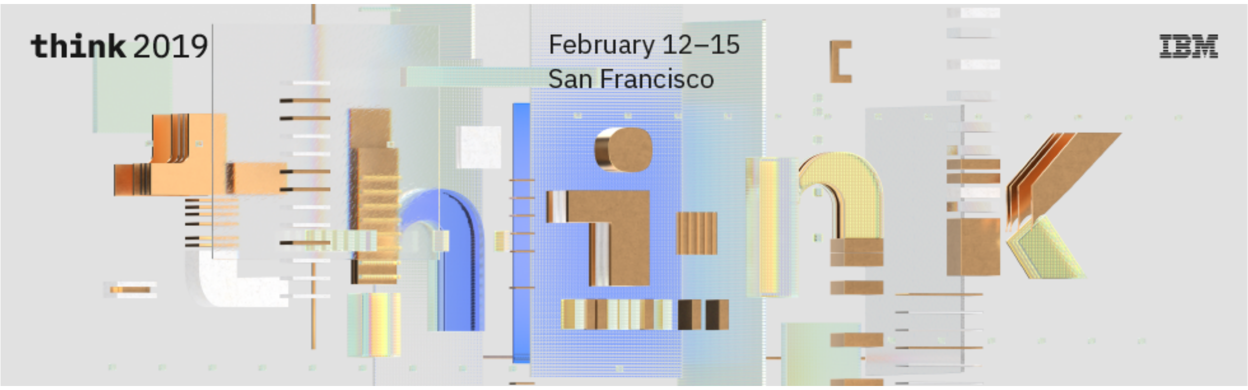


Lab Center – Hands-on Lab

Session 5093
Design a Predictive Model for Self-Serve Analytics with IBM Cloud Private for Data

Sanjit Chakraborty, IBM
Pradeep Kutty, IBM

Table of Contents

Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.
Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.**
IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
IBM products are manufactured from new parts or new and used parts.
In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.
Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Introduction

In this Hand-on Lab (HoL) you will learn how to use IBM® Cloud Private for Data (ICP for Data) to discover, governing, transform, and analyze data. While doing so, you'll get a guided tour of this robust end-to-end solution for all of the data and analytic needs within your enterprise.

You will use two different data sources on Db2 and Informix, which includes sample records of mortgage customer and property. This HoL uses a preconfigured analysis model that you can use to run a basic machine learning simulation. The Mortgage Default model predicts whether or not customers are likely to default on their mortgage loan.

This HoL will guide you how to:

- Create and test a connection between Db2, Informix and ICP for Data
- Add user and role for managing data
- Prompt IBM Cloud Private for Data to discover data assets in its catalog
- Browse and search for data assets
- Implement policies and rules for data governance
- Create, compile, and run a data transformation project and job
- Create an analytics project
- Work with a Jupyter notebook and connect it to the sample database
- Run the notebook and view the results of your analysis
- Use of data virtualization

Through this HoL, **<#>** sign used, please replace it with an unique number to maintain concurrency.

1. Access Credentials

To work through the HoL, you will use Db2 and Informix databases.

1.1. Access credential for Db2 database

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

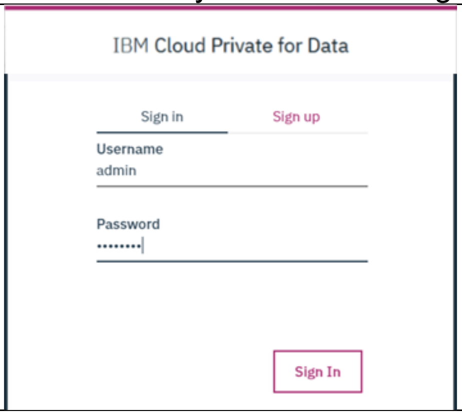
1.2. Access credential for Informix database

JDBC connection credential for Informix:

JDBC Host name	<Same IP address as your web console>
Port number	9088
Database name	MORTGAGEDB
User ID	informix
Password	in4mix
Informix	Version 12.10.FC12W1DE
JDBC connection string	jdbc:informix-sqli://<same IP as Web console>:9088/mortgagedb: INFORMIXSERVER=informix;user=informixt;password=in4mix

1.3. Sign in to ICP for Data web console as Administrator

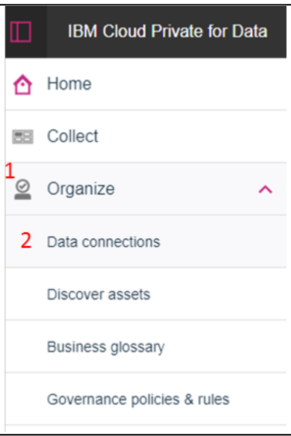
You should use latest version of Firefox or Google Chrome browser to access the ICP for Data web console. Starting from here all instruction needs to execute on ICP for Data web console only. You need to login as admin who has administrator privileges.

 IBM Cloud Private for Data Username admin Password •••••••• 	Sigh in to the ICPD web console as user 'admin' and password is 'password.

2. Create Connection

A data connection allows you to browse through your enterprise data. Create a connection to the data source for Db2 database. For simplicity, let's start with single database. You will add connection to Informix database later.

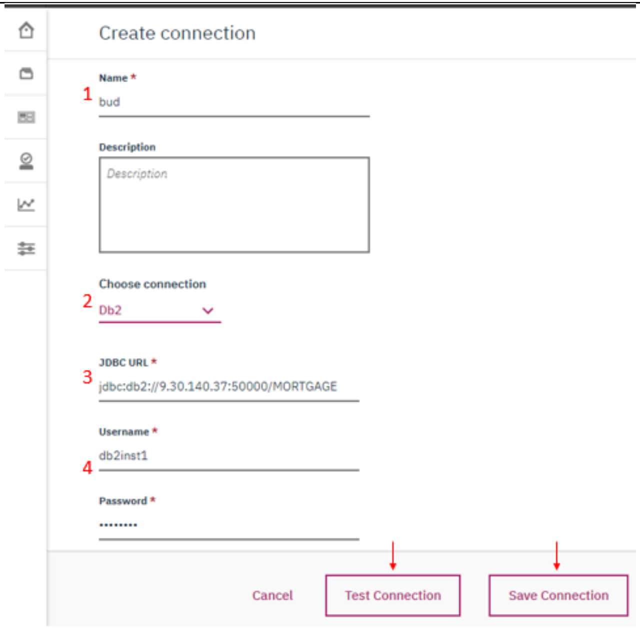
2.1. Navigate to data connection



From **Organize** option on the left pane choose **Data connections**.

Next, on the **Data Connections** window click on the ⊕ Create Connections icon.

2.2. Create connection


Create connection

1 Name * bud

Description

2 Choose connection Db2

3 JDBC URL * jdbc:db2://9.30.140.37:50000/MORTGAGE

4 Username * db2inst1

Password * ••••••••

Fill out the **Create Connection** information according to the information provided in step '2.1. Access credential for DB2. Credential used in following step is just an example.

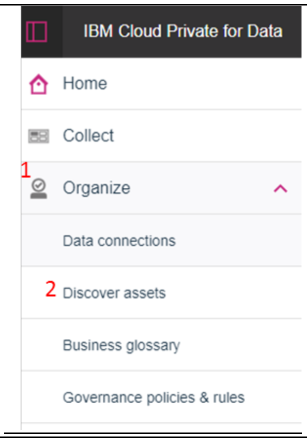
1. For **Choose connection** use the drop-down menu and select 'Db2'.
2. Use 'Bud<#>' as the **Name**
3. **JDBC URL** is 'jdbc:db2://<IP address of master node-1>:50000/MORTGAGE'
4. **Username** is 'db2inst1' and **Password** is 'password'.

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

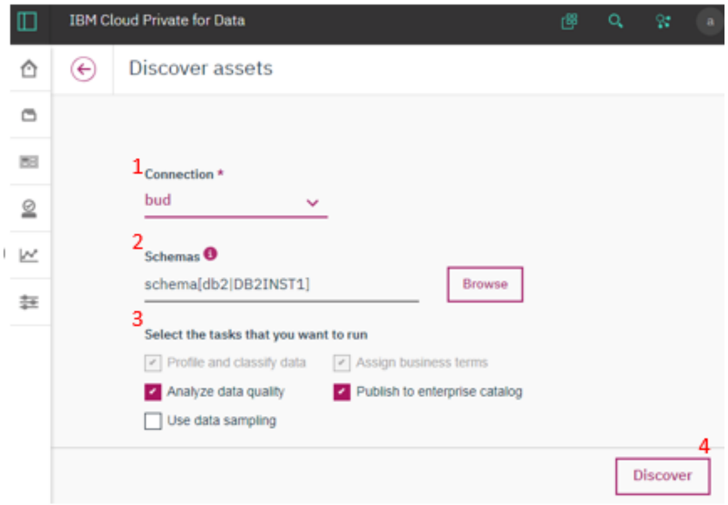
3. Discover Assets

The discover assets enables you to catalog data from data sources to make it easier to search for, govern, and analyze data. Use the data source created above to discover all data assets from Db2 database.

3.1. Navigate to discover assets



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



Connection * bud

Schemas schema[db2|DB2INST1]

Select the tasks that you want to run

☑ Profile and classify data ☑ Assign business terms

☑ Analyze data quality ☑ Publish to enterprise catalog

□ Use data sampling

To discover assets

1. Choose the connection named ‘bud<#>’ from the dropdown menu that you created in previously.

2. Click on **Browse** and select schema **DB2INST1** under **db2**

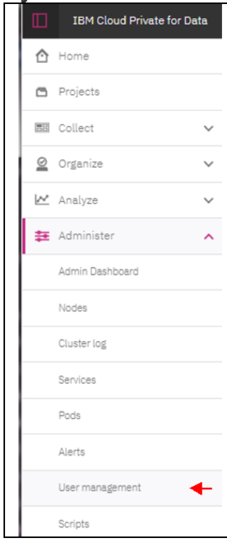
3. Check all boxes under the ‘Select the tasks that you want to run’.

4. Click on **Discover**

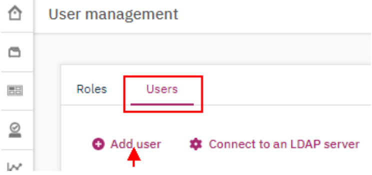
It may take few minutes to complete.

4. Add users

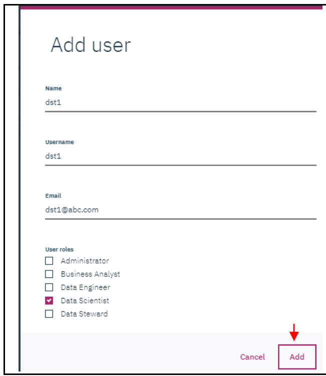
ICP for Data includes several predefined roles with different permissions for different business need. You will create users with these roles to get a feel of ICP for Data eco-system.



From **Administer** option on the left pane, choose **User management**.



Switch tab to ‘Users’ and click on ‘Add user’



Add user

Name
dst1

Username
dst1

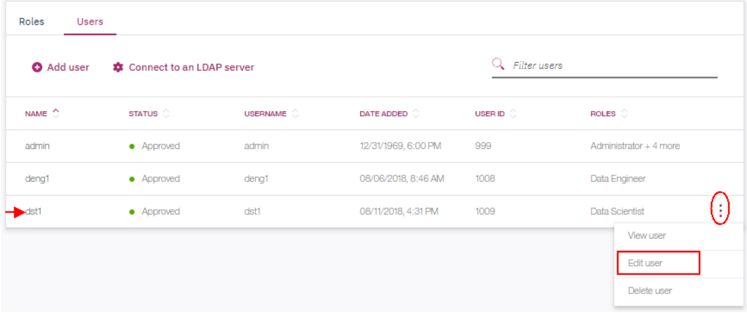
Email
dst1@abc.com

User roles
☐ Administrator
☐ Business Analyst
☐ Data Engineer
☑ Data Scientist
☐ Data Steward

Fill out Add User information for a data scientist

1. ‘Name’ as **dst<#>**
2. Username is **dst<#>**
3. Use a valid email address
4. Chose the user roles as Data Scientist

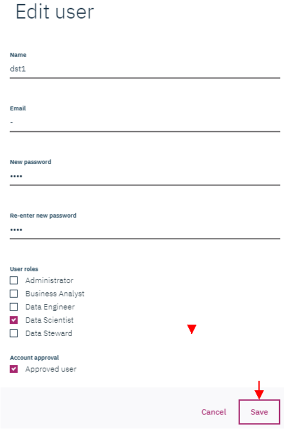
Click on **Add** to confirm the add user



NAME	STATUS	USERNAME	DATE ADDED	USER ID	ROLES
admin	● Approved	admin	12/31/1969, 6:00 PM	999	Administrator + 4 more
deng1	● Approved	deng1	08/06/2018, 8:46 AM	1008	Data Engineer
dst1	● Approved	dst1	08/11/2018, 4:31 PM	1009	Data Scientist

Before hand over user, change the password.

1. Access dst<#> user setting by click on ⋮ icon
2. Choose ‘Edit user’



Edit user

Name: dst1

Email: -

New password: ••••

Re-enter new password: ••••

User roles
☐ Administrator
☐ Business Analyst
☐ Data Engineer
☑ Data Scientist
☐ Data Steward

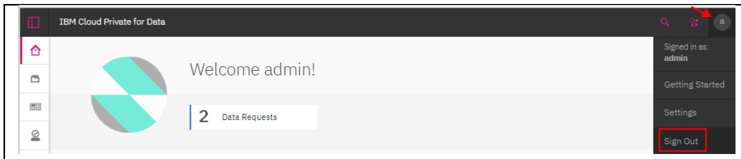
Account approval
☑ Approved user

1. Type password as **dst<#>** 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 **deng<#>** for Data Engineer and **dstw<#>** a data steward.

