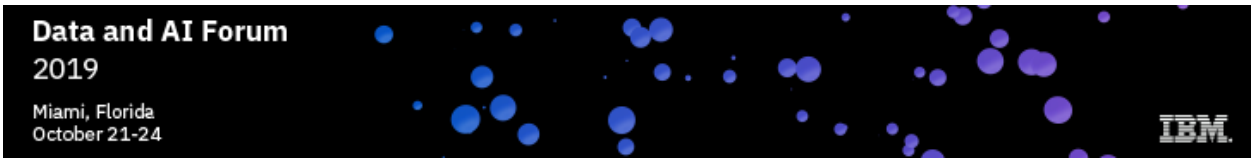


Hands-on Lab: Building an Application is as Simple as “ABC” in Cloud Pak for Data

Sanjit Chakraborty, IBM
Lucy Obus, IBM

Table of Contents

We Value Your Feedback! 4
Acknowledgements and Disclaimers 5
1. Introduction 6
2. Prerequisites 6
3. Access Credentials 7
   3.1. Access credential for Db2 database 7
   3.2. Sign in to Cloud Pak for Data web console as Administrator 8
4. Collect 9
   4.1. Connect to your data source 9
   4.2. Add connection for Db2 Database 9
   4.3. Add connection for Mongo Database 10
5. Discover Assets 11
   5.1. Navigate through discover Db2 assets 11
   5.2. Navigate through discover MongoDB assets 12
6. Data Dictionary 14
   6.1. Download Glossary 14
   6.2. Import Glossary 14
7. Analytics Project 16
   7.1. Download Analytic Project 16
   7.2. Create Analytics Project 16
8. Data Virtualization 17
   8.1. Adding a new data source for Db2 17
   8.2. Adding a new data source for MongoDB 17
   8.3. Select tables for virtualization 17
   8.4. Creating virtual table 18
   8.5. Add virtual table to catalog 19
   8.6. Publish virtualized table 19
   8.7. Access information for virtual table 20
9. RStudio 21
   9.1. Enable RStudio Environment 21
   9.2. Access RStudio Assets 22
   9.3. Shinyapp Package 22
   9.4. Update cluster password in Shinyapp 22
   9.5. Update virtual table name in Shinyapp 23
   9.5. Test the Shinyapp 24

9.6. Externalize the application..... 26

We Value Your Feedback!

- Don’t forget to submit your IBM Data and AI Forum session and speaker feedback! Your feedback is very important to us – we use it to continually improve the conference.
- Access the IBM Data and AI Forum event app to quickly submit your surveys from your smartphone, laptop or conference kiosk.

Acknowledgements and Disclaimers

Copyright © 2017 by International Business Machines Corporation (IBM). No part of this document may be reproduced or transmitted in any form without written permission from IBM.

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

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 according to 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 a 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 is in compliance with 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 in connection with 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 particular**, **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, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services®, Global Technology Services®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli® Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, 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.

1. Introduction to the Demo Challenge:

TravelBid is an e-commerce startup providing discounted hotel rates and rental car reservations. TravelBid has a bulk of hotel and rental car vendors, from which they fulfill travel reservations for their clients.

The company is looking for a data & analytics platform that can help them easily implement an e-commence system using different vendor databases, depending on the hotel or rental car reservation.

TravelBid needs an application interface that allows customers to name their price for hotel rooms and car rentals and returns a list of matches within their price range.

2. Prerequisites

- IBM Cloud Pak for Data instance is already installed and available
- Relational databases (such as Db2 and Mongo) already created and available
- Previously created business glossary and governance policies. Download it from: https://drive.google.com/file/d/1oNQKyu7kDVP65ToPKmFqFRcPyNiSi4KS/view?usp=sharing
- Previously created analytics project. Download it from: https://drive.google.com/file/d/12z0VzLAwPT6JgePvfOsAcNUHpq9eQTTp/view?usp=sharing

3. Access Credentials

To work through the lab, you need access a Db2 and Mortgage databases.

3.1. Access credential for Db2 database

You need JDBC connections to access to Db2 and MongoDB database, which hosted locally on Cloud Pak for Data.

Following is the JDBC connection credential for Db2:

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

JDBC connection credential for Mongo:

JDBC Host name	<Same IP address as your web console>
Port number	27017
Database name	icpd_mongo
User ID	mongodbuser
Password	password
Mongo	Version 4.0.12
JDBC connection string	jdbc:mongo://<same IP as Web Console>:27017/databaseName=icpd_mongo

3.2. Setting up the databases and sample tables

a) Log in to the cluster where ICP for Data is deployed.

b) From your home directory, clone the tutorial sample files:

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

c) Change to the tutorials directory:

```
cd TravelBid/tutorials
```

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

```
./load_samples.sh -t db2-travel-001
```

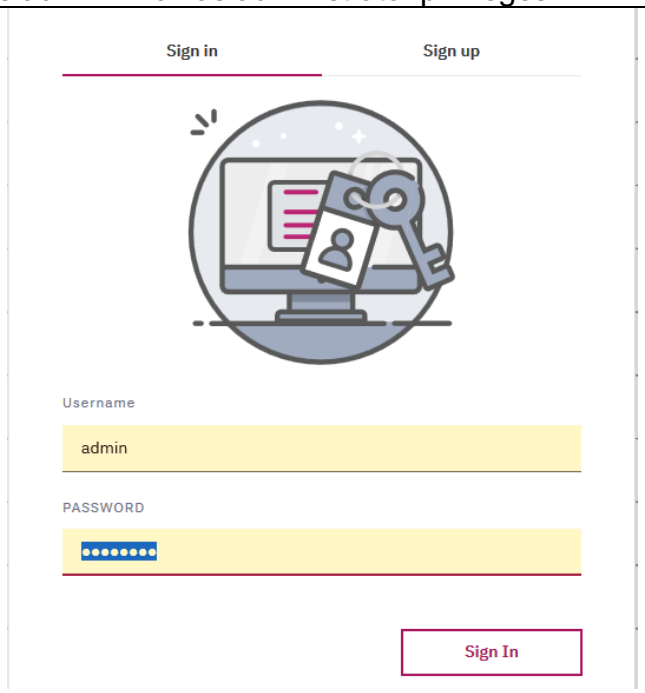
e) Run the following command to load the sample data into a MongoDB database:

```
./load_samples.sh -t mongo-travel-001
```

f) After the data loading process completes, instance of Db2 and MongoDB are hosted on your cluster as a Docker container.

3.2. Sign in to Cloud Pak for Data web console as Administrator

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

	Sign into the Cloud Pak for Data web console: username: admin password: password

4. Collect

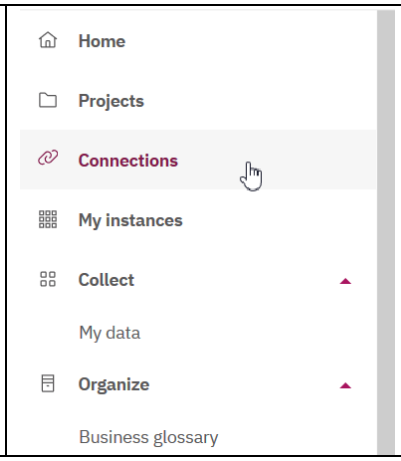
Cloud Pak for data allows you to collect data regardless of where it resides.
In this stage of the lab you'll use integrated connectors to connect to the travel's data sources. After you are connected, you can easily discover data and select the data that you need to analyze.

4.1. Connect to your data source

You will create a JDBC connection to the Db2 database where rental car related data is stored. Connecting to a database is relatively easy - you need only the JDBC link, the username, and password. You'll enter connection details, such as the name of the host where the Db2 database is located, and the credentials to access the database.

- On the left side navigation pane, click **Connections**.
- Use Next, on the **Data Connections** window, click the "Add connection" icon.

Add connection

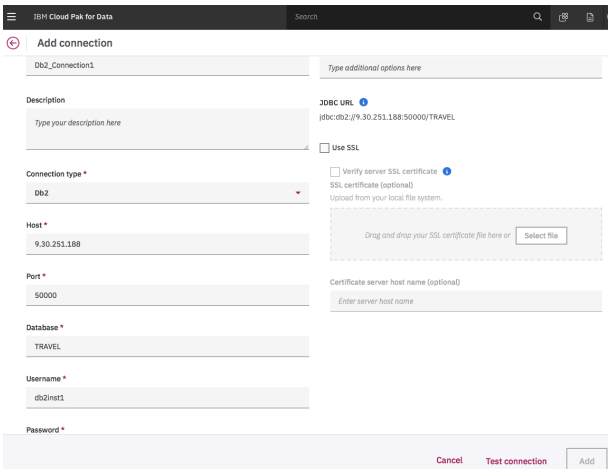


4.2. Add connection for Db2 Database

Fill out the **Add Connection** information according to the information provided in step 3.1 (pg7). Access credential for DB2. Credential used in following step is just an example.

