Lab Guide

Hands-on-Lab

Watson Knowledge Catalog powered by Cloud Pak for Data

Shivam Solanki

Data Scientist

Shivam.raj.solanki@ibm.com

IBM

IBM Watson Knowledge Catalog powers intelligent, self-service discovery of data, models and more, activating them for artificial intelligence, machine learning and deep learning. Access, curate, categorize and share data, knowledge assets and their relationships wherever they reside.

Tutorial

In this tutorial, you will explore the following key capabilities:

- Creating a Governed Knowledge Catalog
- Discovering and Cataloging Data Assets
- Understanding and Socializing Data Assets
- Shopping for Data
- Preparing Data for Analytics and AI
- Protecting Sensitive Information

Introduction

In the insurance industry, claims processing is an area with many inefficiencies and risks to the insurance provider. A significant amount of the risk involved lies in the amount of time it takes to process a claim. The more time required to make the required adjustments, the higher the risk of lawsuits, which are a costly outcome.

Processing insurance claims is an expensive, time consuming and risk-intensive process. Challenges around claims processing become especially intense during natural calamities, when insurers need to process a sudden spike in claims, even to the point of transporting adjusters to the impacted location.

With a data-driven approach, the information gathering process can be expedited tremendously with immediate access to relevant information at the first notice of loss. The use of data analytics and AI can help identify potential claim fraud using machine learning and detailed data analysis.

Using information that's available in the insurance company's enterprise Knowledge Catalog, the business can easily develop a data-driven claims process that:

- Reduces the median time for a claim to be processed.
- Minimizes the risk of fraud.
- Automates as much of the claims and adjustment process as possible, while triaging more complex claims for adjusters to process.

In this use case, the insurance company's goal is to create a dashboard for a claims agent to interact with the information pushed up to the insurance company from the customer's mobile app. To help mitigate the fraud potential of remotely adjusting auto insurance claims, the customer is prompted for details of their vehicle and claim to validate that they are making a legitimate claim. Using IBM Cloud Pak for Data, the business can easily prepare a machine model to assess the authenticity of the claim. The app also needs the customer's account, logistics, policy and claims information to validate that the claim matches the vehicle under the policies coverage.

This tutorial introduces you to the intelligent and collaborative capabilities of the IBM Watson Knowledge Catalog, and the integrated, common fabric of IBM Cloud Pak for Data. These

offerings empower the insurance company's business analysts, data scientists and data professionals to quickly and easily discover, curate, catalog, shape and share data assets in preparation for the analytics and AI processes that will help them achieve their business goals.

Prerequisites

Download Unstructured Files

In the **Discover and Catalog Data Assets** task, you are instructed to add two files to a new Knowledge Catalog using the **Local files** method. You need to download the files, to your desktop or local file system, from this Tutorial Files Box folder and remember where you placed them. Do this **now** before you proceed.

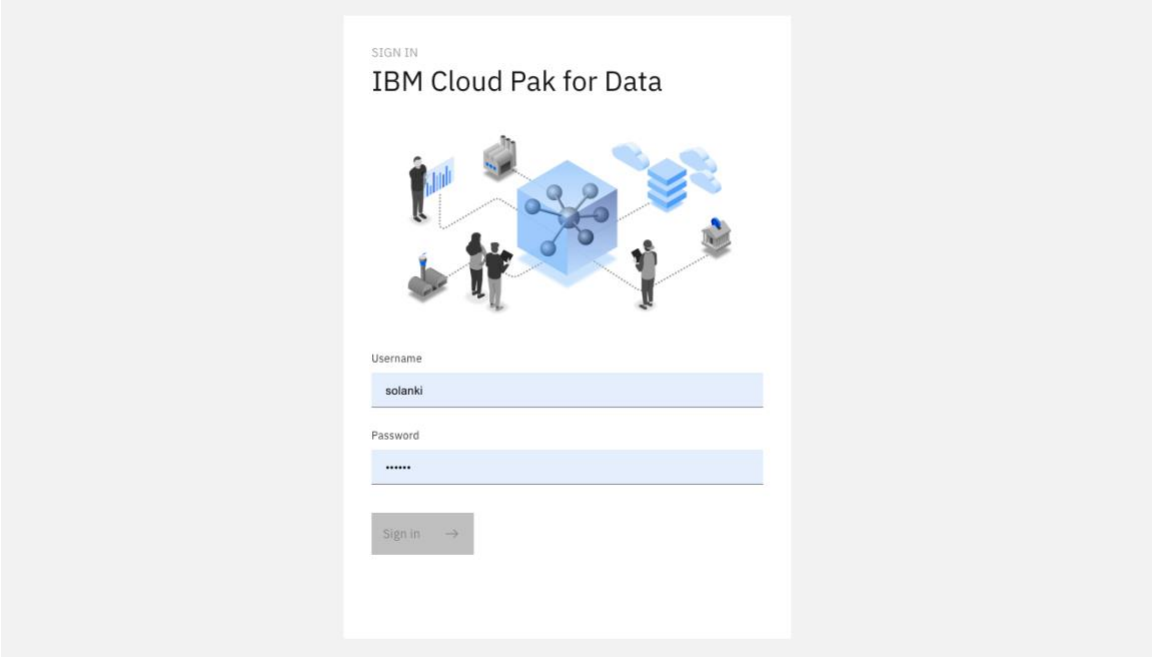
Create an IBM Cloud Pak for Data User

The instructor has already added you and assigned you to all the available roles. This provides you the authority you need to complete the lab and an isolated account to only view what you create. This will shield you from work being done by other users doing the lab in the same environment.

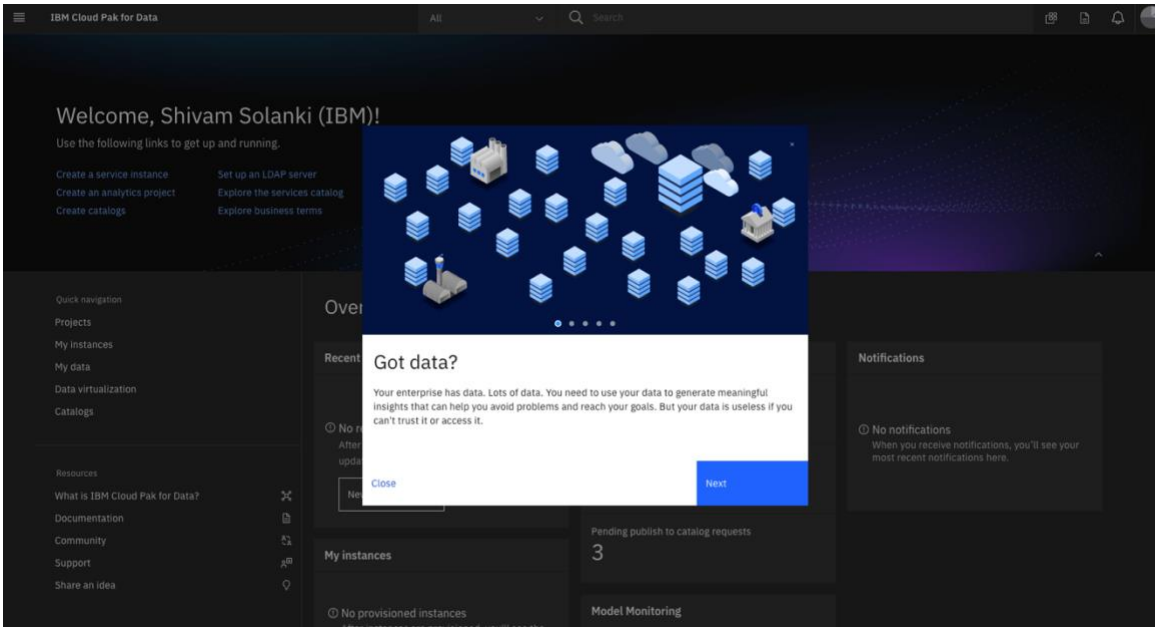
Log in to Cloud Pak for Data

In this section you will log into Cloud Pak for Data using the credentials of the **new** user you just created in the previous step to do the lab.

1. Enter your Last Name as the **Username**. Enter your First Name as the **Password**.



2. Click the **Sign in** button.



You will be brought into the IBM Cloud Pak for Data welcome page.

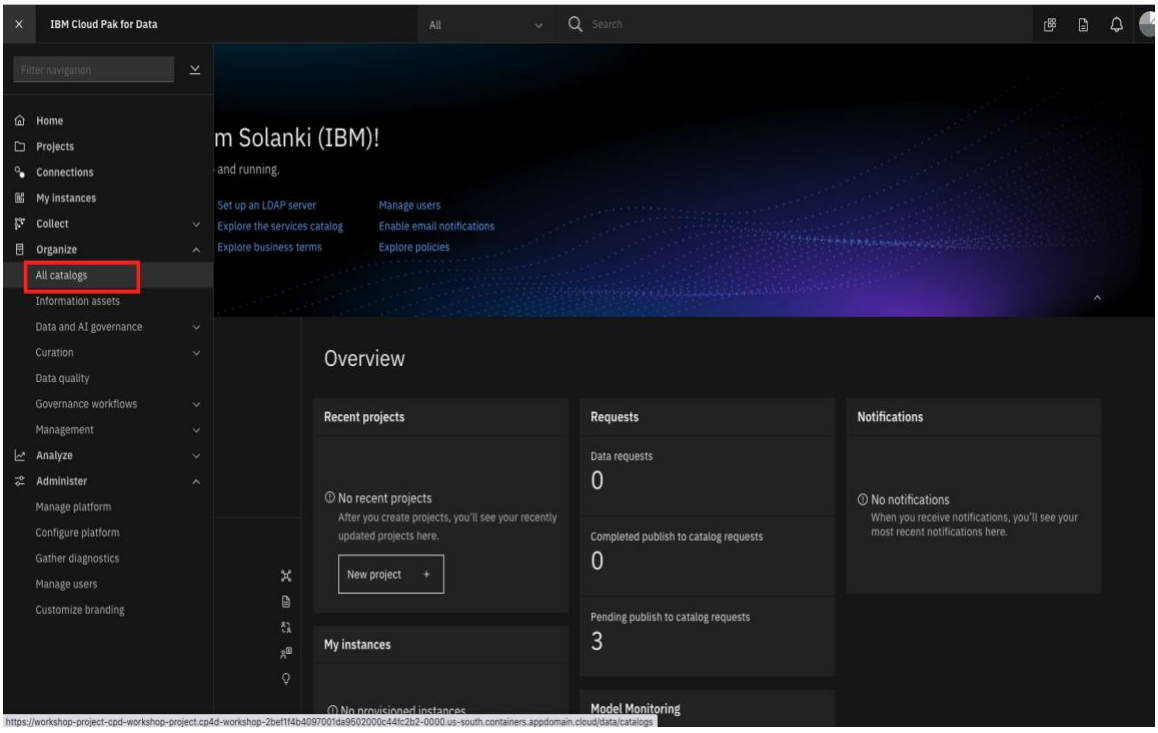
Create a Governed Catalog

In this task, you will create a governed **Knowledge Catalog**. Watson Knowledge Catalog is a secure and collaborative catalog of metadata used to organize and govern information assets. It is tightly integrated with the global business glossary of data governance artifacts that describe and govern the information managed by the catalog, providing self-service capabilities for data professionals to quickly and easily search, find, understand and use data.

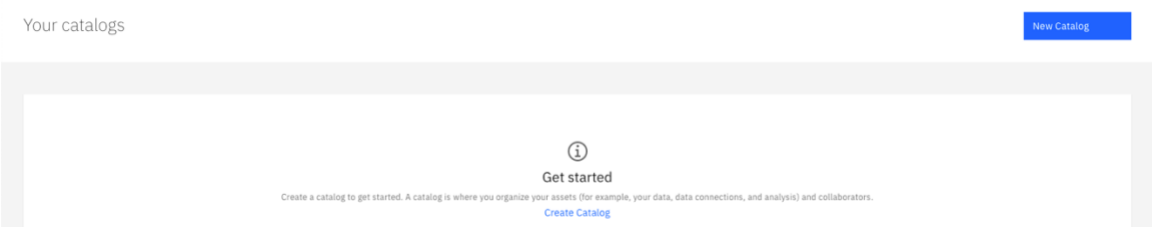
A **Default Catalog** is provided out of the box. However, organizations can create as many catalogs as they need. In this lab, you will create an additional catalog to house the **Auto Insurance** claims fraud analysis information assets that will be used by the analytic project team.

You will learn how to discover, curate and catalog data assets using an additional catalog other than the **Default Catalog** and by using some alternative methods. It will still have integration to the global business glossary and the business policies and rules to govern and protect it.

1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.
2. Select **Organize** → **All Catalogs**.

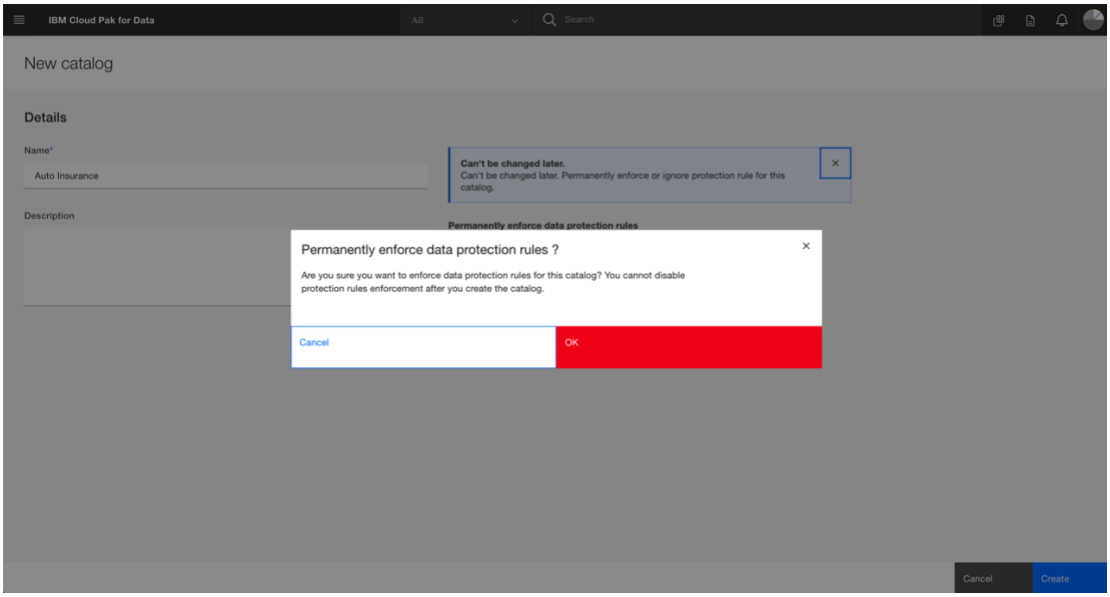


3. Click the **New Catalog** button in the top right corner.



4. Enter a Name of **Auto Insurance**.
5. Enter a Description of **Auto Insurance Knowledge Catalog**.
6. 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.

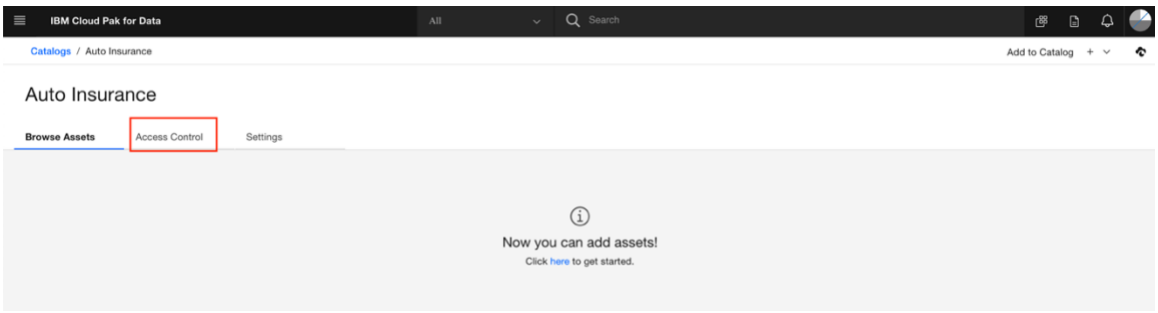


7. Click the **OK** button.

   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 protection rules to restrict access to data based on the contents of the data. These rules 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 protection rules to allow or deny access to a data asset or mask, substitute and redact data at the data asset field level.

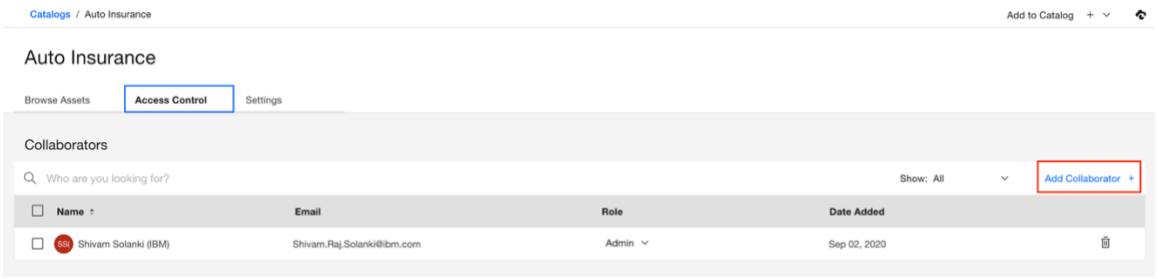
Setting this option for a catalog is a good best practice. Once it is enabled, it cannot be undone, but it does not restrict or impede any functionality, it provides additional security measures to protect data assets.

8. Click the **Create** button. You will see a **Creating Auto Insurance** notification during catalog creation.

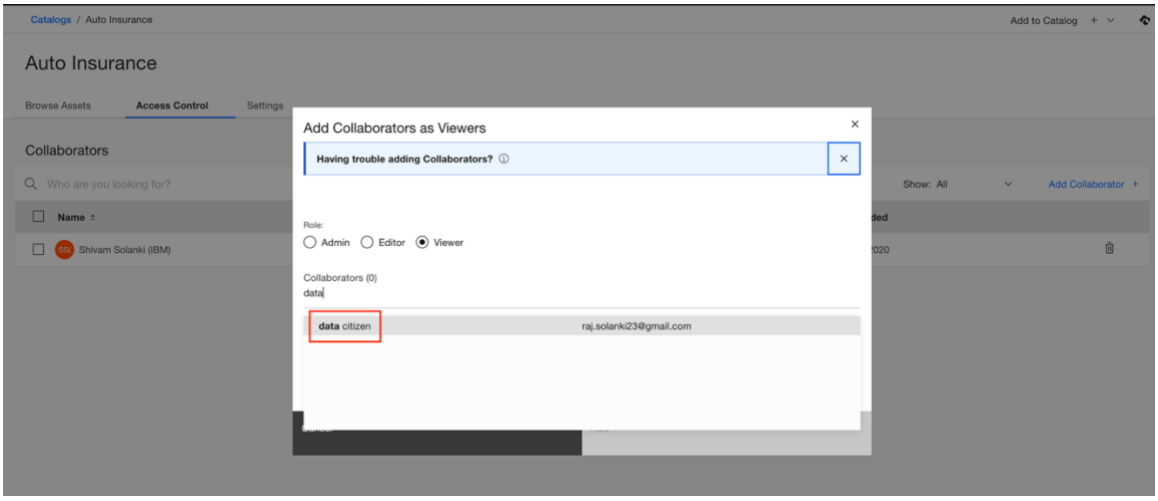


Once completed, you are brought into the newly created catalog. You will now add the **data citizen** user to the catalog as a *Viewer* so they can access the new catalog and use the data assets. You will log in as this user at the end of the lab to see how data protection rules are enforced.

9. Click the **Access Control** tab.

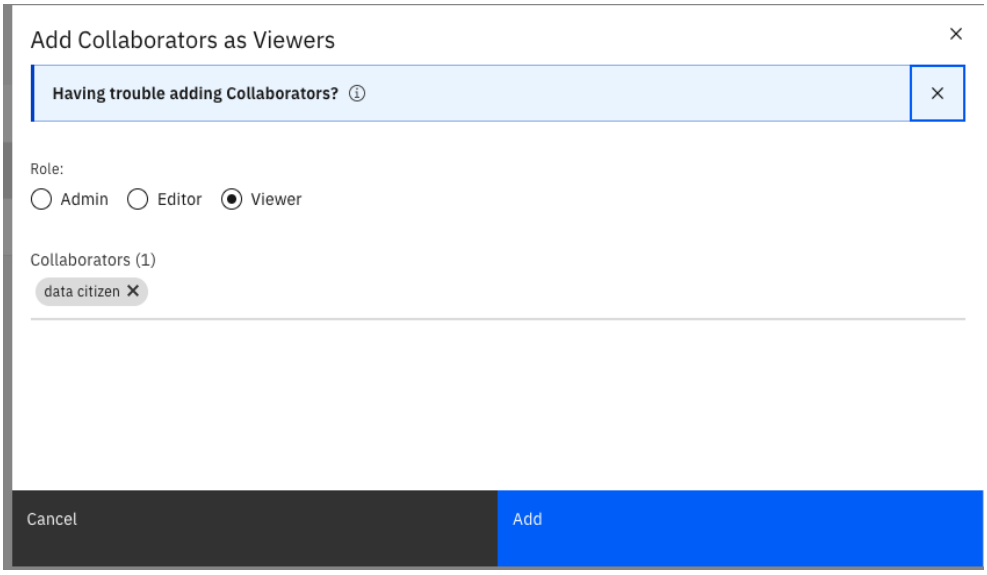


10. Click the **Add Collaborator** button.

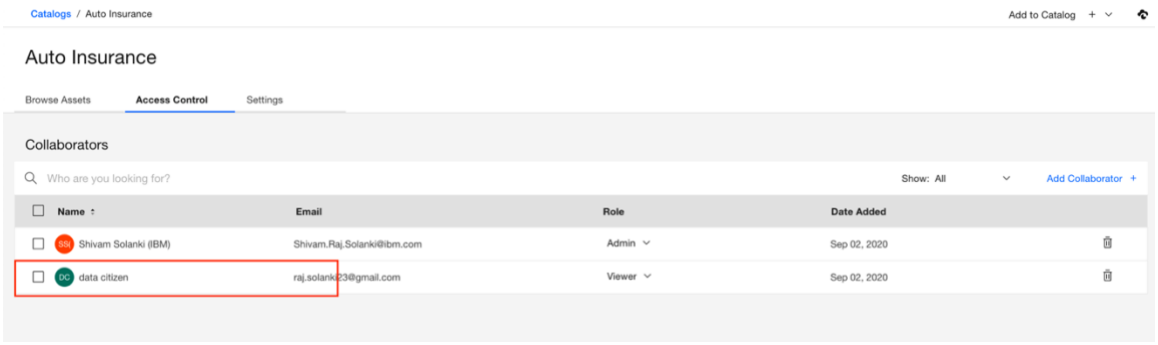


11. Type the word **data** in the search area.

12. Click on the **data citizen** user. The default role of Viewer is automatically assigned. Leave the role set to Viewer.



13. Click **Add**.



You should now see the **data citizen** user added as a **Viewer**.

Create Analytic Projects

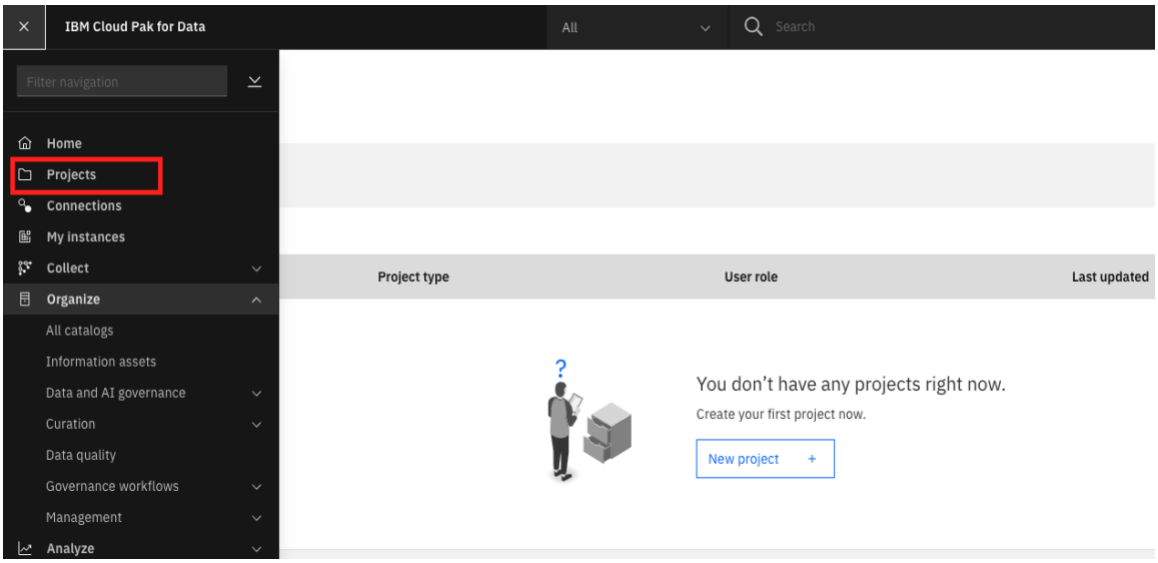
In this task, you will create two Cloud Pak for Data analytics projects.

The first project, which will be named **Auto Insurance**, will be used by the auto insurance analytics team to collaborate and build the analytic and AI assets, notebooks, models, data flows, dashboards, etc. to analyze the auto insurance claims process. You will add auto insurance data assets from the Auto Insurance knowledge catalog to this project and do some shaping of the data using the data refinery to prepare the data for analytical insights.

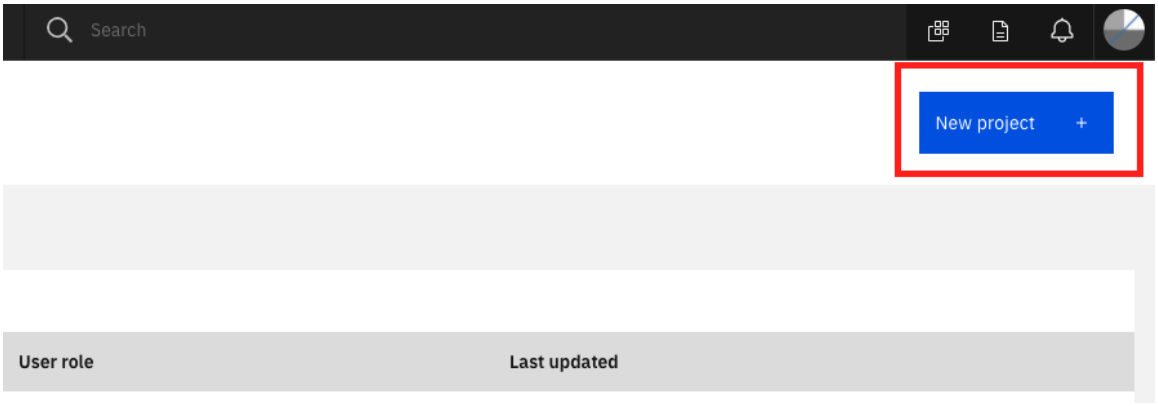
The second project, which will be named **Auto Discovery**, will be used to demonstrate the auto discovery capabilities of Watson Knowledge Catalog. When you catalog a **Connection**, you can choose the option to automatically discover data assets. The discovered assets are added to a Cloud Pak for Data analytics 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 catalog.

Create the Auto Insurance Project

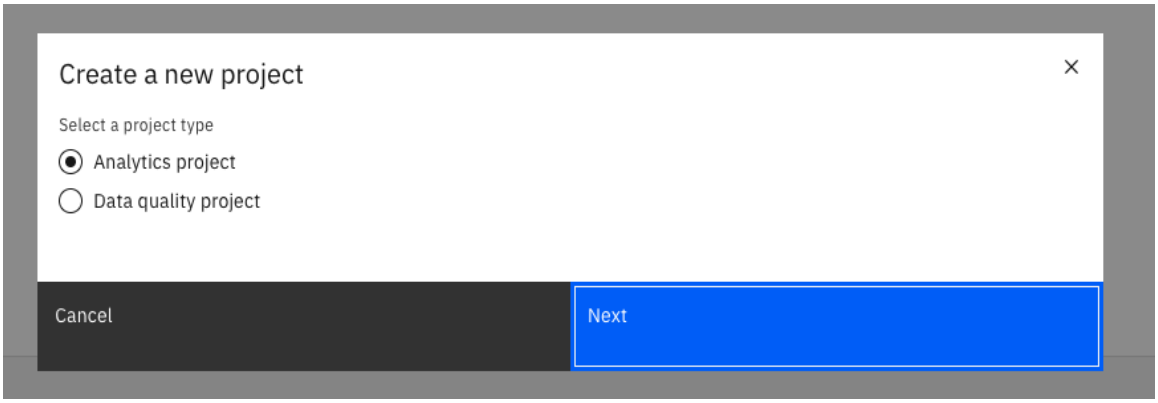
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.
2. Click the **Projects** menu.



