Lab Center – Hands-On Lab

Session 5934

Trust the Art of Insight in the Age of AI through the Use of Data Governance

sanjitc@us.ibm.com

IBM

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 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.

We Value Your Feedback

Don’t forget to submit your Think 2020 session and speaker feedback! Your feedback is very important to us – we use it to continually improve the conference.

Access the Think 2020 agenda tool to quickly submit surveys from your smartphone or laptop.

Table of Contents

1 Introduction

AI enabling assets such as ML models within the enterprise are the next generation of information assets in today's world. Cloud Pak for Data with governance can enables a culture where data governance is an enabler of workforce productivity through self-service as well as a conformance force for regulatory obligations. How does one extend the concepts of data Governance to AI? Is it even possible? Cloud Pak for Data from IBM can make this a reality. This lab shows how the use of foundational Governance services to make this a natural evolution of Enterprise Data Governance.

2 Prerequisites

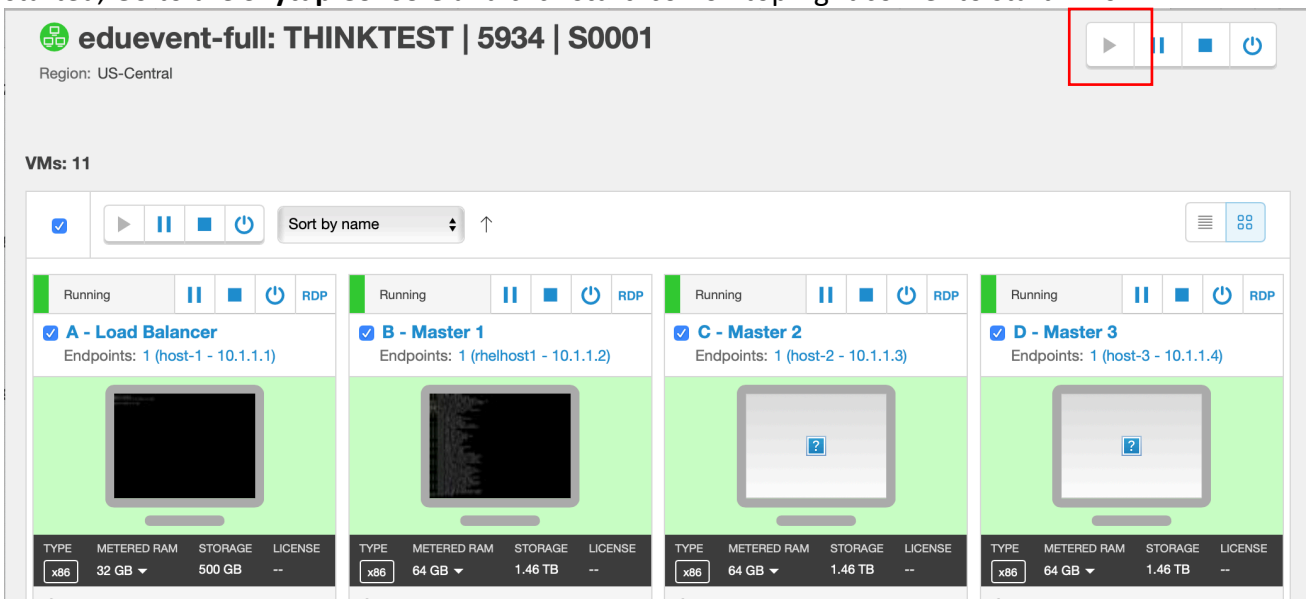
- Access to an operational Cloud Pak for Data (v.2.5 or higher) cluster
- Watson Knowledge Catalog (WKC v.3.0) deployed on the CPD cluster

3 Getting Started

You just need Cloud Pak for Data (CPD), User Interface (UI) for this lab. Each student needs to use their unique URL to access the CPD UI. URL can be found with the Load Balancer entry under Ports column. For example: **https://services-uscentral.skytap.com:xxxxx**

Use the Google Chrome to access the CPD UI.

By default, all VMs should already started and shows as running with green annotation. If VMs not started, Go to the **Skytap Console** and click **start** icon on top right corner to start VMs.

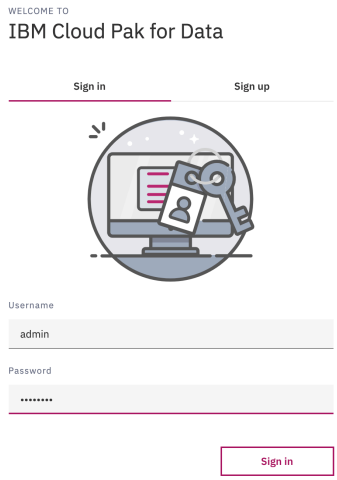


4 Access Credentials

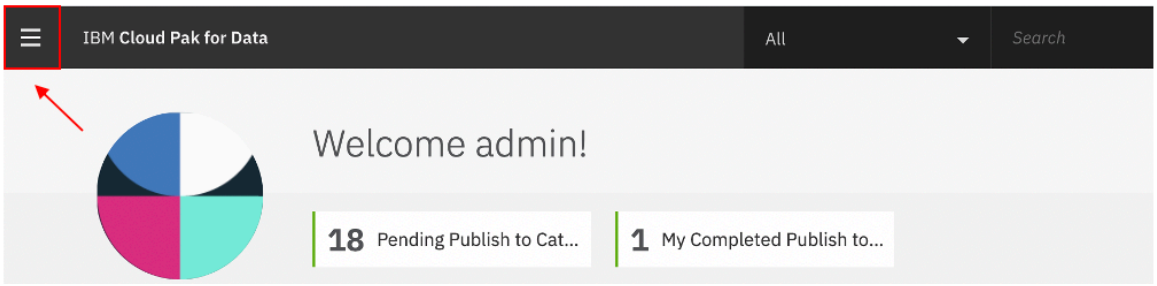
4.1 Sign into Cloud Pak for Data web console as Administrator

You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data UI. Starting from here all instruction need to execute on Cloud Pak for Data UI only.

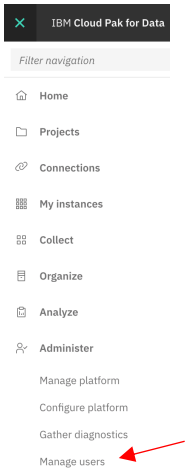
Login as **admin** to give necessary privileges to the individual user **student<id>** that you ae going to use later on this lab.



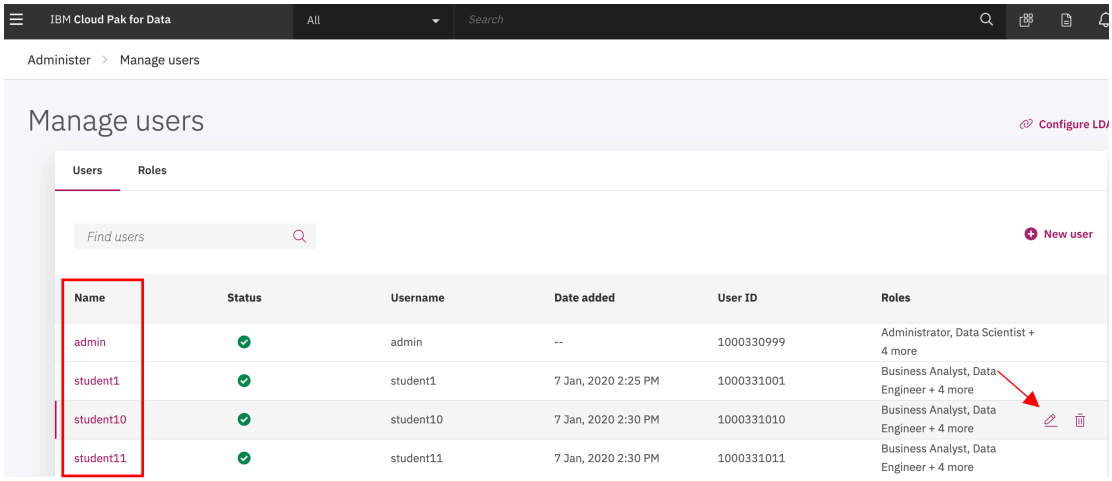
- Enter username **admin**
- Password is **password**
- Click the **Sign In** button



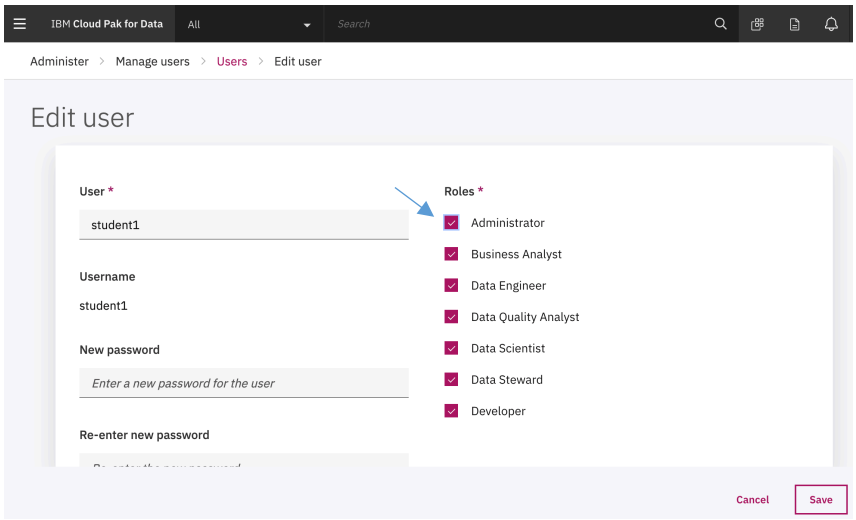
Click on the **Navigation menu** in the top left corner.



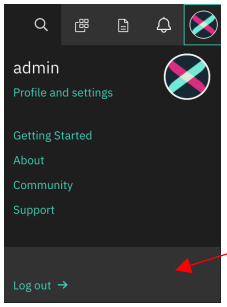
Click **Manage users**



Edit your individual user **student<id>** by using the pen icon on the right side



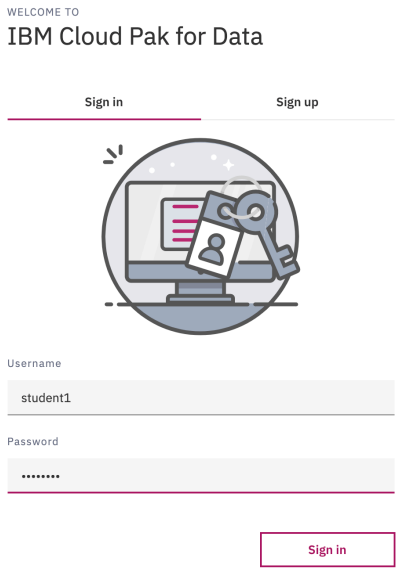
- Grant the **Administrator** role to your individual user **student<id>**
- Click **Save**



Log out from **admin** user

4.2 Sign into Cloud Pak for Data web console as student<id>

Login as your individual student<id> to continue with remaining lab.

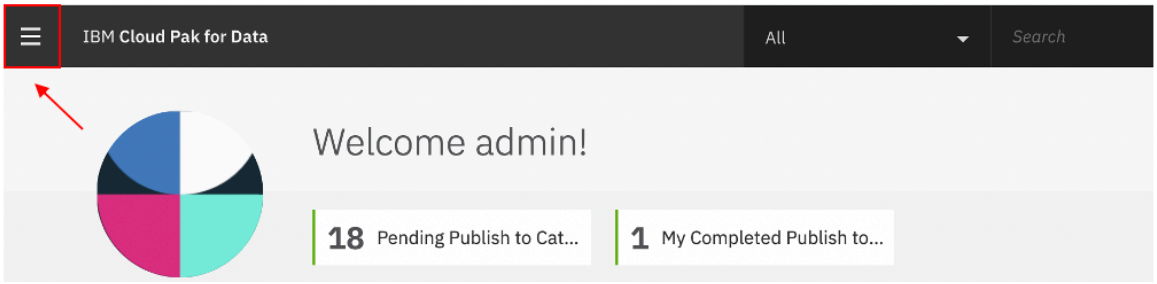


- Enter username **student<id>** (Please use your unique student id)
- Password is **password**
- Click the **Sign In** button

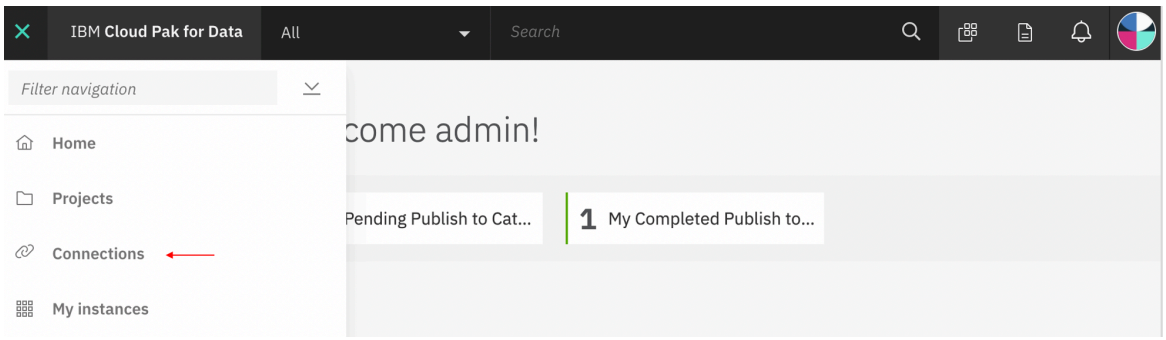
5 Create Global Connection

You need a connection asset for a data source in an analytics project, so you can access data to or from it. A global connection contains necessary information to access respective data source.

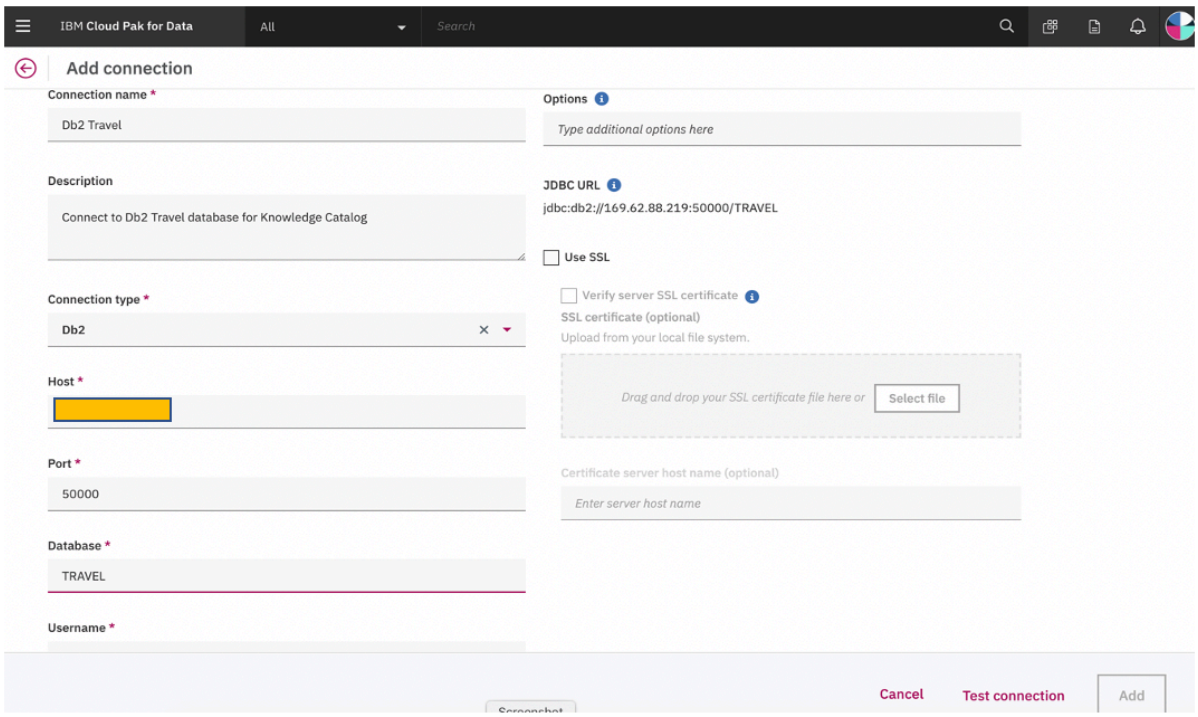
In this task you create a global connection to access necessary data from a Db2 database.



Click on the **Navigation menu** in the top left corner.



- Click **Connections**
- Next click **Add connection**

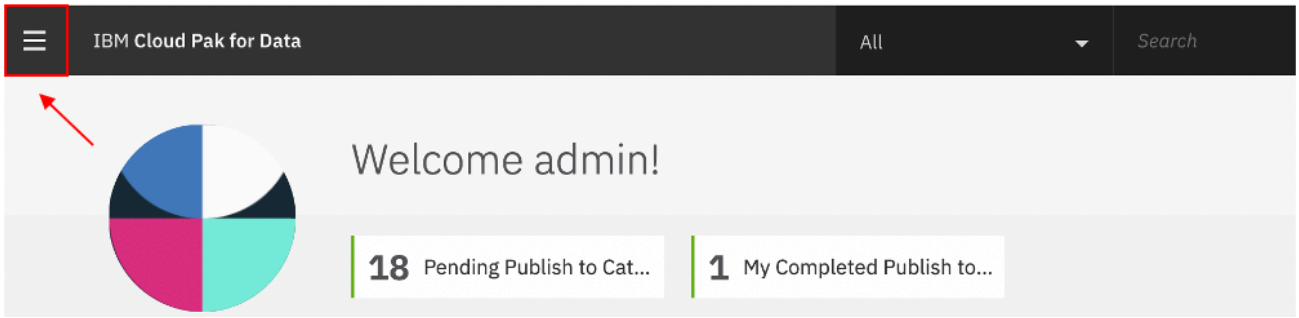


- Enter a Connection Enter a Name of **Db2 Travel**.
- Enter a Description of **Connect to Db2 Travel database for Knowledge Catalog**.
- Connection type of **Db2**.
- Use the **169.62.88.219** in the Hostname or IP Address field.
- Port of **50000**
- Enter a Database of **TRAVEL**.
- Enter a Username of **db2inst1**.
- Use **password** for the Password field.
- Click **Test connection** for verify the connection.
- Click the **Add** button.

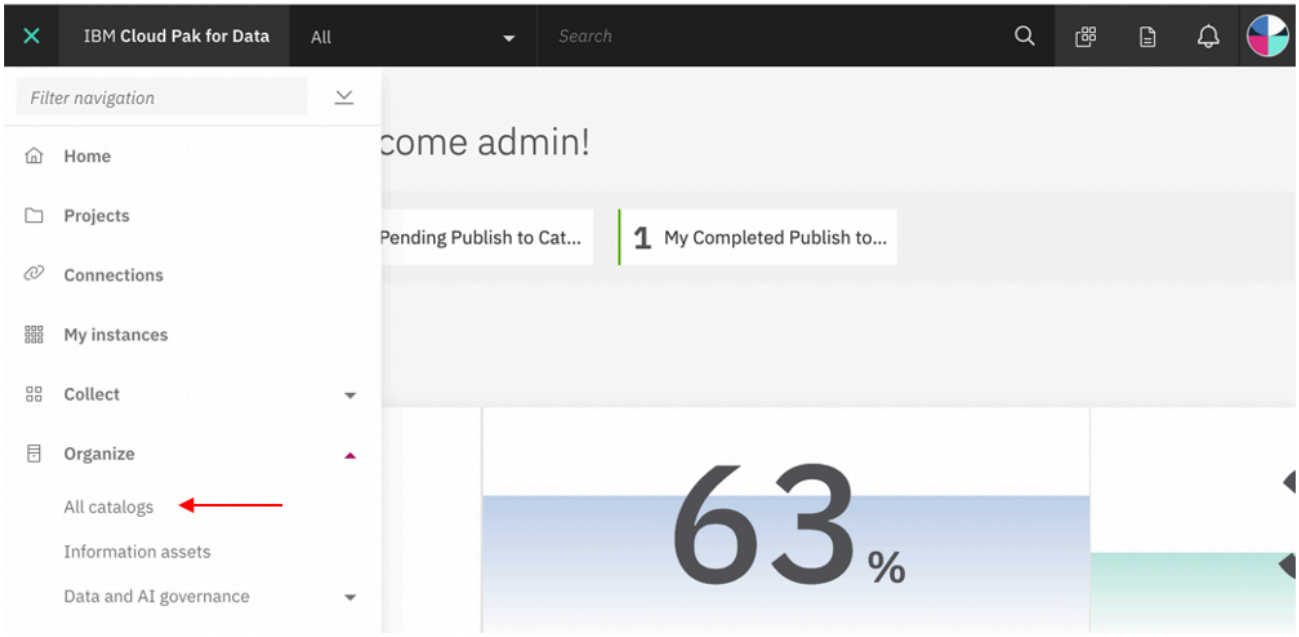
Note: If you are sharing this cluster with other students, connection could be already created by others and you can just use it.

6 Create a Knowledge Catalog

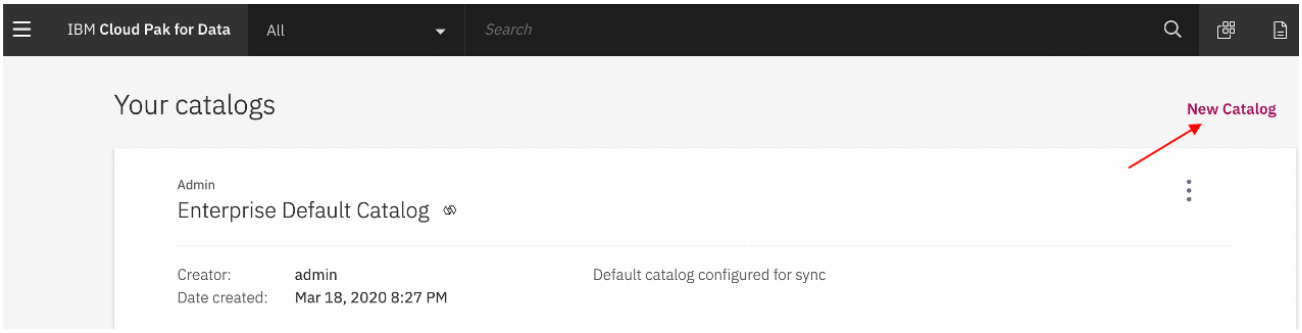
In this task you create a governed Knowledge Catalog. Watson Knowledge Catalog provides a secure enterprise catalog management platform that is supported by a data governance framework. A catalog connects data and knowledge with the people who need to use it. The data governance framework ensures that data access and data quality are compliant with your business rules and standards.



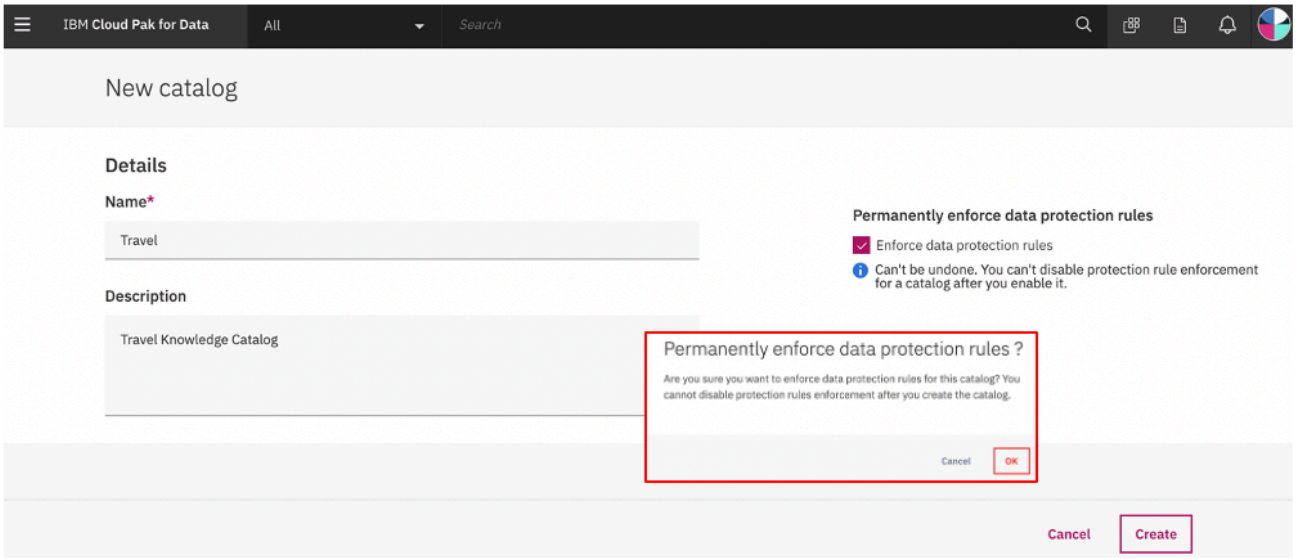
Click on the **Navigation menu** in the top left corner.



Click **Organize > All catalogs** to access catalogs.



Click the **New Catalog** button in the top right corner.



- Enter a Name of **Travel<id>**.
- Enter a Description of **Travel Knowledge Catalog**.
- Select the **Enforce data polices** checkbox.

The **Permanently enforce data protection rules** warning dialog will be displayed, asking if you are sure you want to set this option and informing you that the setting is permanent.

- Click the **OK** button.

By default, access to data assets in a catalog is only restricted by the privacy settings of the data assets. Privacy settings and policy rules can limit which members of the catalog can view and use the assets. You can implement data policies to restrict access to data based on the contents of the data. Data policies help you control data access and ensure that the right people can access the right data. Selecting the option to **Enforce data protection rules** enables the enforcement of data policy rules to allow or deny access to a data asset or redact, substitute or mask data at the column level.

Setting this option for a catalog is a good best practice. Once it is enabled, it cannot be undone, but it does not restrict or impede any functionality, it provides additional security measures to protect data assets.

- Click the **Create** button.