- For **Choose connection** use the drop-down menu and select 'Db2'.
- Use Db2_Connection# as the **Name.** (Replace # with a unique number)
- **Host** is the IP of the master 1 node
- **Port** is 50000
- **Database** TRAVEL
- **Username** is 'db2inst1' and **Password** is 'password'.

- Click on **Test Connection**, once it successful click on **Save.**



4.3. Add connection for Mongo Database

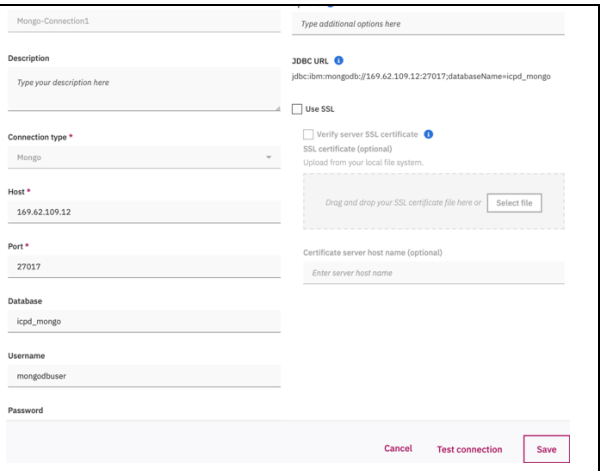
Create a JDBC connection to the MongoDB database where hotel-related data is stored. You need only the JDBC link, the username, and password to connect to a MongoDB database.

Fill out the **Add Connection** information according to the information provided in step 3.1 (pg7). Access credential for MongoDB. Credential used in following step is just an example.

- For **Choose connection** use the drop-down menu and select 'Mongo.
- Use 'Mongo_Local# as the **Name**
- **Host** is the IP of the master 1 node
- **Port** is 27017
- **Database** icpd_mongo
- **Username** is 'mongodbuser' and **Password** is 'password'.

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

Replace # with a unique number

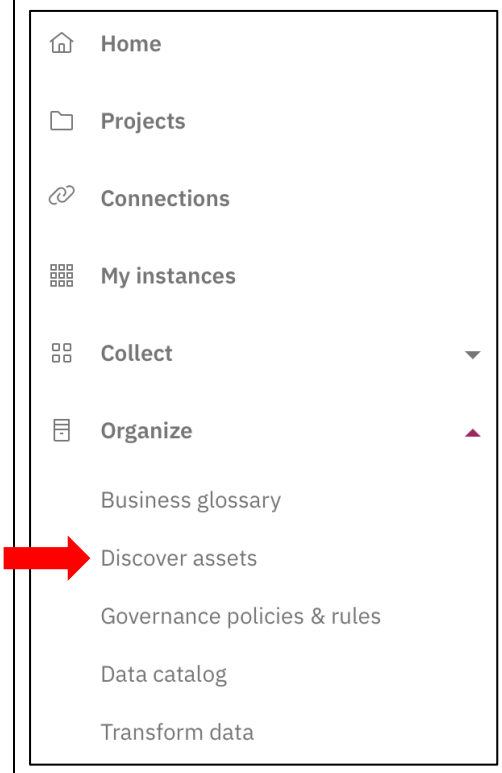


5. Discover Assets

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

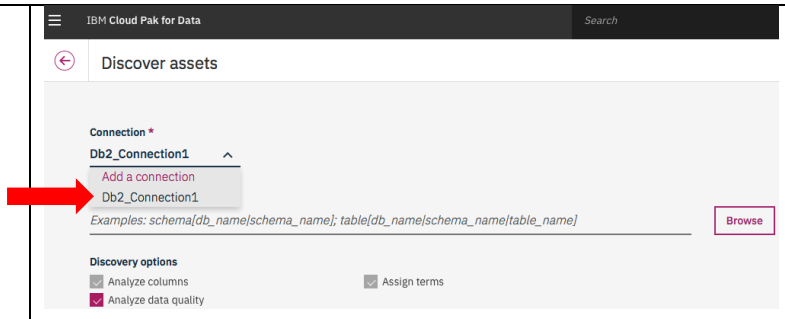
5.1. Navigate through discover Db2 assets

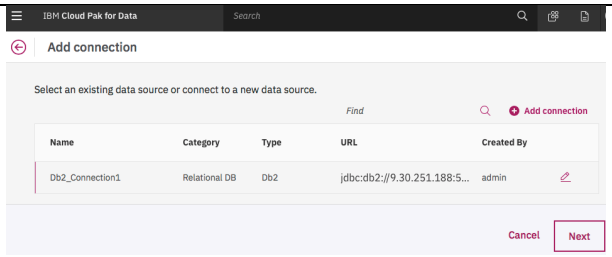
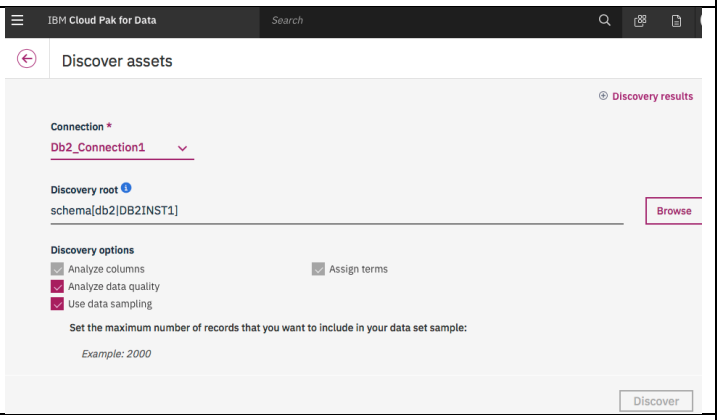
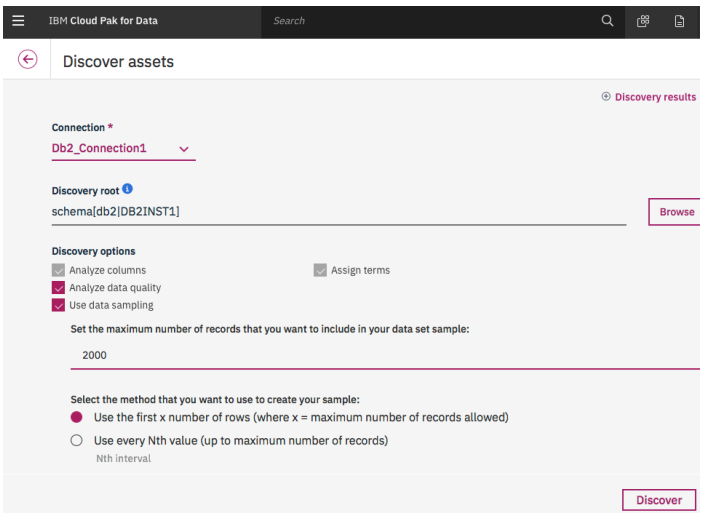
From **Organize** option on the left pane, choose **Discover assets**

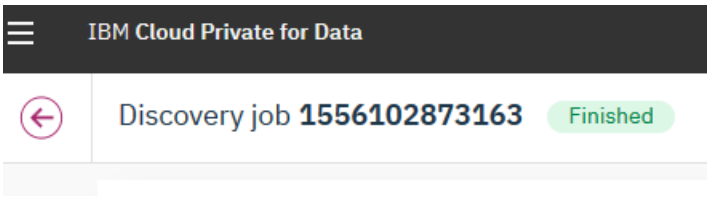


To discover assets:

1. Add a connection
2. Choose the connection named **Db2_Connection#** that you created previously, click **Next**

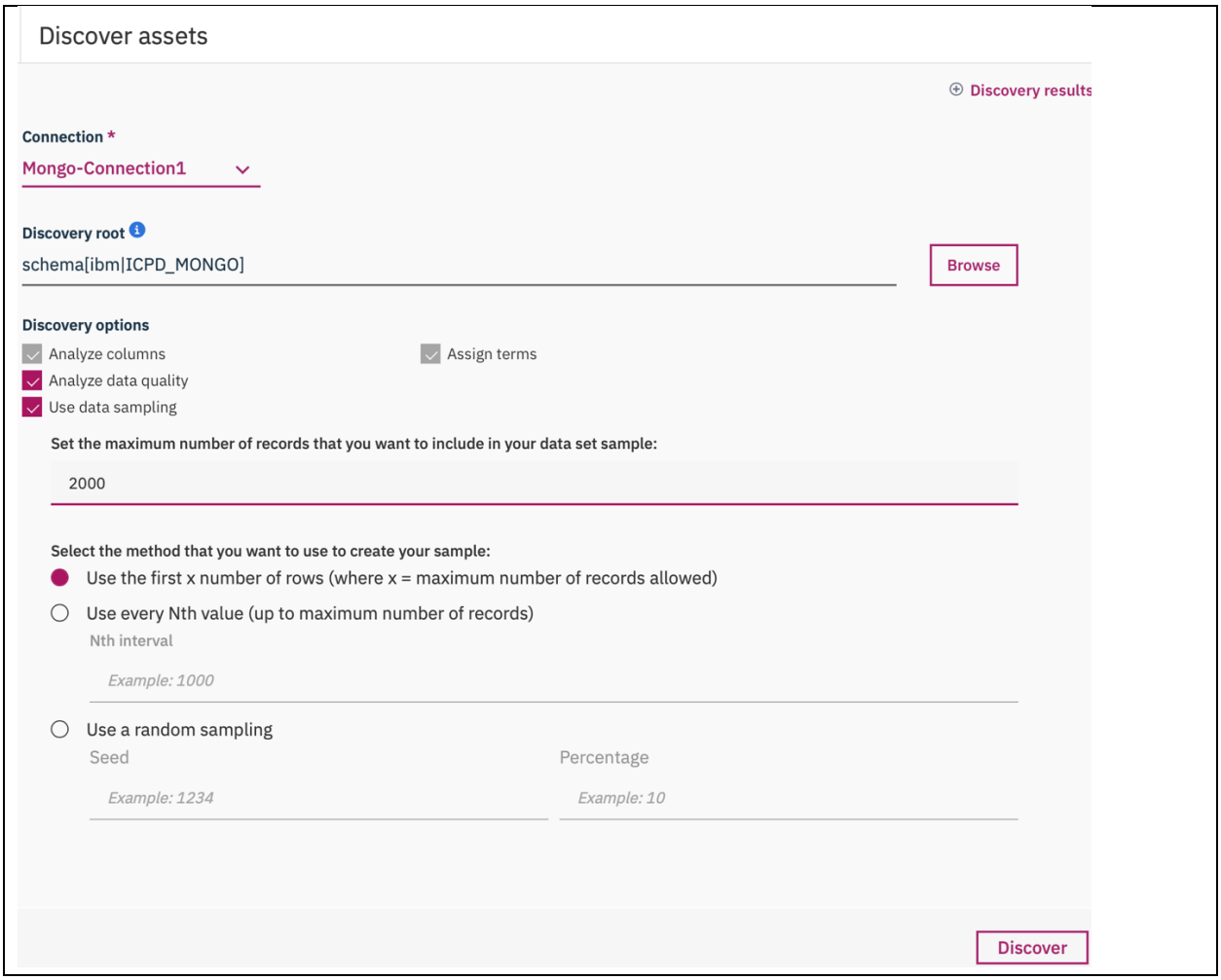


3. Add the **Db2_Connection#** connection Replace # with a unique number	
4. Click on **Browse** and select **'DB2INT1' schema** for discover assets	
5. Check all Discovery options. 6. Set sample dataset as 2000. 7. Click on **Discover** It may take few minutes to complete. Click refresh to get latest status of the discovery job. Wait till discovery job is finished. **Don't forget to document your feedback at **https://pollev.com/turtles****	



5.2. Navigate through discover MongoDB assets

As you discovered Db2 assets in above step, you can also discover assets from MongoDB. Choose the connection named **Mongo_Connection#** (Replace # with a unique number) that you created earlier. For discover roots select **ICPD_MONGO** as the schema.


Discover assets
Discovery results
Connection *
Mongo-Connection1
Discovery root
schema[ibm|ICPD_MONGO]
Browse
Discovery options
Analyze columns
Assign terms
Analyze data quality
Use data sampling
Set the maximum number of records that you want to include in your data set sample:
2000
Select the method that you want to use to create your sample:
Use the first x number of rows (where x = maximum number of records allowed)
Use every Nth value (up to maximum number of records)
Nth interval
Example: 1000
Use a random sampling
Seed
Percentage
Example: 1234
Example: 10
Discover

6. Data Dictionary

Context: A data dictionary contains a business glossary with terms and categories, and information governance policies and rules to ensure data compliance with business objectives. The business glossary is a catalog of assets that defines the character of an enterprise to form a logical structure of your data. Information governance is a quality control discipline for managing, using, improving, and protecting organizational information.

In Cloud Pak for Data you can manually create terms and categories, information governance policies and rules. You can also import glossaries with existing terms, categories, information governance policies and rules. In this lab we will import business glossary and information governance assets related to the travel application.

6.1. Download Glossary

Use the web browser to download the XML file previously created, that contains all business glossary and information governance assets.

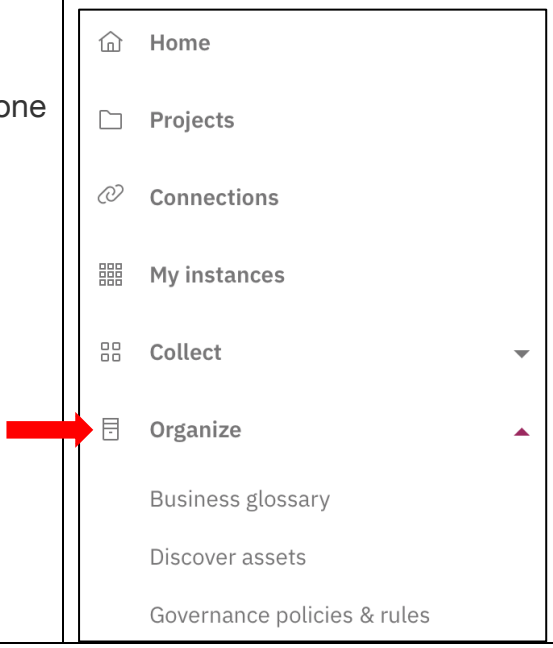
- Go to https://drive.google.com/file/d/1oNQKyu7kDVP65ToPKmFqFRcPyNiSi4KS/view?usp=sharing
- Click on the **Download** button.

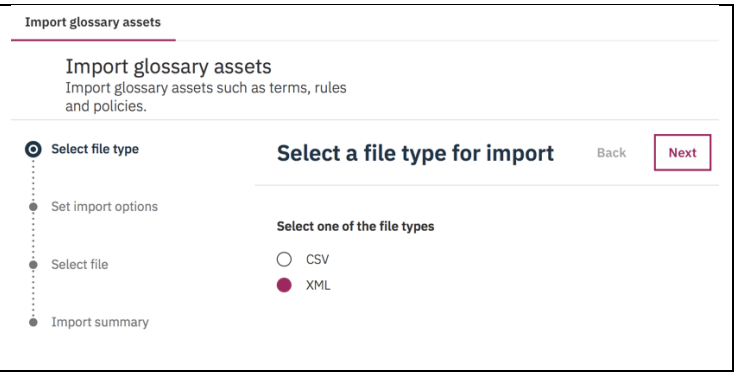
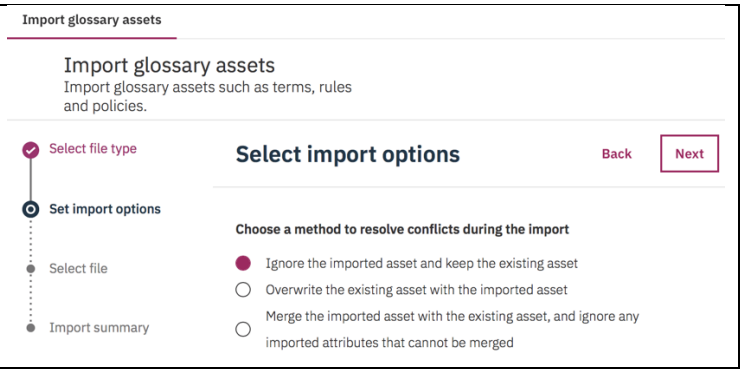
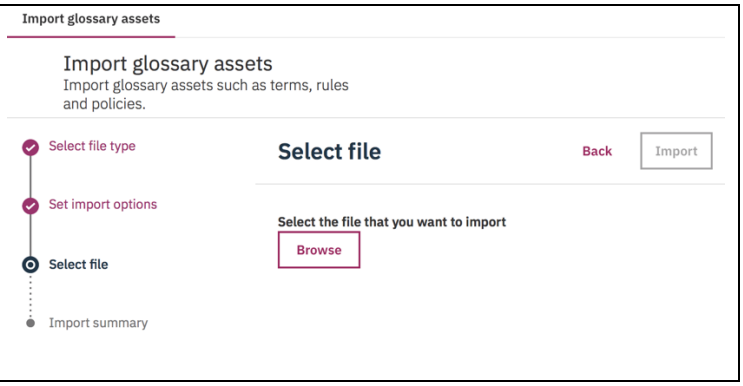
6.2. Import Glossary