3. Click the **New project** button.

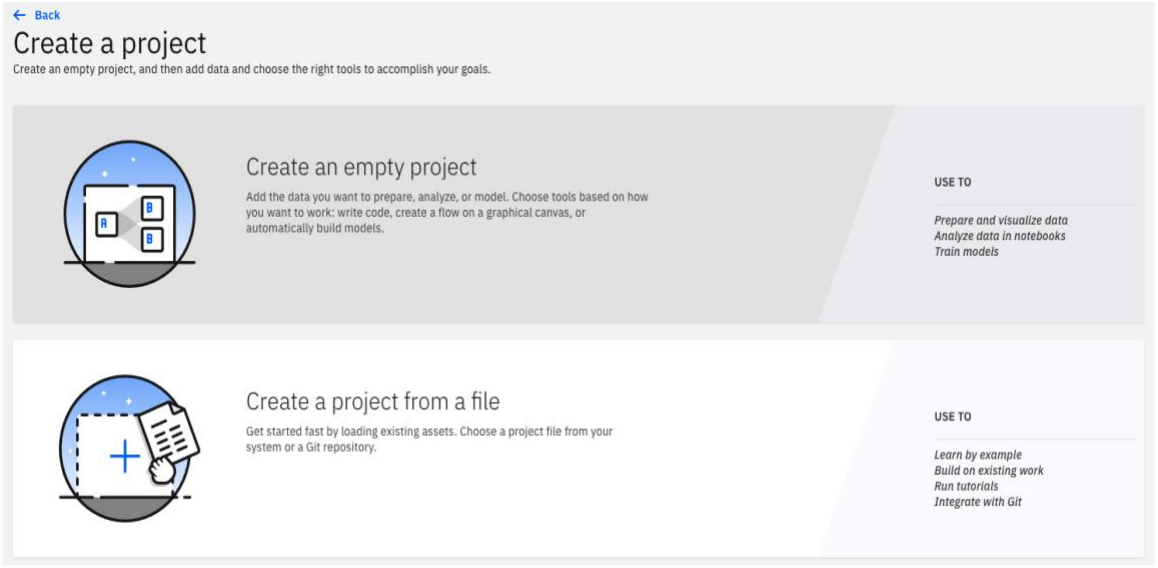


4. Click the **Analytics project** radio button (usually selected by default).

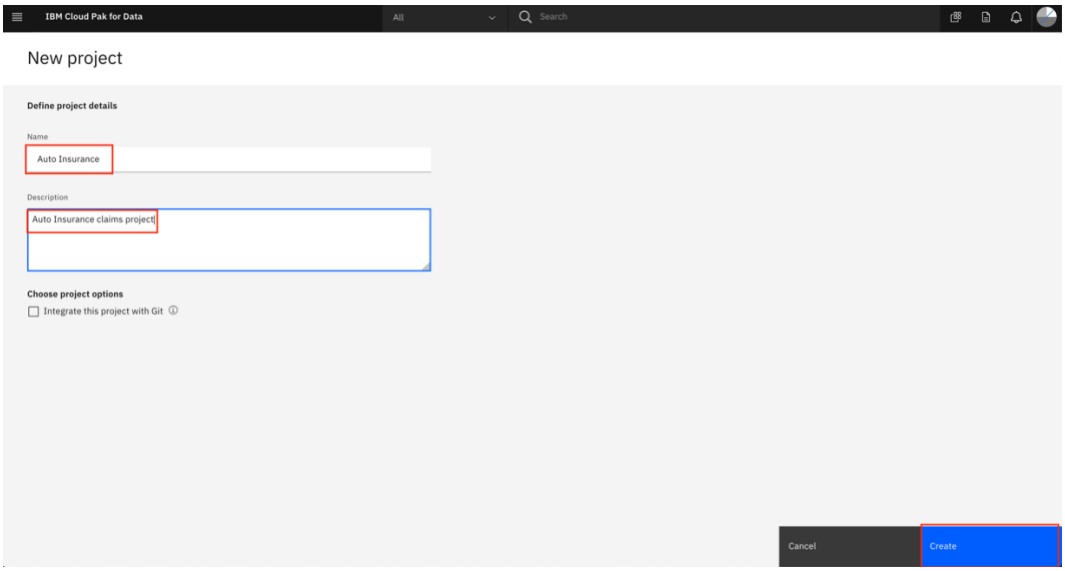


5. Click the **OK** button.

6. Click on **Create an empty project**.

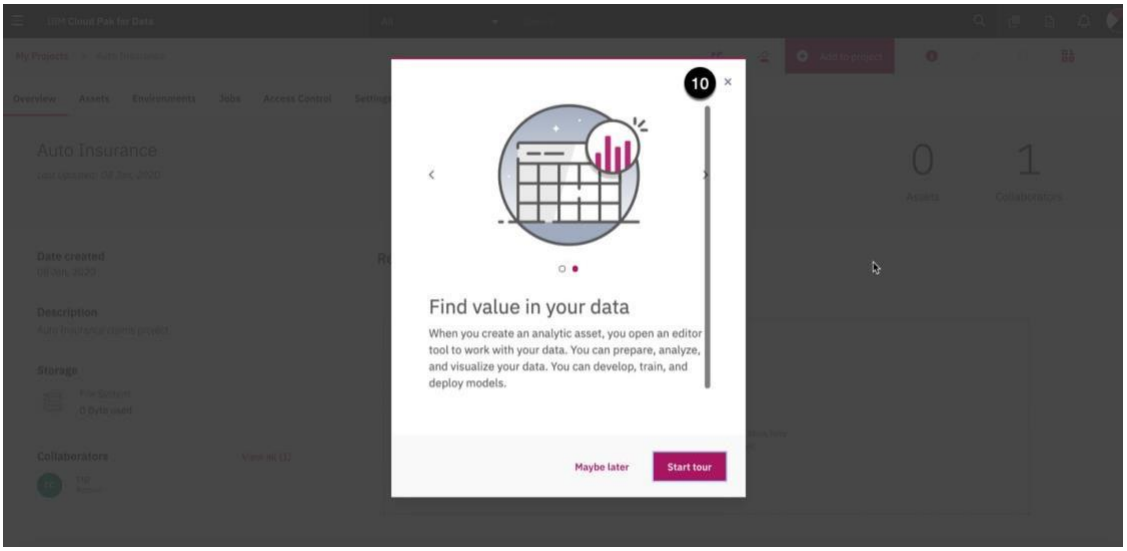


7. Enter a Name of **Auto Insurance** and a Description of **Auto Insurance claims project.** Click the **Create** button.

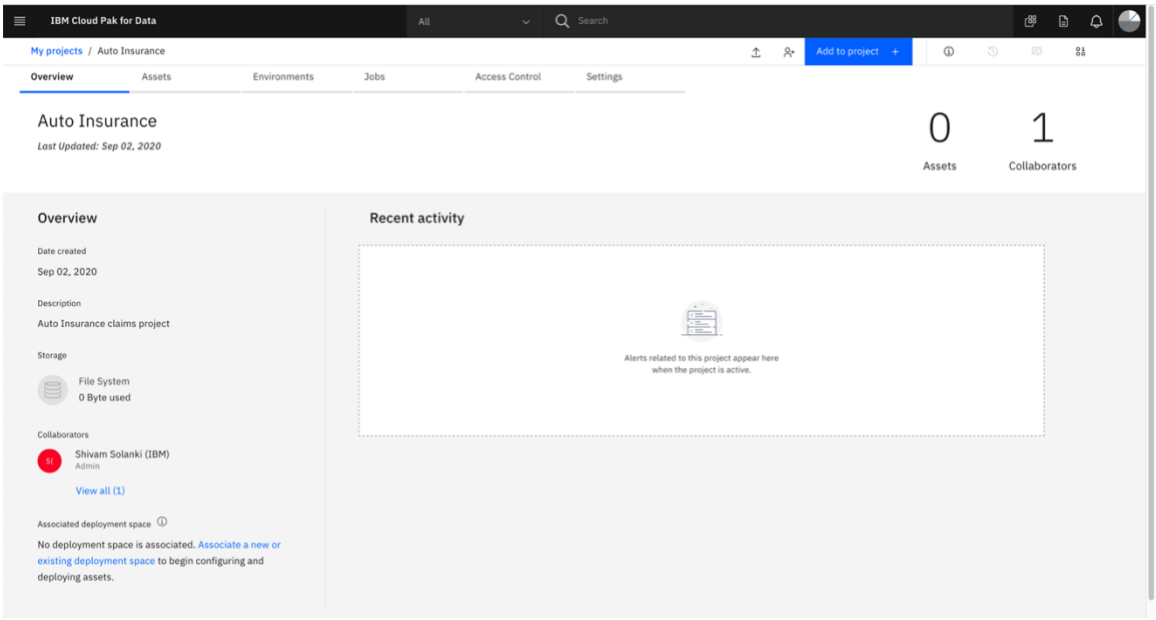


The Create button will turn to **Creating…** so be patient and wait for the project to be created .

If the Getting Started tour dialog appears, click on the **X** in the top right corner to close it.

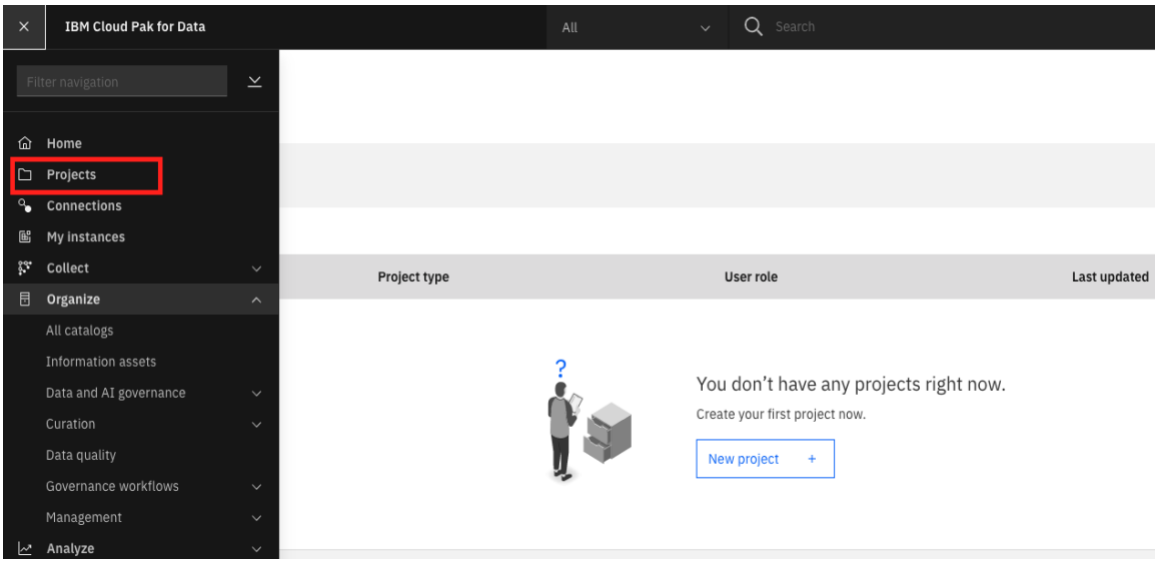


When the project creation is complete, you are brought into your newly created project and you will see the **Overview** section.

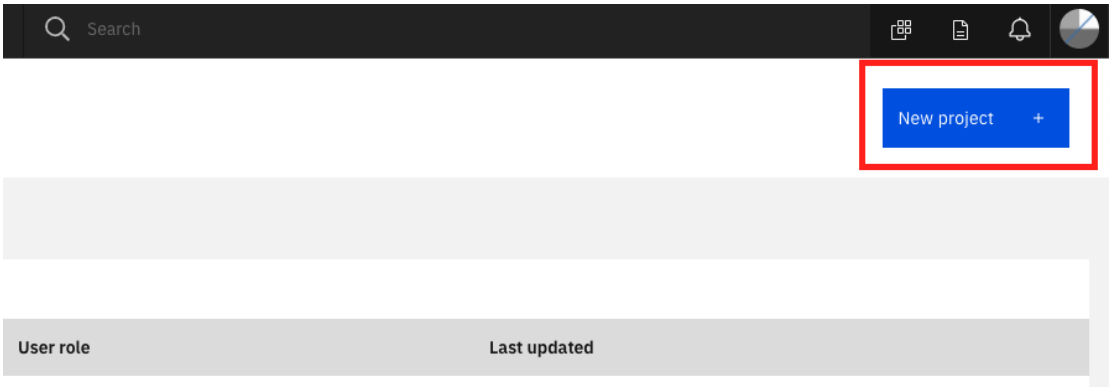


Create the Auto Discovery Project

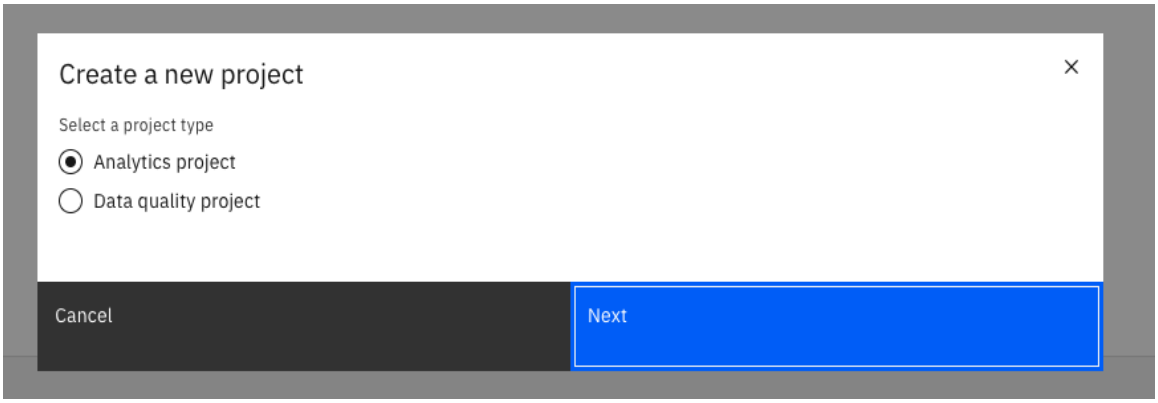
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.
2. Click the **Projects** menu.



3. Click the **New project** button.

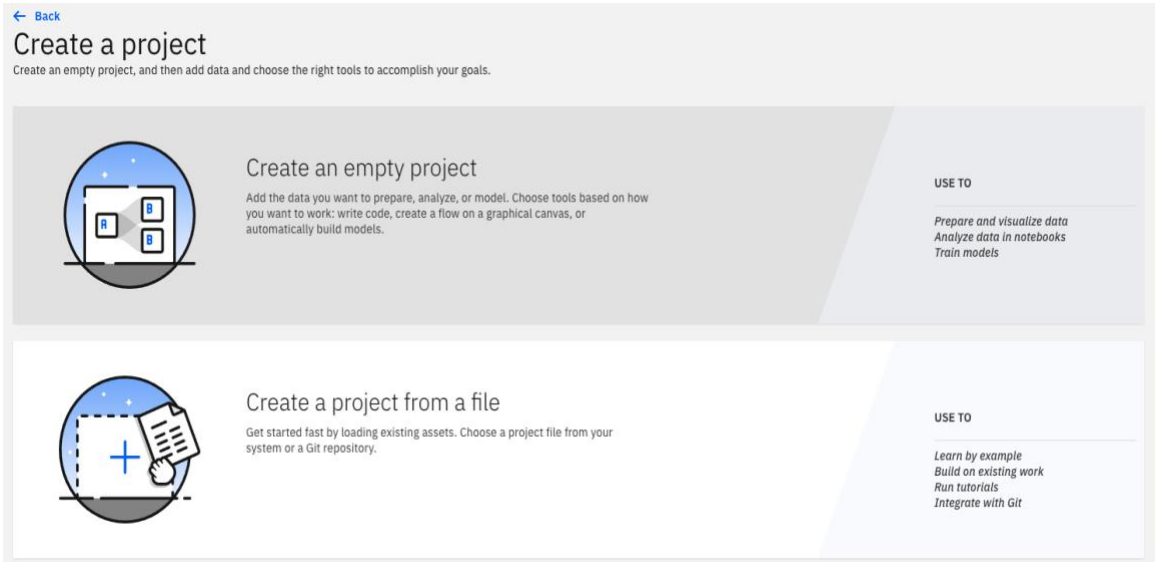


4. Click the **Analytics project** radio button (usually selected by default).

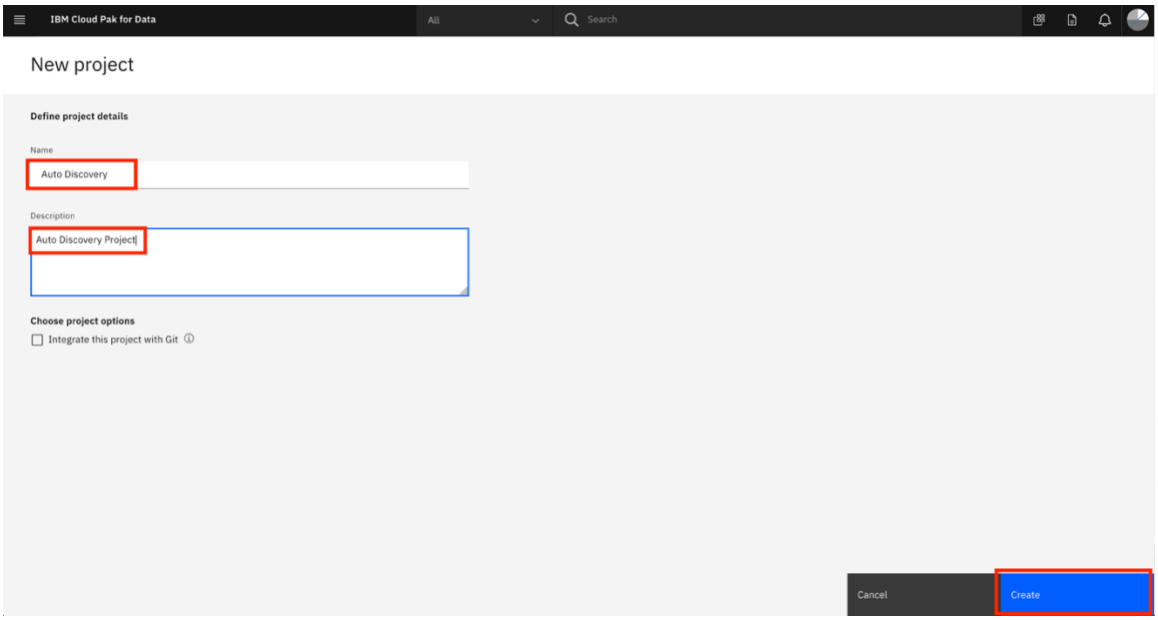


5. Click the **Next** button.

6. Click on **Create an empty project**.

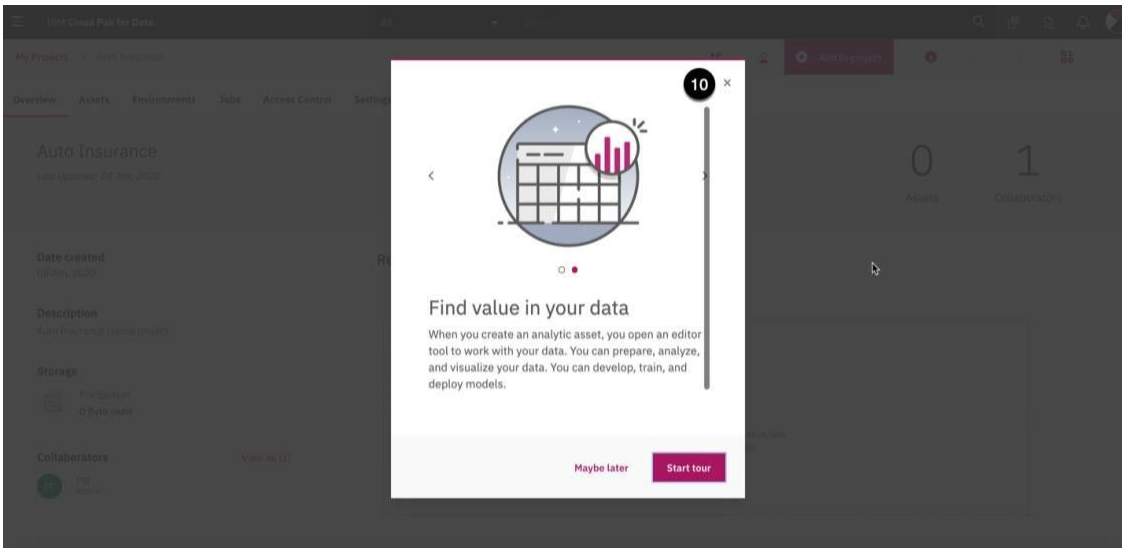


7. Enter a Name of **Auto Discovery**, and a Description of **Auto Discovery project**. Click the **Create** button.

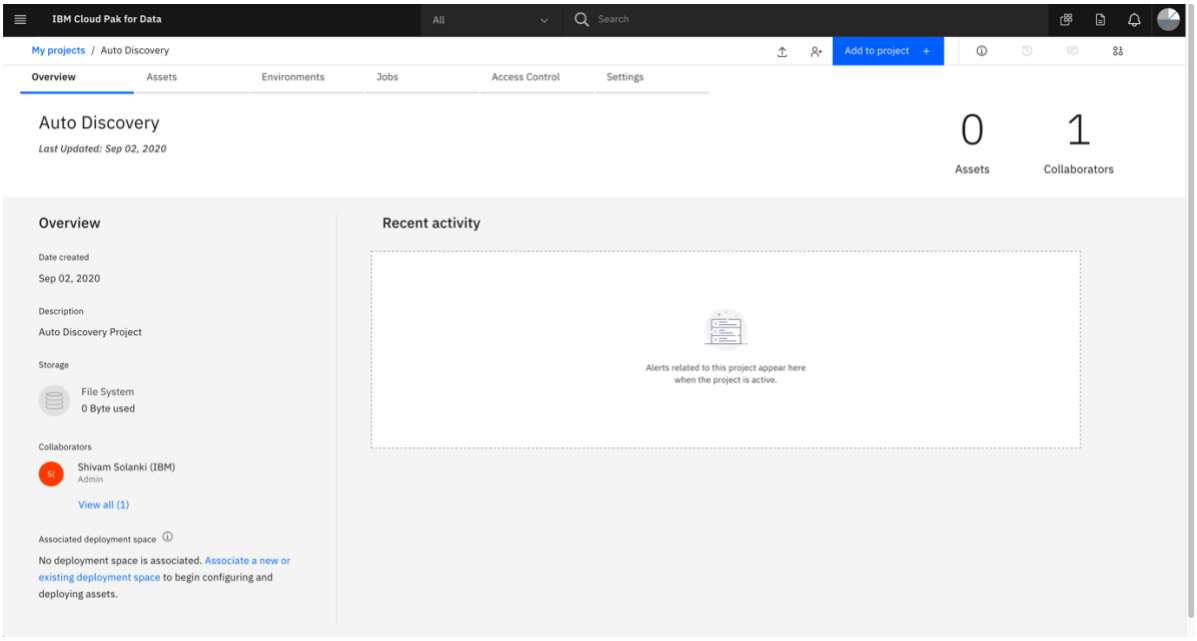


The Create button will turn to **Creating…** so be patient and wait for the project to be created.

If the Getting Started tour dialog appears, click on the **X** in the top right corner to close it.



When the project creation is complete, you are brought into your newly created project and you will see the **Overview** section.

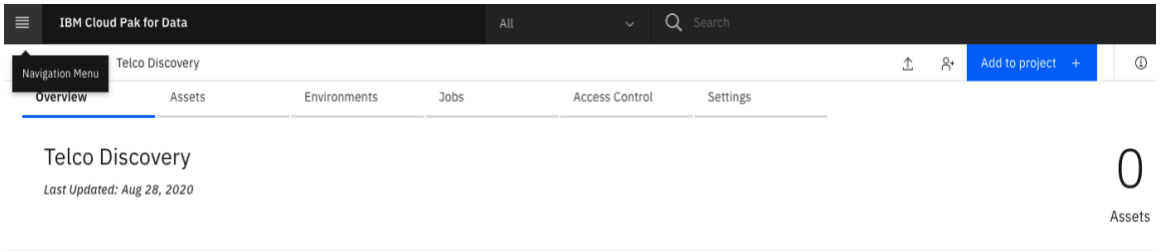

IBM Cloud Pak for Data
All
Search
My projects / Auto Discovery
Add to project +
Overview
Assets
Environments
Jobs
Access Control
Settings
Auto Discovery
Last Updated: Sep 02, 2020
0
Assets
1
Collaborators
Overview
Date created
Sep 02, 2020
Description
Auto Discovery Project
Storage
File System
0 Byte used
Collaborators
Shivam Solanki (IBM)
Admin
View all (1)
Associated deployment space
No deployment space is associated. Associate a new or existing deployment space to begin configuring and deploying assets.
Recent activity
Alerts related to this project appear here when the project is active.

Discover and Catalog Data Assets

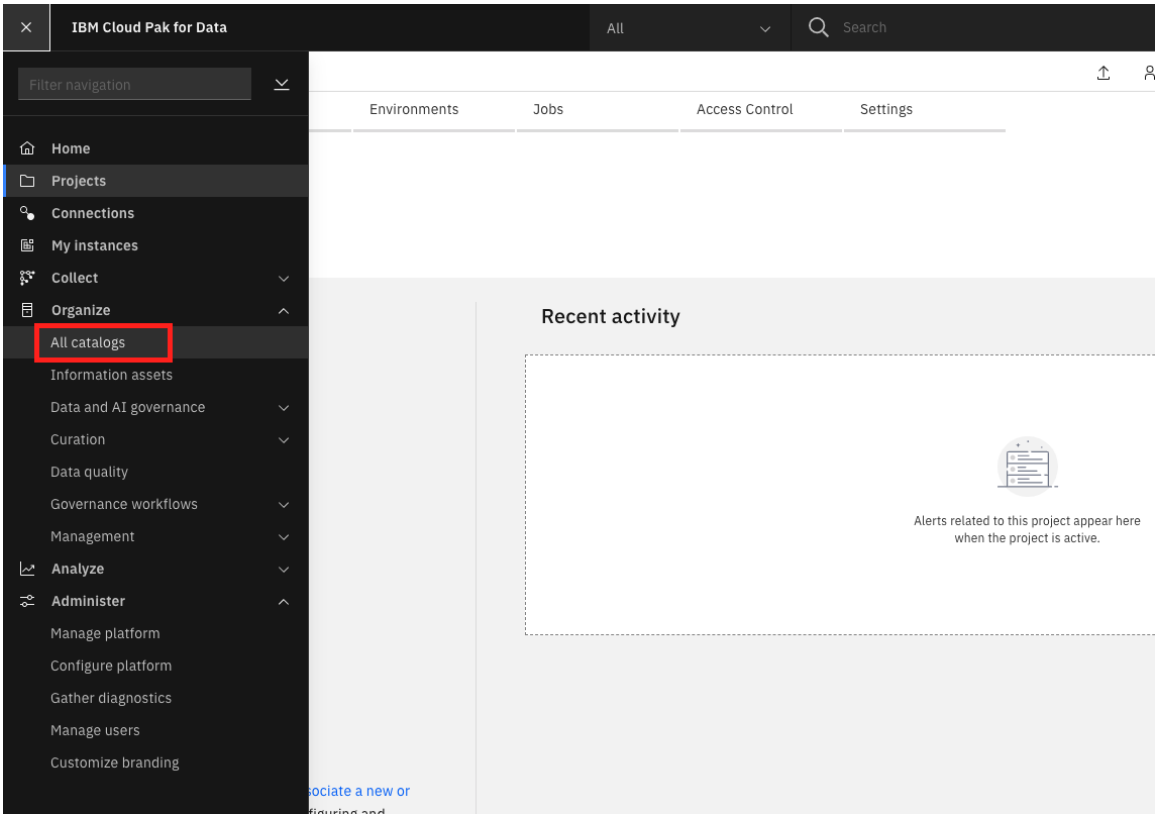
In this task, you will discover and catalog unstructured data assets from the local file system and structured data assets from a **Db2 Warehouse on Cloud** connection that you will create. This will introduce you to the three methods available to discover and catalog data assets; **Local files**, **Connected 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 IBM Cloud Pak for Data, for use during data preparation and within models, dashboards and notebooks.