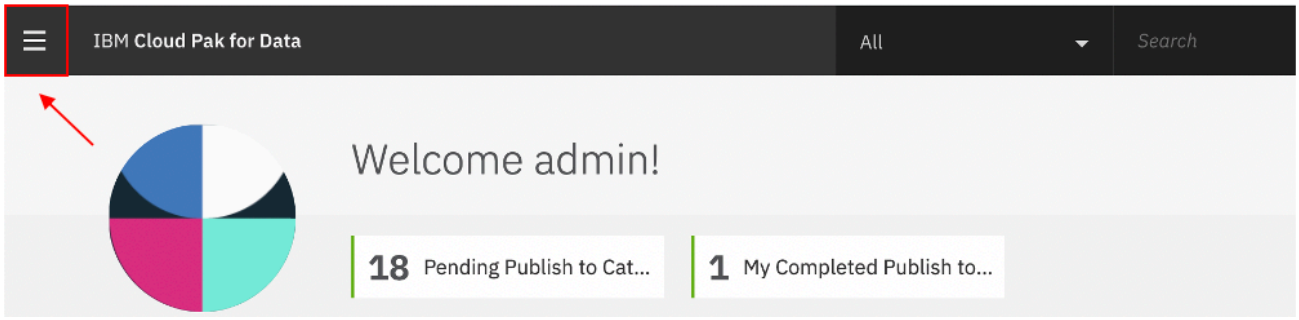
7 Create Analytic Project

In this task you will create two analytical projects.

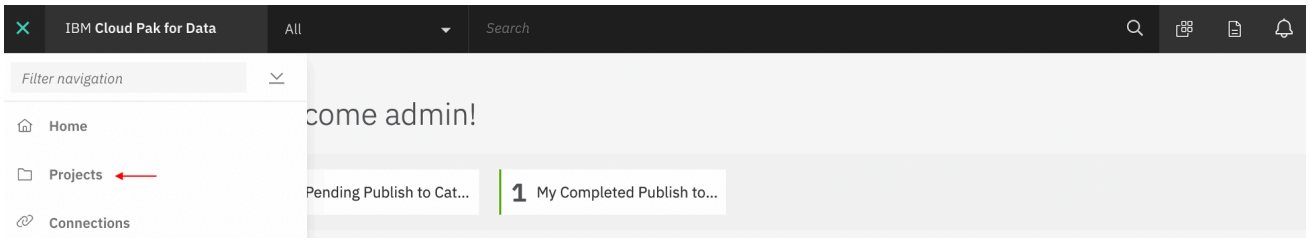
The first project, which will be named **Travel**, will be used for developing a R application, to collaborate and build the analytic and AI assets et. You will add travel related data assets from the **Travel** knowledge catalog to this project and do some shaping of the data using the **Data Refinery** to prepare the data for analytical insights.

The 2nd project, which will be named **Auto Discovery**, will be used to demonstrate the auto discovery capabilities of Watson Knowledge Catalog. When you catalog a **Connection**, you can choose the option to automatically discover data assets. The discovered assets are added to a analytical project as a temporary holding area for review. You will use a separate project for auto discovery, so it does not disrupt the data analytics project. Once data assets are discovered and added to a project, you can review them, determine which assets are relevant and then publish them to a Knowledge Catalog.

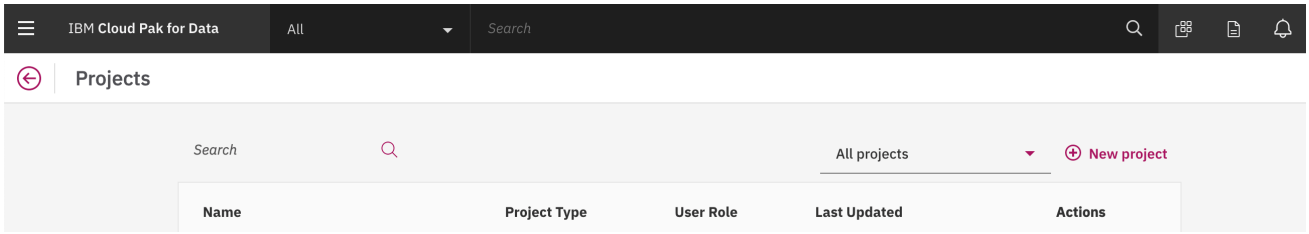
7.1 Create the Travel Analytic Project



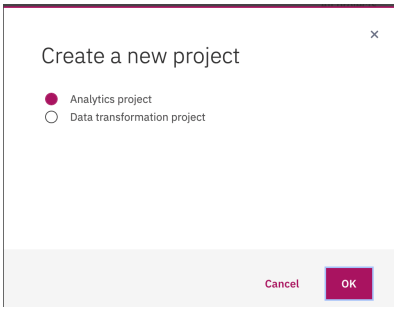
Click on the **Navigation menu** in the top left corner.



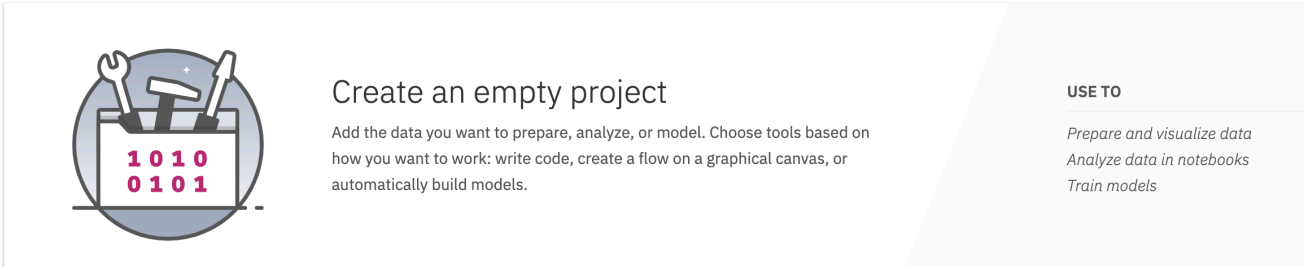
Click **Projects**



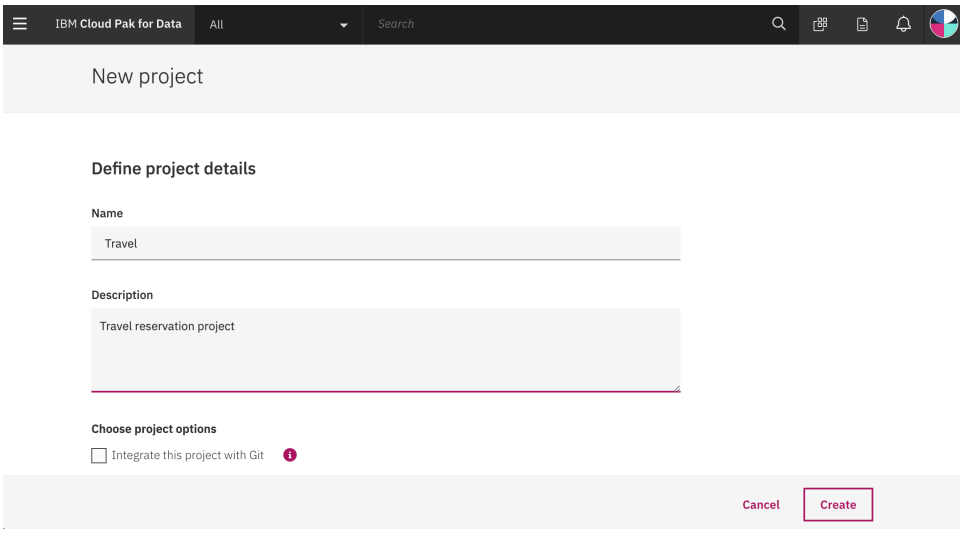
Click **New project**



If **Create a new project** popup window appeared, Choose **Analytics project** and click **OK**.



Click on **Create an empty project**.

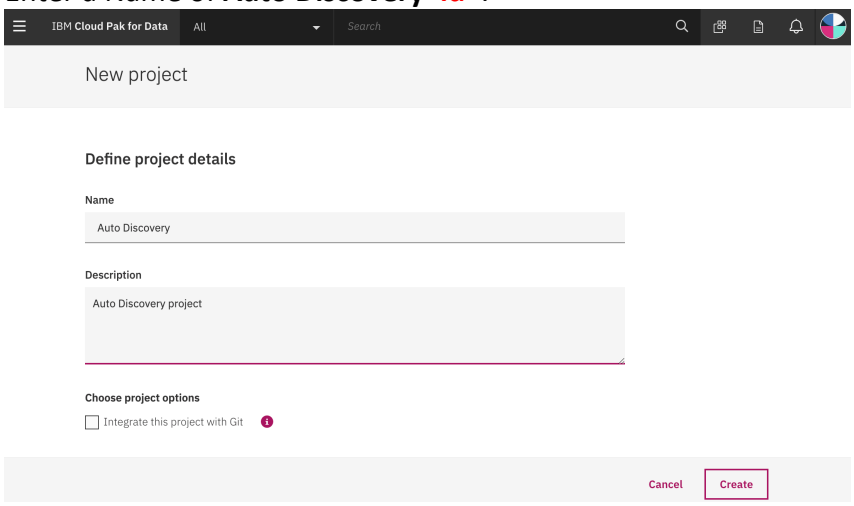


- Enter a Name of **Travel<id>**.
- Enter a Description of **Travel reservation project**.
- Click the **Create** button.

7.2 Create the Auto Discovery Analytic Project

Same as the **Travel** project, create second analytical project called **Auto Discovery**.

- Enter a Name of **Auto Discovery<id>**.



- Enter a Description of **Auto Discovery project**.
- Click the **Create** button.

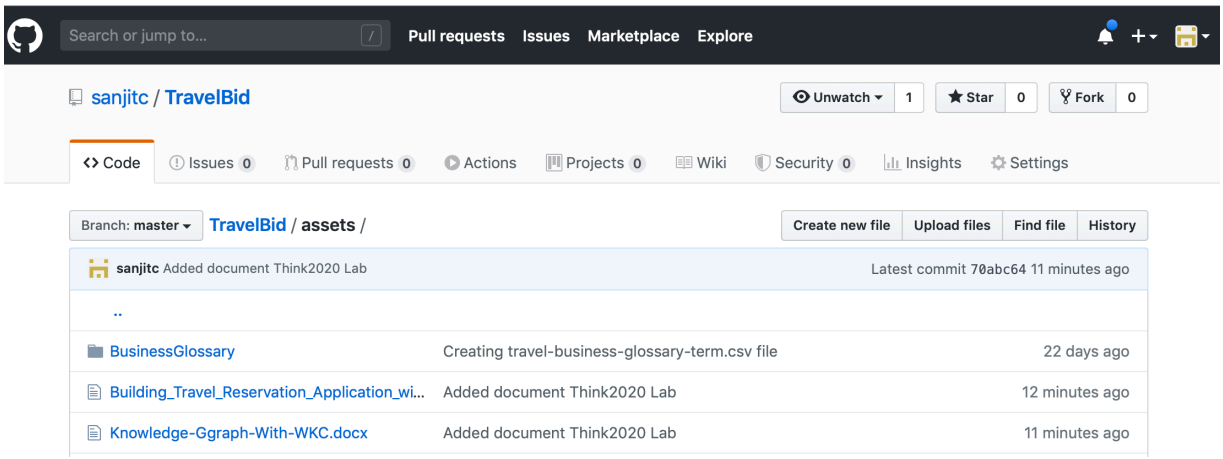
8 Discover and Catalog Data Assets

In this task, you will discover and catalog unstructured data assets from the local file system and structured data assets from an existing **Db2 database**. This will introduce you to the three methods available to discover and catalog data assets; **Local files**, **Connected asset** and **Connection**. You will use these methods to catalog data assets into the newly created Knowledge Catalog and then tag them for users to easily find them, understand their content and make them available throughout Cloud Pak for Data, for use during data preparation and within models, dashboards and notebooks.

8.1 Download Documents Catalog

First download two sample document from GIT in your local machine.

- Go to https://github.com/sanjitc/TravelBid/tree/master/assets



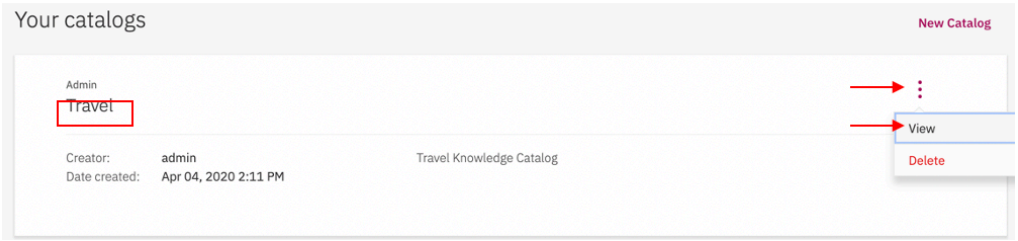
- Download **Building_Travel_Reservation_Application_CPD.pdf** and the **Knowledge-Graph-With_WKC** files. Chose on each individual file and click **Download** tab to start download.

8.2 Add Unstructured Data to Catalog

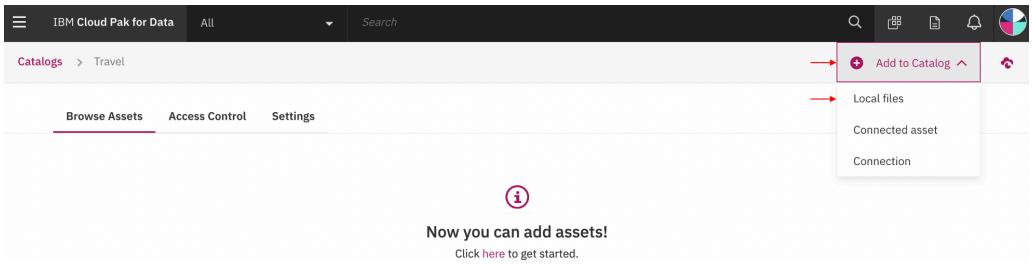
You can see a preview of the contents of many types of data assets and some types of analytical assets in projects and catalogs. These unstructured data can be PDF, text, images, Microsoft Excel documents etc.

You must have Editor or Admin role in a catalog to add a data asset from a file

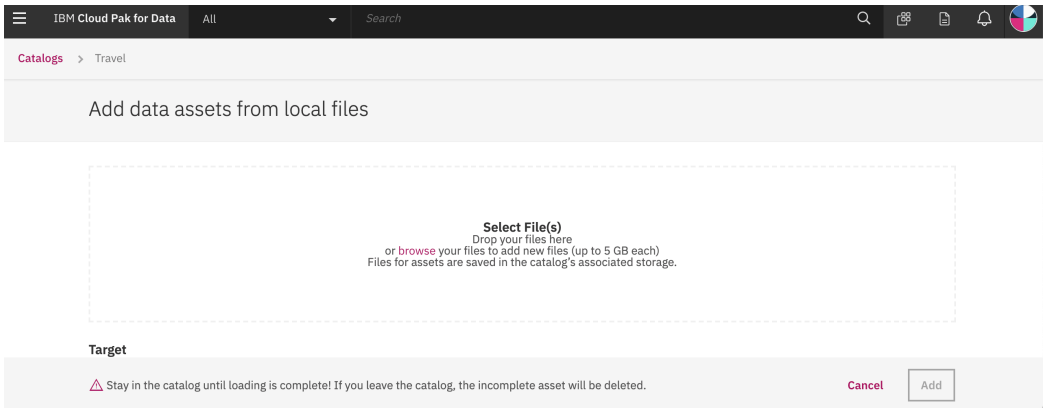
Click on the **Navigation menu** in the top left corner and access catalog from **Organize > All catalogs**.



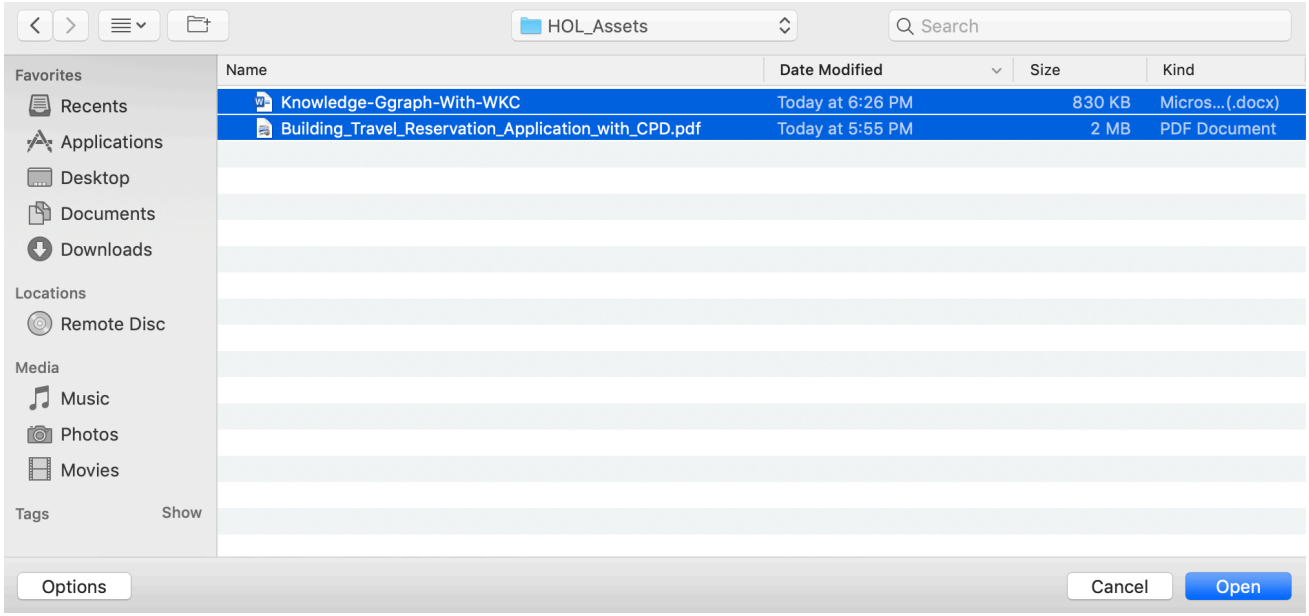
- Click on the **configuration** icon of the **Travel<id>** catalog.
- Select **View**



- To add a local file to catalog, click on **Add to Catalog** > **Local files**.

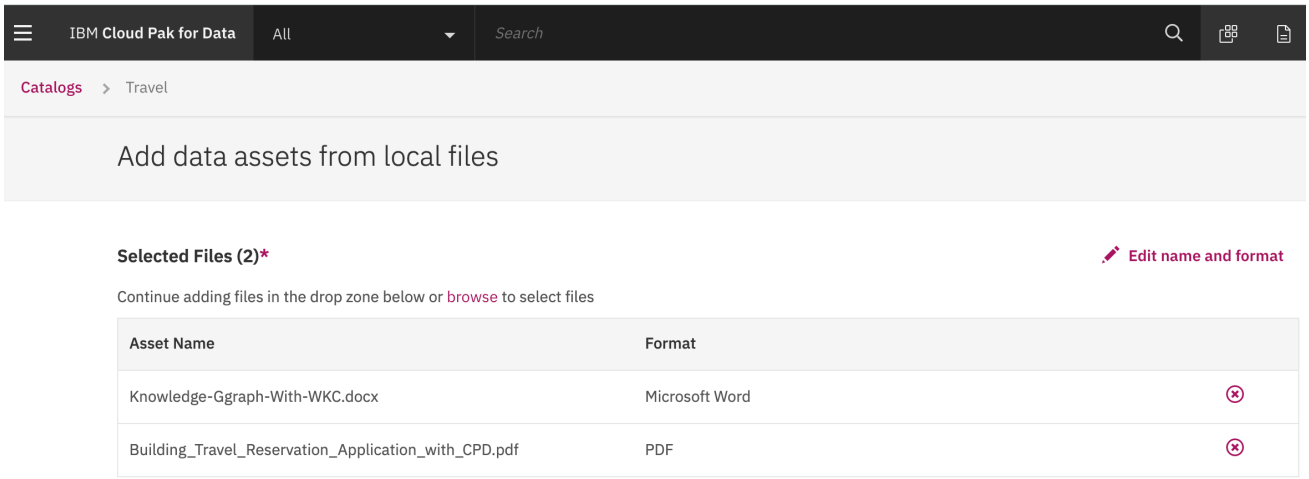


- Click the **browse** link in the **Select File(s)** section to bring up the file selection dialog.



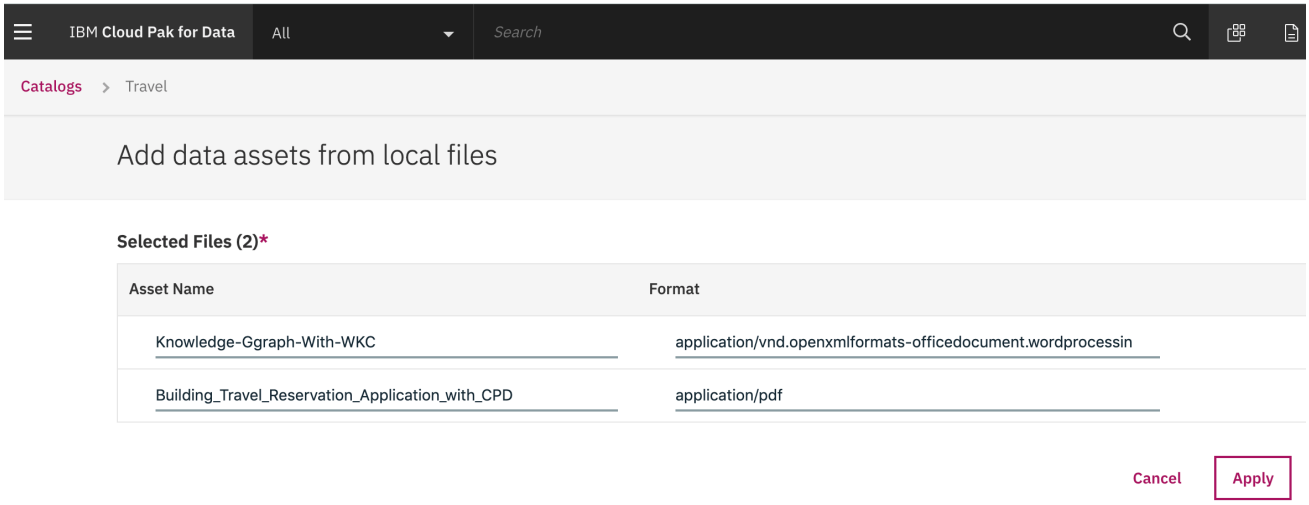
- Locate the **"Building_Travel_Reservation_Application_CPD.pdf"** and the **"Knowledge-Graph-With_WKC"** files under the "**Downloads**" directory on your local machine. Select both using the **Ctrl or Command key** on your keyboard (Ctrl+Click for Windows and Command+Click for MacOS).

- Click the **Open** button to begin cataloging the files.

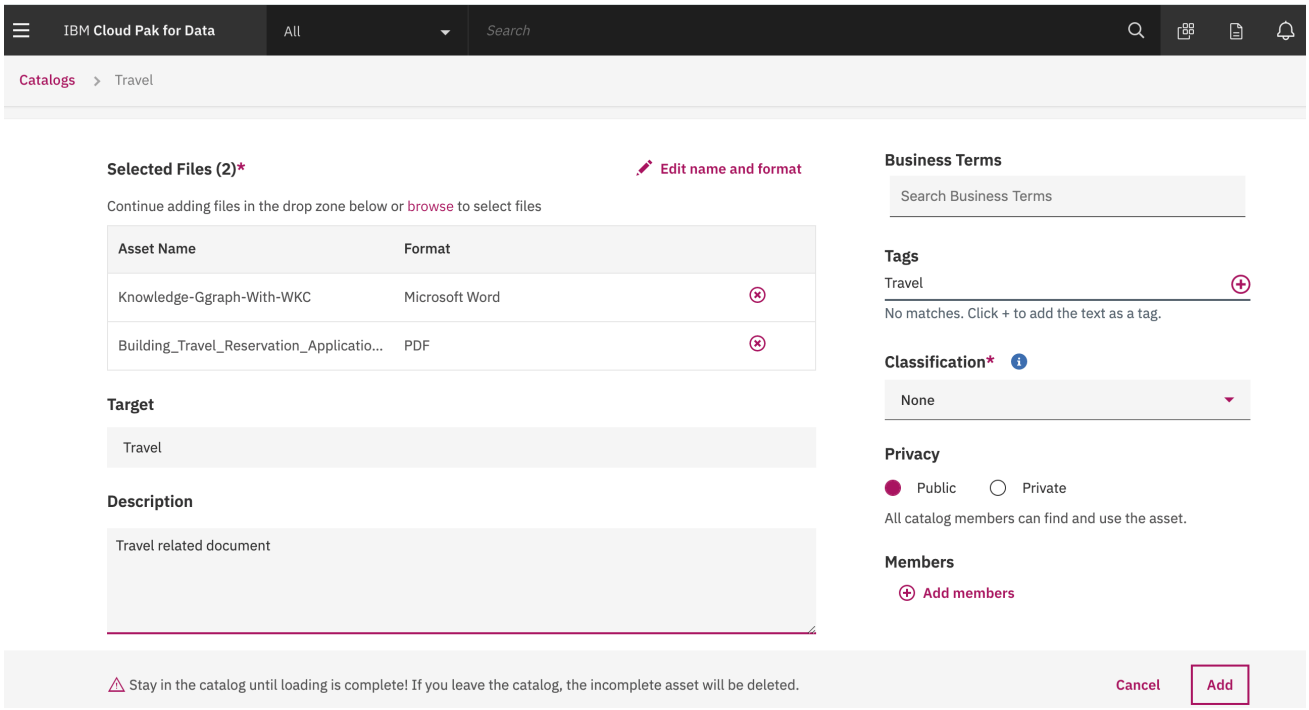


- Click the **pencil icon** next to the **Edit name and format** link.

This allows you to rename the data assets and change their file format. A default file format is inferred for you based on the file extension. In this case, they are PDF and DOCX files, so **PDF** and **Microsoft Word** were auto selected. You **will not** change the format, but you will change their names by removing the file extension.