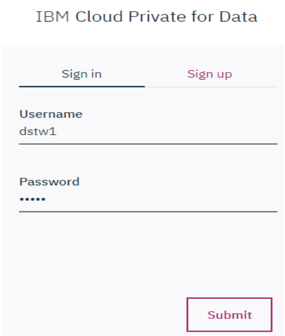
User	Role	Password
• Deng<#>	Data Engineer	deng<#>
• Dstw<#>	Data Stewards	dstw<#>



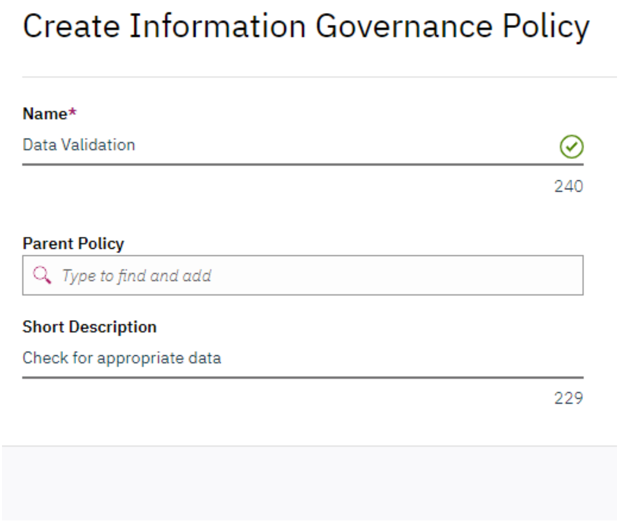
Sign out from user **admin**

5. Implement Policies and Rules

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

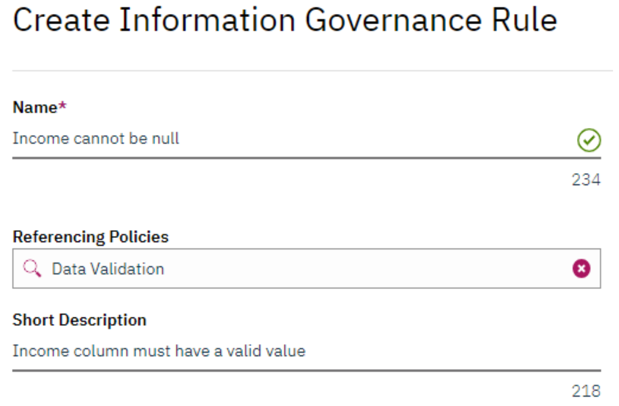
 Username dstw1 Password •••••	Sigh in to the ICPD web console as user ‘dstw<#>’ and password is ‘dstw<#>’ that you created earlier.

5.1. Create a policy

Choose **Organize** from the left pane, then select **Governance policies and rules**. Select **Polices** tab and click on **Create Policy**	
 Name* Data Validation 240 Parent Policy Type to find and add Short Description Check for appropriate data 229	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.

5.2. Create a rule

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



Create Information Governance Rule

Name*
Income cannot be null
234

Referencing Policies
Data Validation

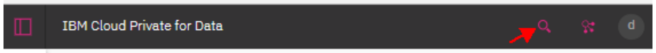
Short Description
Income column must have a valid value
218

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

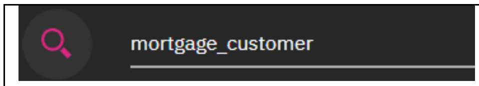
Name: Income cannot be null
Referencing policies: Data Validation
Short Description: Income column must have a valid value

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

5.3. Add rule to metadata

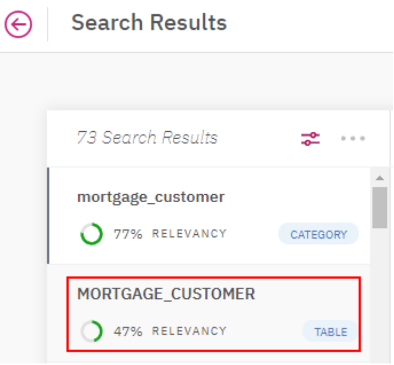


Click on the enterprise search



mortgage_customer

Search for ‘mortgage_customer’ and hit enter



Search Results

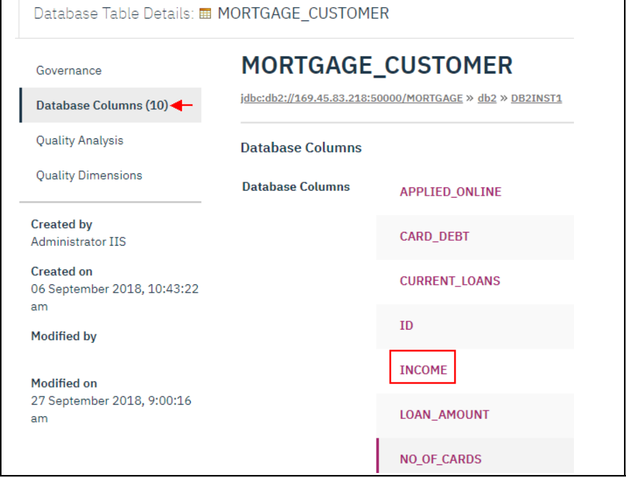
73 Search Results

mortgage_customer
77% RELEVANCY CATEGORY

MORTGAGE_CUSTOMER
47% RELEVANCY TABLE

From the search results select table ‘mortgage_customer’

Click on **Details** tab at the top



Database Table Details: MORTGAGE_CUSTOMER

Governance
Database Columns (10)
Quality Analysis
Quality Dimensions

Created by
Administrator IIS

Created on
06 September 2018, 10:43:22 am

Modified by

Modified on
27 September 2018, 9:00:16 am

MORTGAGE_CUSTOMER

jdbc:db2://169.45.83.218:50000/MORTGAGE » db2 » DB2INST1

Database Columns

Database Columns
APPLIED_ONLINE
CARD_DEBT
CURRENT_LOANS
ID
INCOME
LOAN_AMOUNT
NO_OF_CARDS

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

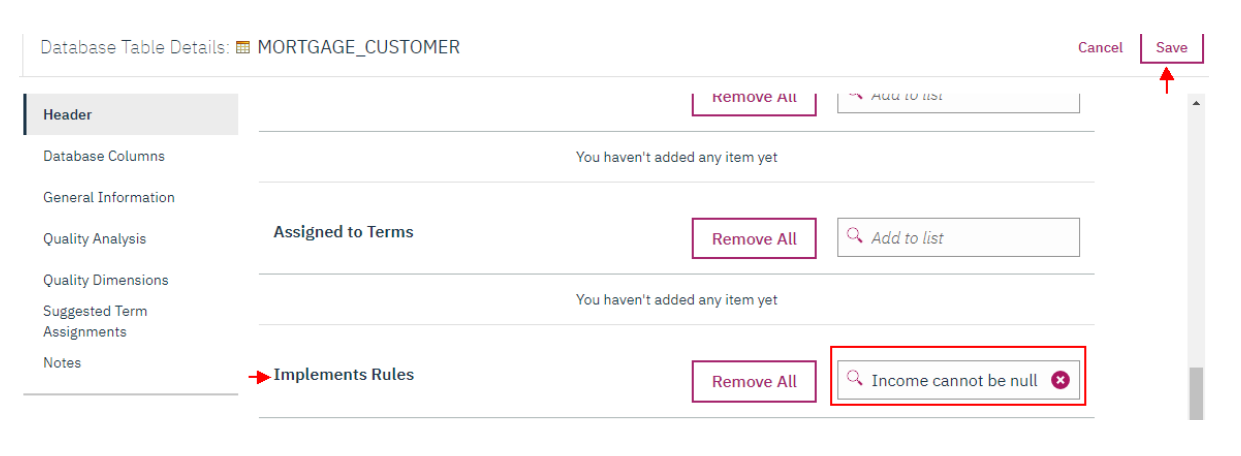
Select INCOME column

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**



Database Table Details: MORTGAGE_CUSTOMER

Cancel Save

Header
Database Columns
General Information
Quality Analysis
Quality Dimensions
Suggested Term Assignments
Notes

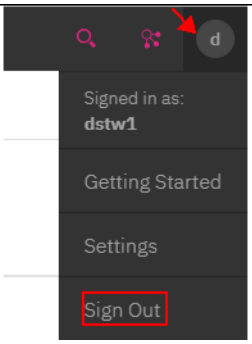
Remove All Add to list

You haven't added any item yet

Assigned to Terms Remove All Add to list

You haven't added any item yet

Implements Rules Remove All Income cannot be null



d

Signed in as:
dstw1

Getting Started

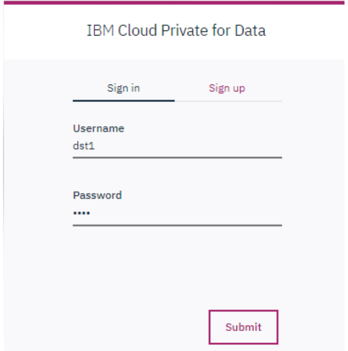
Settings

Sign Out

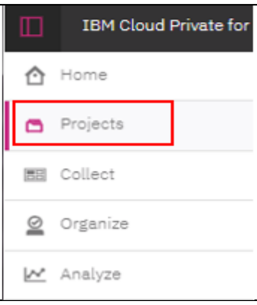
Sign out from user ‘dstw<#>’

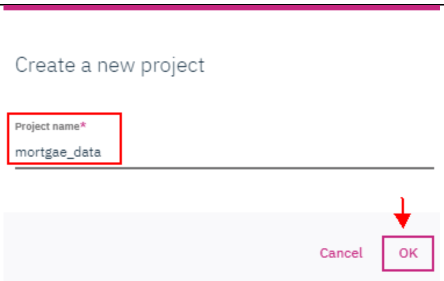
6. Access data as a Data Scientist

Explore the data require for build a model

 Username dst1 Password ••••	Sigh in to the ICPD web console as user ‘dst<#>’ and password is ‘dst<#>’ that you created earlier.

6.1. Create analytic project

 IBM Cloud Private for Home Projects Collect Organize Analyze	Create a new analytical project by ‘Projects’ from right pane. Click on the ⊕ New project icon

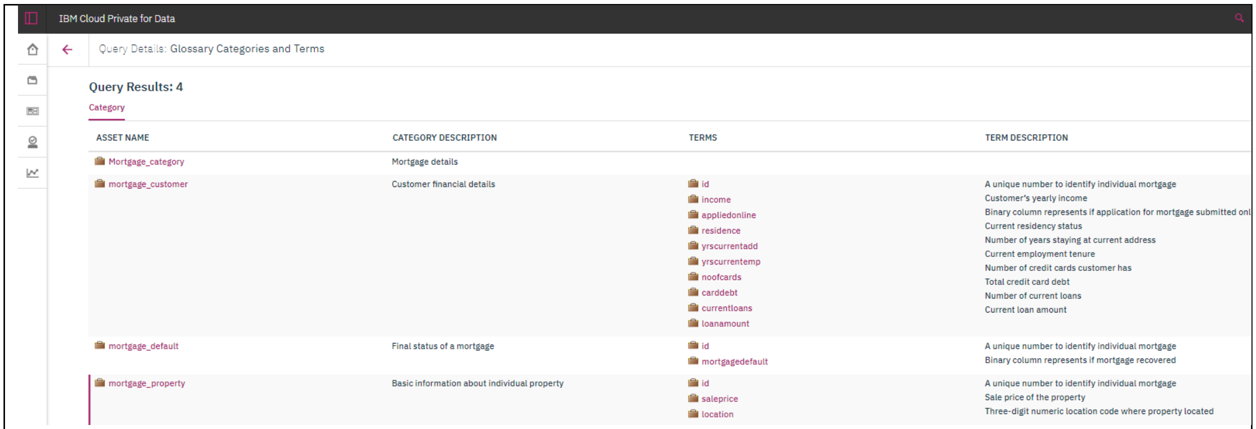
 Project name* mortgae_data	Provide a project name as ‘mortgage_data<#>’ and click **OK** On the next ‘Create project’ window, click on **Create**

6.2. 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.



Query Details: Glossary Categories and Terms

Query Results: 4

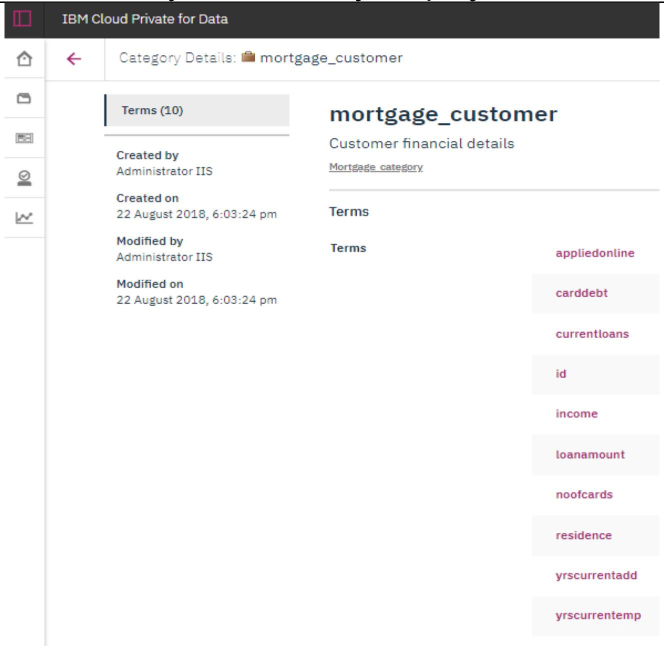
Category

ASSET NAME	CATEGORY DESCRIPTION	TERMS	TERM DESCRIPTION
Mortgage_category	Mortgage details		
mortgage_customer	Customer financial details	id income appliedonline residence yrscurrentadd yrscurrentemp noofcards carddebt currentloans loanamount	A unique number to identify individual mortgage Customer's yearly income Binary column represents if application for mortgage submitted onl Current residency status Number of years staying at current address Current employment tenure Number of credit cards customer has Total credit card debt Number of current loans Current loan amount
mortgage_default	Final status of a mortgage	id mortgagedefault	A unique number to identify individual mortgage Binary column represents if mortgage recovered
mortgage_property	Basic information about individual property	id saleprice location	A unique number to identify individual mortgage Sale price of the property Three-digit numeric location code where property located

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

6.3. Check Asset Details

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



Category Details: mortgage_customer

Terms (10)

Created by
Administrator IIS

Created on
22 August 2018, 6:03:24 pm

Modified by
Administrator IIS

Modified on
22 August 2018, 6:03:24 pm

mortgage_customer

Customer financial details

Mortgage_category

Terms

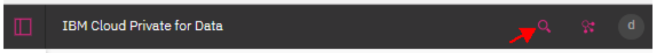
Terms

appliedonline
carddebt
currentloans
id
income
loanamount
noofcards
residence
yrscurrentadd
yrscurrentemp

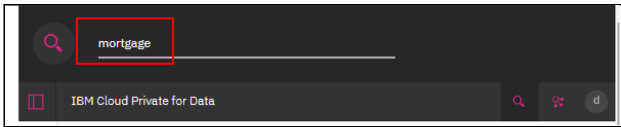
The asset **mortgage_customer** shows different terms associated with it.