Catalog Unstructured Data

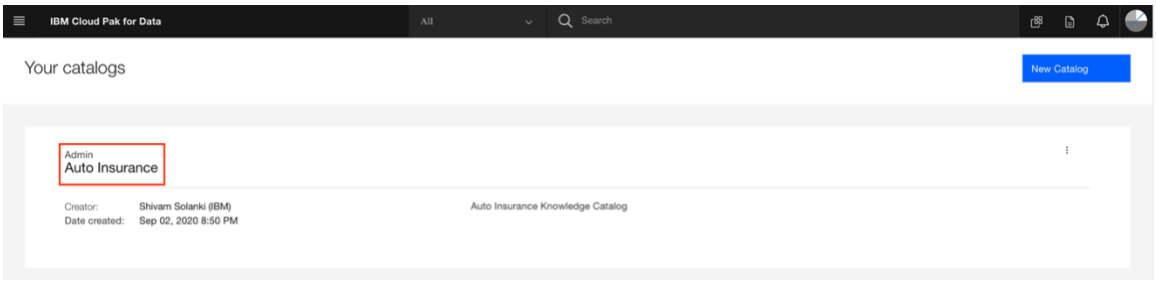
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.



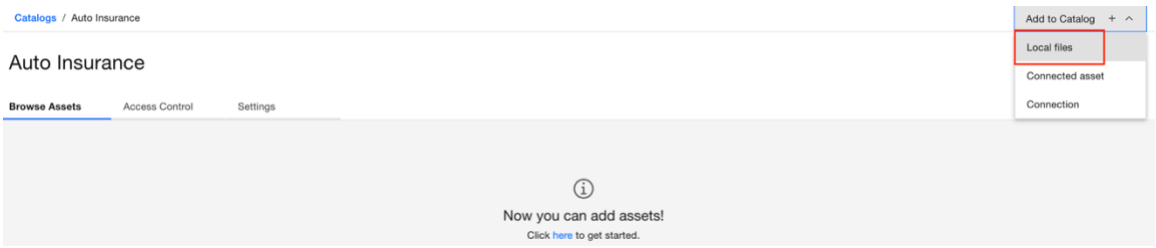
2. From the **Organize** section, select the **All catalogs** menu.



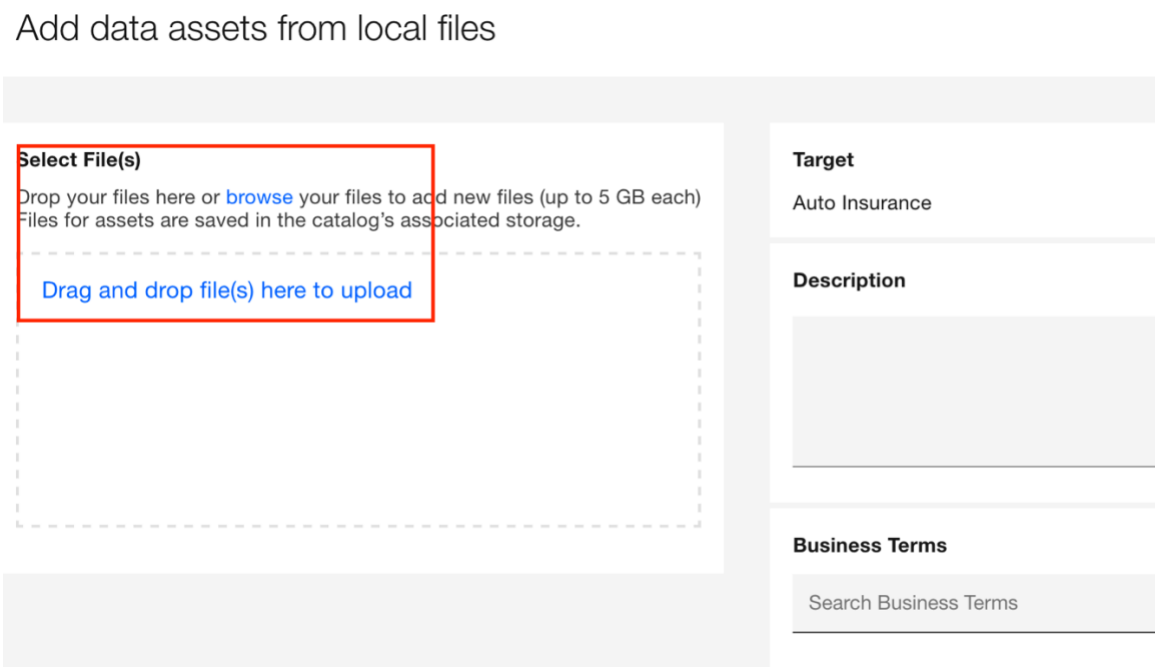
3. Click the **Auto Insurance** catalog.



4. Click **Add to Catalog** → **Local files** from the catalog menu.

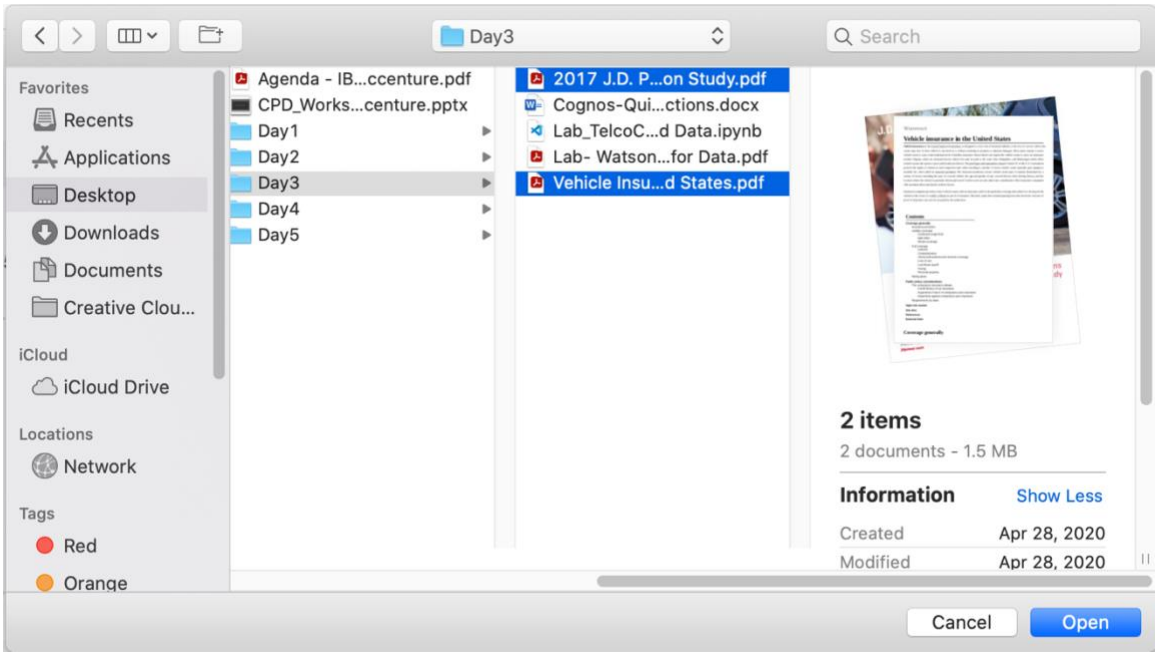


5. Click the **browse** link in the **Select File(s)** section to bring up the file selection dialog.



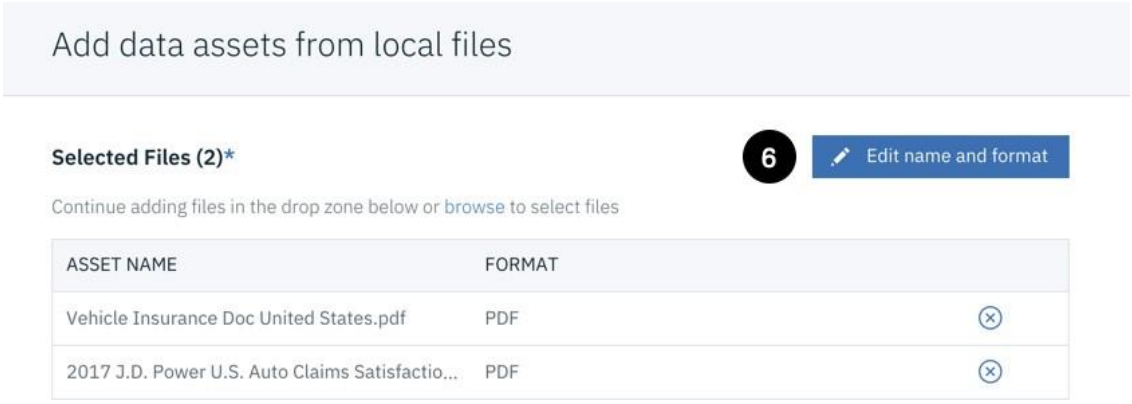
6. The MacOS **Finder** dialog is displayed. If you have a Windows system, it will look different, so depending on what system you are running, adjust to your system's method of selecting files.

   Locate the **"Vehicle Insurance Doc United States.pdf"** and the **"2017 J.D. Power U.S. Auto Claims Satisfaction Survey.pdf"** files on your file system that you were instructed to download. Select both using the **Ctrl or Command key** on your keyboard (CTRL Click for Windows and Command Click for MacOS).



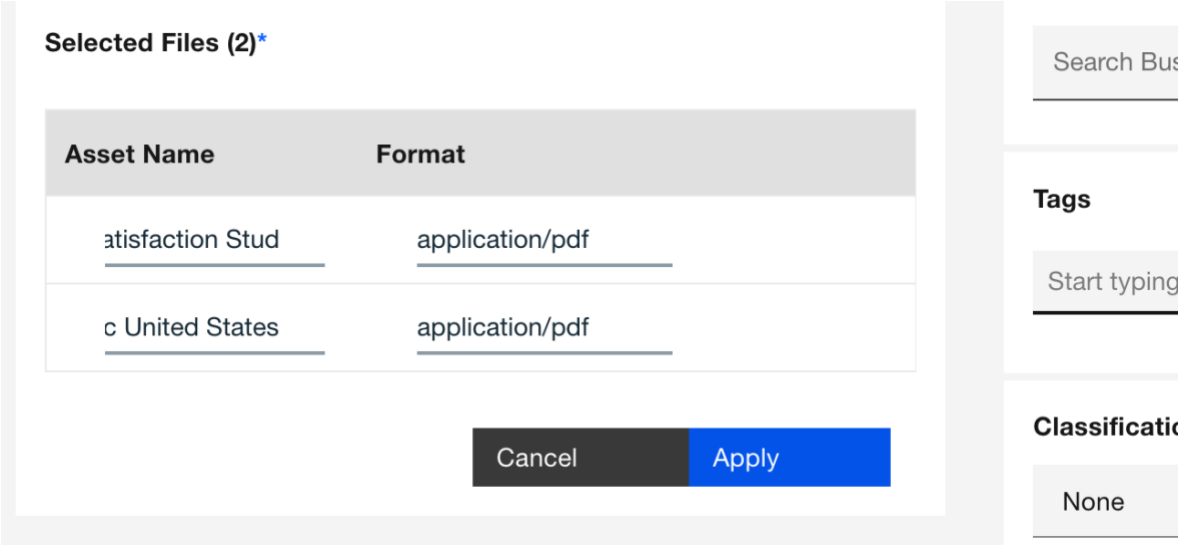
7. Click the **Open** button to begin cataloging the files.

8. Click the pencil icon next to the **Edit name and format** button.



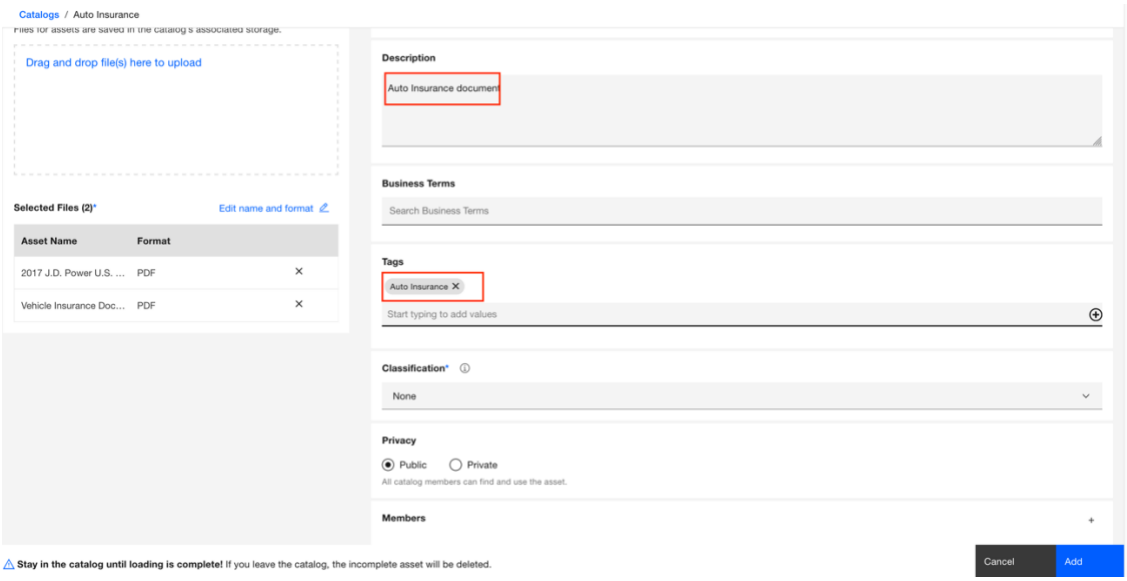
This allows you to rename the data assets and change their file format. A default file format is inferred for you based on the file extension. In this case, they are PDF files, so **PDF** was auto selected. You **will not** change the format, but you will change their names by removing the file extension.

9. Click in the **Asset Name** area of the **“Vehicle Insurance Doc United States.pdf”** file. Go to the end of the filename and remove the **.pdf** extension.



10. Click in the **Asset Name** area of the **“2017 J.D. Power U.S. Auto Claims Satisfaction Survey.pdf”** file. Go to the end of the filename and remove the **.pdf** extension.

11. Click the **Apply** button to save the filename changes.



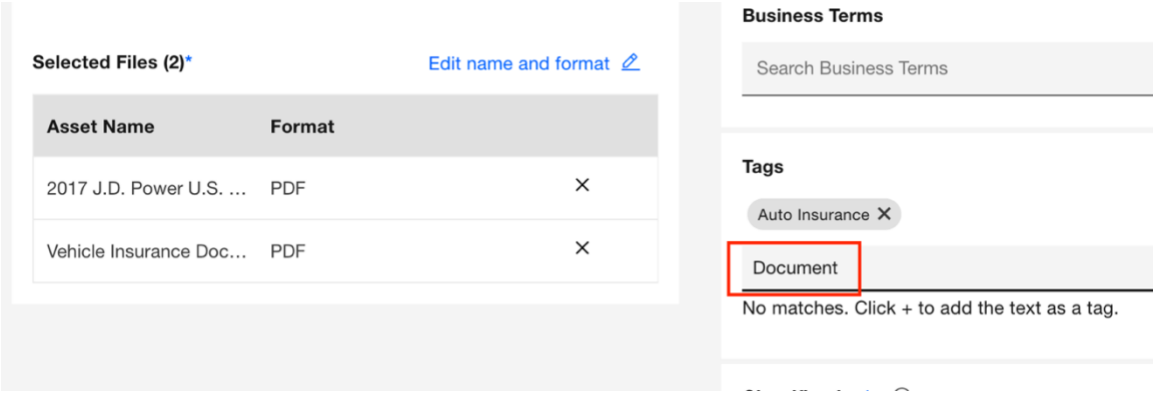
12. Enter a Description of **Auto Insurance document**.

13. Enter a Tag of **Auto Insurance** into the **Tags** area. Click the **+** sign next to the tag to add it.

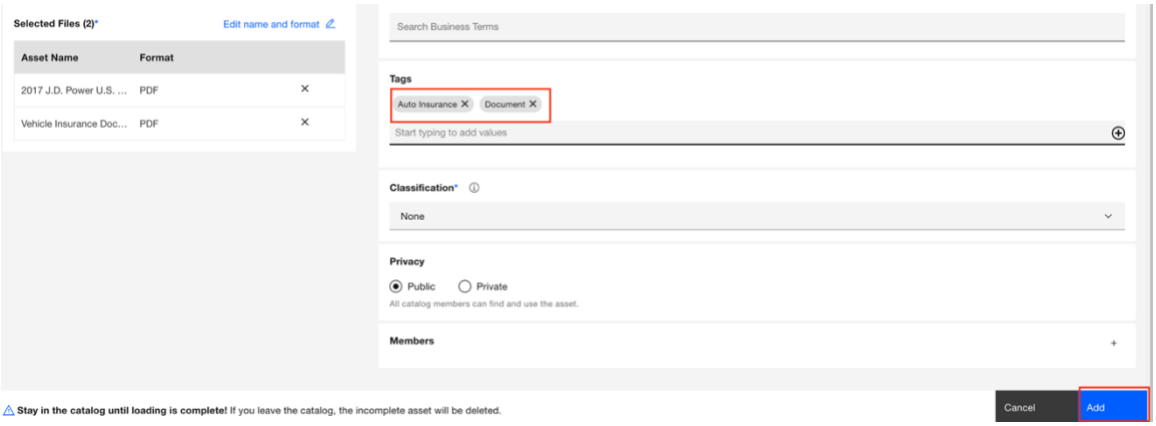
    Each time you enter a tag, you need to click the **+** sign to add the tag. The tags will appear as added tags in the tag area below the tag name. Once a tag is added, it can be used and selected for other data assets. Knowledge Catalog displays all available tags

once they are added to the catalog. You will see this in action when you add the next file to the catalog.

14. Enter a Tag of **Document** into the **Tags** area.



15. Click the **+** sign next to the tag to add it.



The screenshot shows the **Auto Insurance** 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.

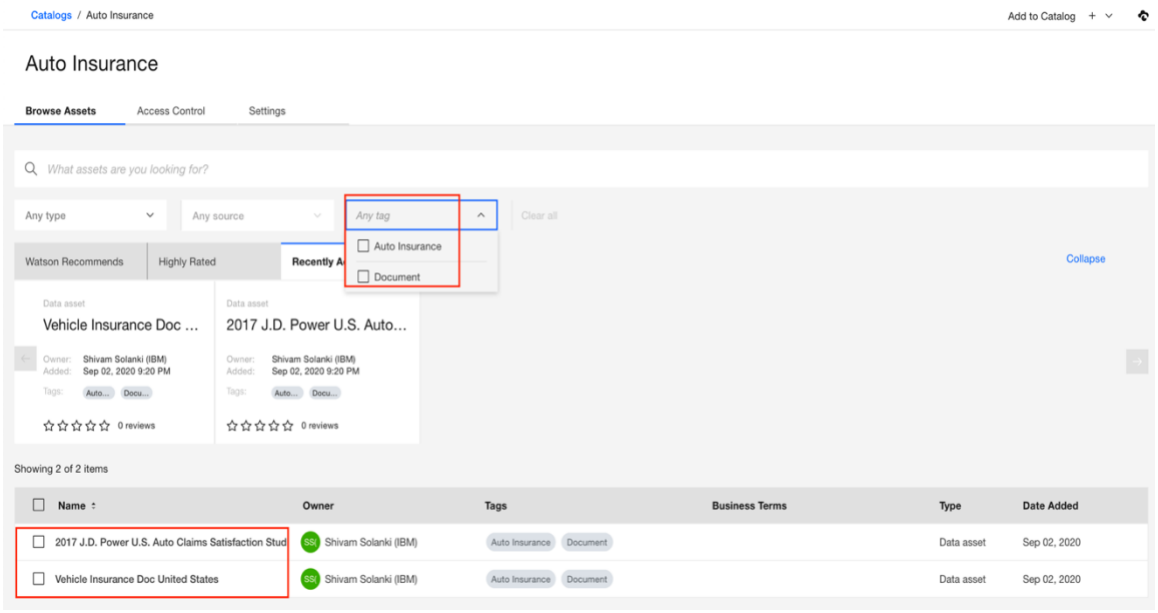
16. Click the **Add** button to catalog the unstructured data assets.

    A message is displayed notifying you that 2 assets are being loaded into the **Auto Insurance** catalog.

17. Click the **X** on the information dialog to close it if it remains open.

18. Click the **Recently Added** tab to view the contents.

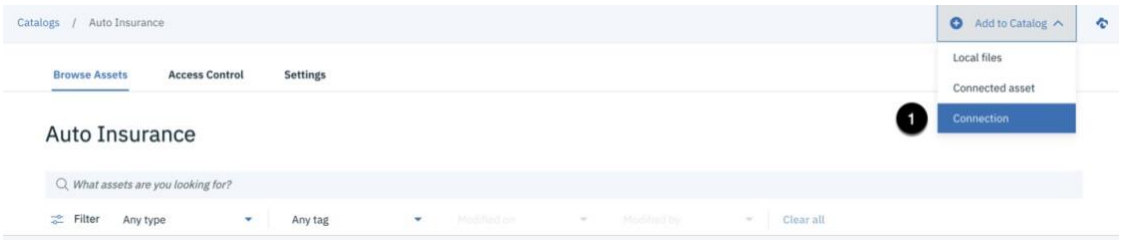
19. Click in the **Any Tag** filter box to view the list of tags.



Upon completion, the data assets will automatically be added to the **Recently Added** 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 **Auto Insurance** and **Document** tags have been added to the **Tags** filter area.

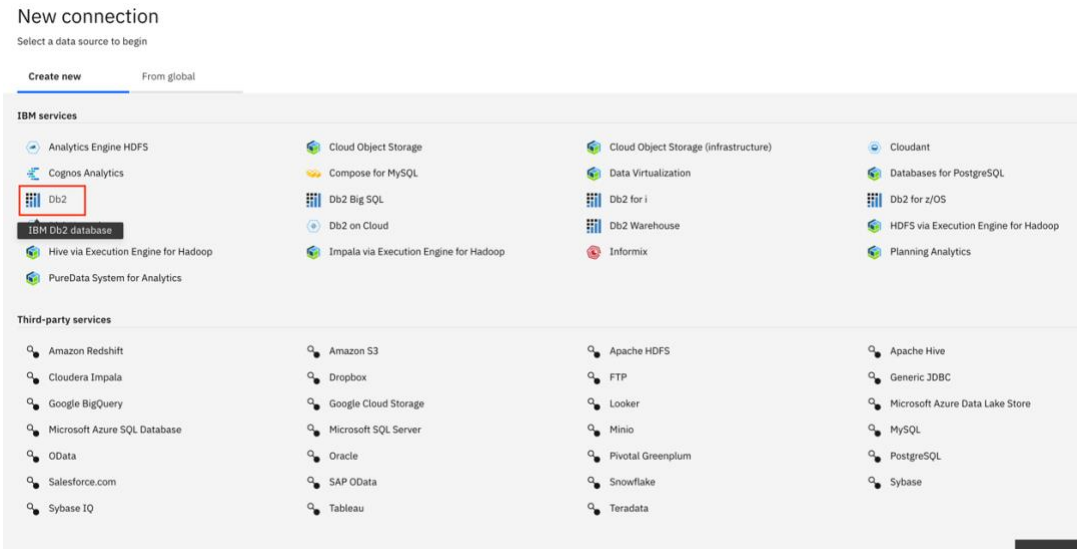
Auto Discover Data Assets

1. Click **Add to Catalog** → **Connection** from the catalog menu.



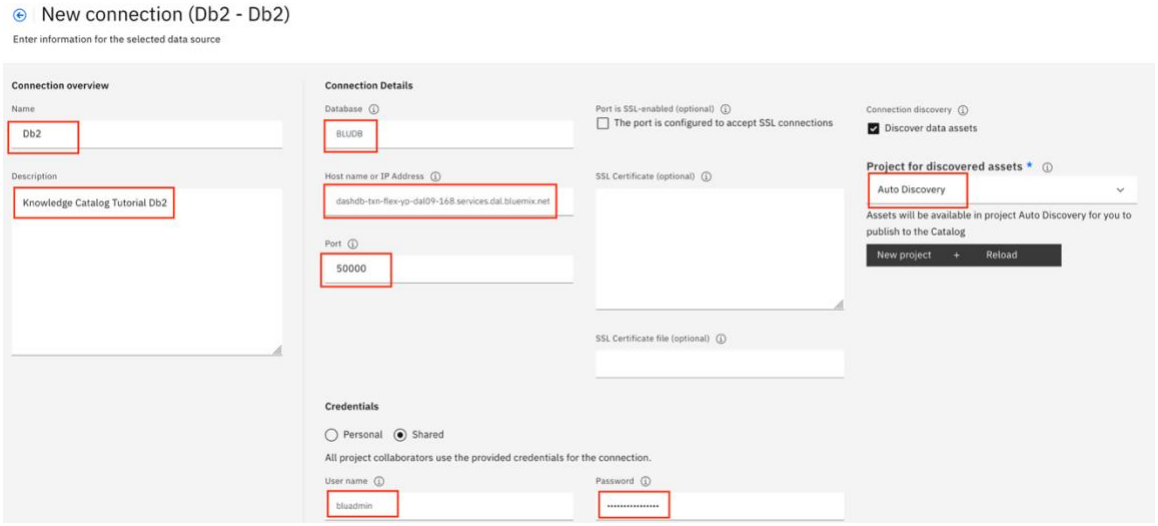
2. Notice that the list of connectors to choose from is quite robust and includes all the IBM services and a generous number of Third-party services as well. Also, connection services are being added on a regular basis, so you may see more than the screenshot this tutorial is displaying.

   Click on the **Db2** connector.



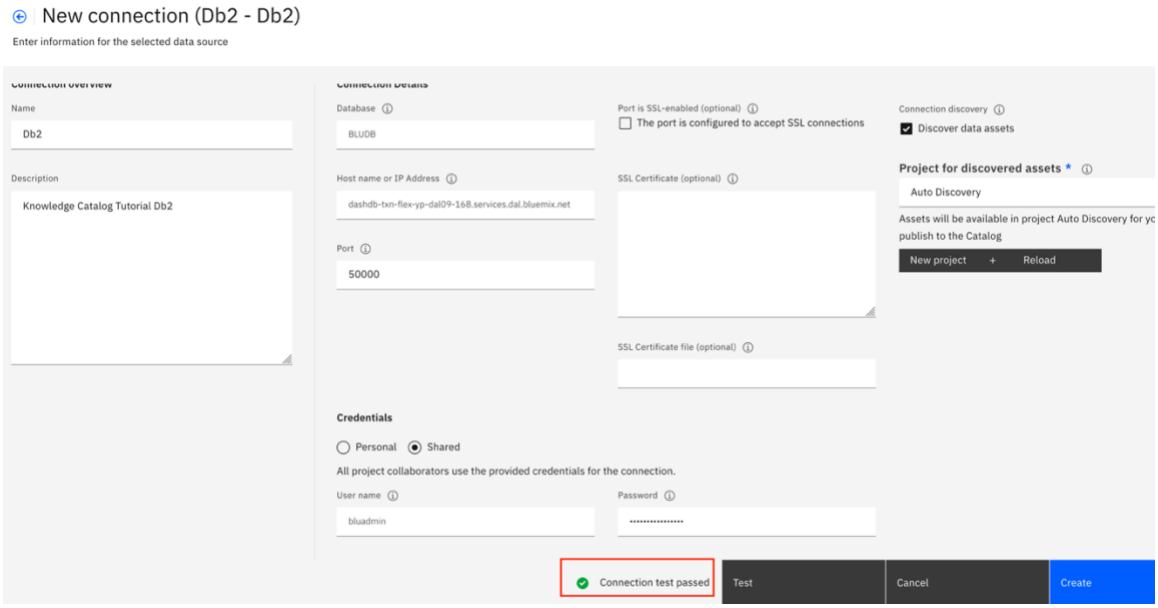
3. Enter the following parameters:

   Name: **Db2**
   Description: **Knowledge Catalog Tutorial Db2**
   Database: **BLUDB**
   Username: **bluadmin**
   Hostname or IP Address field: **dashdb-txn-flex-yp-dal09-168.services.dal.bluemix.net**
   Password: **YmY4ZjIyMTg2YTFl**
   Check the **Discover data assets** check box under **Connection discovery**
   Select the **Auto Discovery** project from the "**Project for discovered assets**" selection list.