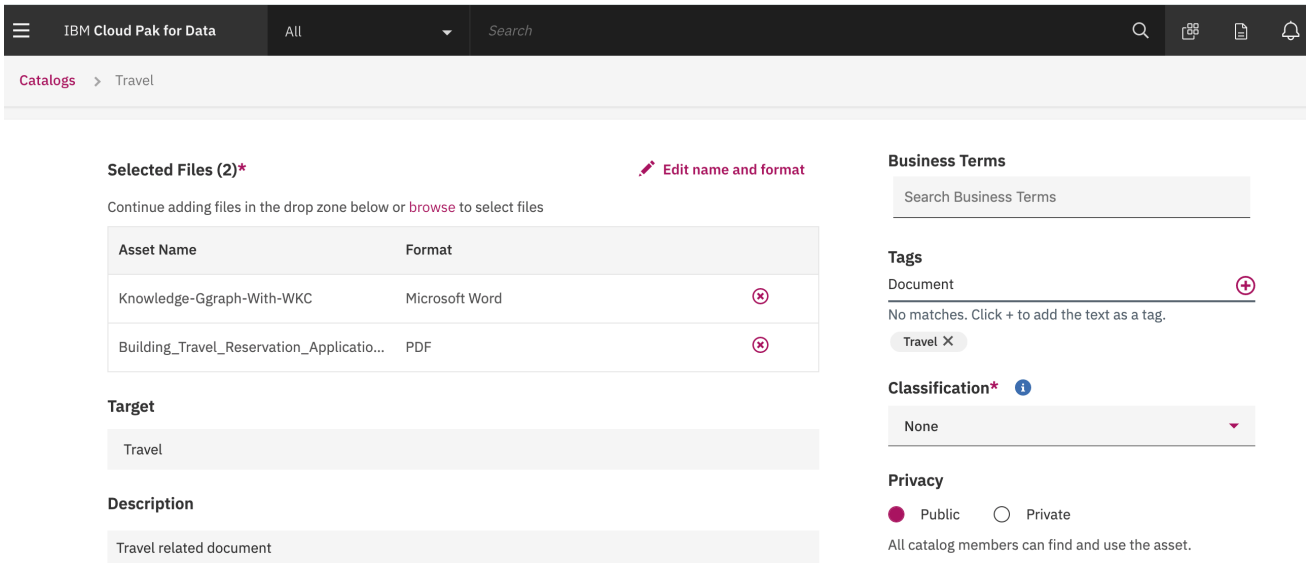


- Click in the **Asset Name** area of the **"Building_Travel_Reservation_Application_CPD.pdf"** file. Go to the end of the filename and remove the **.pdf** extension.
- Click in the **Asset Name** area of the **"Knowledge-Graph-With_WKC.docx"** file. Go to the end of the filename and remove the **.docx** extension.
- Click the **Apply** button to save the filename changes.

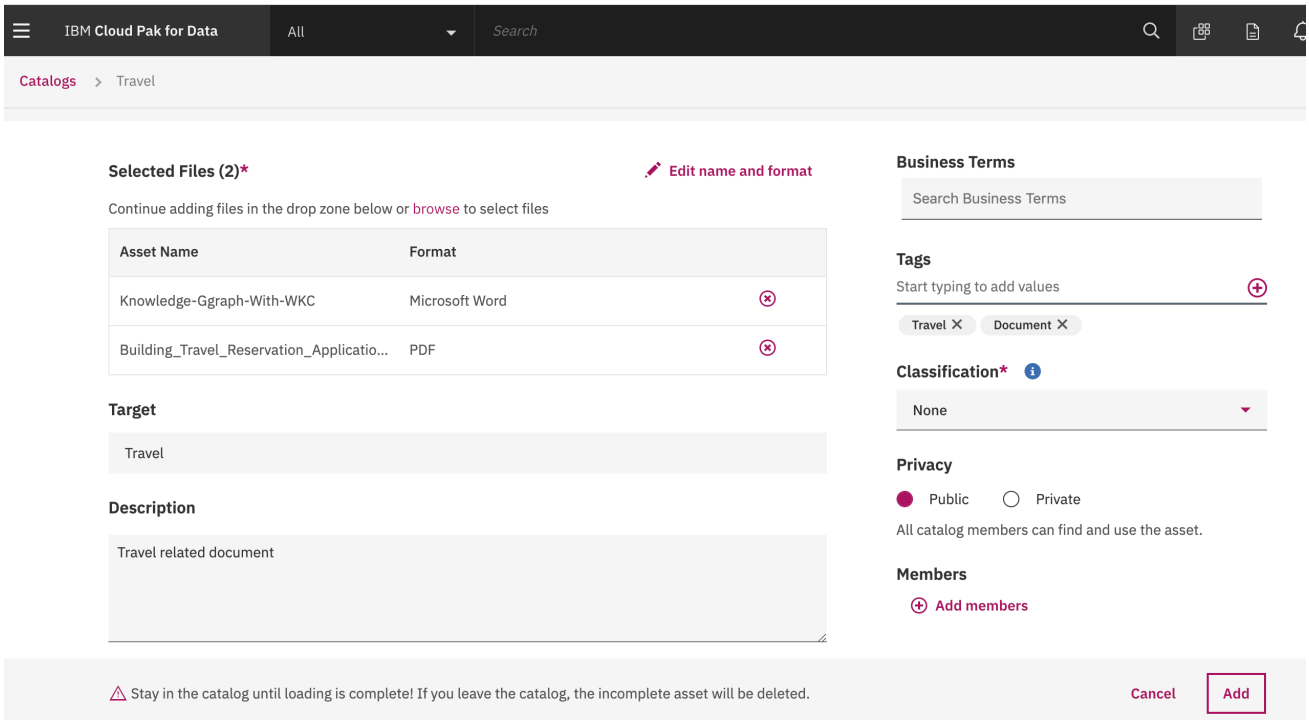


- Enter a Description of **Travel related document**.
- Enter a Tag of **Travel** into the **Tags** area.
- Click the **+** sign next to the tag to add it.

Every time you enter a tag, you need to select the **+** sign to add the tag. The tags will appear as added tags in the tag area below the tag name. Once a tag is added, it can be used and selected for other data assets. Knowledge Catalog displays all available tags once they are added to the catalog. You will see this in action when you add the next file to the catalog.

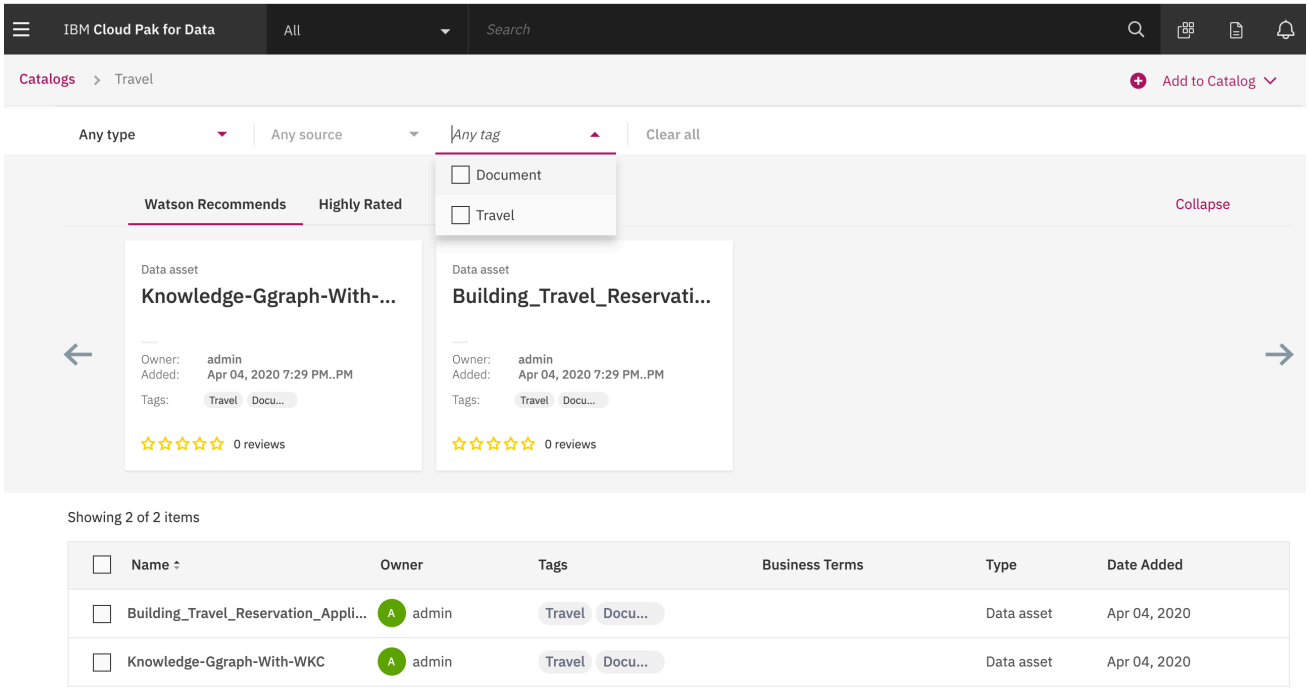


- Enter a Tag of **Document** into the **Tags** area.
- Click the **+** sign next to the tag to add it.



The screenshot shows the **Travel** and **Document** tags that you should have entered for this asset. Make sure you have added them before you proceed to the next step that catalogs them.

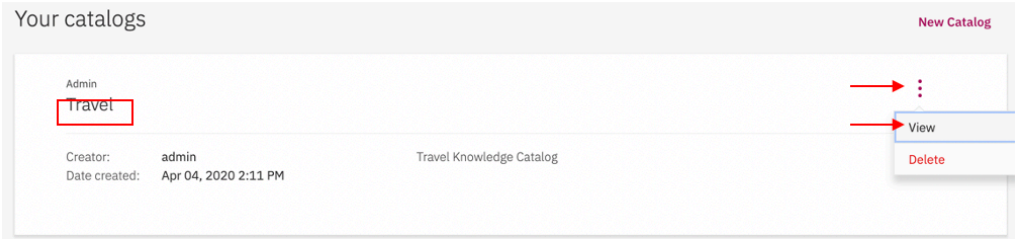
- Click the **Add** button to catalog the unstructured data assets.



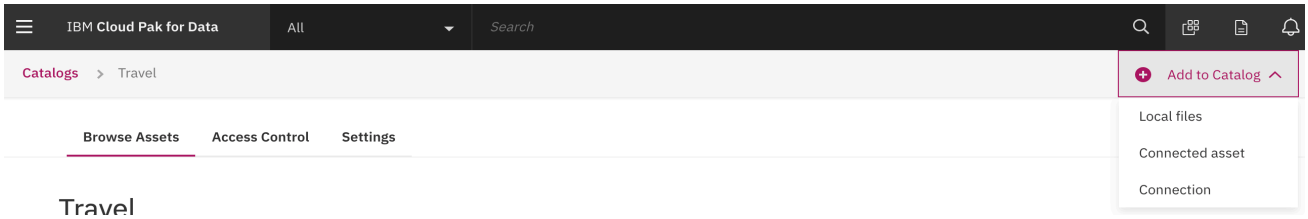
Upon completion, the data assets will automatically be added to the **Browse Assets** section of the catalog asset browser. Scroll down and you will see the two newly added documents in the catalog with the tags you specified. Notice that the **Travel** and **Document** tags have been added to the **Tags** filter area.

8.3 Auto Discover Data Assets

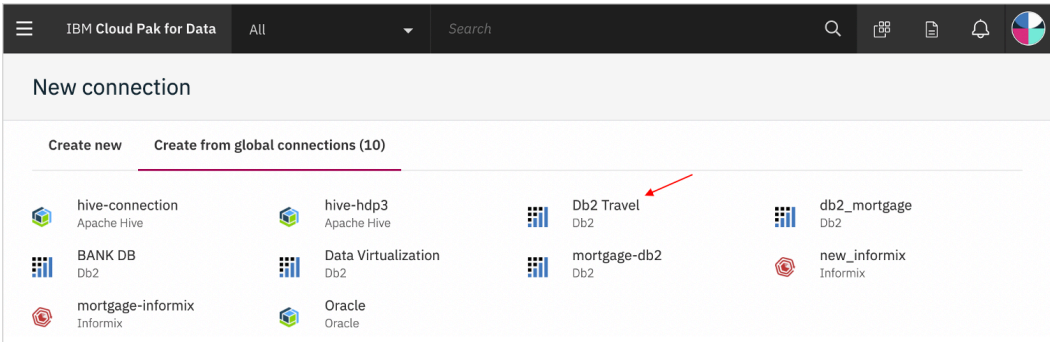
Go to the **Navigation menu** in the top left corner and click on **Travel** catalog from **Organize > All catalogs**.



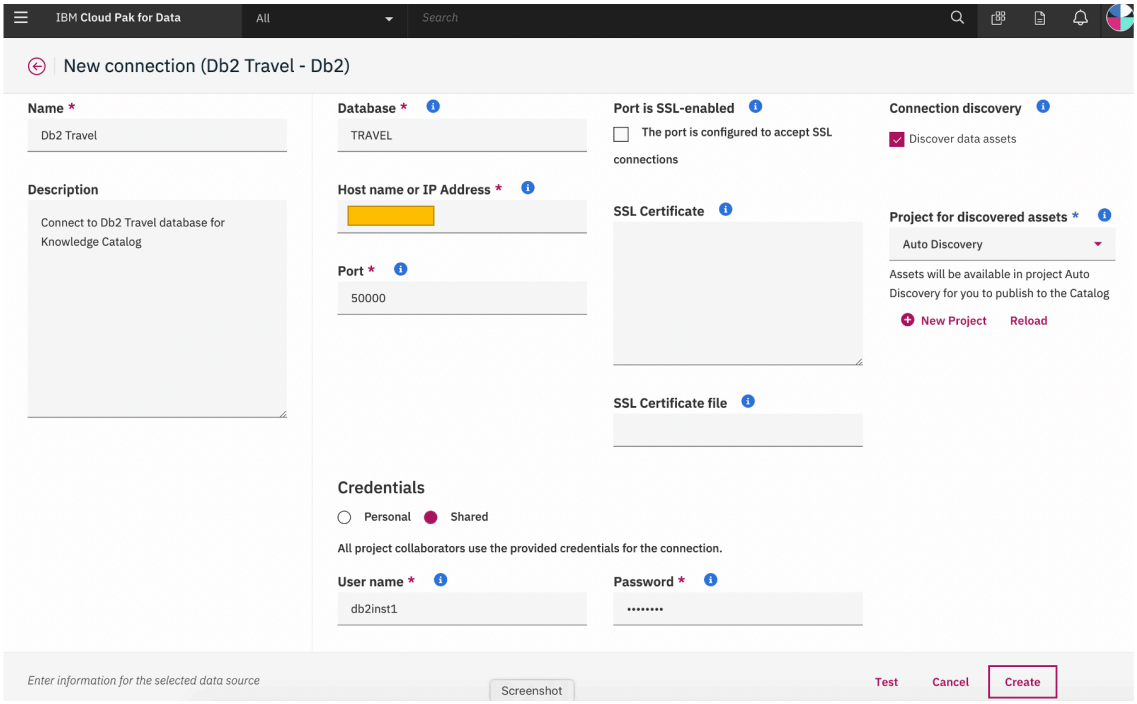
- Click on the **configuration** icon of the **Travel** catalog.
- Select **View**



- Click **Add to Catalog > Connection** from the catalog menu.



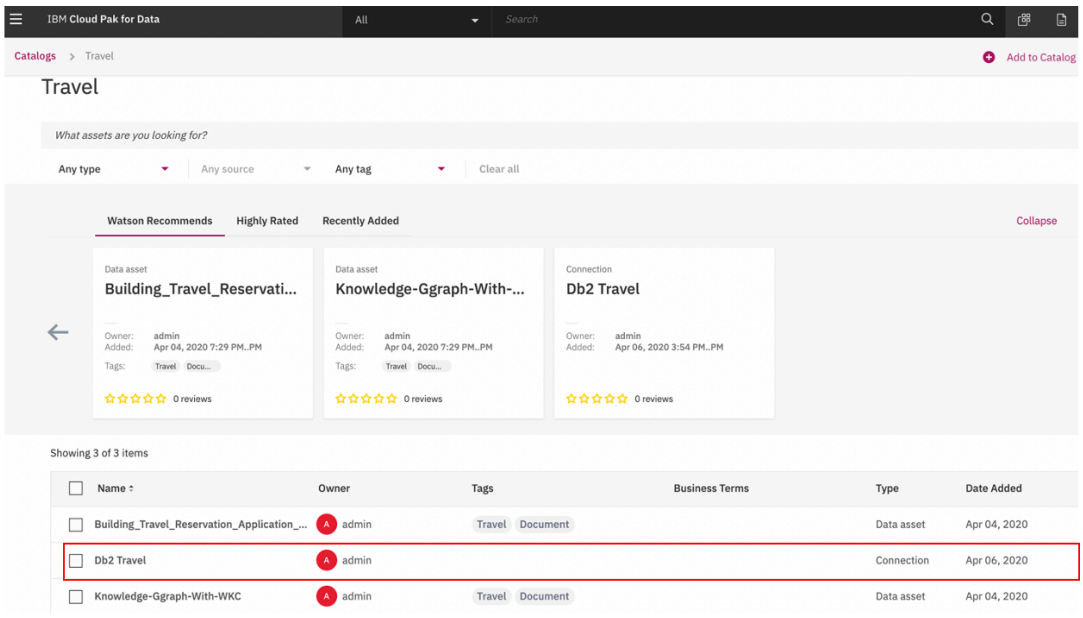
- Choose **Create from global connection** tab
- Click on the **Db2 Travel** connector.



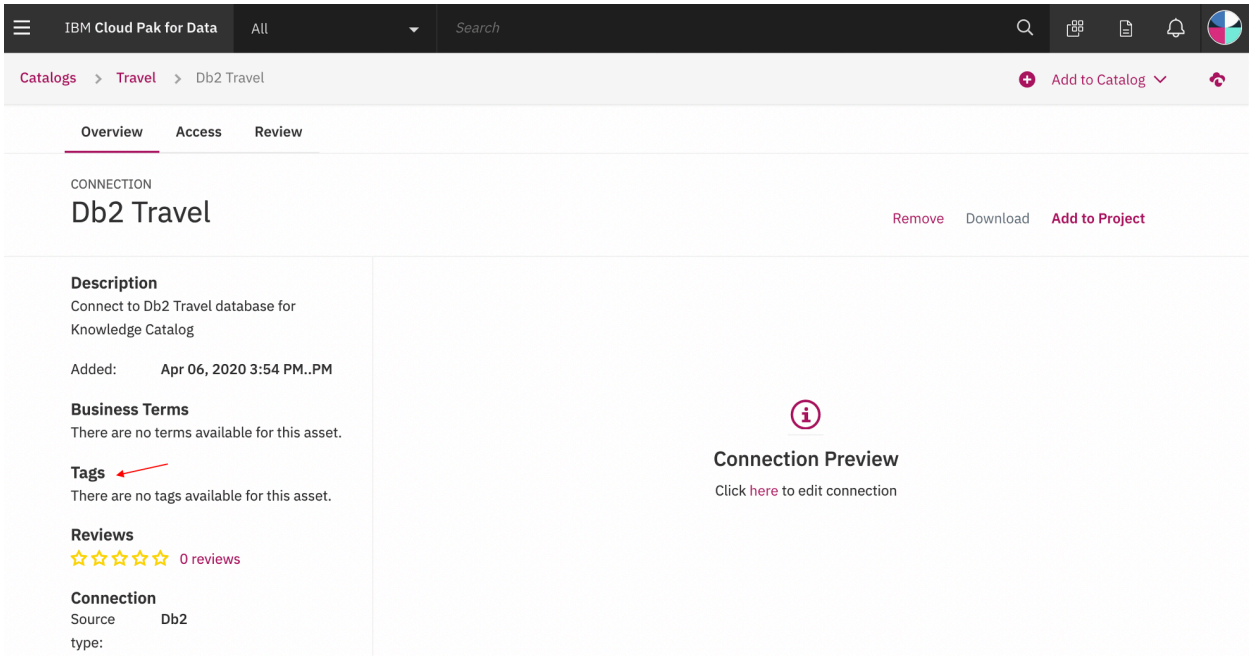
Most of connection information is pre-populated based on the global connection

- Click the **Discover data assets** check box under Connection discovery.
- Select the **Auto Discovery<id>** project from the “Project for discovered assets” selection list.
- Choose **Shared** for Credentials
- Click **Test**
- Click the **Create** button.

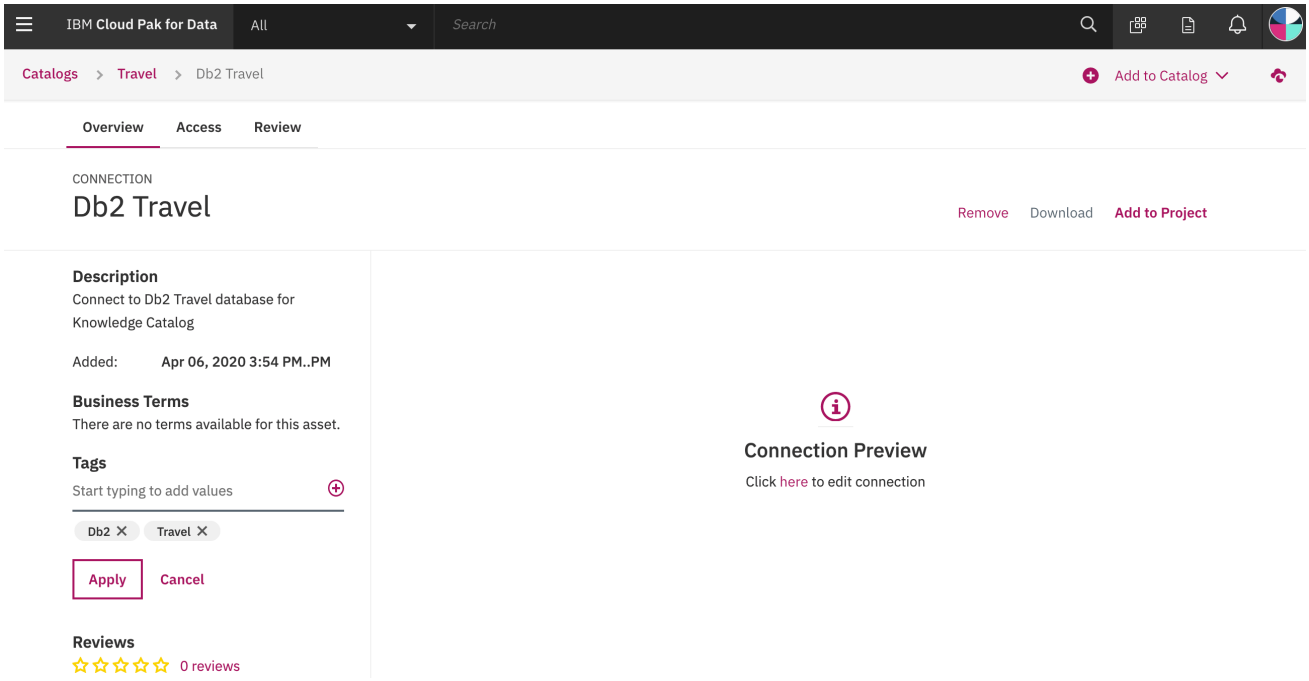
Add a tag to the connection you created



- Click on the connection name **Db2 Travel.**



- Hover next to the **Tags** section and Click the **pencil icon** to add tags to the connection.

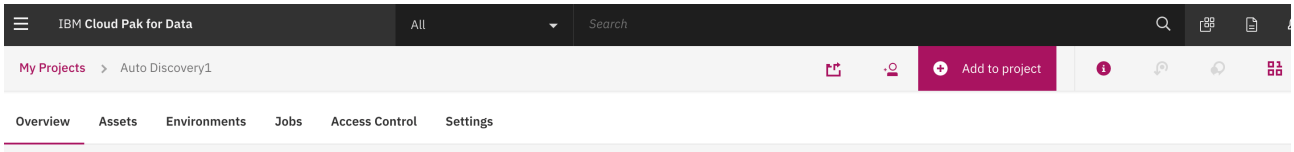


- Click in the **Tags** section and select the **Travel** tag from the list of tags and Click the **+** sign next to the tag to add it.
- Similar way adds another tag **Db2**. But you may need to type the tag name.
- Click the **Apply** button.

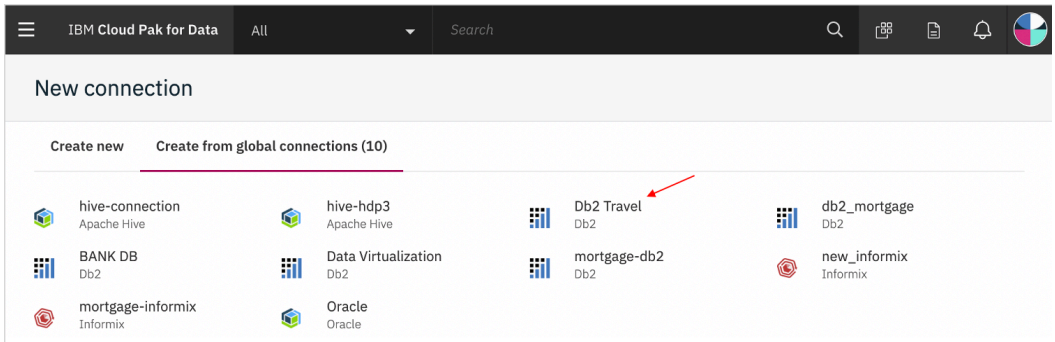
The discovery process has been running as a service in the background, discovering and populating data assets into the **Auto Discovery** project. Let's examine the project to review the discovery results. We are interested in finding relevant travel data, specifically Rentalcar and Hotel data that will be used by the R application development.

