Introduction to Watson Knowledge Catalog

Introduction

This lab will introduce Watson Knowledge Catalog. Watson Knowledge Catalog is a secure enterprise catalog to discover, catalog and govern your data/models with greater efficiency. The catalog is underpinned by a central repository of metadata describing all the information managed by the platform. Users will be able to securely share data with their colleagues more easily, regardless of what the data is, where it is stored, or how they intend to use it. In this way, the intelligent asset catalog will unlock the value held within that data across user groups—helping organizations use this key asset to its full potential.

Objectives

The goal of the lab is for the users to gain familiarity with the features of the Watson Knowledge Catalog. Before exercising the catalog functionality, the user will need to set up the environment needed to demonstrate the catalog features. This will include downloading the data assets, creating a database service, uploading two of the data assets to the database, and adding a user to the Watson Studio account. Once this is completed, we will perform the following catalog tasks:

1. Create a governed catalog
2. Add a member to the catalog
3. Add Data Assets to the catalog
4. Search the catalog
5. Edit/Review/Profile a structured Data Asset
6. Demonstrate access control features.
7. Demonstrate policy creation and enforcement (optional)
8. Push the Data Assets to the project set up in the prerequisites.

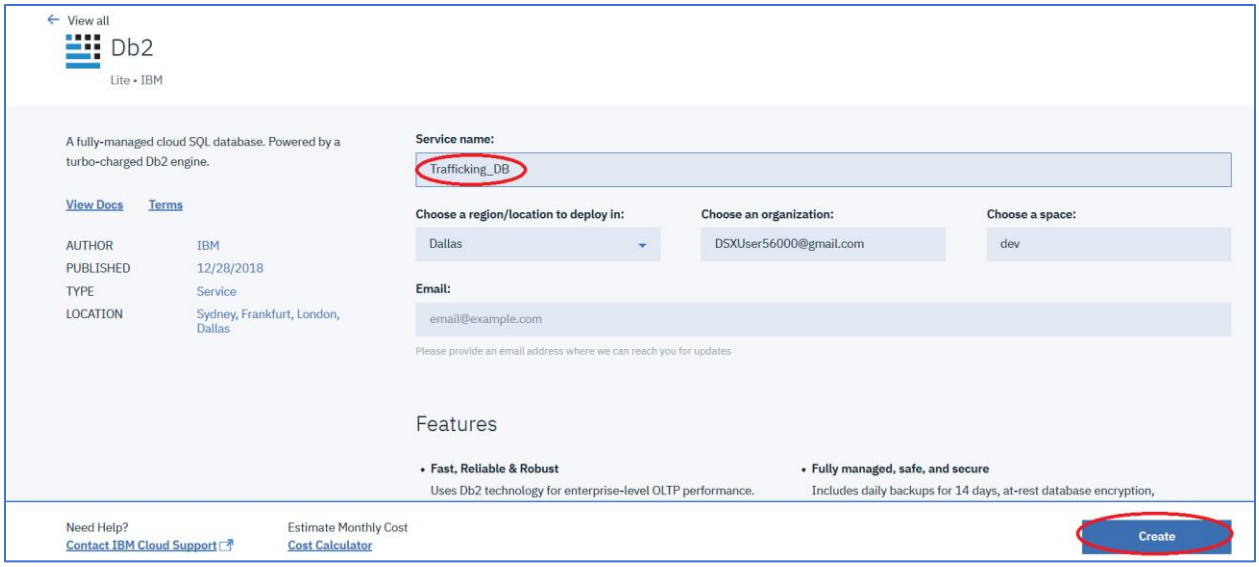
Set Up Environment

Step 1: Download the trafficking_data.zip from the github repository

1. Click on trafficking_data.zip, to download the zipped file.
2. Extract the file contents. You should have three files extracted, (1) female_human_trafficking.csv, (2) Categories.csv, and (3) Occupation.csv.

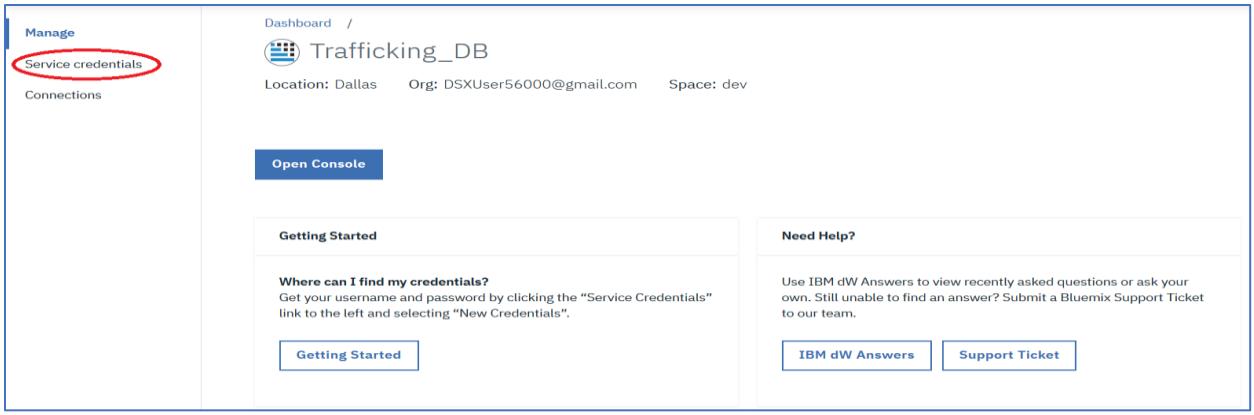
Step 2: Create a DB2 on Cloud service and Load the female_human_trafficking.csv file into the Database.

1. Click on DB2 to create the **DB2 on Cloud** service.
2. Enter Trafficking_DB as the **Service Name** and then click on **Create**.

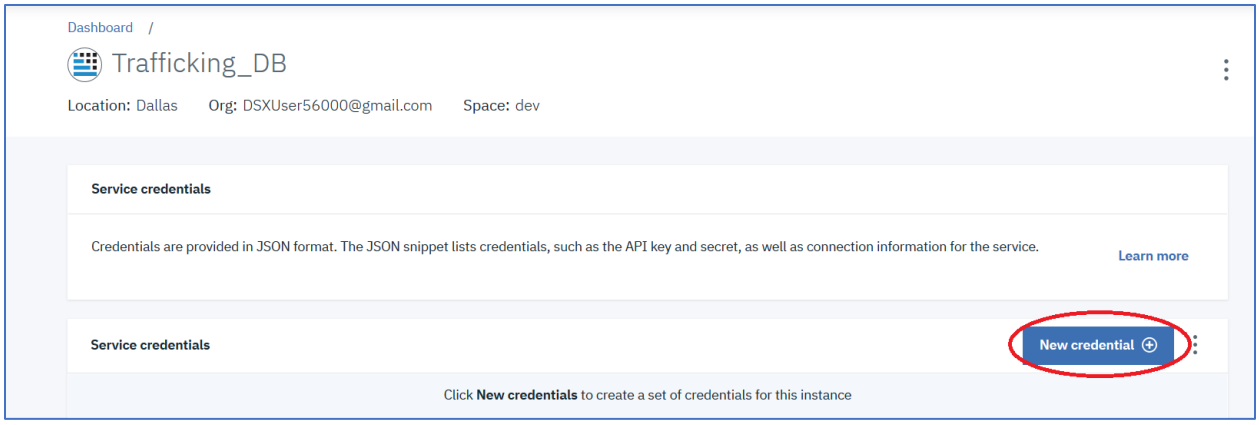


3. Click on Trafficking_DB to open the service.

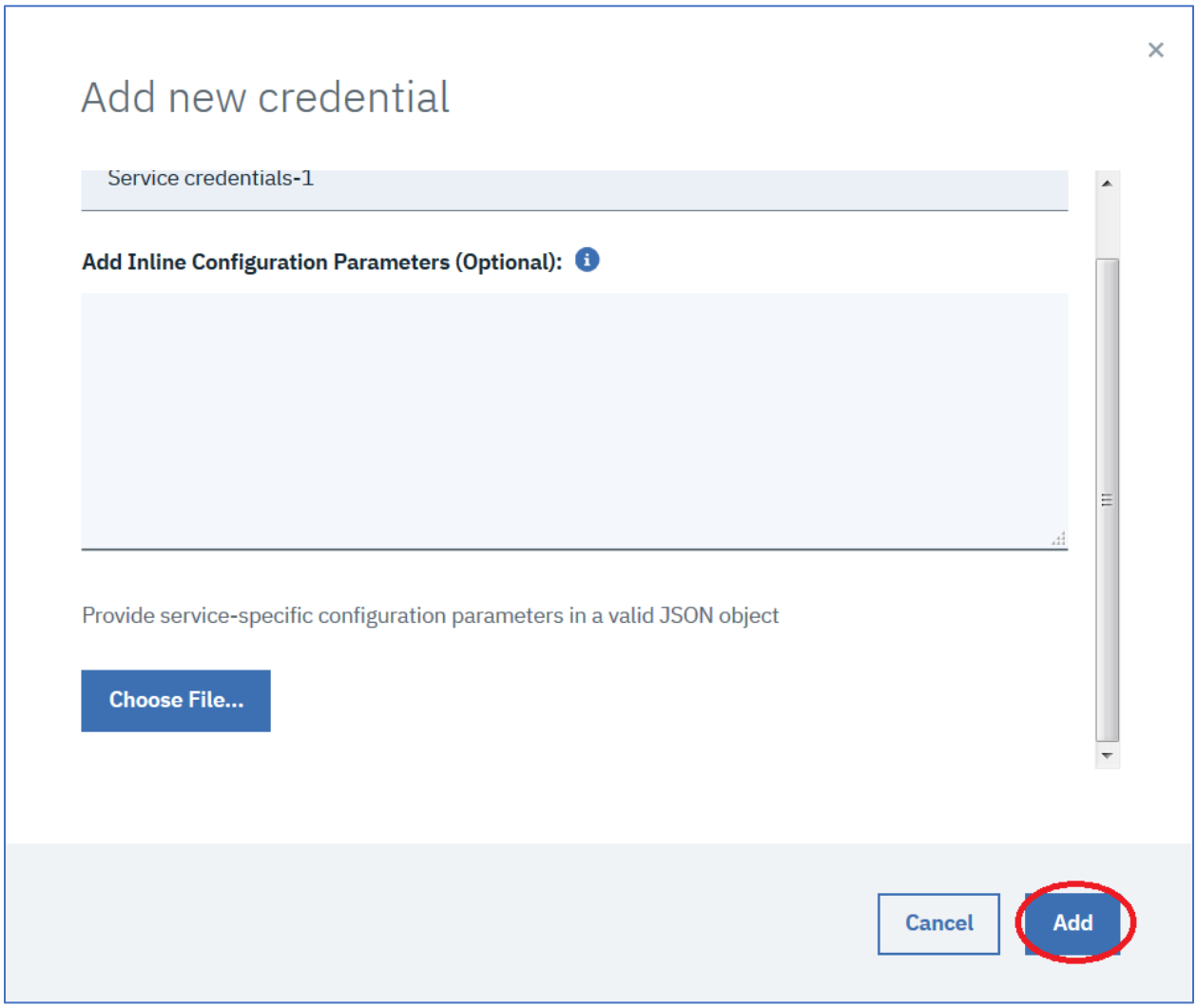
4. We will need the Service Credentials for the DB2 on Cloud service later in the lab. The next few steps will generate the Service Credentials. You will then cut and paste these credentials into a text editor for later use. Click on **Service Credentials**.



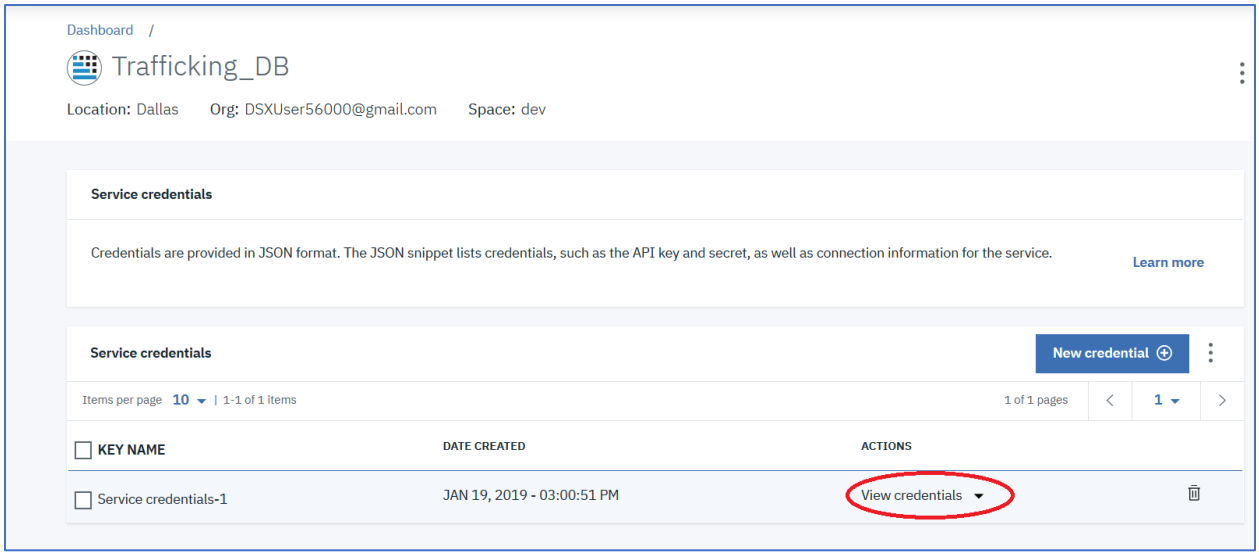
5. Click on **New credential**.



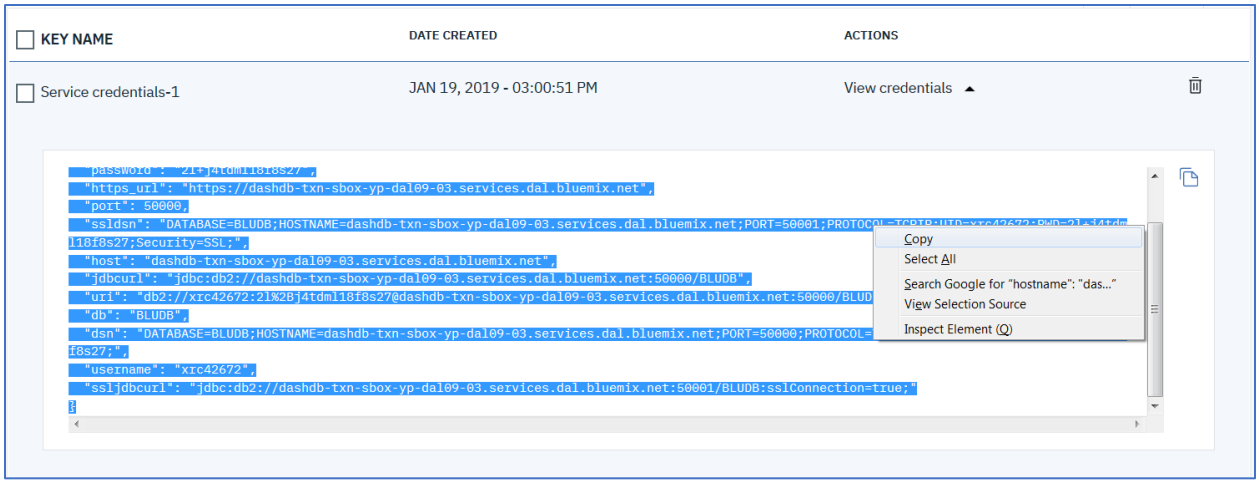
6. Click on **Add**.



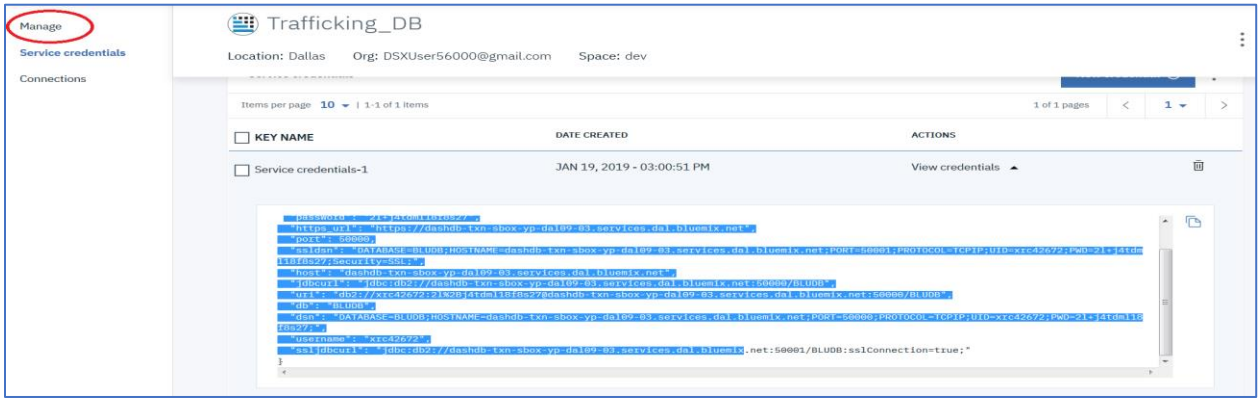
7. Click on **View Credentials**.



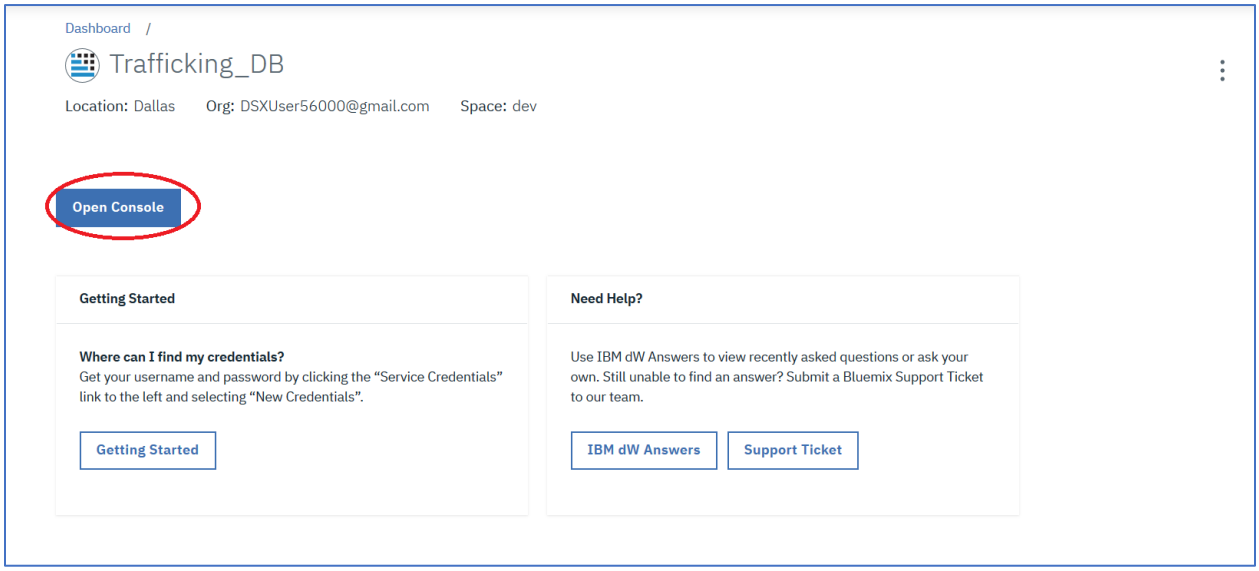
8. Cut and paste the credential information into a text editor. We will need the username, password, host, and db fields to set up a connection to the database service.



