

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.

End-to-End Data Science

The general flow of the End to End Data Science PoT will be guided by the activities shown in Figure 1- End to End Flow. This lab will focus on the Find Governed Data activity.

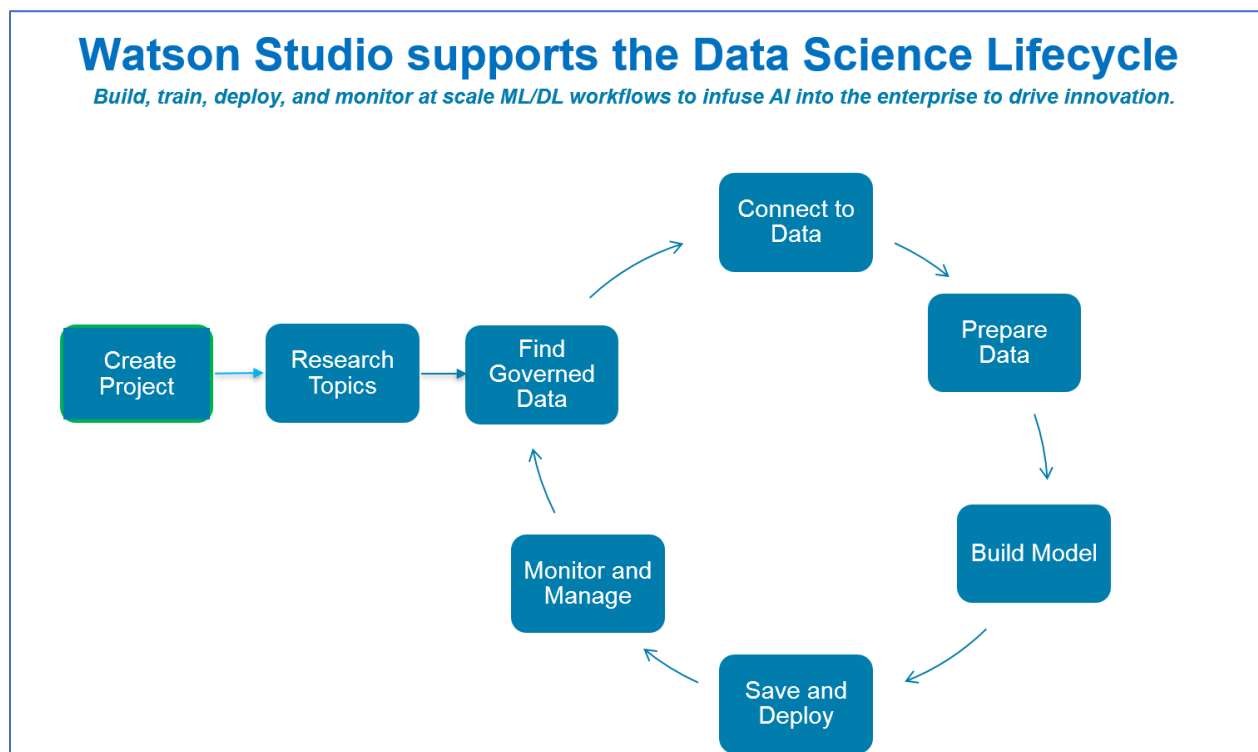


Figure 1- End to End Flow

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 one 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
8. Push the Data Assets to the project set up in Lab-1

Download Files

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 two files extracted, (1) Categories.csv, and (2) Occupation.csv.

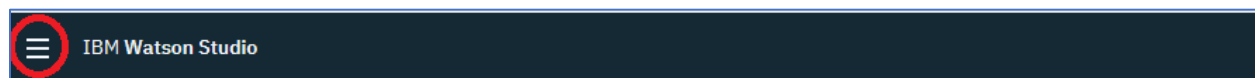
Perform Catalog Tasks

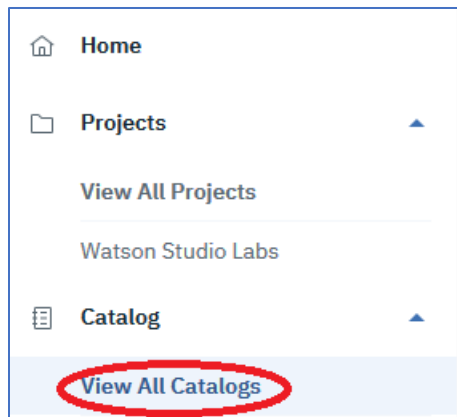
Step 1 – Create a Governed Catalog

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

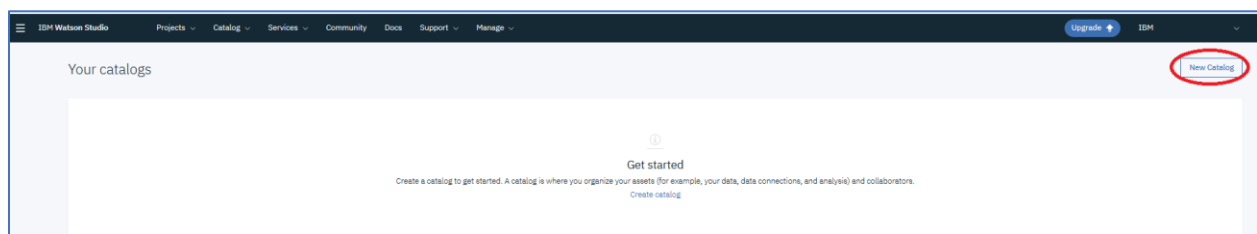


3. Click on  icon and then **View All Catalogs**.

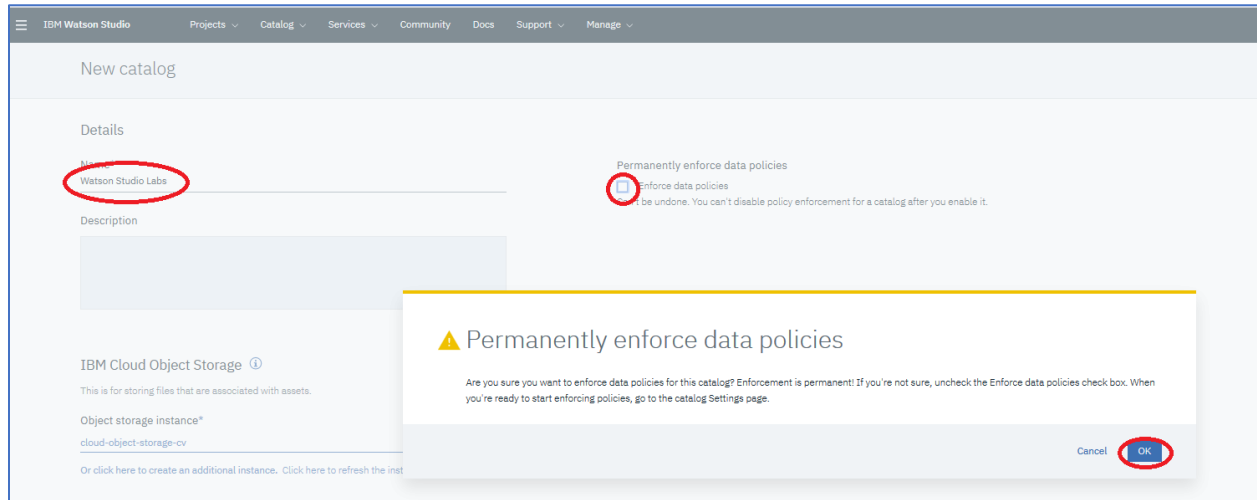




4. Click on **New Catalog**.



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.

New catalog

Details

Name*
Watson Studio Labs

Description
1000

IBM Cloud Object Storage ⓘ
This is for storing files that are associated with assets.

Object storage instance*
cloud-object-storage-cv
Or click here to create an additional instance. Click here to refresh the instances.

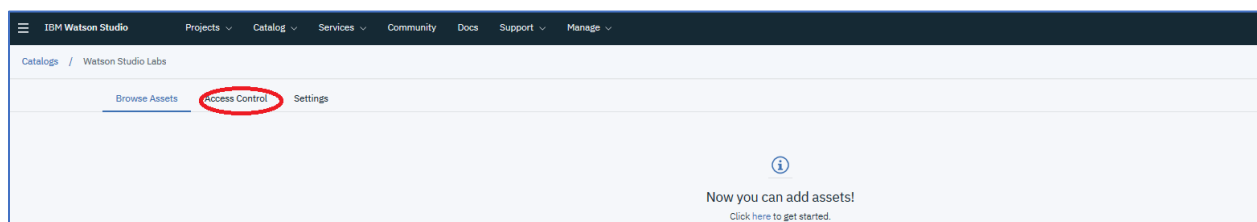
Permanently enforce data policies
☒ Enforce data policies
Can't be undone. You can't disable policy enforcement for a catalog after you enable it.

Cancel Create

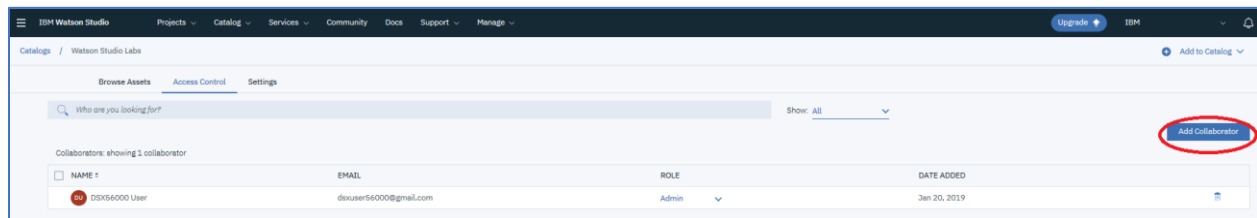
Step 2: Add a Member to the Catalog

To demonstrate the Catalog access control capability, we will add the username that was provided by the lab instructor as a member of the catalog. For illustrative purposes, we are using wsuserbb@gmail.com in this document. Note that the name of this user is “John Doe”.

1. Click on **Access Control**.



2. Click on **Add Collaborator**.



3. Enter John, the list will populate with John Doe and e-mail wsuserbb@gmail.com. Click on the **John Doe** entry in the list box.