Check each **Terms** for additional information.

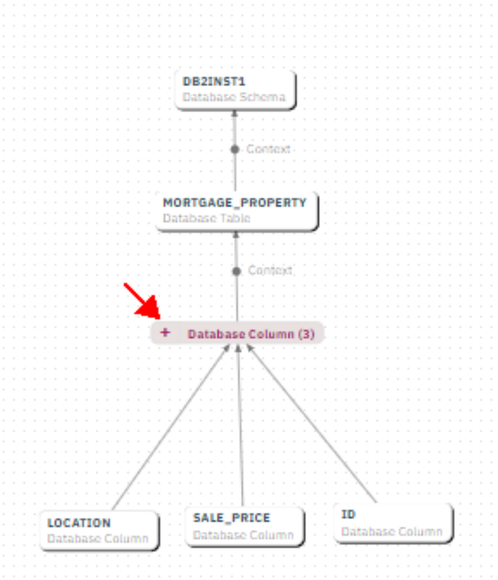
6.4. Enterprise search



Click on the enterprise search



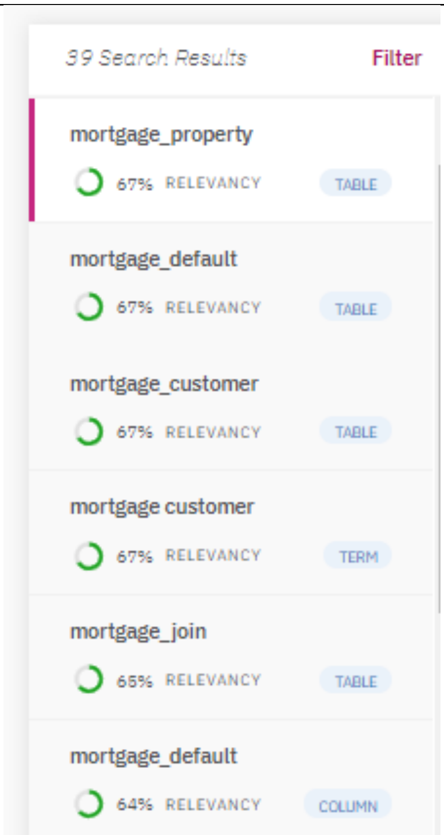
Search for ‘mortgage’ and hit enter



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

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

Same way you can view other mortgage related tables.



39 Search Results

Filter

mortgage_property
67% RELEVANCY TABLE

mortgage_default
67% RELEVANCY TABLE

mortgage_customer
67% RELEVANCY TABLE

mortgage customer
67% RELEVANCY TERM

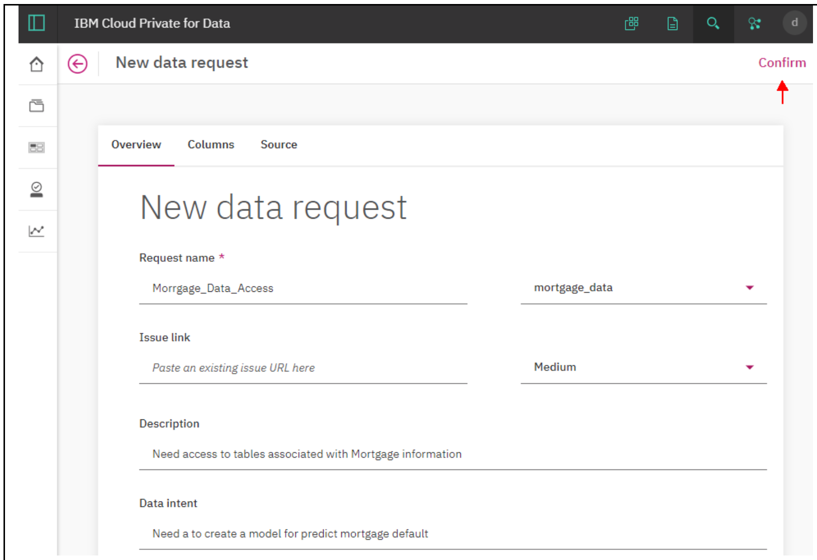
mortgage_join
65% RELEVANCY TABLE

mortgage_default
64% RELEVANCY COLUMN

Go back to the enterprise **Search Result**

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

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



New data request

Request name *
Morrgage_Data_Access
mortgage_data

Issue link
Paste an existing issue URL here
Medium

Description
Need access to tables associated with Mortgage information

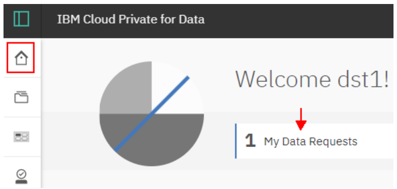
Data intent
Need a to create a model for predict mortgage default

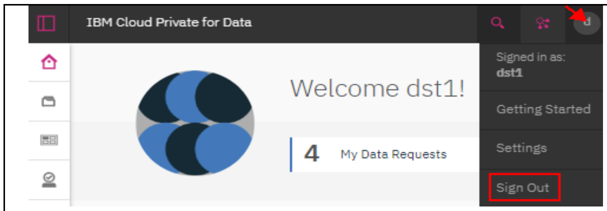
Fill up the **New Data Request** form with detail information as much possible, so a data engineer can provide accurate dataset.

Make sure you choose the right project that you created earlier.

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

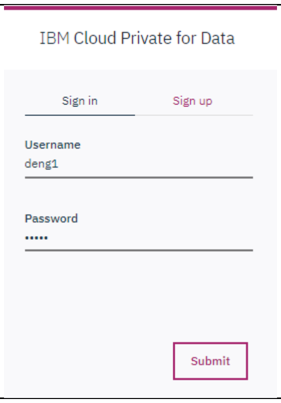
You can go to the home page by clicking on ⌂ icon from left pane and check the status of the data request.





Sign out from user **dst<#>**

7. Review data request



IBM Cloud Private for Data

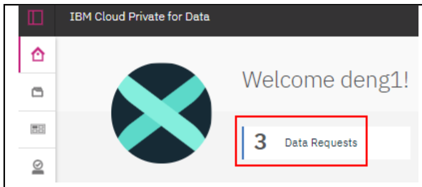
Sign in   Sign up

Username
deng1

Password
•••••

Submit

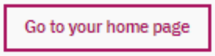
Sigh in to the ICPD web console as user ‘deng<#>’ and password is ‘deng<#>’ that you created earlier.



IBM Cloud Private for Data

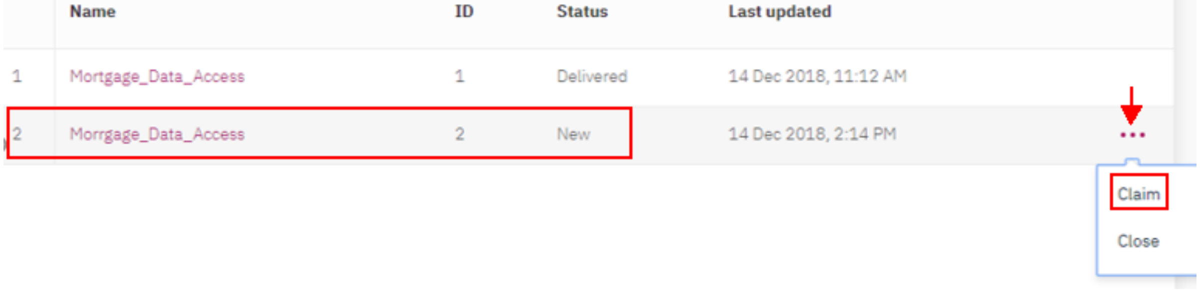
Welcome deng1!

3 Data Requests

After sing in Click on  Go to your home page

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

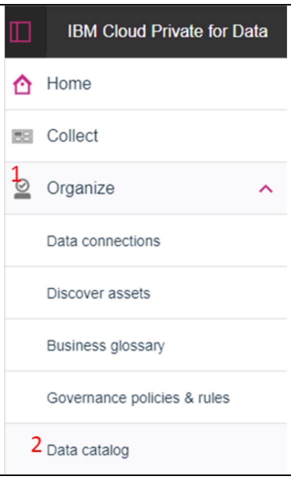
Click on the new data request that submitted by data scientist earlier for review.
After reviewing the request Claim the request.



	Name	ID	Status	Last updated
1	Mortgage_Data_Access	1	Delivered	14 Dec 2018, 11:12 AM
2	Morrgage_Data_Access	2	New	14 Dec 2018, 2:14 PM

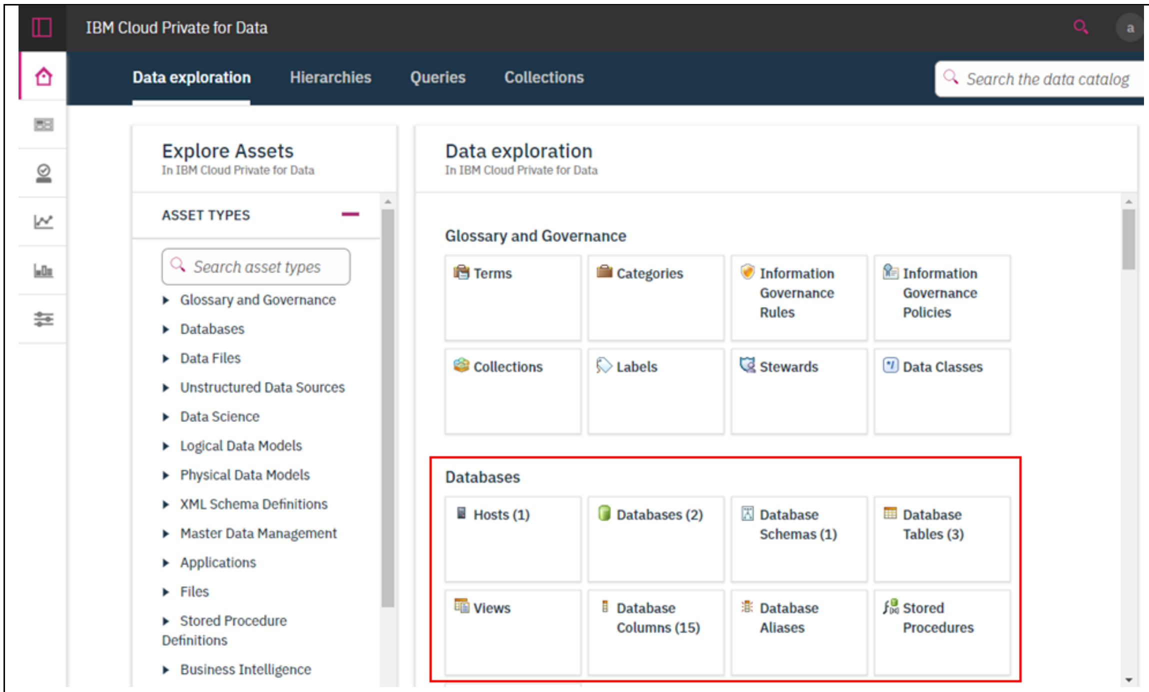
8. Navigate to data catalog

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

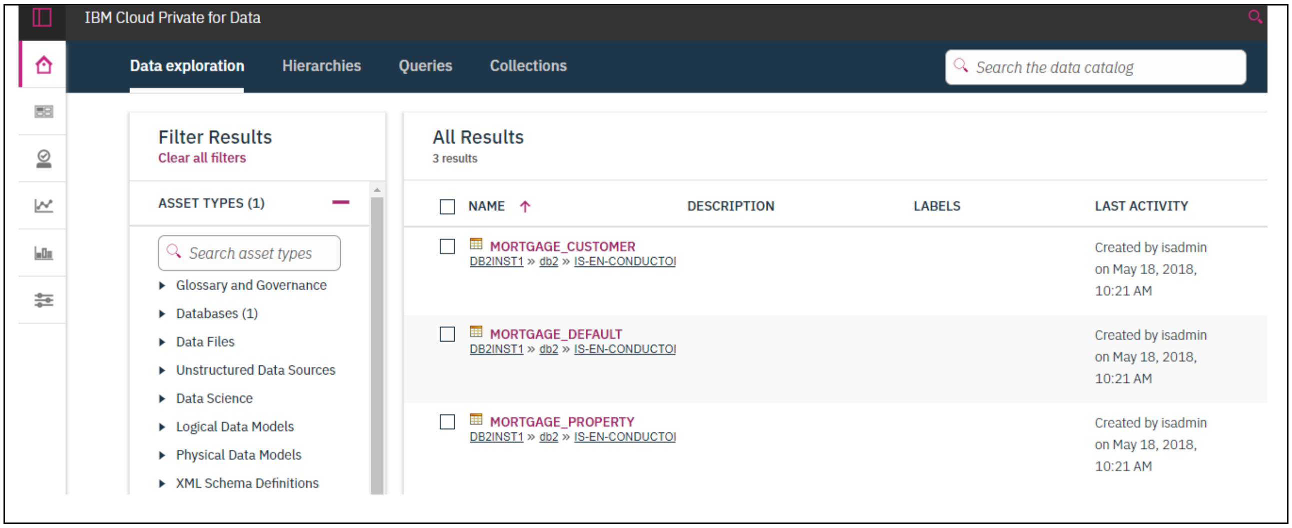


Next go back to **Organize** option on the left pane and choose **Data catalog**.

You can click each individual object under **Database** 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.