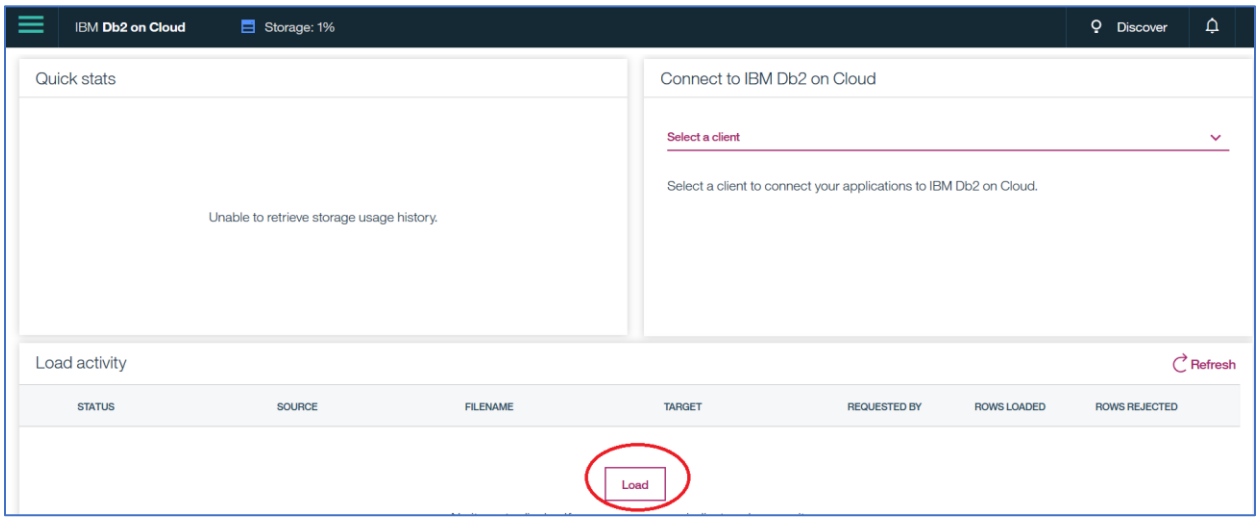
9. We are now going to load the female_human_trafficking file contents into the database service that we just created. The DB2 on Cloud service console provides a built-in load capability. Click on **Manage**.



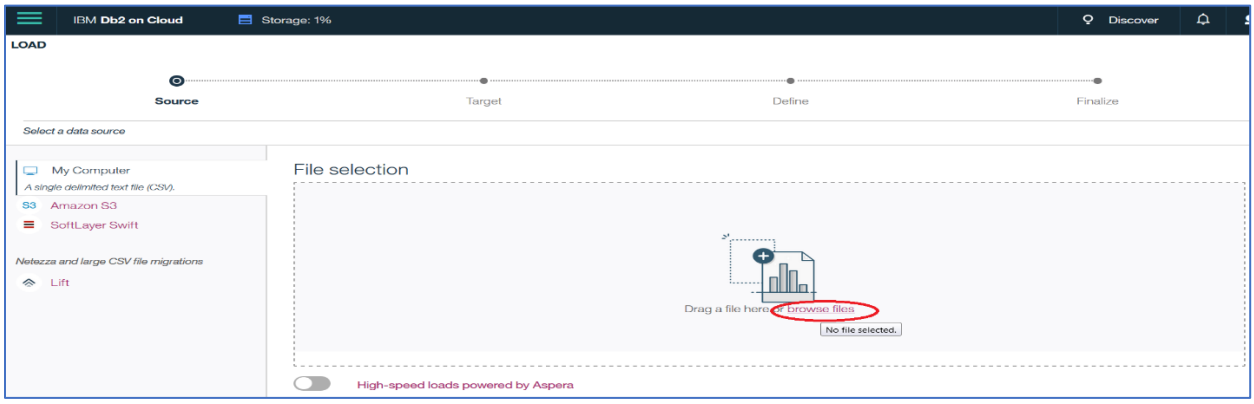
10. Click on **Open Console**



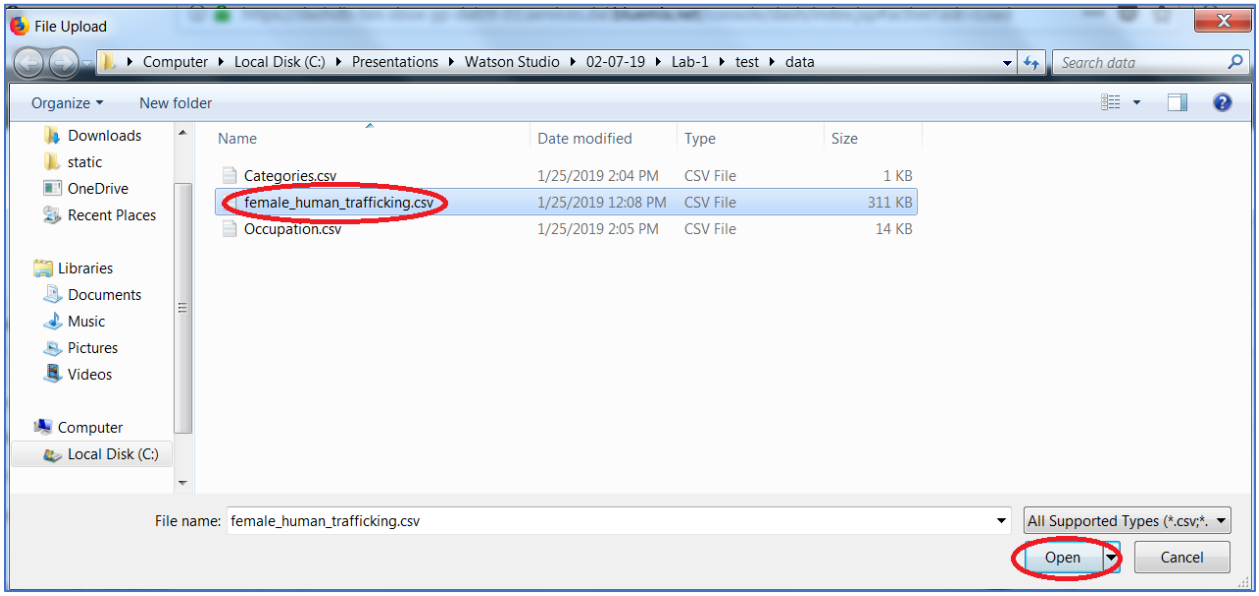
11. Click on **Load**.



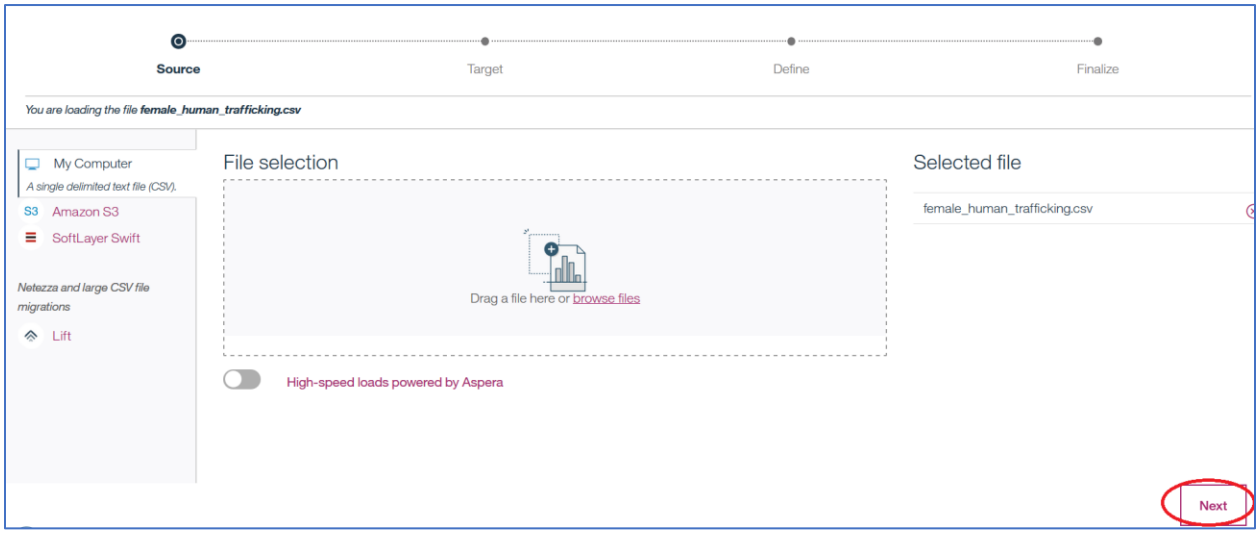
12. Click on **browse files**.



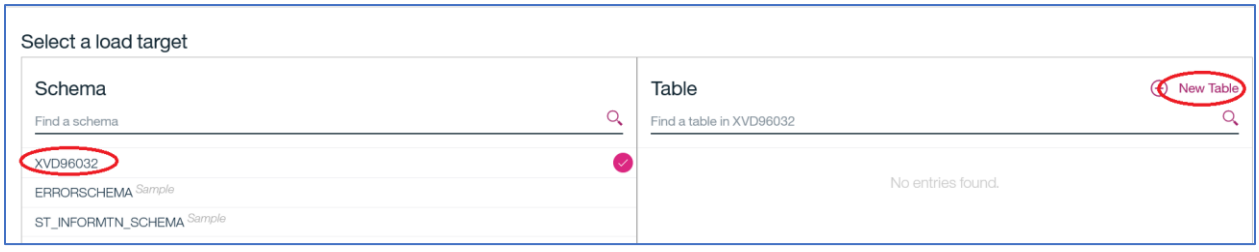
13. Navigate to the folder where you downloaded the trafficking files. Select **Categories.csv,** and then click on **Open**.



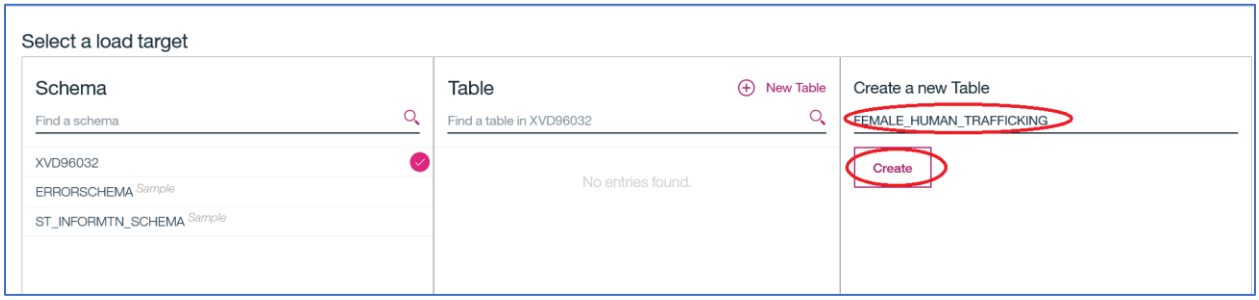
14. Click on **Next**.



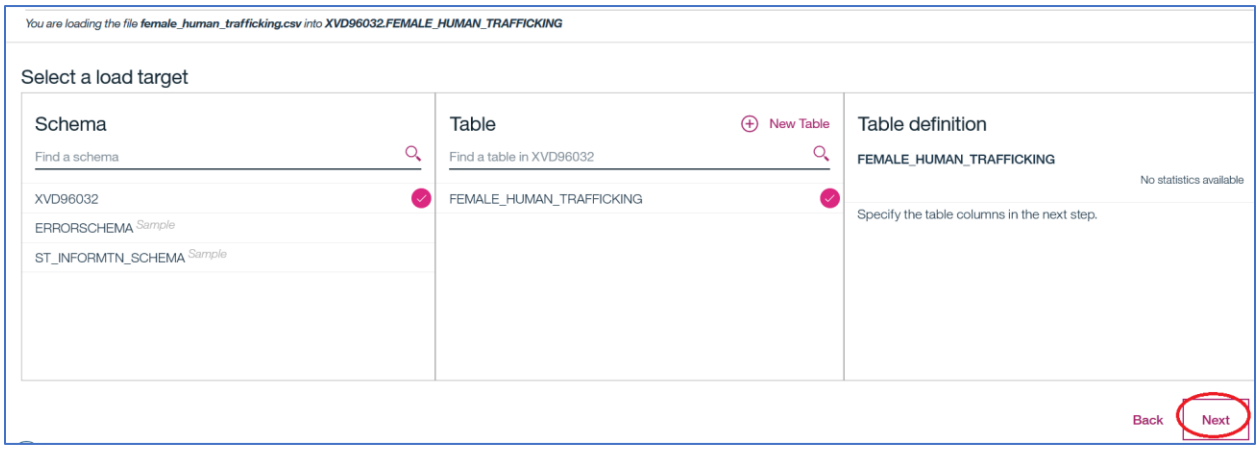
15. Click on the schema that doesn't have Sample appearing next to it. It should have 3 characters followed by a number. After clicking on the schema, click on **New Table**.



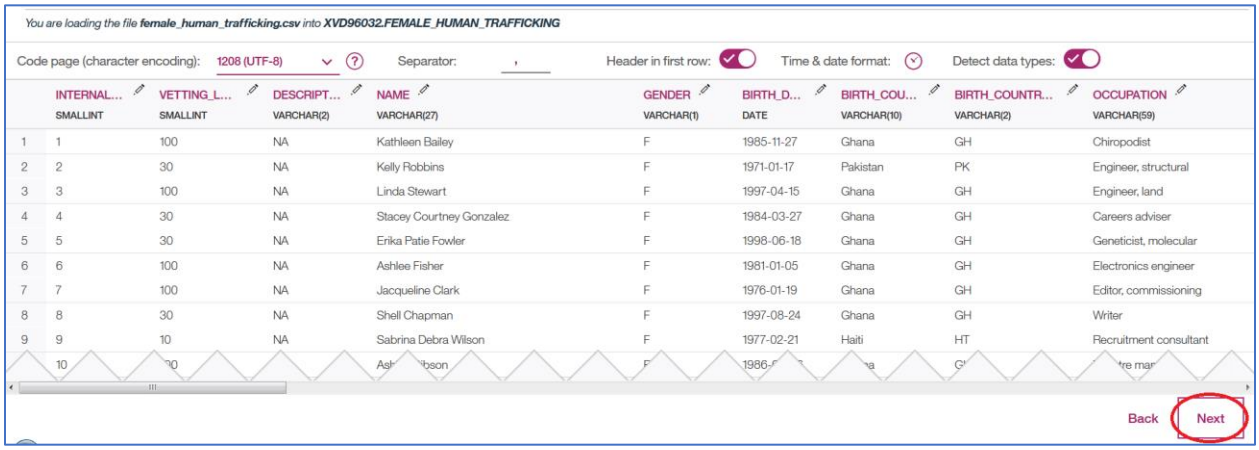
16. Enter **FEMALE_HUMAN_TRAFFICKING** for the name of the database table, and then click **Create** to create the table.



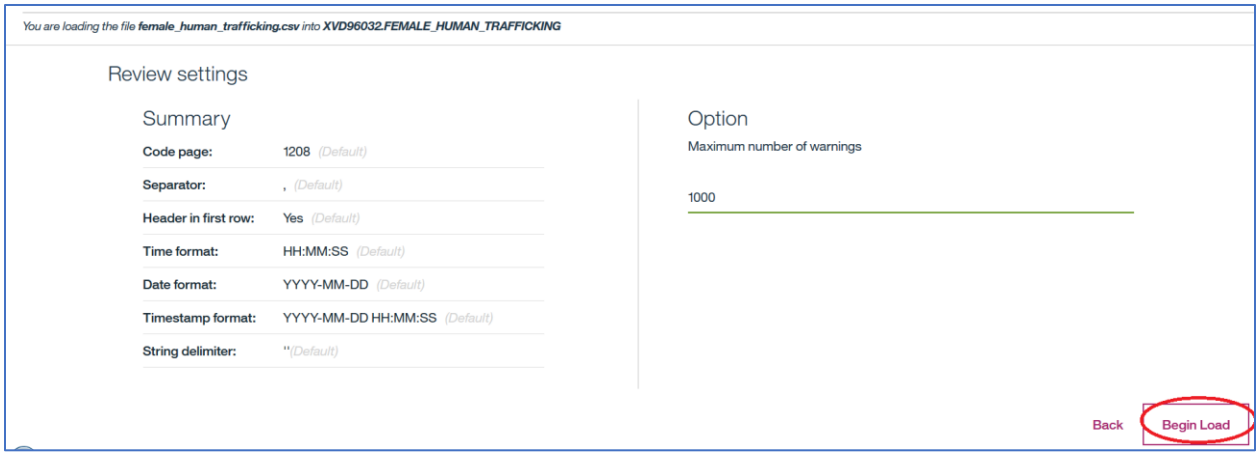
17. Click **Next**.



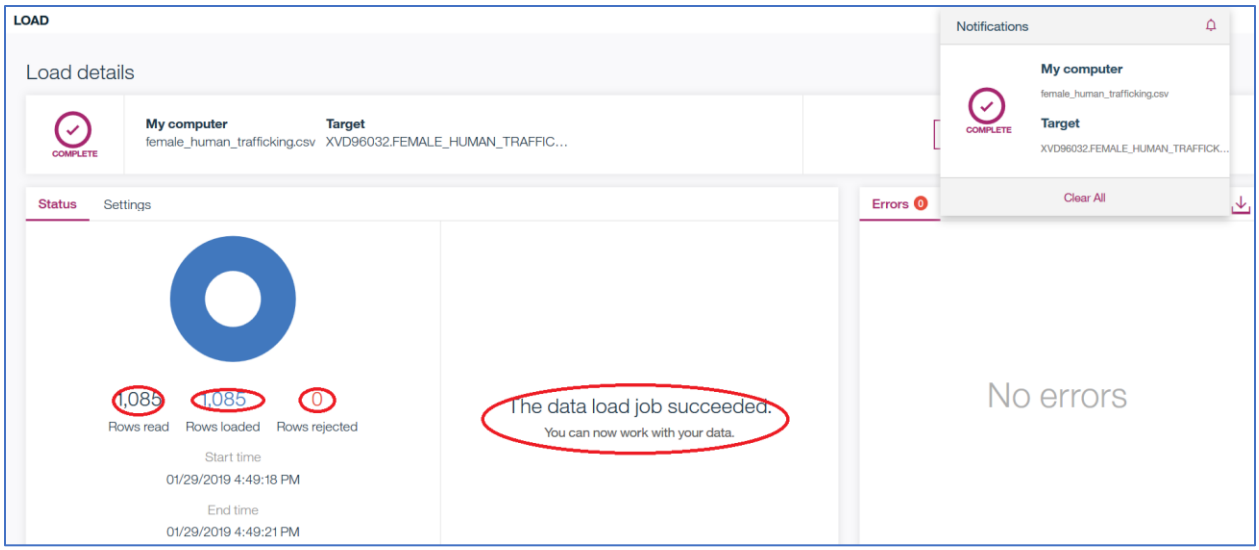
18. Click **Next**.