Add Collaborators as Viewers

Collaborators (0)

John

John Doe wsuserbb@gmail.com

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

Add Collaborators as Viewers

Collaborators (1)

John Doe X

Role: Viewer

Having trouble adding collaborators? ⓘ

Cancel Add

5. “John Doe” is added as a member of the catalog.

Collaborators: showing 2 collaborators

<input type="checkbox"/>	NAME	EMAIL	ROLE	DATE ADDED
<input checked="" type="checkbox"/>	DSX59000 User	dsxuser59000@gmail.com	Admin	Jan 30, 2019
<input checked="" type="checkbox"/>	John Doe	wsuserbb@gmail.com	Viewer	Feb 5, 2019

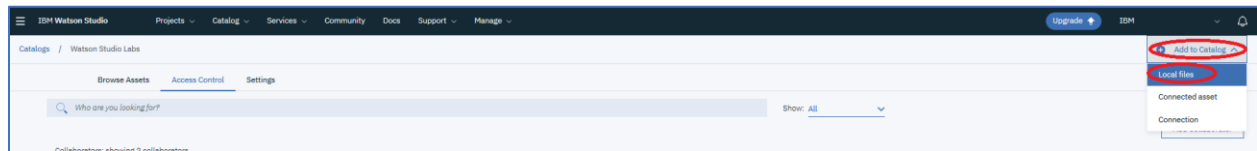
Step 3: Add Data Assets to the Catalog

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. We will add a connected data asset that will point to the FEMALE_HUMAN_TRAFFICKING table.

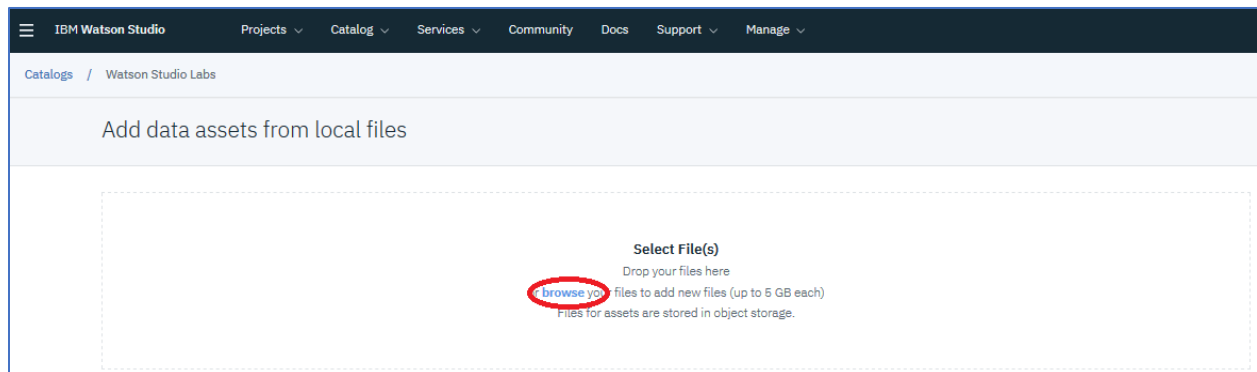
This step will introduce you to the three methods available to discover and catalog data assets; **Local files**, **Connected asset** and **Connection**. You will use these methods to catalog data assets into the newly created Knowledge Catalog and then tag them for users to easily find them, understand 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

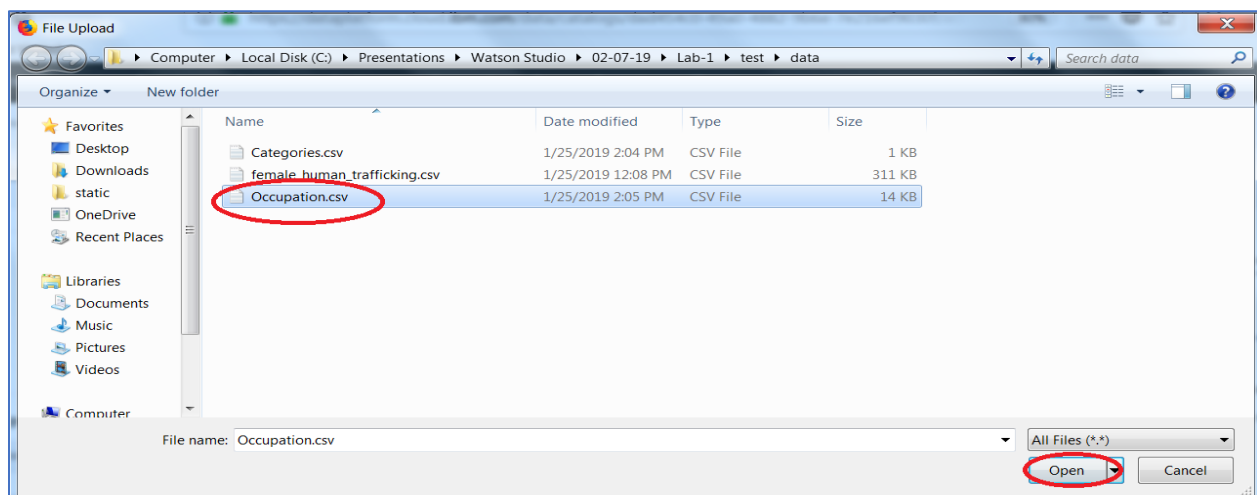
1. Click on **Add to Catalog** and click on **Local Files**.



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**.

Add data assets from local files

Selected Files (1)* [Edit name and format](#)

Continue adding files in the drop zone below or browse to select files

ASSET NAME	FORMAT
Occupation.csv	Microsoft Excel

Target

Watson Studio Labs - Watson Studio Labs_DataCatalog

Description

300

Tags

Start typing to add values

occupations X jobs X

Classification* [?](#)

None

Privacy

☐ Public ☒ Private

Only the asset owner and asset members can find or use the asset.

Members

[Add members](#)

[Cancel](#) [Add](#)

Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.

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.

Browse Assets Access Control Settings

What assets are you looking for?

Watson Recommends Highly Rated Recently Added Collapse

Asset Metadata:

Data Asset
Occupation.csv

Owner: DSX59000 User
Added: Jan 30, 2019 11:38 PM

occup... jobs

☆☆☆☆☆ 0 reviews

Filter

Asset types

☐ Data Asset (1)

Tags

☐ jobs (1)

☐ occupations (1)

Available Assets

Showing 1 of 1 asset

NAME	OWNER	TAGS	TYPE	DATE ADDED
Occupation.csv	DSX59000 User	occupations jobs	Data Asset	Jan 30, 2019

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**.

Add data assets from local files

Selected Files (1)* [Edit name and format](#)

Continue adding files in the drop zone below or browse to select files

ASSET NAME	FORMAT
Categories.csv	Microsoft Excel

Target

Watson Studio Labs - Watson Studio Labs_DataCatalog

Description

300

Tags
Start typing to add values

job types X job categories X

Classification* ⓘ

None

Privacy

☒ Public ☐ Private

All Catalog members can find and use the asset.

Members

ⓘ Add members

Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.

Cancel **Add**

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

Successfully added data asset Categories.csv to Watson Studio Labs.

[Browse Assets](#) [Access Control](#) [Settings](#)

What assets are you looking for?

[Watson Recommends](#) [Highly Rated](#) [Recently Added](#) [Collapse](#)

Data Asset
Categories.csv

Owner: DSX59000 User
Added: Jan 30, 2019 11:49 PM
job ty... | job cate...
☆☆☆☆☆ 0 reviews

Data Asset
Occupation.csv

Owner: DSX59000 User
Added: Jan 30, 2019 11:38 PM
occup... | jobs
☆☆☆☆☆ 0 reviews

Filter ⓘ

Asset types

☐ Data Asset (2)

Tags

☐ job categories (1)
☐ job types (1)
☐ jobs (1)
☐ occupations (1)

Available Assets
Showing 2 of 2 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Categories.csv	DSX59000 User	job types job categories	Data Asset	Jan 30, 2019
Occupation.csv	DSX59000 User	occupations jobs	Data Asset	Jan 30, 2019

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

Available Assets

Showing 2 of 2 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Categories.csv	DSX59000 User	job types job categories	Data Asset	Jan 30, 2019
Occupation.csv	DSX59000 User	occupations jobs	Data Asset	Jan 30, 2019

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

Catalogs / Watson Studio Labs / Occupation.csv

Overview Access Review Profile Lineage

Data Asset

Occupation.csv

Remove Download Add to Project

Description
There is no description available for this asset.

Added: Jan 31, 2019 12:00 AM
Format: Microsoft Excel
Size: 30 KB

Tags
occupations jobs

Reviews
☆☆☆☆ 0 reviews

Connection
Source: Watson Studio Labs_DataCatalog
Source type: Cloud Object Storage

Classification
None

Schema: 2 Columns | 492 Rows
Preview: 492 rows | Last refresh: right now | Refresh

OCCUPATION	Code
Text	Code
Chiroprapist	6
Engineer, structural	2
Engineer, land	2
Careers adviser	15
Geneticist, molecular	7
Electronics engineer	2
Editor, commissioning	4
Writer	4
Recruitment consultant	15
Theatre manager	8
Statistician	2
Merchant navy officer	15
Advertising account executive	9
Administrator, local government	6

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

Catalogs / Watson Studio Labs / Occupation.csv

Overview Access Review Profile Lineage

Data Asset

Occupation.csv

Remove Download Add to Project

Apply Cancel

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**.

IBM Watson Studio

Projects Catalog Services Community Docs Support Manage

Upgrade IBM

Catalogs / Watson Studio Labs

Browse Assets Access Control Settings

What assets are you looking for?

Add to Catalog

Local files

Connected asset

Connection

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

IBM Watson Studio

Projects Catalog Services Community Docs Support Manage

Upgrade IBM

New connection

IBM services

IBM Siginights HDPS

Compose for MySQL

DB2 for i

DB2 Warehouse

PureData for Analytics

Cloud Object Storage

Compose for PostgreSQL

DB2 for z/OS

Informix

IBM Informix database

Cloud Object Storage (Infrastructure)

IBM Db2

DB2 Hosted

Object Storage OpenStack Swift

Cloudant

IBM Db2 Big SQL

DB2 on Cloud

Object Storage OpenStack Swift (Infrastructure)

Third-party services

Amazon Redshift

Dropbox

MySQL

Salesforce.com

Teradata

Amazon S3

FTP

Microsoft Azure Data Lake Store

Oracle

Sybase

Apache Hive

Google BigQuery

Microsoft Azure SQL Database

Pivotal Greenplum

Sybase IQ

Cloudera Impala

Hortonworks HDFS

Microsoft SQL Server

PostgreSQL

Tableau

- Enter “trafficking” for the **Name** of the connection. Click on [Access DB2 Credentials](#) to obtain the DB2 Credentials. You should have an index card that provides your assigned database. If not, ask your instructor.
- Using the credentials for your assigned database, cut and paste the fields under Username, Hostname, Password, and Database. Click on **Create**.