1. Go to **Organize**, and select either **Business glossary**, or **Governance policies & rules**.

2. Depending on what view you opened, click one of the following buttons:
   - **Import Categories**
   - **Import Terms**
   - **Import Rules**
   - **Import Policies**

The same **Import Assets** wizard opens, regardless of which button you click.



• Select the **XML** file type • Click **Next**	
• Choose **Ignore the imported asset and keep the existing asset** method to resolve conflict • Click **Next**	
• Browse the XML file to import that you download previously • Click **Import** • Once import complete, it shows import summary of glossary assets	

7. Analytics Project

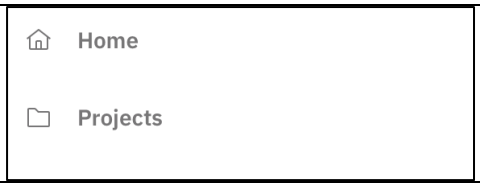
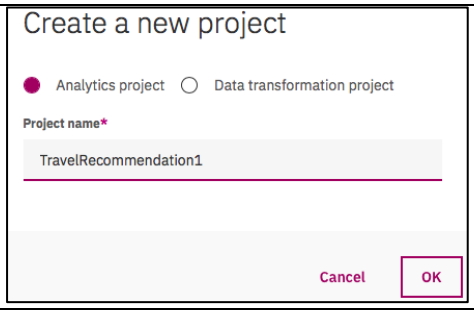
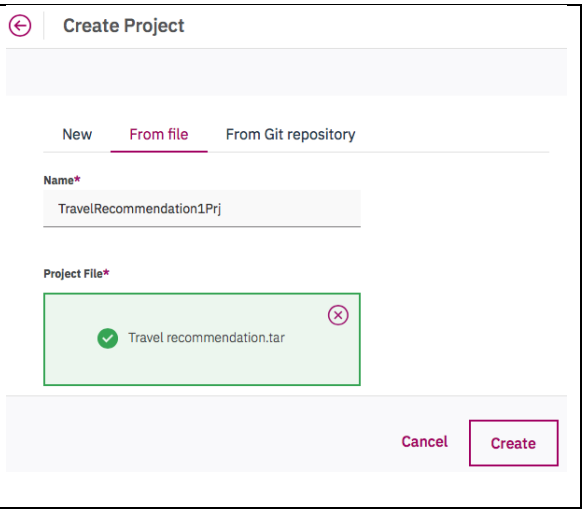
Context: an analytics project is a collection of assets you use to achieve a particular analysis goal. These project assets can include: notebooks, RStudio files, models, data assets, scripts etc. With CPD, you can collaborate with other team members on analytic projects to create machine learning models and visualizations with data from your enterprise. This lab uses preconfigured RStudio scripts that you can use to build a Shiny application.

7.1. Download Analytic Project files

Use the web browser to download the ZGIP file which contains a previously created analytic project file with notebooks and RStudio files.

- Go to https://drive.google.com/file/d/12z0VzLAwPT6JgePvfOsAcNUHpq9eQTTp/view?usp=sharing
- Click on the **Download** button

7.2. Create An Analytics Project

Click on "Folder icon" for **Projects** on left pane and select **New project**	
• Select **Analytics project** • Type **TravelRecommendation#Prj** for project name and click **OK** • Replace # with a unique number	
• Select create project **From file** • Browse the analytic project zip file to import that you download earlier. • Click **Create** • (It may take few minutes) **Some go wrong? Something not make sense? Something you like? Document your feedback at **https://pollev.com/turtles****	

8. Data Virtualization

Context: Data virtualization (DV) integrates data sources across multiple types and locations and turns it into one logical data view. In this case, you have data across two data sources: Db2 and MongoDB. Creating connections to your data sources enables you to quickly view data from across your organization.

8.1. Adding a new data source for Db2

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

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

1. Go to **Collect** > **Virtualized data** > **Menu** > **Data sources**
2. Click **Add** > **New data source > Add connection**
3. Select **Db2-Connection#** that you created earlier and click **Next**

8.2. Adding a new data source for MongoDB

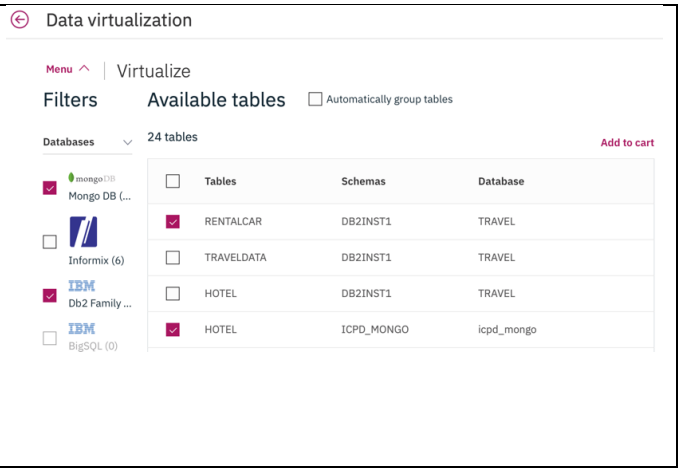
Similar as Db2, define a data connection to MongoDB. Use your existing MongoDB database connection for MongoDB data source.

4. Go to **Collect** > **Virtualized data** > **Menu** > **Data sources**
5. Click **Add** > **New data source > Add connection**
6. Select **Mongo-Connection#** that you created earlier and click **Next**

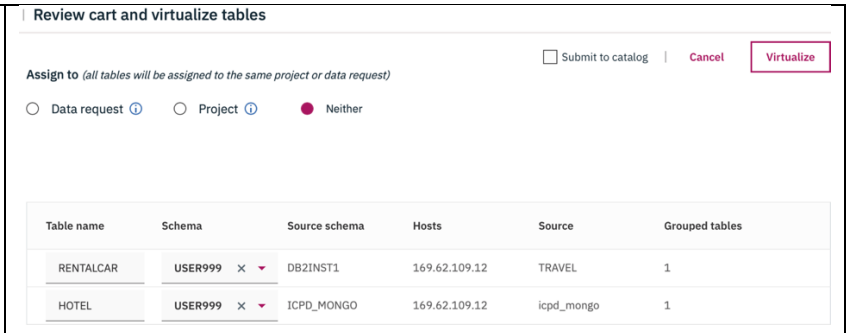
8.3. Select tables for virtualization

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

- Click **Collect** > **Virtualized data** > **Menu** > **Virtualize**
- Select tables **RENTALCAR** from **TRAVEL** database and **HOTEL** from **ICPD_MONGO** database, then click **Add to cart**
- Click **View cart**
- Click **Next**