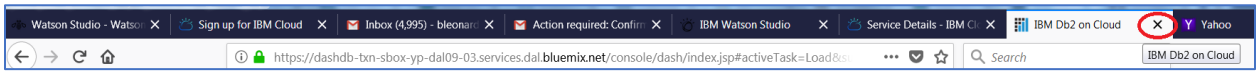
19. Click **Begin Load**.



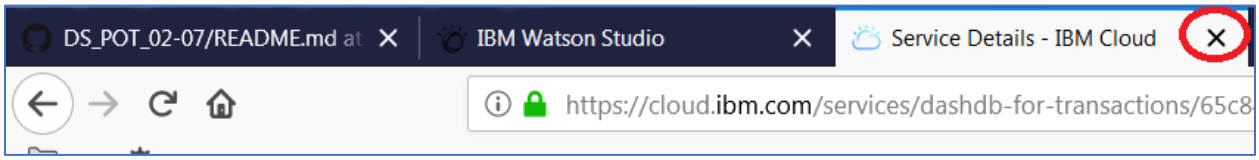
20. The 1085 records in the female_human_trafficking.csv file are successfully loaded. Click on the canvas outside of the Notifications popup.



21. Close the **DB2 on Cloud** browser tab.



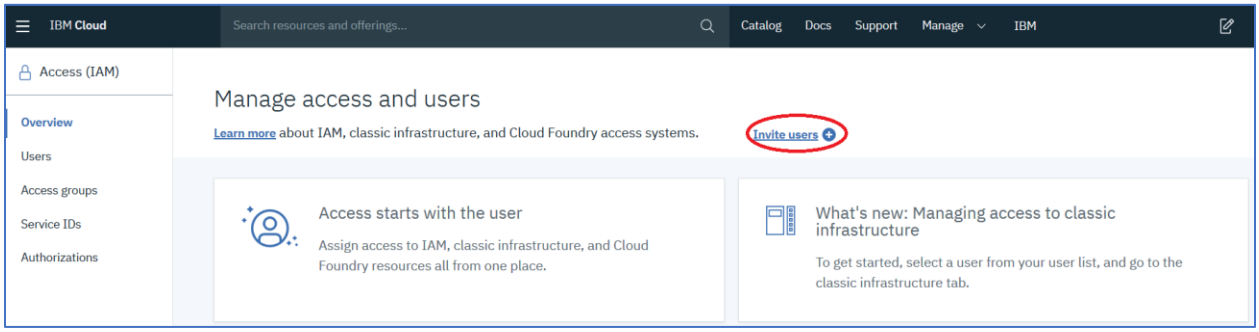
22. Close the **Service Details** browser tab.



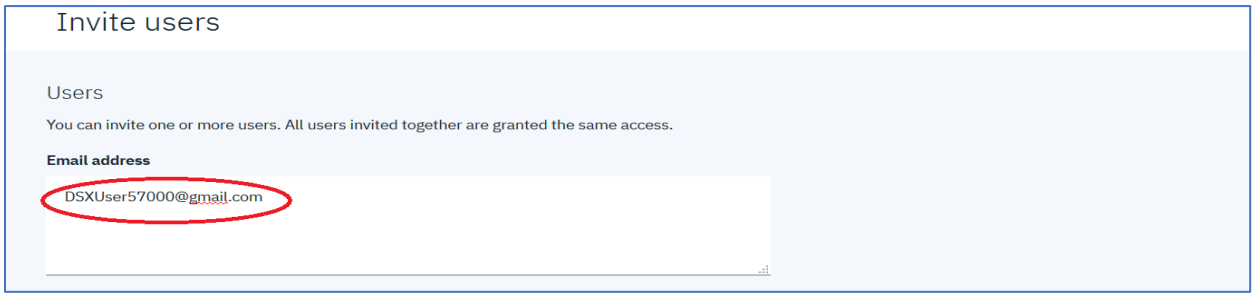
Step 3: Add a User to the Watson Studio account

To demonstrate the access control features of Watson Knowledge Catalog, add a user to your account.

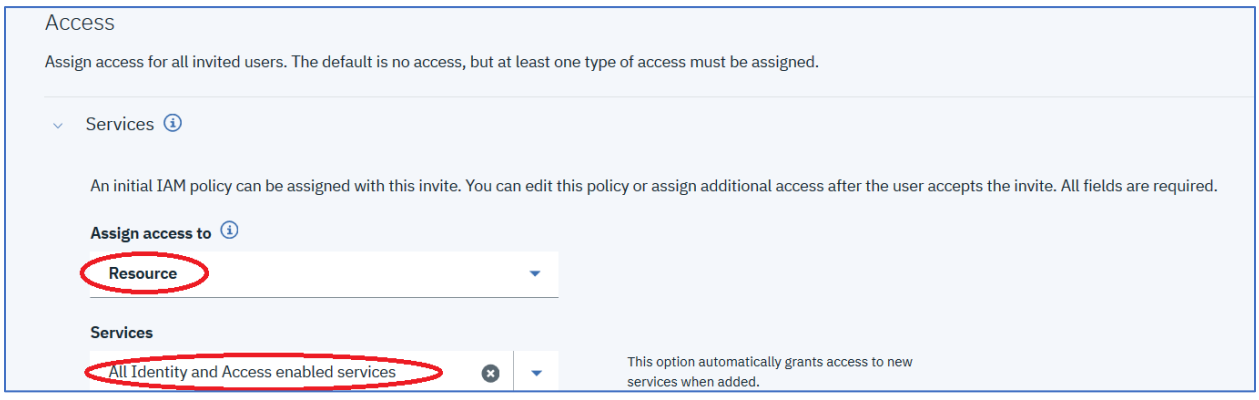
1. Click on Manage Users, and then click on **Invite users**.



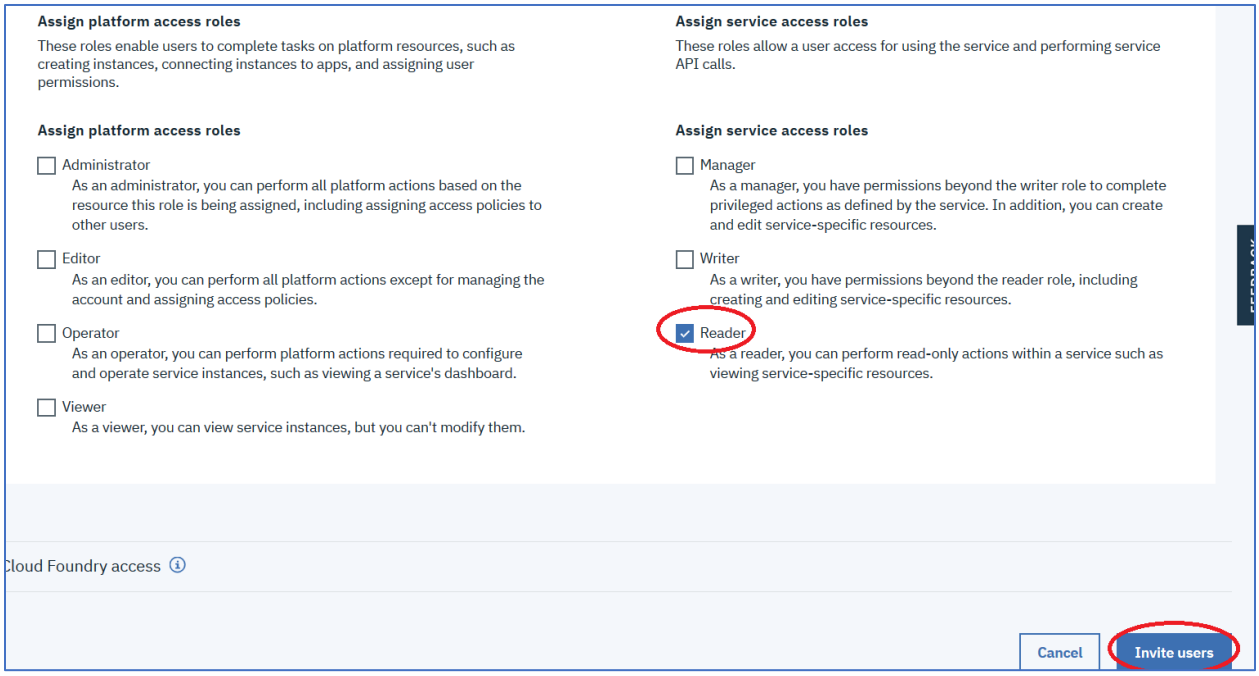
2. In the **Users** section, enter DSX57000@gmail.com for the **Email** address.



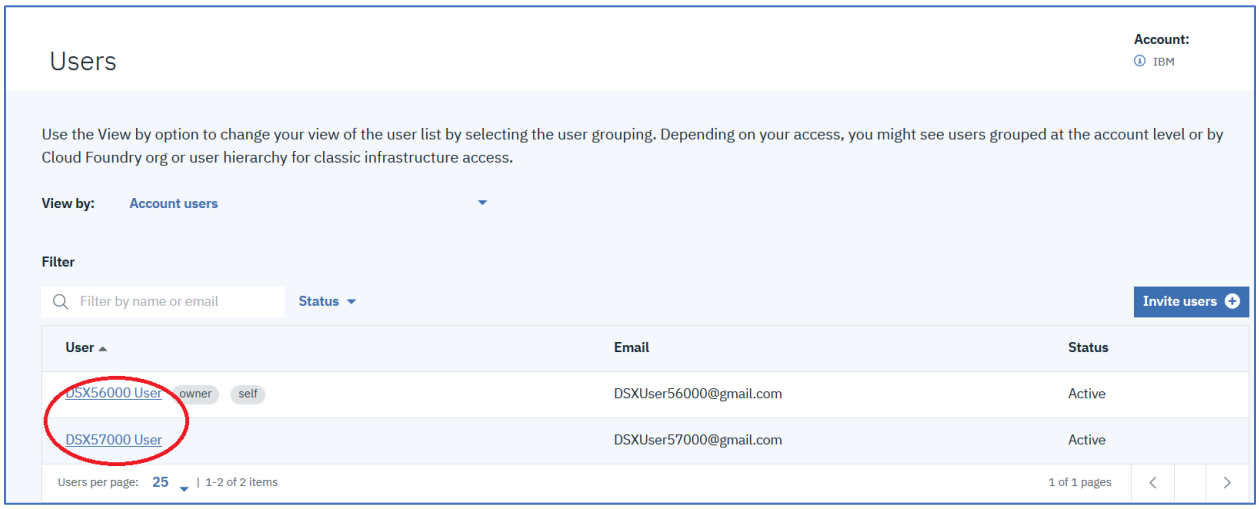
3. In the **Services** section, select Resources for the **Assign access to**, and select All Identity and Access enabled services for **Services.**



4. Scroll down and assign Reader for the **Assign service access roles**, and then click on **Invite users**.



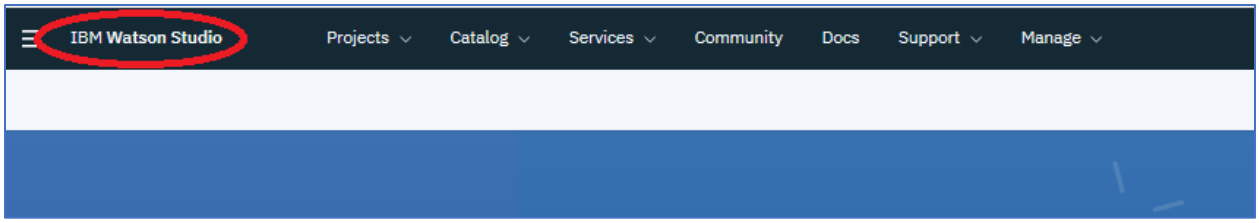
5. DSXUser57000@gmail.com is added as a user in the Watson Studio account.



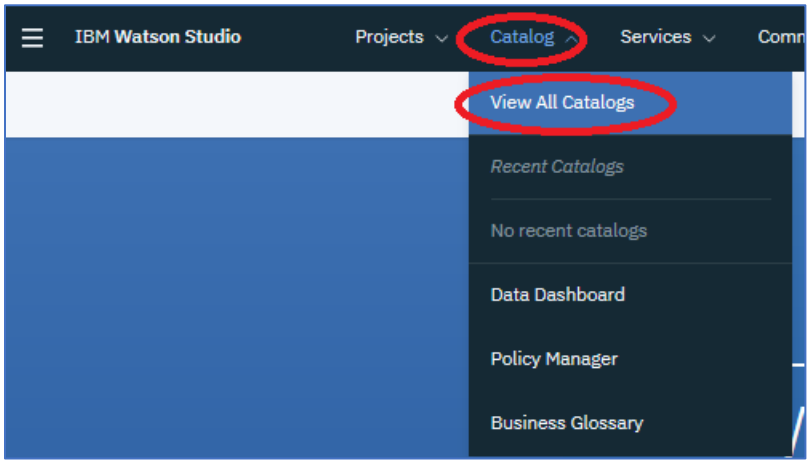
Perform Catalog Tasks

Step 1 – Create a Governed Catalog

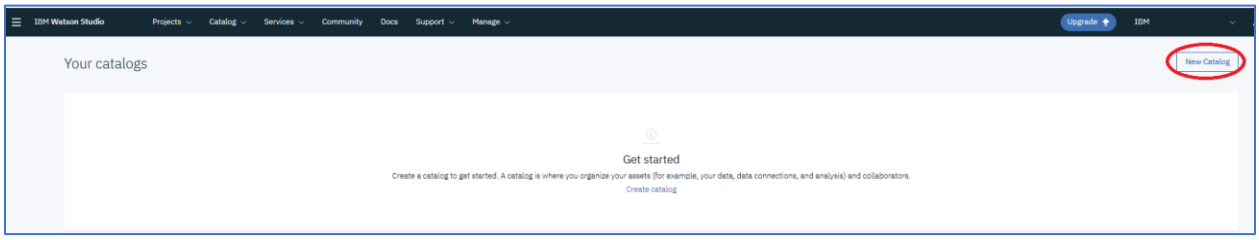
1. Click on the browser tab to return to Watson Studio.
2. Click on Watson Studio to go to the home screen.



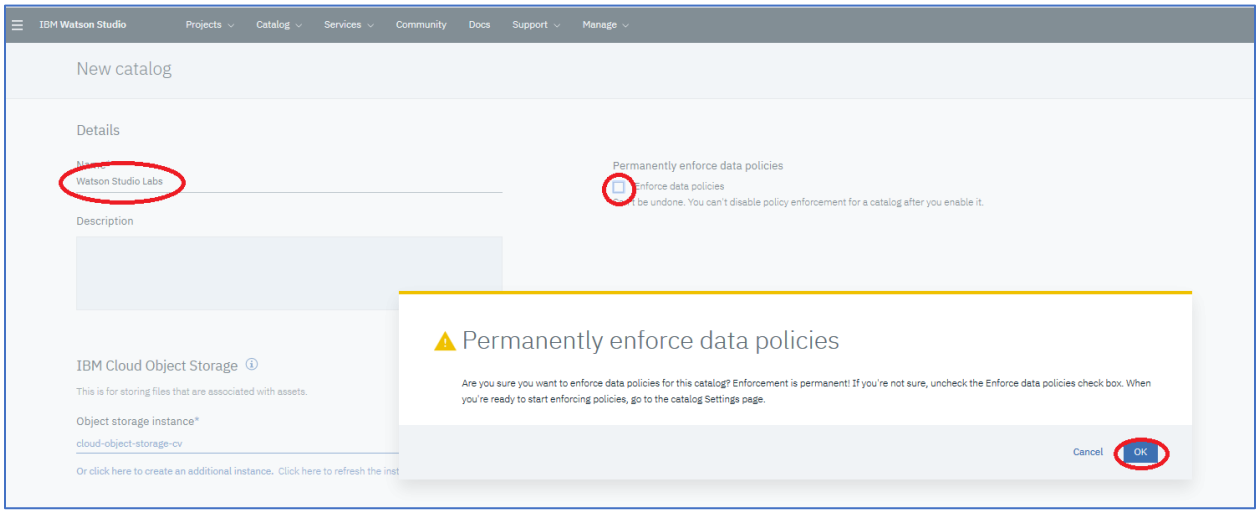
3. Click on **Catalog** and then **View All Catalogs**.



4. Click on **New Catalog**.



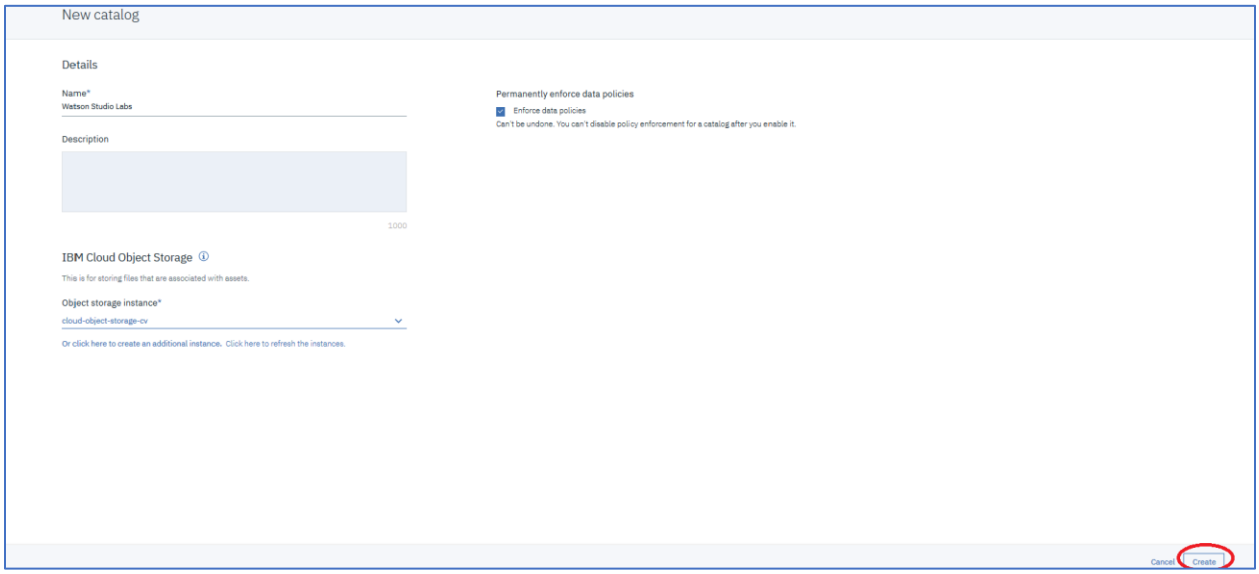
5. Enter **Watson Studio Labs** for the **Catalog** name. Click on **Permanently enforce data policies.** You will get a pop-up as shown below. Click on **OK**.



Note: By default, access to data assets in a catalog are 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 policies** enables the enforcement of data policy rules to allow or deny access to a data asset or mask, anonymize and redact data at the data asset field 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.