Connection overview

Name
trafficking

Description
IBM Db2 fully-managed cloud SQL database

Connection Details

Username *
cxd96032

Hostname or IP Address *
db2db-tm-sbox-yp-dal09-03.services.dal.ibmcloud.net

Secure Gateway
☐ Use a secure gateway

Password *

Database *
BLUDB

Connection discovery
☐ Discover data assets

Enter information for the selected data source

Cancel Create

- 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.

Catalogs / Watson Studio Labs

Browse Assets Access Control Settings

What assets are you looking for?

Watson Recommends Highly Rated Recently Added

Connection trafficking

Owner: DSX59000 User
Added: Jan 31, 2019 12:03 AM

Data Asset Categories

Owner: DSX59000 User
Added: Jan 31, 2019 12:01 AM

Data Asset Occupation

Owner: DSX59000 User
Added: Jan 31, 2019 12:00 AM

Filter

Asset types

☐ Data Asset (2)

☒ Connection (1)

Tags

☐ job categories (1)

☐ job types (1)

☐ jobs (1)

☐ occupations (1)

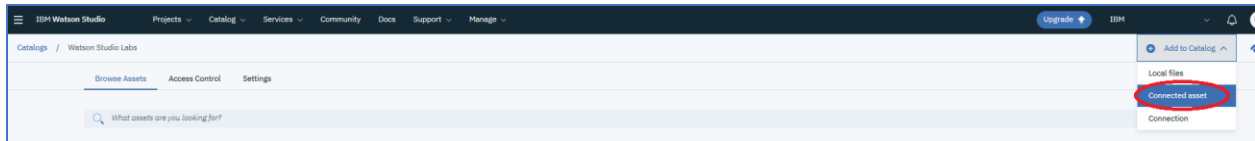
Available Assets

Showing 3 of 3 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Categories	DSX59000 User	job types job categories	Data Asset	Jan 31, 2019
Occupation	DSX59000 User	occupations jobs	Data Asset	Jan 31, 2019
trafficking	DSX59000 User		Connection	Jan 31, 2019

Add Connected Data

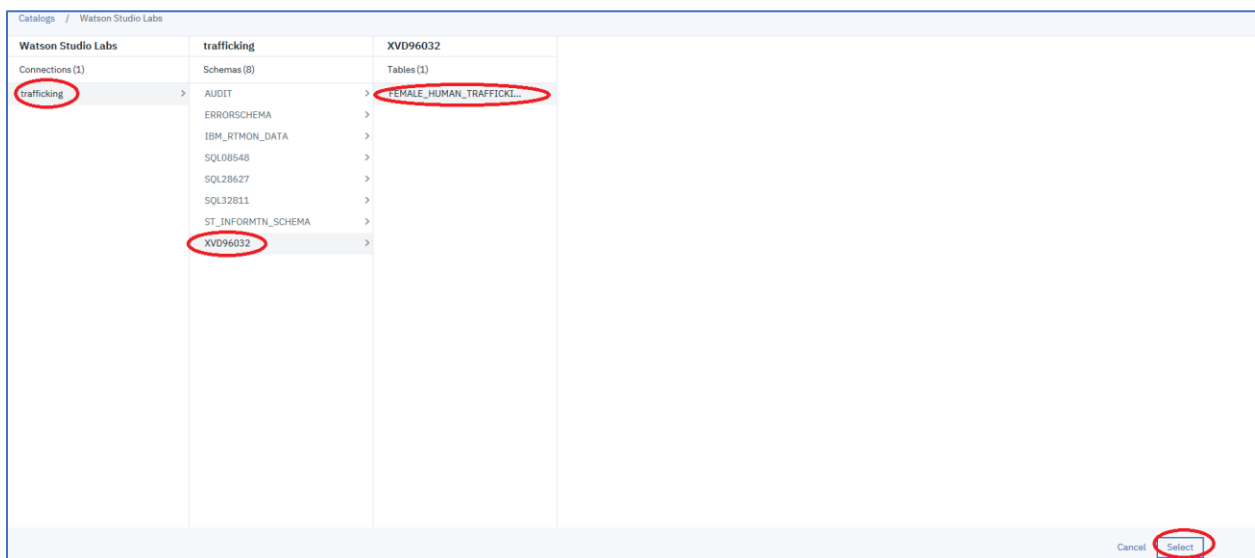
- We have added a connection to the DB2 on Cloud database service. We will now add a Connected Data Asset to point to the FEMALE_HUMAN_TRAFFICKING table that was created earlier. Click on **Add to Catalog** and click on **Connected asset**.



2. Click on **Select Source**.

 This screenshot shows the 'Add asset from connection' form. The form has several fields: 'Source*' with a 'Select Source' button (highlighted with a red circle), 'Name*' with a text input field, and 'Description' with a larger text area. On the right side, there are sections for 'Tags' (with a 'Start typing to add values' prompt), 'Classification*' (set to 'None'), 'Privacy' (with 'Public' and 'Private' radio buttons, where 'Private' is selected), and 'Members' (with an 'Add members' button).

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.

Add asset from connection

Source: Db2 on Cloud Change source

trafficking / JVD96032/FEMALE_HUMAN_TRAFFICKING

Name* female_human_trafficking 76

Description 300

Tags

Start typing to add values

trafficking X female human traffi... X

Classification* (1) Personally Identifiable Information

Privacy

☐ Public ☒ Private

Only the asset owner and asset members can find or use the asset.

Members

+ Add members

Cancel Add

5. Enter John in the **Members** field. The list of matching users is displayed in the member list. Click on **John Doe**.

Add Members

Members (0)

John

John Doe wsuserbb@gmail.com

6. Click on **Add** to add John Doe to the access control list for the female_human_trafficking asset.

Add Members

Members (1)

John Doe X

You can only add people who are in the catalog.

Cancel **Add**

7. Click on **Add**.

Add asset from connection

Source Db2 on Cloud Change source

trafficking : /XVD96032/FEMALE_HUMAN_TRAFFICKING

Name*
female_human_trafficking 76

Description
300

Tags
Start typing to add values
trafficking X female human traffi... X

Classification* ?
Personally Identifiable Information

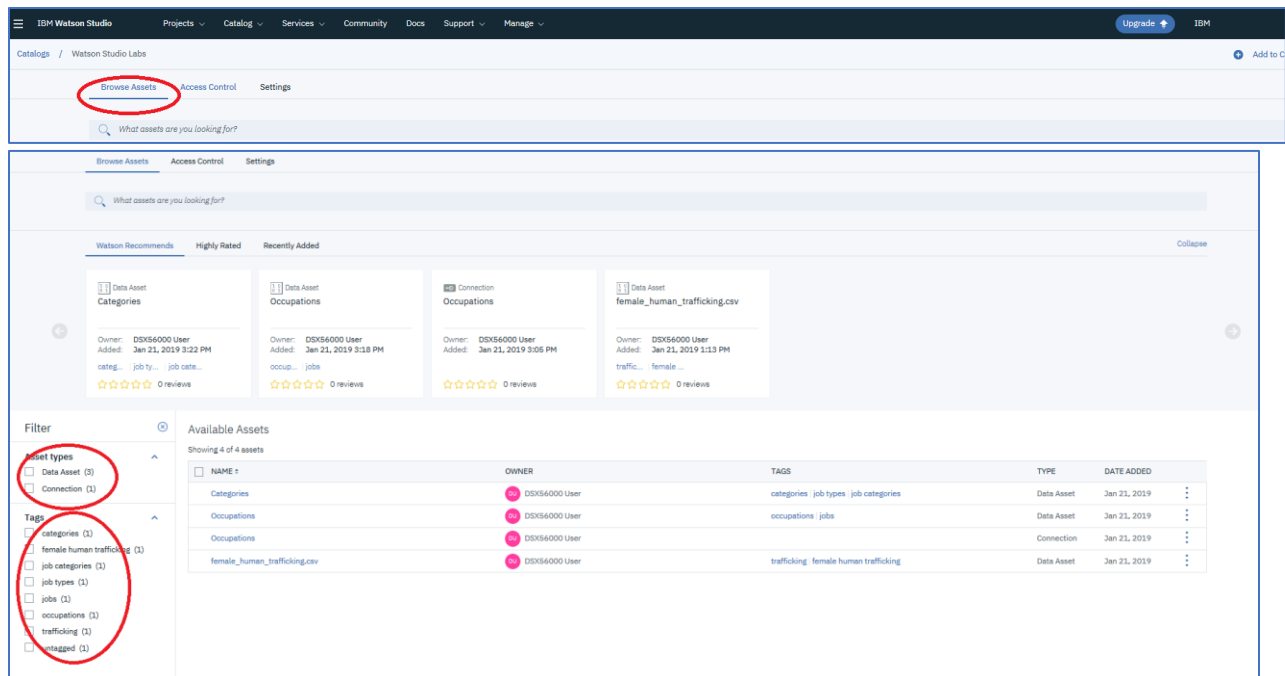
Privacy
☐ Public ☒ Private
Only the asset owner and asset members can find or use the asset.

Members + Add members

NAME
John Doe

Cancel **Add**

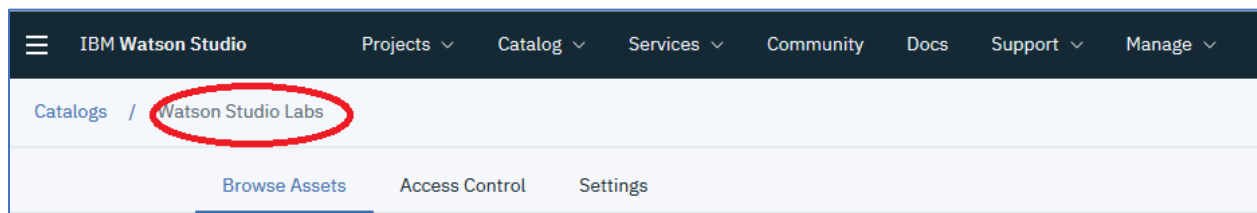
8. All of the assets have now been added to the catalog. The catalog display should appear as below.