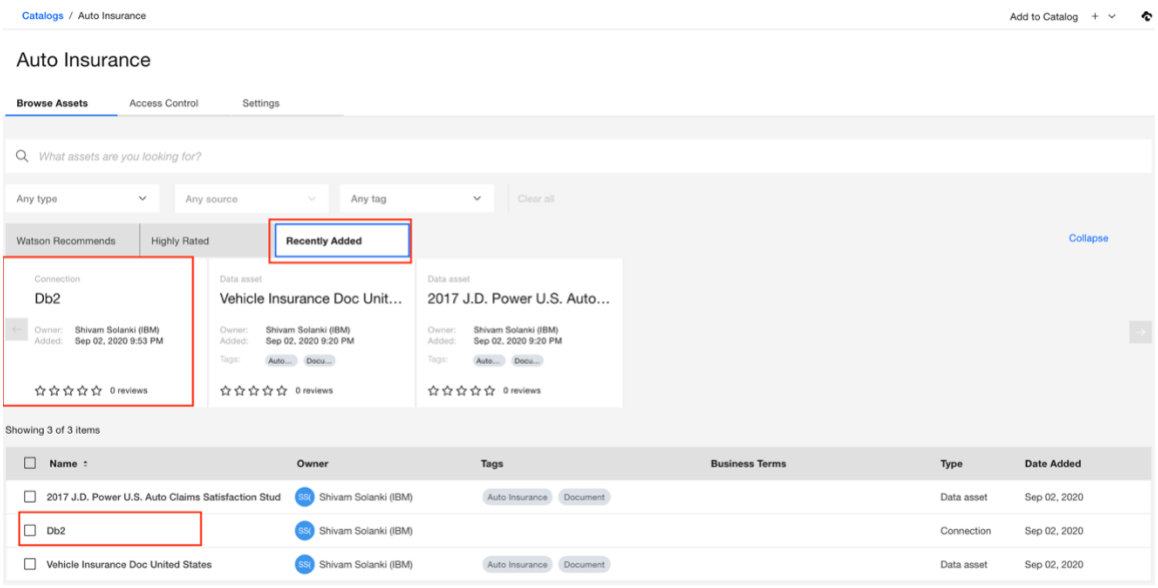
4. Click the **Test** button.

   When you see the green check mark and the message that the **Connection test passed**, click the **Create** button. If it does not pass the test, double check that you entered all the parameters correctly as stated in steps 3-10 above. If it still does not pass the test, notify the instructor.



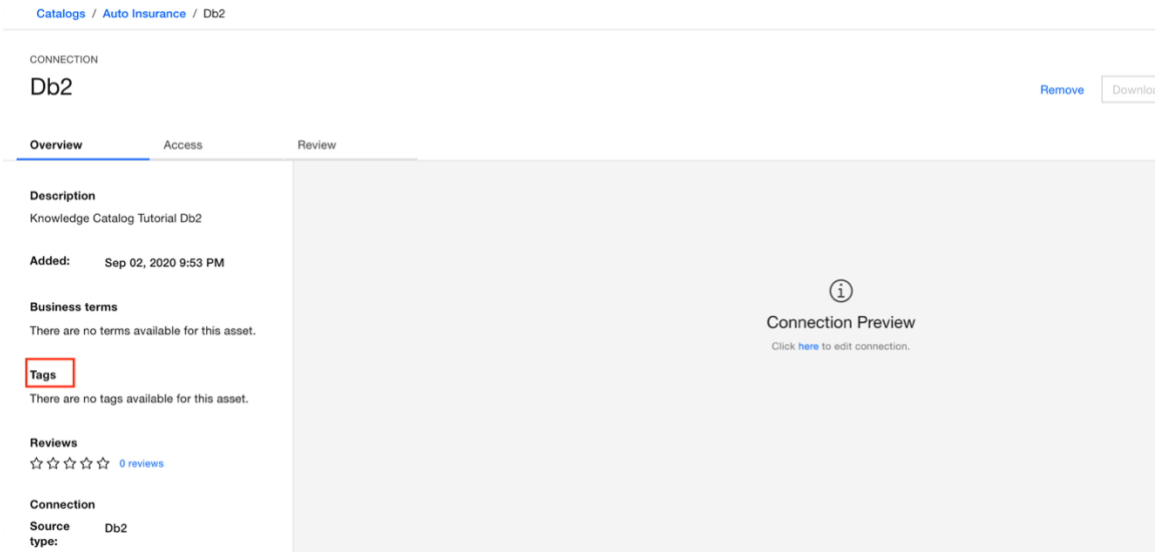
   You will receive a message that Knowledge Catalog is waiting for a response from the connection service (this is the auto discovery service) and a completion and redirection message. You will be brought back to the catalog asset browser and should see your newly added connection in the Data assets list.

5. Click the **Recently Added** section of the asset browser.



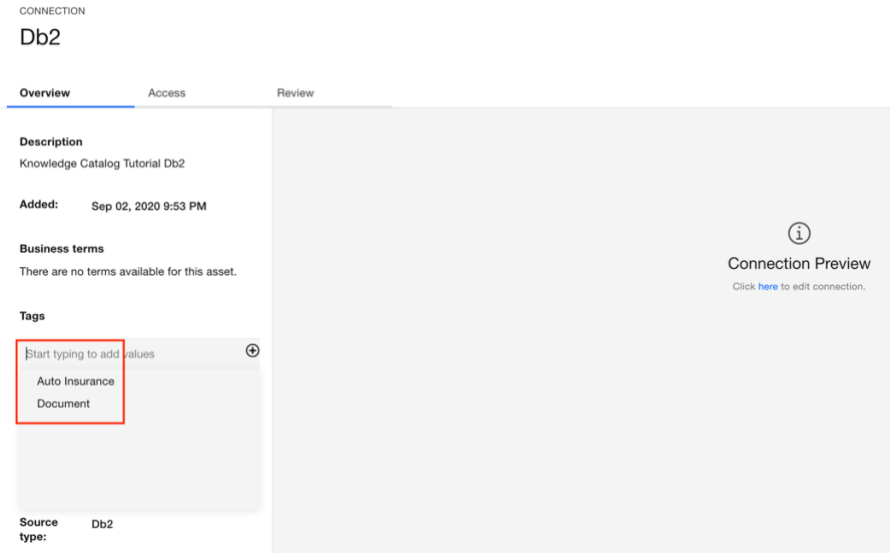
Notice that all the data assets you added appear in this section. As assets are cataloged, they are added to the **Recently Added** section of the catalog asset browser in the sequence they were added, with the most recent appearing first in the list.

6. Click the **Any Type** filter. Notice that the filter has a new asset type of **Connection**.
7. Click on the **Db2 Warehouse** asset in the **Recently Added** section.
8. Hover next to the **Tags** section and click the **pencil icon** to add tags to the connection.



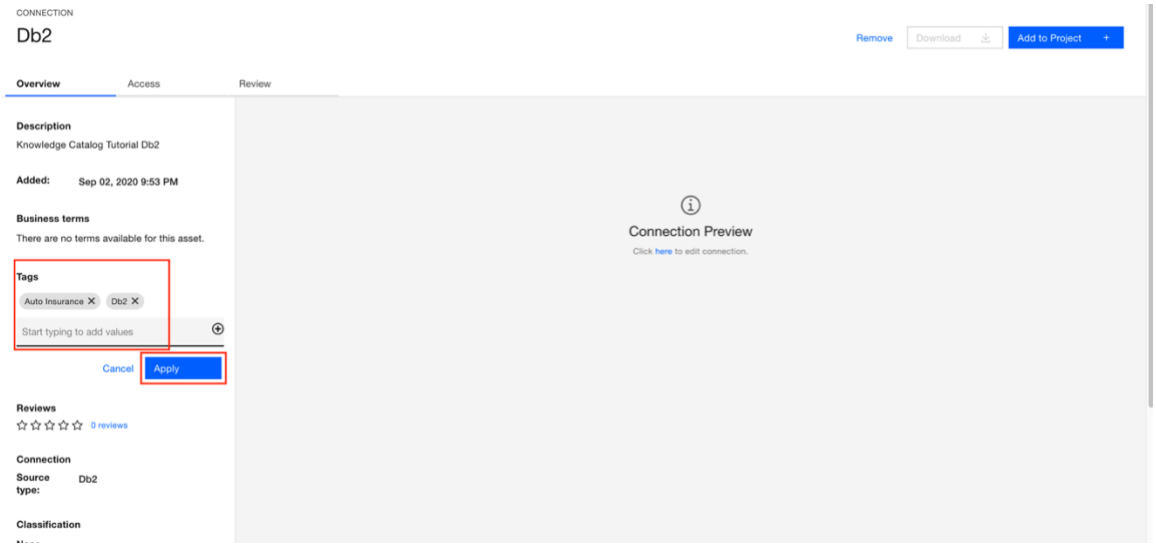
9. Click in the **Tags** section and select the **Auto Insurance** tag from the list of tags.

   Note: If the list does not appear after you click in the Tags section, type in the letter A.



10. Click in the **Tags** section, enter the word **Warehouse** as a tag and click the **+** sign next to the tag to add it.

11. Click the **Apply** button.

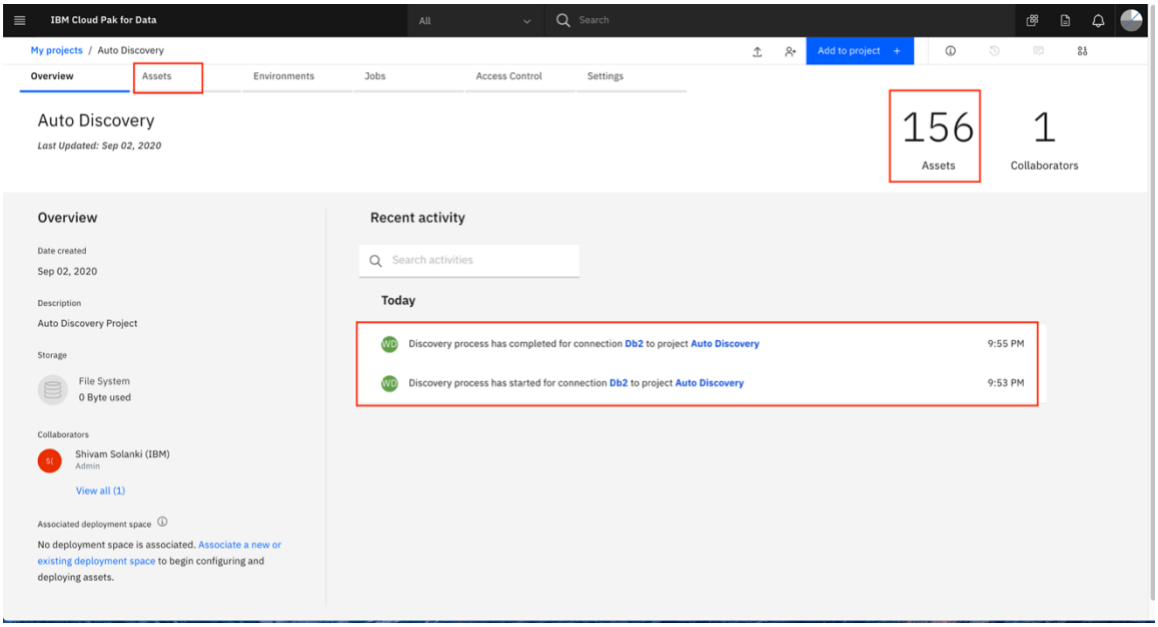


12. The discovery process has been running as a service in the background, discovering and populating data assets into the **Auto Discovery** project. Let's examine the project to review the discovery results. We are interested in finding relevant auto insurance data, specifically Customer, Policy and Claims data that will be used by the auto insurance claims web application.

    Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.

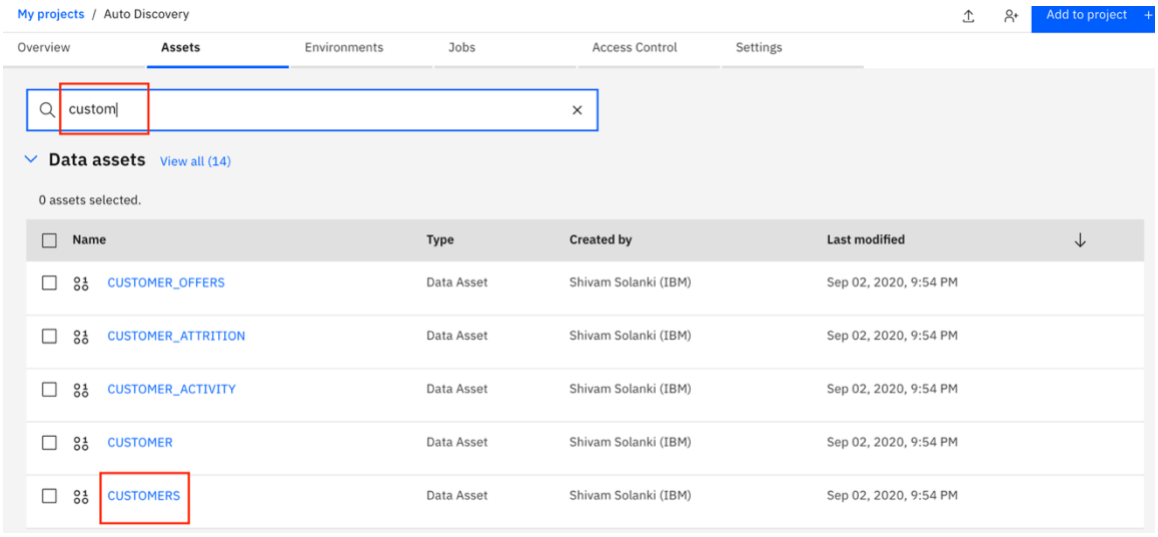
13. Click on the **Projects** menu.

14. From the **Projects** list, select the **Auto Discovery** project.



Notice the number of data assets that were discovered. This discovery, when it was run, auto discovered 156 data assets and added them to the project. Your discovery may be different from this screenshot because this connection is a shared Db2 and data assets are being added and removed all the time by other users. Knowledge Catalog scanned the Db2 Warehouse on Cloud database instance and collected the metadata for all the user data assets that the “bluadmin” user **is authorized** to access.

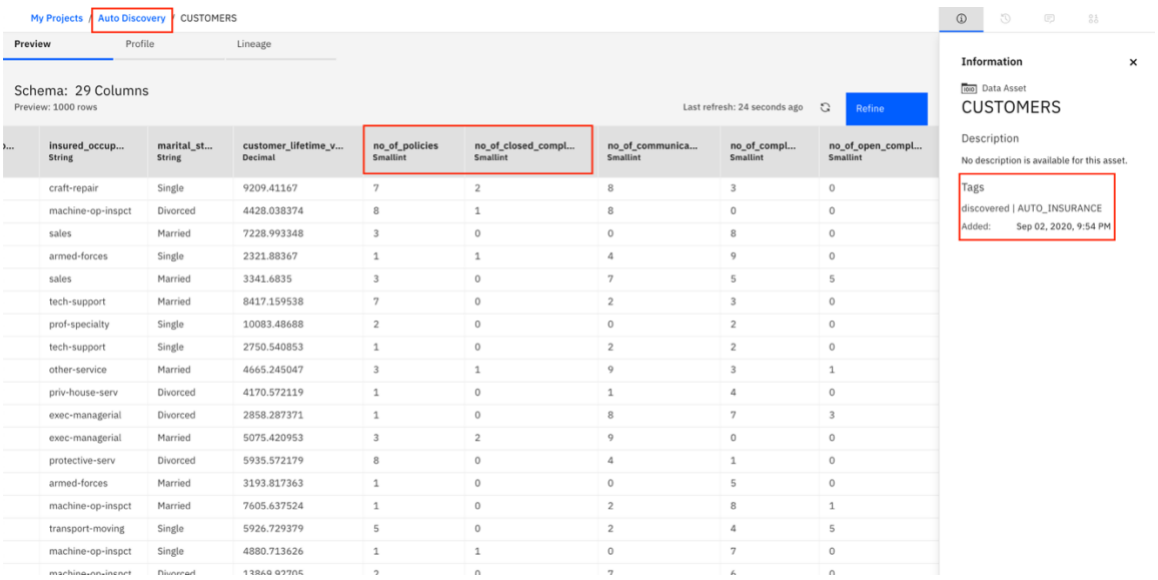
15. Click the **Assets** tab to view the discovered assets.



16. Click in the search area and Enter the letters custom to find all data assets that start with the letter's custom.

    The list is filtered and displays the assets that meet the search criteria. You are looking for auto insurance customers but there are several assets that are related to customer information. The data asset named **CUSTOMERS** looks like it could be the right one. To verify it is related to auto insurance customers, you can preview the first 1000 records of the asset.

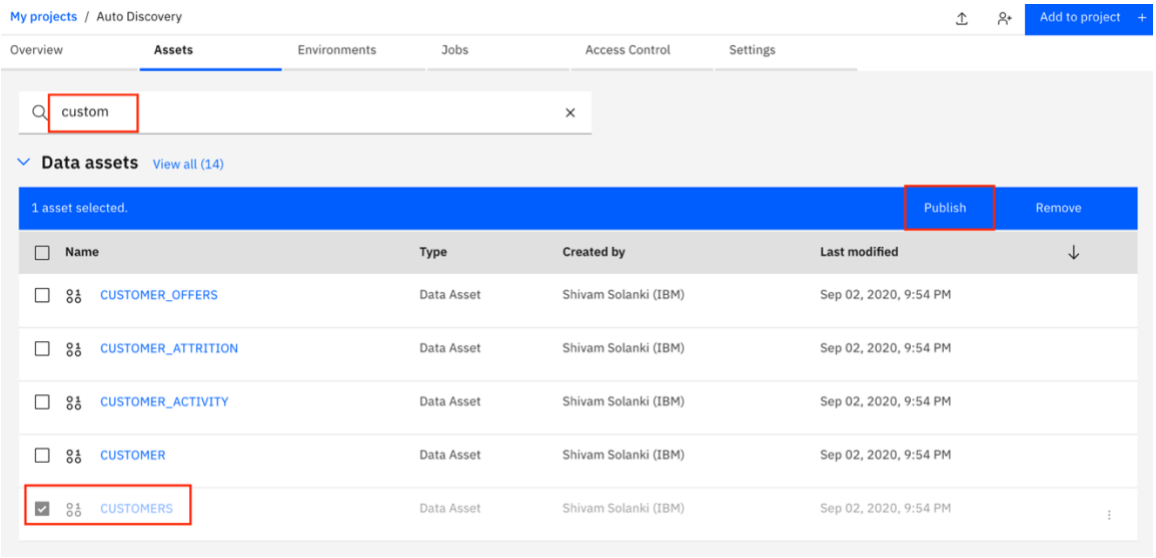
17. Click on the **CUSTOMERS** data asset to open the data previewer.



    The information panel on the right shows a tag of **AUTO_INSURANCE**. This is the schema in the Db2 Warehouse instance that the table came from. Also, if you scroll to the right, you will see that there are columns related to auto insurance like

`number_of_policies` and `number_of_closed_complaints` etc. This is the auto insurance customers data asset that's needed.

The next several steps will demonstrate how you would publish this data asset to the **Auto Insurance** catalog. However, **you will not** publish it from the project. You will catalog it from the **Auto Insurance** catalog in a subsequent step, along with several other tables needed for the auto insurance analysis project, to demonstrate how you can add **Connected assets** from a Connection.



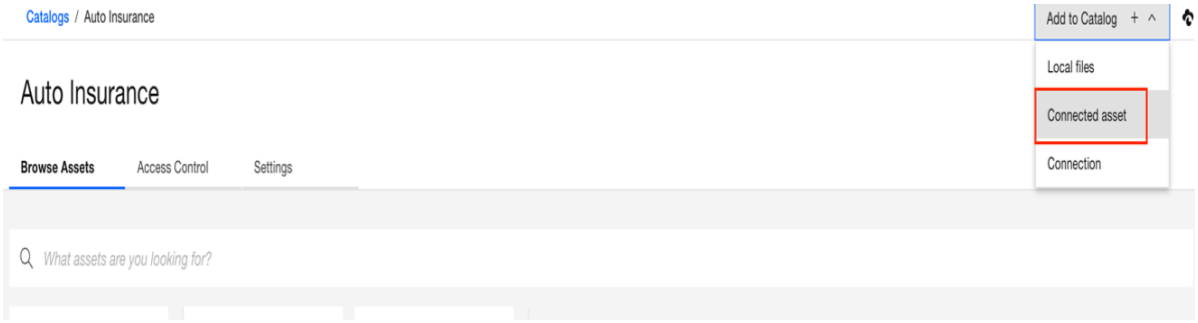
18. Click the **Auto Discovery** link at the top of the page to go back to the Auto Discovery project.
19. Click on the **Assets** tab.
20. Click in the **search area** and Enter the letters **custom** to filter the Data assets list.
21. Click on the **checkbox** to the left of the **CUSTOMERS** data asset.
22. Notice that a **Publish** button appears at the top of the list. **Do not** click the button. This is the button you would select to publish the asset to a Knowledge Catalog, but you **will not** be publishing it to the catalog.

    Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.
23. From the **Organize** section, select the **All catalogs** menu item.
24. Click on the **Auto Insurance** catalog.

Catalog Structured Data

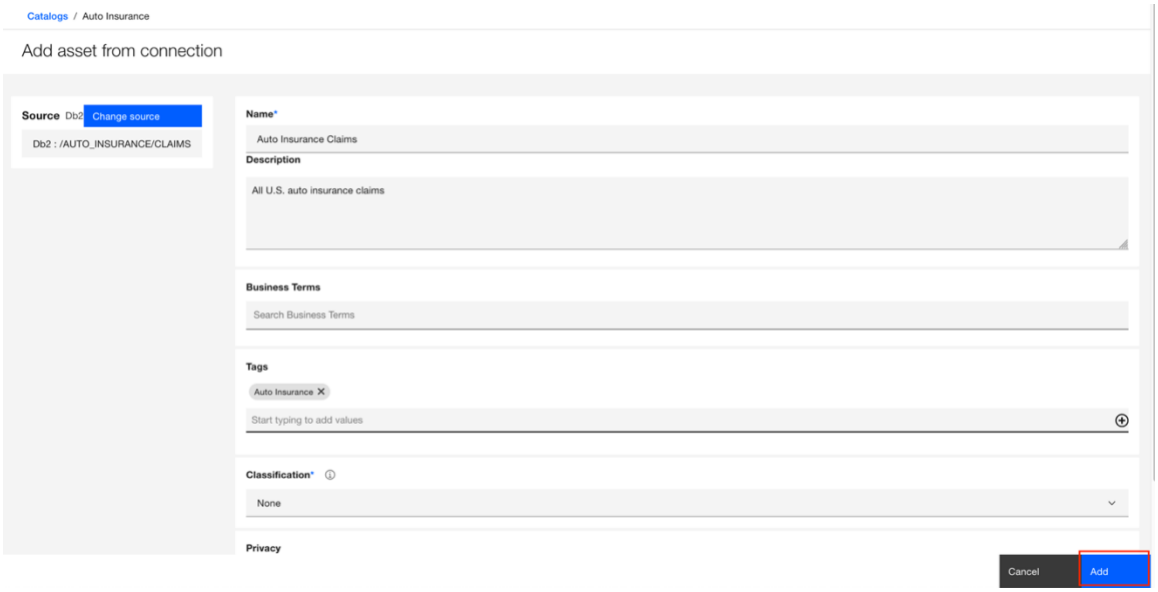
You have cataloged unstructured data files from the local file system and auto discovered data assets from a Db2 Warehouse connection. You will now catalog three tables from the Db2 Warehouse connection; **Claims, Customers** and **Policies** using the **Connected asset** catalog method. These tables are needed for the auto insurance claims analysis processing.

1. Click **Add to Catalog** → **Connected asset** from the Catalog menu.



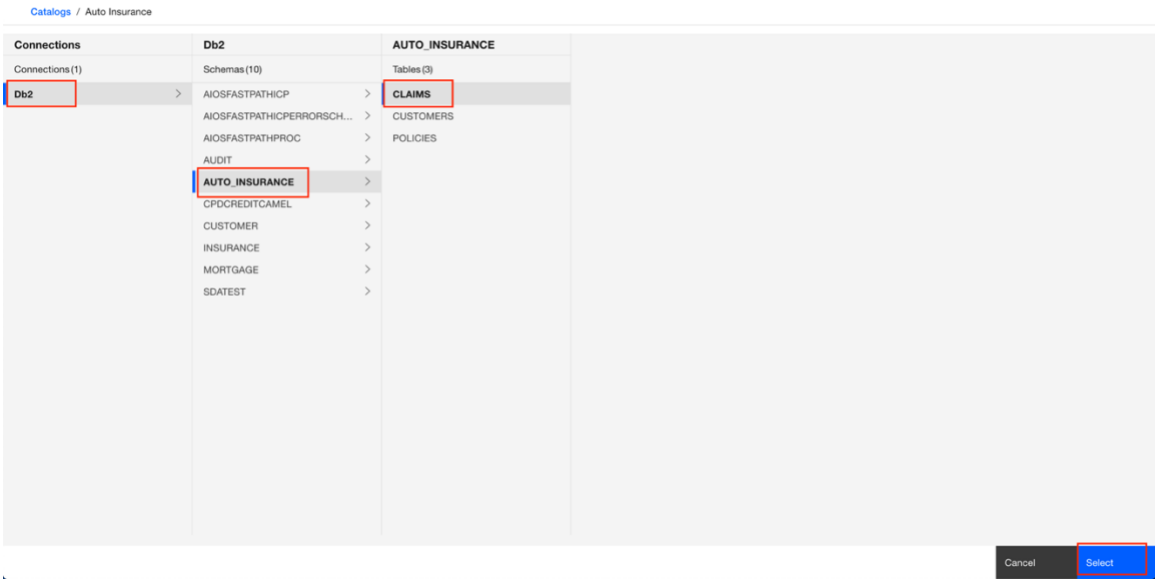
2. Enter a Name of **Auto Insurance Claims** and a Description of **All U.S. auto insurance claims**. Click in the **Tags** area and select the **Auto Insurance** tag from the drop-down list.

   Note: If the list does not appear after you click in the Tags section, type in the letter A.



3. Click the **Select Source** button to choose a Connection to add connected assets from.

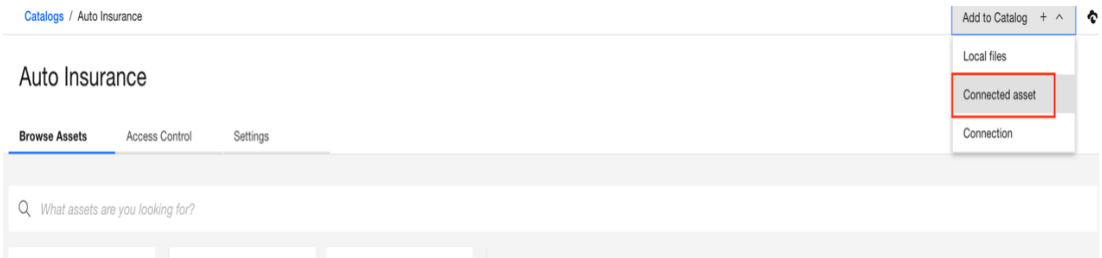
4. Select the **Db2** connection → **AUTO_INSURANCE** schema → **CLAIMS** table. Click the **Select** button.



5. Click the **Add** button.

   You should see the table in the data asset list with the tag you supplied.

6. Click **Add to Catalog** → **Connected asset** from the Catalog menu.



7. Enter a Name of **Auto Insurance Customers** and a Description of **All U.S. auto insurance customers**. Click in the **Tags** area and select the **Auto Insurance** tag from the drop-down list.

   Note: If the list does not appear after you click in the Tags section, type in the letter A.

8. Click the **Select Source** button to choose a Connection to add connected assets from.

9. Click on the **Db2 Warehouse** connection → **AUTO_INSURANCE** schema → **CUSTOMERS** table. Click the **Select** button.

10. Click the **Add** button.

    You should see the table in the data asset list with the tag you supplied.

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

12. Enter a Name of **Auto Insurance Policies** and a Description of **All U.S. auto insurance policies**. Click in the **Tags** area and select the **Auto Insurance** tag from the drop-down list.

