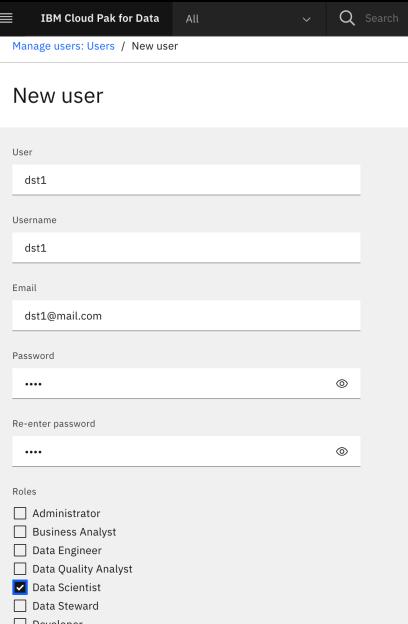
Click on Approve

6. Add users

Create users with different roles.

	<p>From Administer option on the left pane, choose Manage users.</p>
---	--

	<p>Switch tab to ‘Users’ and click on ‘Add user’</p>
--	--

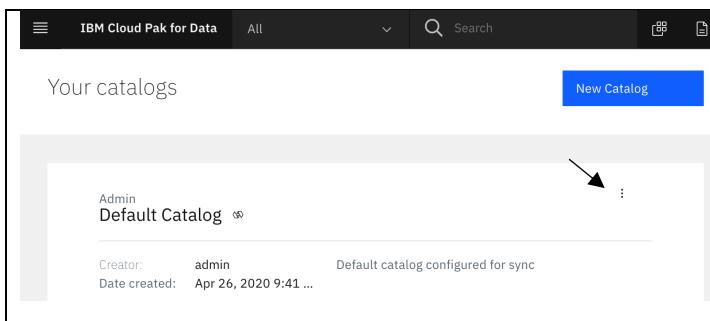
	<p>Fill out Add User information for a data scientist</p> <ol style="list-style-type: none"> 1. User as dst1 2. Username is dst1 3. Use a valid email address 4. Set Password as dst1 5. Choose the user roles as Data Scientist <p>Click on Create to confirm the add user</p>
---	---

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 Steward	dstw1

6.1. Grant Catalog Permission

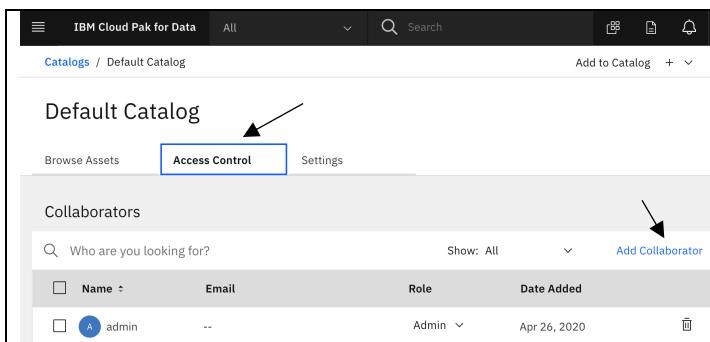
With Watson Knowledge Catalog, you use catalogs to easily find and share your data and other assets. A catalog is like a private community for your organization. It's a way to organize resources for many data science projects: data assets, analytical assets, and the users who need to use the assets.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



Go to **Organize > All catalogs**

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

Choose **View**



Go to **Access Control tab**

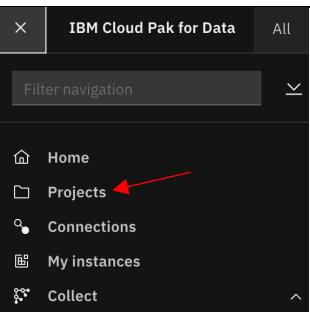
Click on **Add Collaborator**

Add ‘deng1’ and ‘dstw1’ user as collaborator with editor role

Add ‘dst1’ user as collaborator with viewer role

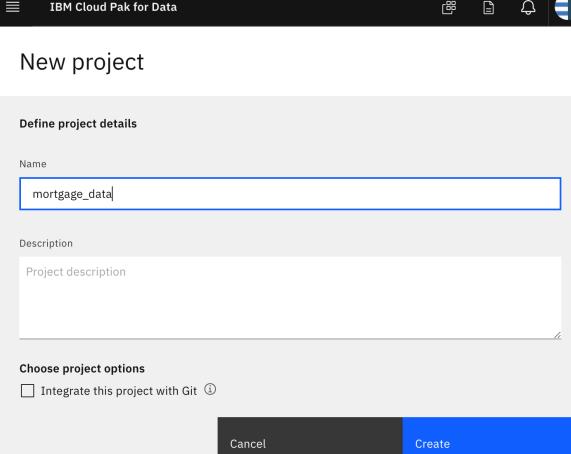
6.2. Create Analytic Project

A project is how you organize your resources to achieve a particular goal. Your project resources can include data, collaborators, and analytic assets like notebooks and models.

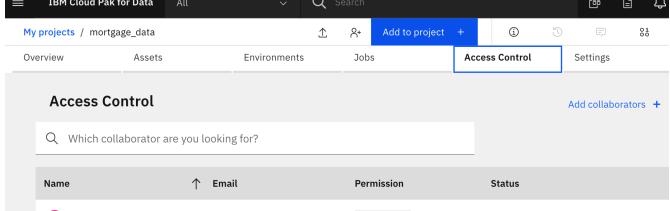


Create a new analytical project by selecting **Projects** on the left pane.