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.

Search: **job** Clear

Filter

Asset types

- ☐ Data Asset (3)
- ☐ Connection (1)

Tags

- ☐ categories (1)
- ☐ female human trafficking (1)
- ☐ job categories (1)
- ☐ job types (1)
- ☐ jobs (1)
- ☐ occupations (1)
- ☐ trafficking (1)
- ☐ untagged (1)

Showing 2 of 2 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Categories	DSX56000 User	categories job types job categories	Data Asset	Jan 21, 2019
Occupations	DSX56000 User	occupations jobs	Data Asset	Jan 21, 2019

- 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.

What assets are you looking for?

Watson Recommends | Highly Rated | Recently Added

Data Asset
Categories

Owner: DSX56000 User
Added: Jan 21, 2019 3:22 PM

category... | job ty... | job cate...

0 reviews

Data Asset
Occupations

Owner: DSX56000 User
Added: Jan 21, 2019 3:18 PM

occup... | jobs

0 reviews

Connection
Occupations

Owner: DSX56000 User
Added: Jan 21, 2019 3:05 PM

0 reviews

Data Asset
female_human_trafficking

Owner: DSX56000 User
Added: Jan 21, 2019 1:13 PM

trafficking... | female...

0 reviews

Filter

Asset types

- ☒ Data Asset (3)
- ☐ Connection (1)

Tags

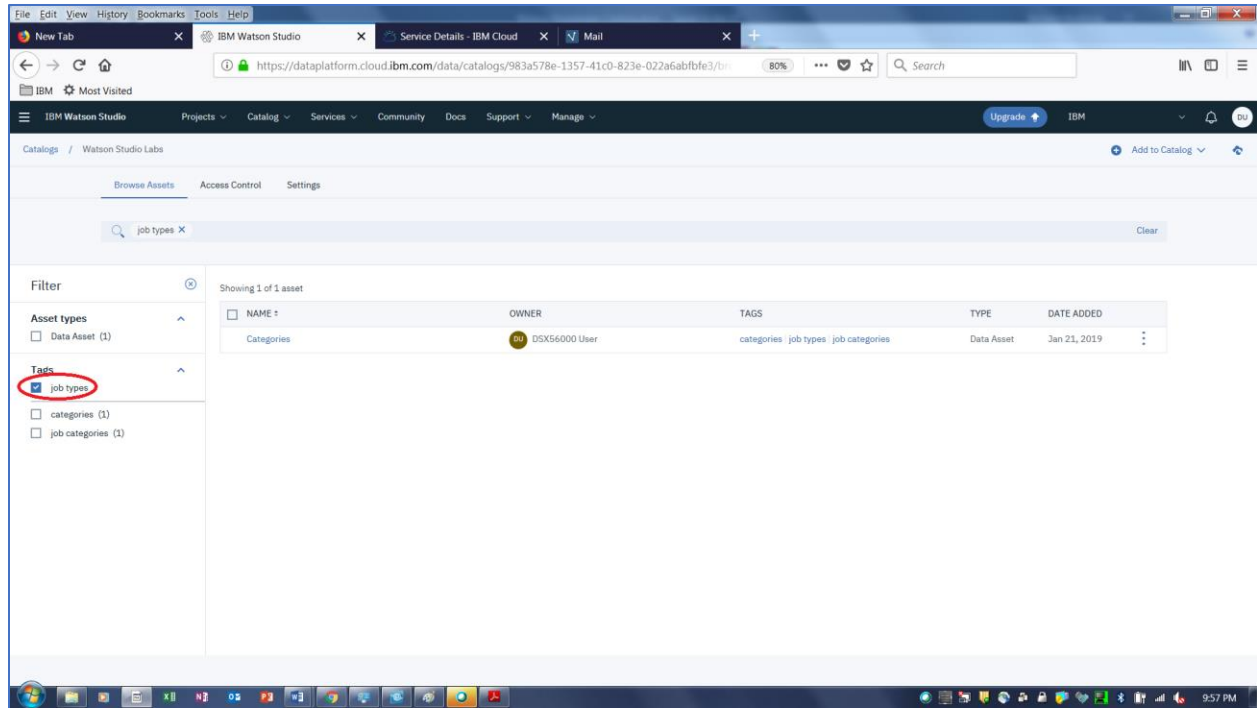
- ☐ categories (1)
- ☐ female human trafficking (1)
- ☐ job categories (1)
- ☐ job types (1)
- ☐ jobs (1)
- ☐ occupations (1)
- ☐ trafficking (1)

Available Assets

Showing 3 of 3 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED
Categories	DSX56000 User	categories job types job categories	Data Asset	Jan 21, 2019
Occupations	DSX56000 User	occupations jobs	Data Asset	Jan 21, 2019
female_human_trafficking	DSX56000 User	trafficking female human trafficking	Data Asset	Jan 21, 2019

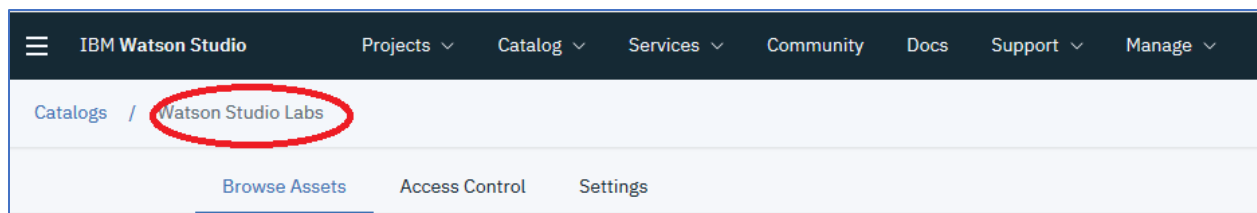
- Click on the Tag **job types** in the Filter area. The Categories Data Asset is listed. Note that the Filter area has been updated to only list Asset Types, and Tags that correspond to the search result.



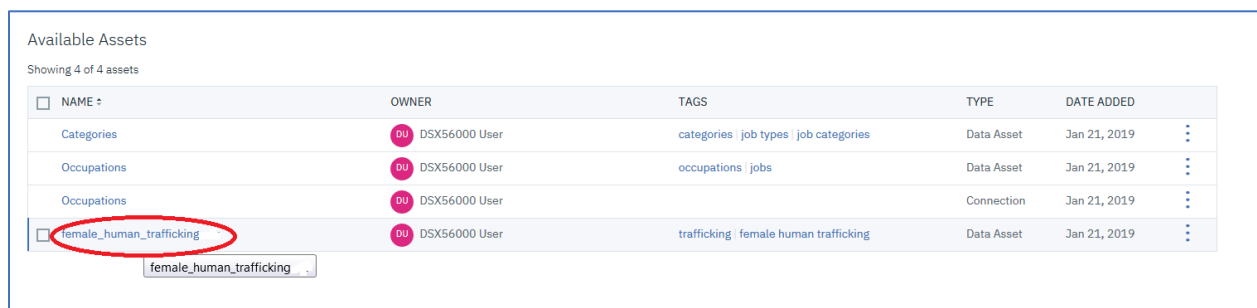
Step 5: Edit/Review/Profile Data Asset

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.



- 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.

Catalogs / Watson Studio Labs / female_human_trafficking

Overview Access Review Profile Lineage

Data Asset
female_human_trafficking

Remove Download Add to Project

Description
Travel data to be examined for trafficking
Added: Jan 21, 2019 1:13 PM
Format: Microsoft Excel

Tags
Tracking female human trafficking

Reviews
☆☆☆☆ 0 reviews

Connection
Source: Watson Studio Labs_DataCatalog
Source type: Cloud Object Storage

Classification
Personally identifiable information
Personally identifiable information (PII) is defined as any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another can be considered PII.

Schema: 26 Columns
Preview: 1000 rows Last refresh: 2 hours ago Refresh

INTERNAL_ID	VETTING_LEVEL	DESCRIPTION	NAME	GENDER	BIRTH_DATE	BIRTH_COUNT...	BIRTH_COUNTRY_CODE	OCCUPATION	ADDRESS	SSN	PASSPORT
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
1	100	NA	Kathleen Bailey	F	1985-11-27	Ghana	GH	Chiroprapist	79421 Jordan On	885-71-9055	240479504
2	30	NA	Kelly Robbins	F	1971-01-17	Pakistan	PK	Engineer, structu	5808 Jacobs Uni	141-10-9199	449064032
3	100	NA	Linda Stewart	F	1997-04-15	Ghana	GH	Engineer, land	824 Kristin Grv, /	011-46-5304	639217816
4	30	NA	Stacey Courtney	F	1984-03-27	Ghana	GH	Careers adviser	322 Hutchinson	691-25-2647	212838449
5	30	NA	Erika Patie Fowle	F	1998-06-18	Ghana	GH	Geneticist, mole	07695 Michael V	165-33-0802	16514173
6	100	NA	Ashlee Fisher	F	1981-01-05	Ghana	GH	Electronics engin	4838 Cassandra	072-16-8742	513428119
7	100	NA	Jacqueline Clark	F	1976-01-19	Ghana	GH	Editor, commissi	1148 Wang Fall	634-03-1462	742709558
8	30	NA	Shelli Chapman	F	1997-08-24	Ghana	GH	Writer	72346 Wilson Fe	004-07-1446	170002642
9	10	NA	Sabrina Debra W	F	1977-02-21	Haiti	HT	Recruitment con	251 John Sh, Fir	233-53-1156	934214048
10	100	NA	Ashley Gibson	F	1986-04-06	Ghana	GH	Theatre manager	17894 Gary Grea	565-33-0366	3699030
11	100	NA	Tammy Sara Gan	F	1995-08-17	Ghana	GH	Statistician	2902 Rasmusser	069-16-0401	136448546
12	10	NA	Sandry Santiago	F	1971-12-16	Ghana	GH	Merchant navy of	9486 Pratt Wall,	778-07-9880	625553653
13	100	NA	Becky Cynthia St	F	1993-11-05	Ghana	GH	Advertising accou	6362 William Cll	612-01-4772	411247380