8.4 Review Discovery Result

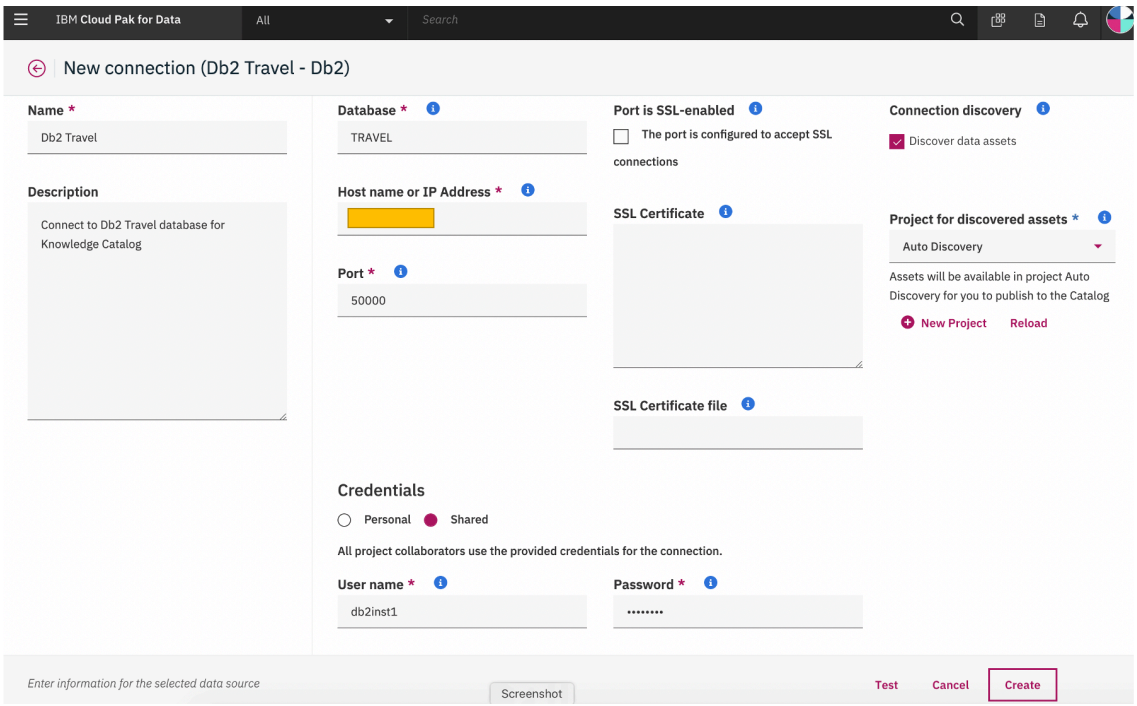
Go to the **Navigation menu** in the top left corner and click on **Projects**. Select **Auto Discovery<id>** from list of projects.



- Click **Add to project** > **Connection**

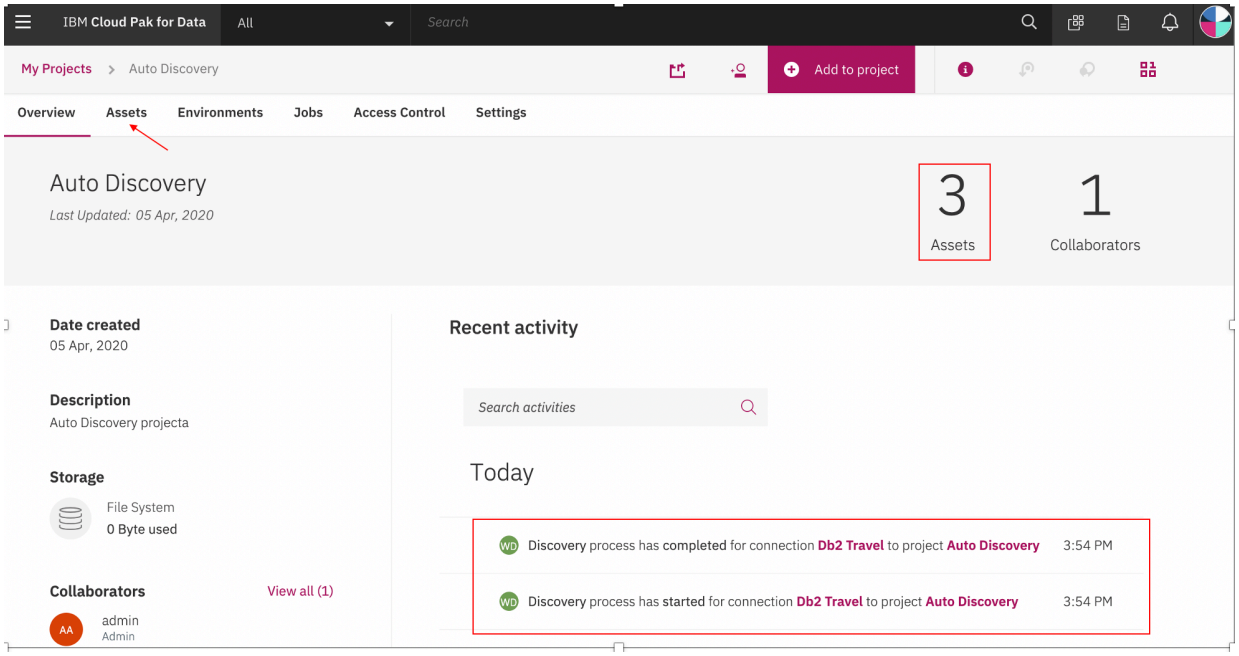


- Choose Create from global connection tab
- Click on the **Db2 Travel** connector.

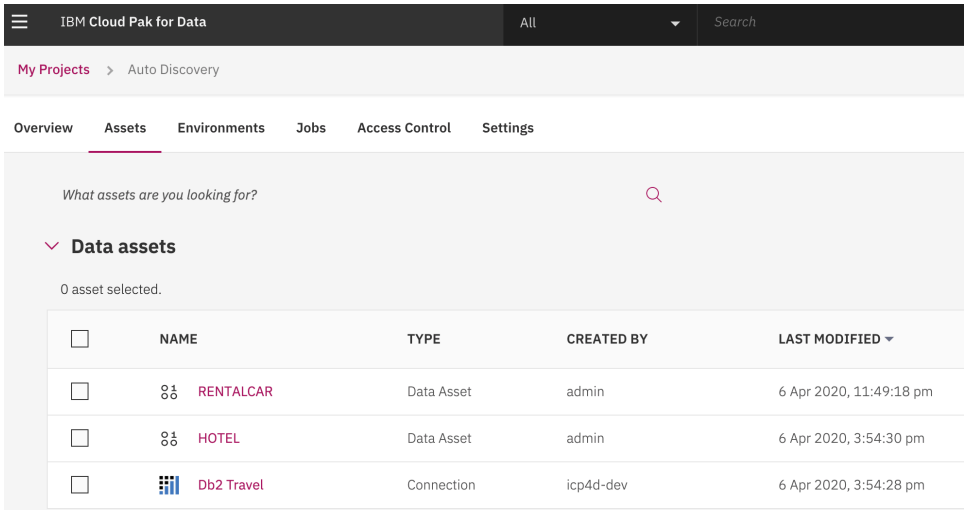


Most of connection information is pre-populated based on the global connection

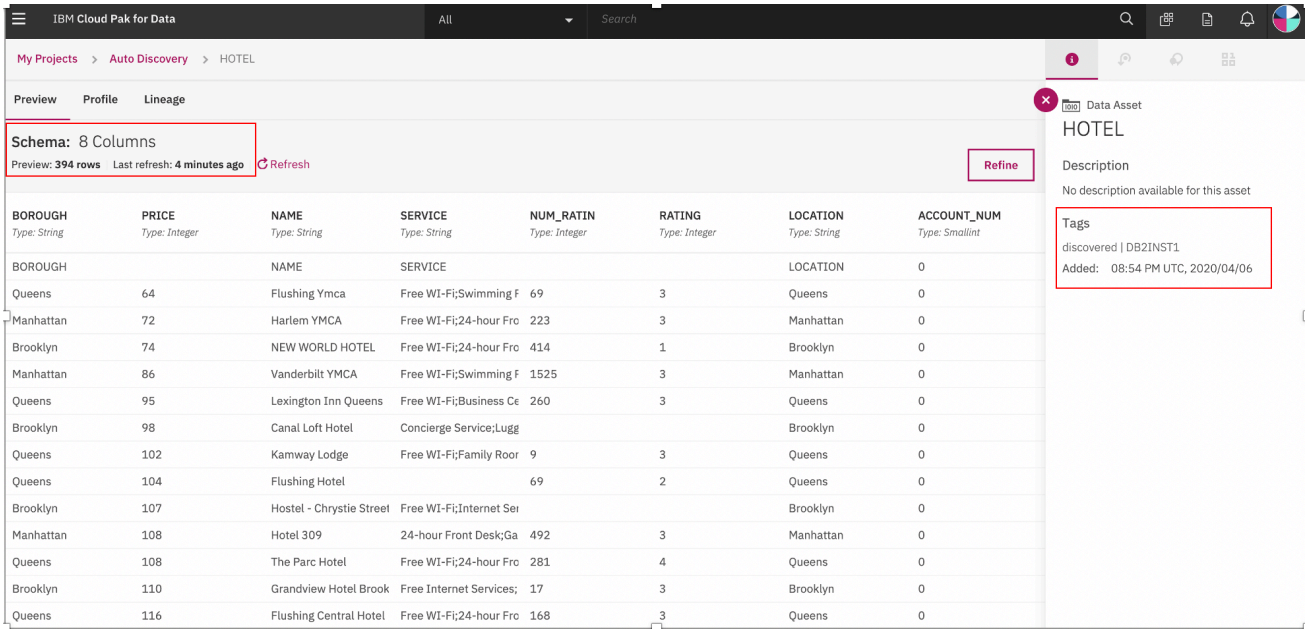
- Click the **Discover data assets** check box under Connection discovery.
- Select the **Auto Discovery<id>** project from the “Project for discovered assets” selection list.
- Choose **Shared** for Credentials
- Click **Test**
- Click the **Create** button.



The auto discovery process discovered 3 data assets and added them to the project. Your discovery may be different from this screenshot because this connection is a shared Db2 database, and data assets are being added and removed all the time by other users. Knowledge Catalog scanned the Db2 database instance and collected the metadata for all the user data assets that user is authorized to access. The **Recent activity** area notifies the status of recent discovery processes.

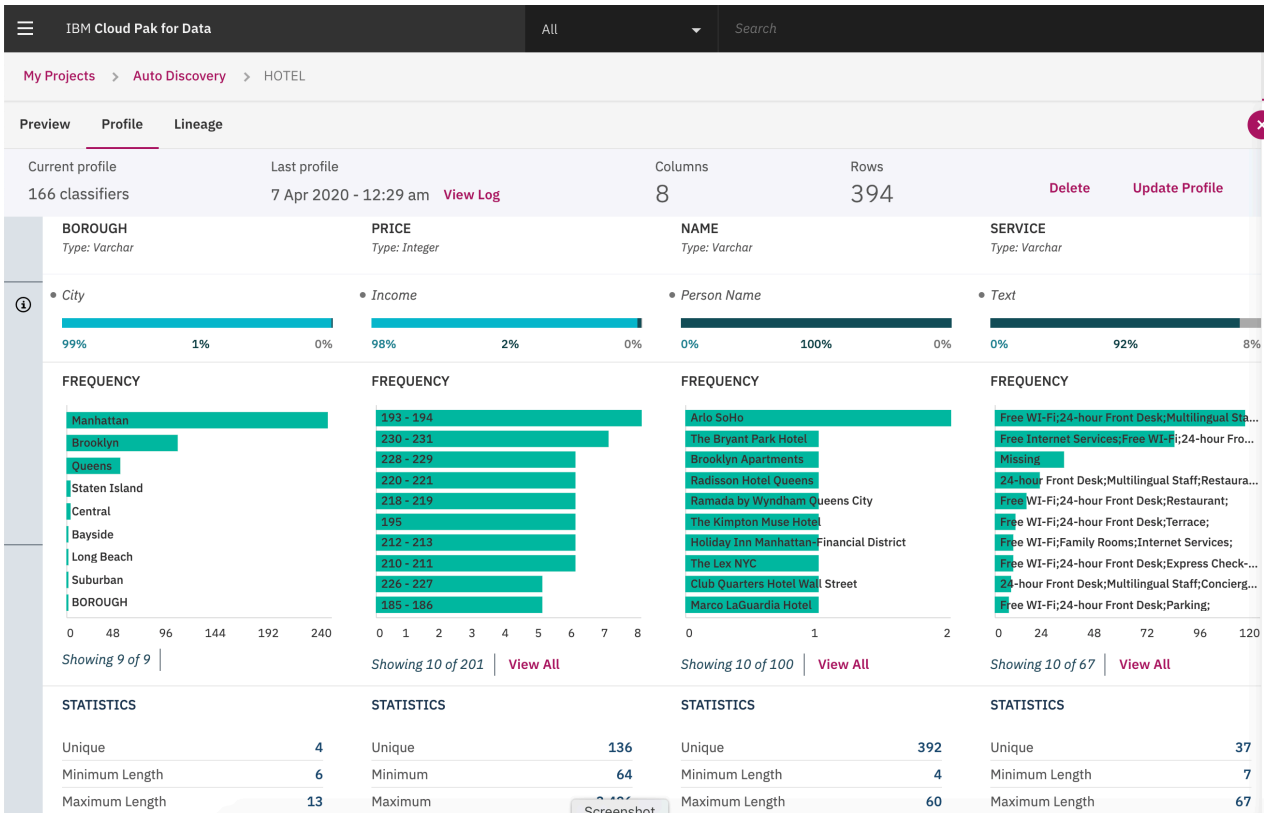


- Click the **Assets** tab to view the discovered assets.
- Click on the **HOTEL** data asset to open the data previewer.



The information panel on the right shows a tag of **DB2INST1**. This is the schema in the Db2 instance that the table came from. The top left section provides information about the total number of columns and rows in the table. In the data section, scroll to the right, to see all columns.

- Click on the **Profile** tab to get frequency and summary statistics for each of your columns. It's a quick way to get an understanding of the data by look at the metrics of data distribution.
- If you are running profile first time on the table, click on **Create Profile**. It will take few minute depends on the volume of data in the table.
- The **Refine** option on left (under **Preview** tab), let you use the Data Refinery tool against the data assets.

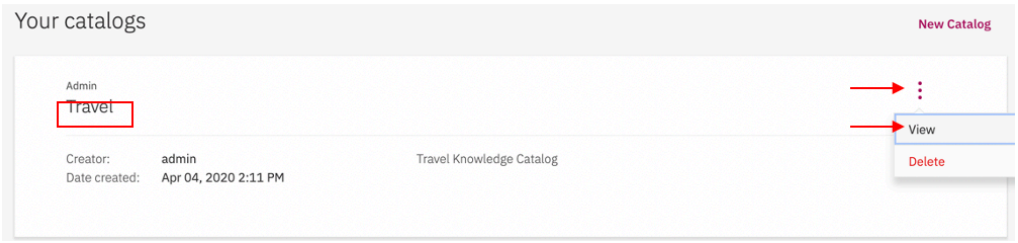


The profile gives an insight of data. It will direct you through the data cleaning requirement by provide frequency and statistics details of each column. Profile can help you to determine if the data refinery flow going in the right direction. For example: BOROUGH column shows most of hotels listed in Manhattan. The NAME column gives an idea of number of unique hotels are there.

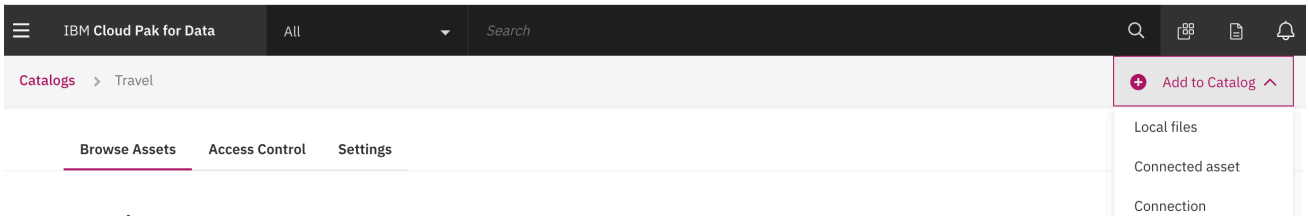
8.5 Add Structured Data to Catalog

You have cataloged unstructured data files from the local file system. Now auto discovered data assets from a Db2 database connection. You will now catalog two tables from the Db2 database connection; **Rentalcar** and **Hotel** using the **Connected asset** catalog method. These tables are needed for build an application on CPD.

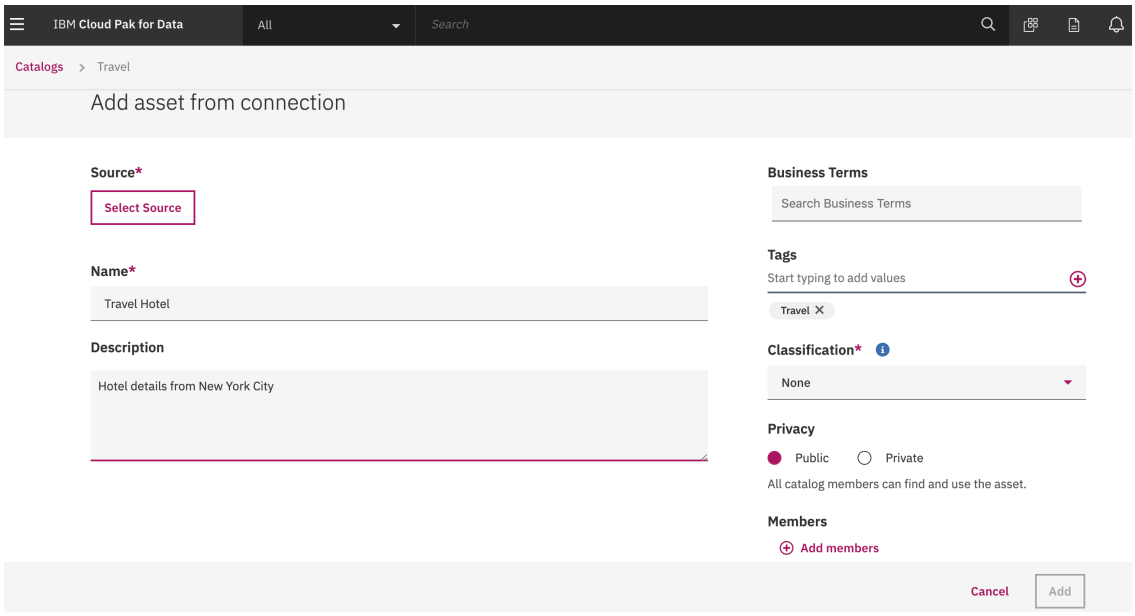
Go to the **Navigation menu** in the top left corner and click on **Travel<id>** catalog from **Organize > All catalogs**.



- Click on the **configuration** icon of the **Travel<id>** catalog.
- Select **View**

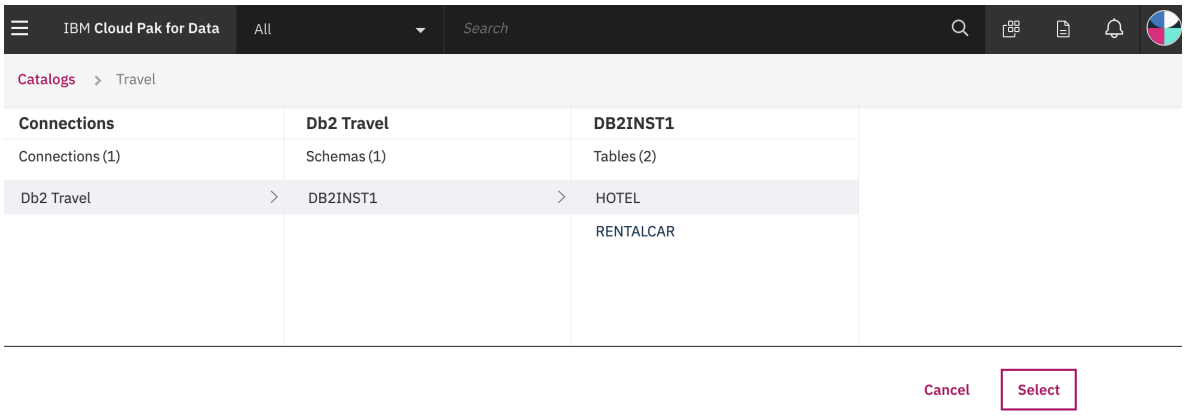


- Click **Add to Catalog > Connected asset** from the Catalog menu.



- Enter a Name of **Travel Hotel<id>**.
- Enter a Description of **Hotel details from New York City**.

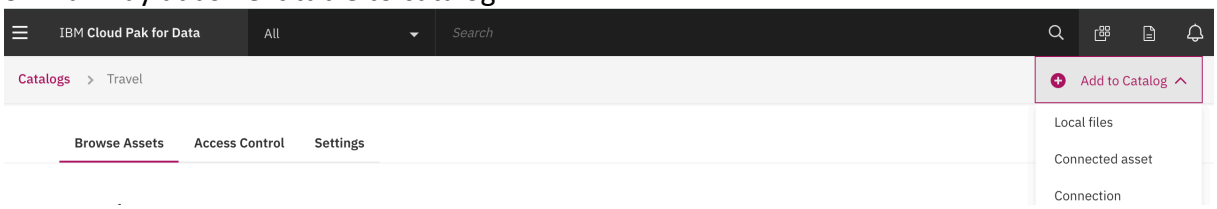
- Click in the **Tags** area and select the **Travel** tag from the drop-down list.
- Click the **Select Source** button to choose a Connection to add connected assets from.



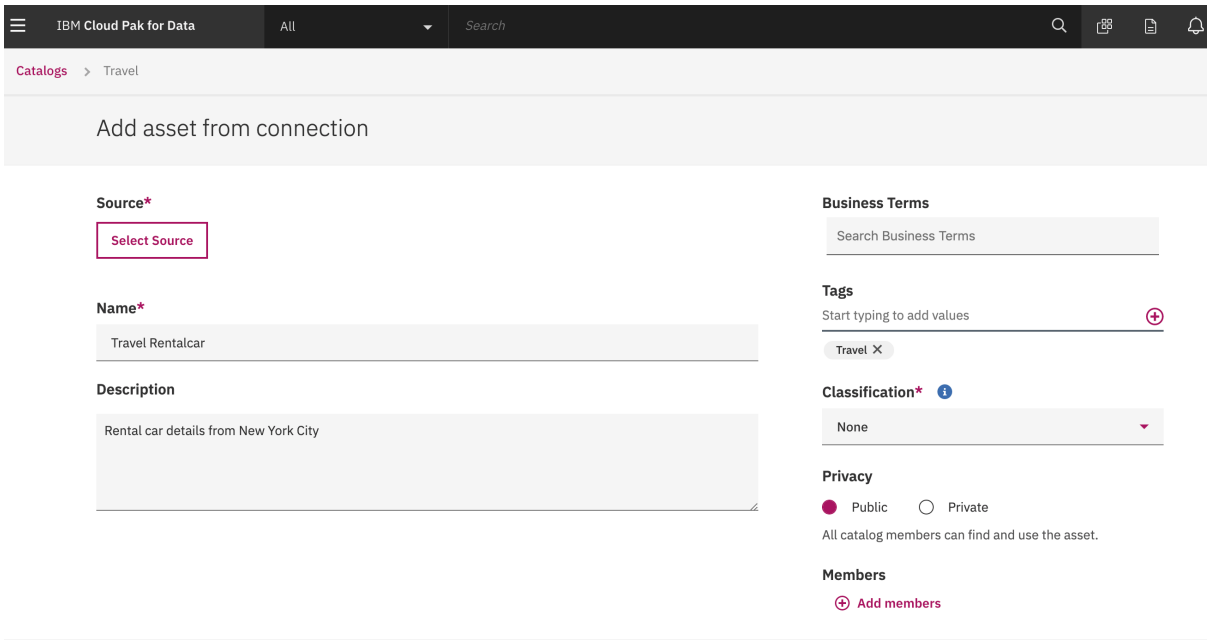
- Click on the **Db2 Travel** connection from the list of connections.
- Click on the **DB2INST1** schema from the list of schemas.
- Click on the **HOTEL** table from the list of tables.
- Click the **Select** button.

On the following window click the **Add** button.
You should see the cataloged table as a data asset with the tag you supplied.

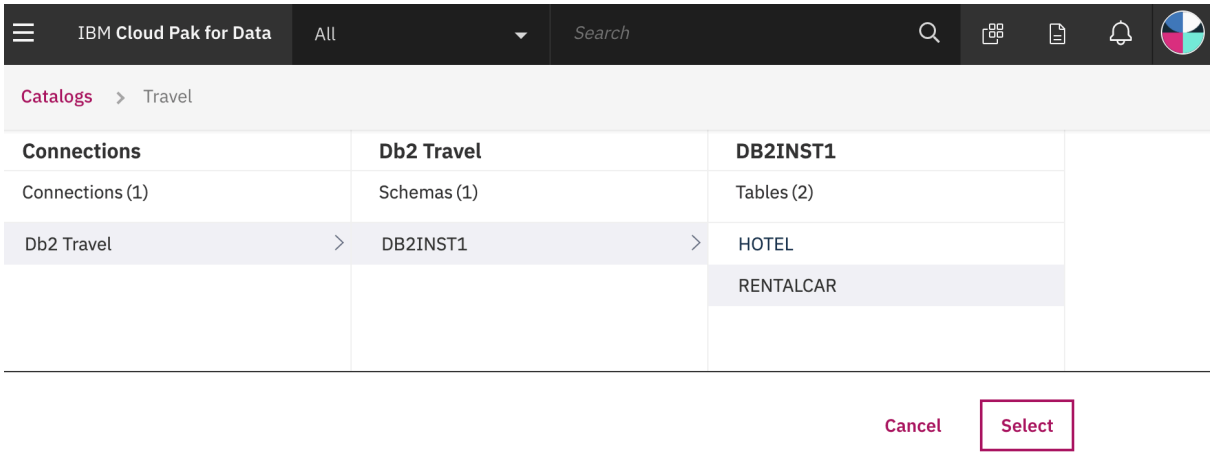
Similar way adds next table to catalog.



- Click **Add to Catalog > Connected asset** from the Catalog menu.



- Enter a Name of **Travel Rentalcar<id>**.
- Enter a Description of **Rental car details from New York City**.
- Click in the **Tags** area and select the **Travel** tag from the drop-down list.
- Click the **Select Source** button to choose a Connection to add connected assets from.



- Click on the **Db2 Travel** connection from the list of connections.
- Click on the **DB2INST1** schema from the list of schemas.
- Click on the **RENTALCAR** table from the list of tables.
- Click the **Select** button.