- Click on the **New Project** icon
- Select project type as **Analytics project**
- Select **Create an empty project**

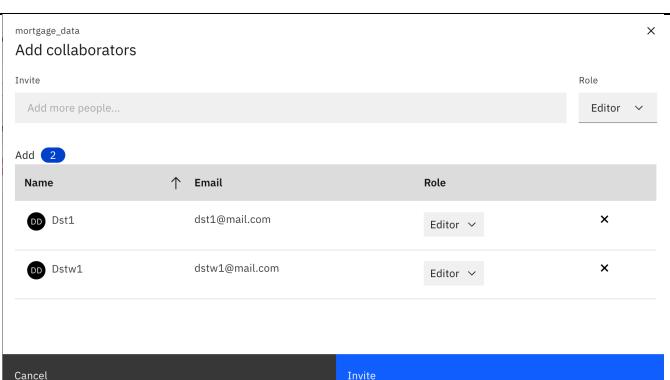


Provide project name **mortgage_data** and click **Create**



Go to **Access Control** tab

Click on **Add Collaborator**



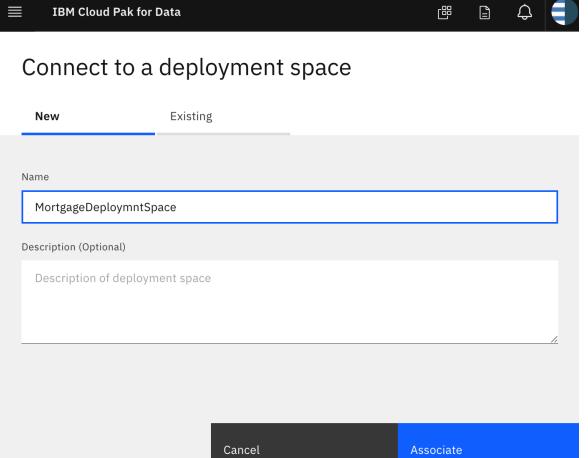
Add **dst1** and **dstw1** user as collaborator with editor role.

Click **Invite**

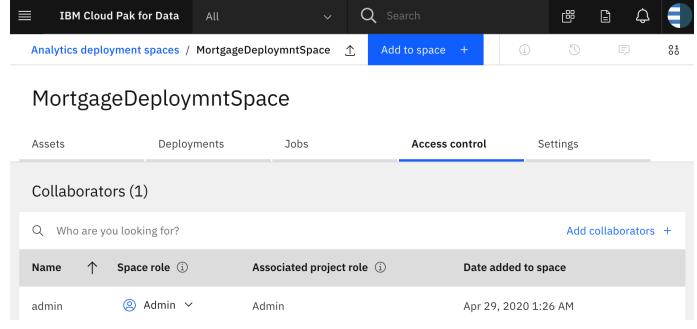
6.3. Create Deployment Space

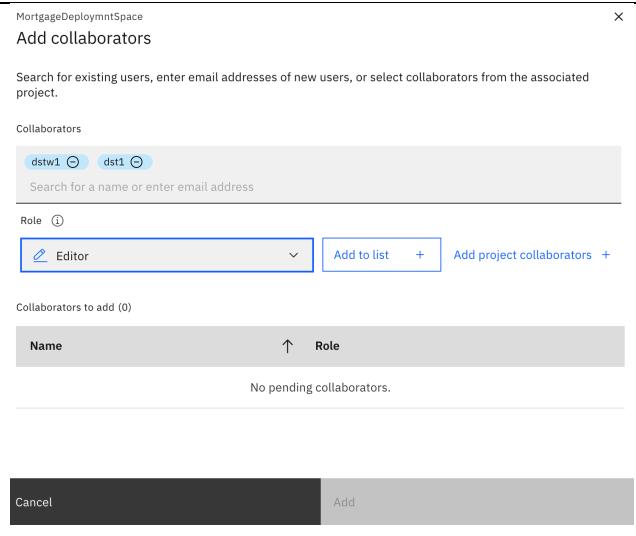
Create a separate deployment space for your project **mortgage_data**.

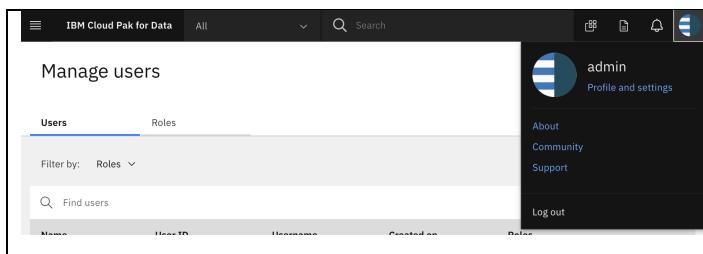
Choose : My Projects > **mortgage_data** > Settings > Associate a deployment space > New

	<p>Name new deployment space as MortgageDeploymntSpace</p> <p>Click on Associate</p>
---	--

Continue with the **mortgage_data** project settings. Select **MortgageDeploymntSpace** under associated deployment space

	<p>Go to Access Control tab</p> <p>Click on Add Collaborator</p>
--	--

	<p>Add dst1 and dstw1 user as collaborator with Editor role.</p> <p>Click Add to list</p> <p>Click Add</p>
---	--



The screenshot shows the 'Manage users' page in the IBM Cloud Pak for Data interface. On the right, a user profile menu is open for the user 'admin'. The menu includes options like 'Profile and settings', 'About', 'Community', 'Support', and 'Log out'. The 'Log out' option is highlighted.

Log out from user **admin**

7. Implement Business Glossary

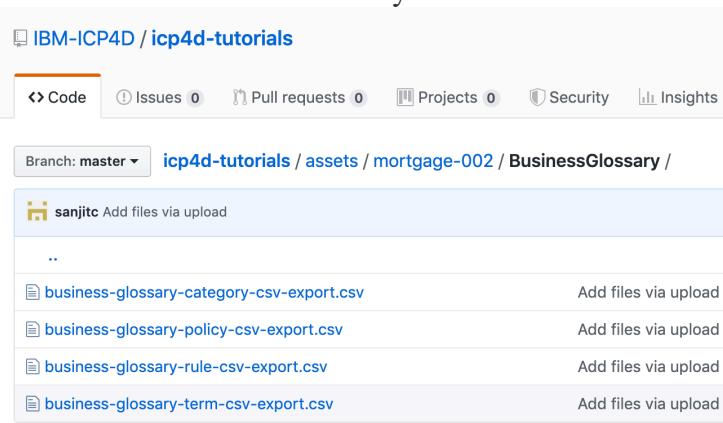
Cloud Pak for Data enables you to structure your enterprise information in a logical way, discover relationships between assets, and keep your data always up-to-date. You can import existing glossary with categories, terms, information governance policies and rules.

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.