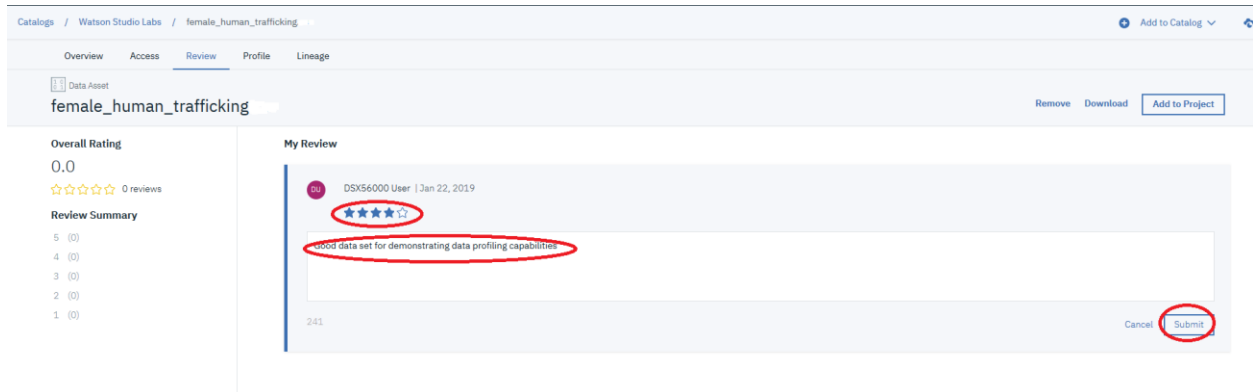
The actions circled in blue will **Remove** the asset from the catalog, **Download** the asset to the local file system, or add the asset to a catalog enabled project (**Add to Project**). Hovering over the **Name** of the data asset, the **Description**, the **Tags**, or the **Classification** will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

- Click on the **Review** tab.

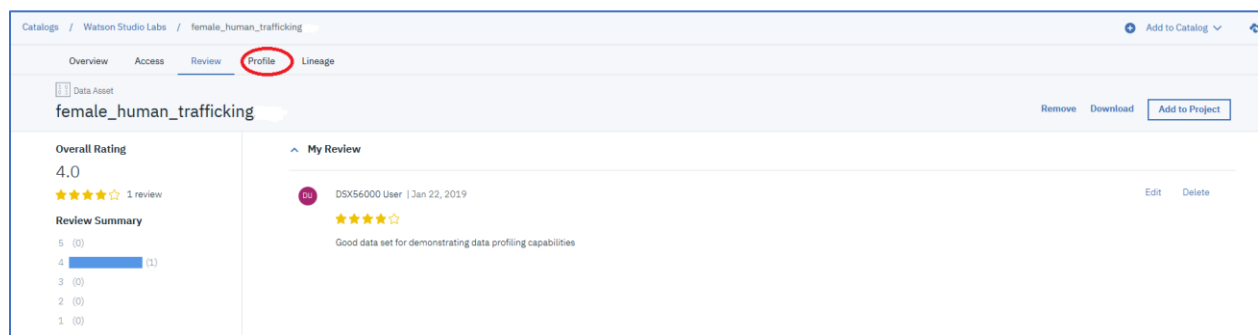
Catalogs / Watson Studio Labs / female_human_trafficking

Overview Access Review Profile Lineage

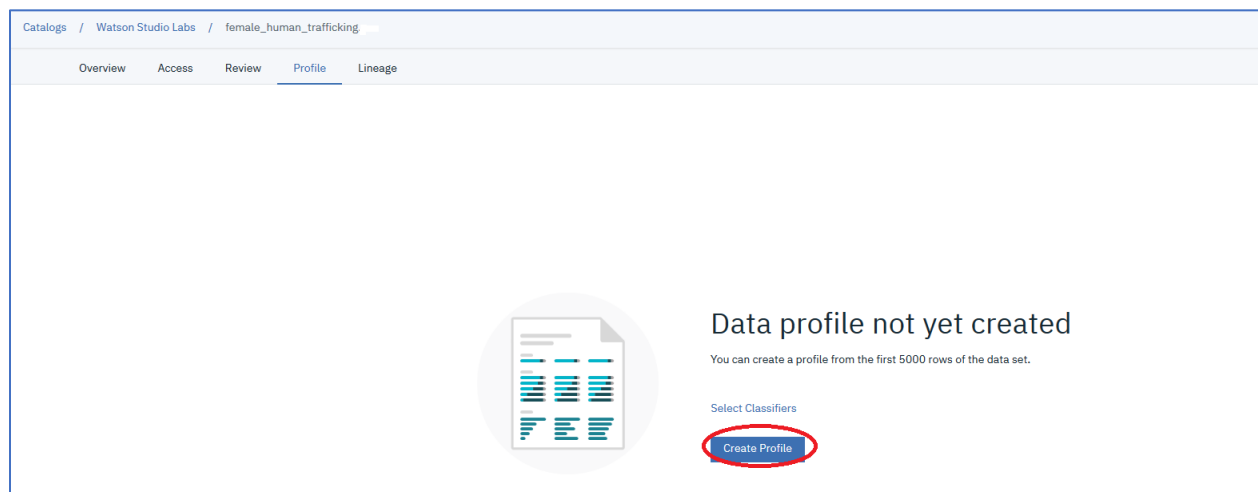
- Click on the fourth star to rate this asset. Enter "Good data set for demonstrating data profiling capabilities" in the **My Review** text field. Click on the **Submit** button.



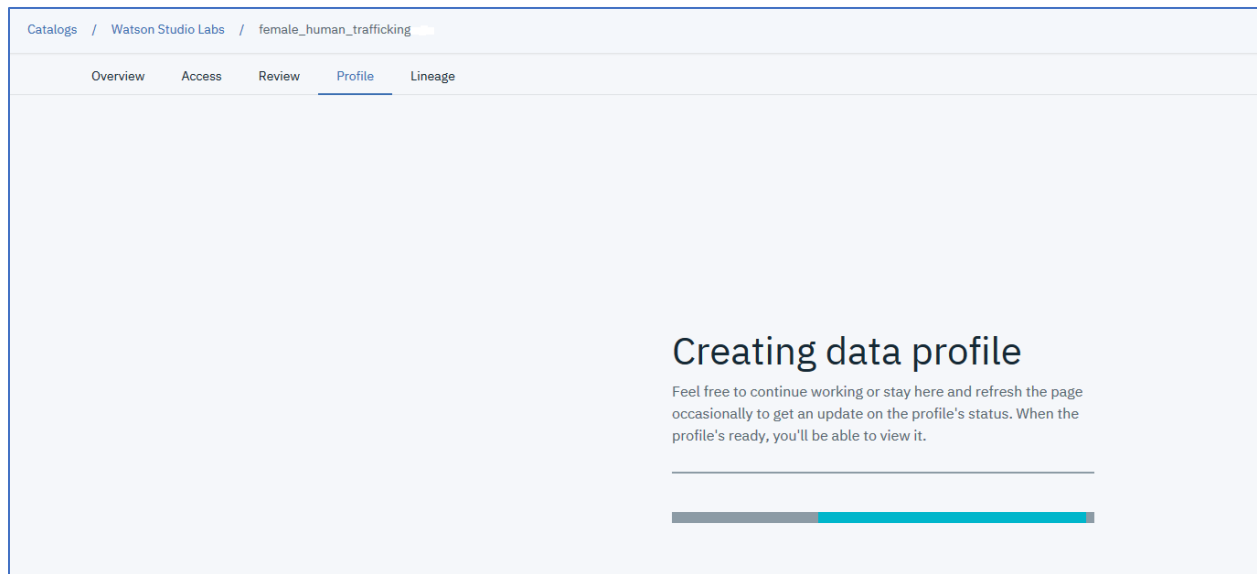
6. The review of the asset is displayed. Click on the Profile tab.



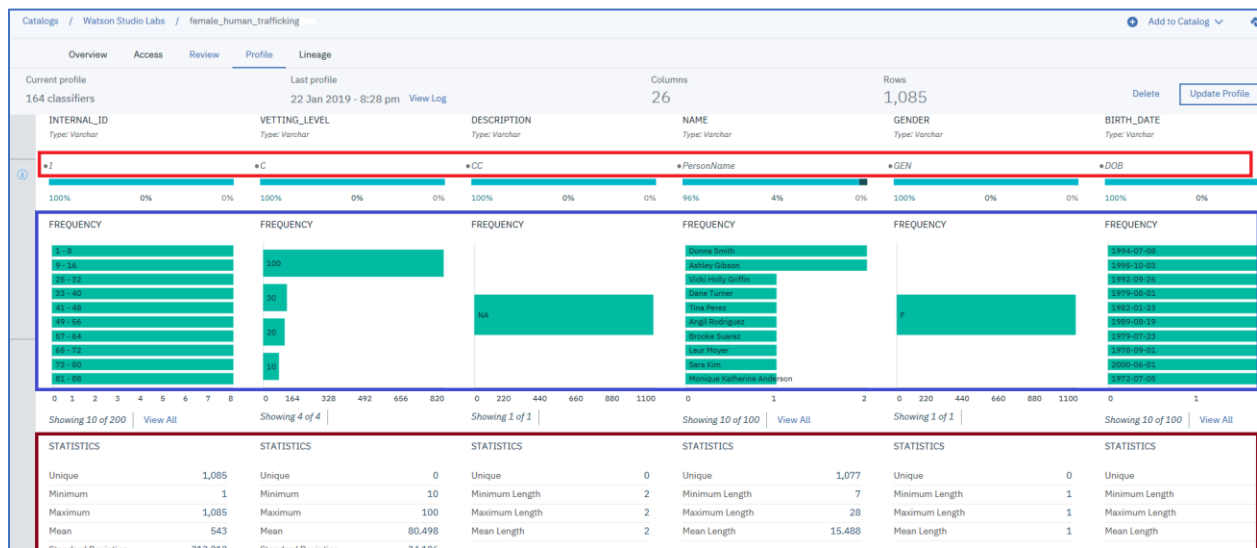
7. Click on **Create Profile**. If a profile was already created, it would be displayed instead of this prompt screen.