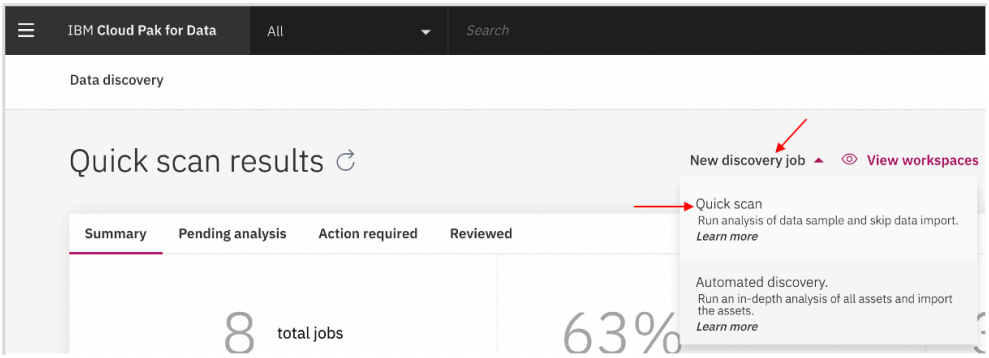
On the following window click the **Add** button.
You should see the cataloged table as a data asset with the tag you supplied.

9 Discover Assets using Quick Scan

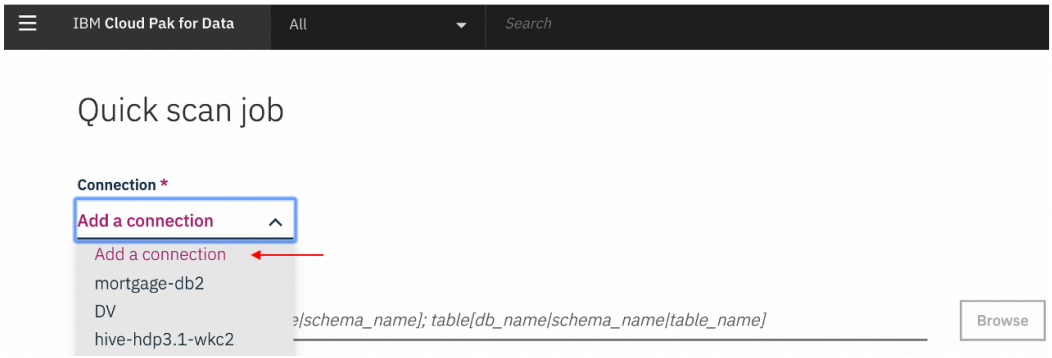
A quick scan can come to great help when you don't know your data very well, and you want to analyze large data sets to see a general overview of the quality of the data. Quick scan performs analyze columns, analyze data quality and automatically assign term.

In this task you will use the **Db2 Travel** global connection to **Travel** database on the Db2.

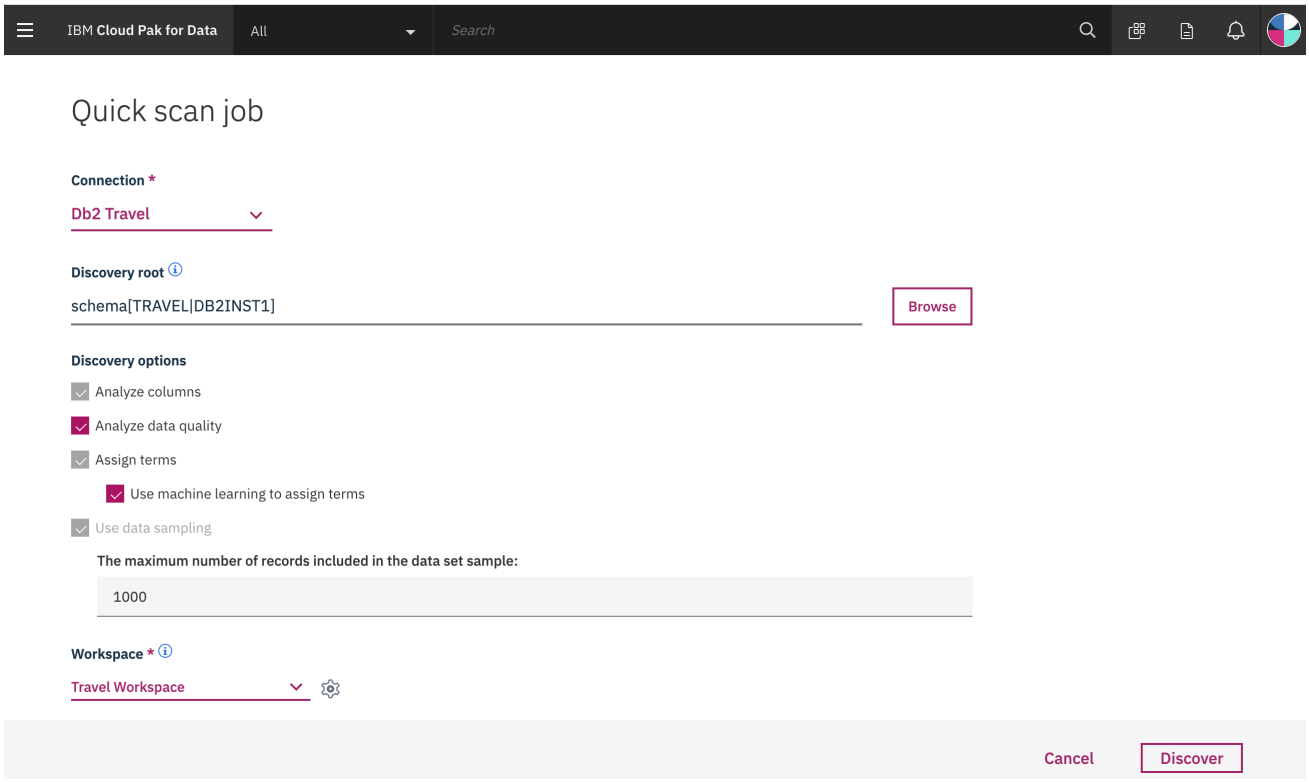
Go to the **Navigation menu** in the top left corner and click on **Organize > Curation > Data discovery**. Select **Auto Discovery** from list of projects.



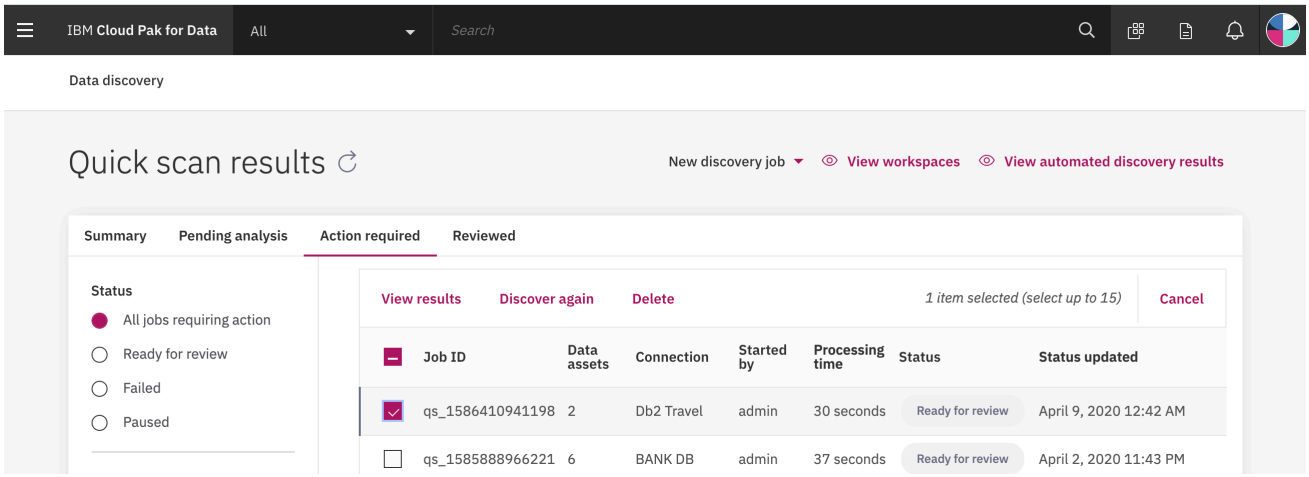
- Click **New discover job** > **Quick scan**



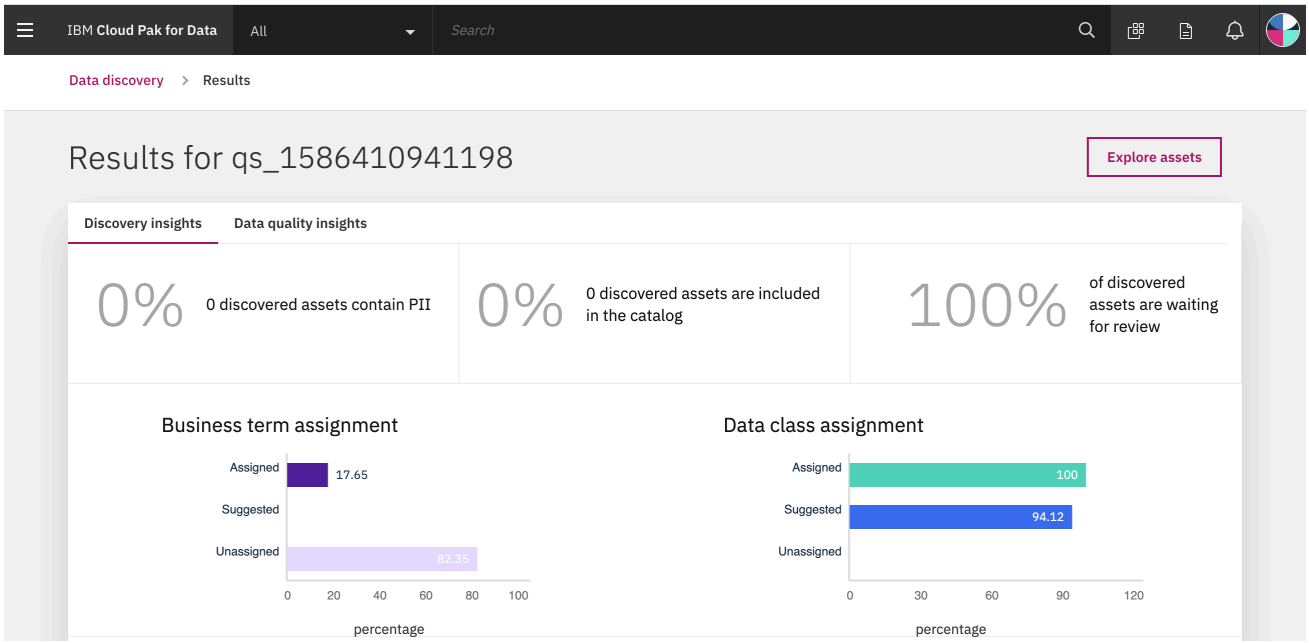
- Click on **Add a connection.**
- Choose connection name **Db2 Travel**.
- Chick **Next** button.



- Click on **Add a connection**
- Choose the connection named **Db2 Travel** that you created previously, click Next
- Select Discover root as **TRAVEL** > **DB2INST1**
- Check necessary Discover options
- Click on Add a workspace under Workspace and named it as **Travel Workspace**. Click Create.
- Click on **Discover**

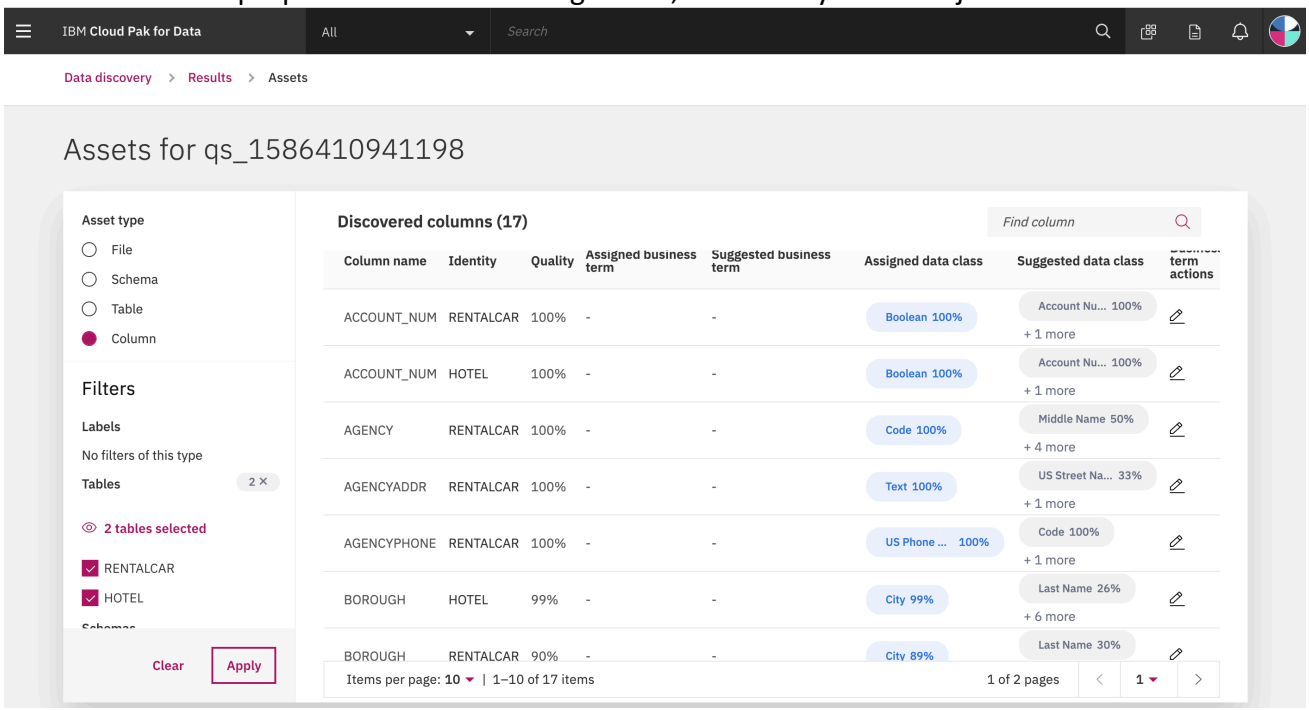


- Click on **Quick scan results**
- Choose **Action required** tab
- Check the quick scan result associated with **Db2 Travel** connection
- Click **View results** tab to explore the discover assets.

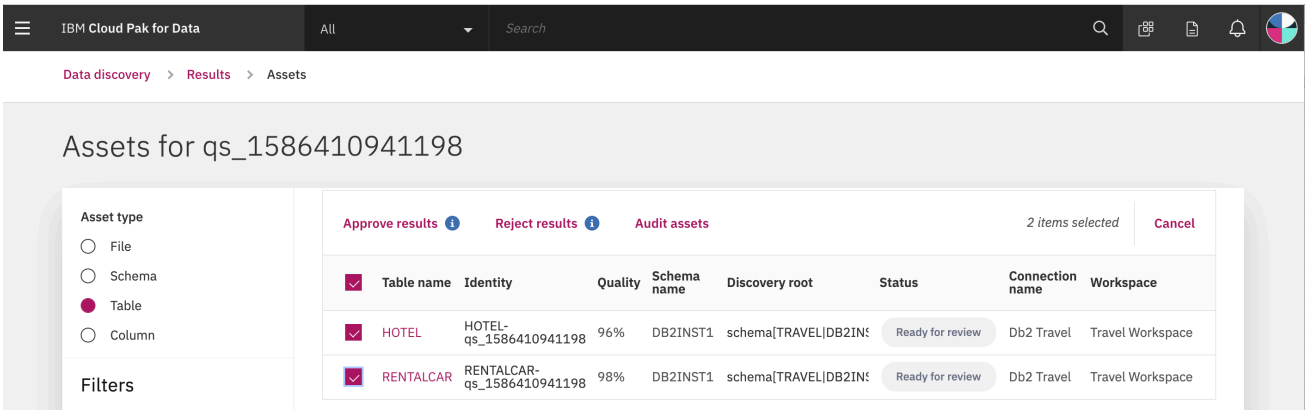


- Review the discovery results using **Explore assets** tab

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



- Select Asset type as **Column**
- Filters necessary tables **RENTALCAR**, **HOTEL** using checkbox
- Click on **Apply**



- Change Asset type as **Table**.
- Select all **Travel** related tables.
- Click on **Approve results**.
- Click on **Approve assets**.

10 Implement Business Glossary

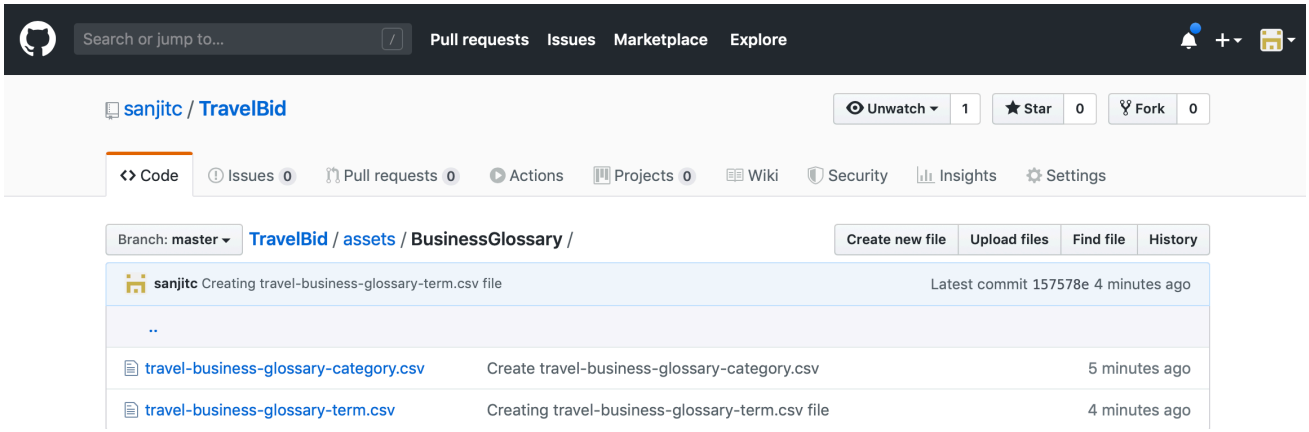
Cloud Pak for Data enables you to structure your enterprise information in a logical way, discover relationships between assets, and keep your data always up-to-date. You can import existing glossary with categories, terms, information governance policies and rules.

In this task, you will import some pre-generated business glossaries, apply rules and polices to govern data.

10.1 Download Business Glossaries

First download pre-generated business glossaries from the GIT to your local machine.

- Go to https://github.com/sanjitc/TravelBid/tree/master/assets/BusinessGlossary



- Download two CSV files and save locally.

Notes:

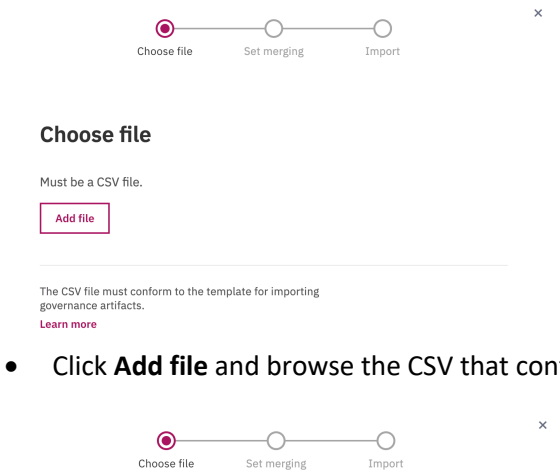
- Go to that particular business glossaries that you want to **download** and click on it.
- You will see "Raw" button on the top right side of the dataset.
    - Press "Alt" and then left click the "Raw" button (on Windows) or
    - Click with two fingers (on Mac)
- Download link file

10.2 Import categories

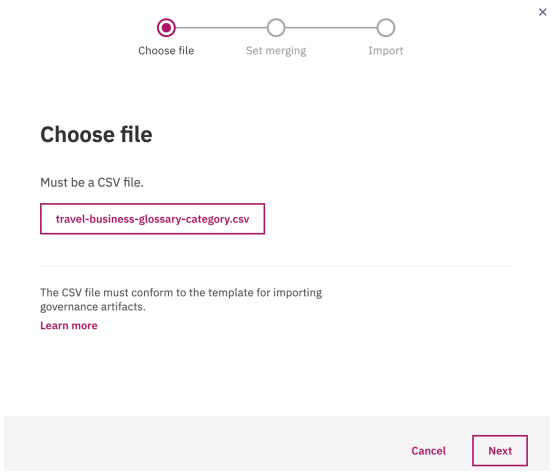
Categories provide the logical structure for governance artifacts so that you can browse and understand the relationships among artifacts. Categories can be organized in a hierarchy based on their meaning and relationships to one another. You can modify governance artifacts outside of the catalog and import them from a file that is in CSV format. Sequence is important when importing business glossaries. Make sure import categories before do the terms.

Go to the **Navigation menu** in the top left corner and click on **Organize > Data and AI governance > Categories**.

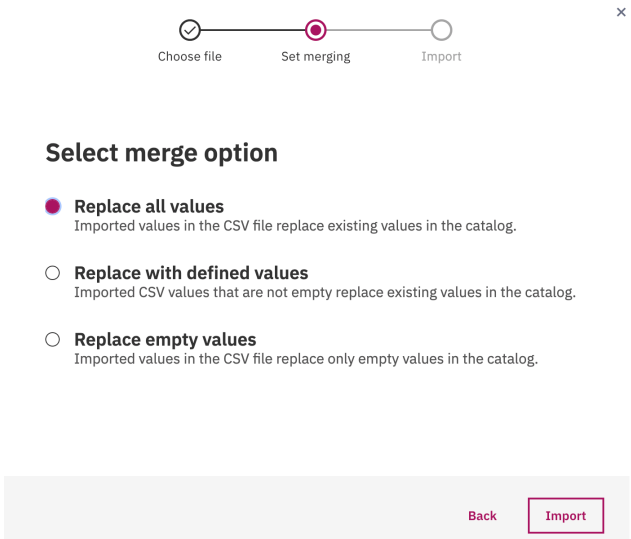
- Click **Import** to import the CSV file contains category information that you downloaded from Git.



- Click **Add file** and browse the CSV that contains category information.



- Choose the **travel-business-glossary-category.csv** file location
- Click **Next**



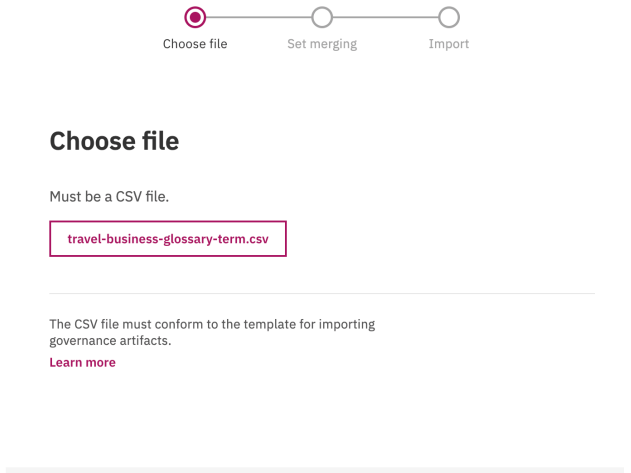
- Select merge option as **Replace all values**.
- Click **Import**.

10.3 Import Terms

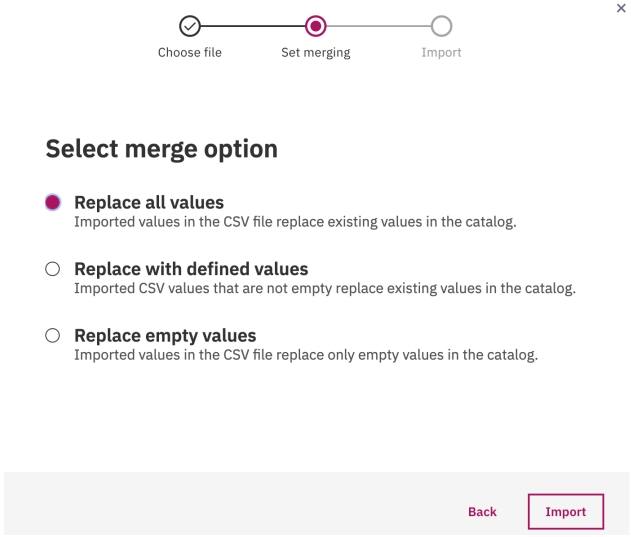
You create business terms to standardize definitions of business concepts so that your data is described in a uniform and easily understood. You can assign one or more business terms to individual columns in relational data sets, to other governance artifacts, or to data assets.

Go to the **Navigation menu** in the top left corner and click on **Organize > Data and AI governance > Business terms**.

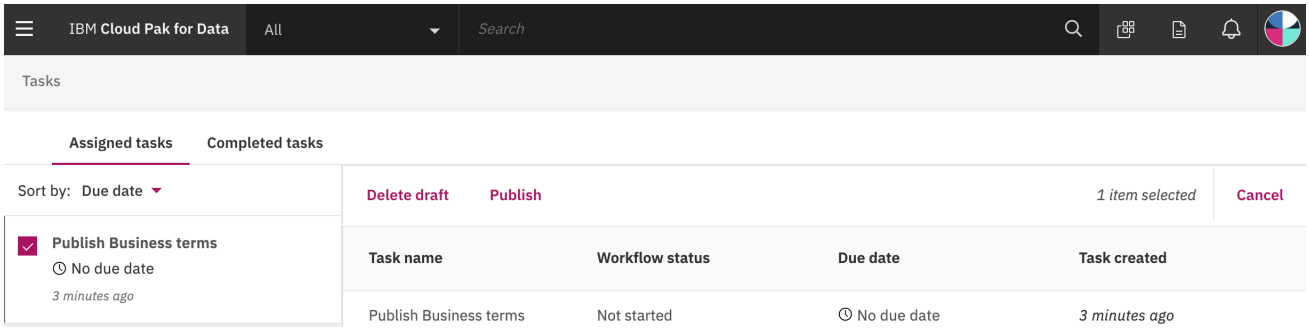
- Click **Import** to import the CSV file contains terms information that you downloaded from Git.
- Next click **Add file** and browse the CSV that contains term information.



- Choose the **travel-business-glossary-term.csv** file location
- Click **Next**



- Select merge option as **Replace all values**.
- Click **Import**.
- On import term completion, click **Go to task**.