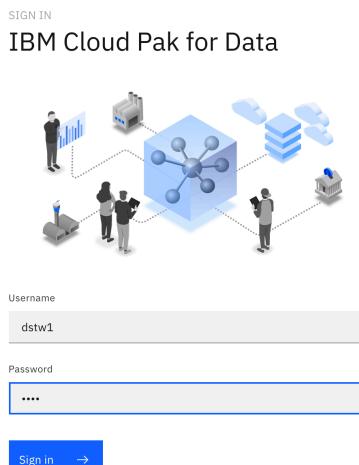


The screenshot shows a GitHub repository interface. At the top, there's a navigation bar with 'Code' selected. Below it, a header bar shows 'Branch: master' and the repository path 'icp4d-tutorials / assets / mortgage-002 / BusinessGlossary /'. A user profile 'sanjitic' is visible. The main content area displays a list of four CSV files under the 'BusinessGlossary' folder. Each file has a small preview icon and the file name followed by an 'Add files via upload' button.

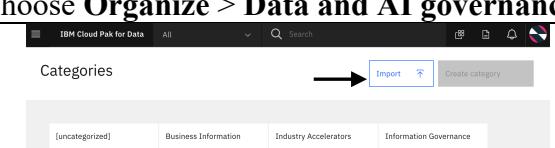
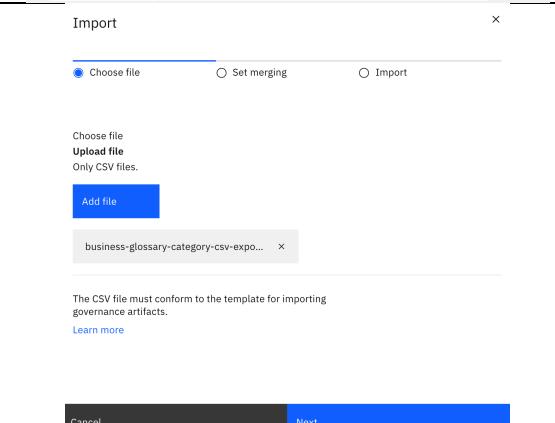
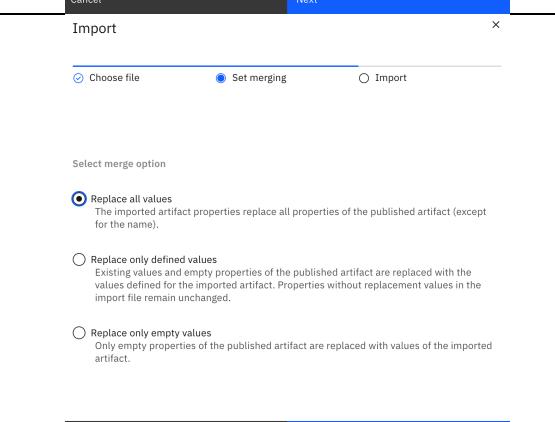
File	Action
business-glossary-category-csv-export.csv	Add files via upload
business-glossary-policy-csv-export.csv	Add files via upload
business-glossary-rule-csv-export.csv	Add files via upload
business-glossary-term-csv-export.csv	Add files via upload

- ❖ Go to that particular business glossaries that you want to **download** and click on it.
- ❖ You will see "Raw" button on the top right side of the dataset.
 - Press "Alt" and then left click the "Raw" button (on Windows) or
 - Click with two fingers (on Mac)
- ❖ Download link file

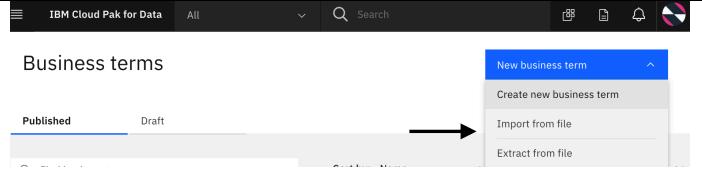
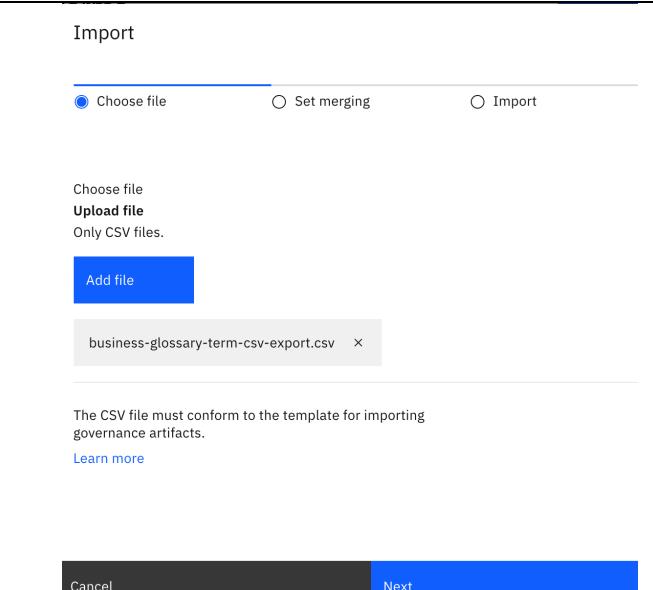
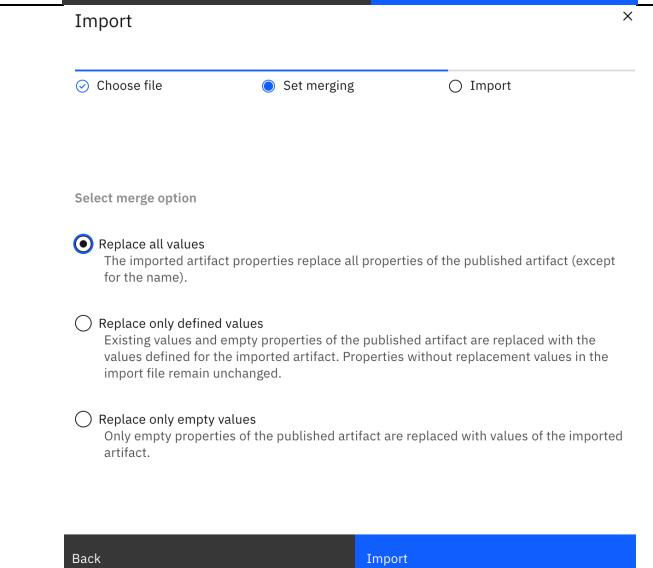
7.2. Import Categories