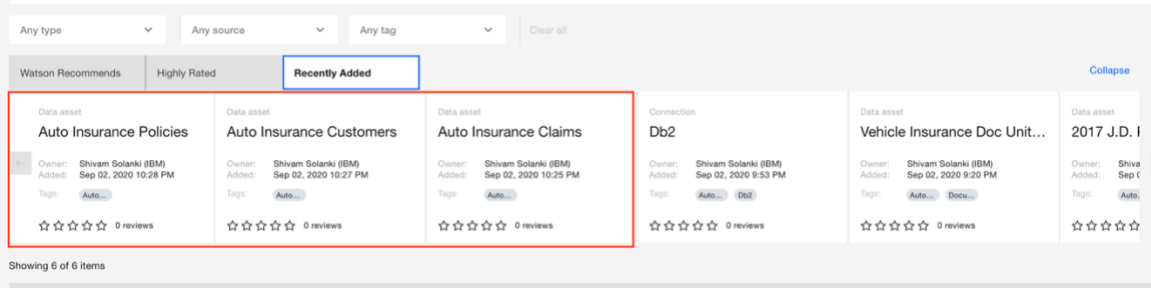
    Note: If the list does not appear after you click in the Tags section, type in the letter A.

13. Click the **Select Source** button to choose a Connection to add connected assets from.

14. Click on the **Db2 Warehouse** connection → **AUTO_INSURANCE** schema → **POLICIES** table. Click the **Select** button.

15. Click the **Add** button.

16. Click the **Recently Added** tab.



You should see the cataloged tables and connection in the **Recently Added** section as a data asset with the tags you supplied.

Understand and Socialize Data Assets

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 data asset to examine the 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.

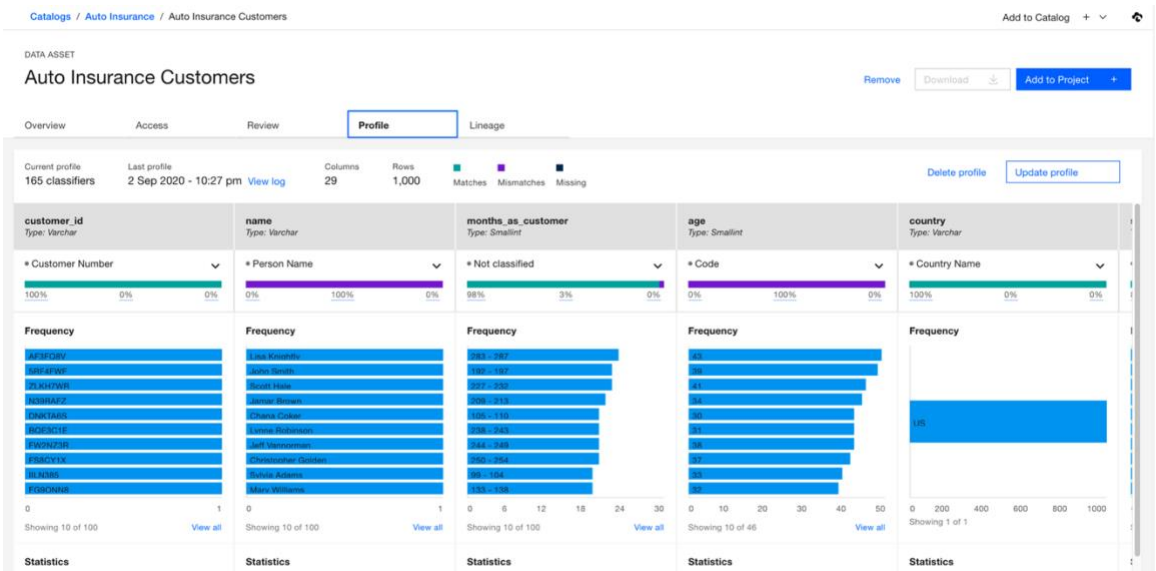
1. Click on the **Recently Added** section of the data asset browser.

2. Click on the **Auto Insurance Customers** asset to view its properties.

   You are brought to the **Overview** section of the asset where you can view a 1000 row sample of the data 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.

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

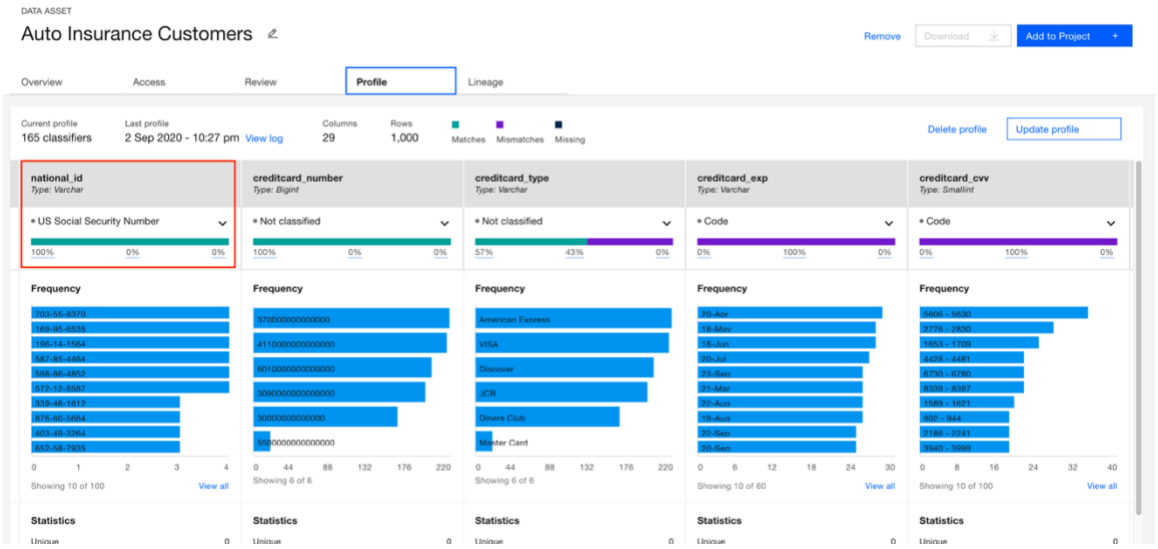
   The profile should automatically appear. If not, and you are presented with a method to create or update the profile, follow the instructions to do so.



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.

Attribute classifiers describe the contents of the data in the column: for example, city, account number or credit card number. Attribute classifiers are necessary to anonymize data with data policies. The attribute classifiers appear for each column on the asset's **Overview** and **Profile** page.

4. Scroll **to the right** until you see the `NATIONAL_ID` column statistics.



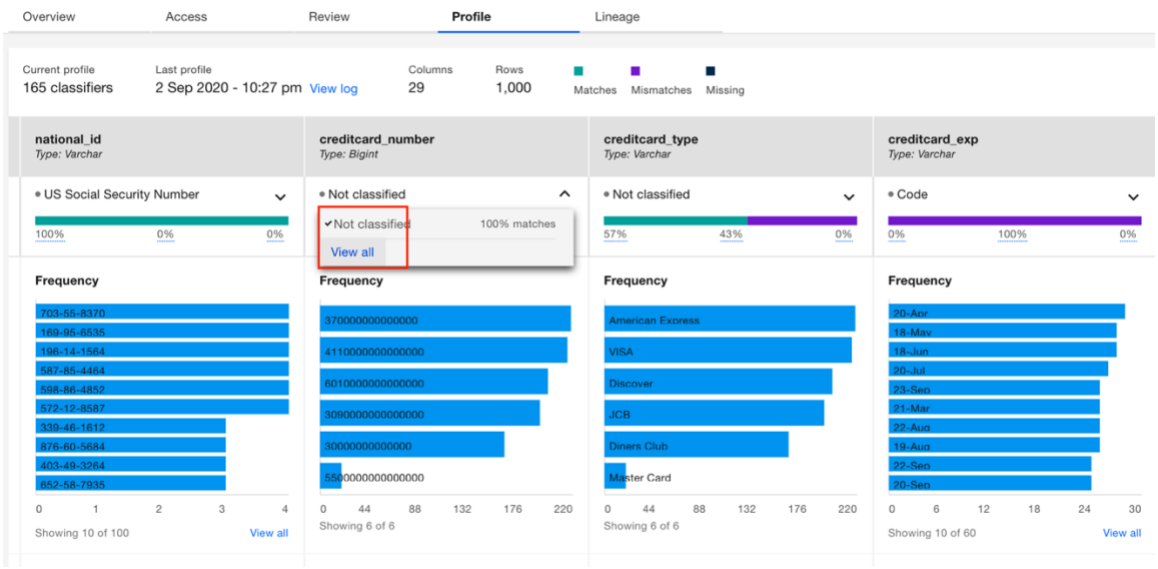
Note: You will notice that the `national_id` column is classified correctly but the `creditcard_number, creditcard_exp, creditcard_exp` and `creditcard_cvv` columns are not. You will also notice they look like sensitive information and should be protected by data protection rules.

There are no data protection rules created in the glossary of this Cloud Pak for Data cluster to protect sensitive information. Even if there were, they would not be enforced and stop you from viewing the data content because you added the data asset to the catalog and are the owner. If another user were to log in and view the data, and rules were active, they would be enforced.

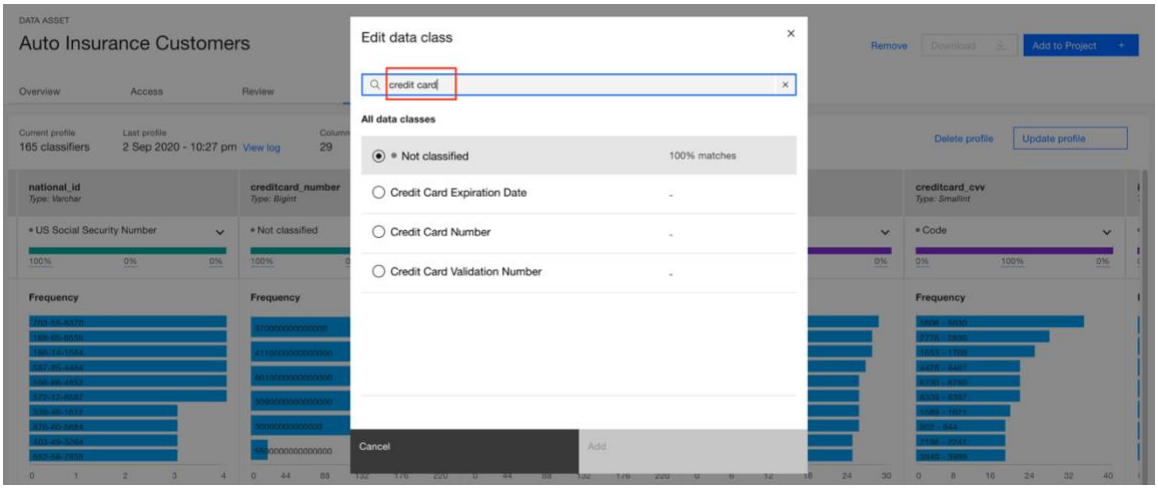
Time permitting, you will get a chance to experience how data protection works at the end of the lab so you can see it in action.

In any case, in order for any data protection rule based on a data class to be enforced, the data class assignments have to be correct. To protect the Credit Card number, Credit Card Expiration Date and Validation Number of a Credit Card, the data classes for the `CREDITCARD_EXP` and `CREDITCARD_CVV` columns have to be set properly.

5. Click on the down arrow next to the **Not Classified** classification of the CREDITCARD_EXP column. Click the **View All** menu item.



6. Type in **credit card** in the search area. Hover next to the **Credit Card Number** data class and click **Add**.



7. Click the **Close** button.

8. Click on the down arrow next to the **Not classified** classification of the CREDITCARD_CVV column.

9. Click the **View All** menu item.

10. Type in **credit card** in the search area.

11. Hover next to the **Credit Card Validation Number** data class and click **Use**.

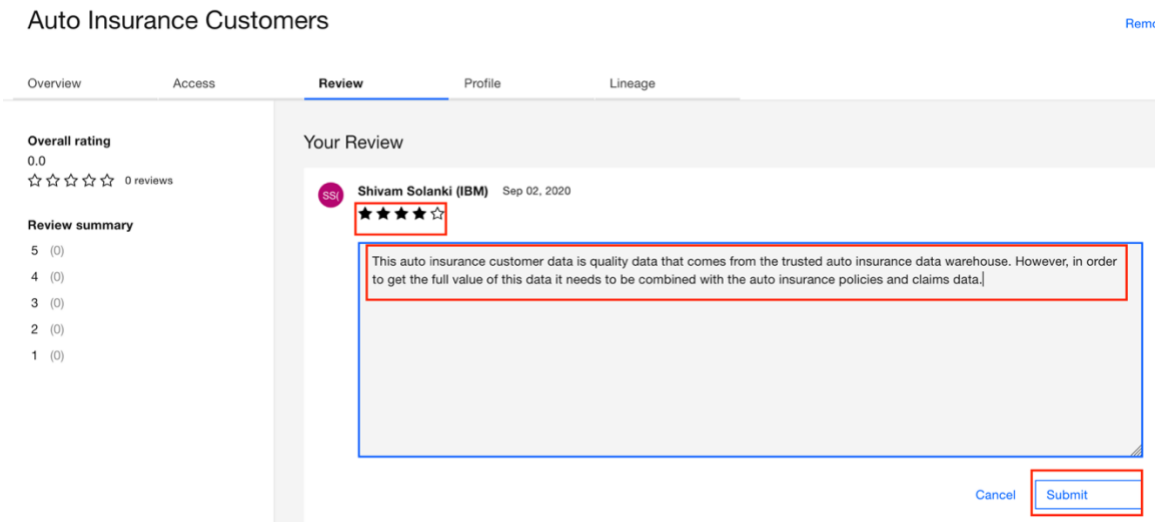
12. Click the **Close** button.

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

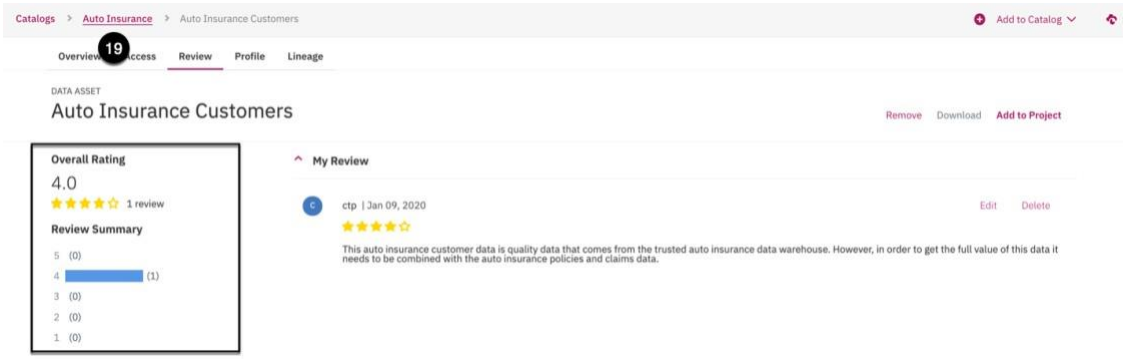
14. Copy and Paste the following bolded text into the **Description**:

    This auto insurance customer data is quality data that comes from the trusted auto insurance data warehouse. However, in order to get the full value of this data it needs to be combined with the auto insurance policies and claims data.

15. Click the **4th star** from the left to give the asset a 4-star rating. Click the **Submit** button.

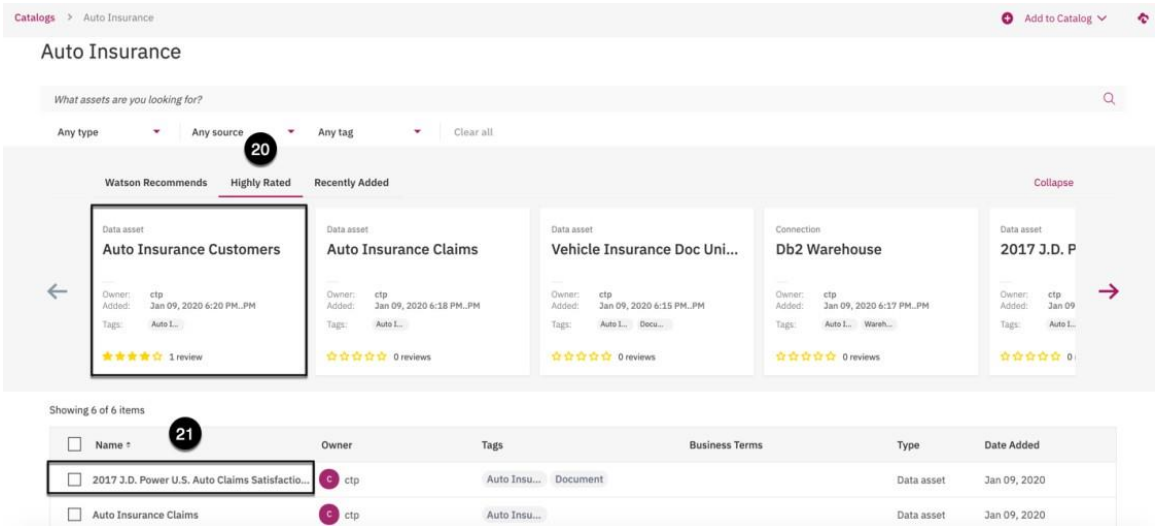


Notice that you now have one review with an Overall Rating of 4.0.

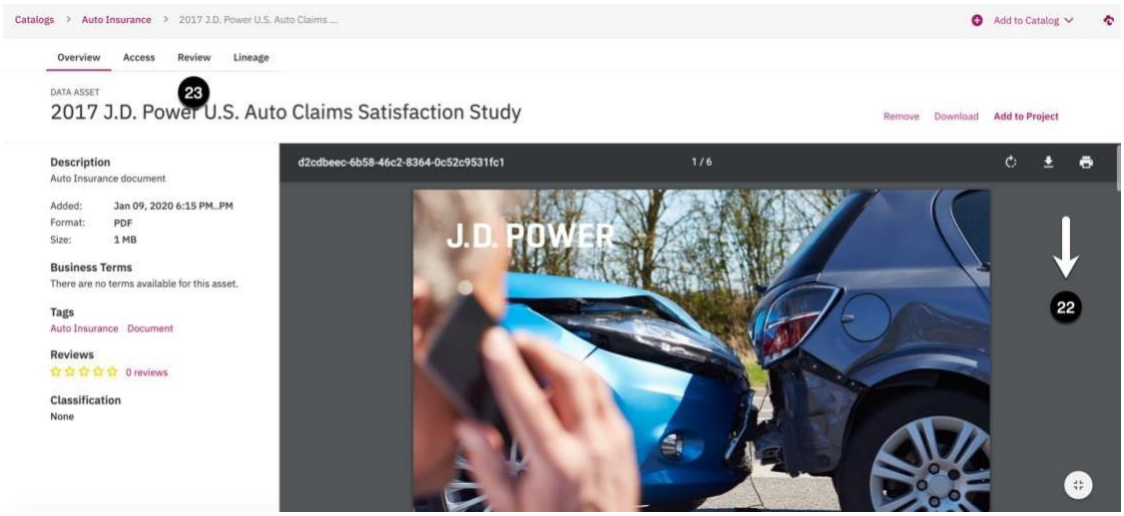


16. Click the **Auto Insurance** link at the top of the page to go back to the catalog asset browser.

17. Click on the **Highly Rated** section and notice that the **Auto Insurance Customers** data asset is the most highly rated data asset with 1 review.



18. From the data asset list below, click on the **2017 J.D. Power U.S. Auto Claims Satisfaction Study** asset to view its properties.



The **Overview** section displays the document and allows you to view its contents. **Notice** that numerous viewing controls appear along with action buttons to print, download and rotate the document. If you do not see the controls, place your cursor inside the document viewing area towards the top.

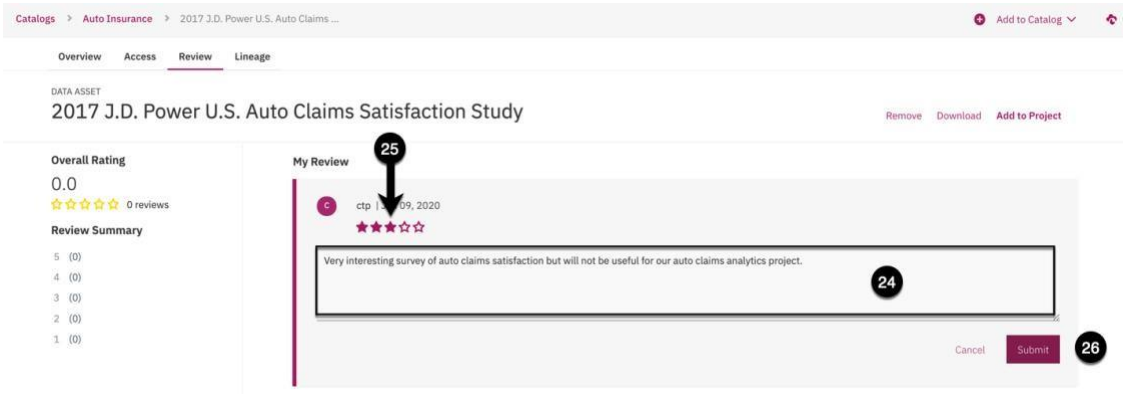
19. **Scroll** down to view the content of the document.

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

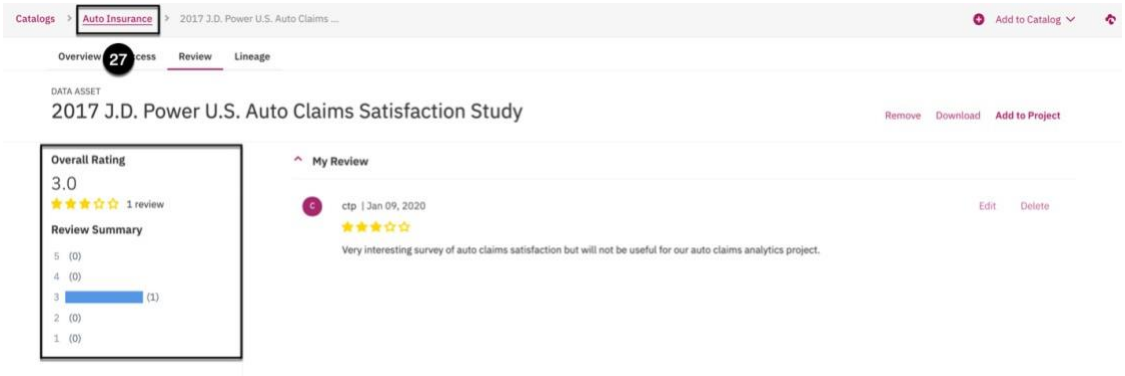
21. Copy and Paste the following bolded text into the **Description**:

    Very interesting survey of auto claims satisfaction but will not be useful for our auto claims analytics project.

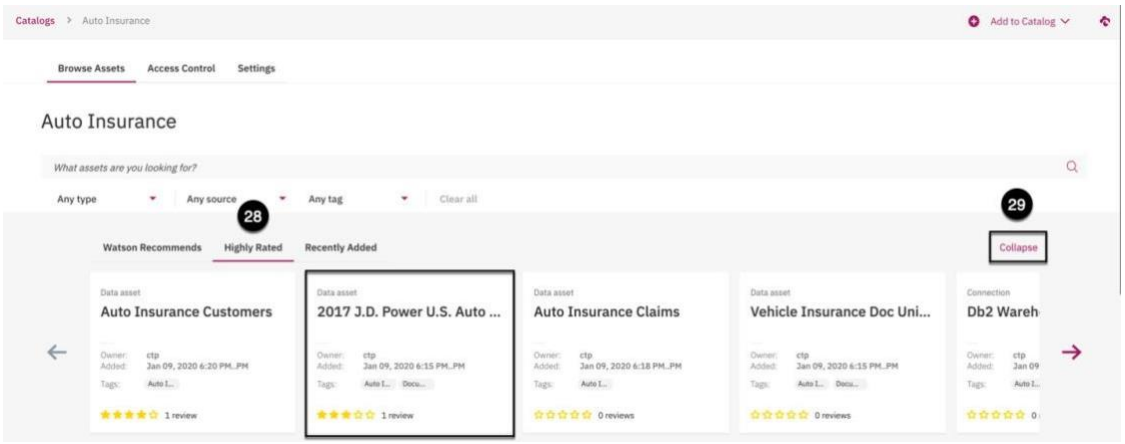
22. Click the **3rd star** from the left to give the asset a 3-star rating. Click the **Submit** button.



Notice that you now have one review with an Overall Rating of 3.0.



23. Click the **Auto Insurance** link at the top of the page to go back to the catalog asset browser.

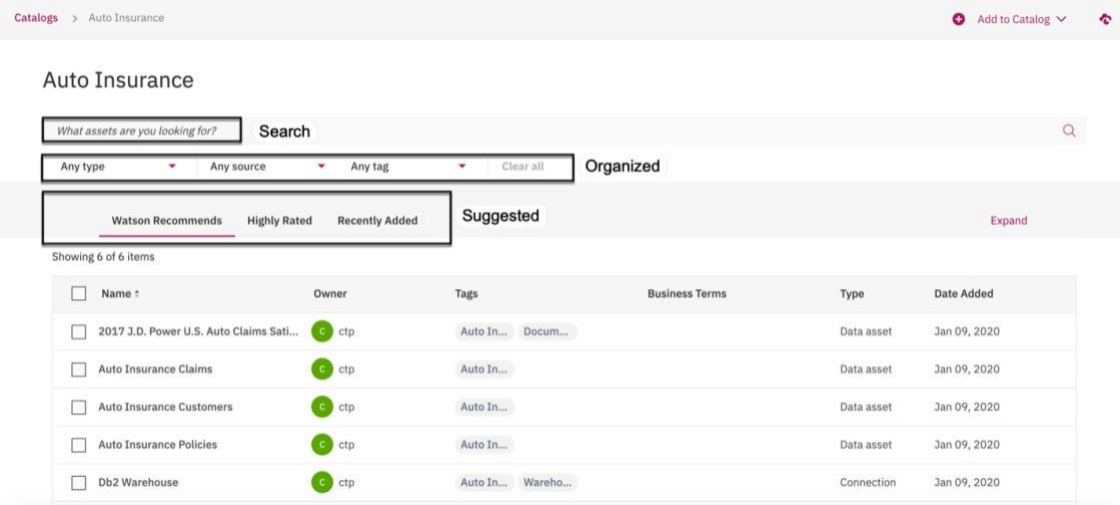


24. Click the **Highly Rated** section.

Notice that the **2017 J.D. Power U.S. Auto Claims Satisfaction Study** data asset is now showing as the 2nd highest rated data asset with 1 review.

25. Click the **Collapse** button to the far right of where the suggestions are to close the area in preparation for the next task.

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.

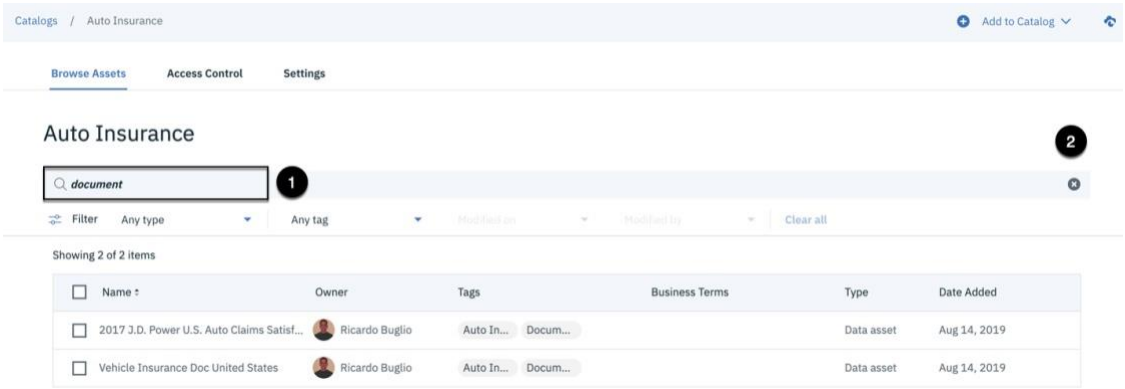
Shop using Suggestions

You just experienced this type of search in the previous task. 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.

Shop using Search

In this section you will shop for data by specifying search criteria using the **Search area** (Where it reads *What assets are you looking for?*) of the Knowledge Catalog asset browser. Note that search criteria are not case sensitive.