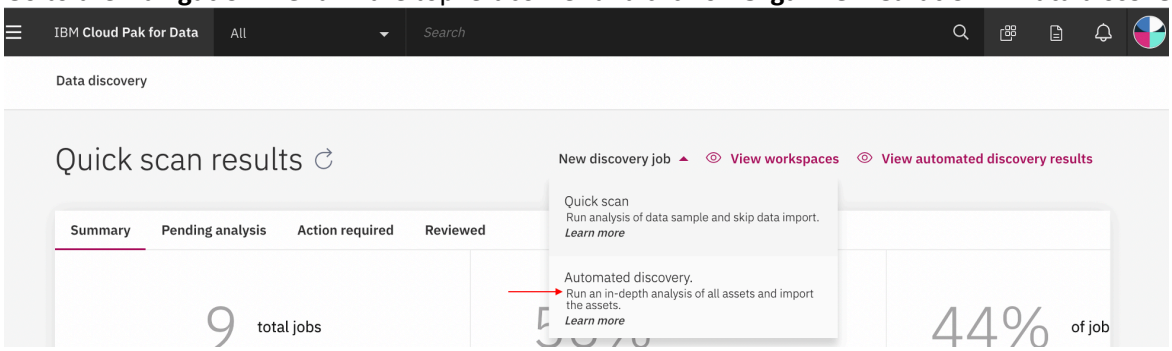
- Select the check box for **Publish Business terms** that you just imported
- Click **Publish** tab

10.4 Automated Discovery

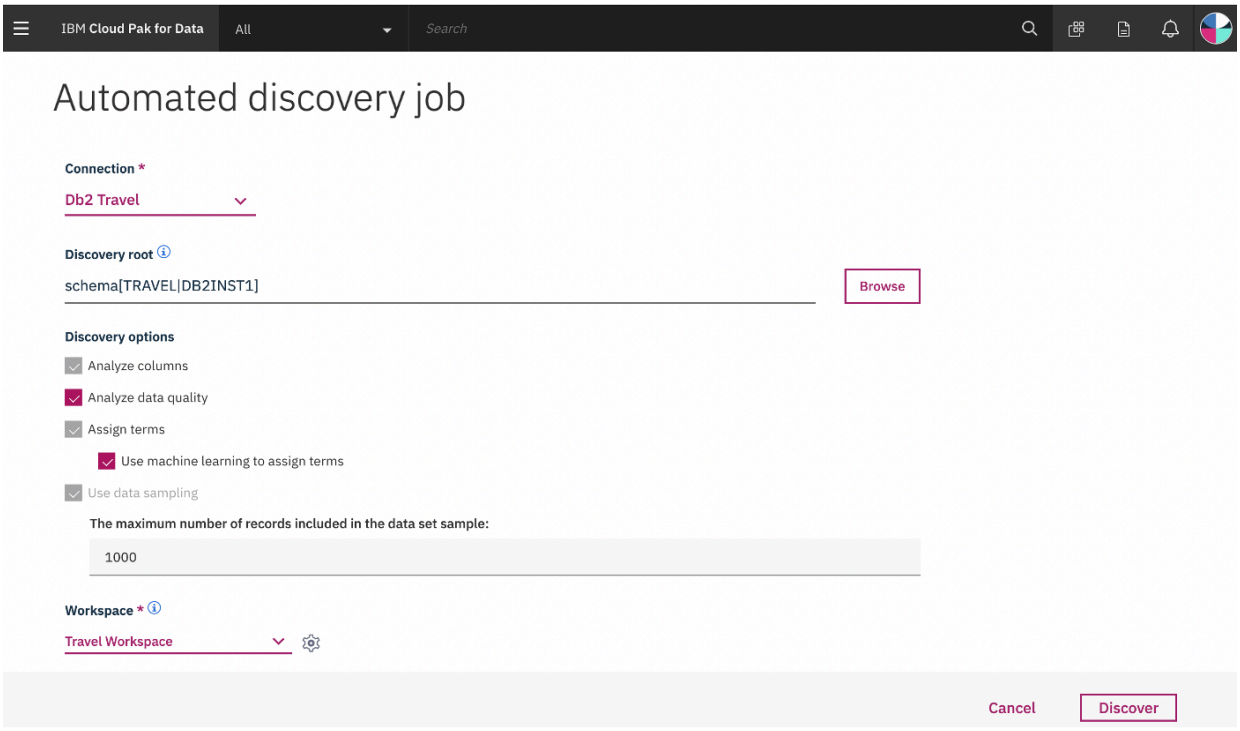
Discover assets to add data to the default catalog. During the discovery the data is imported, analyzed, and classified.

Earlier you ran quick scan, but in this task you will re-run discover assets, so data can imported, analyzed, and classified according the glossary you imported/created earlier.

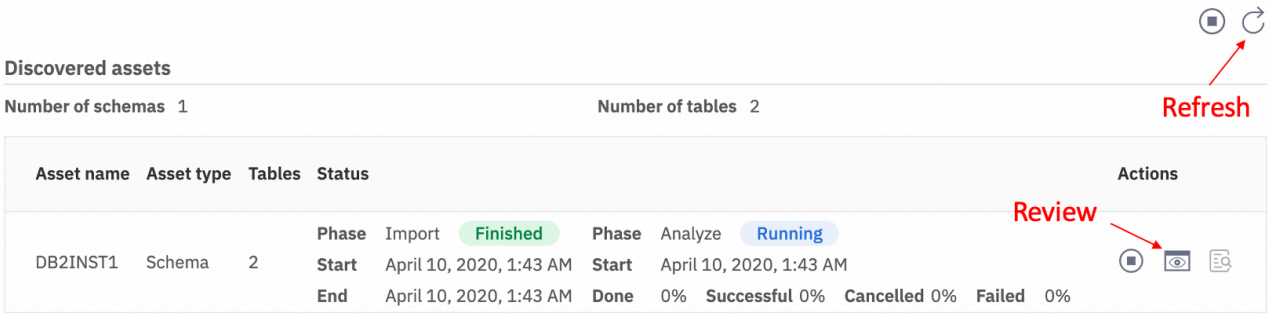
Go to the **Navigation menu** in the top left corner and click on **Organize > Curation > Data discovery**.



- Navigate to **New discover job** > **Automated discovery**



- Choose the connection named **Db2 Travel** that you created previously, click Next
- Select Discover root as **TRAVEL** > **DB2INST1**
- Check necessary Discover options
- Click on Add a workspace under Workspace and named it as **Travel Workspace**. Click Create.
- Click on **Discover**



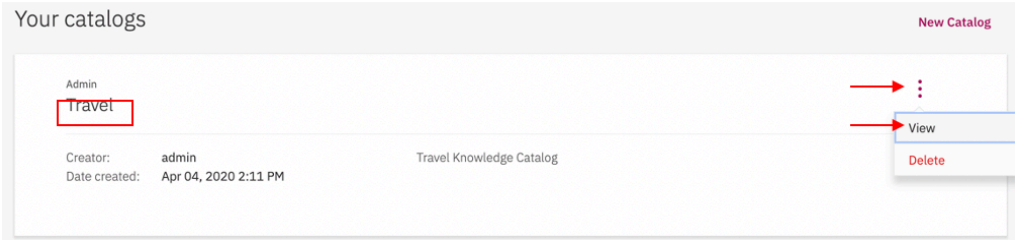
- Discovery job will take some time to complete. Use the **refresh icon** to check the status of the job.
- User **review icon** to discovers how terms automatically assigned to data sets.

11 Data Assets Sentiments

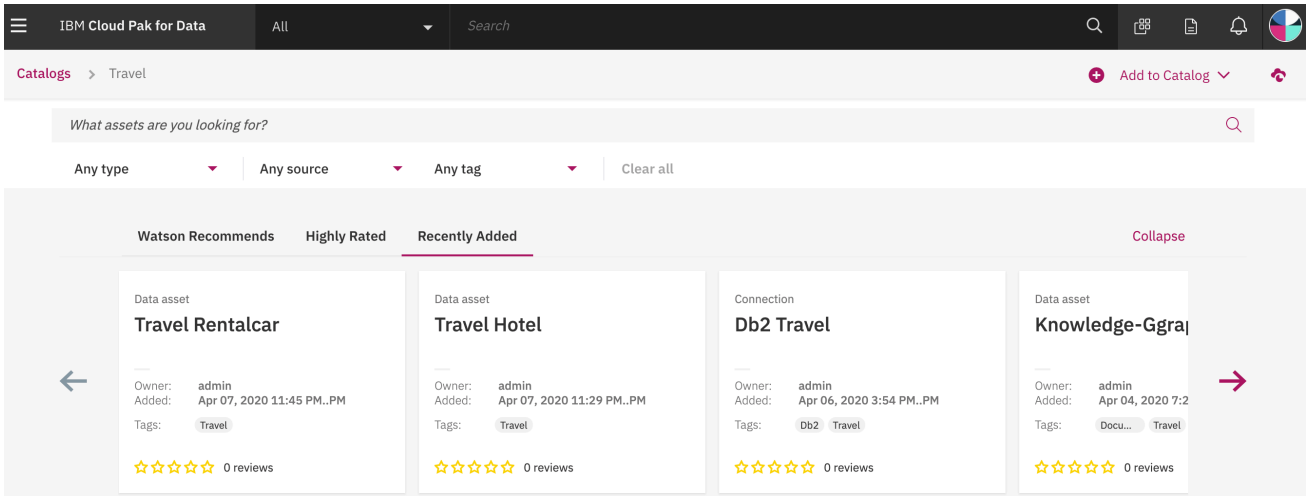
As data assets are cataloged, they are automatically profiled and classified so data consumers can have a better understanding of their content. They can then be enriched using Knowledge Catalog's social capabilities.

In this task, you will visit the **Profile** section of a structured and unstructured data asset, to examine the different profiling and classification features provided. You will also visit the **Review** section to experience how you can rate and review assets to allow others to easily identify and evaluate them based on their ranking and comments.

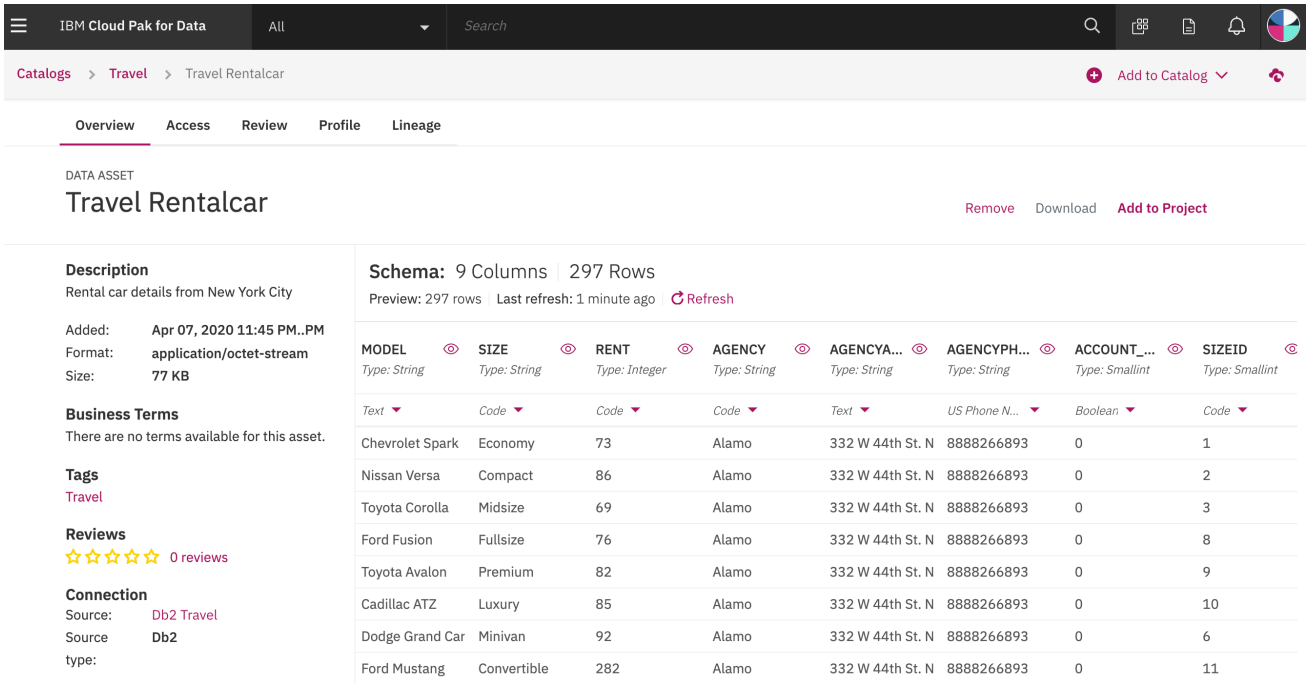
Go to the **Navigation menu** in the top left corner and click on **Travel<id>** catalog from **Organize > All catalogs**.



- Click on the **configuration** icon of the **Travel<id>** catalog.
- Select **View**



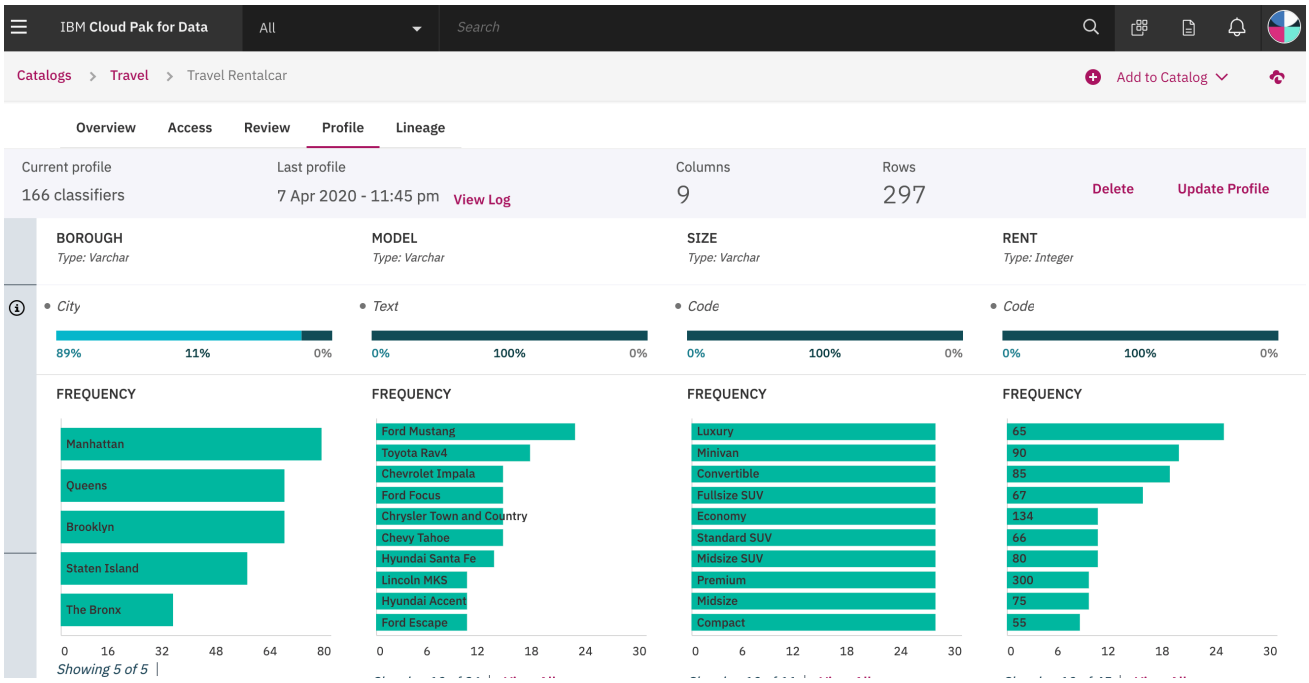
- Click on the **Recently Added** section of the data asset browser.
- Click on the **Travel Rentalcar<id>** asset to view its properties.



The **Overview** section of the asset where you can view sample rows and metadata about the asset, including column level classifications, if it's a data asset. You can modify its name and description, add tags and assign business terms and classifications at the asset or column level.

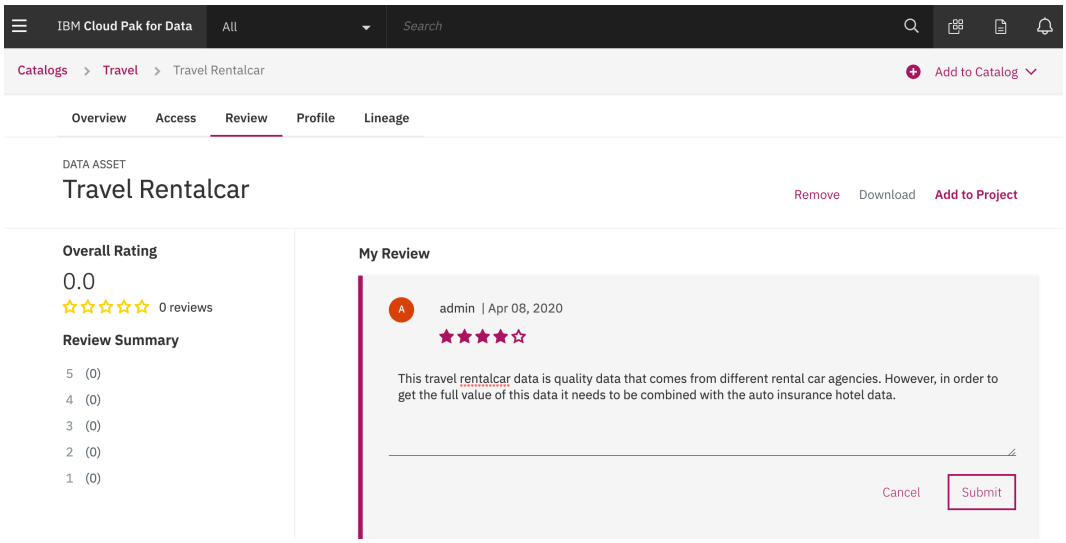
- Click on the **Profile** section of the data asset.

The profile should automatically appear. If not, and you are presented with a method to create or update the profile, follow the instructions to do so.

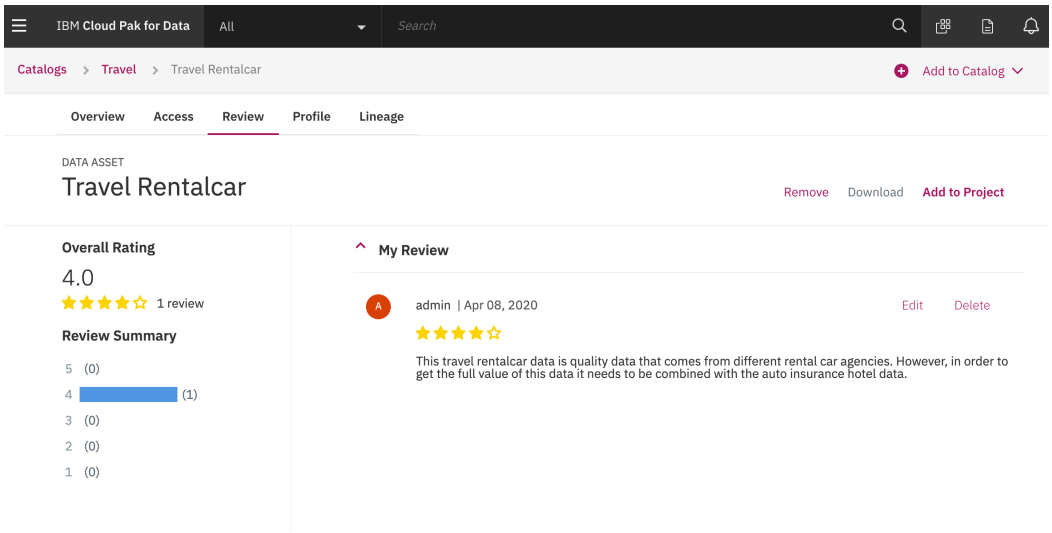


The profile of a data asset that contains relational or structured data, shows information about each column in the data set, based on the first 5,000 rows of data. The profile shows the frequency of the inferred attribute classifiers and statistics about the data for each column.

- Scroll **down** and **to the right** to review all the profiling statistics.
- Click the **Review** section to rate and review the data asset.



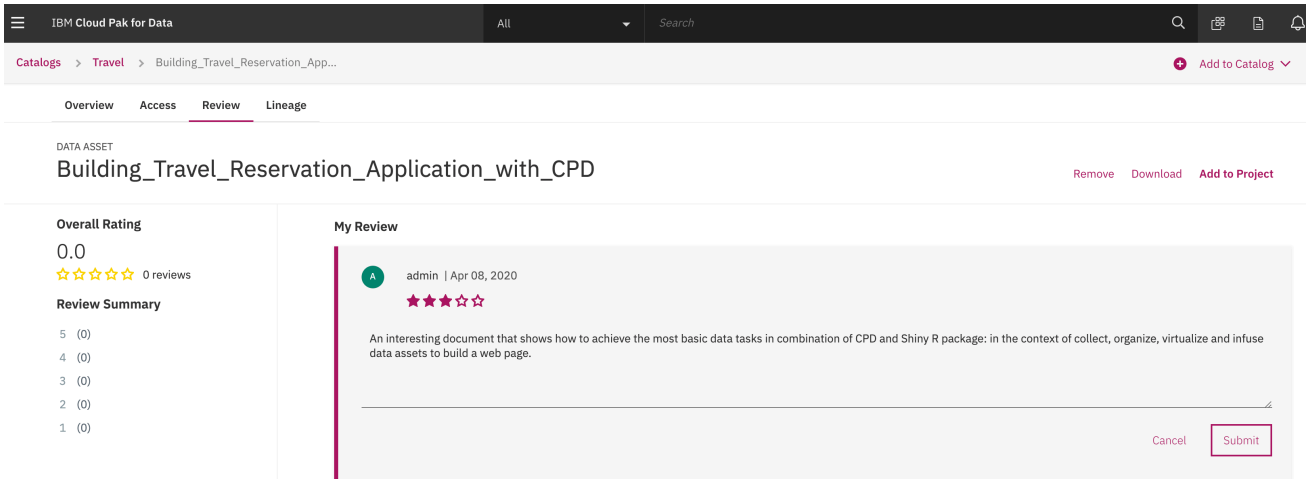
- Copy and Paste the following bolded text into the Description: **This travel rentalcar data is quality data that comes from different rental car agencies. However, in order to get the full value of this data it needs to be combined with the auto insurance hotel data.**
- Click the **4th star** from the left to give the asset a 4 stars rating.
- Click the **Submit** button.



Now you should have one review with an Overall Rating of 4.0.

Let's take a look into one of the documents in the **Travel** catalog. Click on the **Building_Travel_Reservation_Application_with_CPD** data asset that tag as a document to view the properties. As this is a PDF asset, you should able to see the document also. **Scroll** down to view the content of the document.

Click the **Review** section to rate and review the data asset.

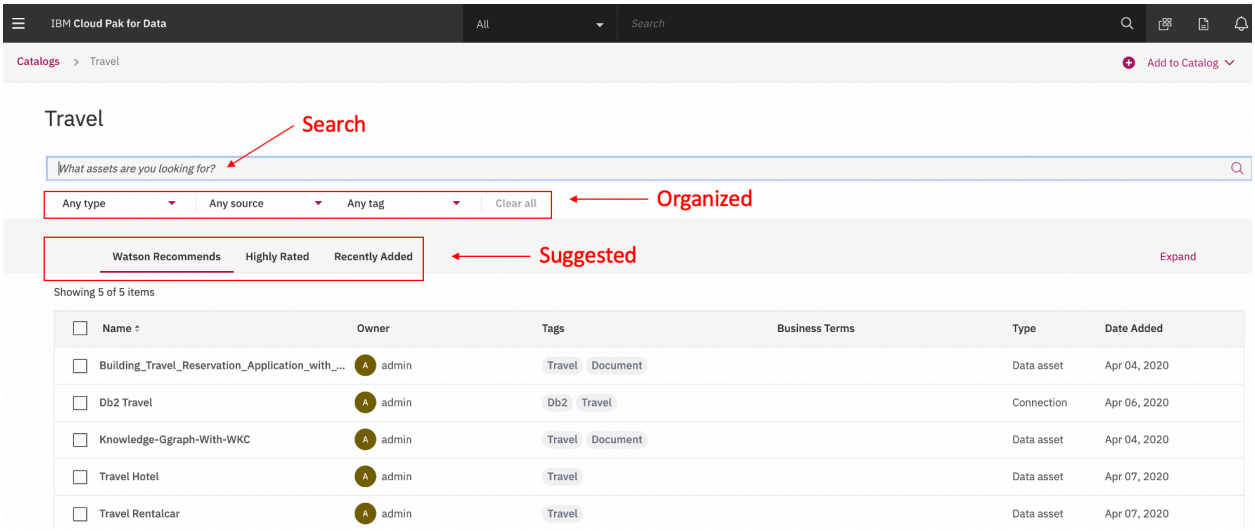


- Copy and Paste the following bolded text into the Description: **An interesting document that shows how to achieve the most basic data tasks in combination of CPD and Shiny R package: in the context of collect, organize, virtualize and infuse data assets to build a web page**.
- Click the **3rd star** from the left to give the asset a 3 stars rating.
- Click the **Submit** button.

12 Shop for Data

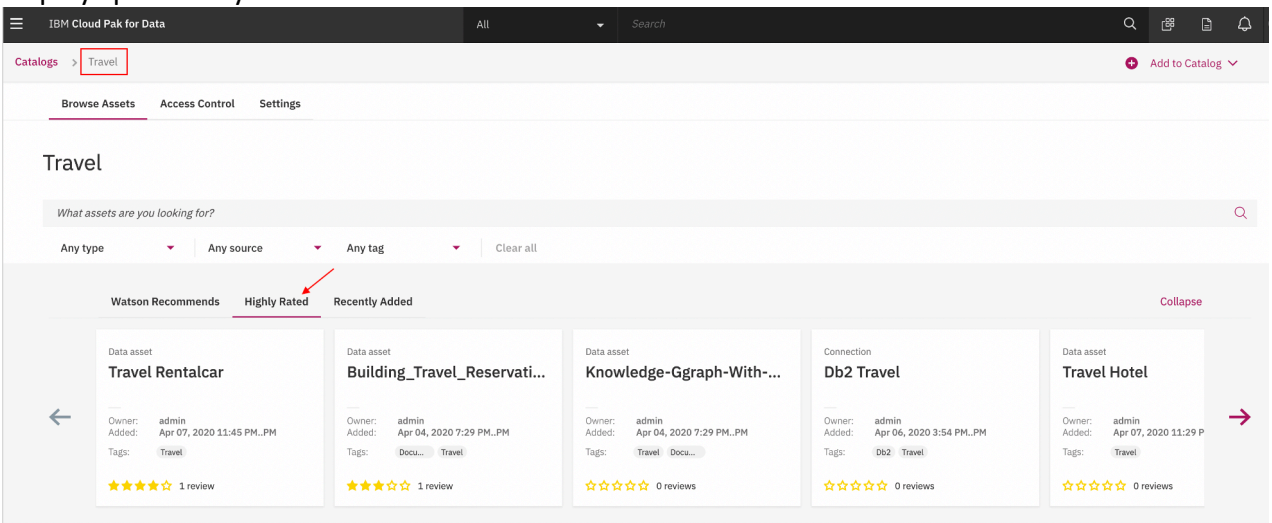
In this task, you will leverage Knowledge Catalog's intelligent **Shop for Data** AI-powered **Search and Suggest** experience that guides you to the most relevant assets in the catalog, based on understanding of relationships between assets, usage of those assets and social connections between the users of those assets.