 <p>The sign-in page for IBM Cloud Pak for Data. It features a header 'SIGN IN' and 'IBM Cloud Pak for Data'. Below the header is a logo depicting three stylized human figures interacting with a central blue cube representing data. The form fields include 'Username' (dstw1) and 'Password' (redacted). A 'Sign in' button with a right-pointing arrow is at the bottom.</p>	<p>Sigh into the Cloud Pak for Data web console as user ‘dstw1’ and password is ‘dstw1’ that you created earlier.</p>
---	---

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

<p>Choose Organize > Data and AI governance > Categories from the left pane.</p>  <p>The 'Categories' page in the Cloud Pak for Data interface. It shows a list of categories like [uncategorized], Business Information, Industry Accelerators, and Information Governance. An 'Import' button is highlighted with a red arrow.</p>		<p>Click on Import to import the CSV file contains category information that you downloaded from Git.</p>
 <p>The 'Import' dialog box. It has tabs for 'Choose file' (selected), 'Set merging', and 'Import'. A file named 'business-glossary-category-csv-expo...' is listed under 'Choose file'. A note says 'The CSV file must conform to the template for importing governance artifacts.' with a 'Learn more' link.</p>		<p>Choose the CSV file location by using Add file tab Click Next</p>
 <p>The 'Import' dialog box again, showing the 'Set merging' tab selected. It lists three options: 'Replace all values' (selected), 'Replace only defined values', and 'Replace only empty values'. Each option has a detailed description below it.</p>		<p>Select merge option as Replace all values Click Import</p>

7.3. Import Terms

<p>Choose Organize > Data and AI governance > Business terms from the left pane.</p> 		<p>Click on New business term > Import from file to import the CSV file contains term information that you downloaded from Git.</p>
		<p>Choose the CSV file location by using Add file tab</p> <p>Click Next</p>
		<p>Select merge option as Replace all values</p> <p>Click Import</p> <p>Review each imported business terms and then publish</p>

7.4. Create a policy

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

<p>Choose Organize > Data and AI governance > Policy from the left pane</p> <p>Select Published tab and click on New Policy > Create Policy</p>	<p>On the Create new policy window create a policy with following information and click on Save as draft:</p> <p>Name: Data Validation Description: Check for appropriate data</p> <p>It will take few minutes to appear under list of available policies.</p> <p>Once new policy available let's publish it.</p>
---	--

7.5. Create a rule

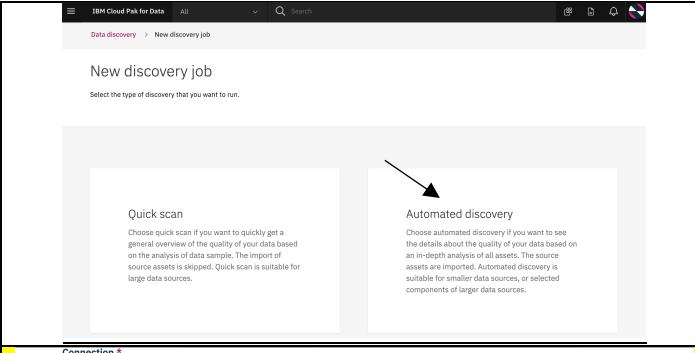
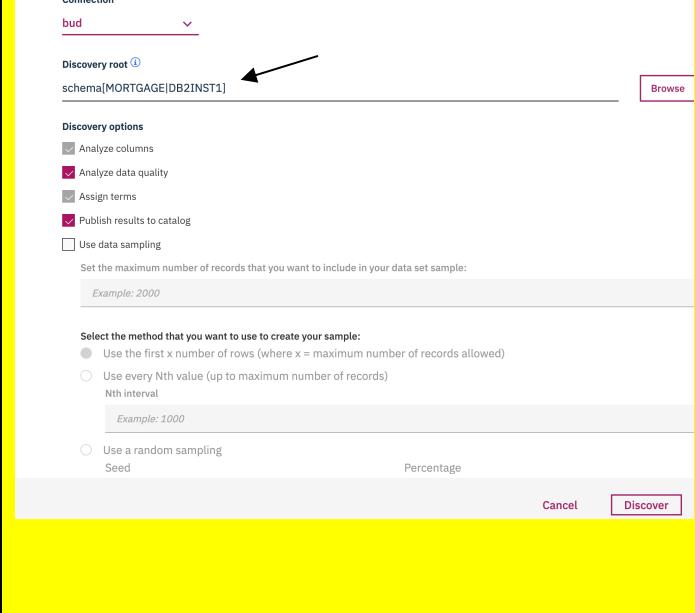
<p>Choose Organize > Data and AI governance > Rule from the left pane</p> <p>Select Published tab and click on New Rule</p> <p>Choose Create new rule</p> <p>Click on Governance rule</p>	<p>On the New governance rule window create a rule with following information and click on Save as draft:</p> <p>Name: Income cannot be null Description: Income column must have a valid value</p> <p>It will take few minutes to appear under list of available rules. Once the new rule is available, publish it.</p>
--	---

	<p>Click on Add policy under Parent policies to assign the Data Validation rule to it.</p> <p>Once the policy available, publish it.</p>
---	---

7.6. Automated Discovery

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

Choose **Organize > Curation > Data discovery** from the left pane.