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



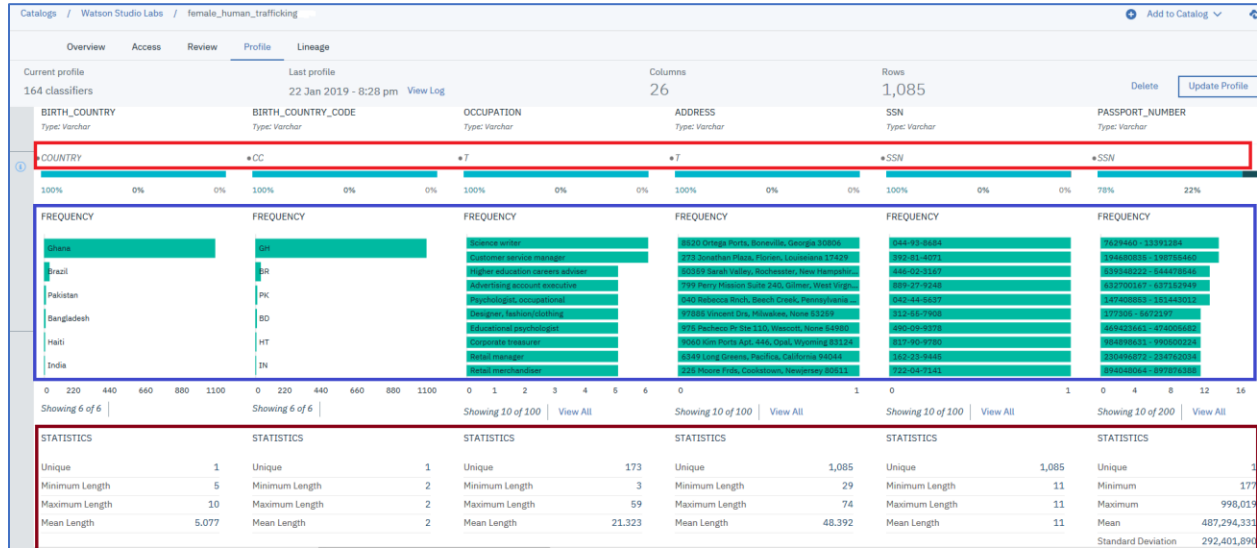
9. The profile information is displayed below using several screenshots, based on horizontal scrolling, to display many of the columns of the data asset. The highlighted red box contains the inferred column classifications. The highlighted blue box contains counts of the number of occurrences of each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



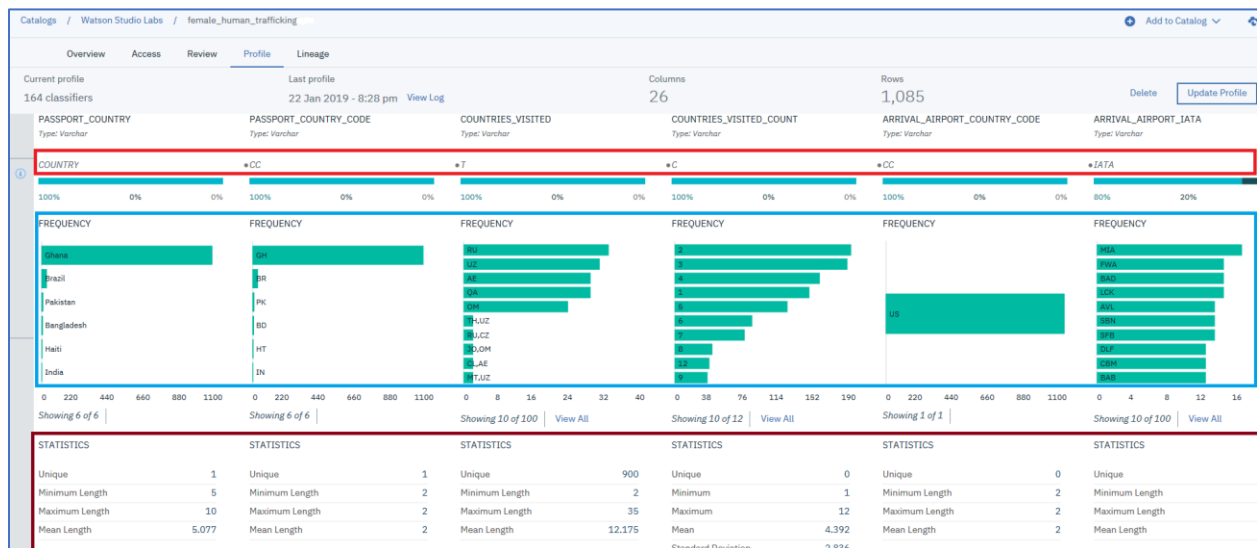
The classifications of the columns are listed below:

- 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 occupation, 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 as we will see below.



- 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 classification of the columns becomes important when we create and enforce security policies.

10. Let's fix the classification for passport. Hover the mouse over the **Passport** Column and click on the ▼ icon.


PASSPORT_NUMBER
Type: Integer
US Social Securit... ▼
240479506
449064032
639217810
212838449
16514173
513428119

11. Click on **View All**.

PASSPORT_NUMBER	PASSPORT_COUNT
Type: Integer	Type: String
US Social Securit... ^	Country Name
<div> <div>✓ US Social Security... 100% matches</div> <div>Not classified 21% matches</div> <div>Canadian Social Ins... 7% matches</div> <div>Routing Transit Num... 3% matches</div> <div>View all</div> </div>	
513428119	Ghana

12. Type "Pass" in the **Search** field. Then click on **Passport Number**.

Select Classifier for SSN

 Pass


Showing 1 of 164 classifiers

Passport Number

Passport number is the unique ID assigned to a travel document, usually issued by the government of a nation, that certifies the identity and nationality of its holder for the purpose of international travel.

13. Click on **Select**.

Select Classifier for SSN

 Pass

Showing 1 of 164 classifiers

Passport Number

Passport number is the unique ID assigned to a travel document, usually issued by the government of a nation, that certifies the identity and nationality of its holder for the purpose of international travel.

Selected: Passport Number

Cancel **Select**

14. The column classification has been changed.

PASSPORT_NUMBER
Type: Integer
Passport Number ▼
240479506
449064032
639218848
212838449
16514173
513428119
742709558
170002642
934214045
3699030
136448540
625553651

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:

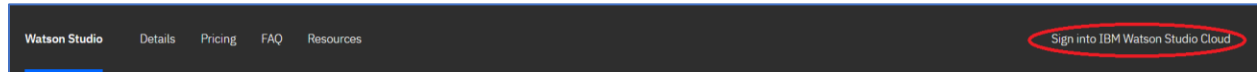
1. Catalog owner (logged on user)
2. E-mail of user provided by lab instructor (used here for illustrative purposes - wsuserbb@gmail.com).

Data Asset Privacy:

1. female_human_trafficking (Private) ACL Members:
 - a. asset owner (logged on user)
 - b. e-mail of user provided by lab instructor (used here for illustrative purposes- wsuserbb@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 wsuserbb@gmail.com, that only 3 Data Assets would appear. The two public Data Assets, and the Private Data Asset where wsuserbb@gmail is on the access control list. This user would not have access to the Occupations asset, 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.

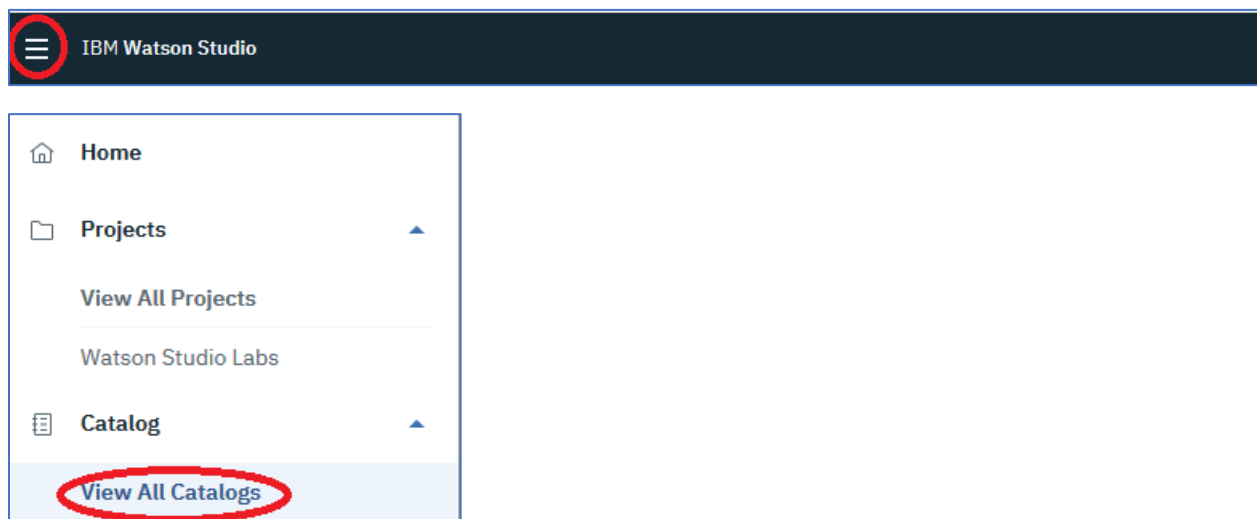
1. Open a different browser (if you are using FireFox, then open Chrome or vice versa).
2. Type in datascience.ibm.com for the URL to navigate to.
3. Click on **Sign into IBM Watson Studio Cloud**



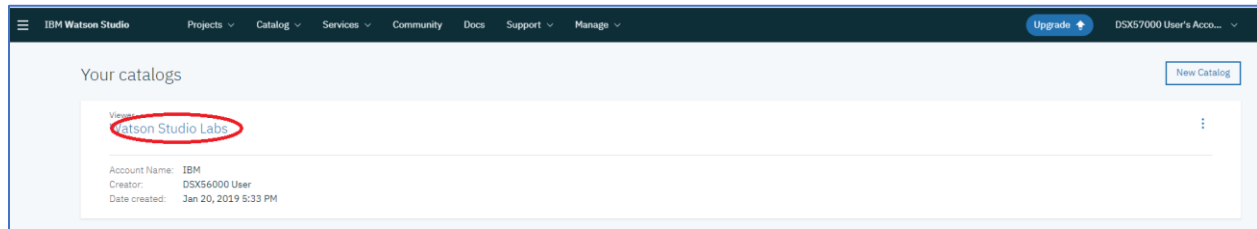
4. Login using wsuserbb@gmail.com for the user name and wsPassw0rd for the password.