You will also use the **Filter** section of the Knowledge Catalog that is automatically built and **Organized** by *Asset Type* and *Tag* as you catalog assets. Tagging is essential when cataloging assets, it expedites the process for consumers to easily search and find what they are looking for.



12.1 Shop Data Assets using Suggestion

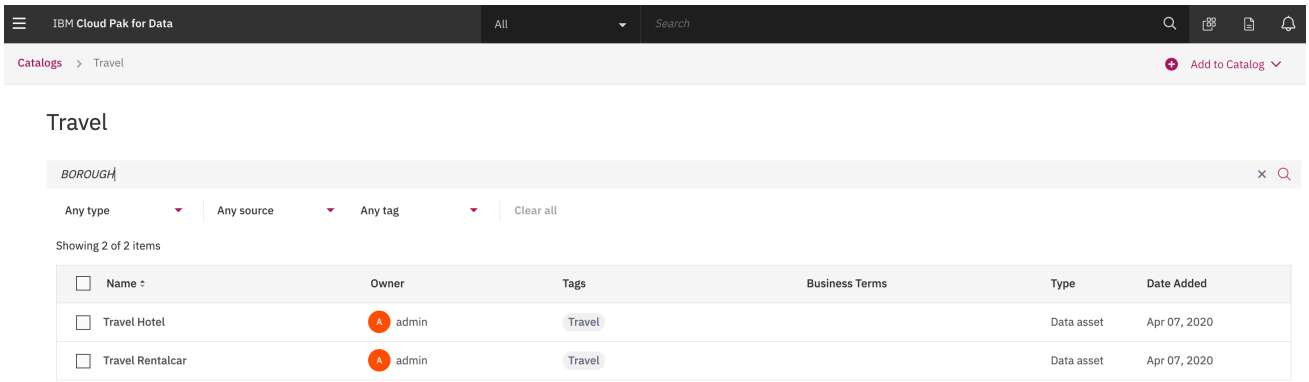
You can easily search for data using the **Watson Recommends**, **Highly Rated** and **Recently Added** suggestion categories to find relevant data. These categories are automatically populated by Knowledge Catalog as you catalog, curate and enrich data assets. For example: Highly Rated assets displays previously reviewed assets.



- Click the **Travel** link at the top of the page to go back to the catalog asset browser.
- Click the **Highly Rated** section.

12.2 Shop Data Assets using Search

In this section you will shop for data by specifying search criteria using the **Search area** of the Knowledge Catalog asset browser. The search criteria are not case sensitive. You search criteria could be a tag, column name, asset name, business terms, description etc.



- Inside the Knowledge Catalog search area, type in **BOROUGH**.
- Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **borough**.
- Click the **x** at the far right of the search area to clear the search.

13 Prepare Data for Analytics Project

Let's talk about the use case to have a better understanding on the analytics project. Imagine TravelBid is a startup, offering discounted hotel and rental car reservations in New York City. They have a bulk of hotel and rental car vendors, from where they fulfill travel reservation for their clients. TravelBid needs to quickly build an e-commerce platform where clients can "name their own price". Depends on client's named price, system will generate list of hotel and car matches within the prices. Plan here to build an interactive GUI based application to address TravelBid's requirement using Shiny R Package on CPD, accessing tables from different relational tables. You noticed earlier there are two travel related data sources. 1) The **HOTEL** table that stores information about hotels in New York City. 2) The **RENTALCAR** table that stores information about different rental car agencies and fleets. Schema of these two tables are as follows:

Table 1: HOTEL

Find

Column Name	Data Type
ACCOUNT_NUM	SMALLINT
BOROUGH	VARCHAR
LOCATION	VARCHAR
NAME	VARCHAR
NUM_RATIN	INTEGER
PRICE	INTEGER
RATING	INTEGER
SERVICE	VARCHAR

Table 2: RENTALCAR

Find

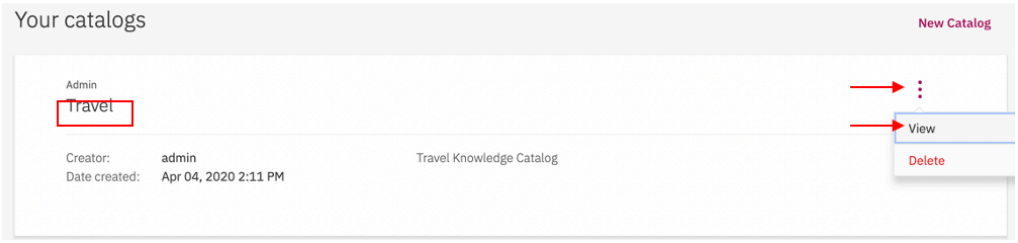
Column Name	Data Type
ACCOUNT_NUM	SMALLINT
AGENCY	VARCHAR
AGENCYADDR	VARCHAR
AGENCYPHONE	VARCHAR
BOROUGH	VARCHAR
MODEL	VARCHAR
RENT	INTEGER
SIZE	VARCHAR
SIZEID	SMALLINT

In this task, you will use the **IBM Data Refinery** to shape data that you cataloged in the **Travel** catalog. You will gain an understanding of the data preparation capabilities of the Data Refinery and how it can help you understand and visualize data in preparation for analytics project.

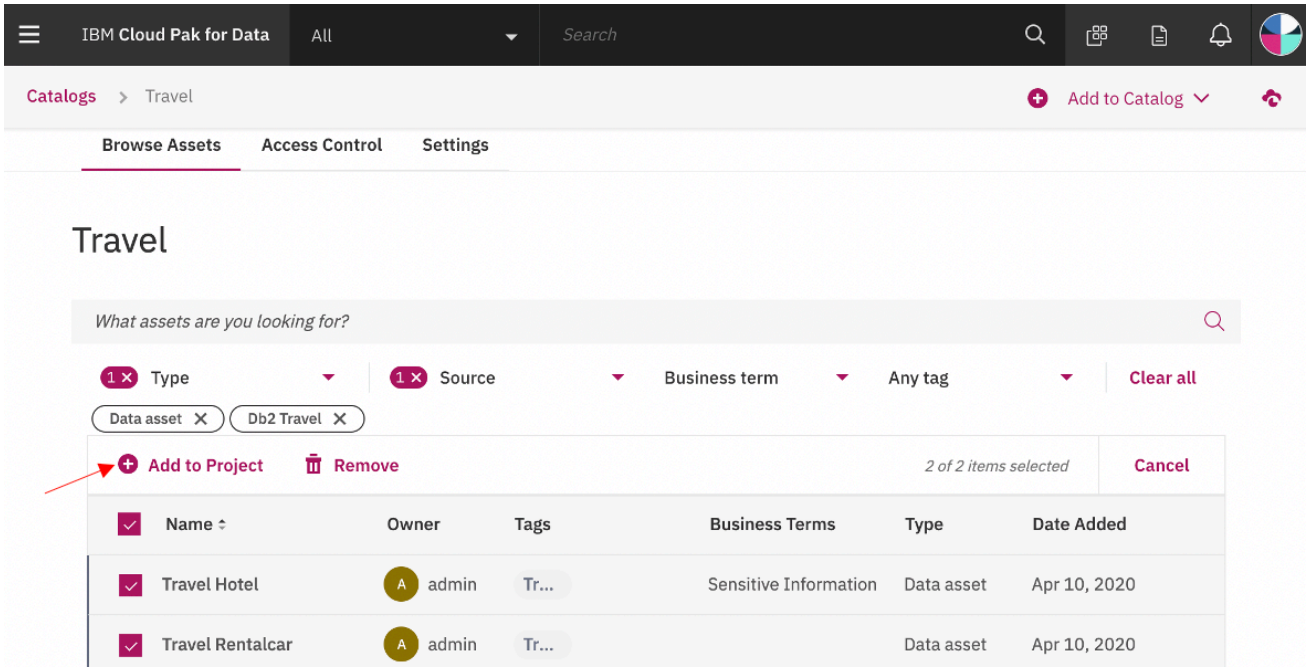
13.1 Add Catalogued Data Assets to Analytics Project

In order to refine data, the data needs to be in a project. You will add **HOTEL** and **RENTALCAR** data assets from the **Travel** catalog to the Travel project to prepare it for analytics and AI.

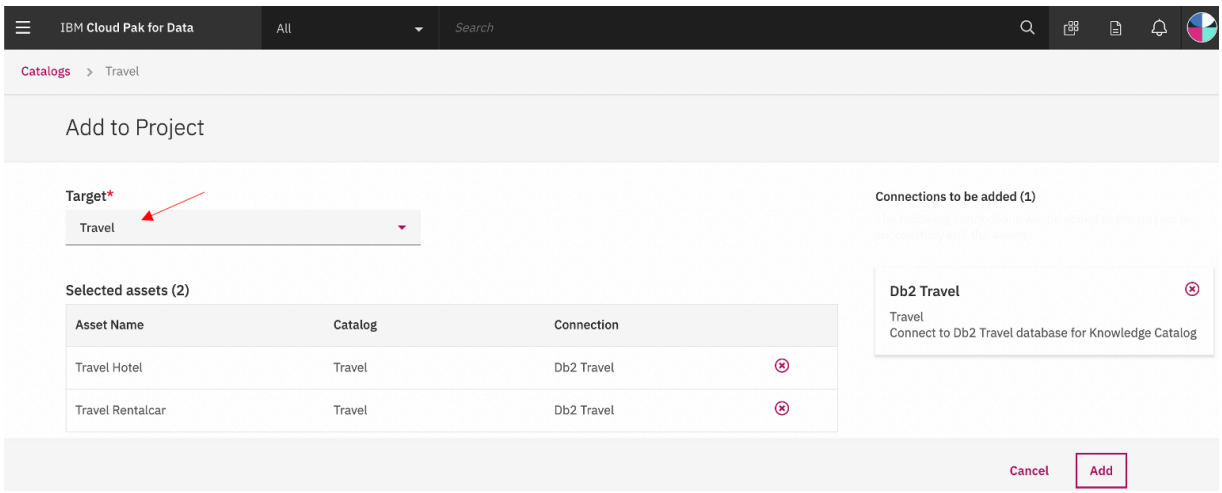
Go to the **Navigation menu** in the top left corner and click on **Travel** catalog from **Organize > All catalogs**.



- Click on the **configuration** icon of the **Travel** catalog.
- Select **View**



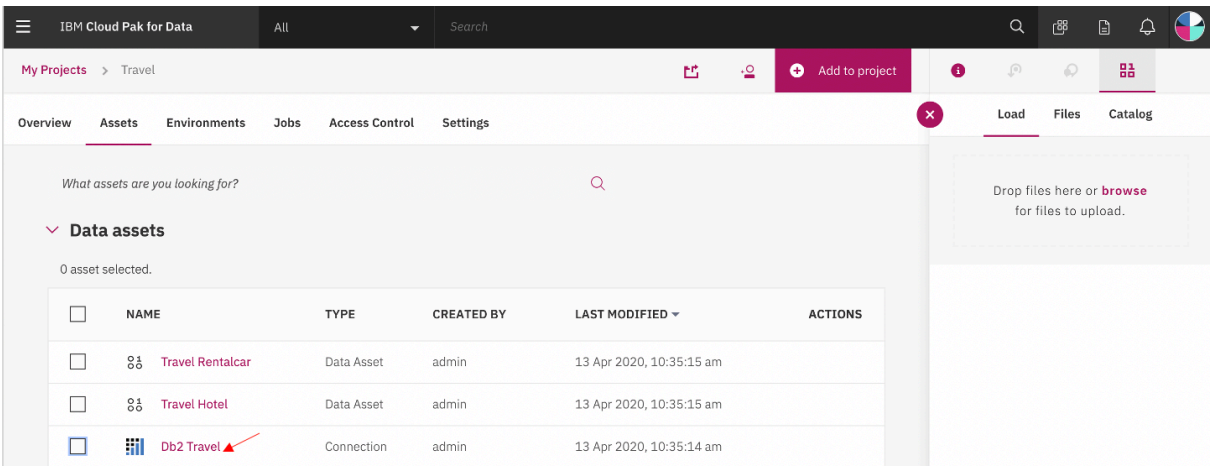
- Click the **check box** next to the **Travel Hotel** data asset.
- Click the **check box** next to the **Travel Rentalcar** data asset.
- Click the **check box** next to the **DB2 Travel** connection.
- Click the **Add to Project** link at the top of the data asset list.



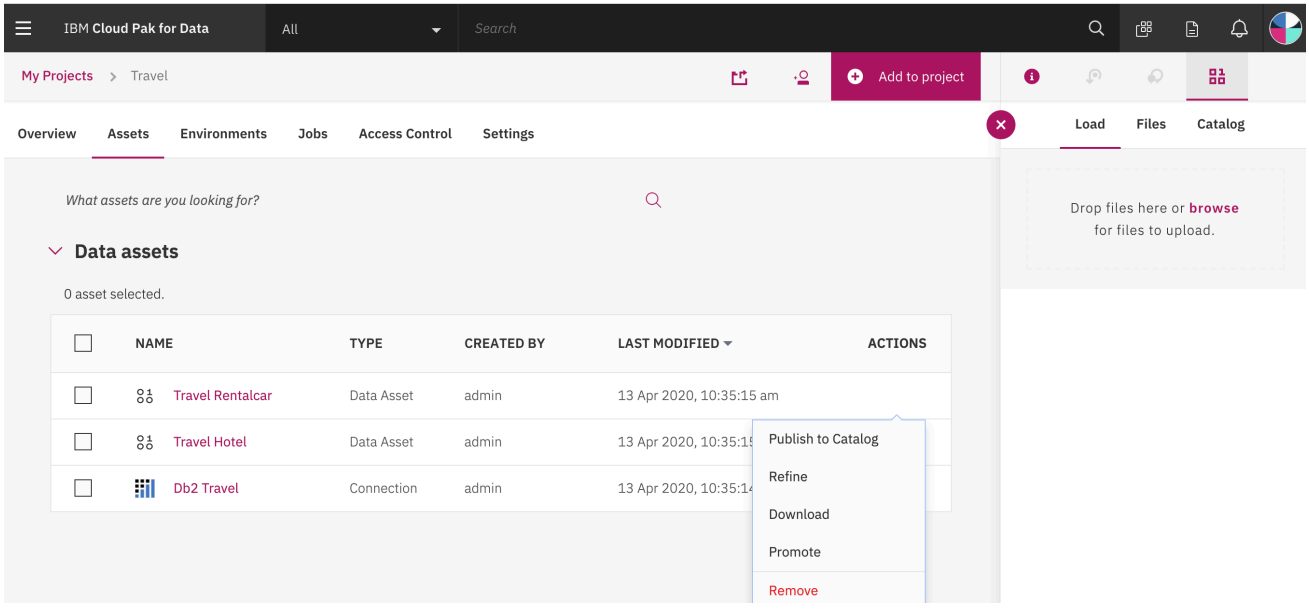
- Select the **Travel<id>** project from the list of Target projects.
- Click the **Add** button.

13.2 Refine Data

Go to the **Navigation menu** in the top left corner and chose **Projects > Travel**

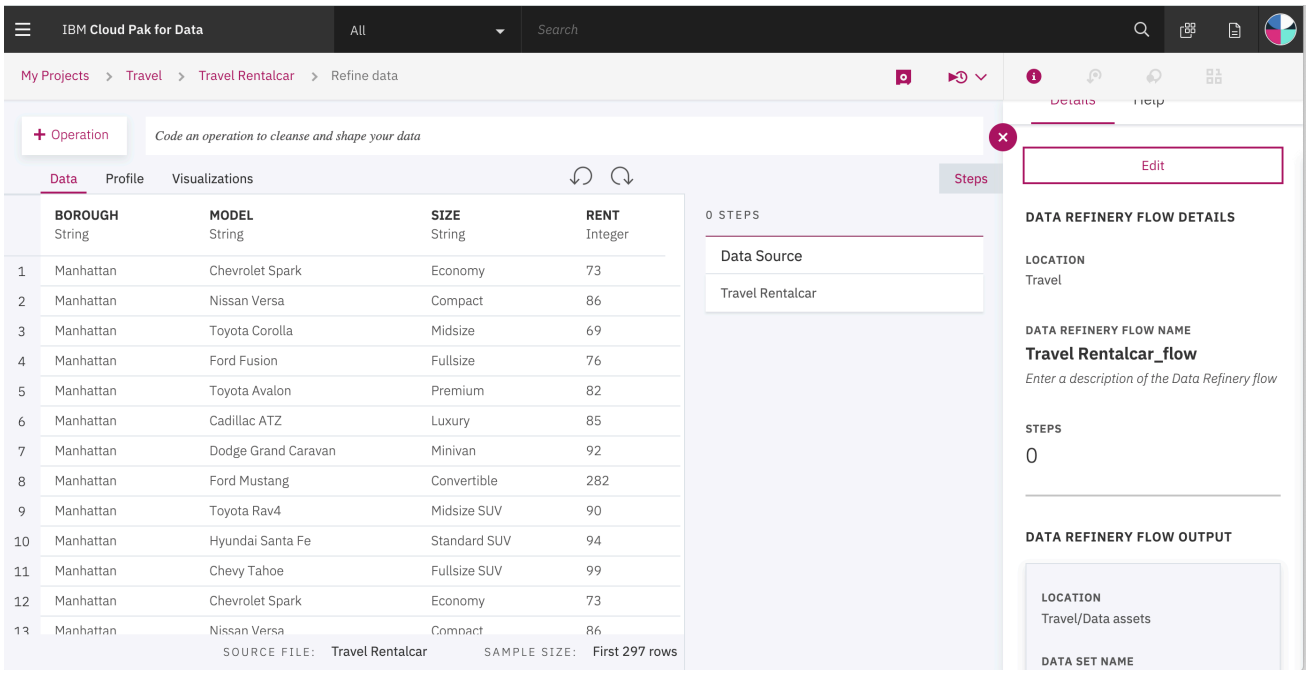


- First test the **Db2 Travel** connection. You should about connect to the database, if not check the username and password.

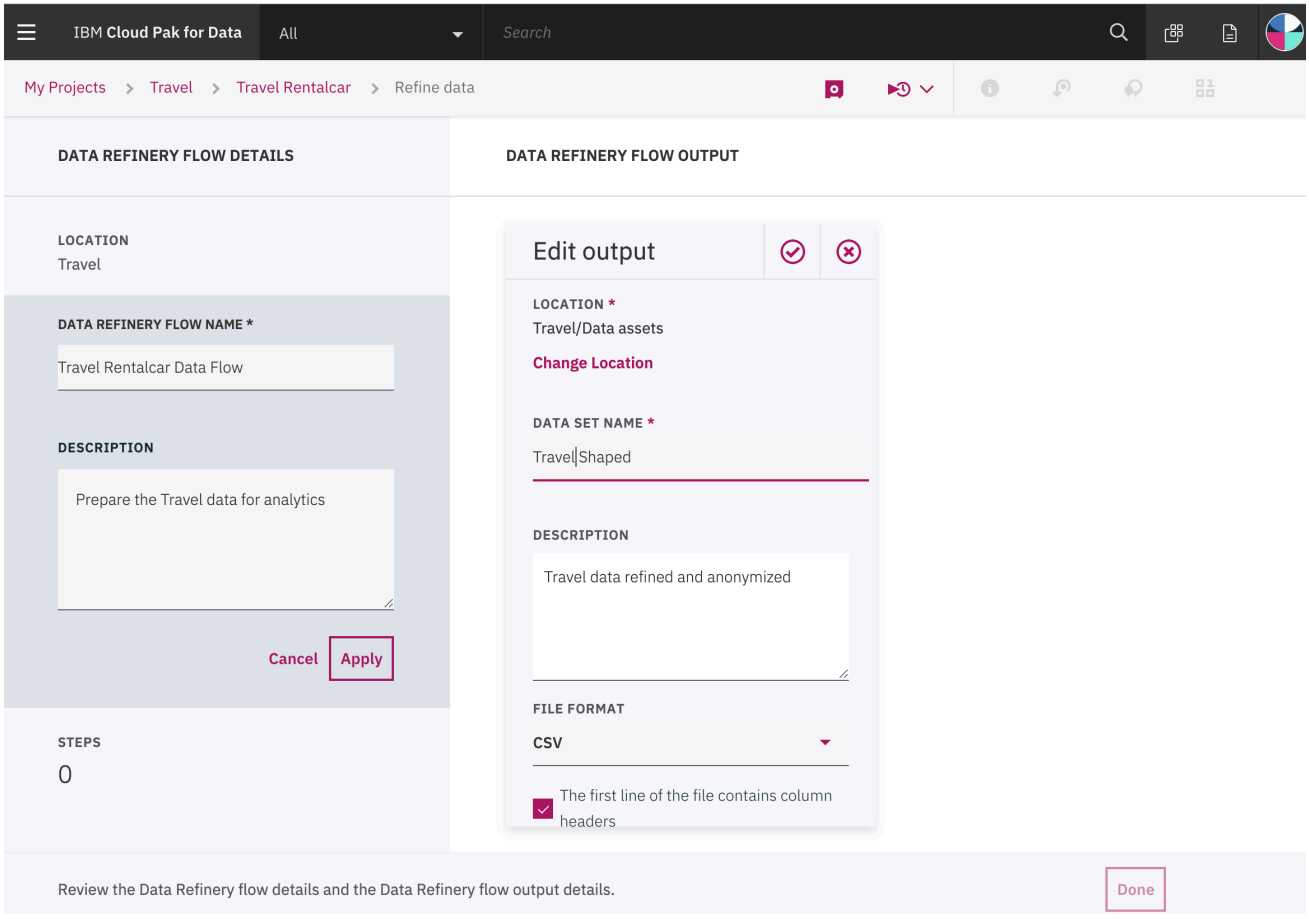


- Click on the **Assets** tab at the top of the project page.
- Select the **ellipses…** to the right of the **Travel Rentalcar** data asset to view the data asset action menu.
- Select the **Refine** menu item.

You are brought into the **IBM Data Refinery** to begin shaping the **Travel Rentalcar** data. In the subsequent steps, you will use some of the Data Refinery operations to shape the rentalcar data you added to the project and create a newly shaped dataset that you will put back to the project as a **CSV** file that will be used for analytics and AI.

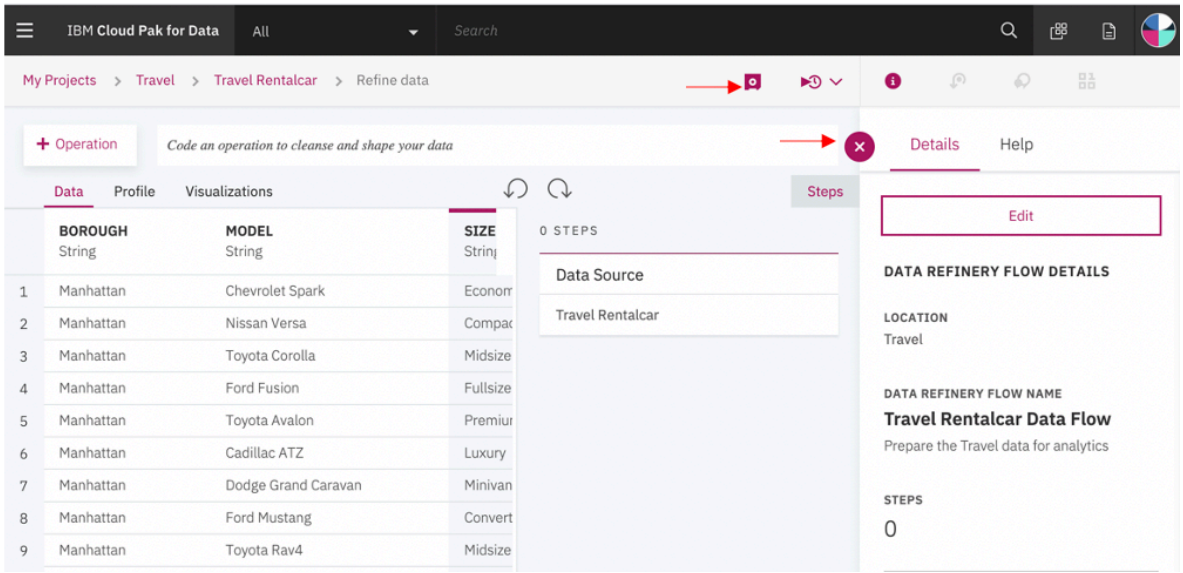


- Click in the **Edit** button to edit the Data Flow details



- Click the **pencil icon** in the DATA REFINERY FLOW NAME area of the DATA REFINERY FLOW DETAILS section.