	<p>To automated discover job</p> <p>Click on Automated discovery</p>
	<p>To discover assets</p> <p>Choose the connection named bud that you created previously.</p> <p>Select Discover root as MORTGAGE > DB2INST1</p> <p>Check necessary Discover options</p> <p>Select Workspace as Mortgage.</p> <p>Click on Discover</p> <p>Wait till import and analyze phase complete.</p>

7.7. Add rule to metadata

Go to **Organize > Information assets**

Search for Database Table name MORTGAGE_CUSTOMER

Assets **Hierarchies**

Explore information assets

Database Table **MORTGAGE_CUSTOMER**

MORTGAGE_CUSTOMER
jdbc:db2://52.117.27.156:50000/MORTGAGE > MORTGAGE > DB2INST1

MORTGAGE_CUSTOMER
jdbc:db2://52.117.27.156:50000/MORTGAGE > MORTGAGE > DB2INST1

Click on MORTGAGE_CUSTOMER data set

Workspaces Catalog data sets

Filter results Data sets

Search data set

Schema Host name Created by Created on Modified by

Data sets
7 results

Add to workspaces Cancel

Name	Quality score	Threshold	First imported	Last published	Terms	Workspaces
MORTGAGE_CUSTOMER			Dec 2, 2019, 12:14 PM			1
MORTGAGE_CUSTOMER	99%	80%	Dec 2, 2019, 2:34 PM	Dec 2, 2019, 2:37 PM	12	1

Database Table details
MORTGAGE_CUSTOMER

Governance Context: jdbc:db2://10.208.125.125:50000/MORTGAGE > db2 > DB2INST1

Database Columns (10)

Created by: admin Created on: 04 June 2019, 11:28:49 am Modified by: InformationServerSystemUser Modified on: 04 June 2019, 11:28:49 am

Database Columns

- APPLIED_ONLINE
- CARD_DEBT
- CURRENT_LOANS
- ID
- INCOME**

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**

The screenshot shows the 'Database Column details' page for the 'INCOME' column. On the left, there's a sidebar with 'Header (1)' selected, containing 'General Information', 'Quality Analysis', 'Suggested Term Assignments', and 'Notes'. The main area has two sections: 'Assigned to Terms' and 'Implements Rules'. In 'Assigned to Terms', there's a search bar with 'Add to list' and a 'Remove all' button. Below it, a message says 'You haven't added any item yet'. In 'Implements Rules', there's a search bar with 'inco' typed in, a clear button 'X', a magnifying glass icon, and a 'Remove all' button. A tooltip below the search bar says 'Income cannot be null'. At the top right are 'Cancel' and 'Save' buttons. A red arrow points from the 'Save' button to the 'Income cannot be null' tooltip.

The screenshot shows the user profile sidebar on the left. It includes a profile picture, the username 'dstw1', and a 'Profile and settings' link. Below that are links for 'About', 'Community', and 'Support'. At the bottom is a 'Log out' link. A red arrow points to the 'Log out' link. To the right of the sidebar, the text 'Log out from user 'dstw1'' is displayed.

8. Access data as a Data Scientist

Explore the data require for build a model

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

8.1. Assets from Glossary

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

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

Your catalogs

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

Click on each assets for additional information.

<input type="checkbox"/> Name	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/> MORTGAGE_CUSTOMER	admin	info_... DB2I...		Data asset	Apr 27, 2020
<input type="checkbox"/> MORTGAGE_DEFAULT	admin	info_... DB2I...		Data asset	Apr 27, 2020
<input type="checkbox"/> MORTGAGE_PROPERTY	admin	info_... DB2I...	Email Address	Data asset	Apr 27, 2020

8.2. Check Asset Details

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

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

ID	INCOME	APPLIED_ON...	RESIDEN...	YRS_CURRENT...	YRS_CURRENT...	NO_OF_CA...	CARD_DE...
100522	43982	Y	O	13	11	2	1055
101756	59944	Y	O	20	11	2	3894
101354	57718	Y	O	25	16	2	1555
100512	45621	Y	O	1	19	1	1878
100537	45081	N	O	14	15	2	713
100458	46645	N	O	19	4	1	884
101430	45066	Y	P	16	15	1	860
101432	44202	N	O	1	23	2	2611
100601	55215	Y	P	1	6	2	1930
101549	44460	Y	O	4	16	1	3467

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

Rating	(0)
5	(0)
4	(0)
3	(0)
2	(0)
1	(0)

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

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

- Header:** IBM Cloud Pak for Data, All, Search, Catalogs / Default Catalog / MORTGAGE_CUSTOMER.
- Top Bar:** DATA ASSET MORTGAGE_CUSTOMER, Remove, Download, Add to Project.
- Tab Navigation:** Overview, Access, Review, **Profile**, Lineage.
- Profile Summary:**
 - Current profile: 165 classifiers, Last profile: 27 Apr 2020 - 5:28 pm, View log.
 - Columns: 10, Rows: 419.
 - Metrics: Matches (green), Mismatches (purple), Missing (dark blue).
- Column Details:**

ID	Type: Integer	INCOME	Type: Integer	APPLIED_ONLINE	Type: Char	RESIDENCE	Type: Char
• Identifier	• Not classified	• Indicator	• Code				
0%	50%	0%	0%				
100%	50%	100%	100%				
0%	0%	0%	0%				
- Frequency:** Bar charts showing the distribution of values for ID, INCOME, APPLIED_ONLINE, and RESIDENCE.
- Statistics:** Summary of unique values for each column.

The screenshot shows the Add to Project dialog with the following details:

- Header:** IBM Cloud Pak for Data, All, Search.
- Title:** Add to Project.
- Target:** mortgage_data (selected from a dropdown).
- Selected assets (1):**

Asset Name	Catalog	Connection
MORTGAGE_CUSTOMER	Default Catalog	bud
- Buttons:** Cancel, Add.

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

The screenshot shows the user profile settings with the following details:

- User Profile:** dst1, Profile and settings.
- Navigation:** About, Community, Support.
- Action:** Log out (highlighted with a red arrow).

Log out from user ‘dst1’

9. Data Virtualization

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

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

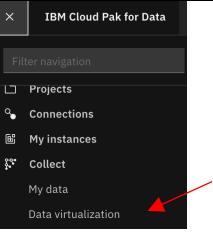
DV allows integrate data sources across multiple types and locations and turns it into one logical data view. In this case, you have data across three different tables. Creating a virtual table you can quickly view data from different tables.