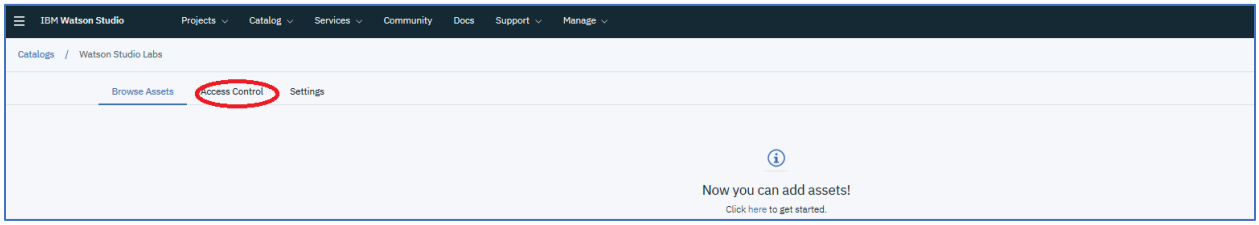
6. Click **Create** to create the catalog.



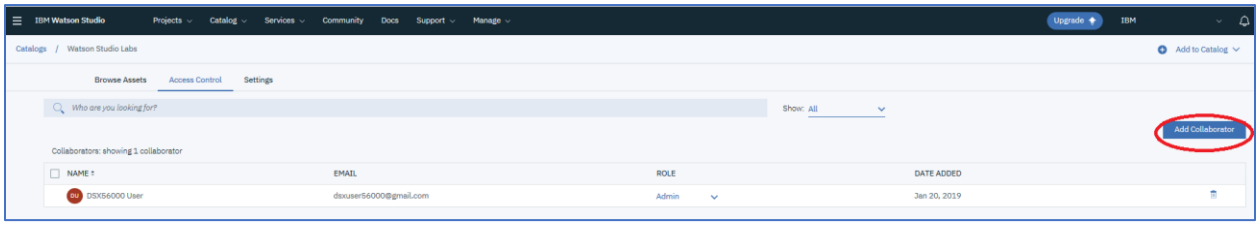
Step 2: Add a Member to the Catalog

To demonstrate the Catalog access control capability, we will add the username DSXUser57000@gmail.com as a member of the catalog. Note that the first name of this user is "DSX57000" and the last name of this user is "User".

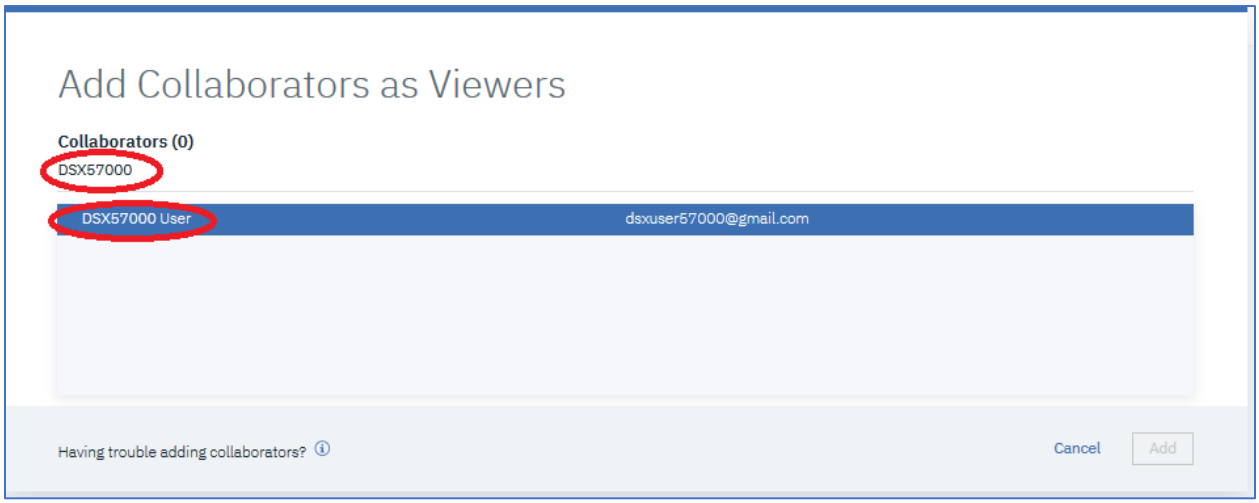
1. Click on **Access Control**.



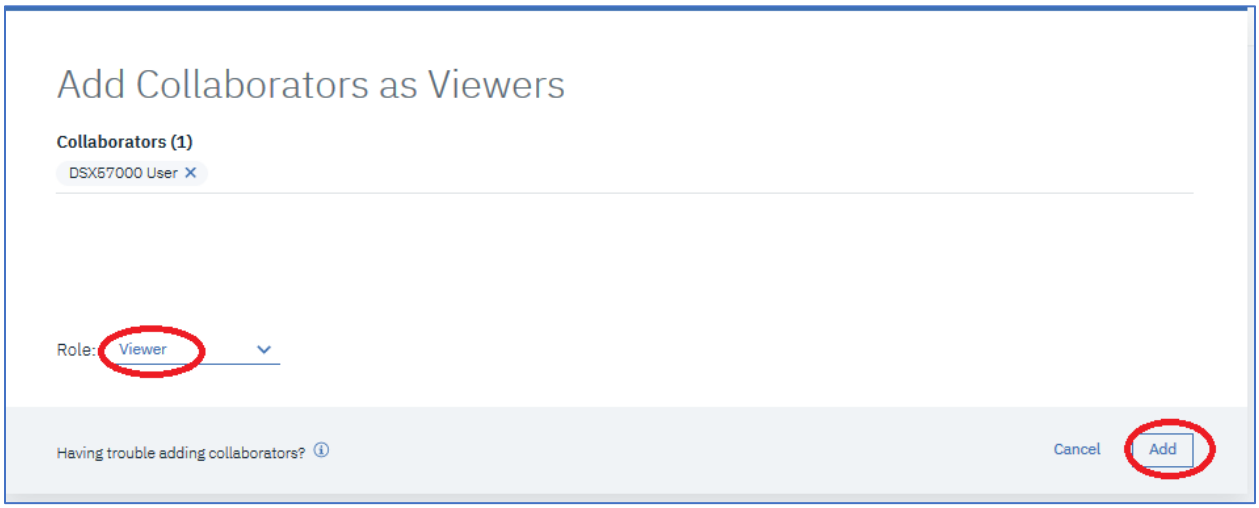
2. Click on **Add Collaborator**.



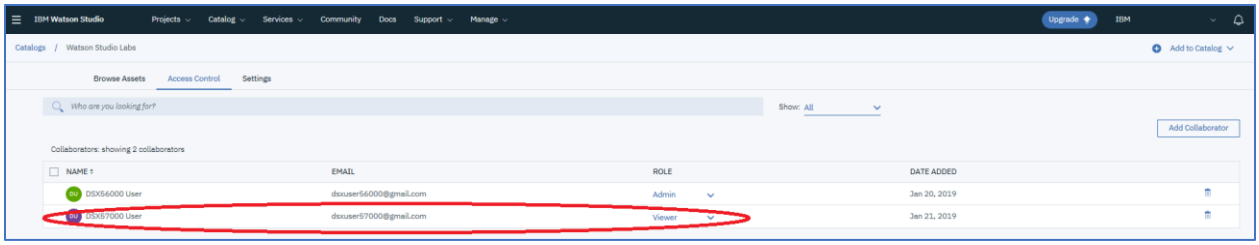
3. Enter DSX57000 (first name of the DSXUser57000@gmail.com user) as the **Collaborator**. DSX57000 should show up in the list box. Click on DSX57000 User.



4. Leave the Viewer as the **Role** and click on **Add**.



5. "DSX57000 User" is added to the catalog.



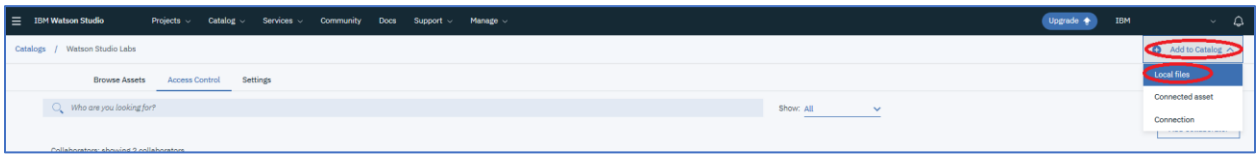
Step 3: Add Data Assets to the Catalog

In this step, we will add two local files, a connection, and a connected data asset. The local files we will add are the "Occupation.csv" file, and the "categories.csv" file. We will set up a connection to the DB2 on Cloud database service that we created earlier (this is where we need the credentials). We will add a connected data asset that will point to the FEMALE_HUMAN_TRAFFICKING table.

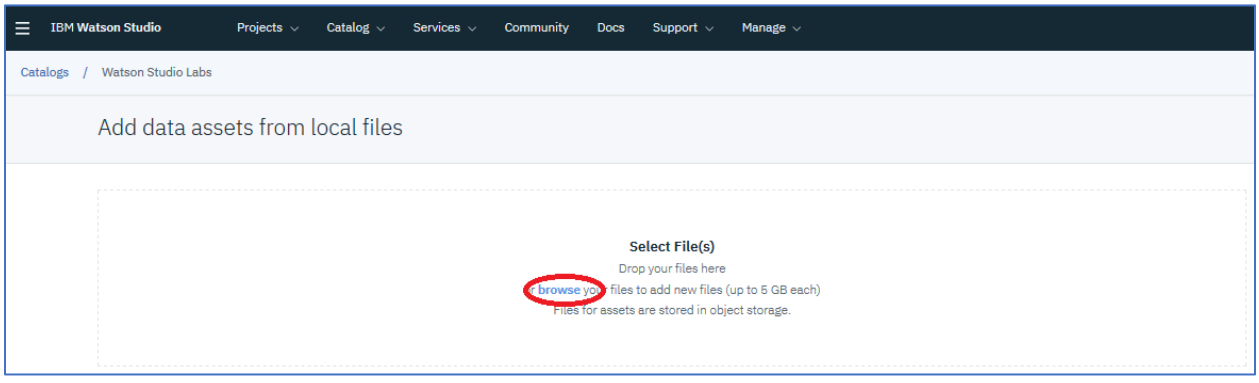
This step will introduce you to the three methods available to discover and catalog data assets; **Local files, Connected asset and Connection**. You will use these methods to catalog data assets into the newly created Knowledge Catalog and then tag them for users to easily find them, understand their content and make them available throughout IBM Watson Studio, for use during data preparation and within models, dashboards and notebooks as we will see in our lab exercises.

Add Local Files

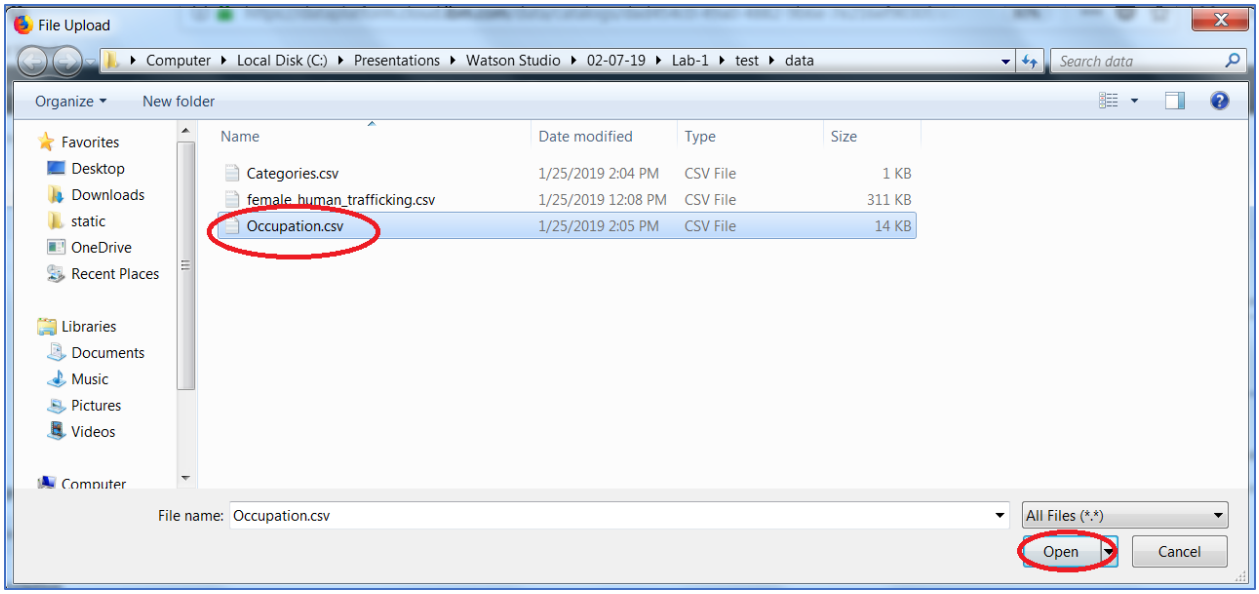
1. Click on **Add to Catalog** and click on **Local Files**.



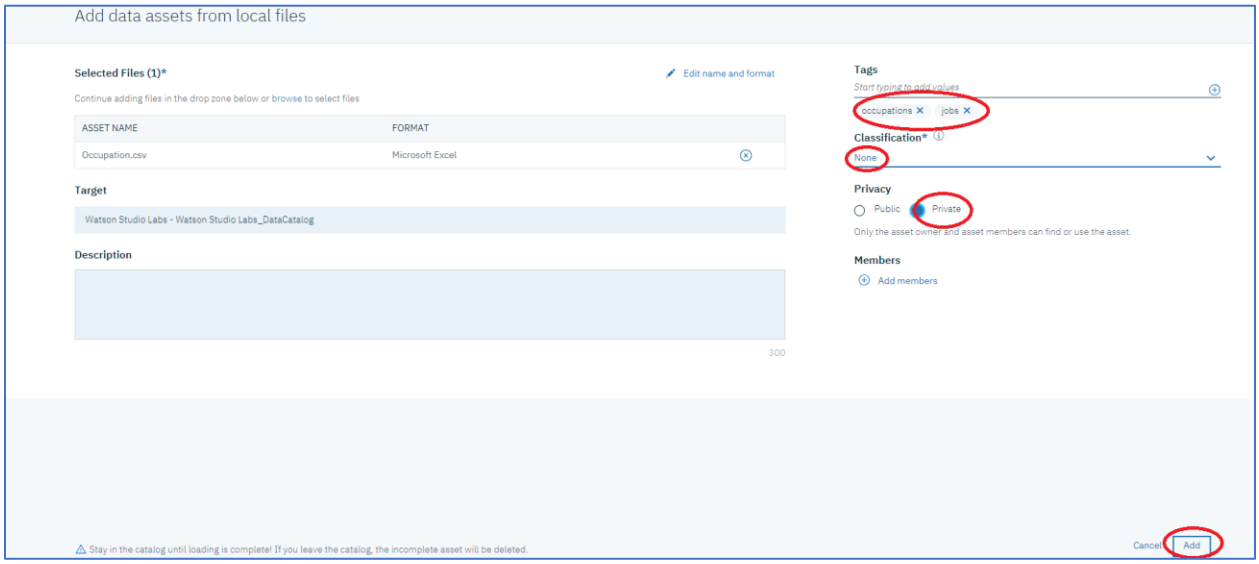
2. Click on **browse** to navigate to the folder where the "Occupation.csv" file is stored.



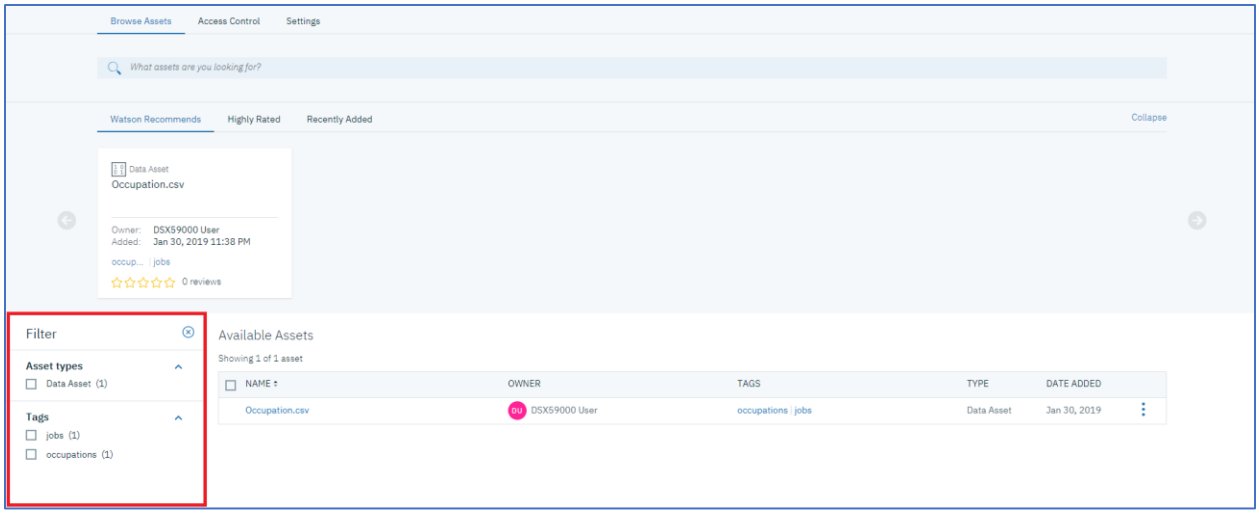
3. Click on **Occupation.csv** and then click **Open**.



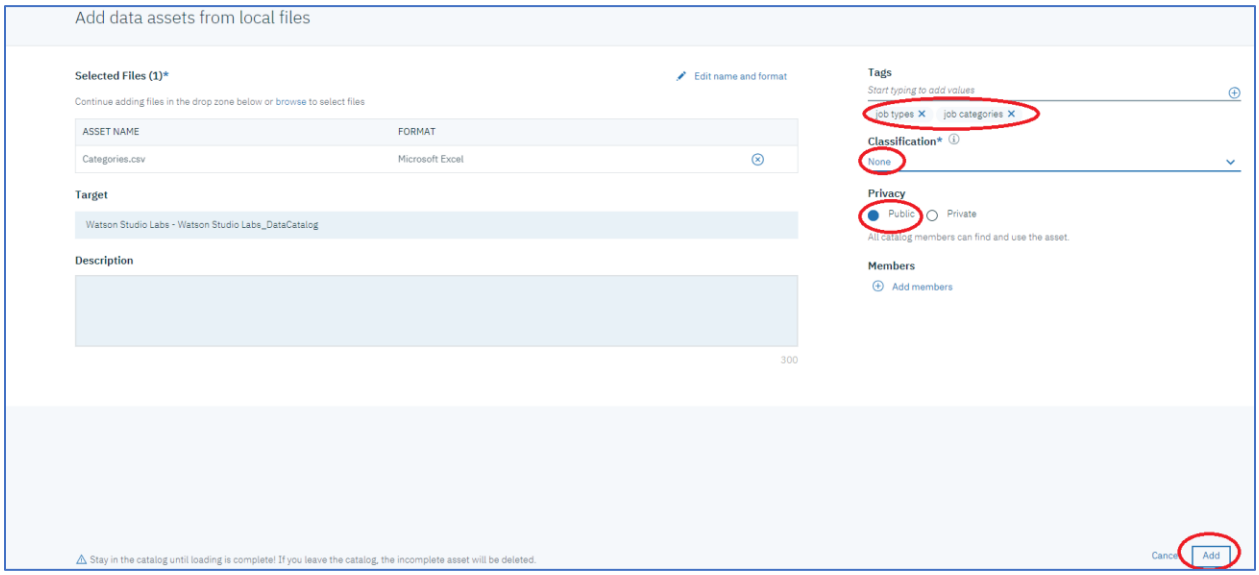
4. We can add tags to the newly added data asset to more easily search for it. Add two tags, enter “occupations”, click on the + icon, and enter “jobs” and click on the + icon. Leave the **Classification** as None. Click on Private for the **Privacy.** Click on **Add**.