1. Inside the Knowledge Catalog search area type in **document**.



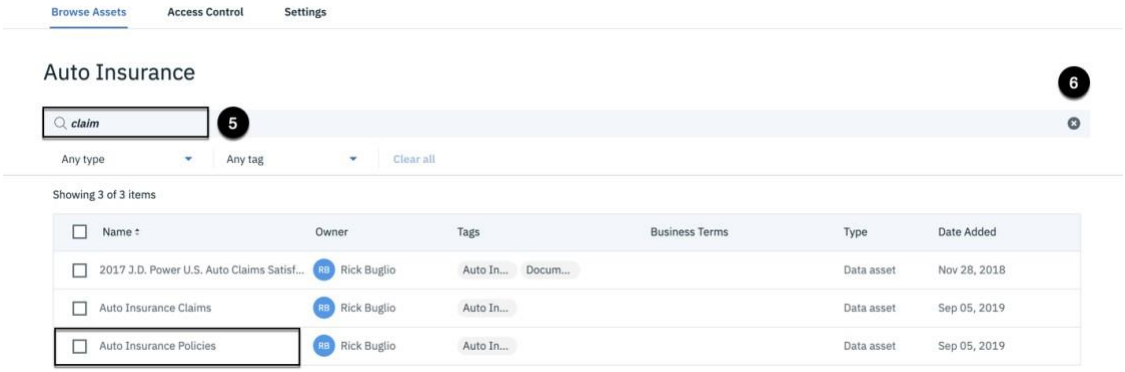
Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **document**.

2. Click the **x** at the far right of the search area to clear the search.

3. Inside the Knowledge Catalog search area type in **db2**.

   Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **db2**.

4. Click the **X** at the far right of the search area to clear the search.



5. Inside the Knowledge Catalog search area type in **claim**.

   Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **claim**.

   Why is the **Auto Insurance Policies** table in the result set? This is an example of the search finding an asset that has a column that contains the consecutive characters **claim** in its name. This table has three columns named **LAST_CLAIM**, **DENIED_CLAIMS** and **CLAIMS_FILED** that meet the criteria. And remember, search is not case sensitive. You will see these columns when you prepare the data in the next task.

6. Click the **X** at the far right of the search area to clear the search.

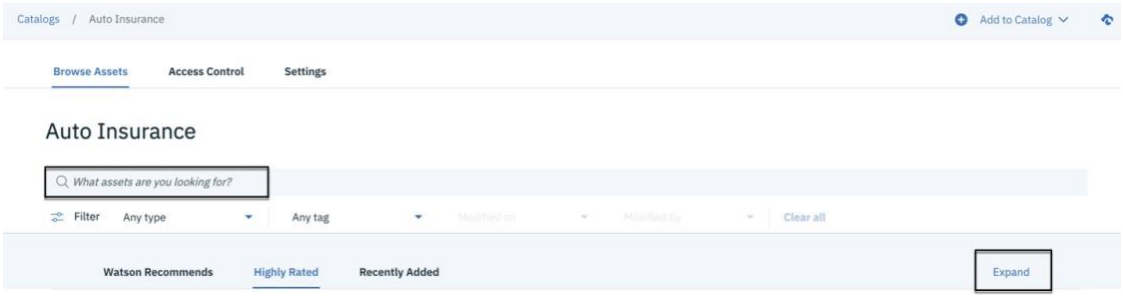
Using the same search method, type in the following searches in the **search area**. Clear the search area after each search to get the correct results:

- Enter the characters **all** in the search area and view the results.
- Enter the characters **U.S** in the search area and view the results.
- Enter the characters **study** in the search area and view the results.
- Enter the characters **unit** in the search area and view the results.

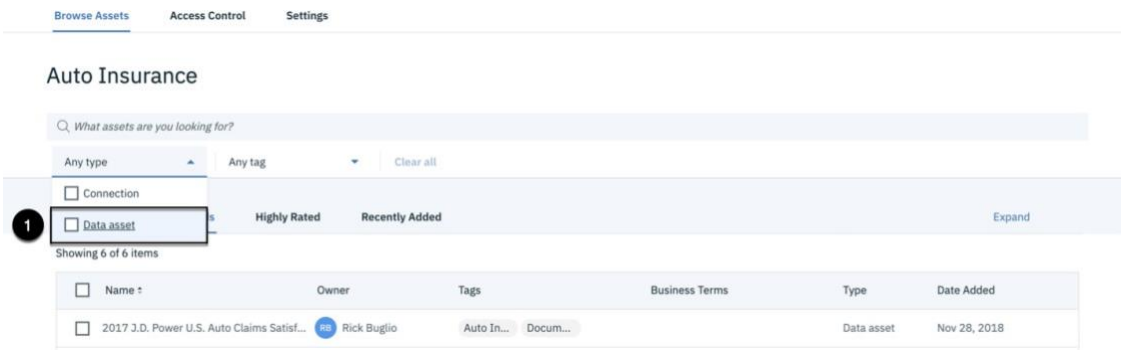
Time permitting, you can experiment on your own and type in different criteria in the **search area** to get more experience with how search works.

Search using Filters

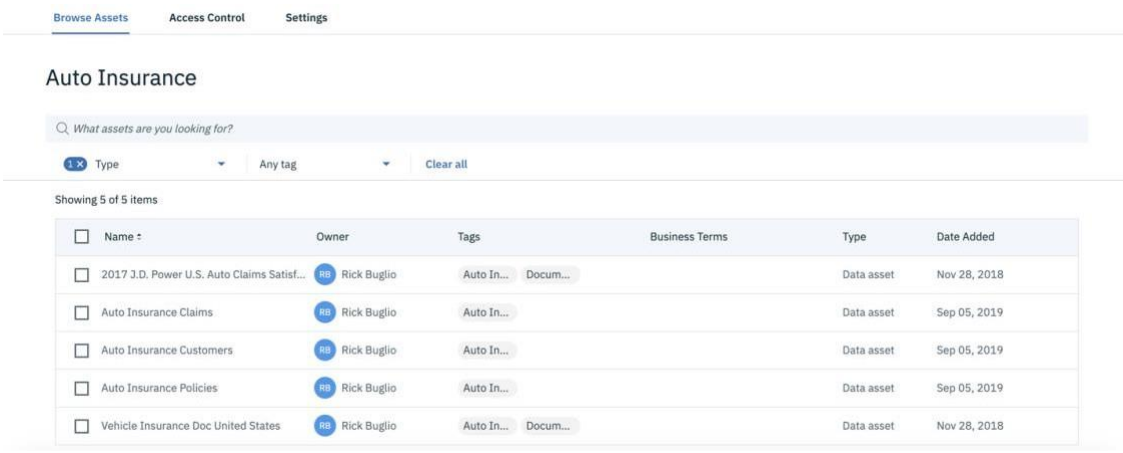
In this section, you shop for data using the **Filter** area that is automatically built by Knowledge Catalog as assets are added to the catalog. You can use one to many filters in combination with each other to get the desired results you are looking for. You may also get an empty search result depending on the combinations you specify.



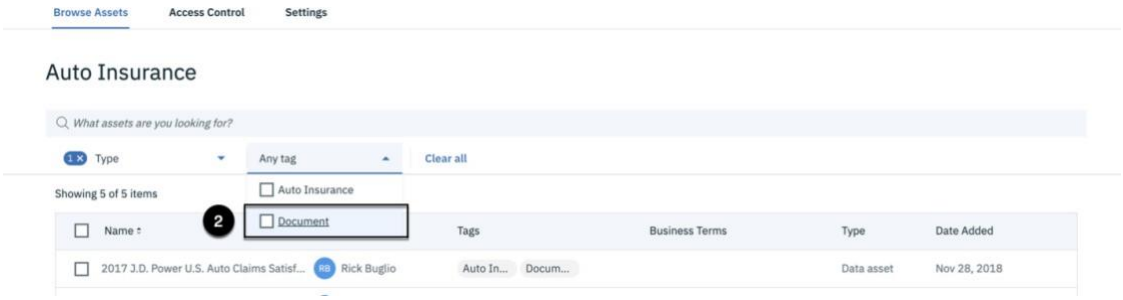
Before you begin, make sure the **search area** is cleared out and that the suggestion categories are collapsed. You should see an **Expand** button if they are collapsed, like in the screenshot above. If you see a **Collapse** button, select it to collapse the section to gain more viewing real estate.



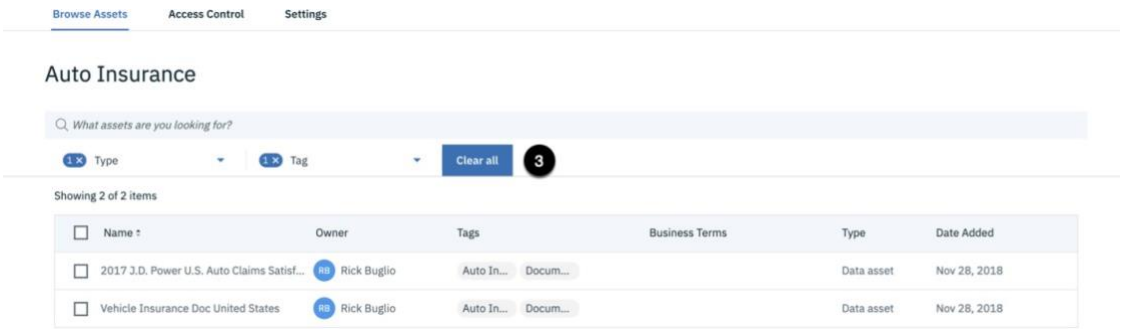
1. In the **Filter** area of the catalog browser, click in the **Type** filter area and select the **Data asset** type and view the results.



Only the 5 assets that have an asset type of **Data asset** are displayed; 2 files and 3 tables. The other types of assets that can be cataloged are **Connections**, **Models**, **Notebooks** and **Dashboards**.



2. Click in the **Tag** filter area and select the **Document** tag and view the results.



Only **Data assets** that have a tag of **Document** are displayed.

3. Click the **Clear All** button to clear all filters.

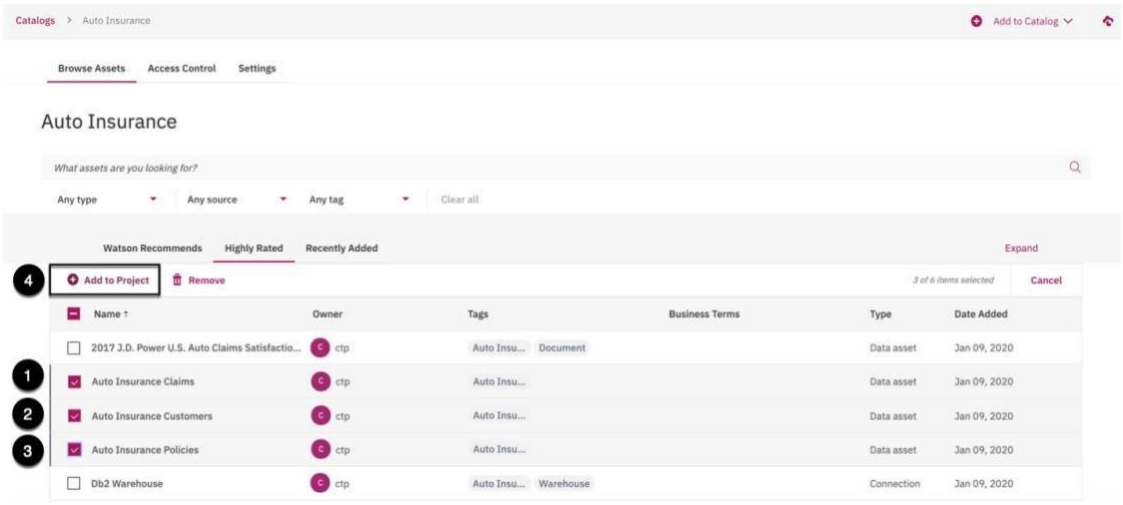
Prepare Data for Analytics and AI

In this task, you will take the structured data you cataloged in the **Auto Insurance** catalog into the **Auto Insurance** analytics project you created and prepare the data for analytics and AI. You will gain an understanding of the data preparation capabilities within a Cloud Pak for Data analytic project and how it can also help you understand and visualize the data before and after preparation.

Add Cataloged Data Assets to a Project

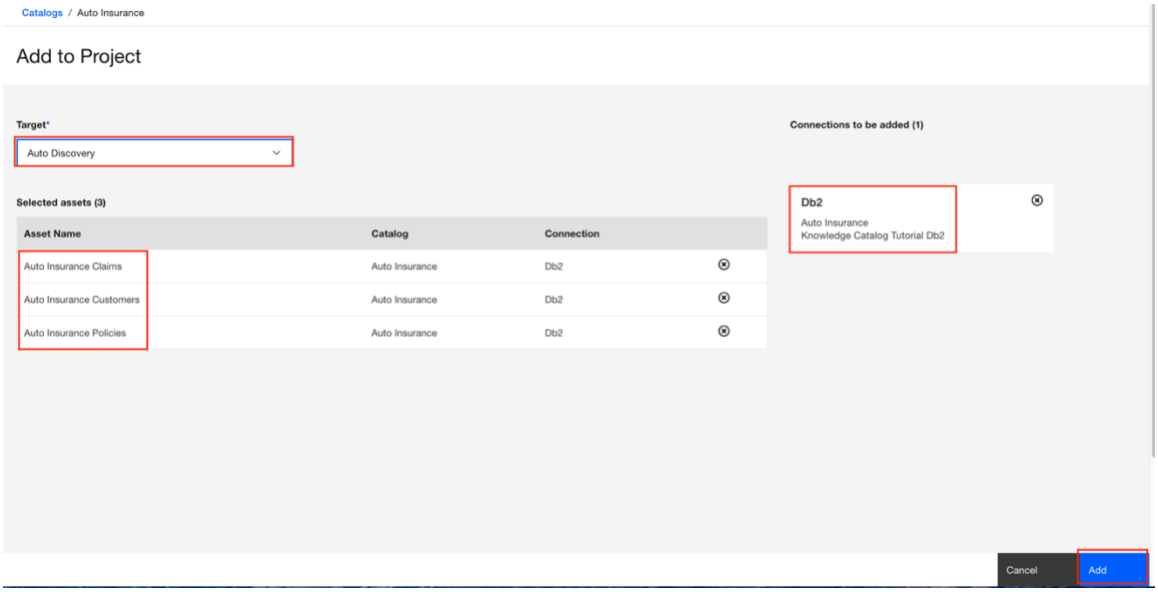
In order to refine data, the data needs to be in a project. You will add three auto insurance data assets from the **Auto Insurance** catalog to the Auto Insurance project to prepare it for analytics and AI. There are two ways to add cataloged assets to a project; from the catalog and from a project. You will add the cataloged assets to the **Auto Insurance** project from the **Auto Insurance** catalog.

1. Click the **check box** next to the **Auto Insurance Claims** data asset.



2. Click the check box next to the Auto Insurance Customers data asset.
3. Click the check box next to the Auto Insurance Policies data asset.

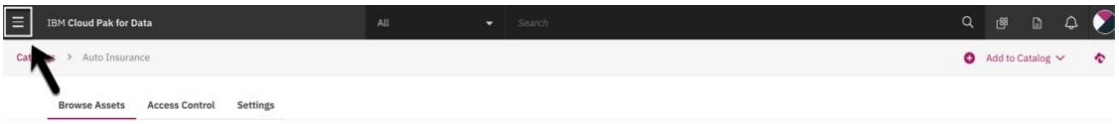
4. Click the **Add to Project** button at the top of the data asset list.



Notice on the right that the **Db2 Warehouse** connection was also included to be added to the Auto Insurance project; The connection is needed by the project to access the data from the Db2 Warehouse.

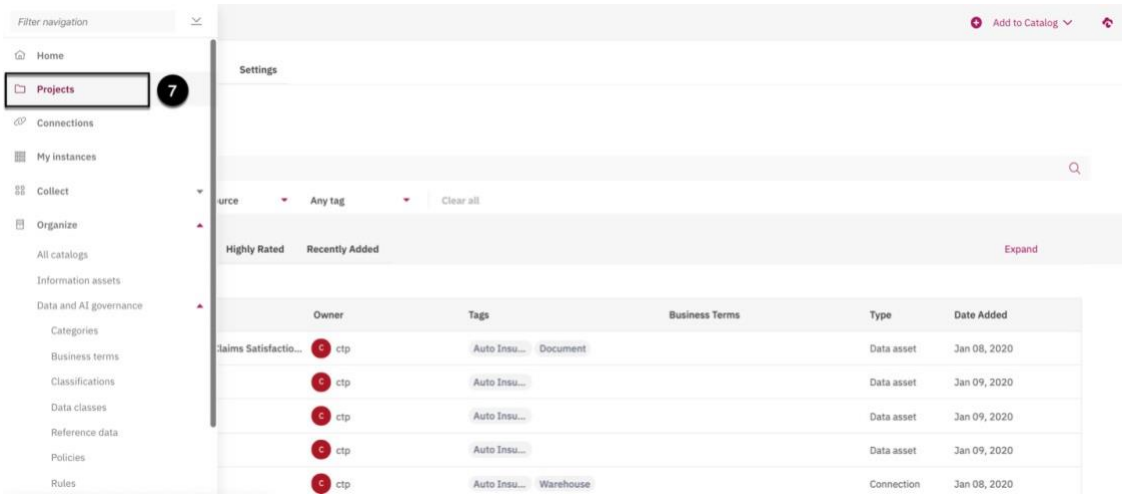
5. Select the **Auto Insurance** project from the list of Target projects.

6. Click the **Add** button.

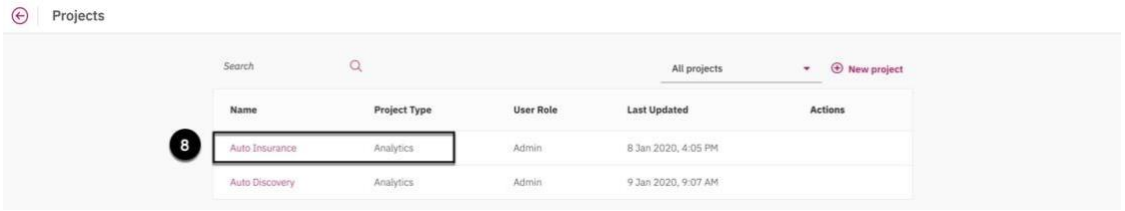


You are brought back into the **Auto Insurance** Knowledge Catalog after the data assets are added to the project. A message at the top of the catalog will inform you that the assets were successfully added to the project.

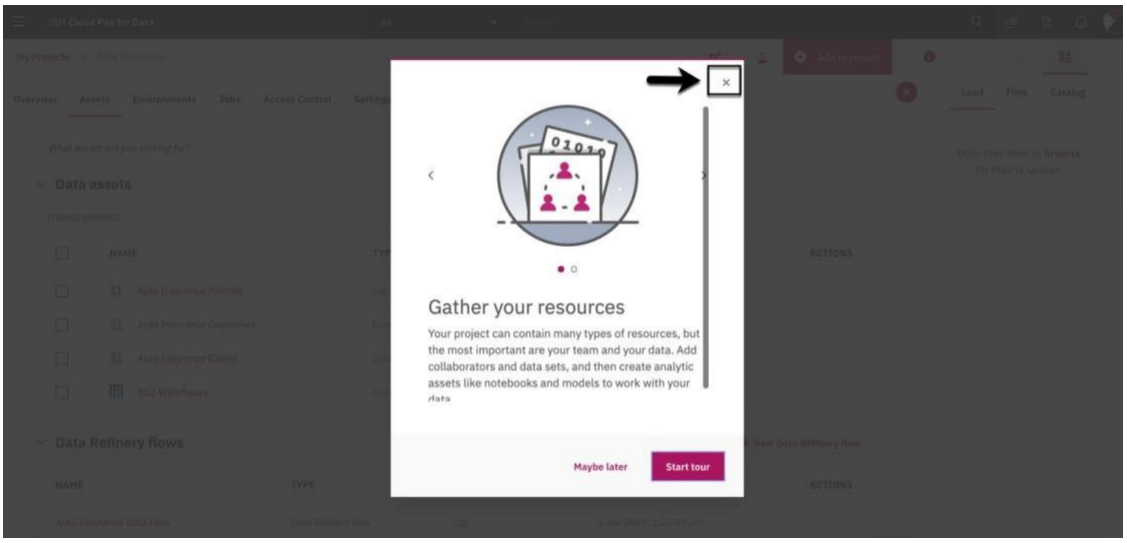
7. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.



8. Click the **Projects** menu.



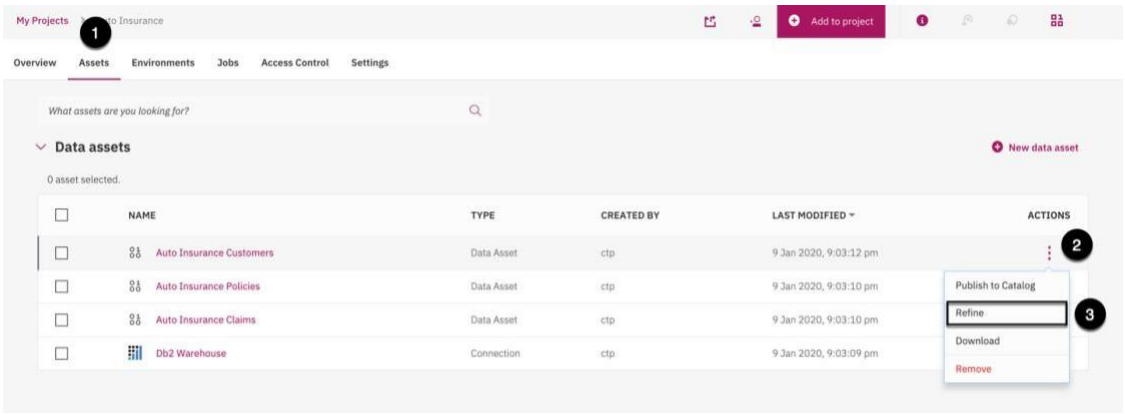
9. From the **Projects** list, select the **Auto Insurance** project.



If you see the Getting Started dialog appear, click on the **X** in the top right corner to close it.

Refine the Data

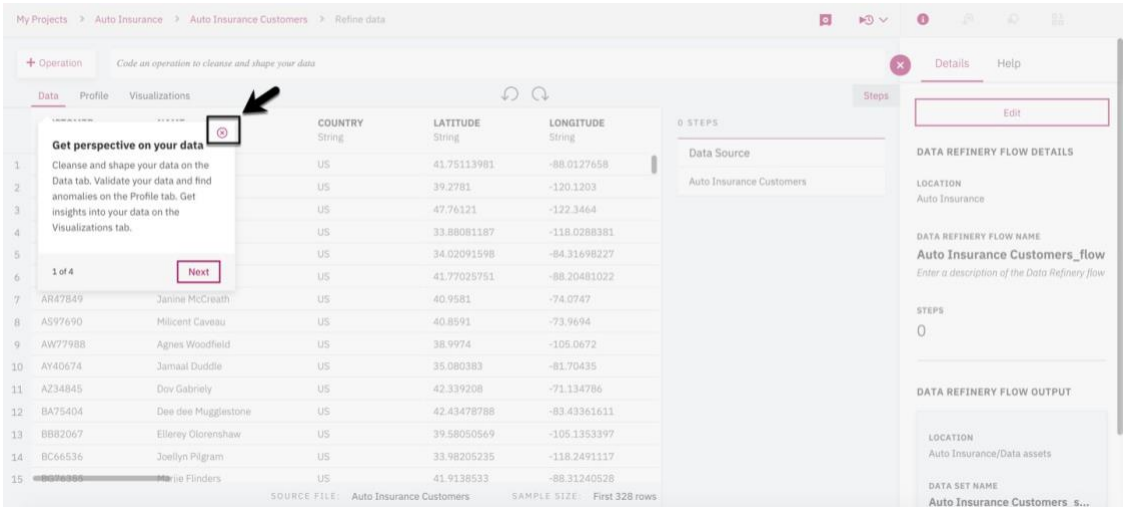
1. Click on the **Assets** tab at the top of the project page.



2. Select the **ellipses…** to the right of the **Auto Insurance Customers** data asset to view the data asset action menu.

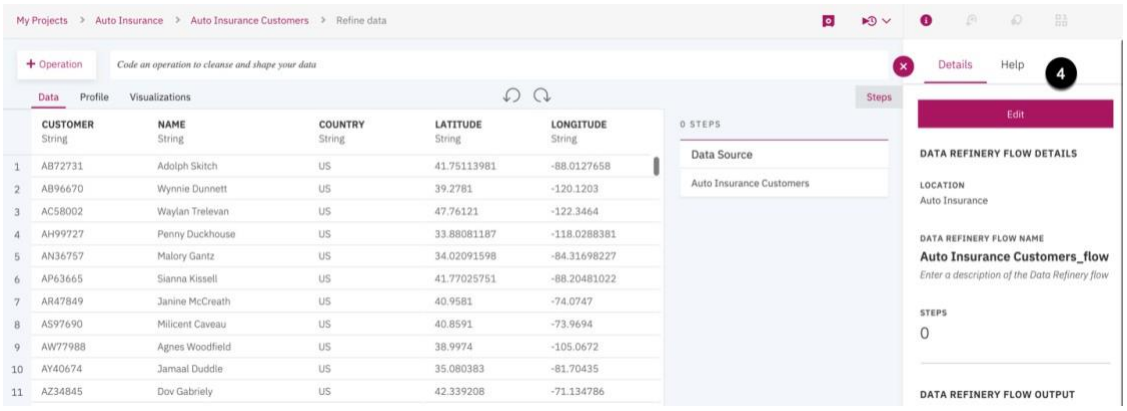
3. Select the **Refine** menu item.

   You are brought into the **data preparation** component of the analytic project to begin shaping the **Auto Insurance Customers** data. In the subsequent steps, you will use some of the data preparation operations to shape the auto insurance 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 by the analytics project team.

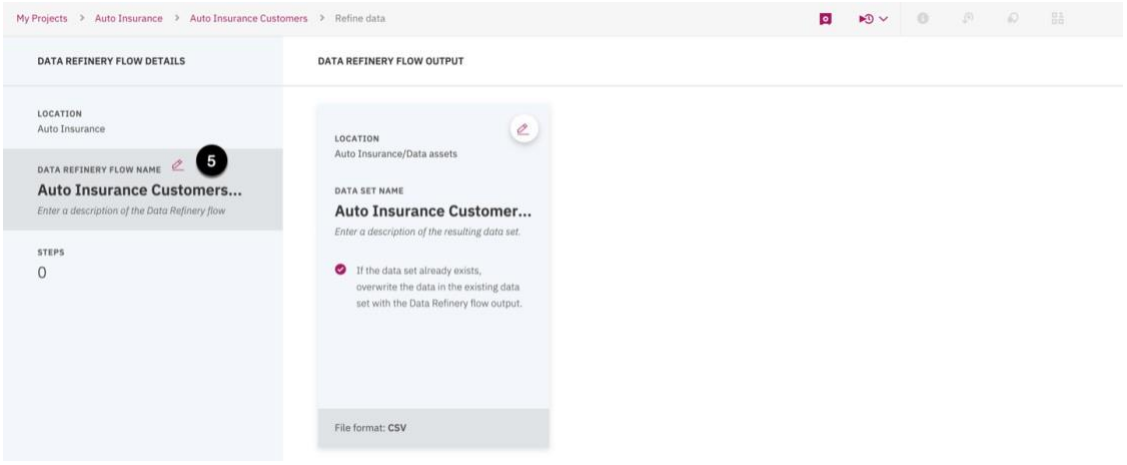


   If you see the Getting Started dialog appear, click on the **X** in the top right corner to close it.

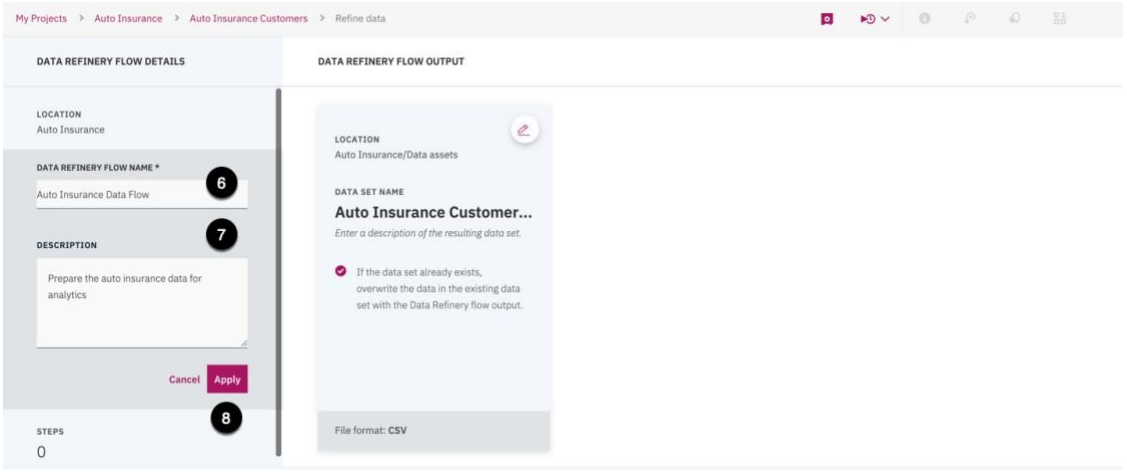
4. Click in the **Edit** button to edit the Data Flow details.