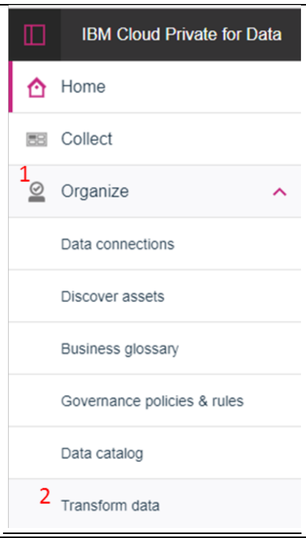


All Results
3 results

□	NAME ↑	DESCRIPTION	LABELS	LAST ACTIVITY
□	MORTGAGE_CUSTOMER DB2INST1 » db2 » IS-EN-CONDUCTOI			Created by isadmin on May 18, 2018, 10:21 AM
□	MORTGAGE_DEFAULT DB2INST1 » db2 » IS-EN-CONDUCTOI			Created by isadmin on May 18, 2018, 10:21 AM
□	MORTGAGE_PROPERTY DB2INST1 » db2 » IS-EN-CONDUCTOI			Created by isadmin on May 18, 2018, 10:21 AM

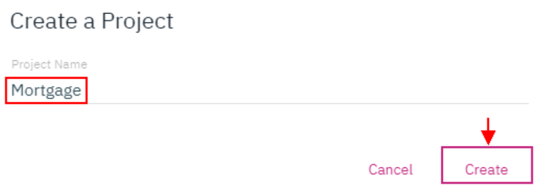
9. Transform Data

Transform data provides enriched and tailored information for your enterprise. You can create, edit, load, and run transformation ETL jobs from ICP for Data.



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

9.1. Create a Project


Create a Project

Project Name: Mortgage

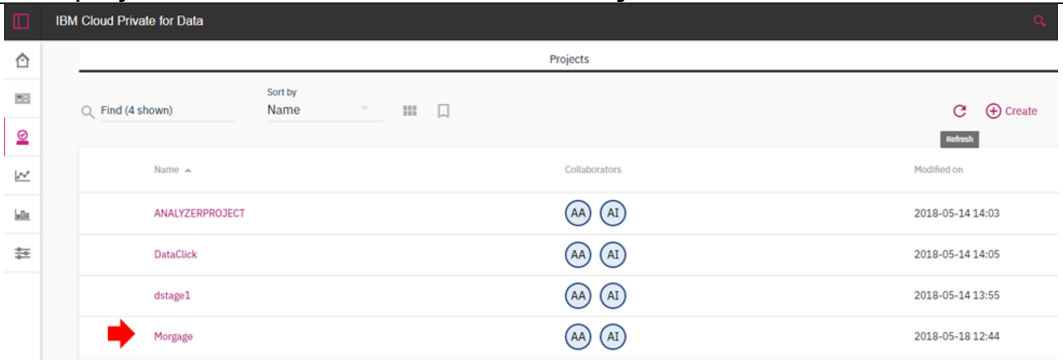
Next create a project by clicking on ⊕ Create icon on top left corner.

On **Create a Project** window use the **Project Name** as ‘Mortgage<#>’.

Click on **Create**

It may take few minutes to complete.

Once project created it will list under the **Projects.**

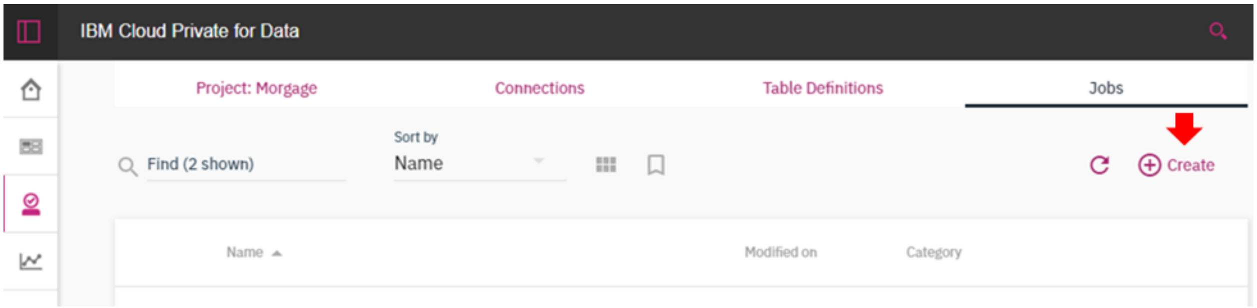


Name	Collaborators	Modified on
ANALYZERPROJECT	AA AI	2018-05-14 14:03
DataClick	AA AI	2018-05-14 14:05
dstage1	AA AI	2018-05-14 13:55
Morgage	AA AI	2018-05-18 12:44

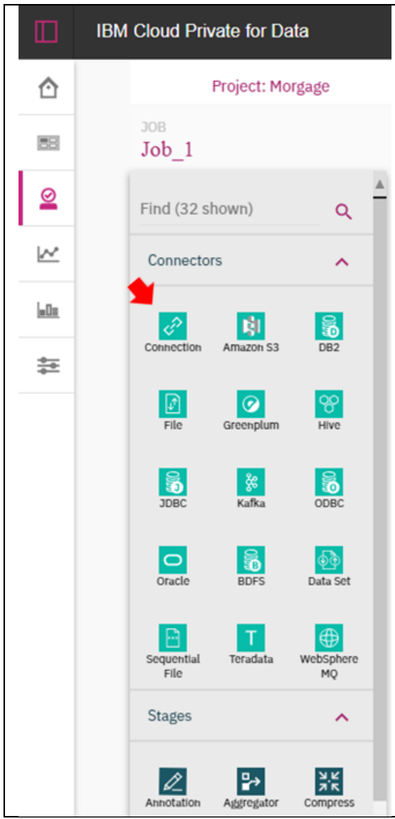
Click on the project Name **Mortgage<#>**

9.2. Create a job

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

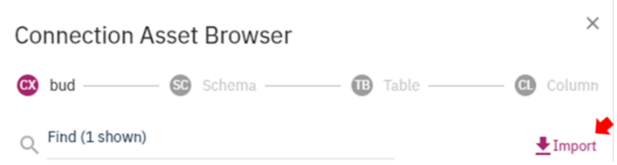


9.3. Add tables from asset browser



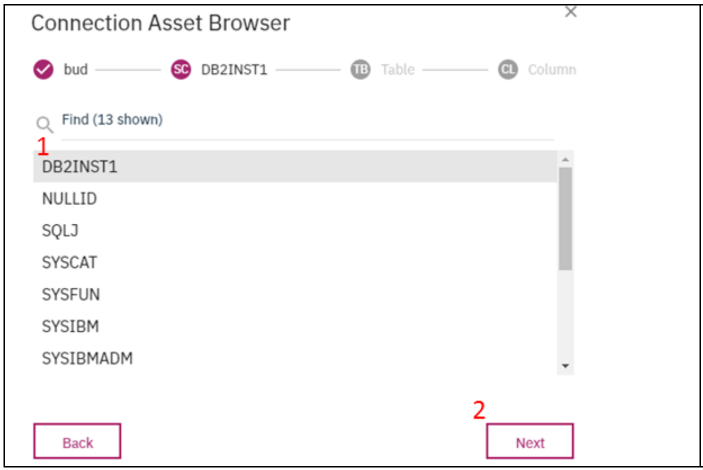
The create job operation will open a palette on the left.

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

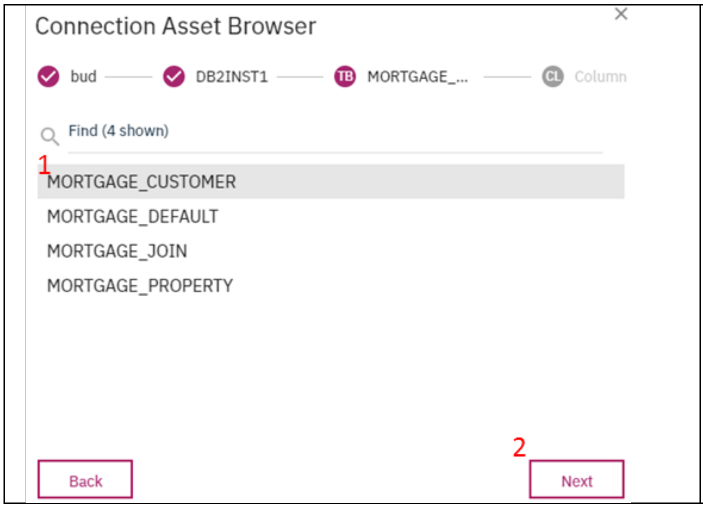


On the **Connection Asset Browser** window,
Click on the **Import** to use the connection that you created earlier on step 4.2.

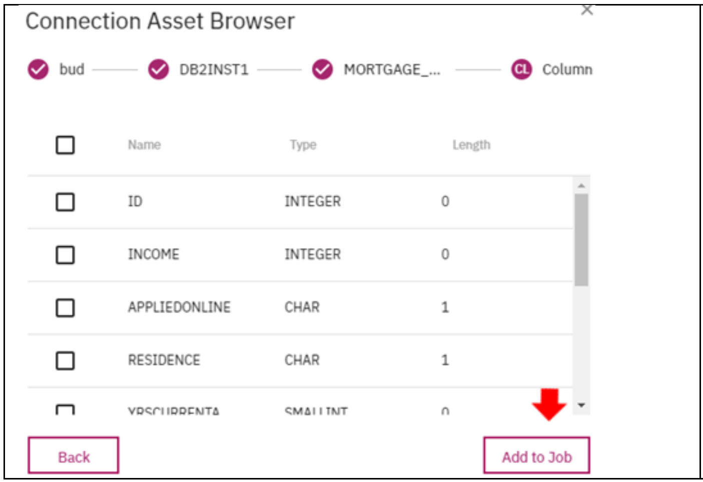
If connection name is already exist just select it and click **Next**.



1. Chose the schema named ‘DB2INST1’
2. Click **Next**

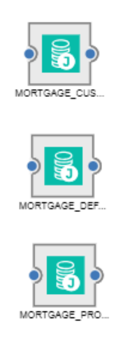


1. Choose table named ‘MORTGAGE_CUSTOMER’
2. Click **Next**

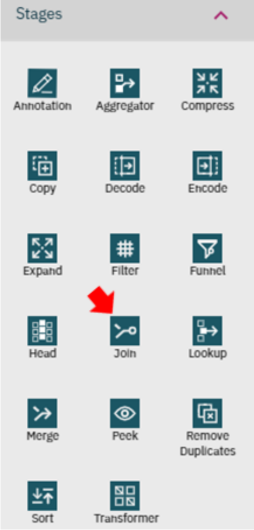


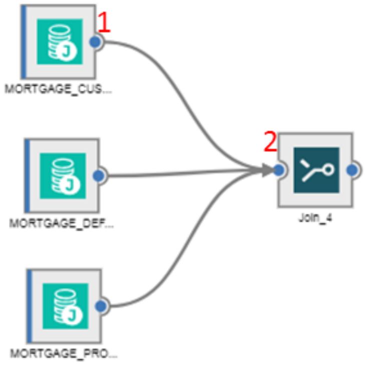
☐	Name	Type	Length
☐	ID	INTEGER	0
☐	INCOME	INTEGER	0
☐	APPLIEDONLINE	CHAR	1
☐	RESIDENCE	CHAR	1
☐	YRSCURRENTA	SMALLINT	0

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

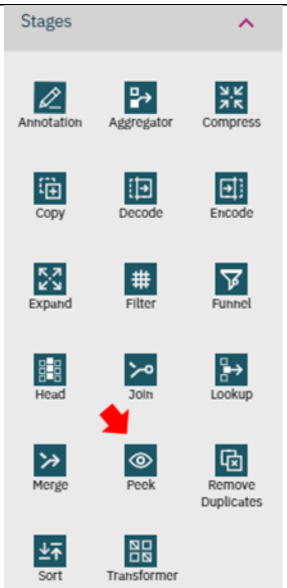
 	Repeat the above step in 5.4. for add tables 'MORTGAGE_DEFAUT' and 'MORTGAGE_PROPERTY' to the job. Once all three tables added to the job, you should have three tiles on right pane.

9.4. Join tables

 Stages Annotation Aggregator Compress Copy Decode Encode Expand Filter Funnel Head Join Lookup Merge Peek Remove Duplicates Sort Transformer	Now join the data. Click on the Join icon from the palette on left, drag it on the right pane and click once again.

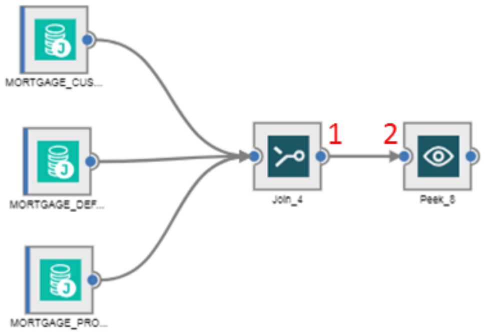
 1 2 MORTGAGE_CUS... MORTGAGE_DEF... MORTGAGE_PRO... Join_4	Connect one table tile at a time to the join tile. To connect tiles, click on two blue dots one at a time.

9.5. Preview output from join



Add peek to look into join output.

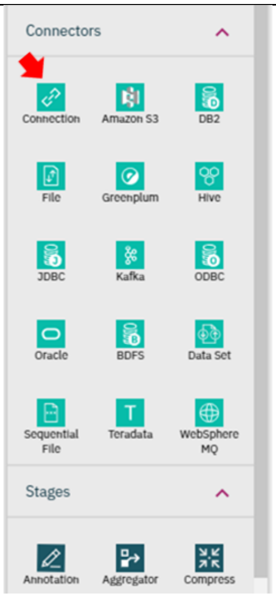
Click on the Peek icon from the palette on left, drag it on the right pane and click once again.



Connect join to the peek tile.

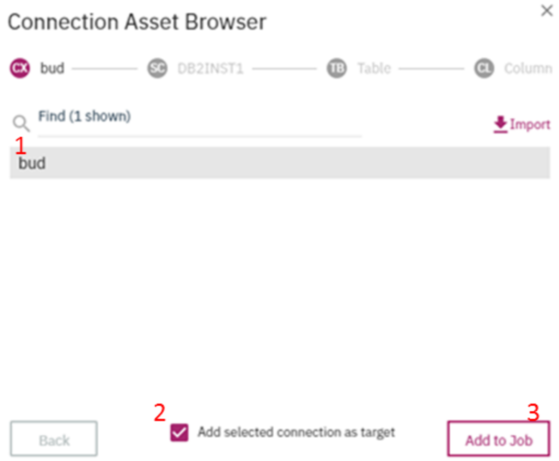
To connect tiles, click on two blue dots one at a time.