A white login page with the heading 'Log in to IBM'. On the left, there are links: 'Don't have an IBMid? Create an account', 'Log in with your company credentials (SSO)', and 'Need help? Contact the IBMid Helpdesk'. On the right, there are input fields for 'IBMId' (containing 'wsuserbb@gmail.com') and 'Password' (containing 'wsPassw0rd'). There are links 'Forgot IBMid?' and 'Forgot password?'. At the bottom right, there is a 'Remember me' checkbox and a blue 'Log in' button. The username, password field, and 'Log in' button are circled in red.

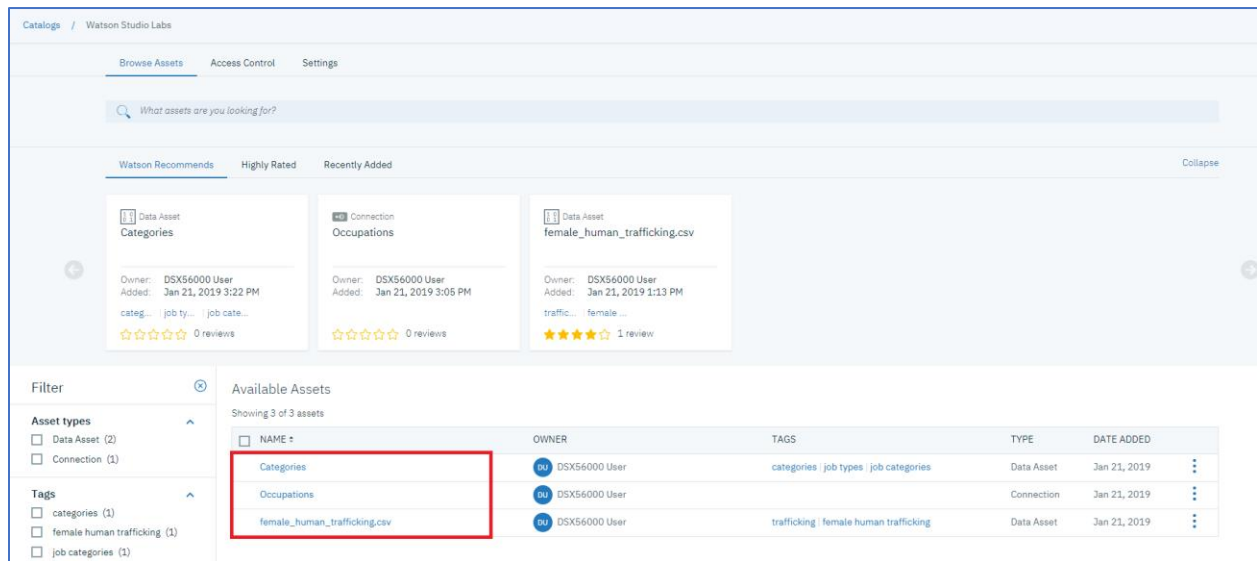
5. Click on  icon. Click on **Catalog** and then **View All Catalogs**



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




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



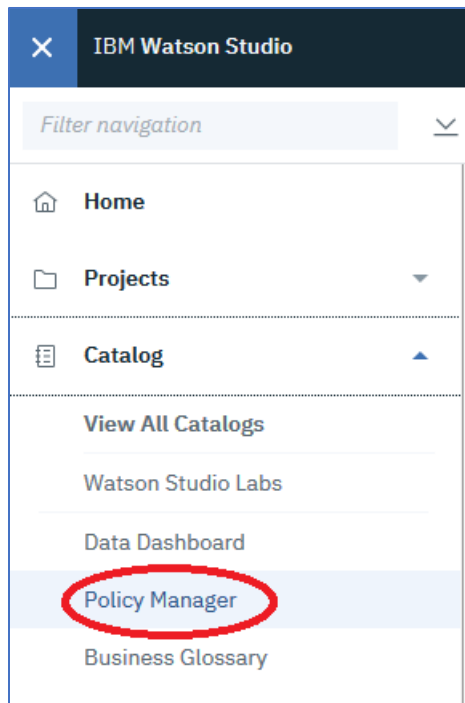
- Leave this browser at this screen for Step 7.

Step 7: Create and Enforce Policy

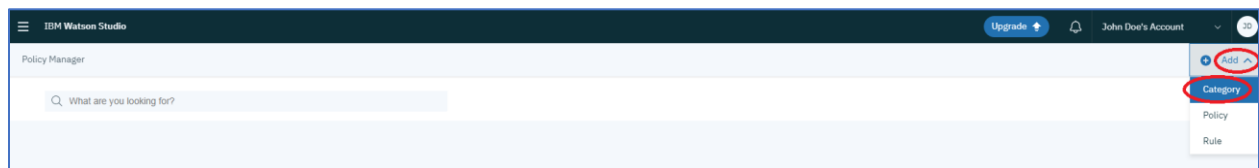
In this step, we will create a policy that will anonymize the SSN and Passport fields when user wsuserbb@gmail.com tries to access the female_human_trafficking data asset.

- Return to the browser where your Watson Studio id is logged in.
- Click on the  icon. Click on **Catalog** and then **Policy Manager**.





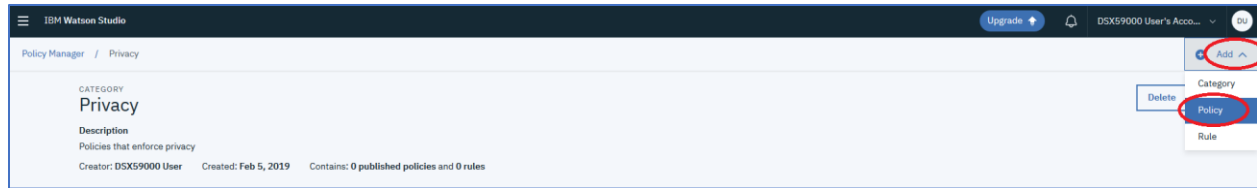
3. Click on **Add** and then **Category**.



4. Type in the Privacy for the **Name** of the Category, and optionally a **Description**, and click on **Create**.

The image shows the 'New category' form in IBM Watson Studio. The form has a 'Details' section with two fields: 'Name*' and 'Description'. The 'Name*' field is labeled 'Privacy' and is highlighted with a red circle. The 'Description' field contains the text 'Policies that enforce privacy' and is also highlighted with a red circle. At the bottom right of the form, there are two buttons: 'Cancel' and 'Create'. The 'Create' button is highlighted with a red circle.

5. Click on **Add** and then **Policy** to add a policy to the Privacy Category.



6. Type in PII for the Policy **Name**, type in “Limit access to PII data” for the Policy **Description**, click on **create a new rule**.

7. Enter “Redact SSN” for the rule **Name**, enter “Redact a Government ID” for the rule **Description**, for the rule select the following:

If **Asset Attribute Informed Classification** contains any **Government Identities**

And

If **User Name** contains any **John Doe**

then **anonymize data** where **in columns containing Government Identities**

Name *
Redact SSN

Type *
Access

Business definition *
Redact a Government ID

Criteria *

CONDITION 1

If Asset Attribute Inferred Classification contains any
Government Identities X

And

CONDITION 2

If User Name contains any
John Doe X

Action *

then anonymize data in columns containing
Government Identities X

8. Select how to anonymize data: **Substitute**, and then click **Create**.

Action *

then anonymize data in columns containing
Personal Information X

Select how to anonymize data:

Redact

BEFORE
452-821-1120

Replace data with Xs.

Substitute

BEFORE
452-821-1120

Replace data with values that don't match the original format.

Mask

BEFORE
452-821-1120

Replace data with similarly formatted values.

Cancel Create

9. Click on **Publish**.

New policy

Details

Name *
PII

Type *
Access

Category *
Privacy

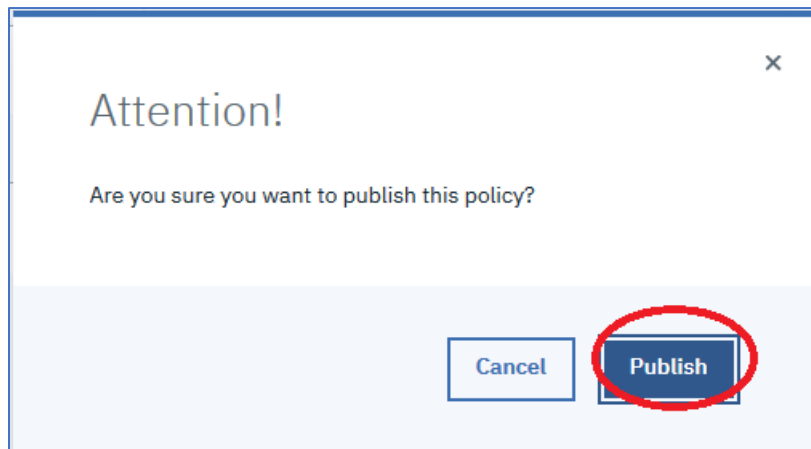
Description
Limit access to PII data

Rules (1)

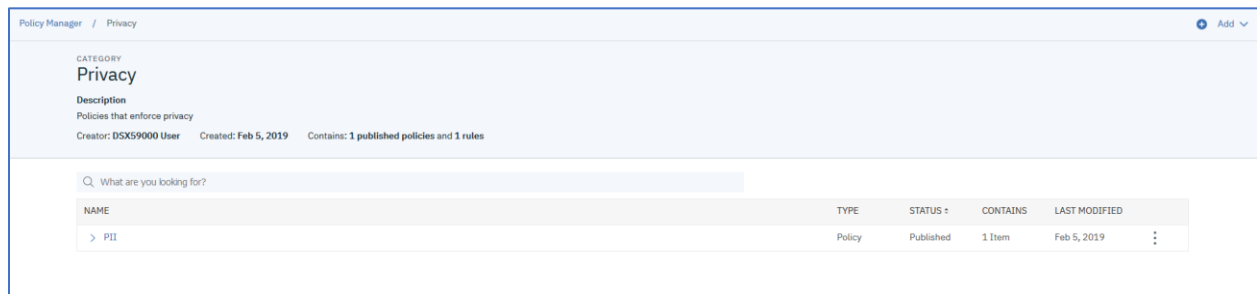
Redact SSN
Redact SSN Column

Cancel Save as draft Publish

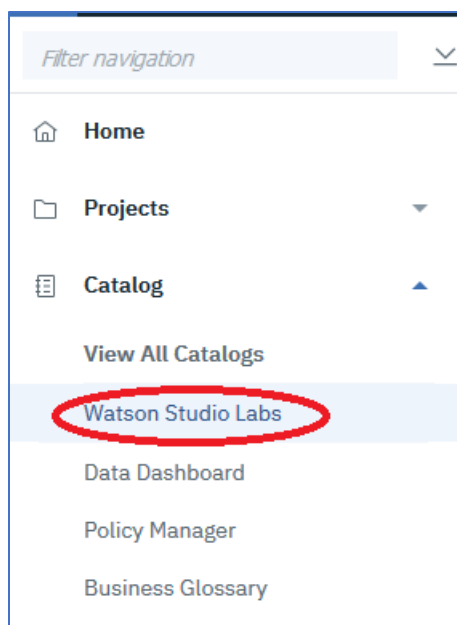
10. Click on **Publish**.