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



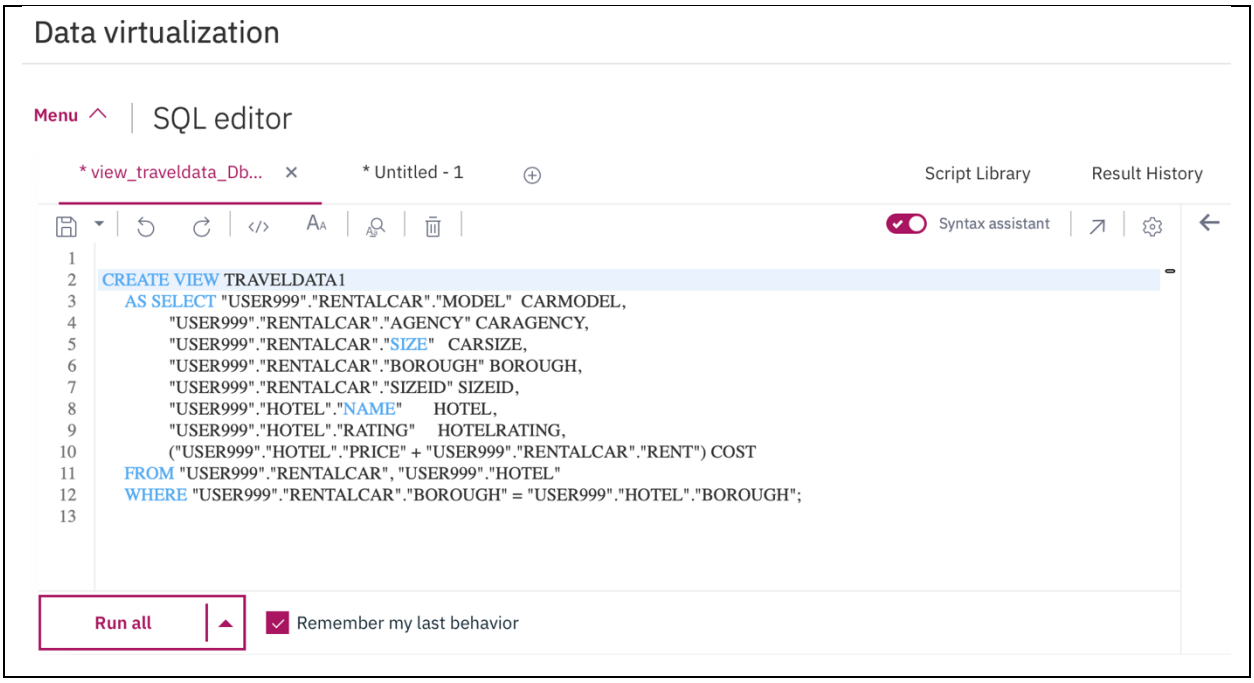
8.4. Creating virtual table

You can create a new virtual table based on existing tables under **My data** section. You can use “drag and drop” or write your own SQL to create the view.

- Click **Collect** > **Virtualized data** > **Menu** > **SQL editor** to access the editor.
- Copy the following SQL statement and paste it on the editor
- Replace # with a unique number for TRAVELDATA# (on the 1st line)
- Click on **Run all**

(If you are login as a different user than admin, you need to change the schema according in the SQL.)

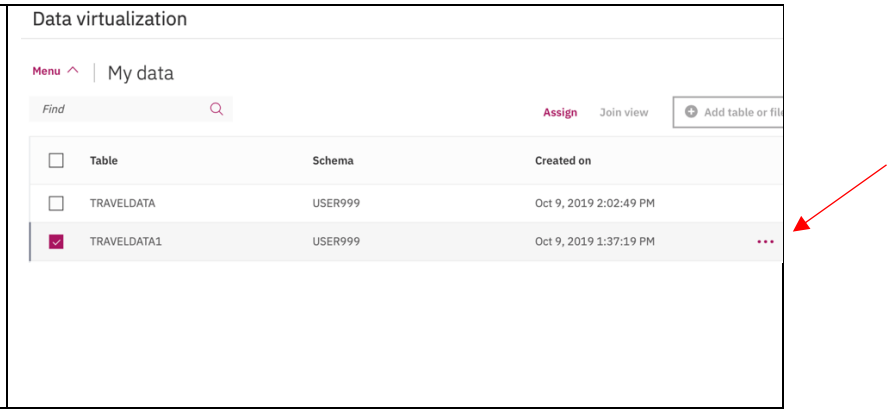
```
CREATE VIEW TRAVELDATA#
     AS SELECT "USER999"."RENTALCAR"."MODEL"  CARMODEL,
               "USER999"."RENTALCAR"."AGENCY" CARAGENCY,
               "USER999"."RENTALCAR"."SIZE"   CARSIZE,
               "USER999"."RENTALCAR"."BOROUGH" BOROUGH,
               "USER999"."RENTALCAR"."SIZEID" SIZEID,
               "USER999"."HOTEL"."NAME"       HOTEL,
               "USER999"."HOTEL"."RATING"     HOTELRATING,
               ("USER999"."HOTEL"."PRICE" + "USER999"."RENTALCAR"."RENT") COST
     FROM "USER999"."RENTALCAR", "USER999"."HOTEL"
     WHERE "USER999"."RENTALCAR"."BOROUGH" = "USER999"."HOTEL"."BOROUGH";
```



8.5. Add virtual table to catalog

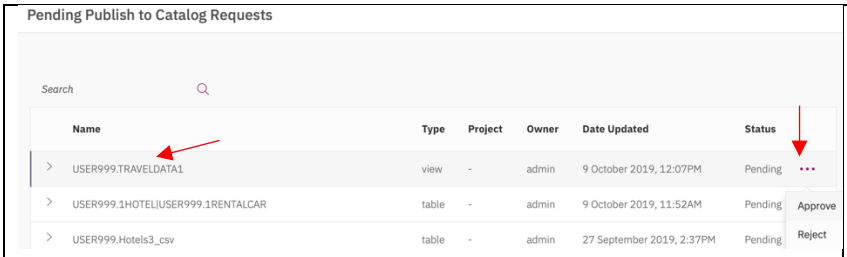
Once you create a virtual table, you can add it to the catalog, making it easily searchable.

- Click **Collect** > **Virtualized data** > **Menu** > **My data** to find the virtual table just created.
- Mark the checkbox associated with virtual table
- Chose **Submit to catalog** from table action
- Click on **Confirm**



8.6. Publish virtualized table

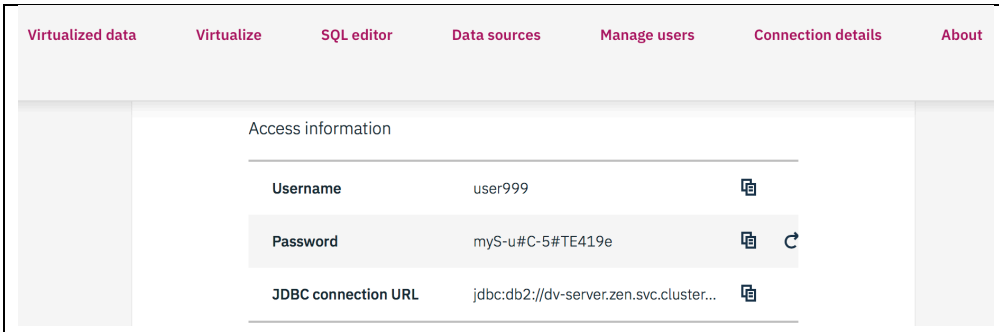
A data steward needs approve the published request before the asset is added to the enterprise data catalog. You signed in as user ‘admin’, it should allow to publish the virtual table.



- Click on access the **Home** page
- Click on **Pending Publish to Catalog Requests**
- Click on icon on left for virtual table **TRAVELDATA#** that you created
- Click on **Approve**

8.7. Access information for virtual table

To access virtual table from external application, you need the JDBC connection information. Click on **Collect** > **Virtualized data** > **Menu** > **Service details** to find out access information. You will use this information later in the RStudio.



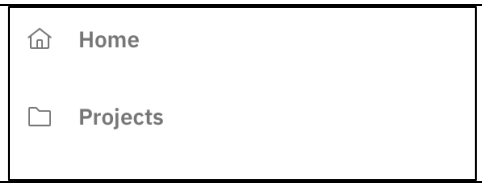
9. RStudio

Context: R is a popular statistical analysis and machine-learning package that enables data management and includes tests, models, analyses, and graphics, and enables data management. RStudio provides an IDE for working with R. This feature is available only if your administrator configures the RStudio Server add-on. For this lab RStudio Server add-on is already installed.