9.6. Store output from join

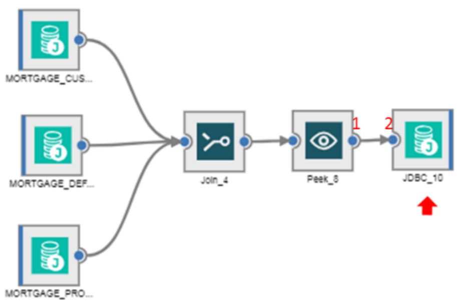


Next save the persistent data from join to the target at Db2 database.

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



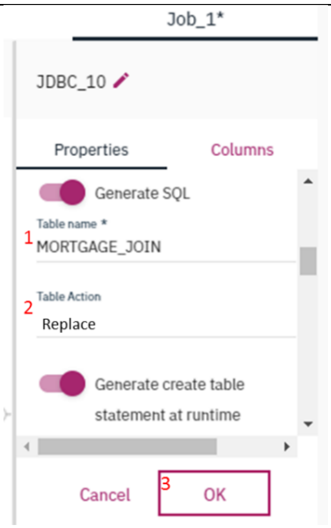
1. On the **Connection Asset Browser** window, click on connection that you created earlier in step 4.2.
2. Use check box **Add selected connection as target**
3. Click **Add to Job**



Join the target table tile with the peek

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

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



Generate SQL

1 Table name * MORTGAGE_JOIN

2 Table Action Replace

Generate create table statement at runtime

On the Job properties pane

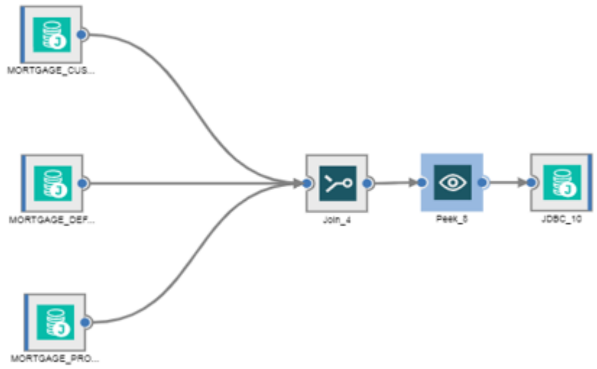
1. Use ‘MORTGAGE_JOIN<#>’ as target **Table name**
2. From the **Table Action** dropdown menu chose ‘Replace’
3. Click on **OK**

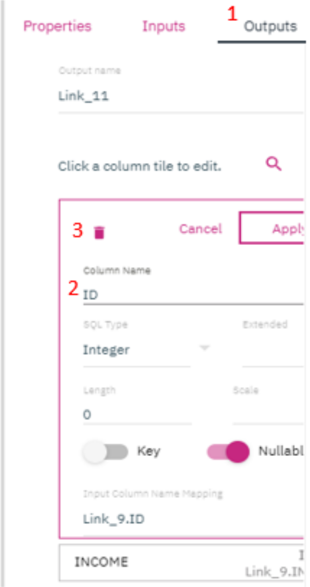
9.7. Transform output data

Let’s go back to the Peek tile and 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**.





Output name Link_11

Click a column tile to edit.

2 Column Name ID

SQL Type Integer

Extended

Length 0

Scale

Key

Nullabl

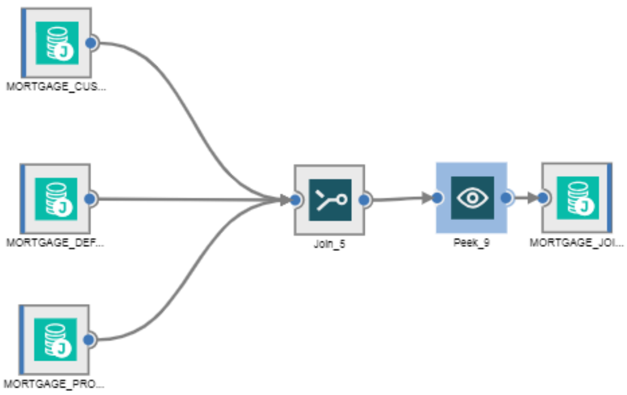
Input Column Name Mapping Link_9.ID

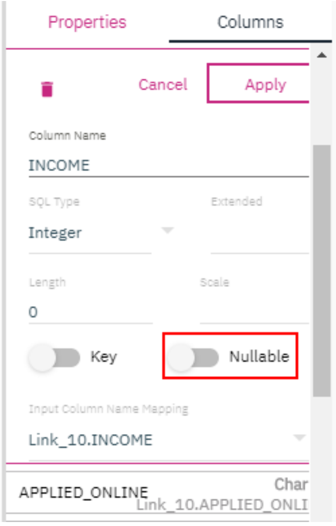
INCOME Link_9.IN

9.8. Apply governance rule

Go back to the **Peek** tile and 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**





Column Name: INCOME

SQL Type: Integer

Extended:

Length: 0

Scale:

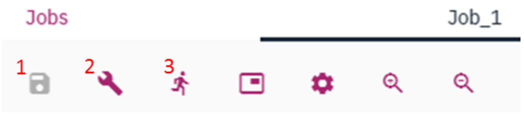
Key

Nullable

Input Column Name Mapping: Link_10.INCOME

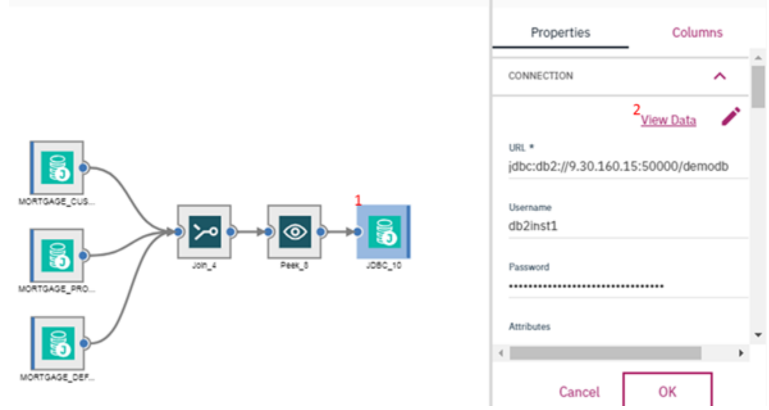
APPLIED_ONLINE Char Link_10.APPLIED_ONLI

9.9. Execute job

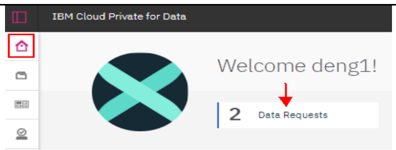


1. Click on the 💾 icon to save the job. You can use the default job name.
2. Use 🔧 icon for compile the job
3. Next click on 🏃 icon to run the job that will bring a **Job Run Options** window. Don’t change anything, just click on **Run**. Run may take few minutes to complete.
4. Click on the ⟳ icon to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

9.10. Preview output data

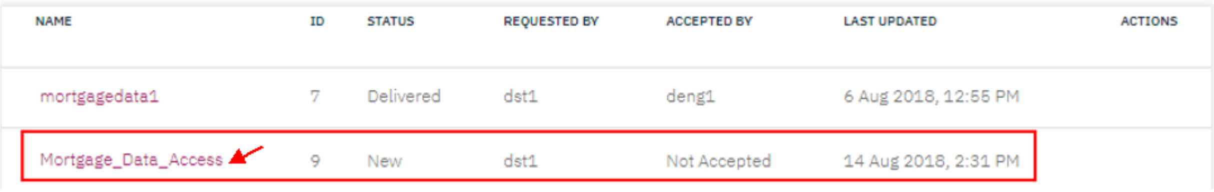
 MORTGAGE_CUS... MORTGAGE_PRO... MORTGAGE_DEF... Join_4 Peek_8 1 JDBC_10 Properties Columns CONNECTION 2 View Data Attributes Cancel OK URL * jdbc:db2://9.30.160.15:50000/demodb Username db2inst1 Password •••••••••••••••••••••••••••••••••	Let's take a quick look into the final data. 1. Click on the new 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.

10. Deliver Dataset



Go to the home page by clicking on ⌂ icon from left pane and check the **Data Requests** tab.

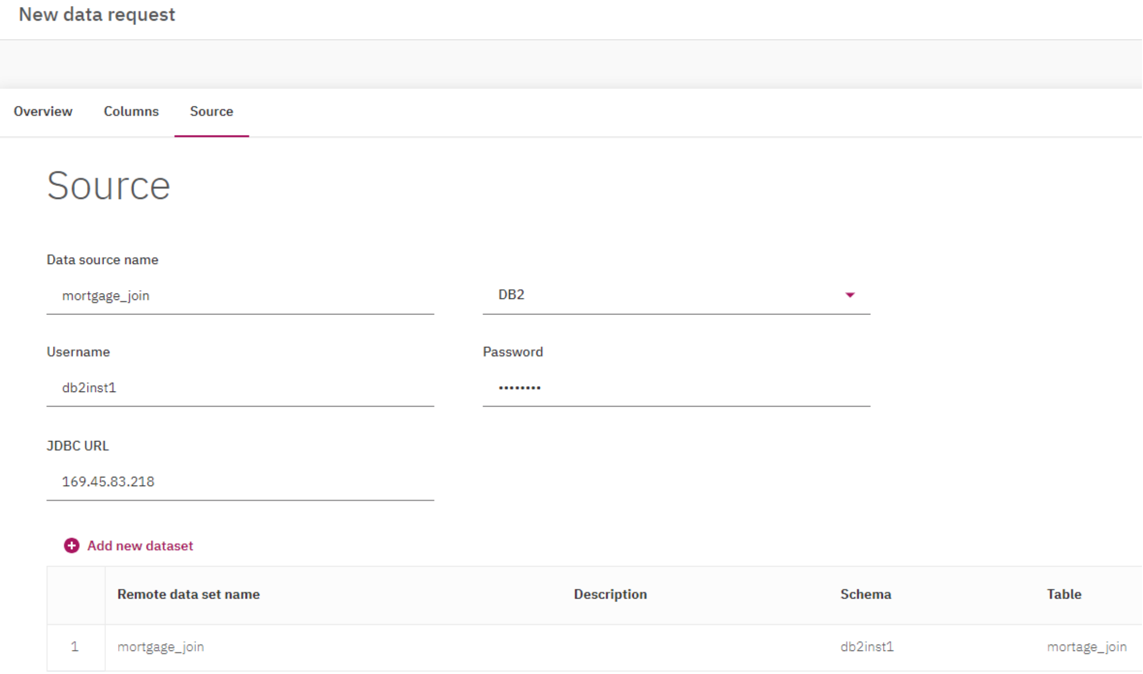
Click on the data request for update that submitted by data scientist earlier.



NAME	ID	STATUS	REQUESTED BY	ACCEPTED BY	LAST UPDATED	ACTIONS
mortgagedata1	7	Delivered	dst1	deng1	6 Aug 2018, 12:55 PM	
Mortgage_Data_Access	9	New	dst1	Not Accepted	14 Aug 2018, 2:31 PM	

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.



Source

Data source name: mortgage_join

DB2

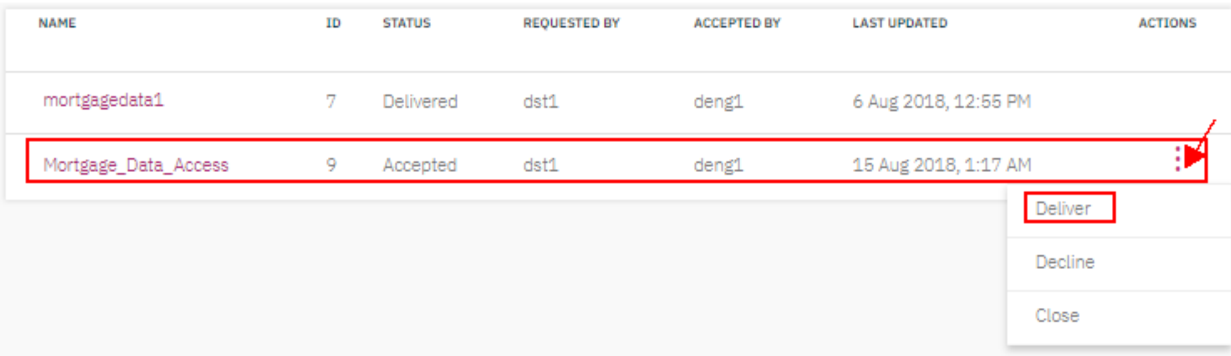
Username: db2inst1

Password: ••••••••

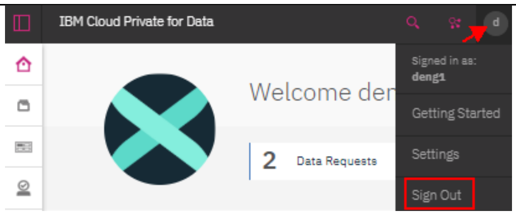
JDBC URL: 169.45.83.218

	Remote data set name	Description	Schema	Table
1	mortgage_join		db2inst1	mortage_join

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



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

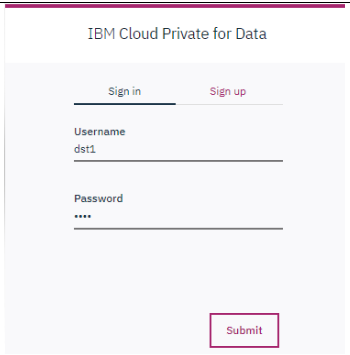


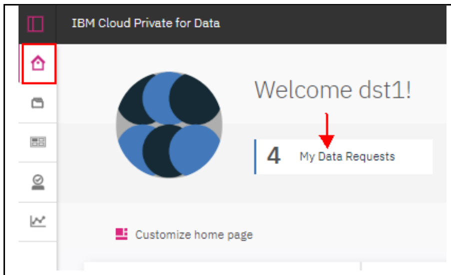
Sign out from user **deng<#>**

11. Build Model

With ICP for Data, you can collaborate with other team members on analytic projects to create visualizations and machine learning models with data from your enterprise. This HoL uses a preconfigured analysis model that you can use to run a basic machine learning simulation. The Mortgage Default model predicts whether or not customers are likely to default on their mortgage loan.