11. The PII policy has been added.



12. Click on the  icon, then click on **Catalog** and then **Watson Studio Labs**.



13. Switch back to the browser with the wsuserbb@gmail.com logged in.
14. Click on the **female_human_trafficking** asset.

Available Assets

Showing 3 of 3 assets

<input type="checkbox"/>	NAME	OWNER	TAGS	TYPE	DATE ADDED
<input type="checkbox"/>	Categories.csv	DSX59000 User		Data Asset	Jan 31, 2019
<input checked="" type="checkbox"/>	female_human_trafficking	DSX59000 User	trafficking female human trafficking	Data Asset	Jan 31, 2019
<input type="checkbox"/>	trafficking	DSX59000 User		Connection	Jan 31, 2019

15. The female_human_trafficking asset is displayed. Note that 2 columns are anonymized.

Data Asset

female_human_trafficking

Description: There is no description available for this asset.

Added: Jan 31, 2019 10:02 AM
Format: application/octet-stream
Size: 347 KB

Tags: trafficking | female human trafficking

Reviews: 0 reviews

Connection: Source: Watson Studio Labs_DataCatalog
Source type: Cloud Object Storage

Classification: Personally Identifiable Information
Personally identifiable information (PII) is defined as any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another can be considered PII.

Schema: 26 Columns | 1085 Rows | 2 Columns anonymized

Preview: 1000 rows | Last refresh: 22 seconds ago | Refresh

DATE	BIRTH_COUNT...	BIRTH_COUNTRY_CODE	OCCUPATION	ADDRESS	SSN	PASSPORT_NUMBER
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
th	Country Name	Country Code	Text	Text	US Social ...	Passport Number
15	Ghana	GH	Engineer, land	824 Kristin Grv, I	afe55d1d355c3:	1c9da91e1e20863dd850
19	Ghana	GH	Editor, commissi	1148 Wang Fall :	77a0daa42ec7d	12d38855ed107e7cc5dd
16	Ghana	GH	Merchant navy ol	9486 Pratt Wall,	669061087d6d1	c43ed0283a3def7031d8:
17	Ghana	GH	Paramedic	0890 Johnson Tr	997b59e501b2e	179abee5ba608418154d
18	Ghana	GH	Surveyor, buildin	2315 Brittany Cr	70329b83b40c:	84524ccc3c5c6590600e
24	Ghana	GH	Waste managem	88811 Donald Ps	d2f2236f52407:	a730ae13f5ed96f71e904
23	Ghana	GH	Doctor, general p	9150 Donald Rpx	d2c2d41163d8f:	ced1617be1d70e44421c
02	Ghana	GH	Forest/woodland	1355 Lopez Villa	62007942c2b0c:	8c8debda401b6b6d954b
12	Ghana	GH	Land/geomatics :	86792 Amy Vlg,	08f8dd9f9ba89:	a43f1d6c9cacdfa82a1a1
10	Ghana	GH	Oncologist	108 Erin Via, Nev	f8b871f6e058e:	f289be62078ebbe457c6:
07	Ghana	GH	Veterinary surger	79572 Schmidt E	f2006c1d30df3:	624a9605774a0cfd98aa

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 Labs** to go to the Catalog home screen.

IBM Watson Studio

Projects Catalog Services Community Docs Support Manage

Catalogs / **Watson Studio Labs**

Browse Assets Access Control Settings

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

Catalogs / Watson Studio Labs

Data Asset

Occupations

Owner: DSX56000 User
Added: Jan 23, 2019 11:11 AM

occup... | jobs

☆☆☆☆☆ 0 reviews

Data Asset

Categories

Owner: DSX56000 User
Added: Jan 21, 2019 3:22 PM

categ... | job ty... | job cate...

☆☆☆☆☆ 0 reviews

Connection

Occupations

Owner: DSX56000 User
Added: Jan 21, 2019 3:05 PM

☆☆☆☆☆ 0 reviews

Data Asset

female_human_trafficking

Owner: DSX56000 User
Added: Jan 21, 2019 1:13 PM

traff... | female ...

☆☆☆☆☆ 1 review

Filter

Asset types

- ☐ Data Asset (3)
- ☐ Connection (1)

Tags

- ☐ categories (1)
- ☐ female human trafficking (1)
- ☐ job categories (1)

Available Assets

Selected 4 of 4 assets

NAME	OWNER	TAGS	TYPE	DATE ADDED	Add to Project	Remove
Categories	DSX56000 User	categories job types job categories	Data Asset	Jan 21, 2019		
Occupations	DSX56000 User		Connection	Jan 21, 2019		
Occupations	DSX56000 User	occupations jobs	Data Asset	Jan 23, 2019		
female_human_trafficking	DSX56000 User	trafficking female human trafficking	Data Asset	Jan 21, 2019		

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

Catalogs / Watson Studio Labs

Add to Project

Target*

Watson Studio Labs

Selected assets (4)

ASSET NAME	CONNECTION
Categories	Occupations
Occupations	Occupations
Occupations	Occupations
female_human_trafficking	Watson Studio Labs_DataCatalog

Connections to be added (1)

The following connections will be added to the project to successfully add the assets.

Occupations
IBM Db2 fully-managed cloud SQL database

Cancel Add

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

IBM Watson Studio

Projects Catalog Services Community Docs Support Manage

Catalogs / Watson Studio Labs

View All Projects

Recent Projects

Watson Studio Labs

Data Asset female_hun

Data Asset Occupations

Data Asset Categories

Connection Occupations

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

IBM Watson Studio

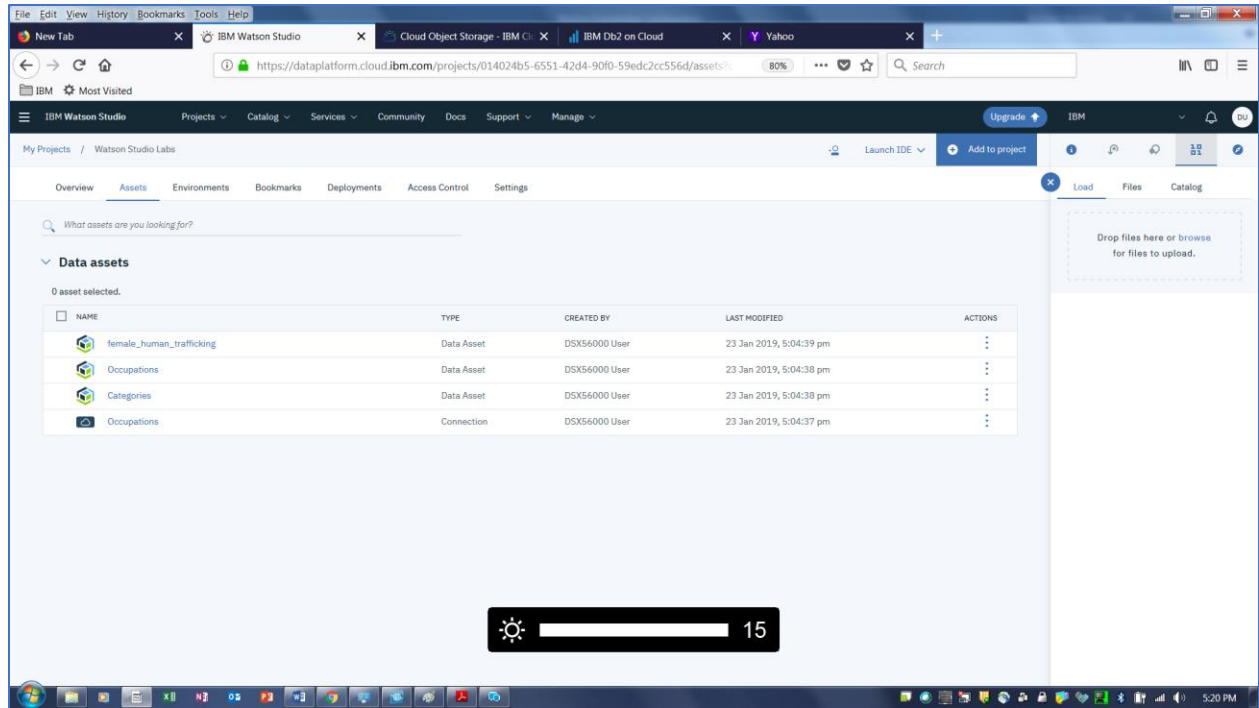
Projects Catalog Services Community Docs Support Manage Upgrade

My Projects / Watson Studio Labs

Launch IDE Add to project

Overview Assets Environments Bookmarks Deployments Access Control Settings

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



▼ Data assets

0 asset selected.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	female_human_trafficking	Data Asset	DSX56000 User	23 Jan 2019, 5:04:39 pm	⋮
<input type="checkbox"/>	Occupations	Data Asset	DSX56000 User	23 Jan 2019, 5:04:38 pm	⋮
<input type="checkbox"/>	Categories	Data Asset	DSX56000 User	23 Jan 2019, 5:04:38 pm	⋮
<input checked="" type="checkbox"/>	Occupations	Connection	DSX56000 User	23 Jan 2019, 5:04:37 pm	⋮