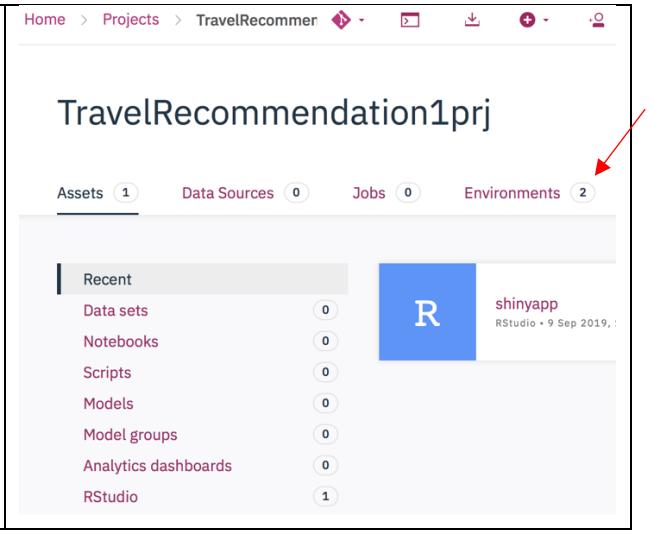
9.1. Enable RStudio Environment

By default, CPD includes Jupyter Notebook Server with Python 3.6. As you need other development environment you can optionally install and enable other development environments. In this lab you need enable the RStudio Server for your analytic project.

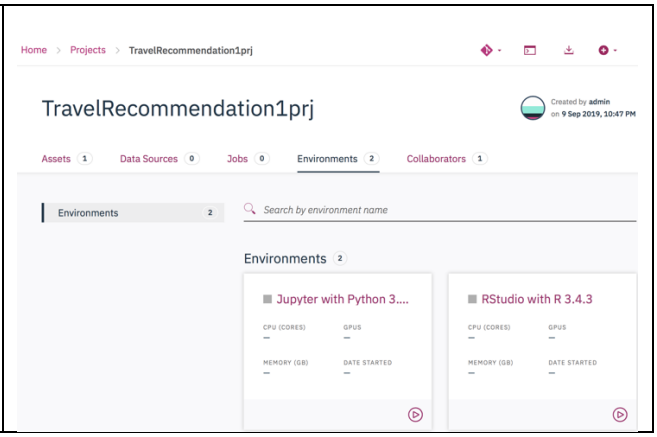
- Click on “Folder icon” for **Projects** on left pane and select your existing **TravelRecommendation#Prj** project (Replace # with a unique number)



- Once inside your project, select **Environments** tab



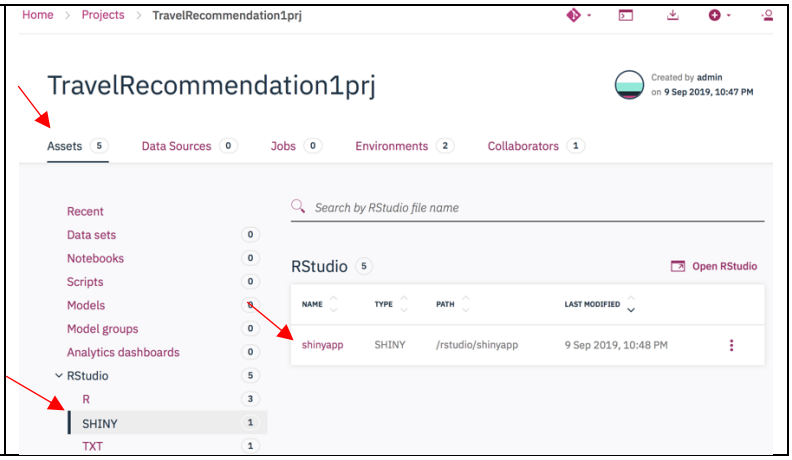
- If the RStudio is not already started, used the ▷ icon on the **RStudio with R 3.4.x** environment tile to start the environment.



9.2. Access RStudio Assets

Once RStudio environment started, you can have access to all RStudio assets associated to the analytic project. There are some R scripts and previously build Shiny package added to the project. You will use the Shiny package called **shinyapp** in this lab.

- Click on **Assets**
- Expand asserts under **RStudio**
- Click on **SHINY** to list the Shiny packages
- Click on **shinyapp** package to open it in RStudio



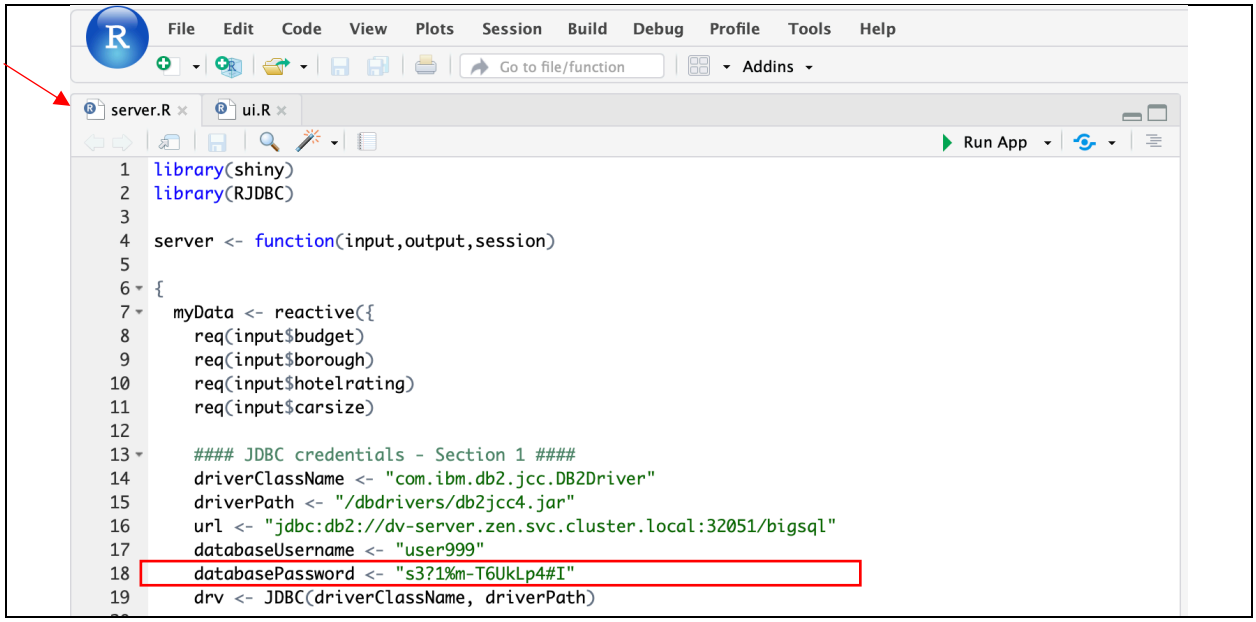
9.3. Shinyapp Package

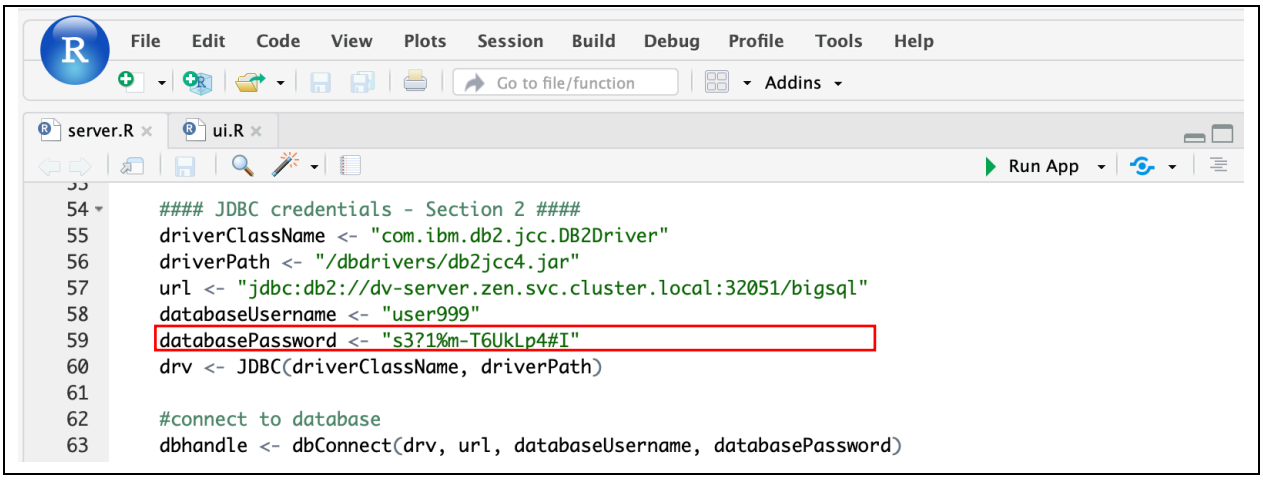
The shinyapp opened in RStudio IDE. App consists of two R scripts. The **ui.R** use for the frontend application and **server.R** connect use for connecting to database and retrieve the dataset.