The object of this model is to show the functionality of ICP for Data, not the prediction accurecy. One can use lot more data and build a compmex algorithem to get better accurecy.

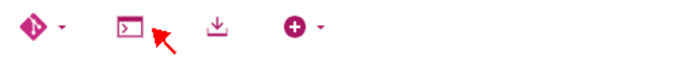
 Username dst1 Password ••••	Sigh in to the ICPD web console as user ‘dst<#>’ and password is ‘dst<#>’ that you created earlier.

 Welcome dst1! 4 My Data Requests	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

11.1. Navigate to analytics project

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

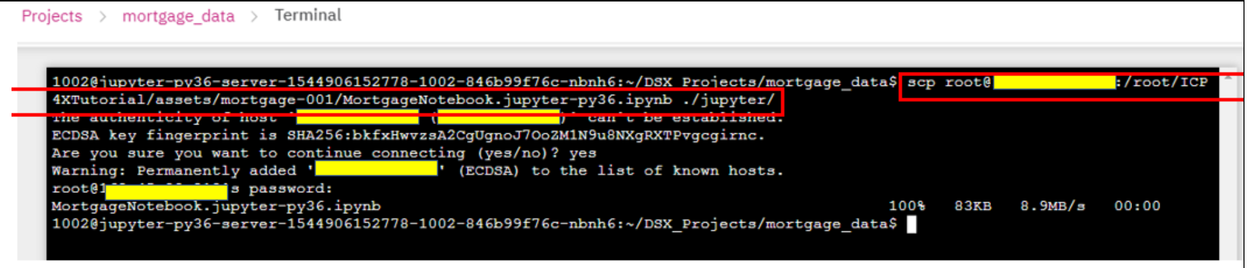
11.2. Create a model



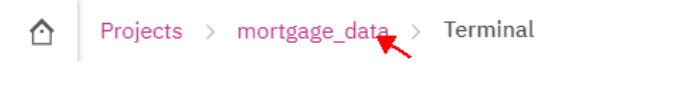
Next, choose the **Launch Terminal with Python** from top right corner.

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.

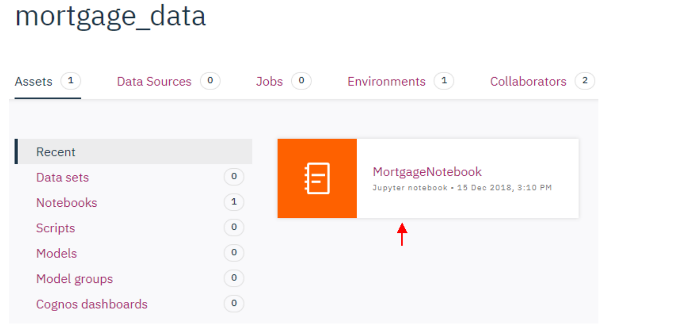
(This step is needed just for this HoI to create a model easily. In real life a data scientiest will not have access of root on master-1 node.)



```
1002@jupyter-py36-server-1544906152778-1002-846b99f76c-nbnh6:~/DSX_Projects/mortgage_data$ scp root@[redacted]:/root/ICP
4XTutorial/assets/mortgage-001/MortgageNotebook.jupyter-py36.ipynb ./jupyter/
The authenticity of host '[redacted] ([redacted])' can't be established.
ECDSA key fingerprint is SHA256:bkfxHwvzsA2CgUgnoJ7OoZM1N9u8NXgRXTPvgcgirnc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '[redacted]' (ECDSA) to the list of known hosts.
root@1[redacted]s password:
MortgageNotebook.jupyter-py36.ipynb                                   100%   83KB   8.9MB/s   00:00
1002@jupyter-py36-server-1544906152778-1002-846b99f76c-nbnh6:~/DSX_Projects/mortgage_data$
```



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

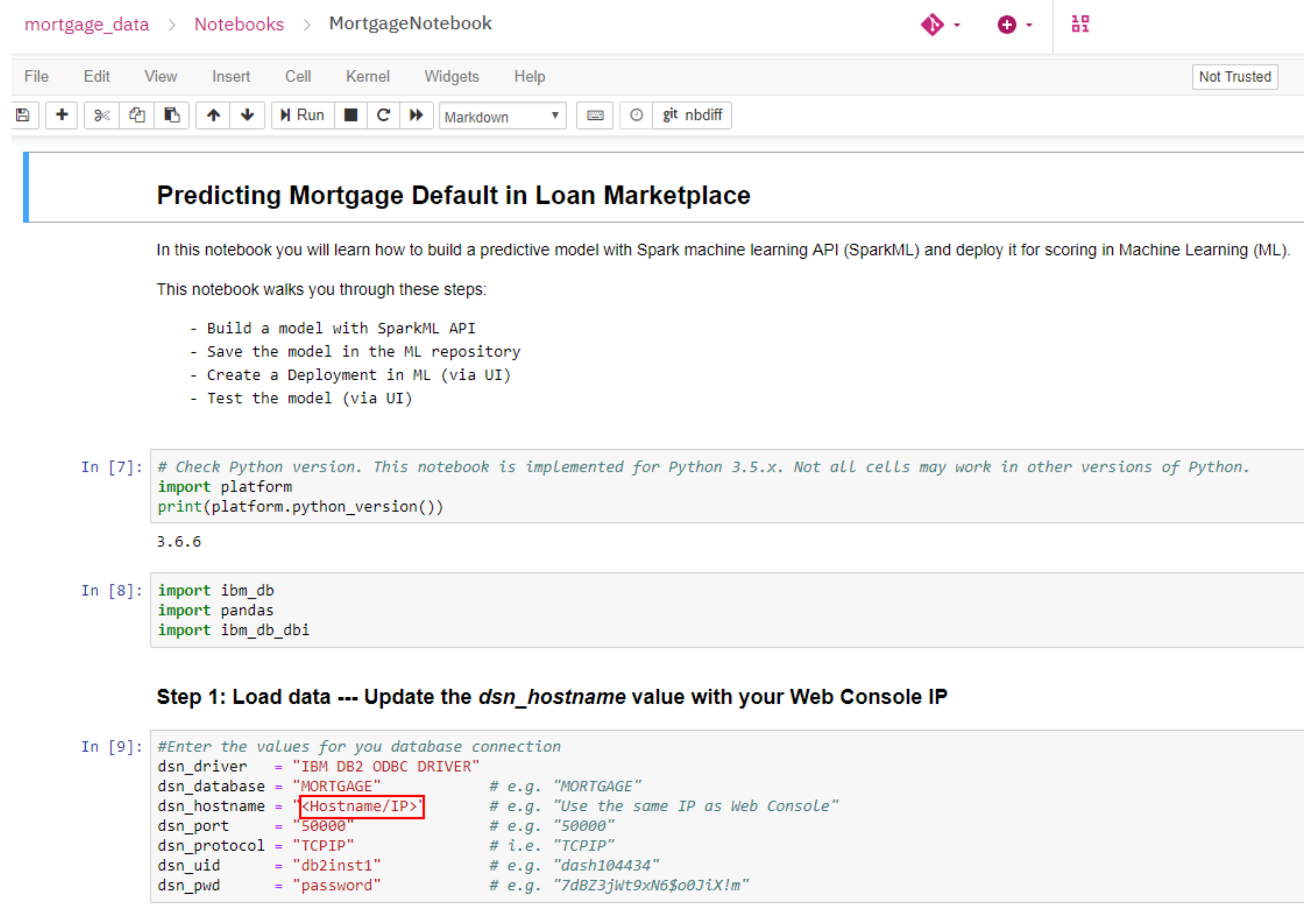


Open the predefined notebook called **MortgageNotebook**.

11.3. 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.



Predicting Mortgage Default in Loan Marketplace

In this notebook you will learn how to build a predictive model with Spark machine learning API (SparkML) and deploy it for scoring in Machine Learning (ML).

This notebook walks you through these steps:

```
- Build a model with SparkML API
- Save the model in the ML repository
- Create a Deployment in ML (via UI)
- Test the model (via UI)
```

In [7]:
```
# Check Python version. This notebook is implemented for Python 3.5.x. Not all cells may work in other versions of Python.
import platform
print(platform.python_version())
```

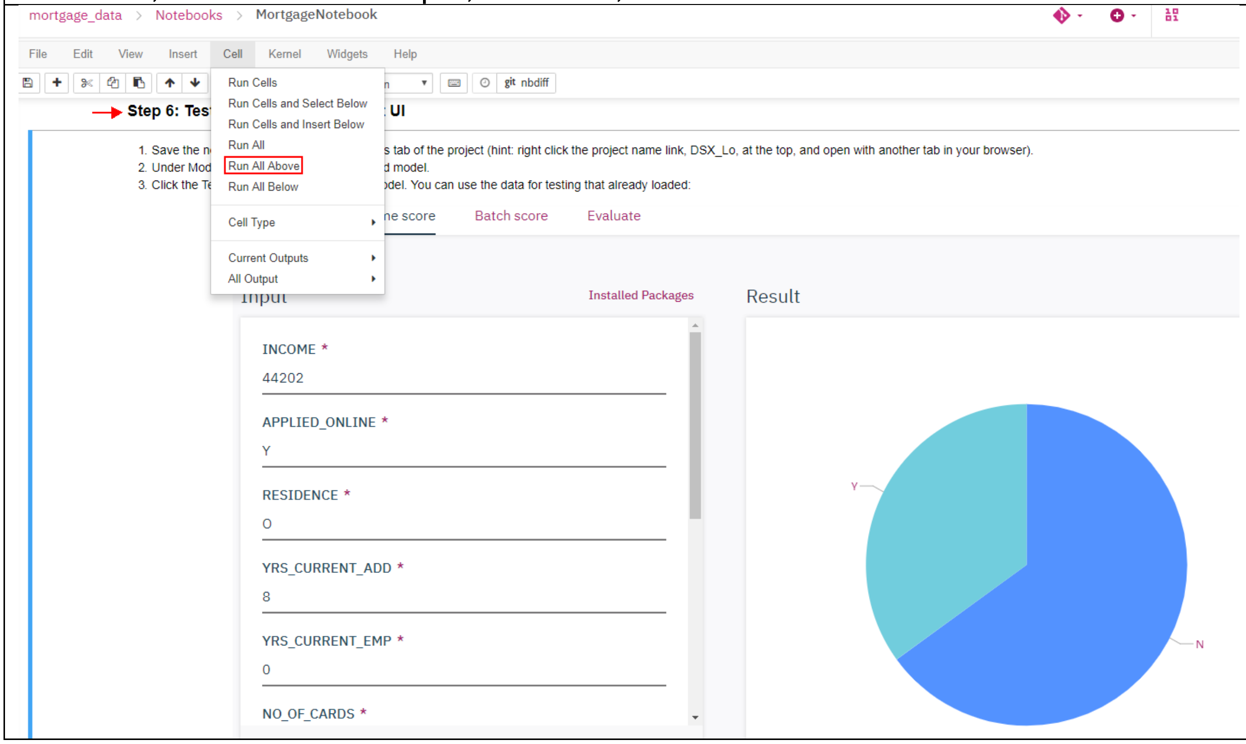
```
3.6.6
```

In [8]:
```
import ibm_db
import pandas
import ibm_db_dbi
```

Step 1: Load data --- Update the *dsn_hostname* value with your Web Console IP

In [9]:
```
#Enter the values for you database connection
dsn_driver   = "IBM DB2 ODBC DRIVER"
dsn_database = "MORTGAGE"            # e.g. "MORTGAGE"
dsn_hostname = "<Hostname/IP>"        # e.g. "Use the same IP as Web Console"
dsn_port     = "50000"               # e.g. "50000"
dsn_protocol = "TCPIP"               # i.e. "TCPIP"
dsn_uid      = "db2inst1"            # e.g. "dash104434"
dsn_pwd      = "password"            # e.g. "7dBZ3jWt9xN6$o0JiX!m"
```

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**.



1. Save the n… s tab of the project (hint: right click the project name link, DSX_Lo, at the top, and open with another tab in your browser).
2. Under Mod… d model.
3. Click the Te… odel. You can use the data for testing that already loaded:

INCOME * 44202

APPLIED_ONLINE * Y

RESIDENCE * O

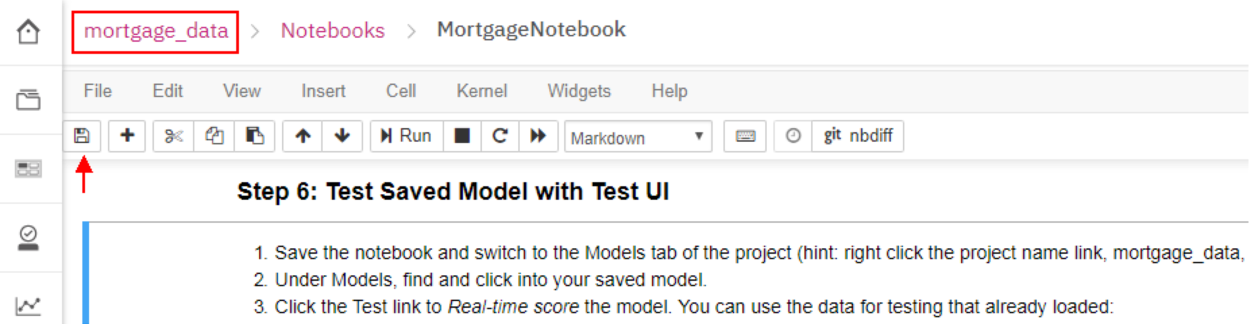
YRS_CURRENT_ADD * 8

YRS_CURRENT_EMP * 0

NO_OF_CARDS *

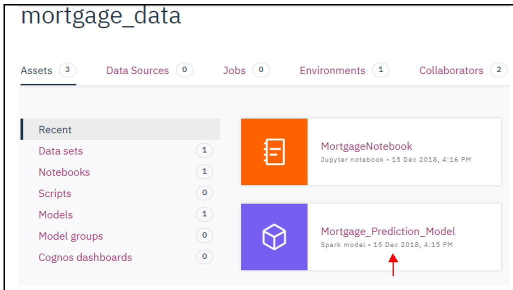
11.4. Test the model

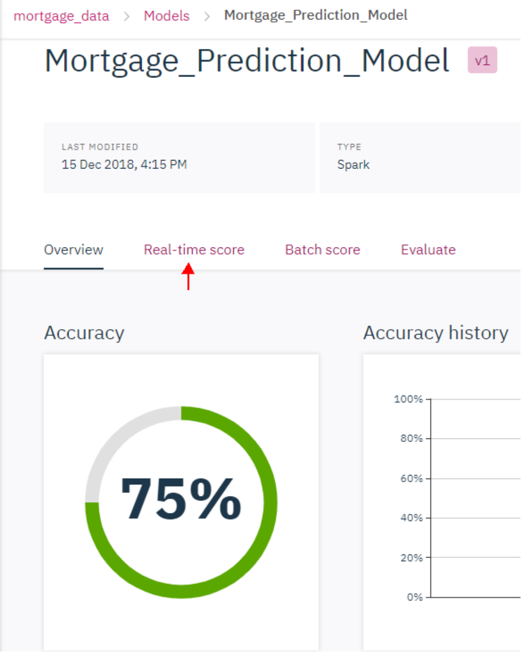
Save the notebook and switch to the Models tab of the project (hint: right click the project name link, **mortgage_data**, at the top, and open with another tab in your browser).