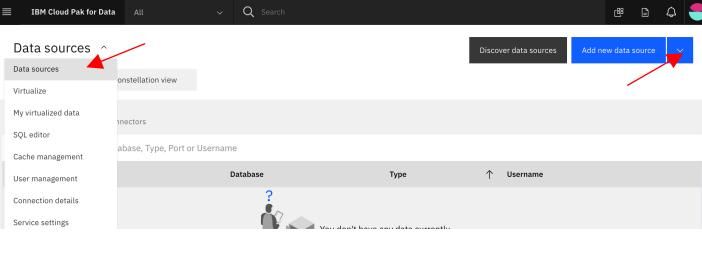
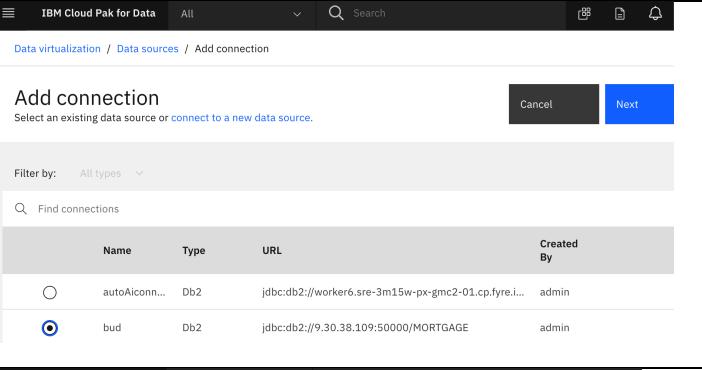
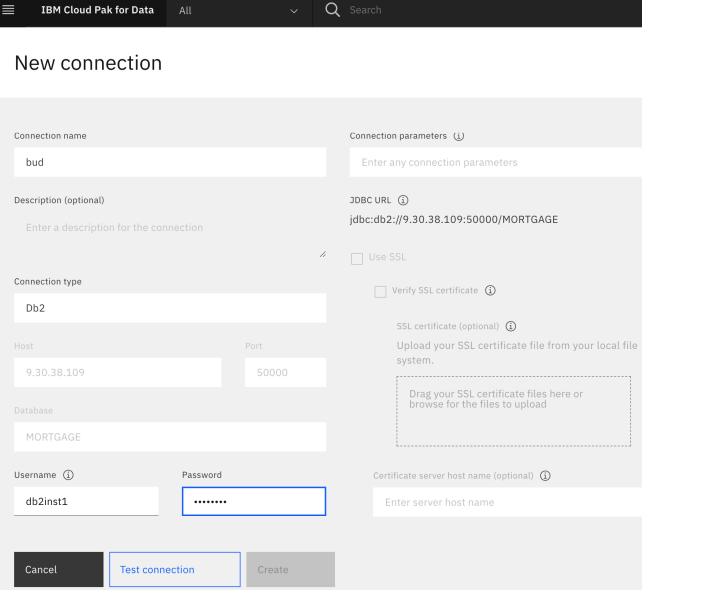
 <p>SIGN IN IBM Cloud Pak for Data</p> <p>Username: deng1 Password:</p> <p>Sign in →</p>	<p>Sigh into the Cloud Pak for Data web console as user ‘deng1’ and password is ‘deng1’ that you created earlier.</p>
---	---

9.1. Adding a new data source for Db2

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

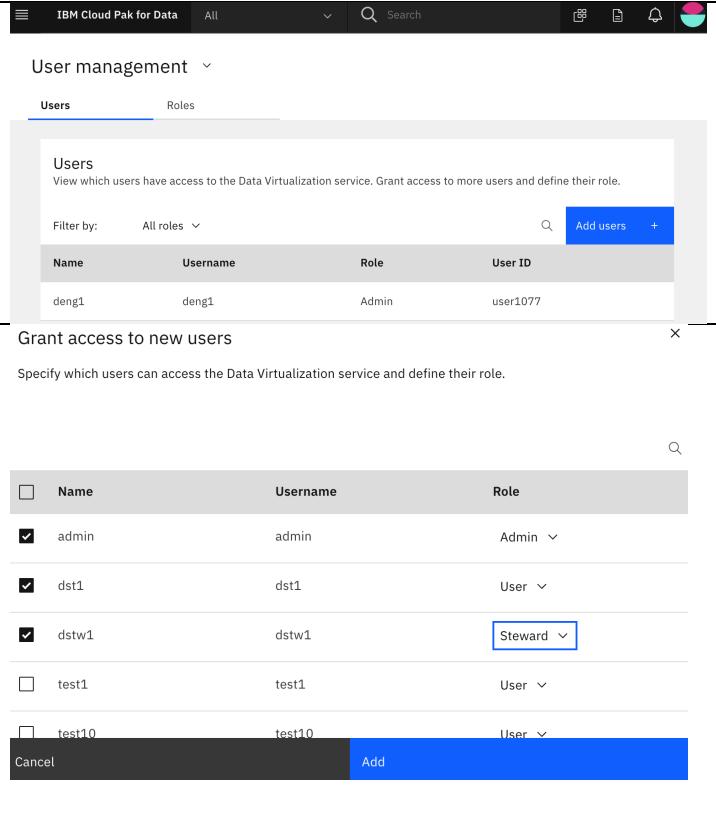
Define a data connection to Db2. Use your existing Db2 database connection for Db2 data source.

	<p>Go to Collect > Data Virtualization</p>
---	--

	<p>DV Menu > Data sources</p> <p>Click Add new data source > From existing connections</p>
	<p>Select bud that you created earlier</p> <p>Click Next</p>
	<p>Fill out the username and password information in the Add Connection</p> <p>Username is ‘db2inst1’ Password is ‘password’.</p> <p>Next click on Test Connection, once it successful click on Save Connection.</p>

9.2. Adding Users to Data Virtualization

Access to DV is administered by CPD user management functions. Access within DV is managed by Db2 authorizations. You need to provide necessary privileges to the other users so they can access virtual tables created on DV. The user **deng1** automatically granted admin role because it created the data source.