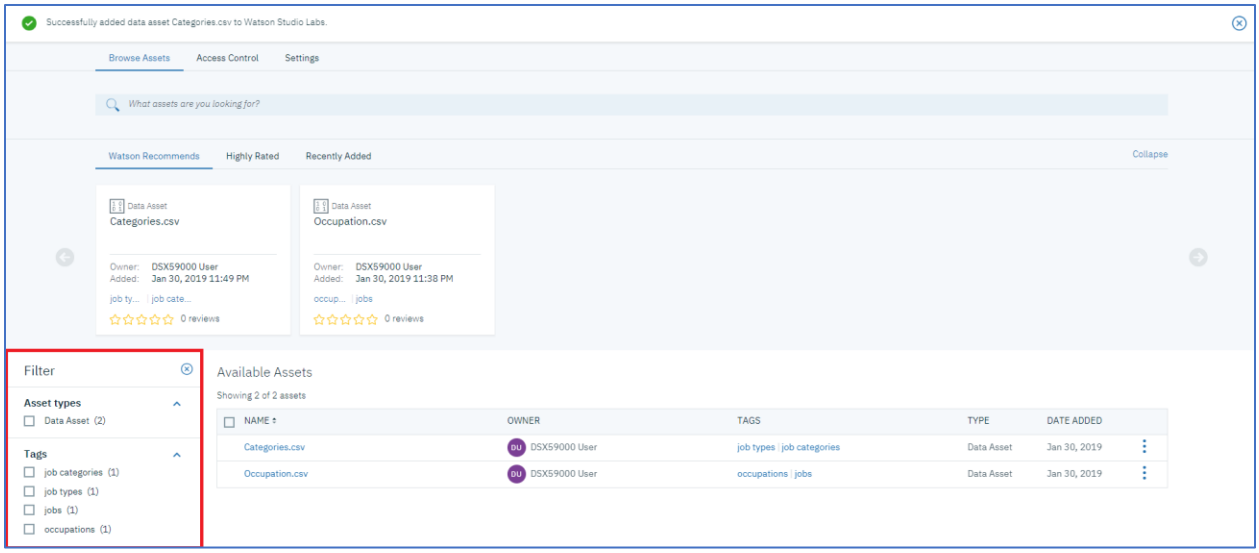
5. The Occupation.csv asset metadata is displayed. Note a number of features on the panel. At the top are 3 tabs: **Watson Recommended**, **Highly Rated**, and **Recently added**. The square below contains assets that are displayed when one of these tabs are selected. On the left side in the **Filter** area are shown filters by **Asset Type**, and **Tags**. As we add data assets, the tags associated with those assets will be displayed, and an additional asset type will be recorded. The number next to each tag entry indicates how many assets contains the specific tag. The number next to the Asset Type entry indicates how many assets exist for the specific type.



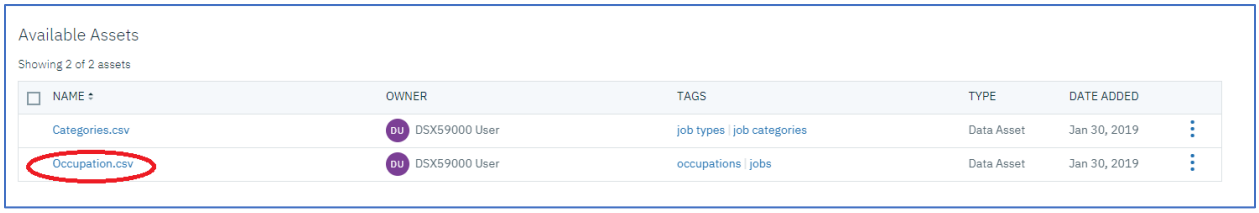
6. Follow the same procedure (1-4) above to add the local file "categories.csv" file to the catalog. For step 4, enter the tag "job types" and then click on the "+" icon, enter the tag "job categories" and then click on the "+" icon, leave the **Classification** as None, leave the **Privacy** as Public, and then click on **Add**.



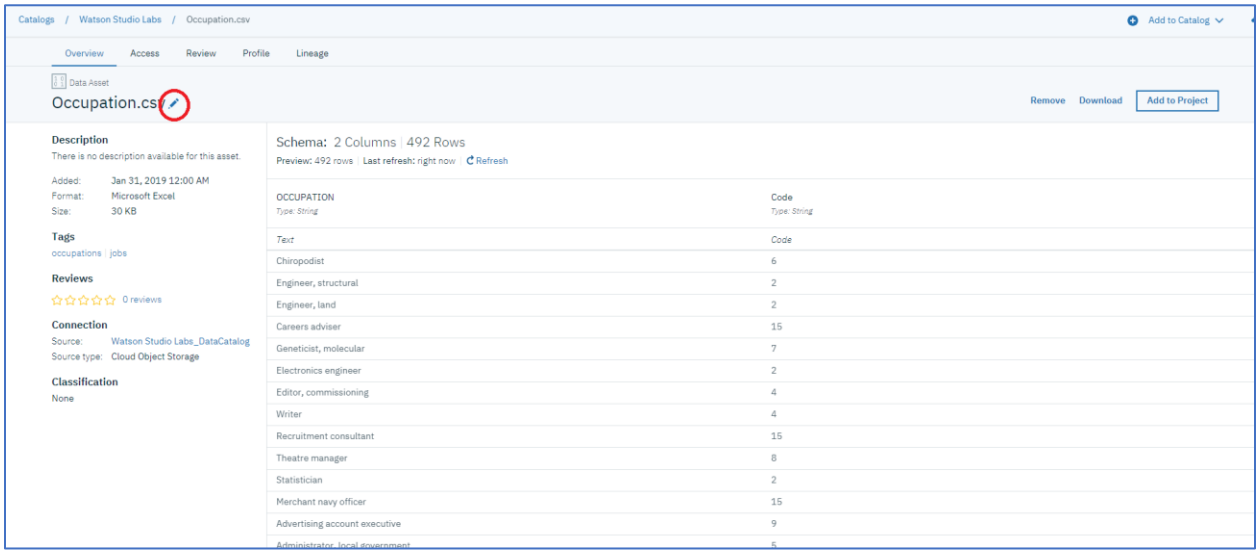
7. The categories.csv asset is displayed. Note the **filter** area information has been updated to reflect the addition of this asset.



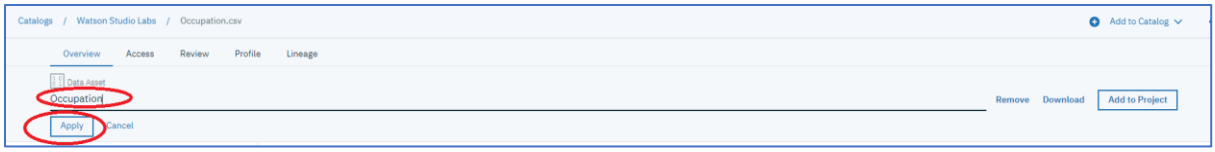
8. Change the name of the Occupations.csv asset to be just Occupation by clicking on the **Occupation.csv** asset.



9. Hover the mouse on the **Occupation.csv**. Click the “pencil” icon when it appears.



10. Change the name to Occupation and click **Apply.**

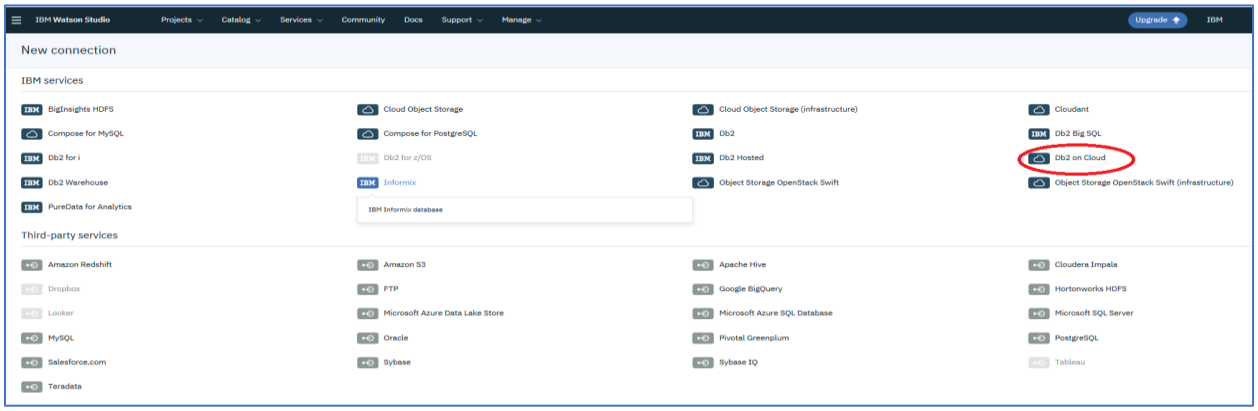


11. Click on **Watson Studio Labs** to return to the catalog display.
12. Follow steps 8-11 to change the Categories.csv asset name to be Categories.

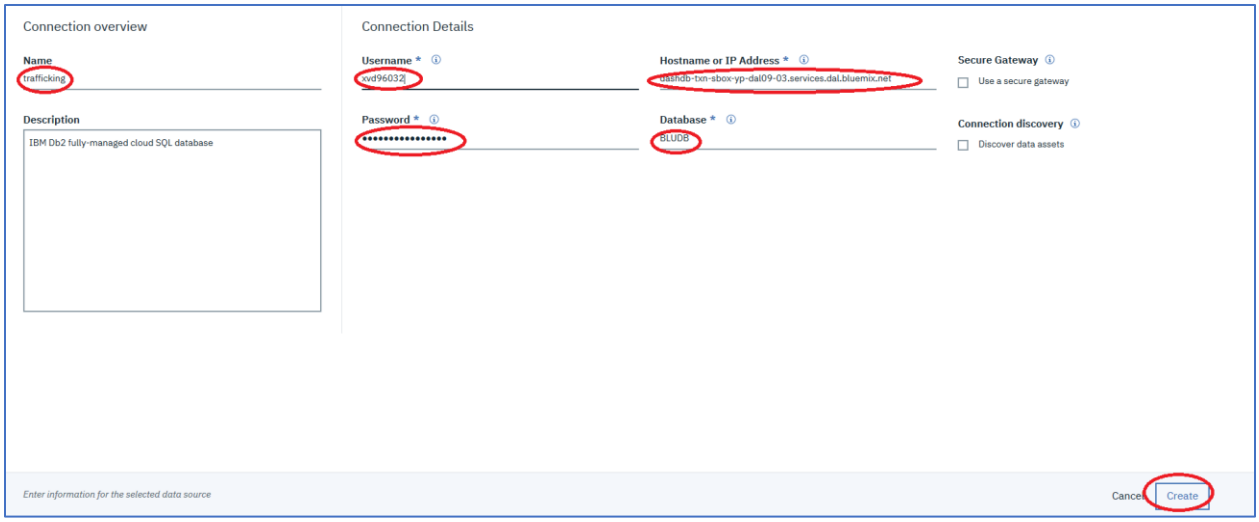
Add Connection

1. Click on **Add to Catalog**, and then click on **Connection**.

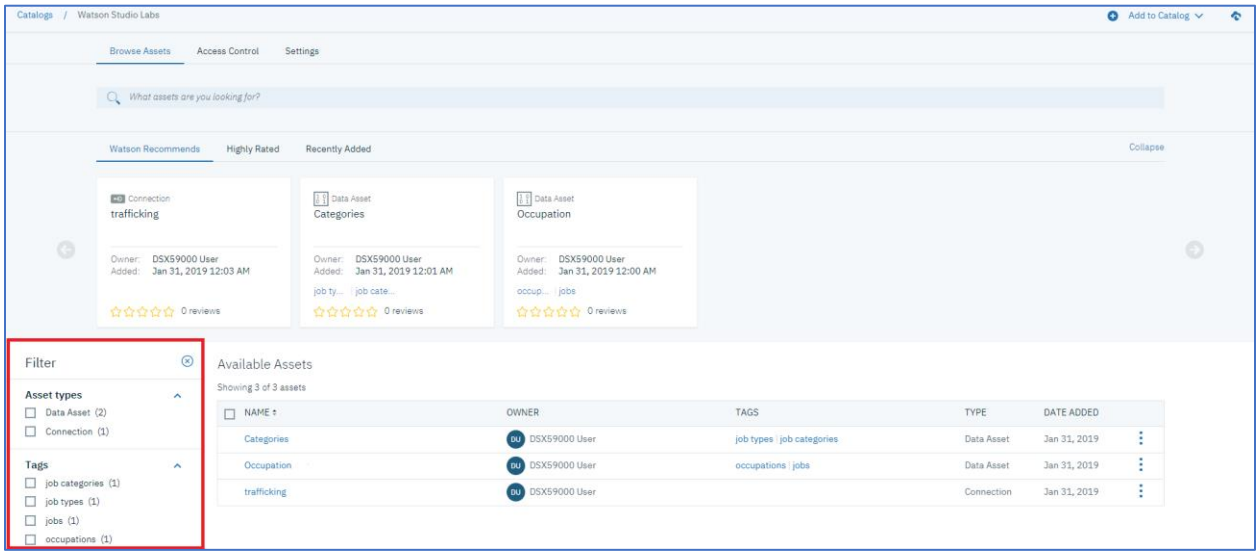
2. Click on **DB2 on Cloud**.



3. Enter “trafficking” for the **Name** of the connection. Cut and paste the credentials Username, Host, Password, and Database that you saved earlier into the corresponding fields. Click on **Create**.



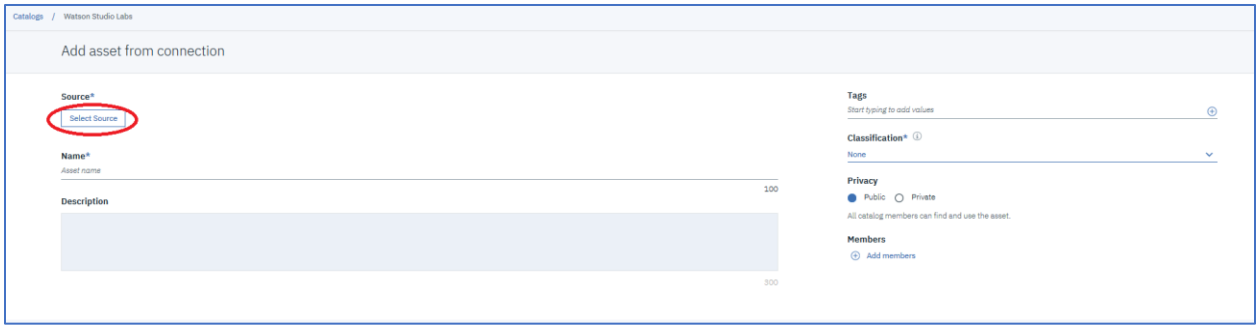
4. The connection asset is displayed. Note that a new Asset Type (Connection) is added in the **Filter-Asset Type** area, as well as in the **Type** description for the data asset. Also, note an untagged entry has been created in the **Filter-Tags** area.



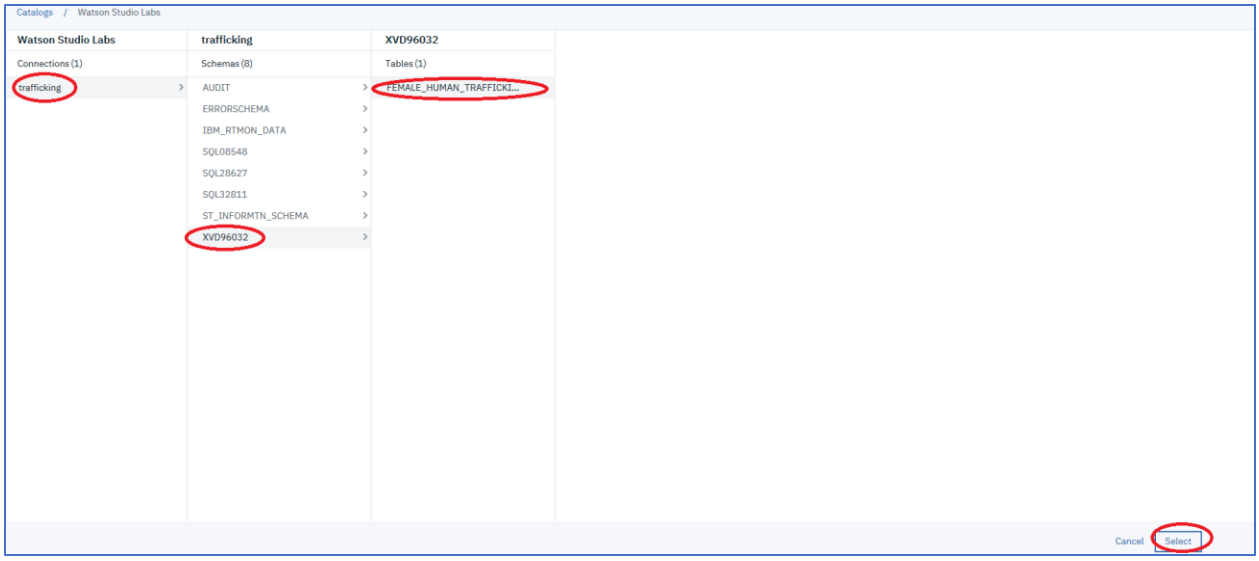
Add Connected Data

1. We have added a connection to the DB2 on Cloud database service. We will now add a Connected Data Asset to point to the FEMALE_HUMAN_TRAFFICKING table that was created earlier. Click on **Add to Catalog and click on Connected asset**.

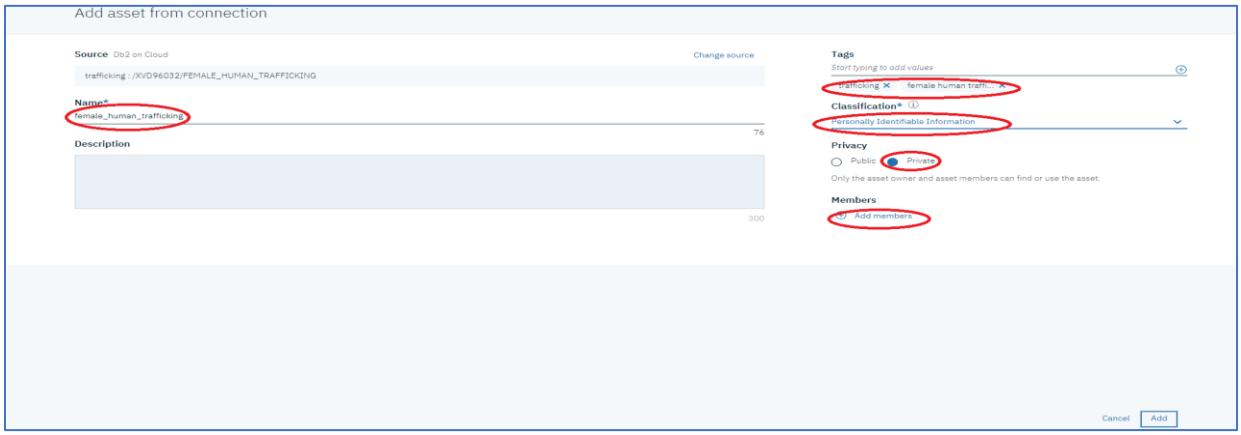
2. Click on **Select Source**.