Step 6: Test Saved Model with Test UI

1. Save the notebook and switch to the Models tab of the project (hint: right click the project name link, mortgage_data,
2. Under Models, find and click into your saved model.
3. Click the Test link to *Real-time score* the model. You can use the data for testing that already loaded:

 mortgage_data Assets 3 · Data Sources 0 · Jobs 0 · Environments 1 · Collaborators 2 Recent Data sets 1 Notebooks 1 Scripts 0 Models 1 Model groups 0 Cognos dashboards 0 MortgageNotebook Jupyter notebook • 15 Dec 2018, 4:16 PM Mortgage_Prediction_Model Spark model • 15 Dec 2018, 4:15 PM	Chose the **Mortgage_Prediction_Model**



mortgage_data > Models > Mortgage_Prediction_Model

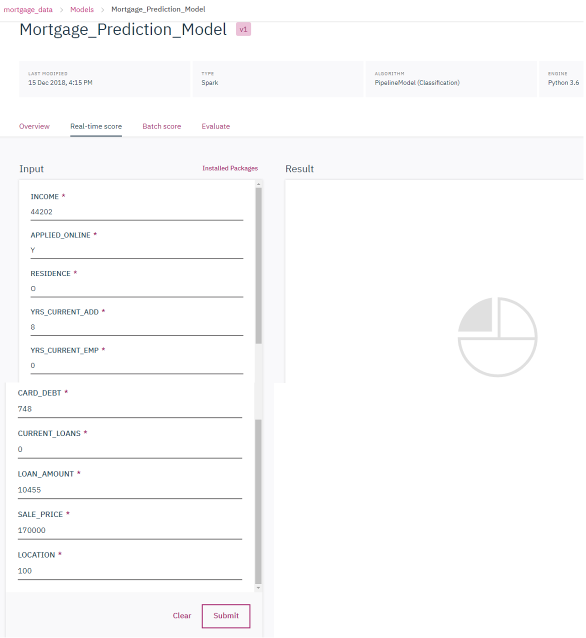
Mortgage_Prediction_Model v1

LAST MODIFIED
15 Dec 2018, 4:15 PM

TYPE
Spark

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.



mortgage_data > Models > Mortgage_Prediction_Model

Mortgage_Prediction_Model v1

LAST MODIFIED
15 Dec 2018, 4:15 PM

TYPE
Spark

ALGORITHM
PipelineModel (Classification)

ENGINE
Python 3.6

INCOME *
44202

APPLIED_ONLINE *
Y

RESIDENCE *
0

YRS_CURRENT_ADD *
8

YRS_CURRENT_EMP *
0

CARD_DEBT *
748

CURRENT_LOANS *
0

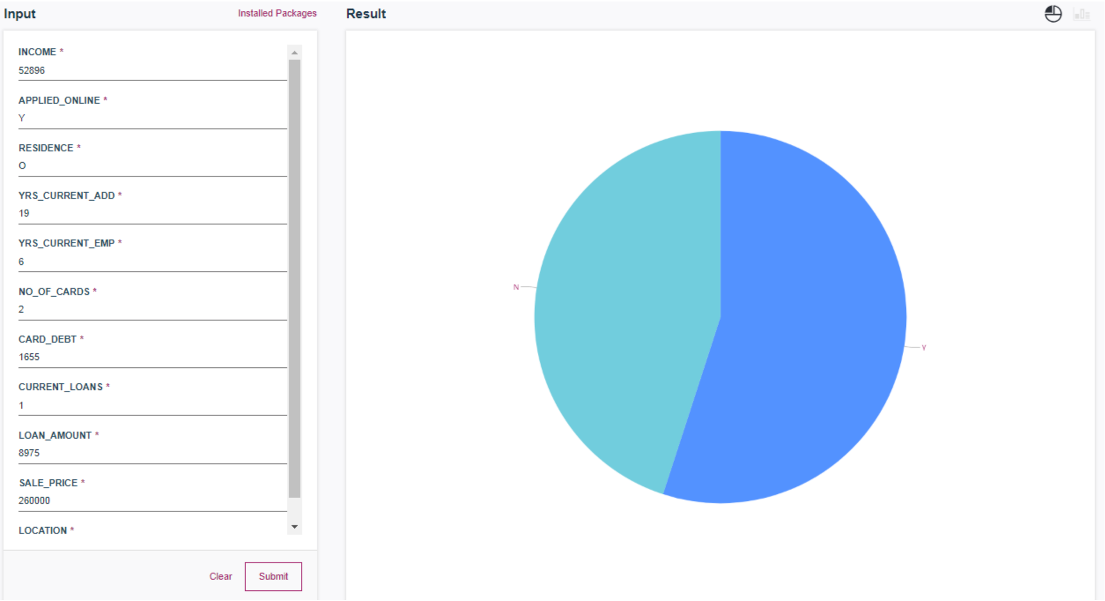
LOAN_AMOUNT *
10455

SALE_PRICE *
170000

LOCATION *
100

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.



INCOME *
52896

APPLIED_ONLINE *
Y

RESIDENCE *
O

YRS_CURRENT_ADD *
19

YRS_CURRENT_EMP *
6

NO_OF_CARDS *
2

CARD_DEBT *
1655

CURRENT_LOANS *
1

LOAN_AMOUNT *
8975

SALE_PRICE *
260000

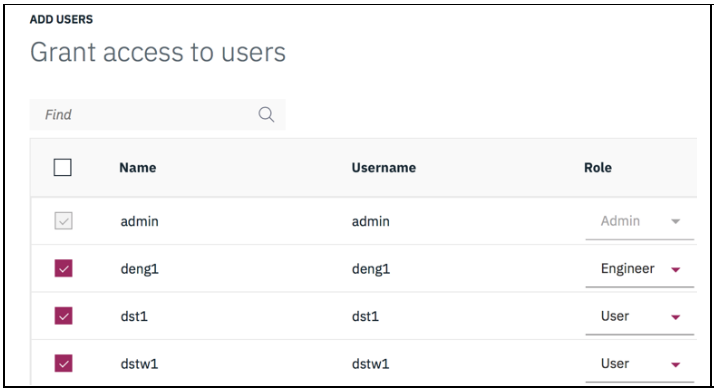
LOCATION *

12. Data Virtualization

Data virtualization (DV) integrates data sources across multiple types and locations and turns it into one logical data view. Creating connections to your data sources enables you to quickly view across your organization's data. This virtual data platform enables real-time analytics without moving data, duplication, ETLs, and additional storage requirements, so processing times are greatly accelerated. This brings real-time insightful results to decision-making applications or analysts more quickly and dependably than existing methods. DV is an add-on feature in ICP for Data. It's already provisioned in this HoL.

12.1. Giving users access to data virtualization

In order for a user to have access to the data virtualization service, you must assign them to appropriate data virtualization roles.



ADD USERS

Grant access to users

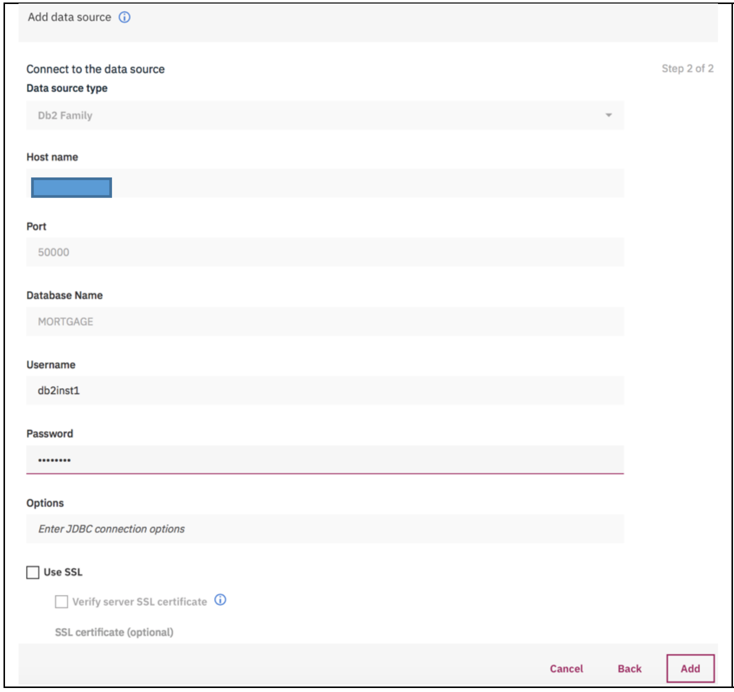
☐	Name	Username	Role
☑	admin	admin	Admin
☑	deng1	deng1	Engineer
☑	dst1	dst1	User
☑	dstw1	dstw1	User

1. Select **Collect** > **Virtualize data** from left pane
2. Select **Menu** > **Manage users** > **Add users** from top
3. Check all users that you created earlier and keep their default role.
4. Click **Add**

12.2. Adding an existing data source

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 has the ability to create virtual tables. DV agents connect to relational data sources using JDBC protocol. In this HoL you will add two data sources, one for Db2 and other one for Informix.

As you already created data connection to Db2 earlier, you need to add same to DV as an existing data source.



Connect to the data source — Step 2 of 2

Data source type: Db2 Family

Host name: [redacted]

Port: 50000

Database Name: MORTGAGE

Username: db2inst1

Password: ••••••••

Options: Enter JDBC connection options

□ Use SSL

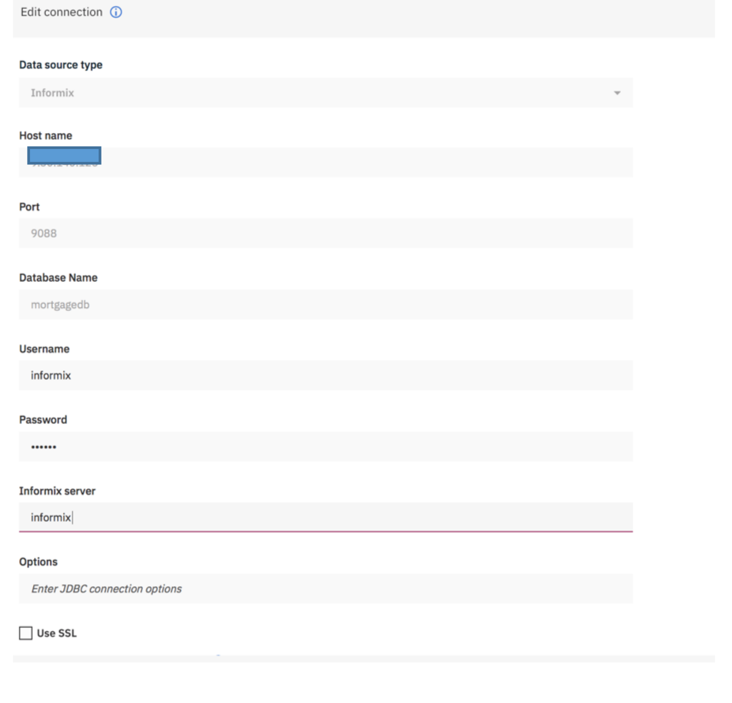
□ Verify server SSL certificate ⓘ

SSL certificate (optional)

1. Go to **Collect** > **Virtualized data** > **Menu** > **Data sources**
2. Click **Add data source** to see a list of data sources that can be added to data virtualization. Select **Db2** data source that have been defined in the Organize menu earlier.
3. In the **Add data source** panel use **db2inst1** and **password** for username and password.
4. Click **Add**

12.3. Adding a new data source

Let's add a new data source for the Informix.



Data source type: Informix

Host name: [redacted]

Port: 9088

Database Name: mortgagedb

Username: informix

Password: ••••••

Informix server: informix

Options: Enter JDBC connection options

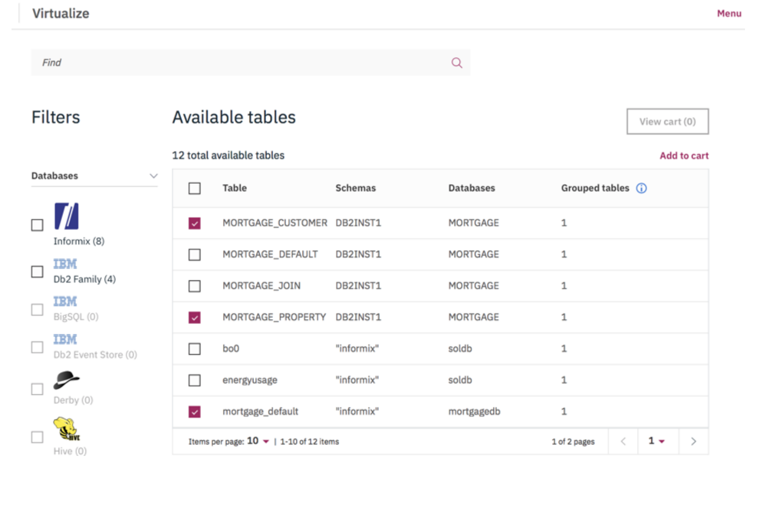
□ Use SSL

1. Go to **Collect** > **Virtualized data** > **Menu** > **Data sources**
2. Click **Add data source** > **New data source**
3. Update data source with following information:

   Data source type = Informix
   Host name = <IP of node 1>
   Port = 9088
   Database Nam = mortgagedb
   Username = informix
   Password = in4mix
   Informix server = informix
4. Click **Add**

12.4. Creating virtualized table

The most common mechanism for virtualizing data is to create a "view" or virtual table. You can create a virtual table to segment data from one or more tables. Such segmentation can be vertical (either a subset or superset of columns based on a selection of chosen columns) or horizontal (an explicit set of rows or records based on a conditional expression) or both. You can then run queries against the resulting virtual table.