DV Menu > User management Click Add users	 <p>User management</p> <p>Users</p> <p>View which users have access to the Data Virtualization service. Grant access to more users and define their role.</p> <table border="1"> <thead> <tr> <th>Name</th> <th>Username</th> <th>Role</th> <th>User ID</th> </tr> </thead> <tbody> <tr> <td>deng1</td> <td>deng1</td> <td>Admin</td> <td>user1077</td> </tr> </tbody> </table> <p>Grant access to new users</p> <p>Specify which users can access the Data Virtualization service and define their role.</p> <table border="1"> <thead> <tr> <th><input type="checkbox"/></th> <th>Name</th> <th>Username</th> <th>Role</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/></td> <td>admin</td> <td>admin</td> <td>Admin</td> </tr> <tr> <td><input checked="" type="checkbox"/></td> <td>dst1</td> <td>dst1</td> <td>User</td> </tr> <tr> <td><input checked="" type="checkbox"/></td> <td>dstw1</td> <td>dstw1</td> <td>Steward</td> </tr> <tr> <td><input type="checkbox"/></td> <td>test1</td> <td>test1</td> <td>User</td> </tr> <tr> <td><input type="checkbox"/></td> <td>test10</td> <td>test10</td> <td>User</td> </tr> </tbody> </table> <p>Cancel Add</p>	Name	Username	Role	User ID	deng1	deng1	Admin	user1077	<input type="checkbox"/>	Name	Username	Role	<input checked="" type="checkbox"/>	admin	admin	Admin	<input checked="" type="checkbox"/>	dst1	dst1	User	<input checked="" type="checkbox"/>	dstw1	dstw1	Steward	<input type="checkbox"/>	test1	test1	User	<input type="checkbox"/>	test10	test10	User
Name	Username	Role	User ID																														
deng1	deng1	Admin	user1077																														
<input type="checkbox"/>	Name	Username	Role																														
<input checked="" type="checkbox"/>	admin	admin	Admin																														
<input checked="" type="checkbox"/>	dst1	dst1	User																														
<input checked="" type="checkbox"/>	dstw1	dstw1	Steward																														
<input type="checkbox"/>	test1	test1	User																														
<input type="checkbox"/>	test10	test10	User																														
Grant Admin role to user admin Grant User role to user dst1 Grant Steward role to user dstw1																																	
Click Add																																	

9.3. Select tables for virtualization

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

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

Table	Schema	Database	Hostname: Port	Columns
MORTGAGE_DEFAULT	DB2INST1	MORTGAGE	9.30.38.109:50000	2
MORTGAGE_PROPERTY	DB2INST1	MORTGAGE	9.30.38.109:50000	3
MORTGAGE_CUSTOMER	DB2INST1	MORTGAGE	9.30.38.109:50000	10
MORTGAGE_JOIN	DB2INST1	MORTGAGE	9.30.38.109:50000	12

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

Table	Schema	Source schema	Databases/File Path	Hostname: Port	Grouped tables	
MORTGAGE_CUSTOMER	USER1077	X	DB2INST1	MORTGAGE	9.30.38.109:50000	1
MORTGAGE_DEFAULT	USER1077	X	DB2INST1	MORTGAGE	9.30.38.109:50000	1
MORTGAGE_PROPERTY	USER1077	X	DB2INST1	MORTGAGE	9.30.38.109:50000	1

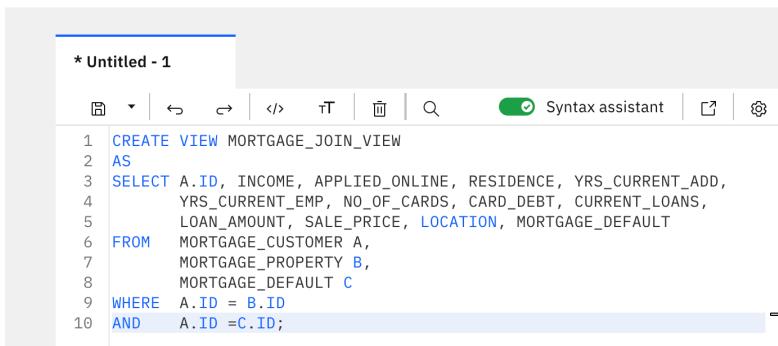
9.4. Creating Virtual Table

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 > Data virtualization > DV Menu > SQL editor > Create New** 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;
```

SQL editor ▾