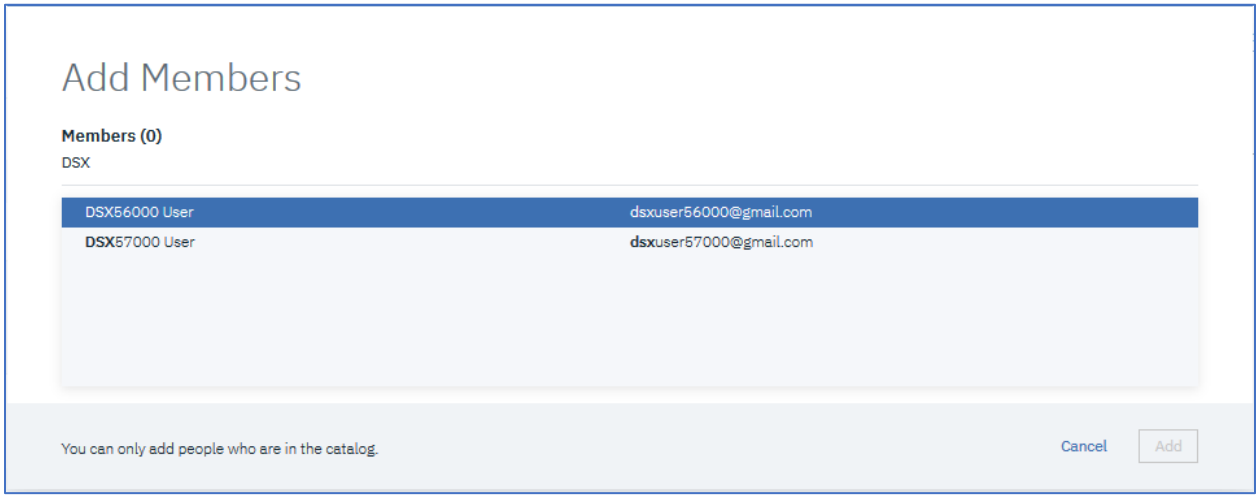
3. Click on the **trafficking** connection, click on the **Schema name**, click on FEMALE_HUMAN_TRAFFICKING, and then click **Select**.



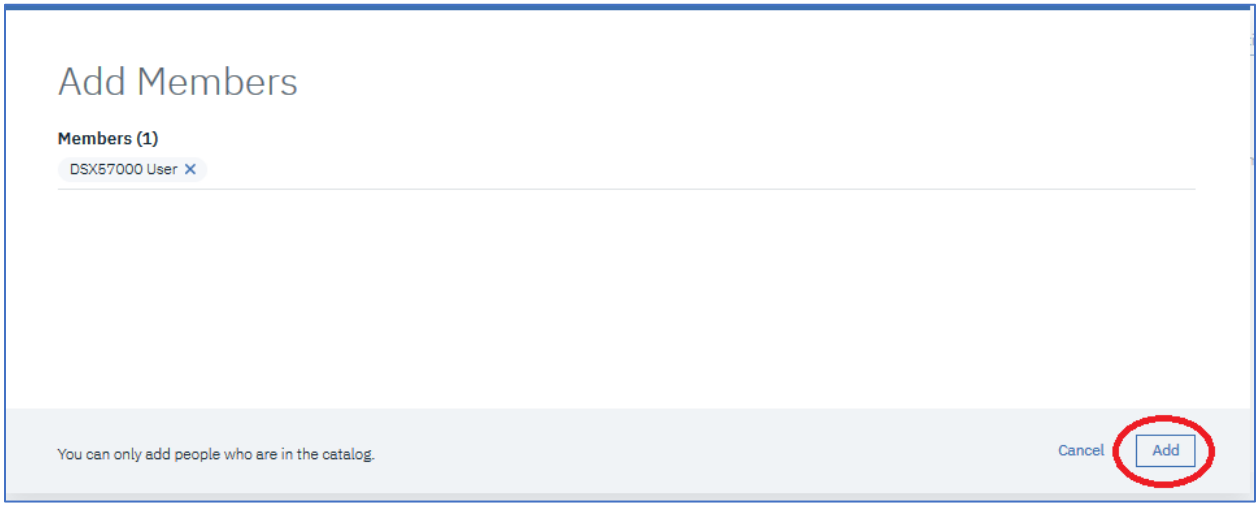
4. Enter female_human_trafficking for **Name**, enter "trafficking" for **Tags** , click the "+" icon, enter "female human trafficking for **Tags**, click the "+" icon. Select "Personally Identifiable Information" for **Classification**, click on **Private**, for Privacy, and click on **Add members**. The privacy setting of "Private" specifies that only catalog members on the access control list for this data asset will have access to the data asset.



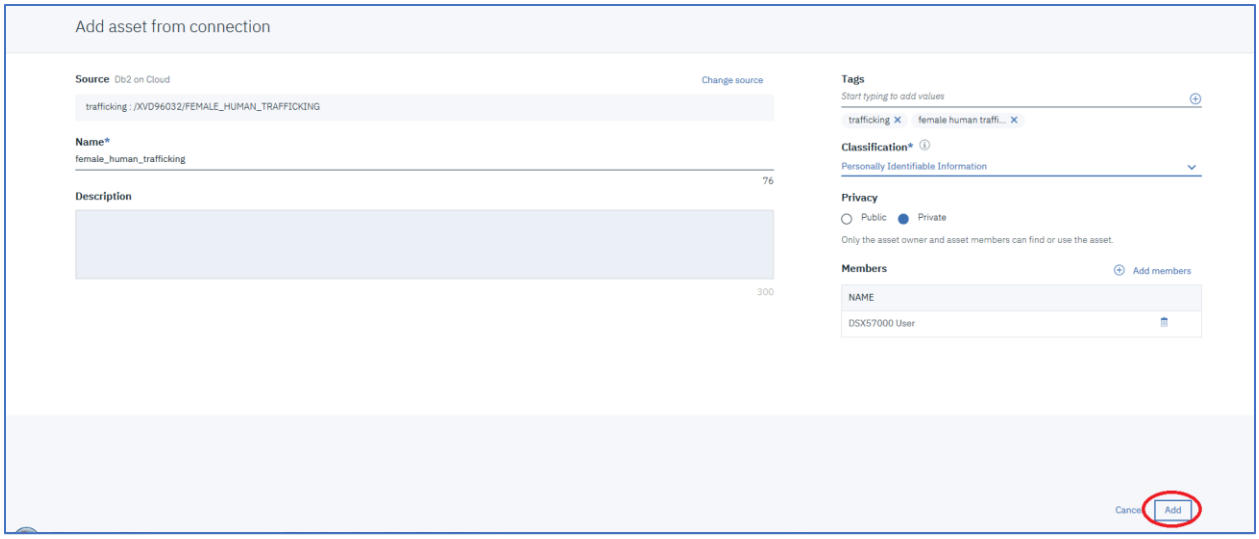
5. Enter DSX in the **Members** field. The list of matching users is displayed in the member list. Select the DSX57000 user.



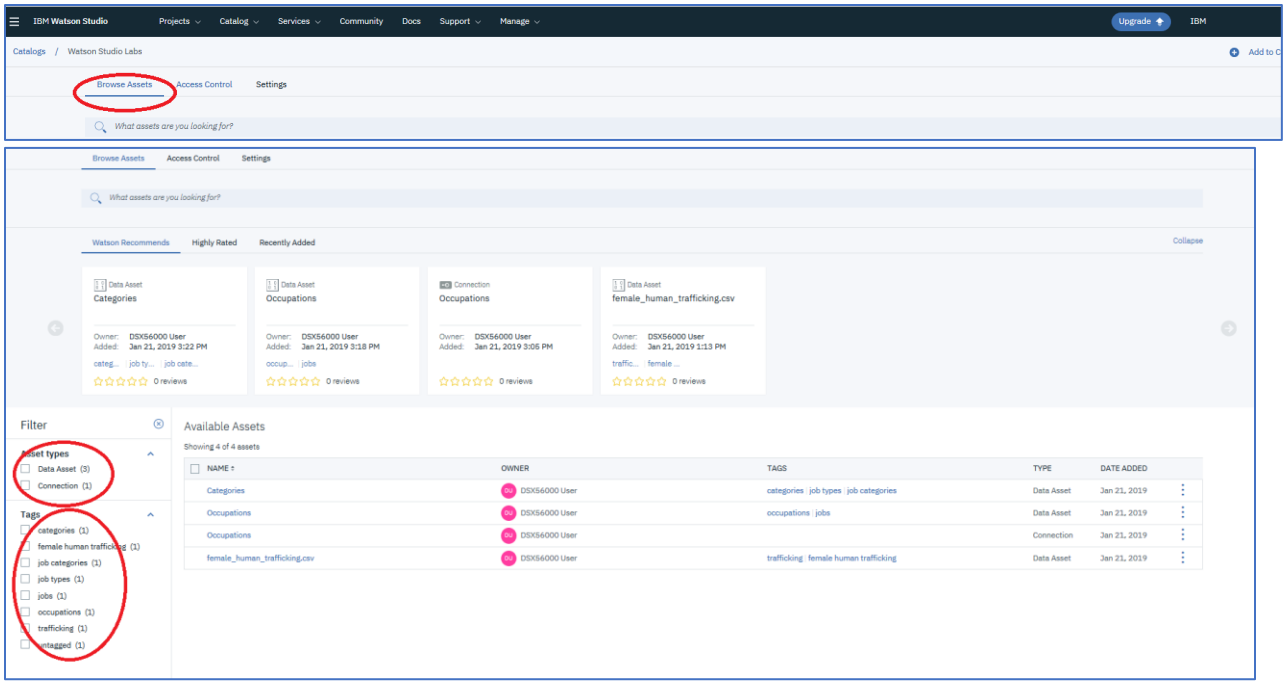
6. Click on **Add** to add DSX5700 User to the access control list for the female_human_trafficking asset.



7. Click on **Add**.



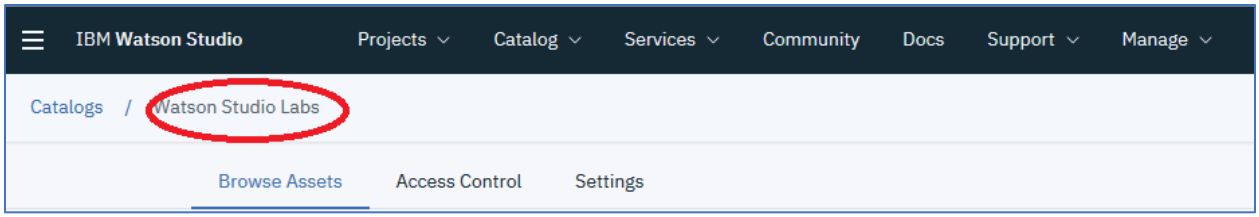
8. All of the assets have now been added to the catalog. The catalog display should appear as below.



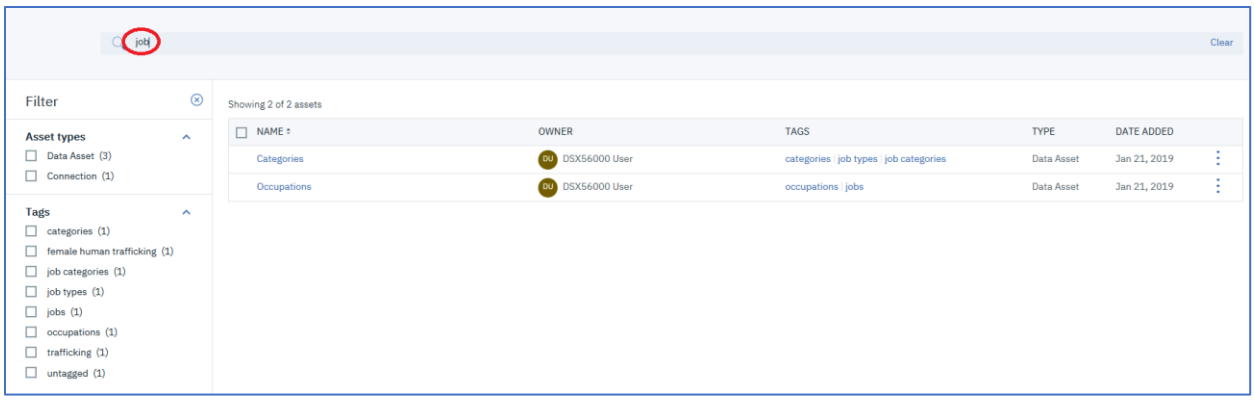
Step 4: Search the Catalog

Several options exist for searching the assets in the catalog. We can enter text in the search input text field and the system will match against asset names and tags. We can then further filter the search results by selecting the Asset Type, and/or the Tags. Or we can directly select the Asset Type or Tags without entering anything in the search text field.

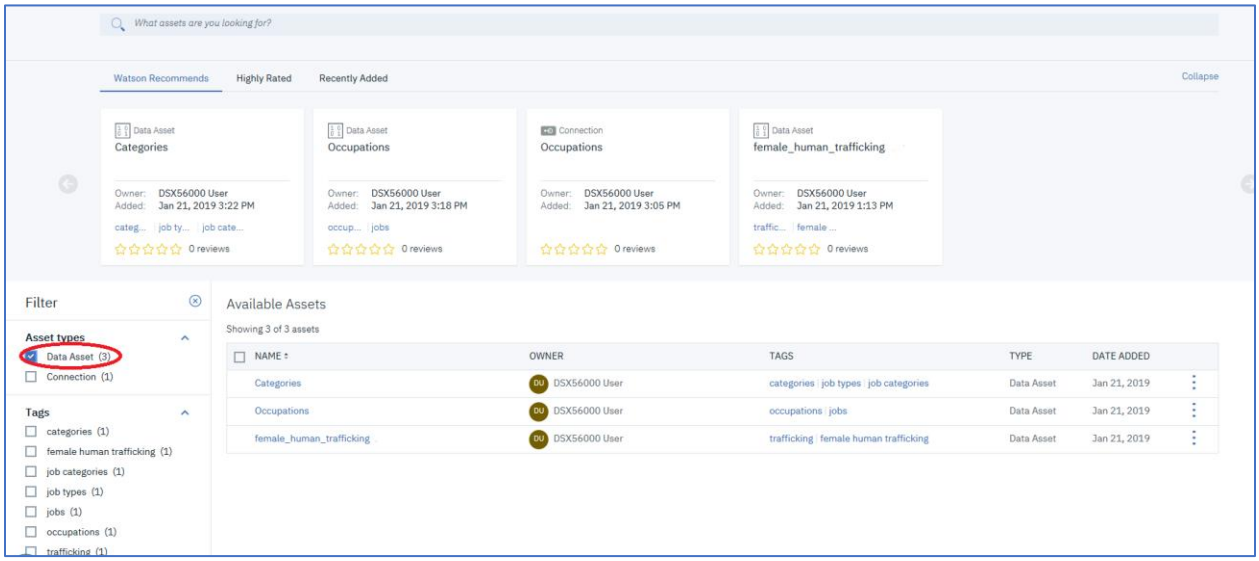
1. Click on **Watson Studio Labs** to return to the Catalog home screen.



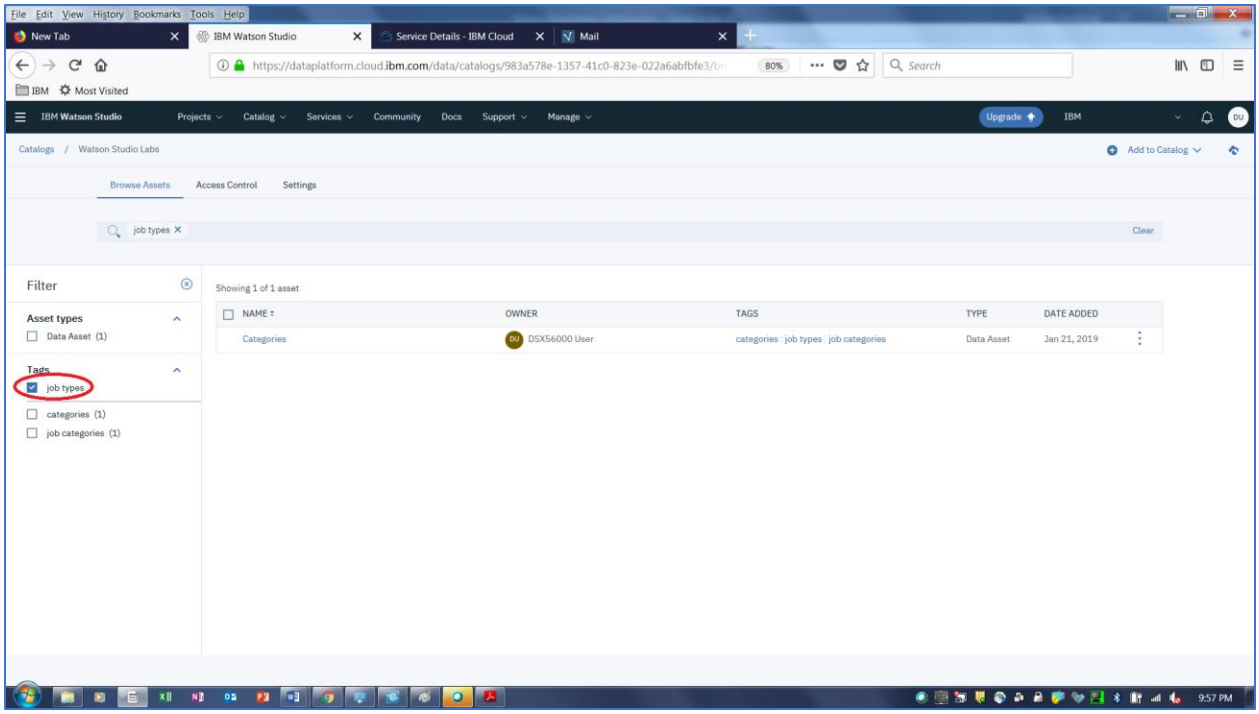
2. Enter “job” in the **search text field**. Data assets are displayed that contain the word or letters **job** in their Name, Description, or tags.



3. Clear the **search text field.** Click on the check box next to **Data Asset** in the Filter area. The search results list the 3 Data Assets. The 1 Connected asset, and the 2 uploaded files. Note: Other asset types that can be cataloged include, Connections, Models, Notebooks, and Dashboards.



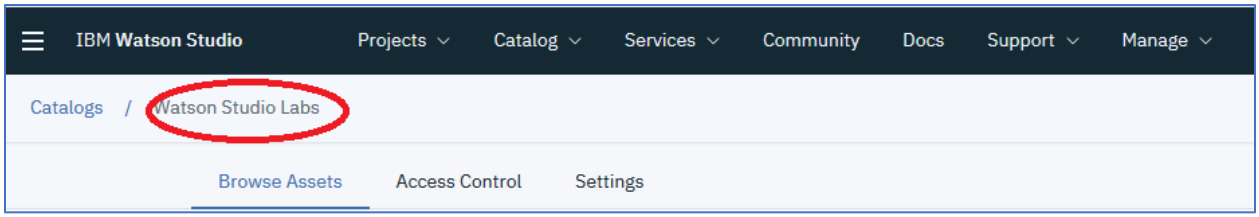
4. Click on the Tag **job types** in the Filter area. The Categories Data Asset is listed. Note that the Filter area has been updated to only list Asset Types, and Tags that correspond to the search result.