Available tables

12 total available tables

□	Table	Schemas	Databases	Grouped tables
☑	MORTGAGE_CUSTOMER	DB2INST1	MORTGAGE	1
□	MORTGAGE_DEFAULT	DB2INST1	MORTGAGE	1
□	MORTGAGE_JOIN	DB2INST1	MORTGAGE	1
☑	MORTGAGE_PROPERTY	DB2INST1	MORTGAGE	1
□	bo0	"informix"	soldb	1
□	energyusage	"informix"	soldb	1
☑	mortgage_default	"informix"	mortgagedb	1

Items per page: 10 | 1-10 of 12 items   1 of 2 pages   1

1. Click **Collect** > **Virtualized data** > **Menu** > **Virtualize**

2. Select three tables **MORTGAGE_CUSTOMER**, **MORTGAGE_PROPERTY** from MORTGAGE database and **mortgage_default** from mortgaged, then click **Add to cart**

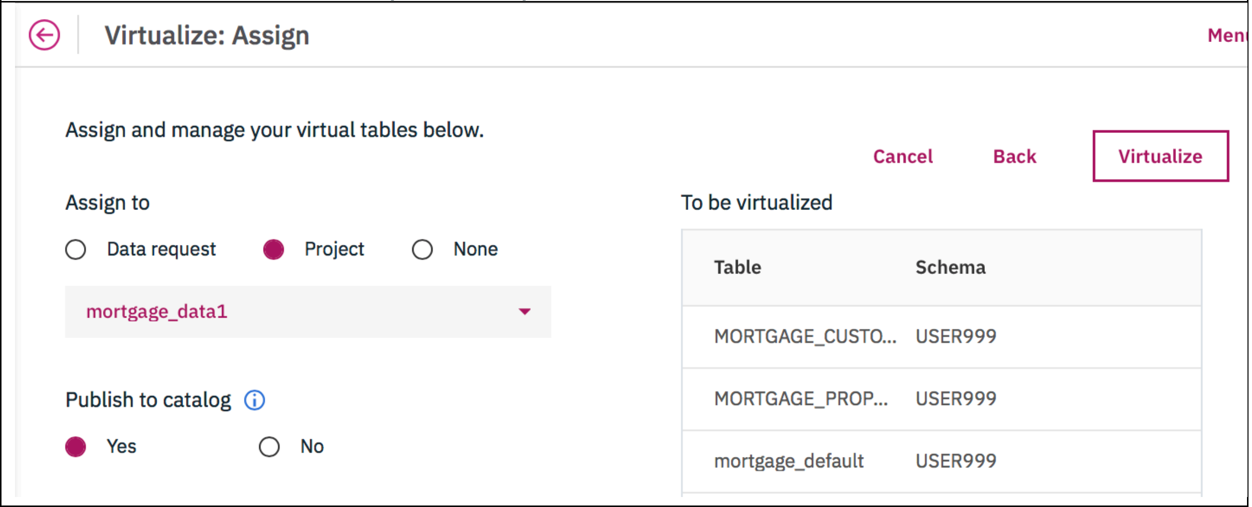
3. Click **View cart**

4. Click **Next**

1. Select **project** to assign virtualized table to your analytics project. Then, choose the **mortgage_data<#>** project.

2. Choose **publish to catalog** for include virtualized table to the data catalog. This operation will create a publishing request, a data steward must approve the request before the asset is added to the enterprise data catalog.

3. Click **Virtualize** to complete the process



Assign and manage your virtual tables below.

Assign to

○ Data request ● Project ○ None

mortgage_data1

Publish to catalog

● Yes ○ No

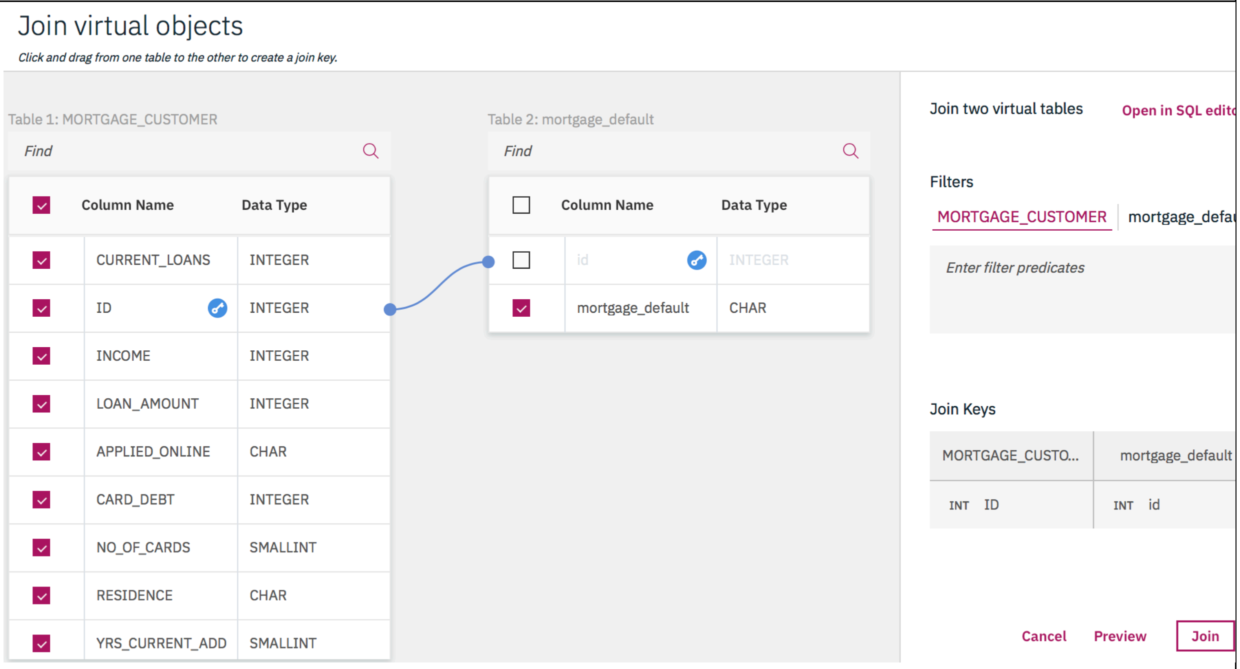
To be virtualized

Table	Schema
MORTGAGE_CUSTO...	USER999
MORTGAGE_PROP...	USER999
mortgage_default	USER999

12.5. Creating joined virtual table

You can create a new virtual table based on existing virtual tables.

1. Click **Collect** > **Virtualized data** > **Menu** > **Virtualized data** to see your virtualized tables.
2. Check **MORTGAGE_CUSTOMER** and **mortgage_default** virtual tables for join.
3. Click on **Join view**
4. Uncheck the ID column from mortgage_default table for reduction redundancy
5. Click and drag from one **ID** column to another to create a join key. Both join keys must be of the same data type.
6. Click **Join**



Join virtual objects

Click and drag from one table to the other to create a join key.

Table 1: MORTGAGE_CUSTOMER

Find

☑	Column Name	Data Type
☑	CURRENT_LOANS	INTEGER
☑	ID	INTEGER
☑	INCOME	INTEGER
☑	LOAN_AMOUNT	INTEGER
☑	APPLIED_ONLINE	CHAR
☑	CARD_DEBT	INTEGER
☑	NO_OF_CARDS	SMALLINT
☑	RESIDENCE	CHAR
☑	YRS_CURRENT_ADD	SMALLINT

Table 2: mortgage_default

Find

☐	Column Name	Data Type
☐	id	INTEGER
☑	mortgage_default	CHAR

Join two virtual tables

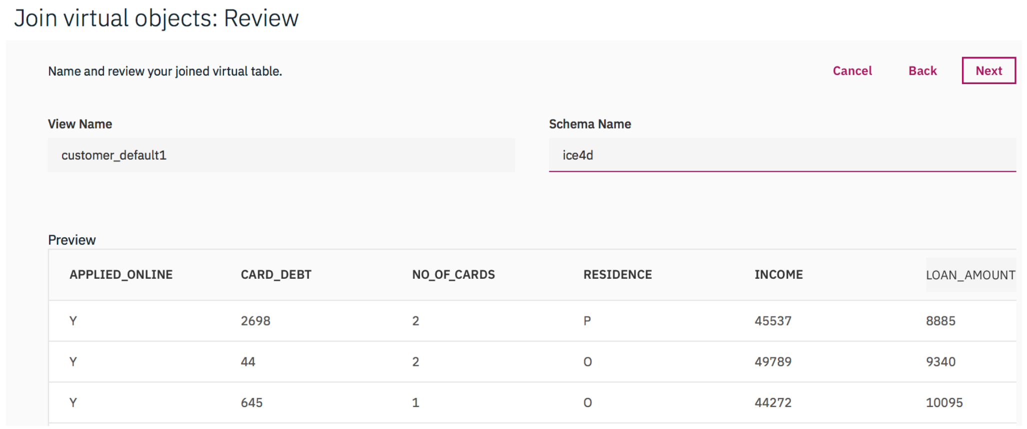
Filters

Enter filter predicates

Join Keys

MORTGAGE_CUSTO...	mortgage_default
INT ID	INT id

Name the view as **customer_default<#>** and schema as **icp4d,** then click **Next**



Join virtual objects: Review

Name and review your joined virtual table.

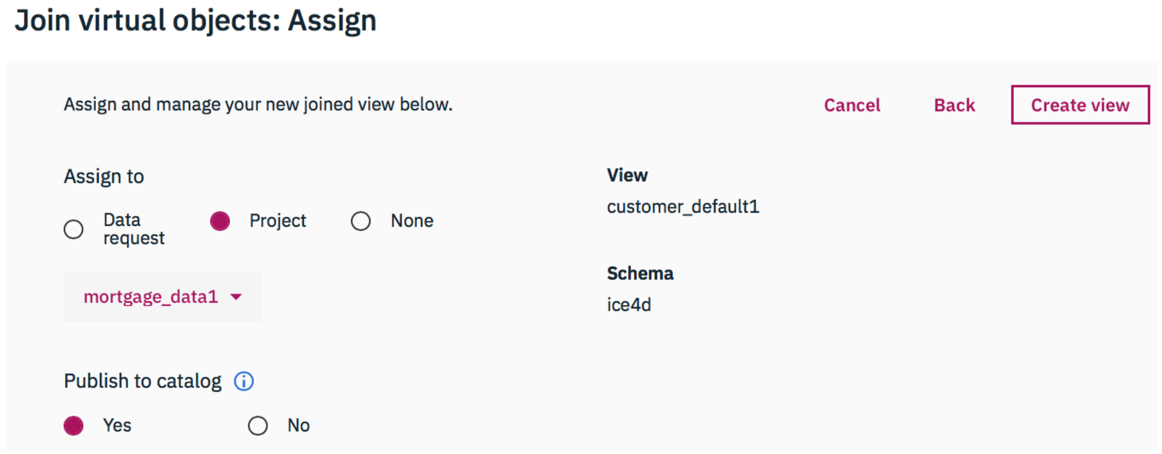
View Name: customer_default1

Schema Name: ice4d

Preview

APPLIED_ONLINE	CARD_DEBT	NO_OF_CARDS	RESIDENCE	INCOME	LOAN_AMOUNT
Y	2698	2	P	45537	8885
Y	44	2	O	49789	9340
Y	645	1	O	44272	10095

1. Select **project** to assign virtualized table to your analytics project. Then, choose the **mortgage_data<#>** project.
2. Click **Create view**



Join virtual objects: Assign

Assign and manage your new joined view below.

Assign to

○ Data request ● Project ○ None

mortgage_data1 ▾

Publish to catalog ⓘ

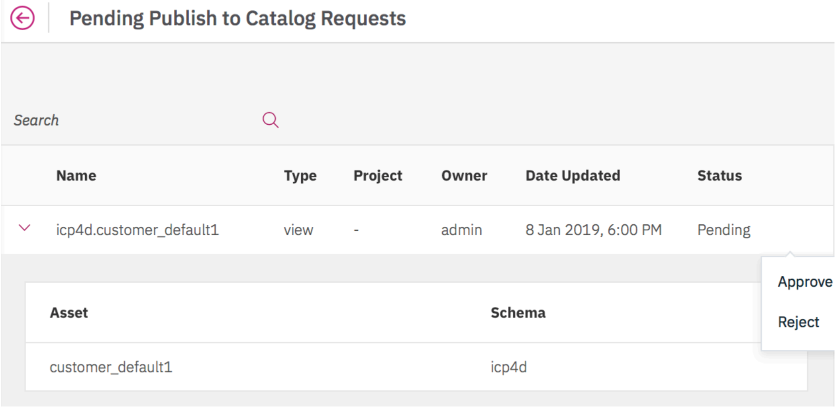
● Yes ○ No

View
customer_default1

Schema
ice4d

12.6. Publish virtualized table

A data steward needs approve the published request before the asset is added to the enterprise data catalog.



Pending Publish to Catalog Requests

Search

Name	Type	Project	Owner	Date Updated	Status
icp4d.customer_default1	view	-	admin	8 Jan 2019, 6:00 PM	Pending

Asset	Schema
customer_default1	icp4d

1. Login as **dstw1**
2. Click on ⌂ access the **Home** page
3. Click on **Pending Publish to Catalog Requests**
4. Click on ⋮ icon on left for virtual table **customer_default<#>** that you created
5. Click on **Approve**

We Value Your Feedback!

- Don’t forget to submit your Think 2019 session and speaker feedback! Your feedback is very important to us – we use it to continually improve the conference.
- Access the Think 2019 agenda tool to quickly submit your surveys from your smartphone, laptop or conference kiosk.