```
* Untitled - 1
CREATE VIEW MORTGAGE_JOIN_VIEW
AS
SELECT A.ID, INCOME, APPLIED_ONLINE, RESIDENCE, YRS_CURRENT_ADD,
       YRS_CURRENT_EMP, NO_OF_CARDS, CARD_DEBT, CURRENT_LOANS,
       LOAN_AMOUNT, SALE_PRICE, LOCATION, MORTGAGE_DEFAULT
FROM   MORTGAGE_CUSTOMER A,
       MORTGAGE_PROPERTY B,
       MORTGAGE_DEFAULT C
WHERE  A.ID = B.ID
AND    A.ID = C.ID;
```

- Click **Collect > Data virtualization > DV 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 manage access window select **All data virtualization users**
- Click **Grant access to all**



Manage Access: MORTGAGE_JOIN_VIEW

Grant access to
 Specific users All data virtualization users

[Users \(0\)](#) [Roles \(1\)](#)

9.5. Add virtual table to catalog

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

<ul style="list-style-type: none"> • Click Collect > Data virtualization > DV Menu > My virtualized data to find the virtual table just created. • Mark the checkbox associated with virtual table • Chose Submit to catalog from table action • Click on Confirm 	
--	--

9.6. Publish virtualized table

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

<ul style="list-style-type: none"> • Click on access the Home page • Check Requests section under Overview • Click on Pending Publish to Catalog Requests • Click on check mark icon on left for virtual table MORTGAGE_JOIN_VIEW that you created • Click on Approve 	
---	--

9.7. Access information for virtual table

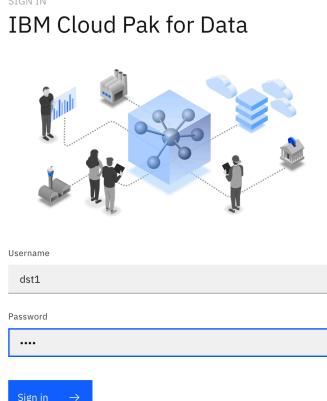
To access virtual table from external application, you need the JDBC connection information. Click on **Collect > Data Virtualization > DV Menu > Service settings** to find out access information. You will use this information later in the building model section.

The screenshot shows the 'Service settings' page for the 'data-virtualization' instance. It has two main sections: 'Access information' and 'About this instance'. In the 'Access information' section, there are three rows: 'User ID' (user1077), 'Password' (redacted), and 'JDBC connection URL' (jdbc:db2://dv-server.zen.svc.cluster.local:32051/bigsql). The 'About this instance' section contains four rows: 'Service name' (data-virtualization), 'Service version' (1.4.0.0), 'Created on' (Apr 29, 2020), and a 'Copy' button for the service name.

The screenshot shows the user profile menu on the left side of the screen. It includes a profile picture for 'deng1', a 'Profile and settings' link, and a 'Log out' link at the bottom. A red arrow points to the 'Log out' link. To the right of the menu, the text 'Sign out from user deng1' is displayed.

10. Build Model

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

 <p>SIGN IN</p> <p>IBM Cloud Pak for Data</p> <p>Username dst1</p> <p>Password</p> <p>Sign in →</p>	<p>Sign into the Cloud Pak for Data web console as user ‘dst1’ and password is ‘dst1’ that you created earlier.</p>
--	---

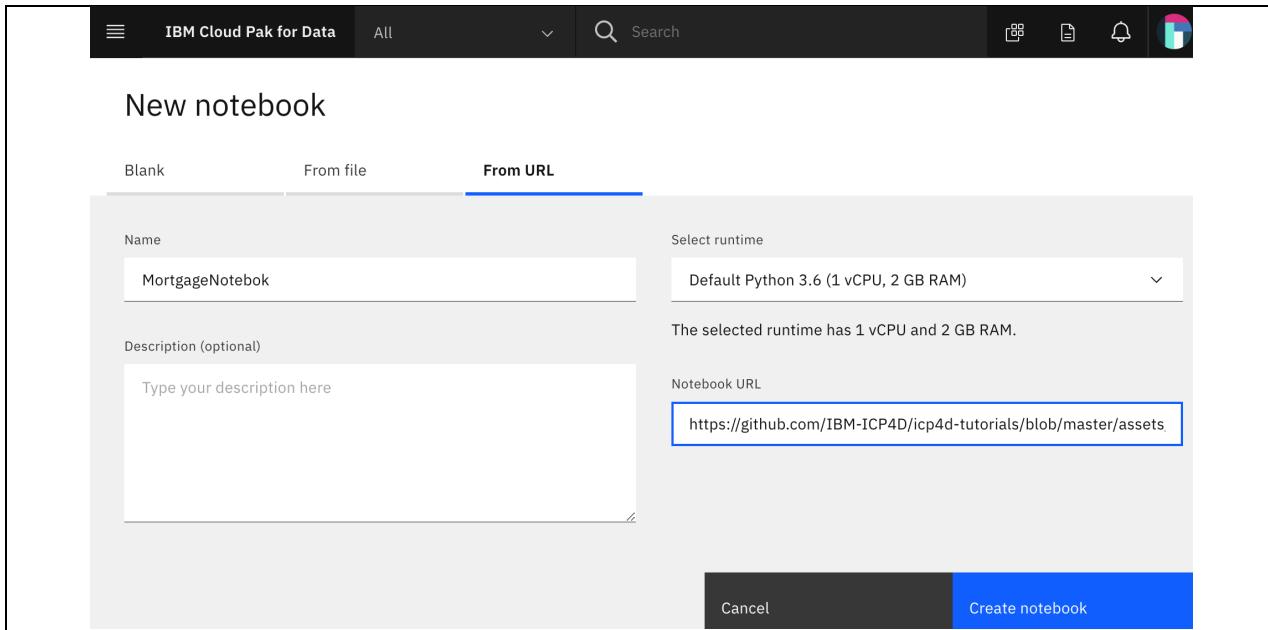
10.1. Navigate to analytics project

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

10.2. Create notebook

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

- Go to : My Projects > **mortgage_data** > Add to project
- Choose asset type as Notebook
- Choose **From URL** tab
- 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**



10.3. 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.

If it's not already 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: Install** section first. Once all package installed make sure restart the Python kernel before move on next step.

```

In [1]: # Check Python version
import platform
print(platform.python_version())
3.6.10

In [2]: # Uninstall the older Watson Machine Learning client
!pip uninstall watson-machine-learning-client -y

# Install the WML client
!pip install watson-machine-learning-client-v4

# Verify WLM Client version
!pip list | grep watson

```

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?#@1T_674MfPI5" # e.g.
```

Run all cells between step 2 and 6.

You may need to change the MORTGAGE_JOIN_VIEW schema name in step 3, according to your environment.

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

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

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



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

Before exit, save the notebook .

10.4. 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

MORTGAGE PREDICTION Deployed Online

Enter input data

Body

```
{ "input_data": [] }
```

Predict

MORTGAGE PREDICTION

Created Apr 30, 2020 6:21 PM

Updated Apr 30, 2020 6:21 PM

Deployment ID c9814de4-c300-4e2c-9da8-0b03...

Nodes 1

Description No description provided.

Associated asset MORTGAGE PREDICTION MODEL c7ef58e7-541e-4493-b915-9c4fd...

```
{
  "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 IBM Cloud Pak for Data interface with the 'Test' tab selected for the 'MORTGAGE PREDICTION' model. On the left, the 'Enter input data' section contains a JSON body with the following data:

```

    "Body": [
      1,
      8600,
      320000,
      110
    ]
  }
}

```

Below this is a 'Predict' button. To the right, the 'Result' section displays the predicted fields:

```

 0 {
 1   "predictions": [
 2     {
 3       "fields": [
 4         "INCOME",
 5         "APPLIED_ONLINE",
 6         "RESIDENCE",
 7         "YRS_CURRENT_ADD",
 8         "YRS_CURRENT_EMP",
 9         "NO_OF_CARDS",
 10        "CARD_DEBT",
 11        "CURRENT_LOANS",
 12        "LOAN_AMOUNT",
 13        "SALE_PRICE",
 14        "LOCATION",
 15        "PREDICTION"
 16      ]
 17    }
 18  ]
 19 }

```

On the far right, there is a detailed view of the model's metadata, including its creation date (Apr 30, 2020 6:21 PM), deployment ID (c9814de4-c300-4e2c-9da8-0b03...), and a single node entry.