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



6. Rename the Data Flow to **Auto Insurance Data Flow** with the proper case, and spaces between the words.

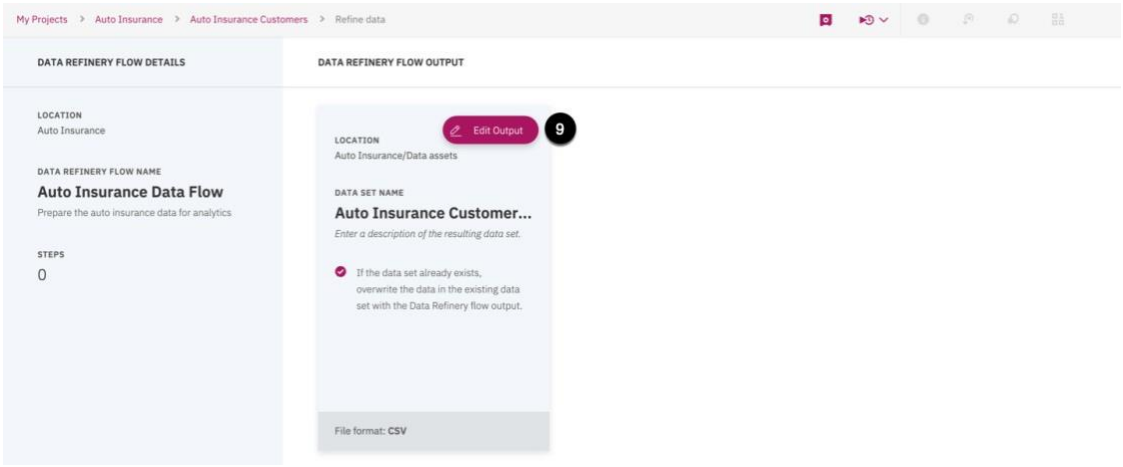


7. Copy and paste, or enter, this bolded text **Prepare the auto insurance data for analytics** into the DESCRIPTION field.

8. Click the **Apply** button.

   Hover over the **pencil icon** in the LOCATION area of the DATA REFINERY FLOW OUTPUT section.

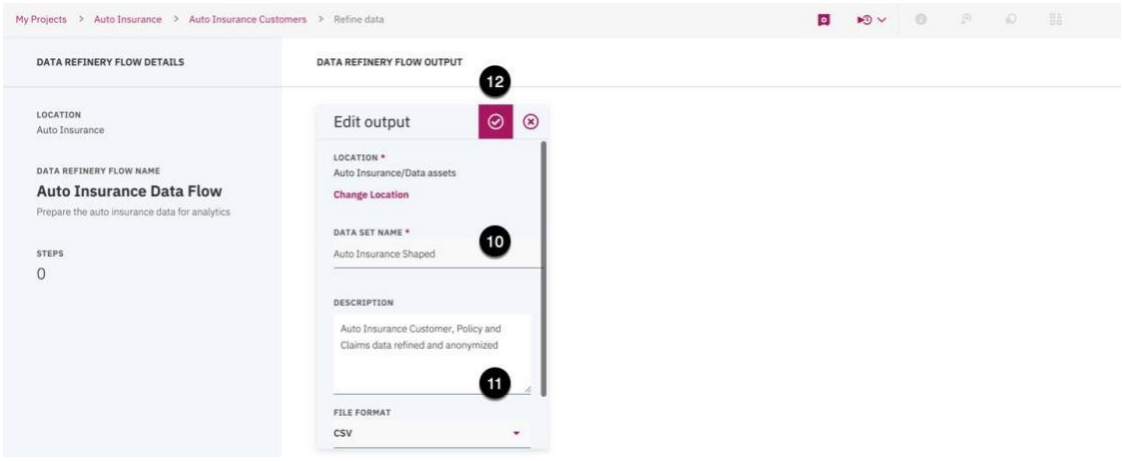
9. Click the **Edit Output** button to change the DATA SET NAME.



   In this section, you can change the data flows output target location. You can choose any connector, supported as a target connector, that is available as part of the Cloud Pak for Data common fabric. However, only connectors defined to the project you are in, that can be targets, will be displayed to select from.

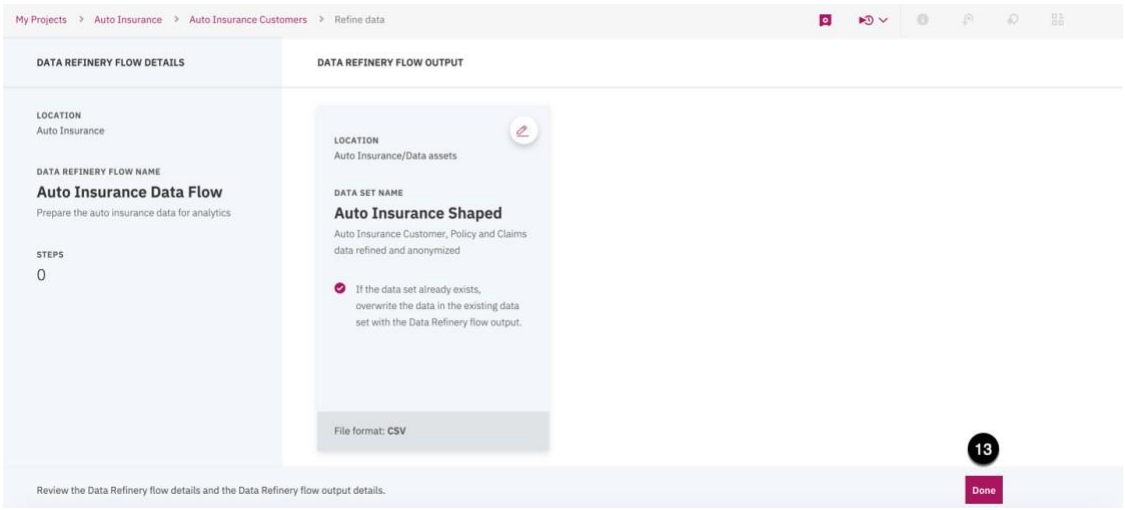
   For this tutorial, you will change the DATA SET NAME but **not** the LOCATION. The target location will be the default location, the **Auto Insurance** project.

10. Rename the Data Set to **Auto Insurance Shaped** with the proper case, spaces between the words, and removal of the **.csv** extension.



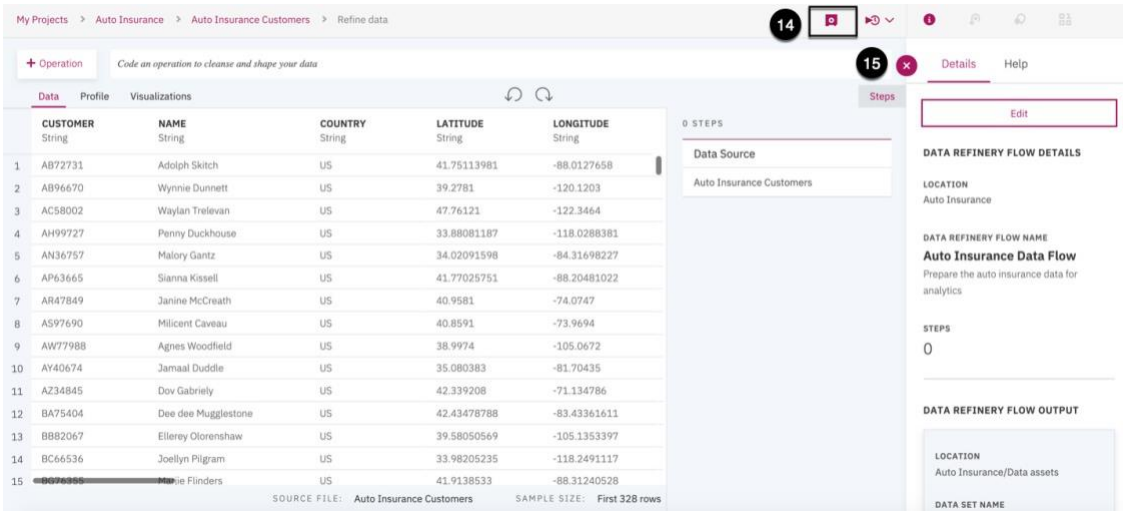
11. Copy and paste, or enter, this bolded text **Auto Insurance Customer, Policy and Claims data combined and refined** into the DESCRIPTION field.

12. Click the **Save** button (looks like a check mark) on the toolbar to save the changes.



13. Click the **Done** button.

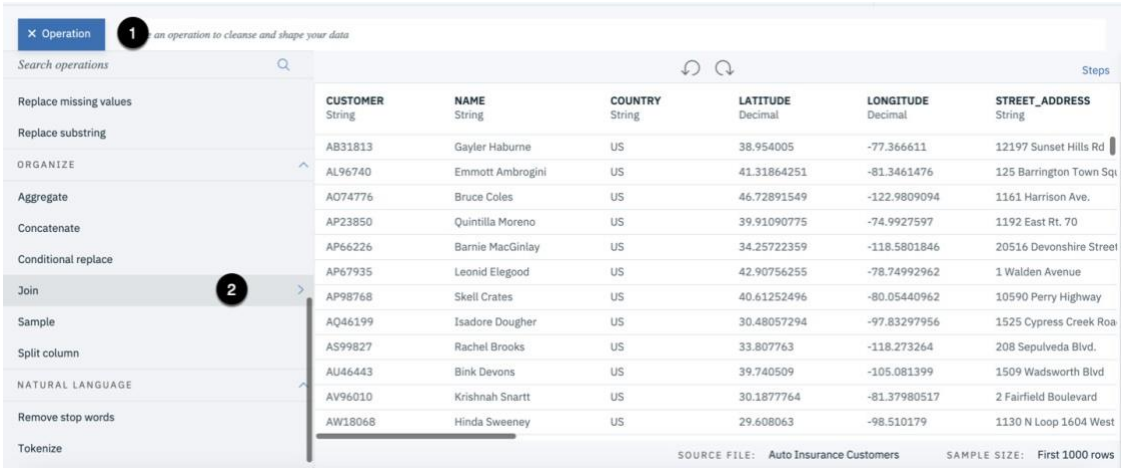
14. Click the **Save** button on the toolbar to save the Data Flow.



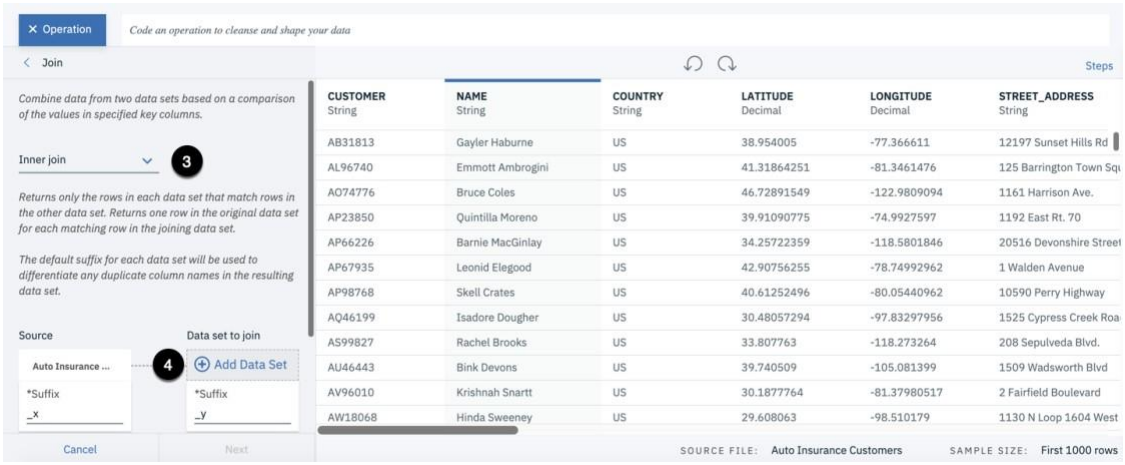
15. Click the **X** on the Details panel to close the panel and maximize the shaper real estate.

Combine Data

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

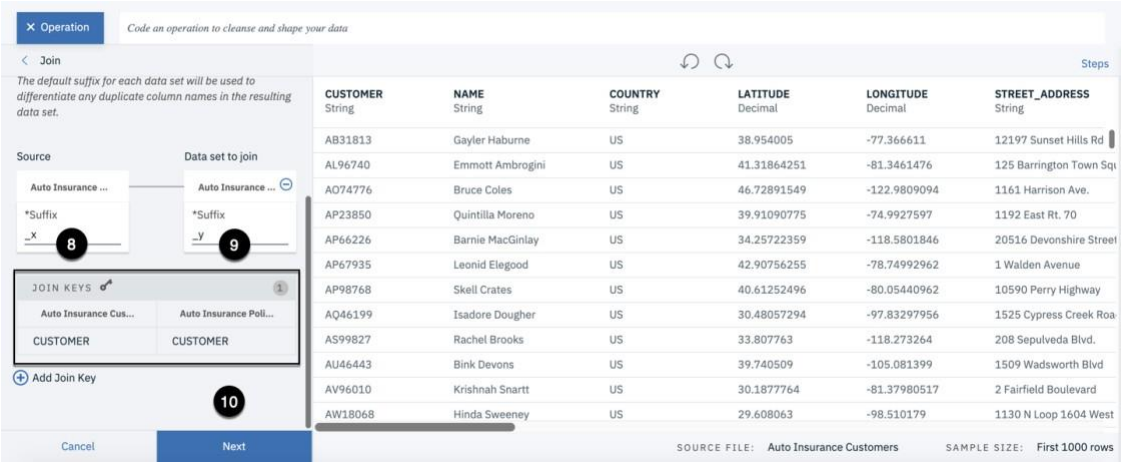


2. Scroll down and click the **Join** operation.



3. Select the **Inner join** method from the join method list.
4. Click the **+ Add Data Set** button in the **Data set to join** section.
5. Click on the **Data assets** section.
6. Select the **Auto Insurance Claims** data asset.
7. Click the **Apply** button.

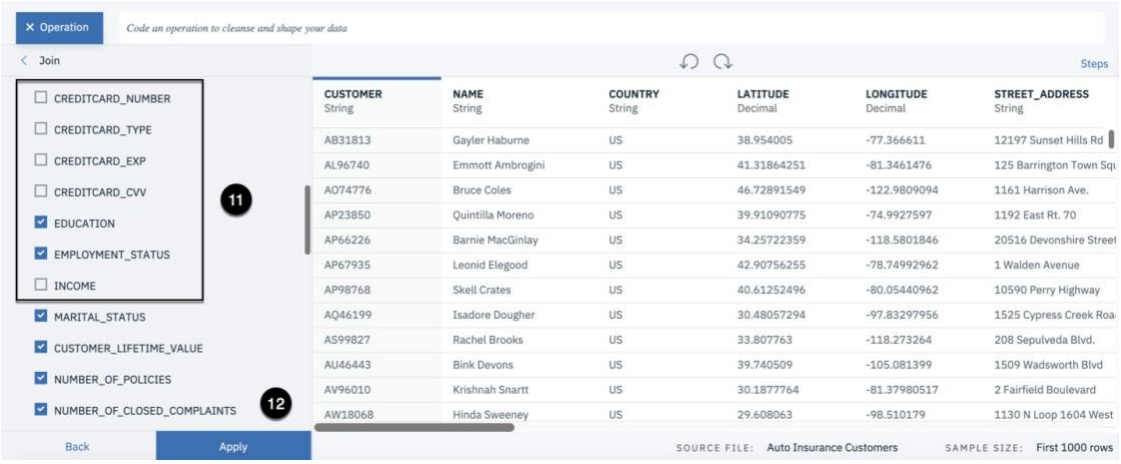
8. Scroll down in the Join properties area until you see the JOIN KEYS section. Click in the **Auto Insurance Customers** JOIN KEYS column selection list on the left and select the **customer_id** column as the join key column.



9. Click in the **Auto Insurance Claims** JOIN KEYS column selection list on the right and select the **customer_id** column as the join key column.

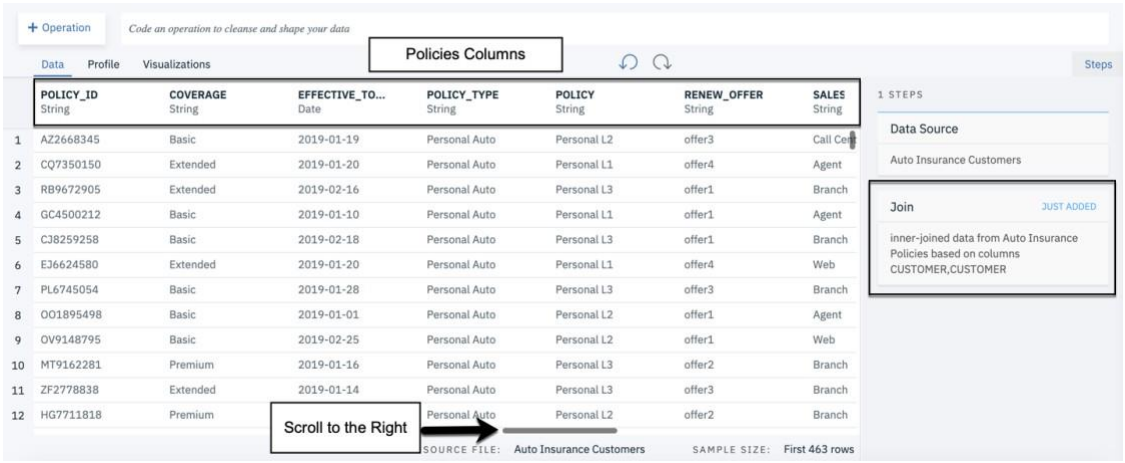
10. Click the **Next** button.

11. Scroll down the column list and **uncheck** the following columns:
`CREDITCARD_NUMBER, CREDITCARD_TYPE, CREDITCARD_EXP, CREDITCARD_CVV`
Unchecking columns excludes them from the join result.



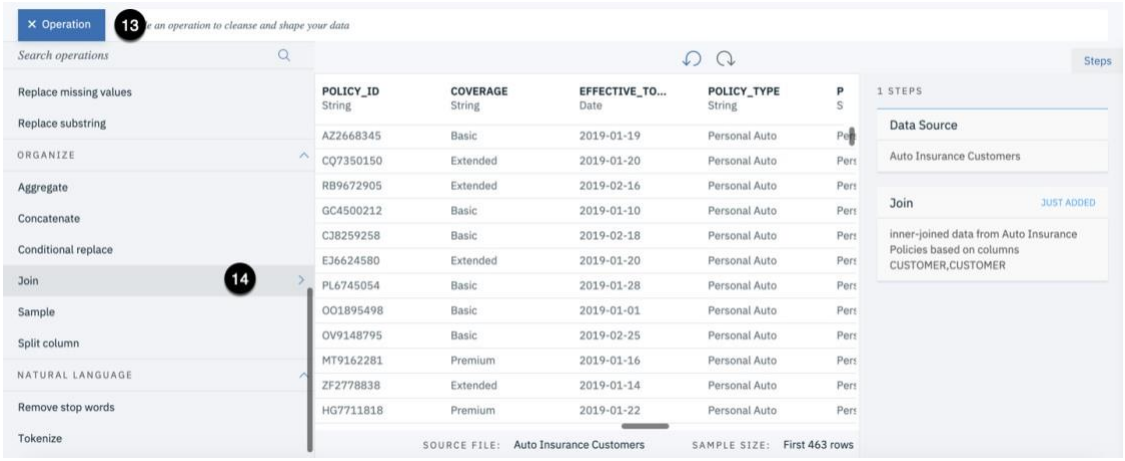
12. Click the **Apply** button.

    The join should complete successfully. **Scroll** to the right to see that the Auto Insurance Policies table columns are now appended at the end of the Auto Insurance Customers table in the shaper.



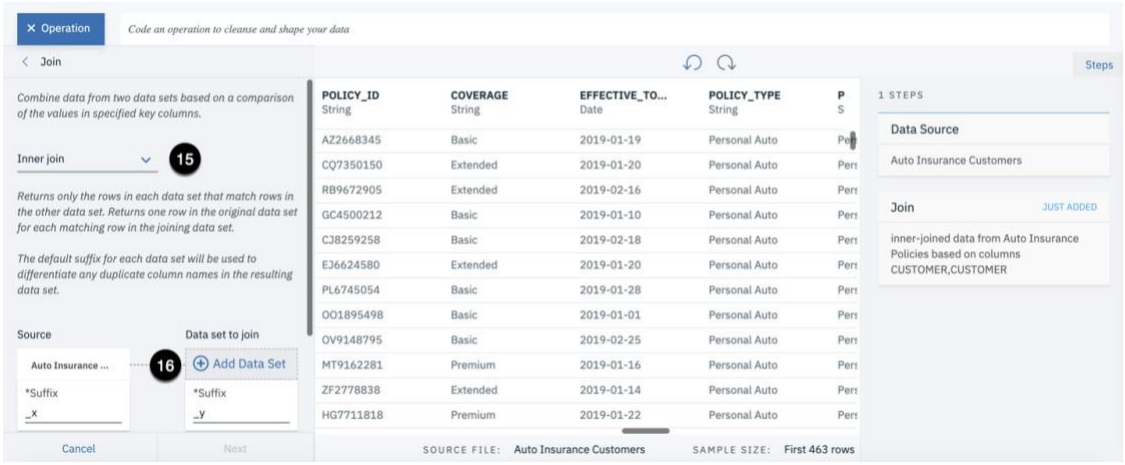
    Notice that the Steps panel appears with the Join as the first shaping step. The Steps panel lets you view your shaping operations, in the order they are performed, and allows for the modification and removal of steps to back out shaping operations done in error or no longer needed.

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



14. Scroll down and Click the **Join** operation.

15. Select the **Inner join** method from the join method list.



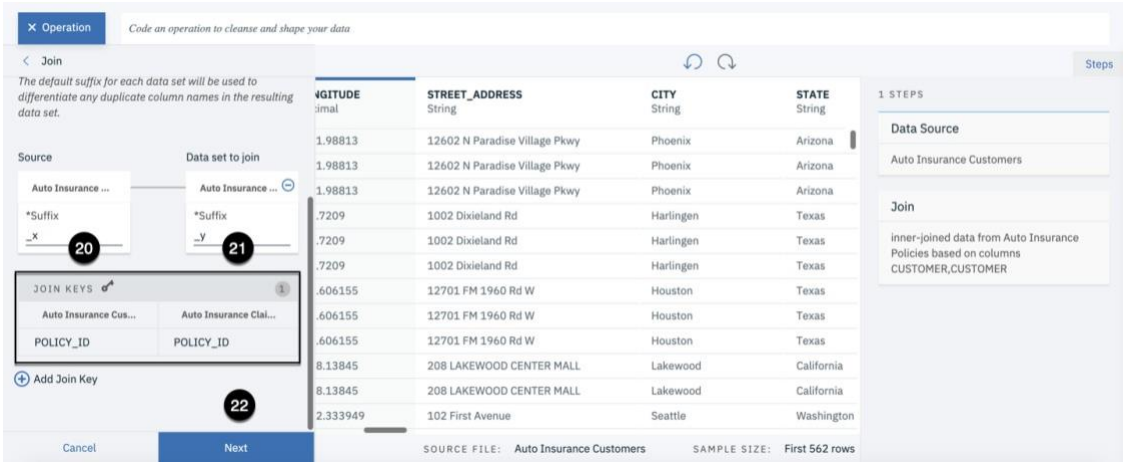
16. Click the **+ Add Data Set** button in the **Data set to join** section.

17. Click on the **Data assets** section.

18. Select the Auto Insurance Policies data asset.

19. Click the **Apply** button.

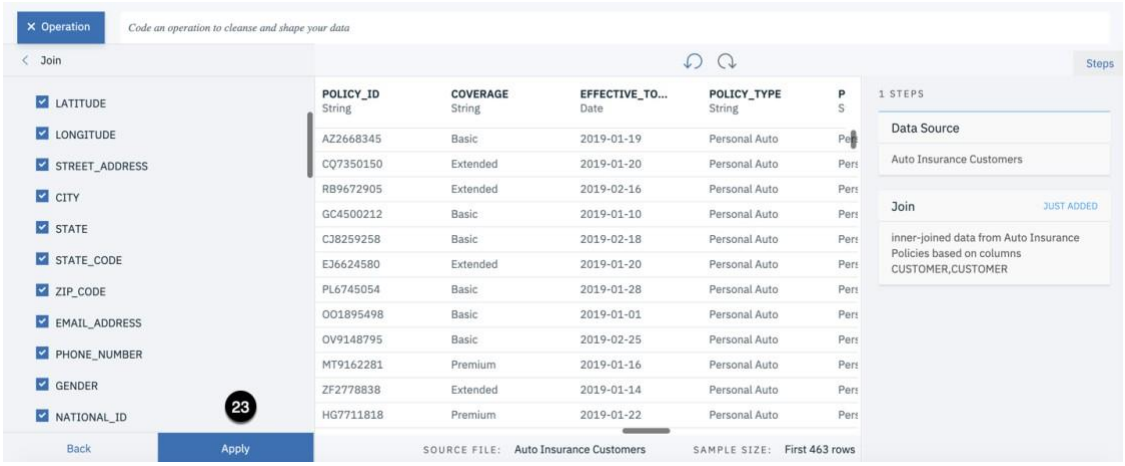
20. Scroll down in the Join properties area to view a full list of columns. Click in the **Auto Insurance Customers** JOIN KEYS column selection list on the left and select the **policy_id** column as the join key column.



21. Click in the **Auto Insurance Claims** JOIN KEYS column selection list on the right and select the **policy_id** column as the join key column.

22. Click the **Next** button.

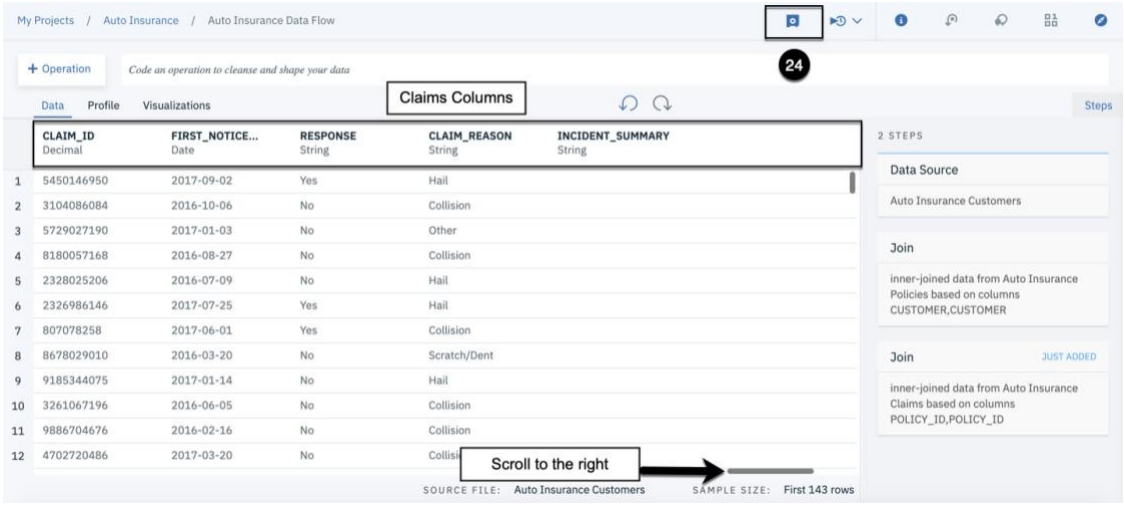
23. Click the **Apply** button.



The join should complete successfully. **Scroll** to the right to see that the Auto Insurance Claims table columns are now appended at the end of the Auto Insurance Customers and Policies table in the shaper.

Notice that the Steps panel appears with the two Joins shaping operations.