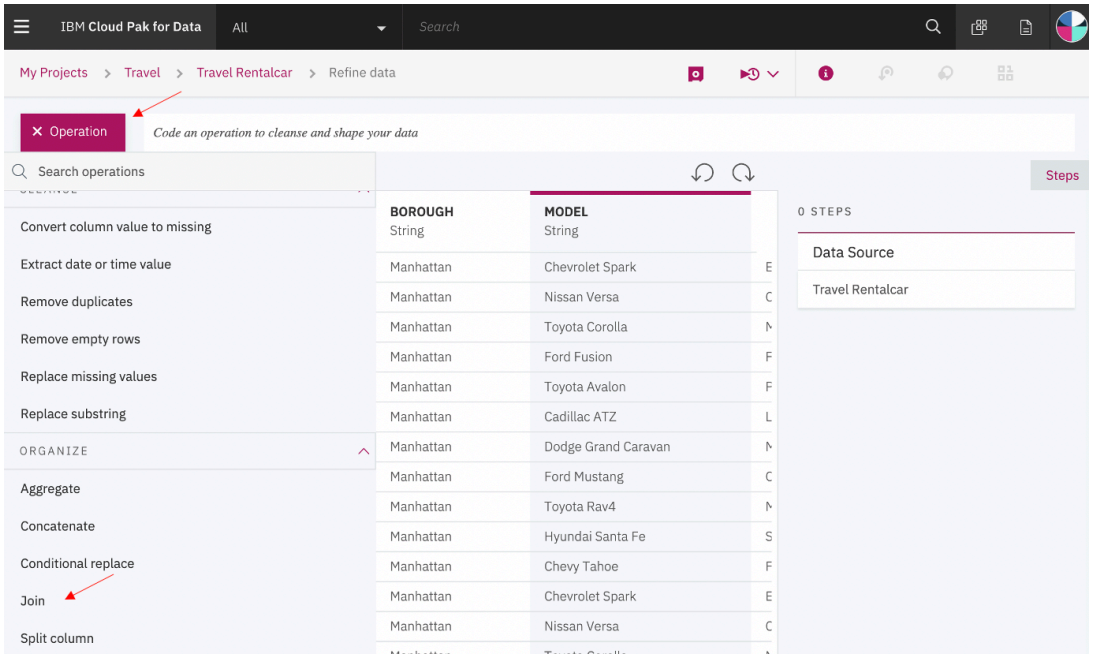
- Rename the Data Flow to **Travel Rentalcar Data Flow**.
- Use **Prepare the Travel data for analytics** for the DESCRIPTION field.
- Click the **Apply** button.
- Hover over the **pencil icon** in the LOCATION area of the DATA REFINERY FLOW OUTPUT section.
- Click the **Edit Output** button to change the DATA SET NAME.
- Rename the Data Set to **Travel Shaped<id>**.
- Use **Travel data refined and anonymized** into the DESCRIPTION field.
- Click the **Checked** icon on the toolbar to save the changes.
- Click the **Done** button.



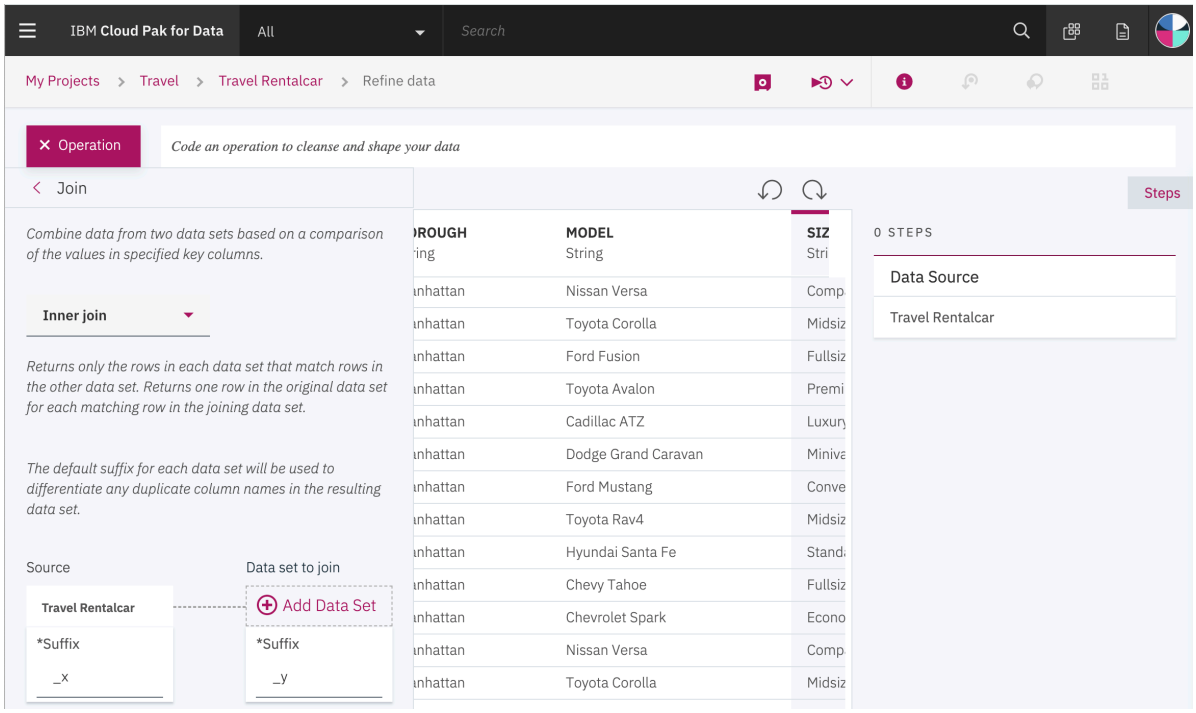
- Click the **Save** button on the toolbar to save the Data Flow.
- Click the **X** on the Details panel to close the panel and maximize the shaper real estate.

13.3 Combine Data

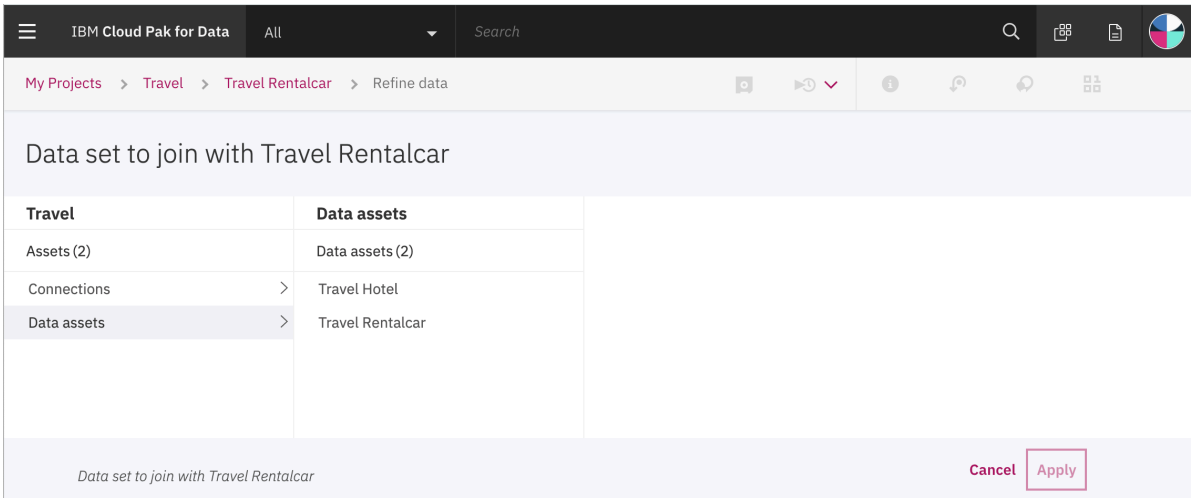
You will merge Rentalcar and Hotel tables together and create a single data asset for analytics and AI.



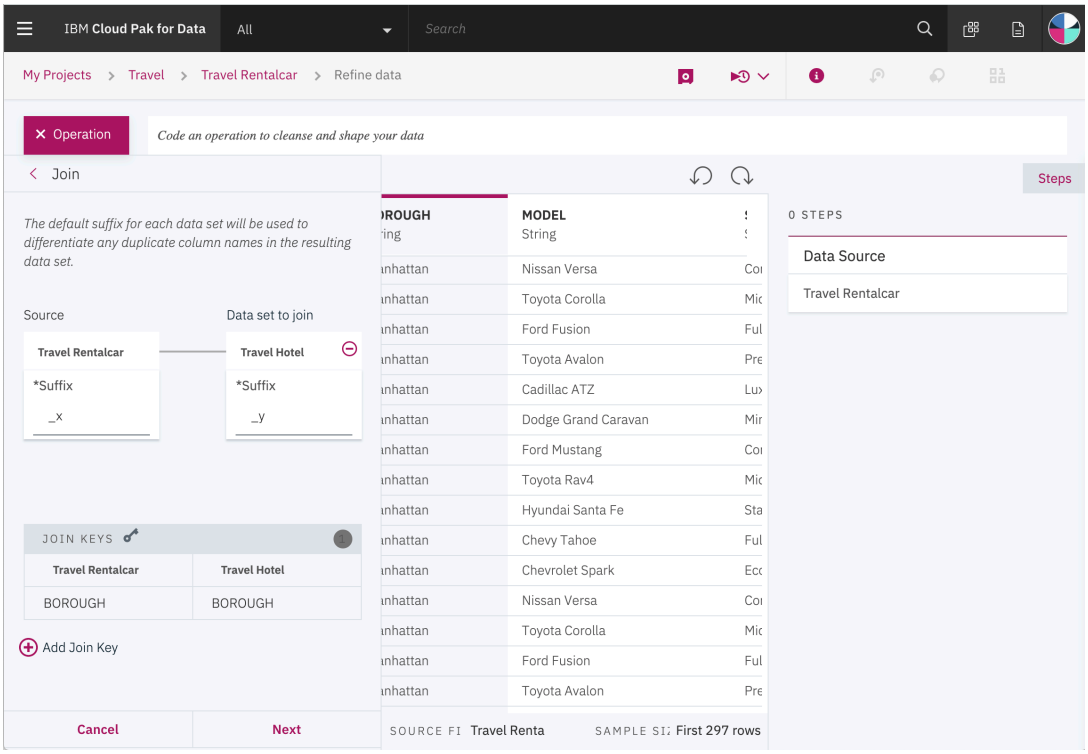
- Click the **Operation** button to view the shaping operations menu.
- Scroll down and click the **Join** operation.



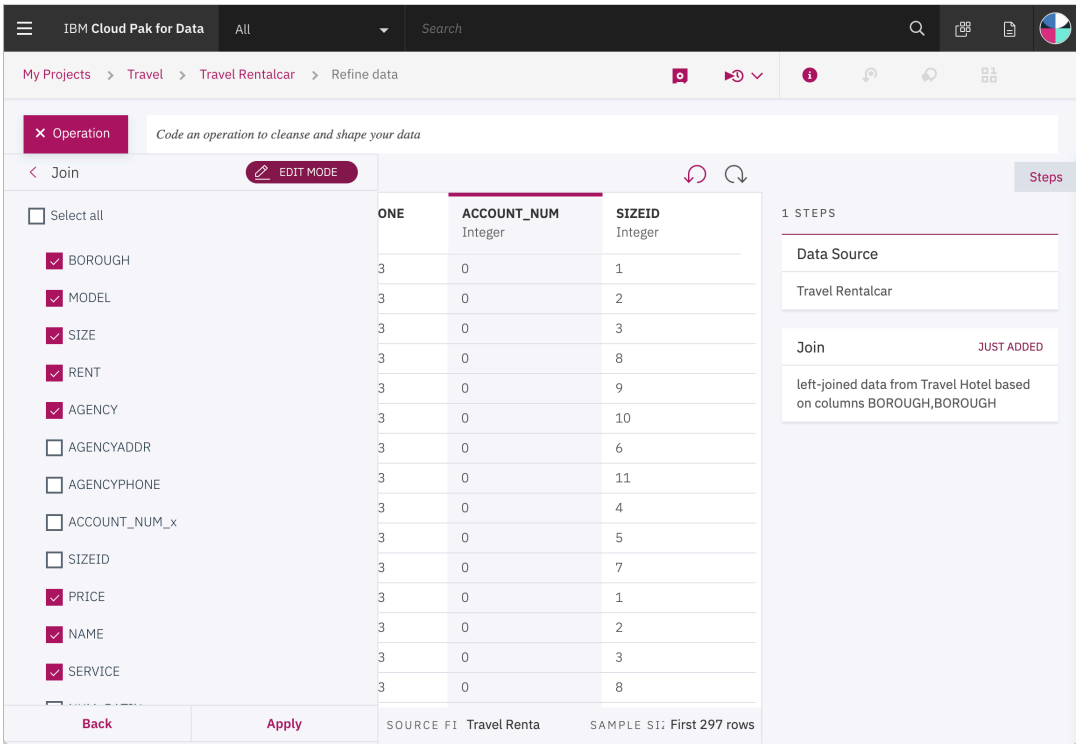
- Select the **Inner join** method from the join method list.
- Click the **+ Add Data Set** button in the **Data set to join** section.



- Click on the **Data assets** section.
- Select the **Travel Hotel** data asset.
- Click the **Apply** button.



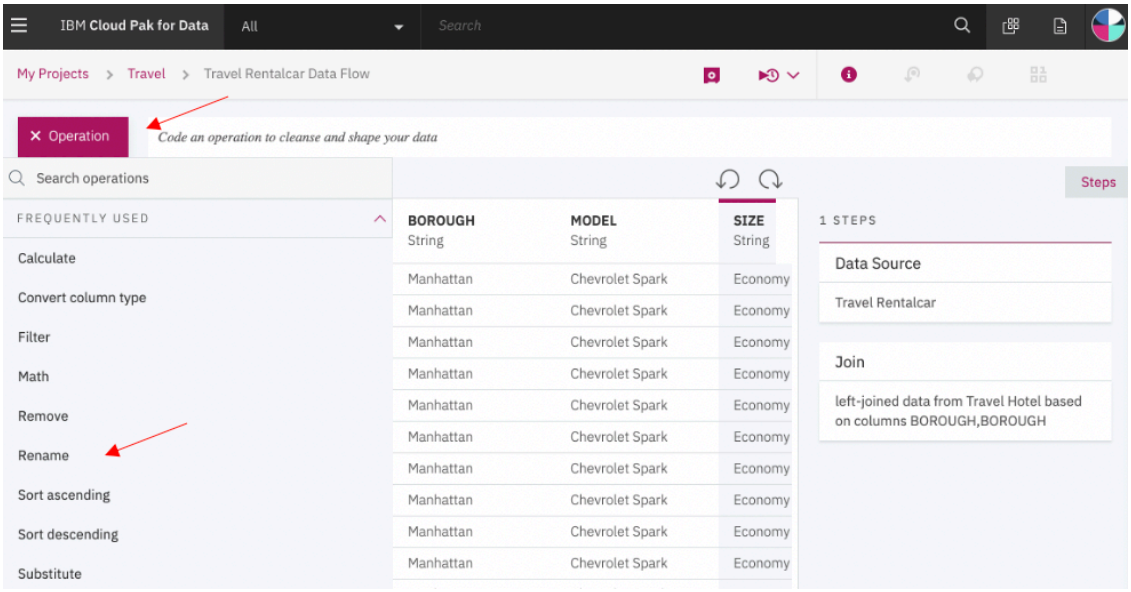
- Scroll down in the Join properties area until you see the JOIN KEYS section.
- Click in the **Travel Rentalcar** JOIN KEYS column selection list on the left and select the **BOROUGH** column as the join key column.
- Click in the **Travel Hotel** JOIN KEYS column selection list on the right and select the **BOROUGH** column as the join key column.
- Click the **Next** button.



- Scroll down the column list and keep **checked** the following columns:
- BOROUGH, MODEL, SIZE, RENT, AGENCY, PRICE, NAME, SERVICE, RATING
- Unchecking columns excludes them from the join result.
- Click the **Apply** button.
- Click the **Save** button on the toolbar to save the Data Flow.

13.4 Rename Column

Some of the column names are ironies. Let’s change them to a meaningful name.



- Click the **Operation** button to view the shaping operations menu.

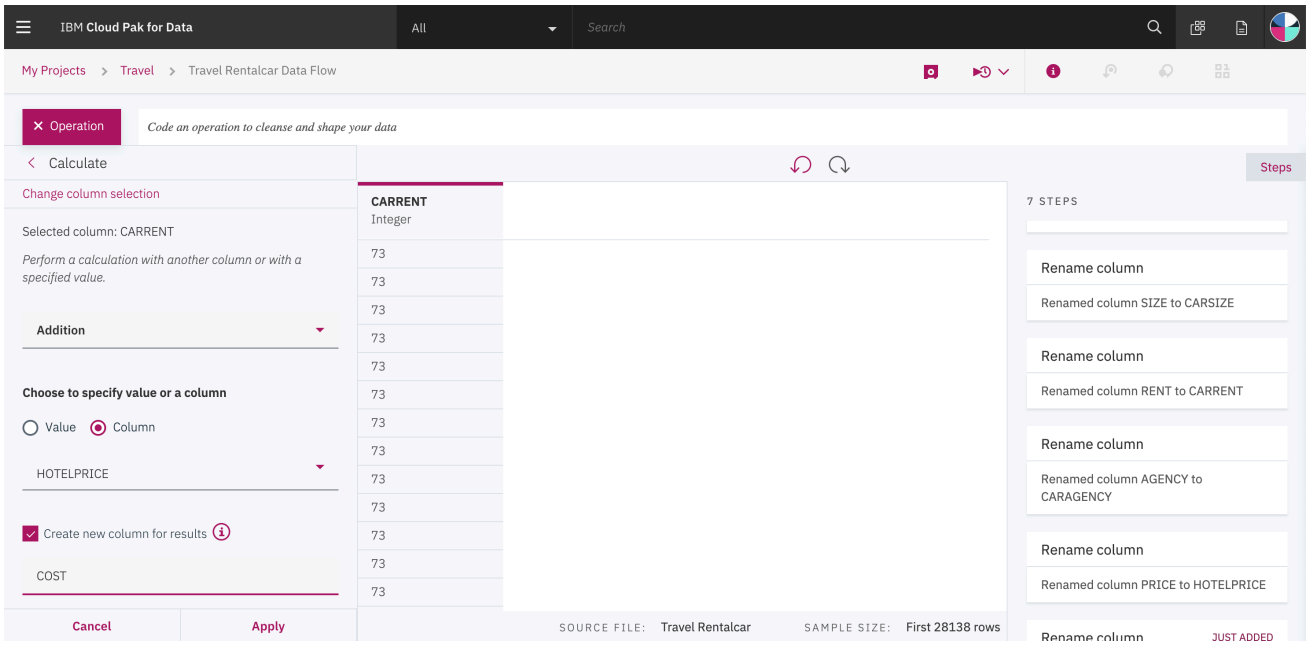
- Click the **Rename** operation.
- Then **select column name** as **MODEL** > Click **Next** > **Rename column name** with **CARMODEL** > Click on **Apply**.
- Click the **Save** button on the toolbar to save the Data Flow.

Similar way you can rename following columns:

- SIZE with CARSIZE
- RENT with CARRENT
- AGENCY with CARAGENCY
- PRICE with HOTELPRICE
- NAME with HOTELNAME

13.5 Create a New Column

In the travel dataset there are **RENT** column represents daily rental cost of a car, where as **PRICE** represent daily hotel room rate. You need to create new column named **COST** that stores value of RENT + PRICE.

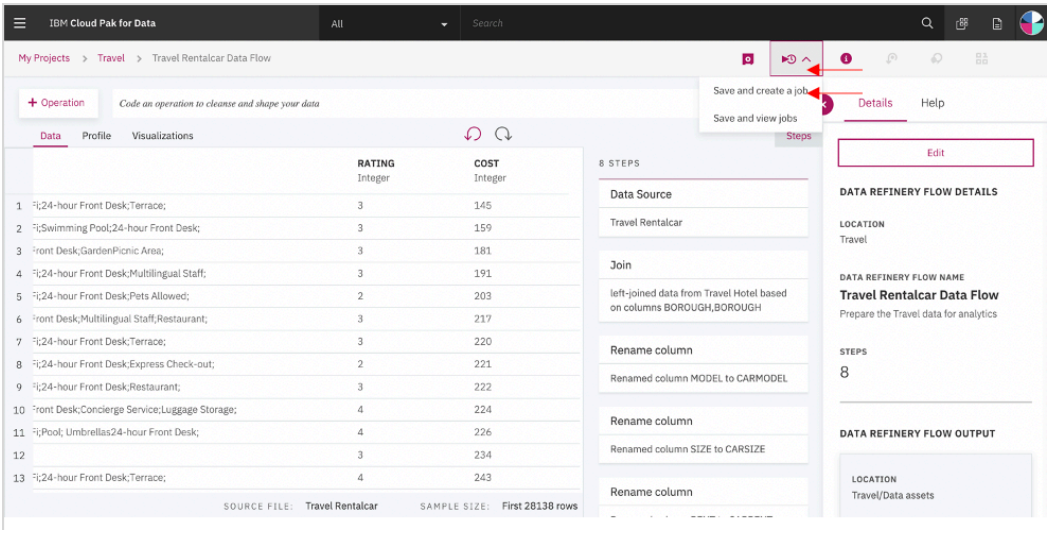


- Click the **Operation** button to view the shaping operations menu.
- Scroll down and click the **Calculate** operation.
- Select column **CARRENT**
- Click **Next**
- Choose **Addition** for Performa a calculation…
- Specify **Column** radio button
- Select column **HOTELPRICE**
- Check the box for Create new column for results.
- Use **COST** as the name of the new column for result

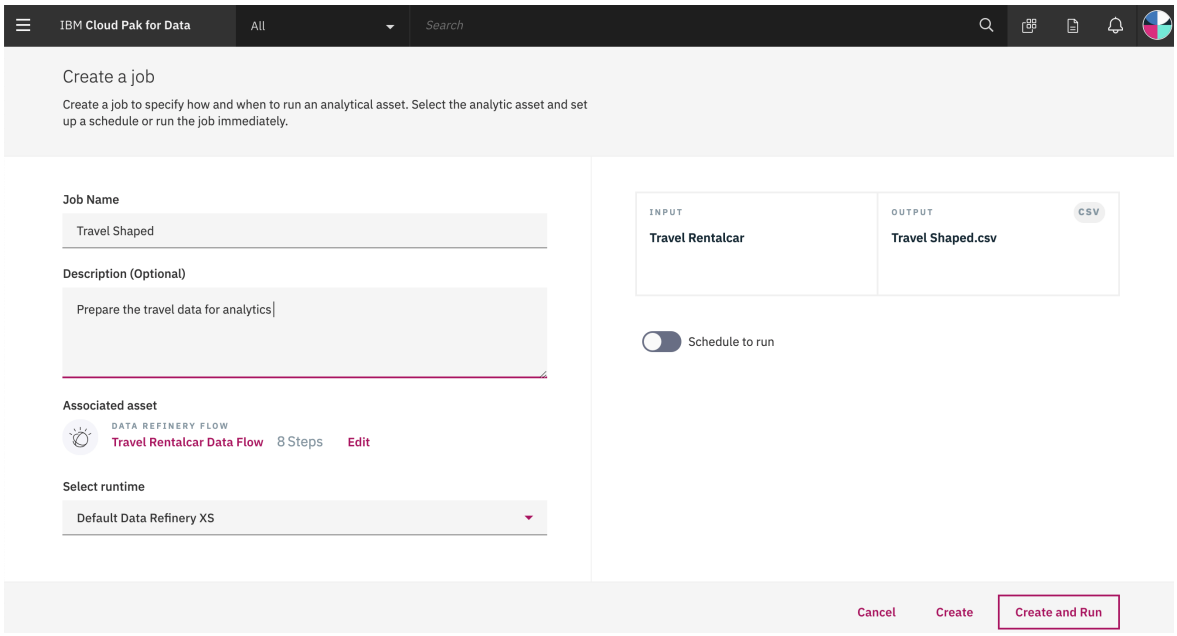
- Click **Apply**
- Click the **Save** button on the toolbar to save the Data Flow.

13.6 Run Data Flow

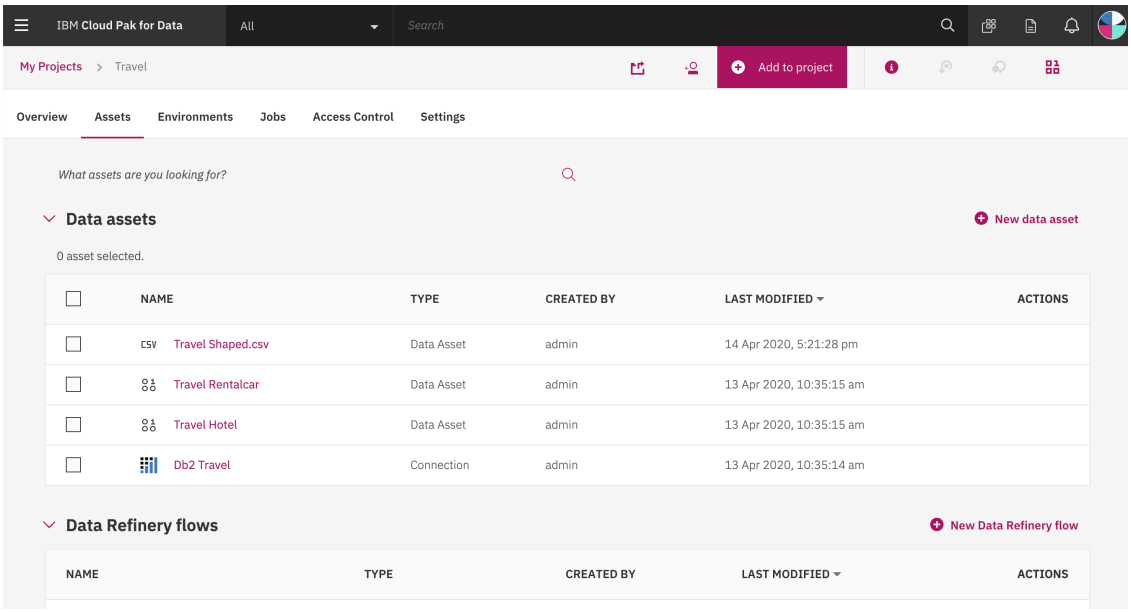
In order to process the shaping operations, you need to create a **Job** and run it. The job will use the data flow's output data set name, target location and format type to place and create the data flow output. Based on the changes you specified, the job will create a CSV file named **Travel Shaped** in your **Travel** project.



- Click the **Jobs** button on the toolbar.
- Select the **Save and create a job** menu item.



- Enter a Job Name of **Travel Shaped<id>** with the proper case, and spaces between the words.
- Use description as **Prepare the travel data for analytics**.
- Use the **Default Data Refinery XS** runtime, it should be pre-selected.
- Click the **Create and Run** button.



- Click on the **Travel** project navigation link on the toolbar to get back to the sections of the project.
- Click on the **Assets** tab to view the project assets.

14 Congratulation!

You have created the new data asset named **Travel Shaped**. This is the CSV dataset the Data Refinery generated based on your data shaping operation. As a data engineer you can handed it over to the data scientist for their data science project.