Click on the **server.R** tab to access the associated R code. You need to change the **databasePassword** value with your cluster password. Review step 8.7 to find out the cluster password.

9.4 Update cluster password in Shinyapp

There are two places in the Server.R file where you need to replace this **databasePassword** value.

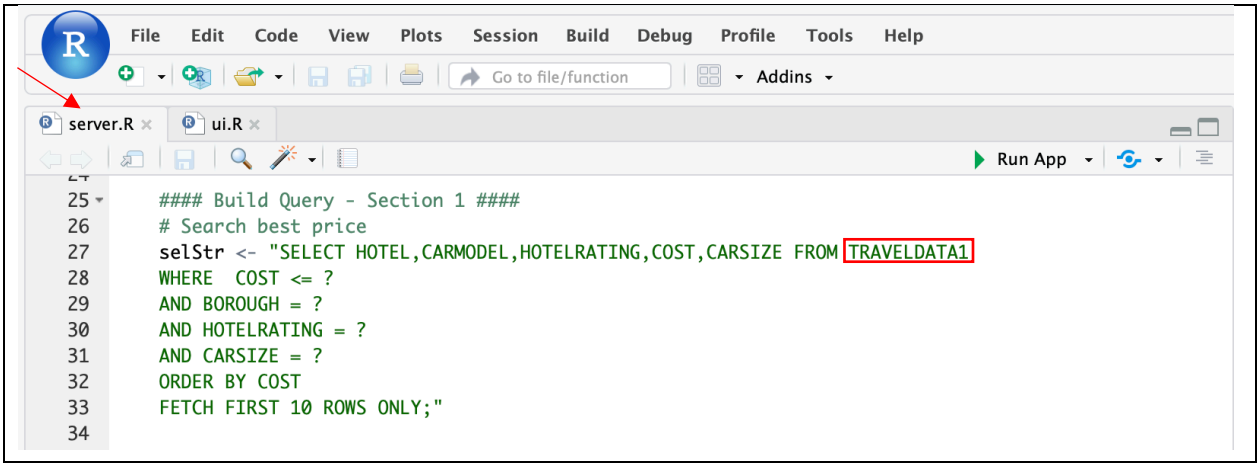


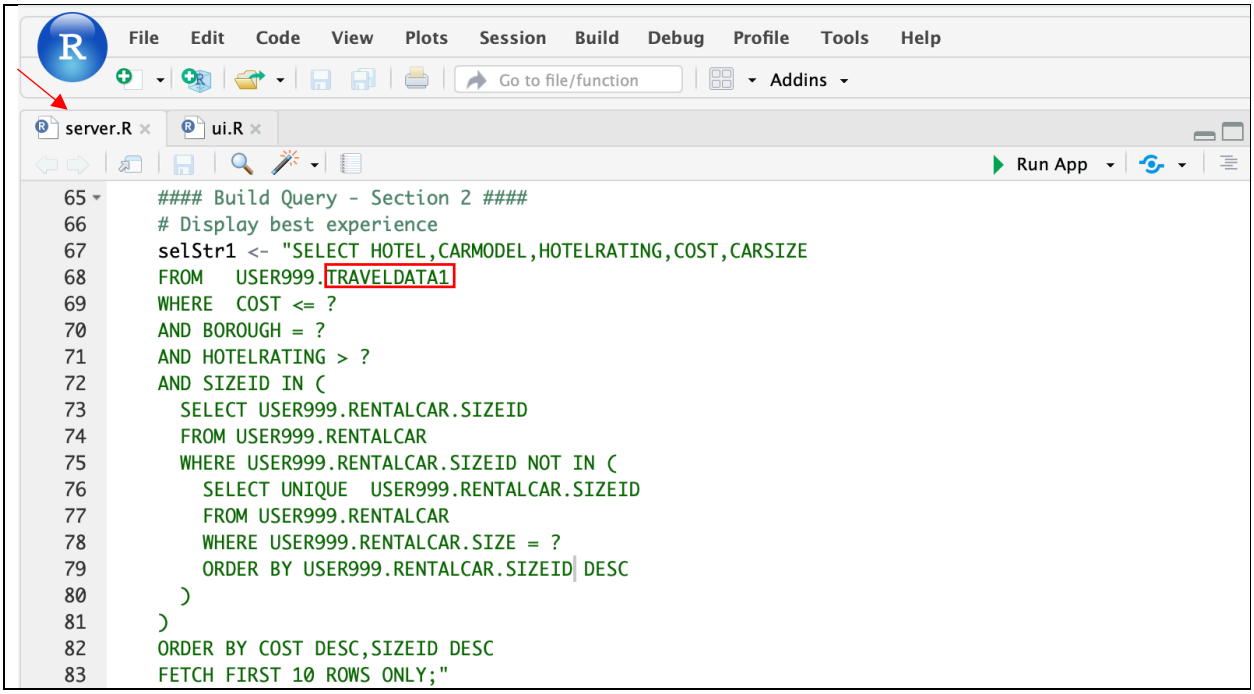


Save the server.R file.

9.4. Update virtual table name in Shinyapp

There are two queries in the Server.R file. In these two queries you need to update the virtual table name that you crated in DV.



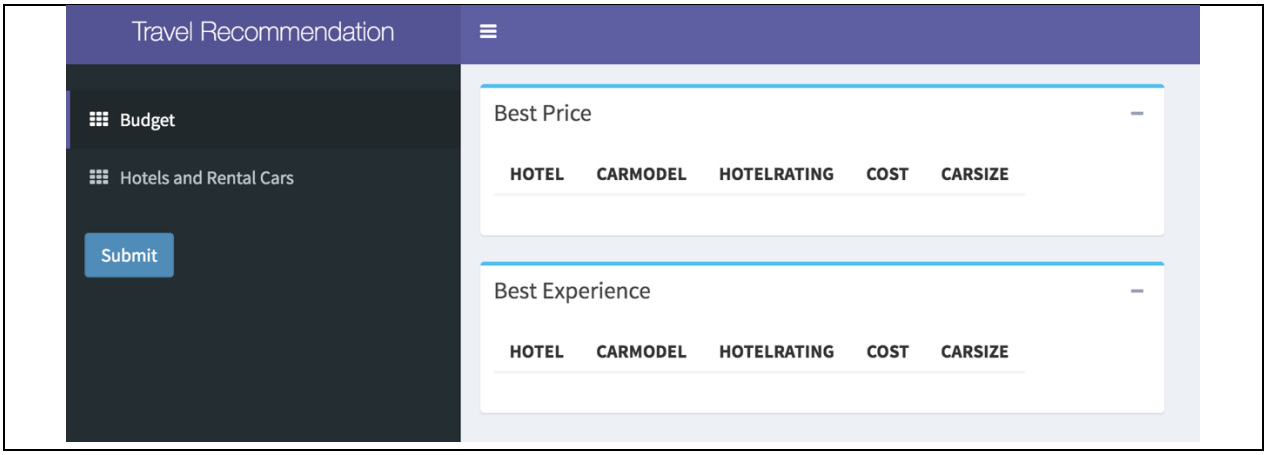


Save the server.R file.

9.5. Test the Shinyapp

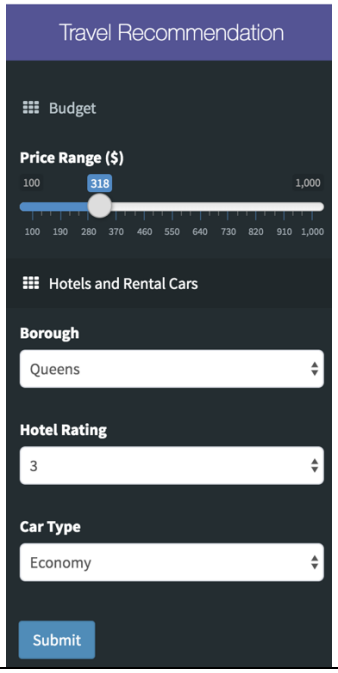
Click on **Run App** to spawn the application.