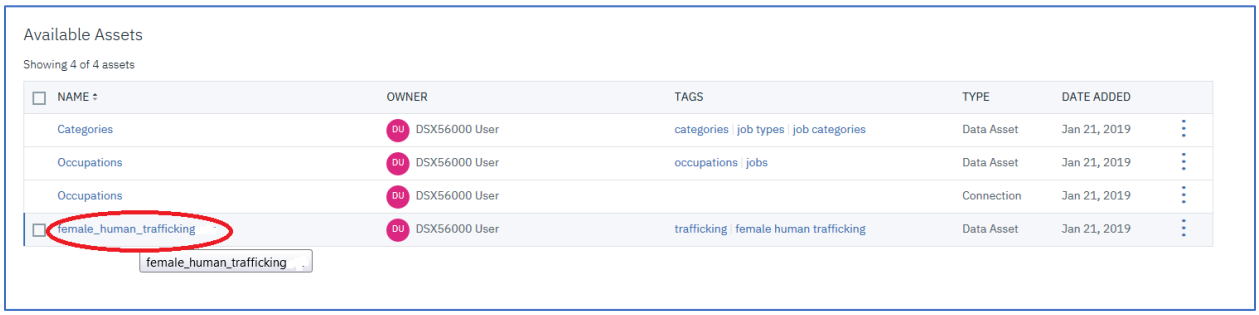
Step 5: Edit/Review/Profile Data Asset

In this step, we will drill down on the female_human_trafficking data asset to show how to edit the metadata describing the asset, to provide a rating and review of the asset, and to profile the asset.

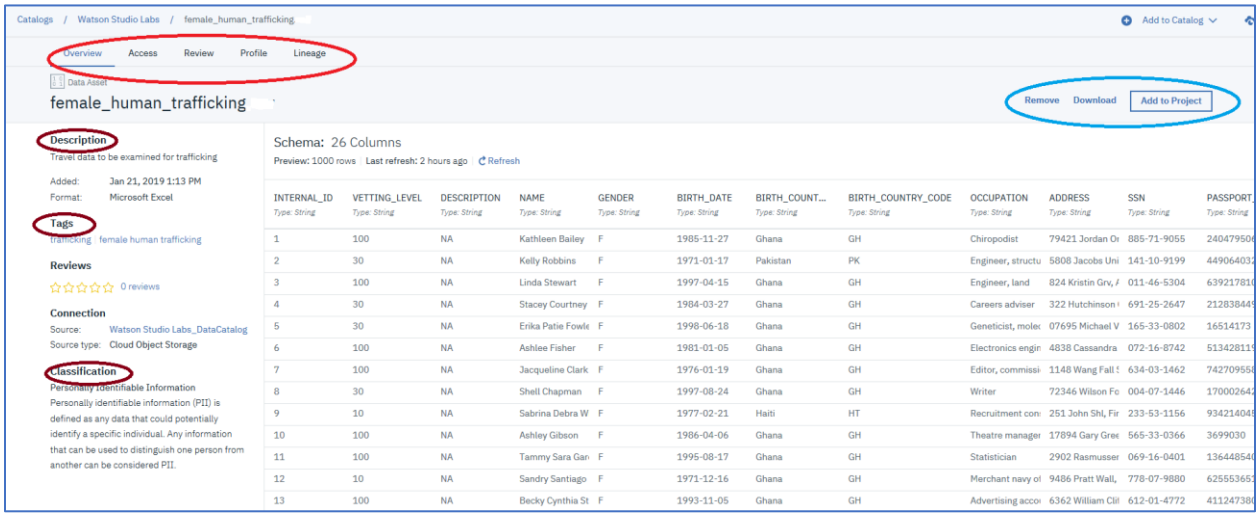
1. Click on Watson Studio Labs to return to the Catalog home screen.



2. Click on the female_human_trafficking asset.

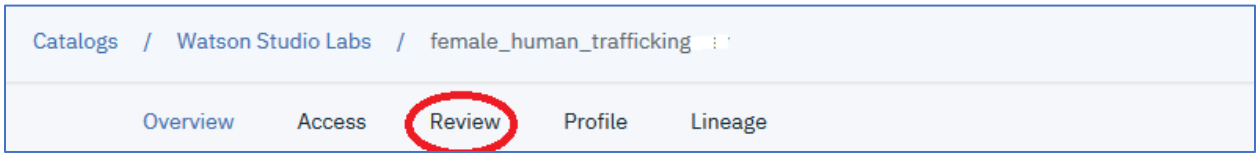


3. The metadata and data contents of the asset are displayed as shown below. Toward the top of the screen, are 5 tabs, circled in red below. The panel shown below corresponds to the **Overview** tab. The **Access** tab allows the user to edit the access control list for this asset. The **Review** tab enables the user to provide a rating and descriptive review of the asset. These social features aid in searching for relevant and useful assets. The **Profile** tab invokes an auto-profiling capability where the columns of the data asset are classified based on machine learning scoring into one of a number of classifiers that are built into the platform. This capability will be shown below. Finally, the **Lineage** tab tracks the changes made to the data asset over its lifecycle. This capability is not available in the lite catalog offering so will not be exercised in this lab.

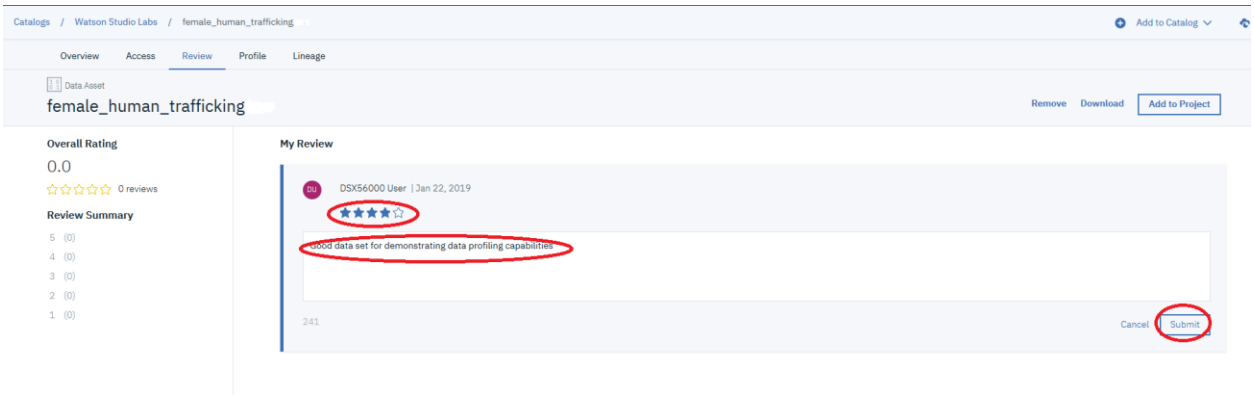


The actions circled in blue will **Remove** the asset from the catalog, **Download** the asset to the local file system, or add the asset to a catalog enable project (**Add to Project)**. Hovering over the name of the data asset, the **Description**, the **Tags**, or the **Classification** will popup and Edit icon that allows the user to edit these entities.

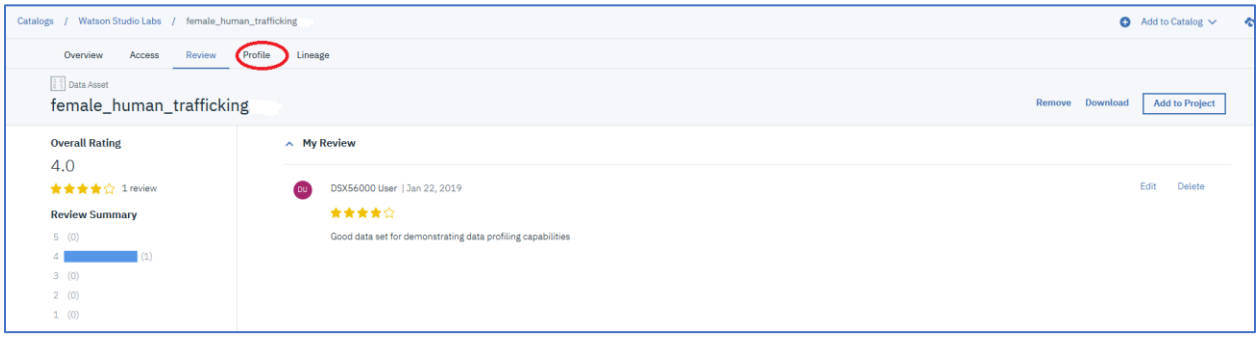
4. Click on the **Review** tab.



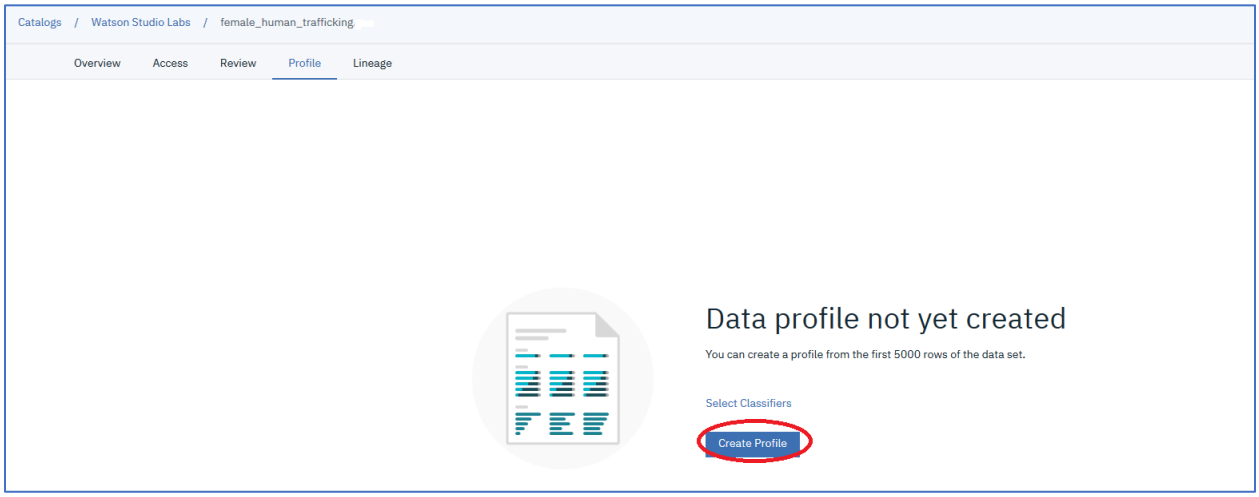
5. Click on the fourth star to rate this asset. Enter “Good data set for demonstrating data profiling capabilities” in the **My Review** text field. Click on the **Submit** button.



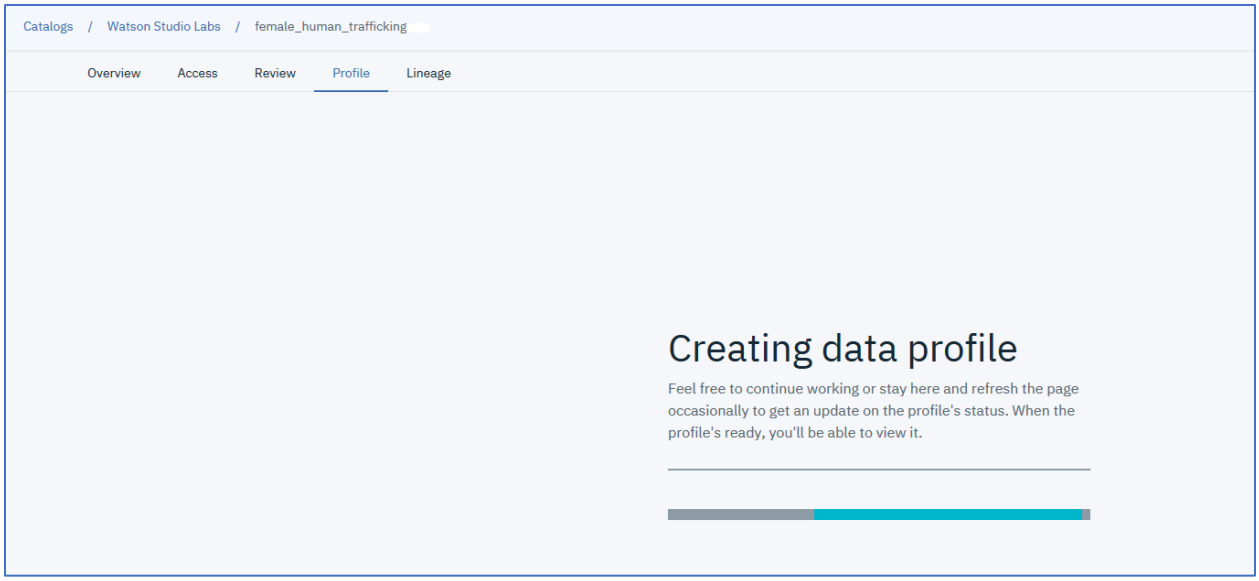
6. The review of the asset is displayed. Click on the Profile tab.



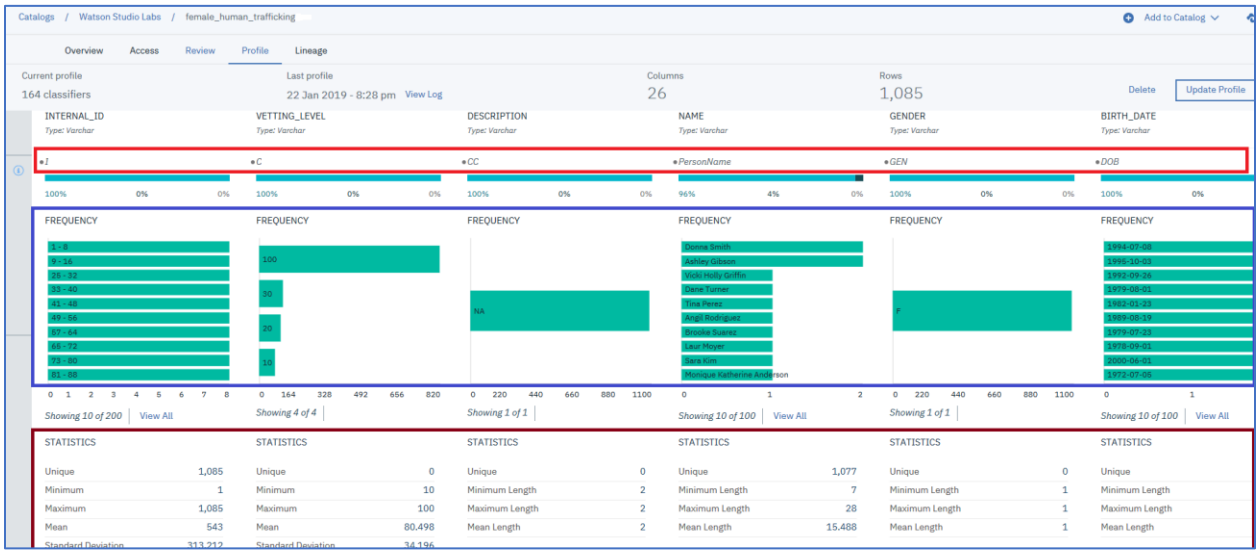
7. Click on **Create Profile**. If a profile was already created, it would be displayed instead of this prompt screen.



8. Wait a minute or so, and then click on browser refresh.



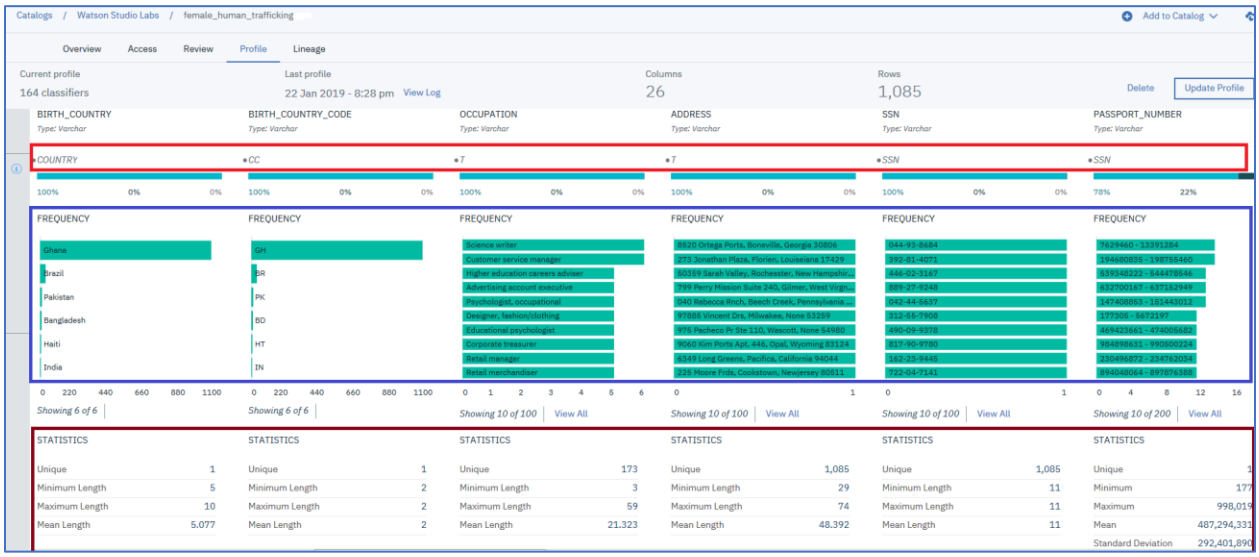
9. The profile information is displayed below in several screenshots based on horizontal scrolling to display many of the columns of the data asset. The highlighted red box contains the column classifications. The highlighted blue box contains counts of the number of occurrences of each value (for discrete fields), and a histogram for continuous fields. The darker red box contains statistics calculated from the contents of the columns.



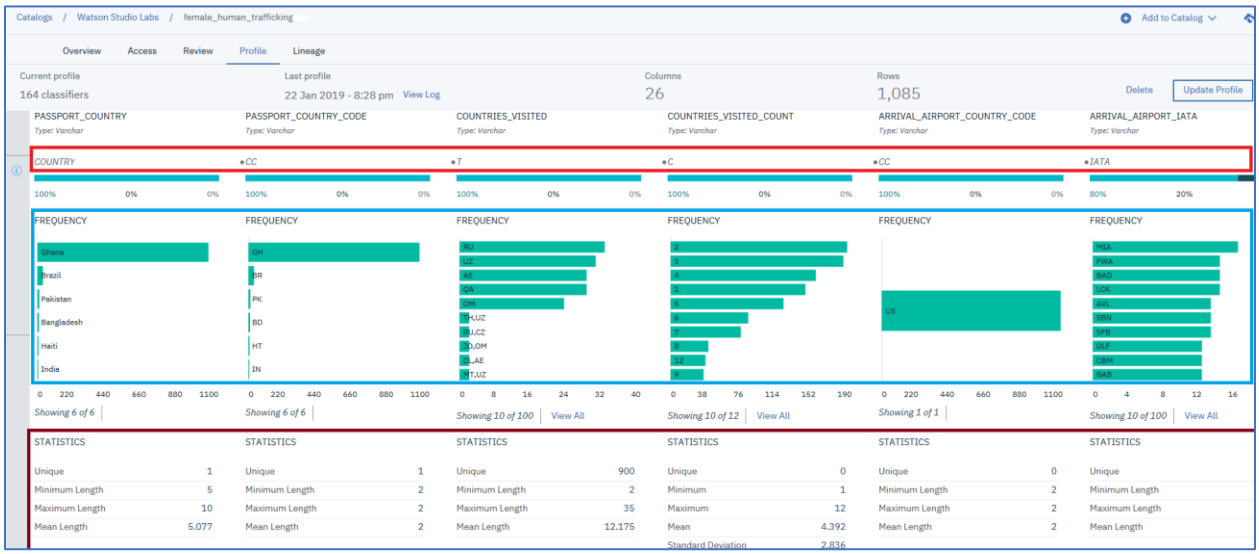
The classifications of the columns are listed below:

- INTERNAL_ID – I (Identifier)
- VETTING_LEVEL – C (Code)
- DESCRIPTION – CC (Country Code) – this is an error due to the description input as NA. We can correct this misclassification.

- NAME: PersonName – the classification accurately determines that this field is the name of a person.
- BIRTH_DATE: DOB – the classification accurately determines that this field represents a date of birth.



- BIRTH_COUNTRY: COUNTRY- correctly classified as a country
- BIRTH_COUNTRY_CODE – CC (Country Code) - correctly classified
- OCCUPATION – T (Text field) – there is not a classifier for occupations so it determines that this is a text string.
- ADDRESS: T (Text Field)
- SSN:SSN – correctly classifies the column as a social security number.
- PASSPORT_NUMBER – incorrectly classifies as a social security number. This can be corrected, although it may be useful to treat as a SSN for privacy reasons.



- PASSPORT_COUNTRY: COUNTRY- correctly classified as a country

- PASSPORT_COUNTRY_CODE: CC(Country Code) - correctly classified
- COUNTRIES_VISITED – T (Text field)
- COUNTRY_VISITED_COUNT – C (Code) – this is an error. It's not a code. This can be re-classified if necessary.
- ARRIVAL_AIRPORT_COUNTRY_CODE: COUNTRY –
- ARRIVAL_AIRPORT_IATA: IATA – the classification accurately determines that this field represents an IATA.

There are more columns in the data set but this gives a representative sample of the profiling capability. The classifications of the columns becomes important when we create and enforce security policies.

Step 6: Access Control

All Catalog members can access assets with a privacy setting of Public. Assets that have a privacy setting of Private are secured based on the asset access control list. A summary of the access control settings for our Watson Studio Labs catalog and the contained Data Assets is shown below.

Watson Studio Lab Catalog members:

1. Catalog owner (logged on user)
2. DSXUser57000@gmail.com

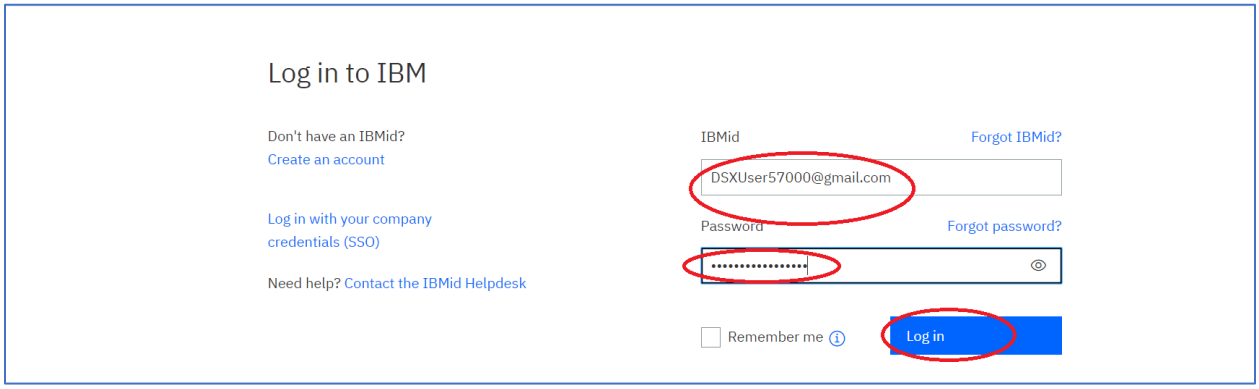
Data Asset Privacy:

1. female_human_trafficking (Private) ACL Members:
   a. asset owner (logged on user)
   b. DSXUser57000@gmail.com
2. Occupations (Private) ACL Members:
   a. asset owner (logged on user)
3. Categories (Public)
4. trafficking (Connection) (Public)

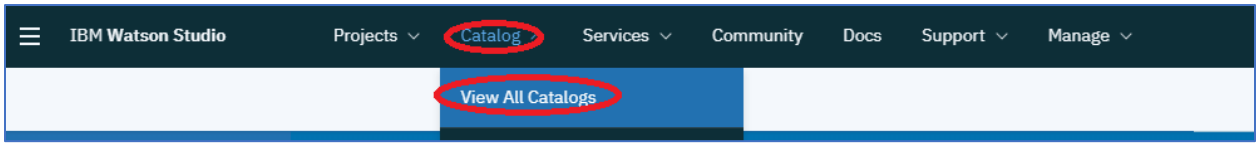
We would expect that if we log in as DSXUser57000@gmail.com, that only 3 Data Assets would appear. The two public Data Assets, and the Private Data Asset where DSXUser57000@gmail is on the access control list. This user would not have access to the Occupations asseet, since it is a Private asset, and the user is NOT on the access control list for this asset. To verify complete the following steps.

1. Open a different browser (if you are using FireFox, then open Chrome or vice versa).
2. Type in datascience.ibm.com for the URL to navigate to.
3. Click on **Sign into IBM Watson Studio Cloud**

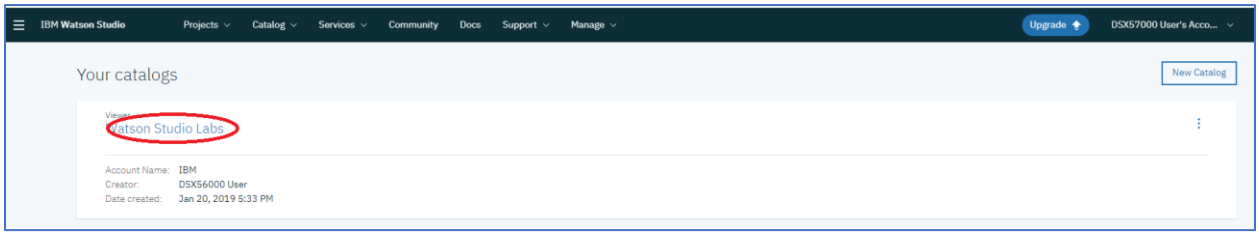
4. Login using DSXUser57000@gmail.com for the user name and DSXPassw0rd57000! for the password.



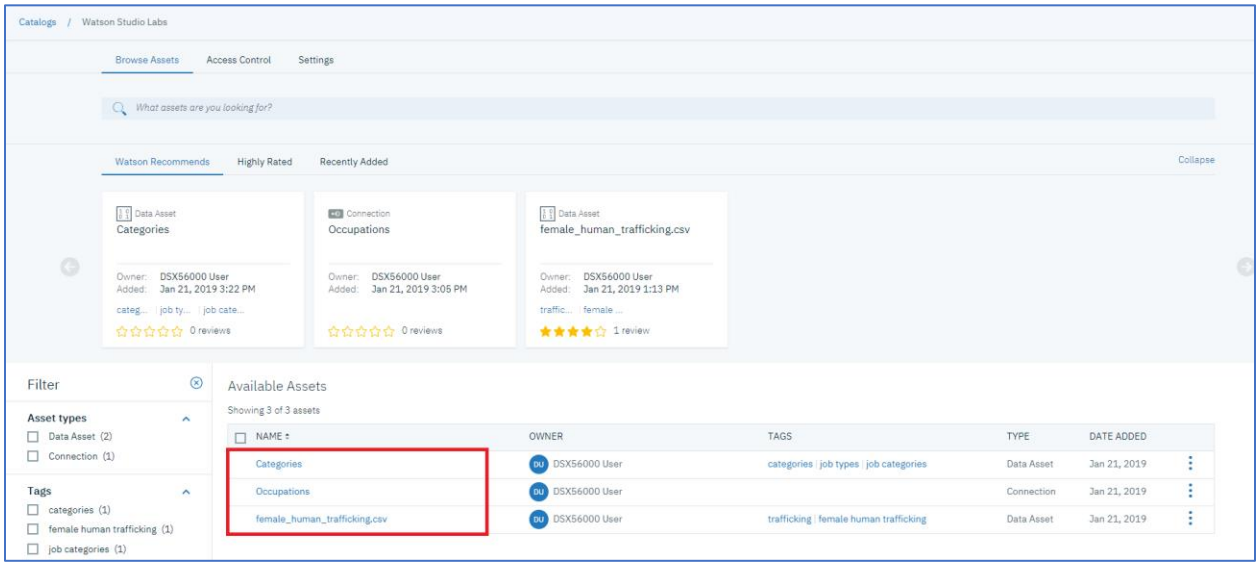
5. Click on **Catalog** and then **View All Catalogs**.



6. Click on the **Watson Studio Labs** catalog that corresponds to your username. Note that there will be 1 catalog per lab participant.



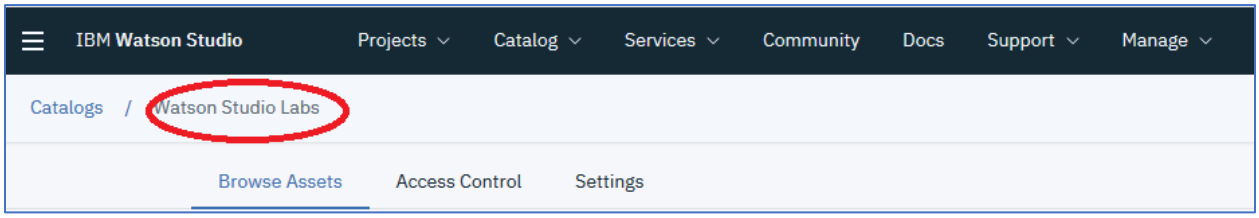
7. The Watson Studio Labs catalog is displayed. Three Data Assets are listed as expected.



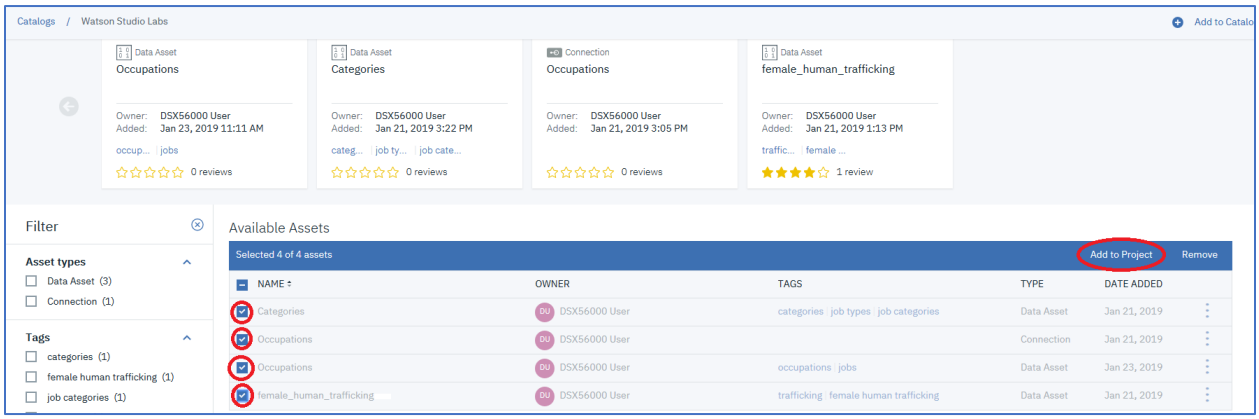
Step 8: Push the data assets to the project

This step will add the cataloged data assets to our Watson Studio project to be used in subsequent labs.

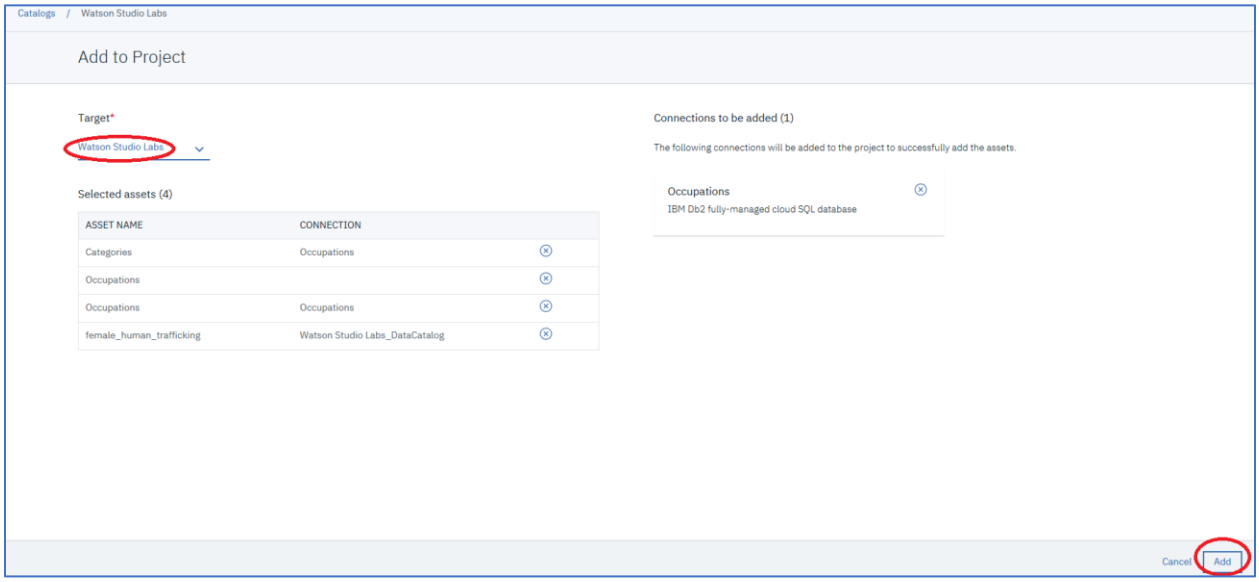
1. Click on Watson Studio Lab to go to the Catalog home screen.



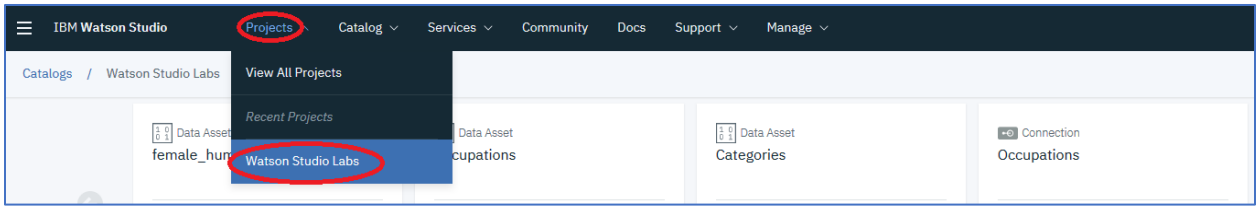
2. Hover the cursor to the left of each data asset and then click on the checkbox. Click on **Add to Project**.



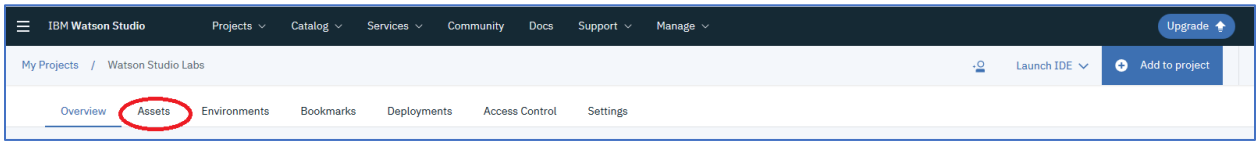
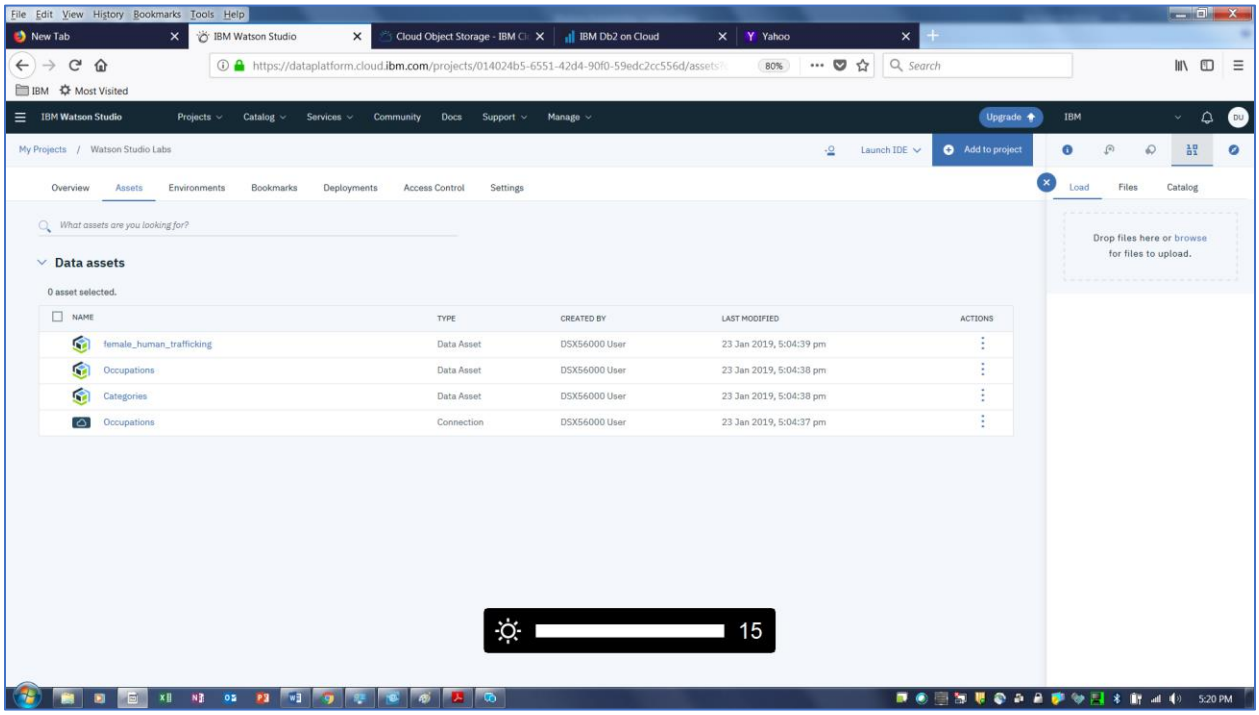
3. Select the Watson Studio Labs as the **Target**. Click on **Add.**



4. Click on **Projects** and then **Watson Studio Labs**.



5. Click on the **Assets** tab.



6. The Data assets are now incorporated in the project.