The shinyapp will open the Travel Recommendation application in a new window. The left pane collects input from the user. Depending on the input, the right pane displays suggested hotel and car information.

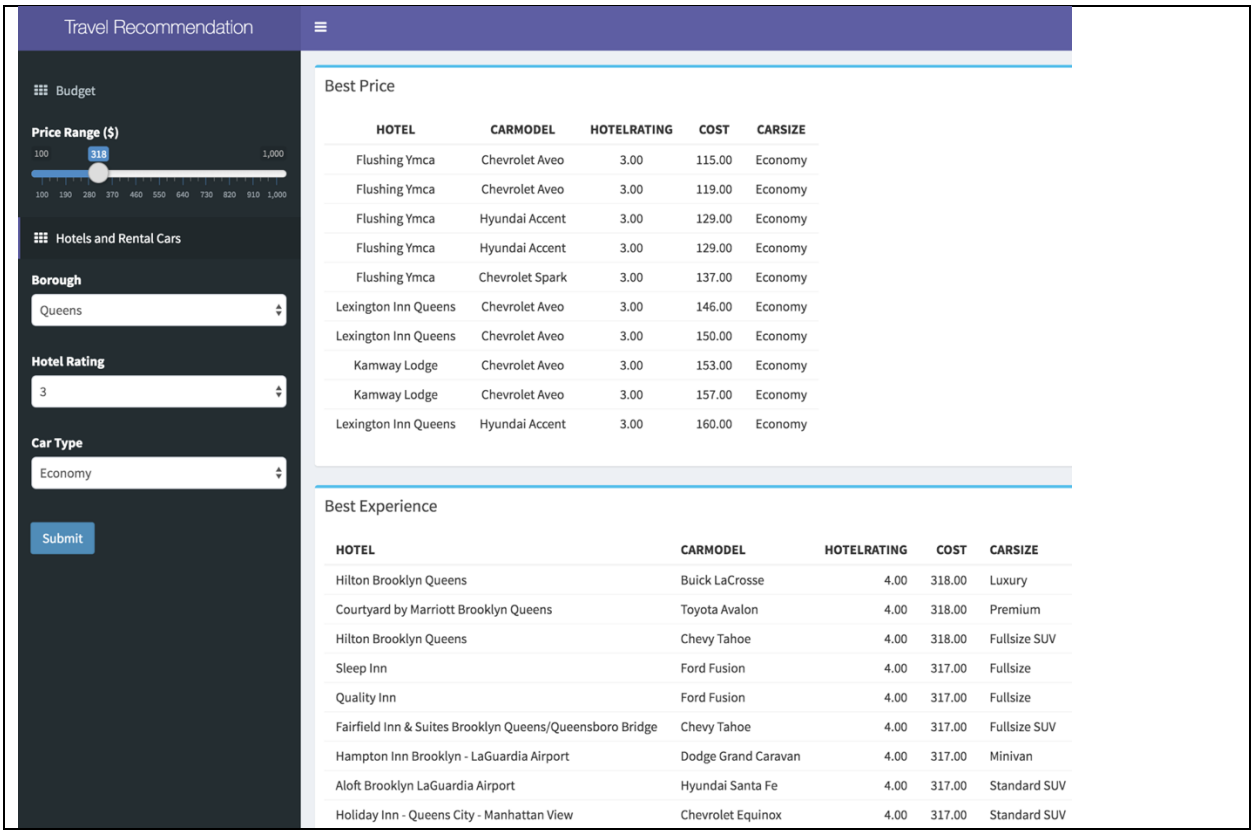


The user can provide their daily budget for hotel and rental cars.
The application also collects information about the place they want to stay, preferred hotel rating and car type.

- Provide a **Price Range** using slider
- Choose **Borough** where you want to stay in New York
- Choose desired **Hotel Rating** from drop down menu
- Similarly select a **Car Type**
- Click **Submit**

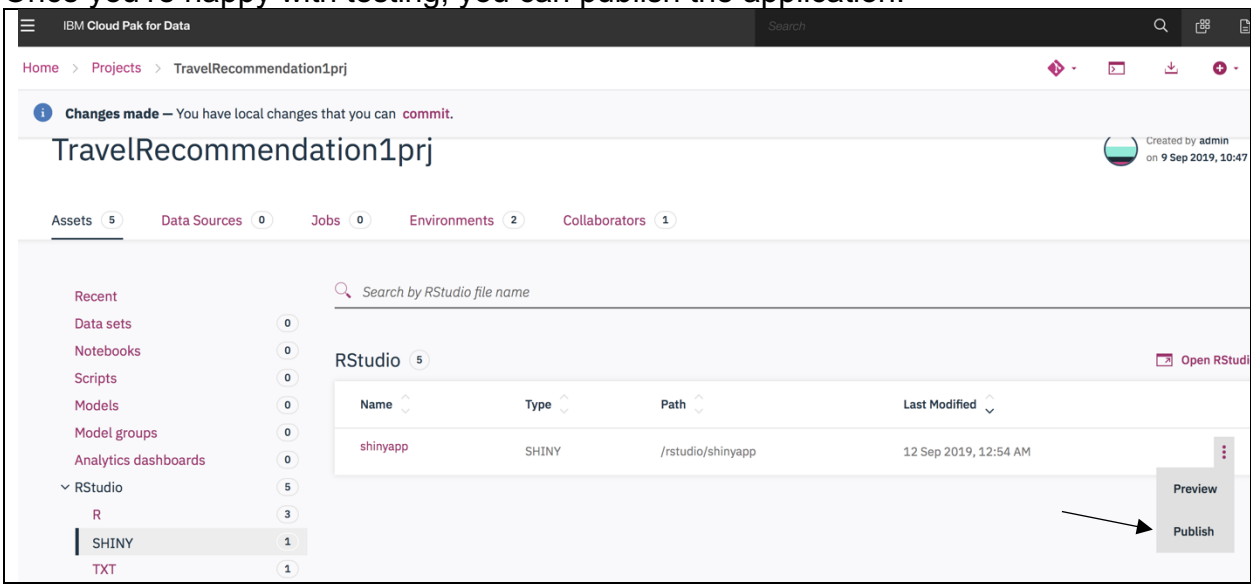


Depending on the input, the application will generate a list of cheapest possible hotel and car combination under Best Price. The Best Experience section show highest comfort hotel and car combination.

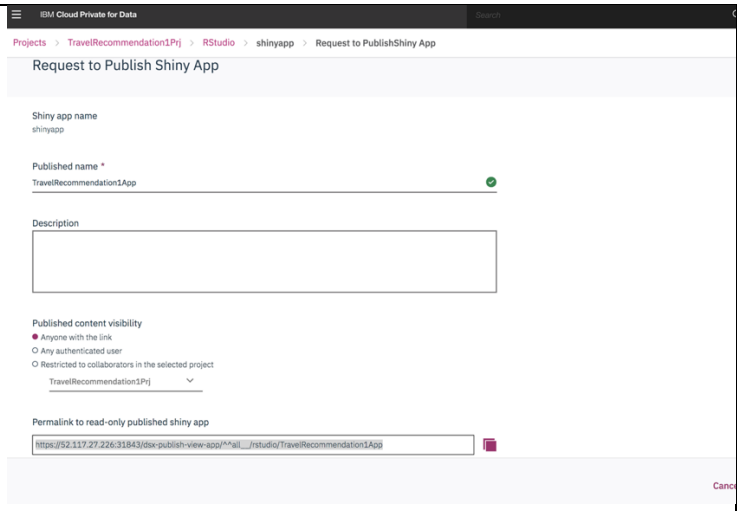


9.6. Externalize the application

Once you're happy with testing, you can publish the application.



- Give a unique published name, e.g., TravelRecomendataion<#> APP. System will automatically generates a permalink.
- Click on **Publish**.



Cloud Pak for Data has a built-in web server under Watson Studio. Once the application is ready, you can deploy it on the platform and anyone with the published web link can easily access it on the internet. Access to this app can be governed and managed in CPD.

https://52.117.27.226:31843/dsx-publish-view-app/^^all__/rstudio/TravelRecommendation1App

You can find out your published application from UI using following link:

```
https://<Cluster IP>:31843/#/publishedAssets/?_k=vmhmgo
```