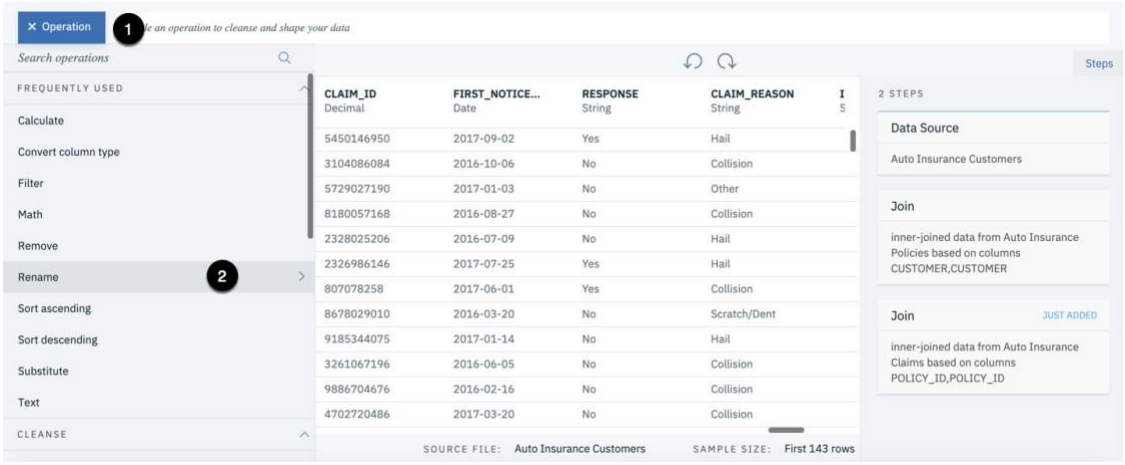
24. Click the **Save** button on the toolbar to save the data flow.



Frequently saving a data flow is a good best practice and ensures you will not lose any of your work. Auto saving will be implemented in a future release.

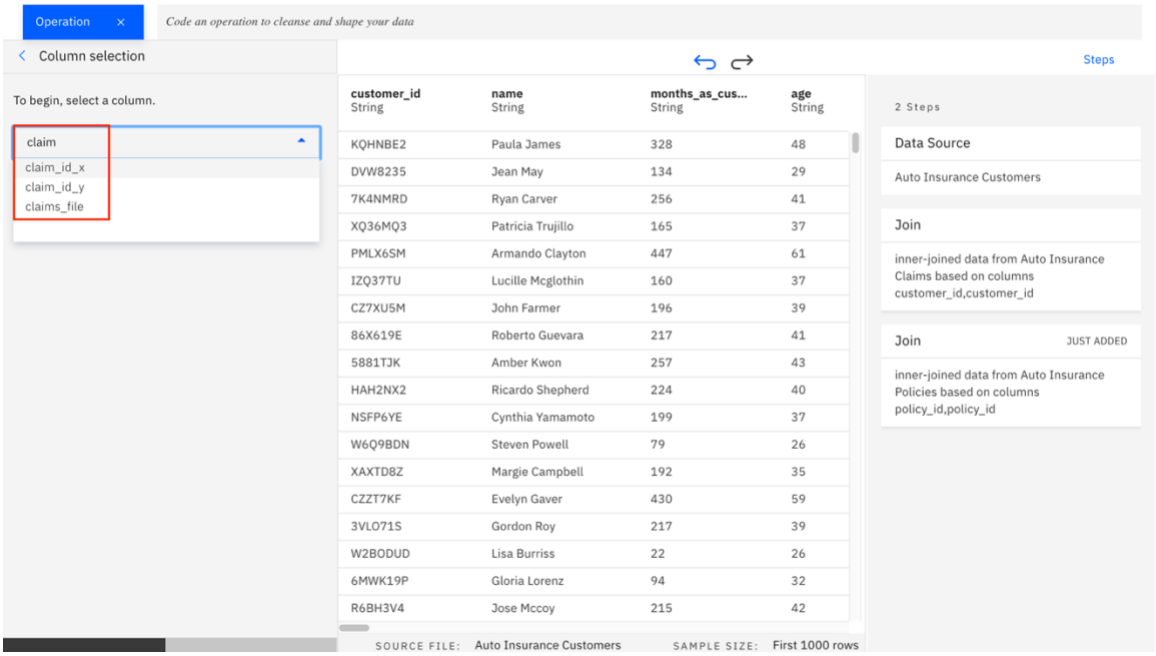
Rename and Remove Data

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



2. Click the **Remove** operation.

3. Click in the column selection list area and start typing the letters **claim**. Select the **Claim_id_x** column from the list.

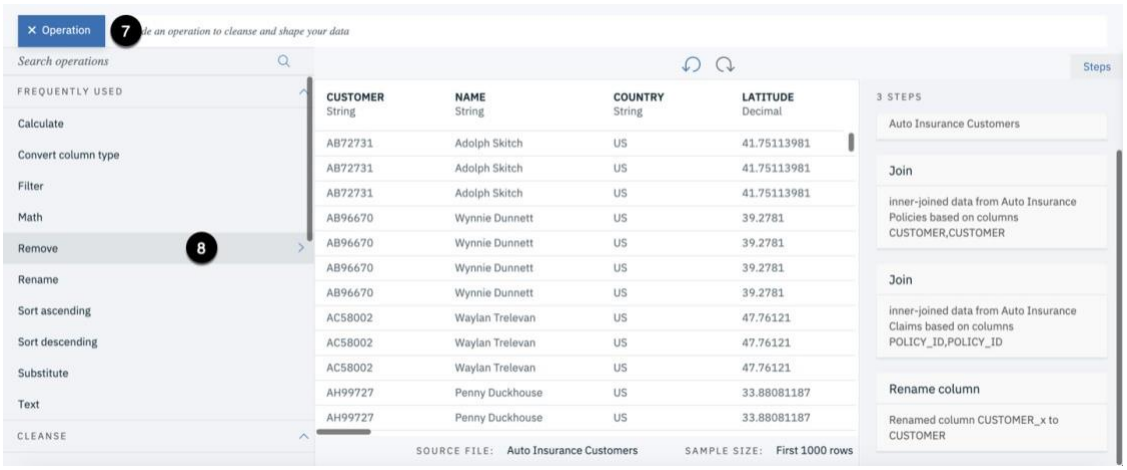


4. Click the **Next** button.

5. Click the **Apply** button.

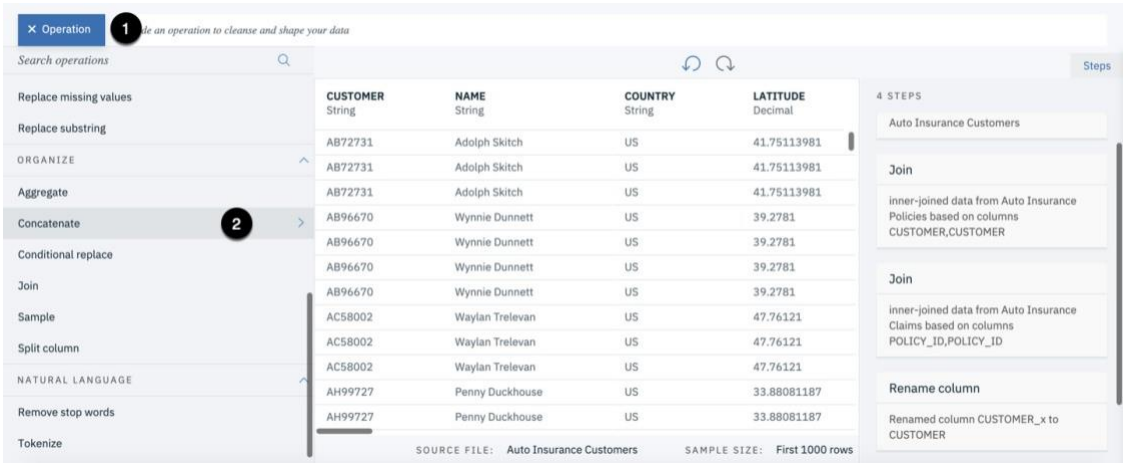
6. Go to the toolbar and Click the **Save** button to save the data flow.

7. Then remove **Claim_id_x** as well.

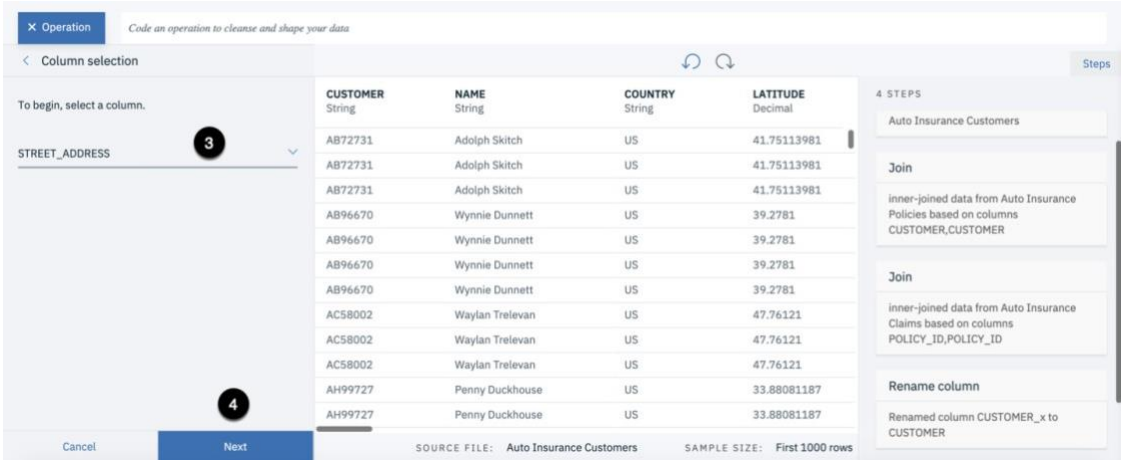


Create New Data

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

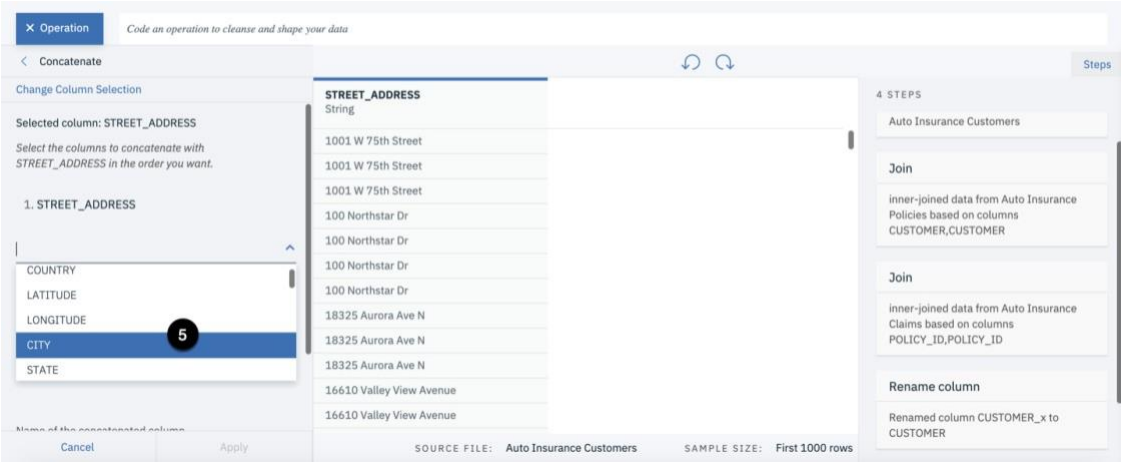


2. Scroll down to the ORGANIZE section and click the **Concatenate** operation.
3. Select the **street_address** column from the column selection list.

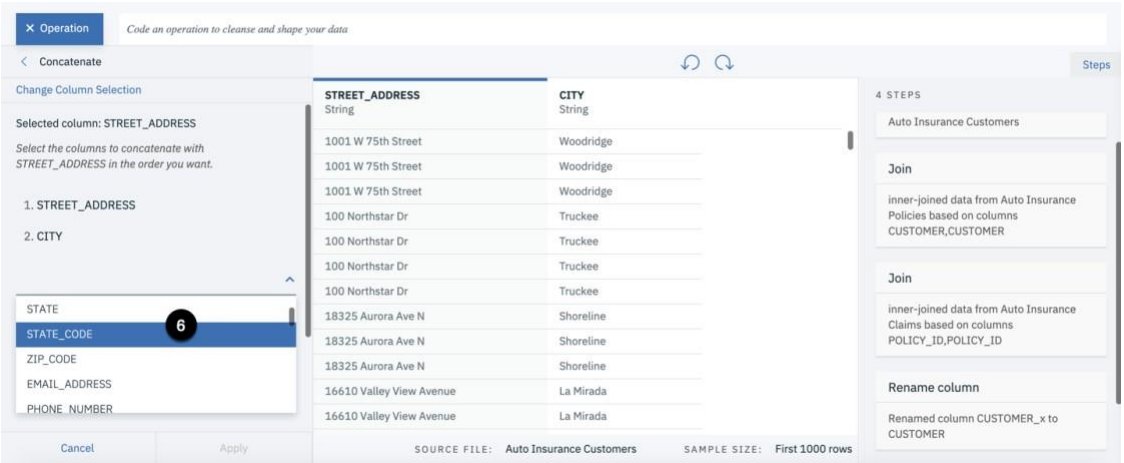


4. Click the **Next** button.

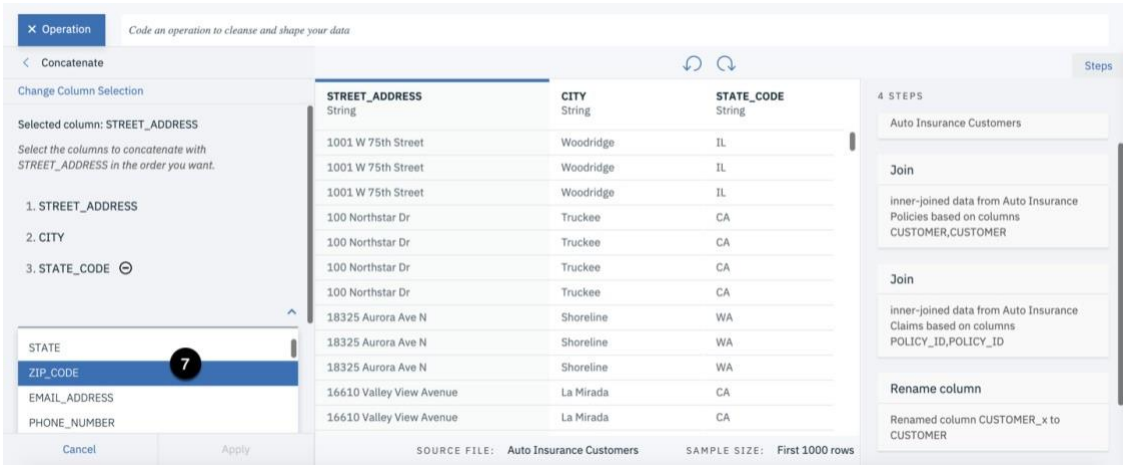
5. Click in the *Select Column* area and select the **city** column from the list of columns to concatenate.



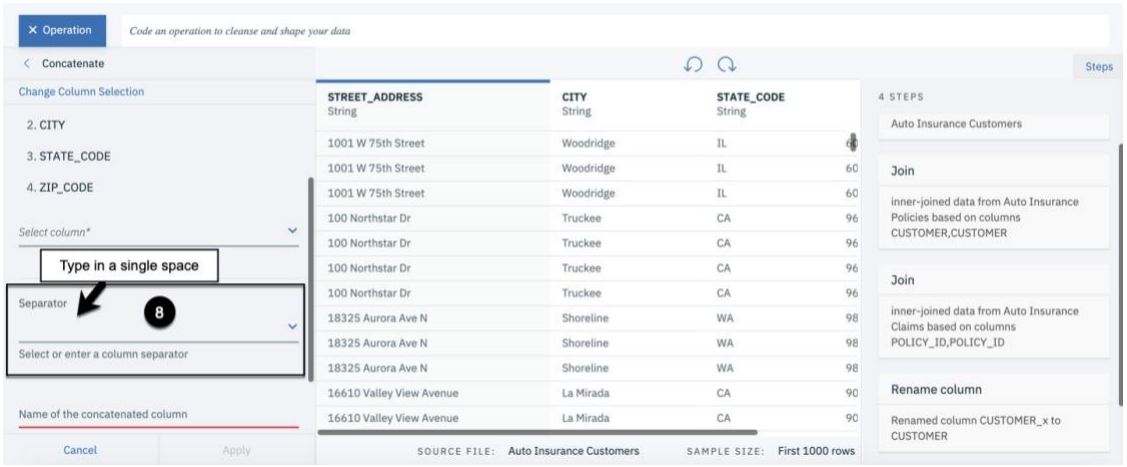
6. Click in the *Select Column* area and select the **state_code** column from the list of columns to concatenate.



7. Click in the *Select Column* area and select the **insured_zip** column from the list of columns to concatenate.

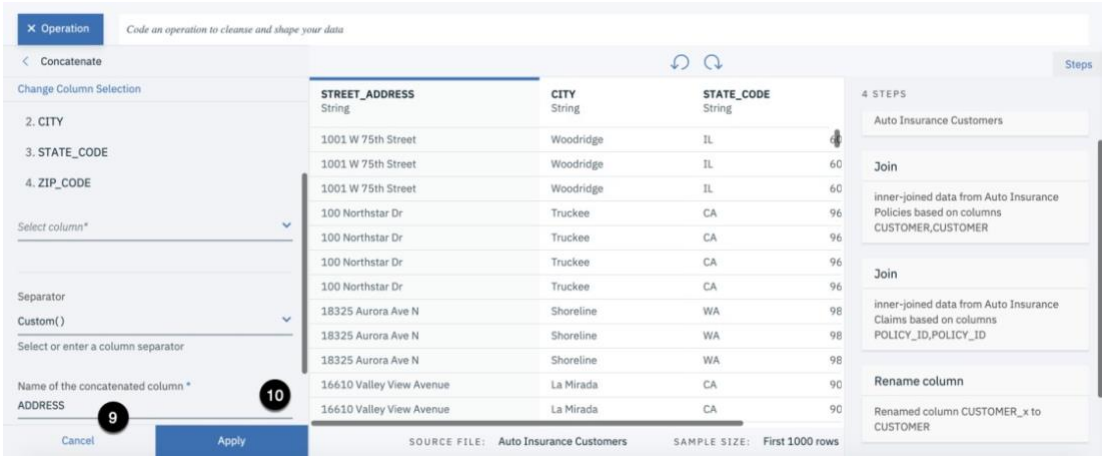


8. Click in the Separator area and **type in a single space** character.



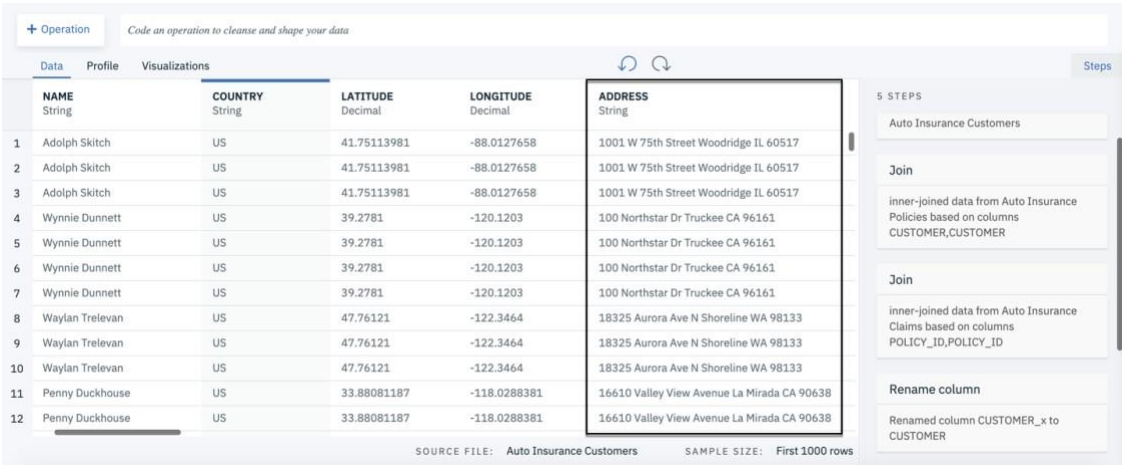
Note: Before proceeding to the next step, to name the new concatenated column, you should see a **Custom( )** entry appear in the Separator field. This was added by Data Refinery because you typed a **space** in the Separator field.

9. Enter **address** as the Name of the concatenated column.



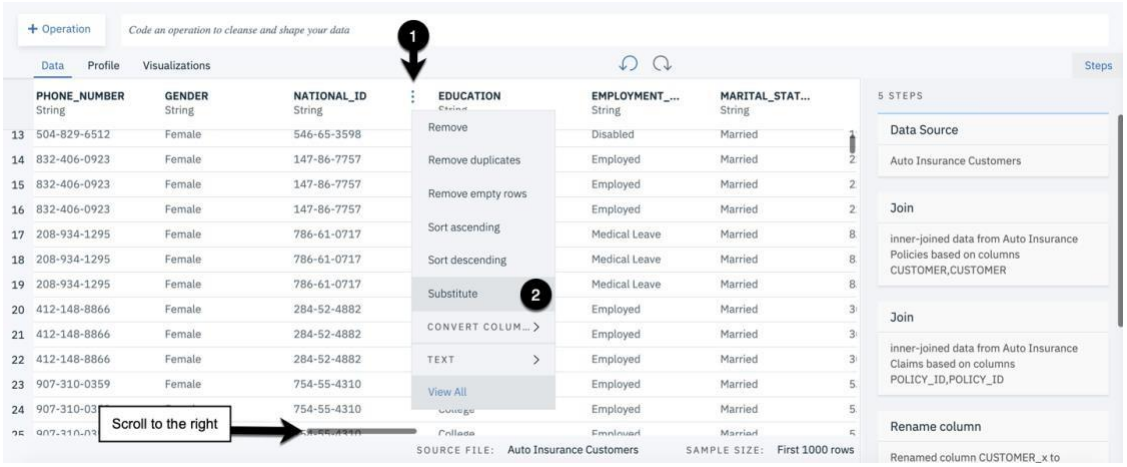
10. Click the **Apply** button.

    As a result of the concatenate operation, you will now see that you have a new column named **ADDRESS** that contains the concatenation of the columns **STREET_ADDRESS**, **CITY**, **STATE_CODE** and **POSTAL_CODE** separated by a space.



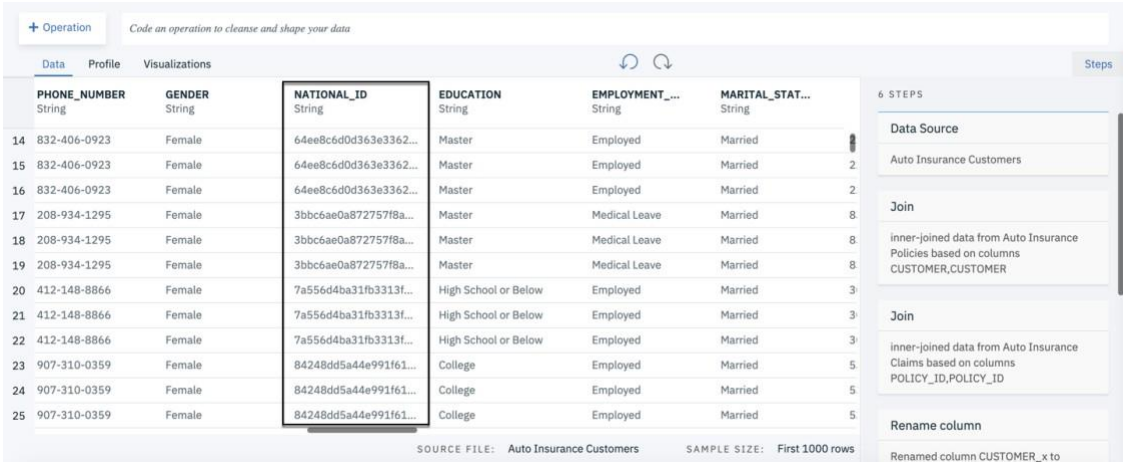
Anonymize Data

1. Scroll to the right and locate the **national_ID** column. Select the **ellipses…** in the top right corner of the **national_id** column to view the column action menu.



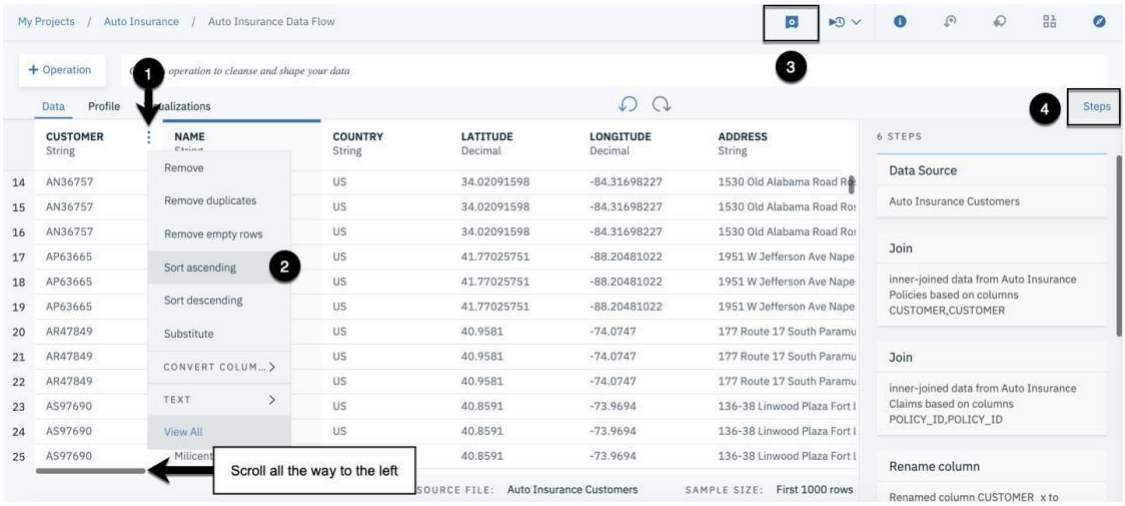
2. Select the **Substitute** menu item.

   The **NATIONAL_ID** column contains a U.S. SSN, which is classified as sensitive information that business users should not have access to. The **Substitute** operation anonymizes the data and replaces the original value with a unique and consistent substituted value to protect the privacy of the information. This column was intentionally included in the join to demonstrate how this operation works. This is another way within IBM Cloud Pak for Data, combined with the data governance capabilities of Knowledge Catalog, to protect sensitive, confidential or personally identifiable information.



Sort Data

1. Scroll all way to the left to the **customer_id** column. Select the **ellipses…** in the top right corner of the **CUSTOMER** column to view the column action menu.



2. Select the **Sort ascending** menu item.

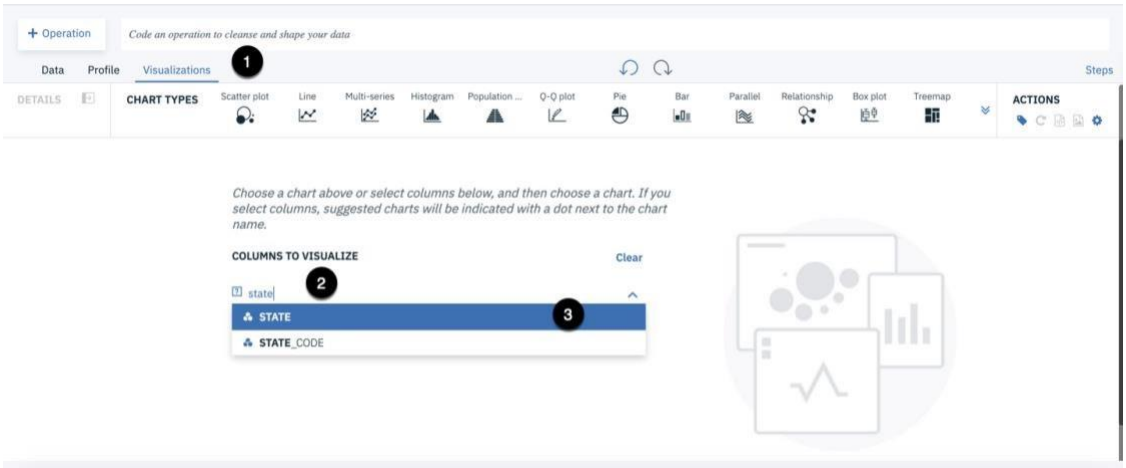
3. Go to the toolbar and select the **Save** button.

4. Click the **Steps** button to hide the steps for more real estate to get ready for the next section.

Understand Data

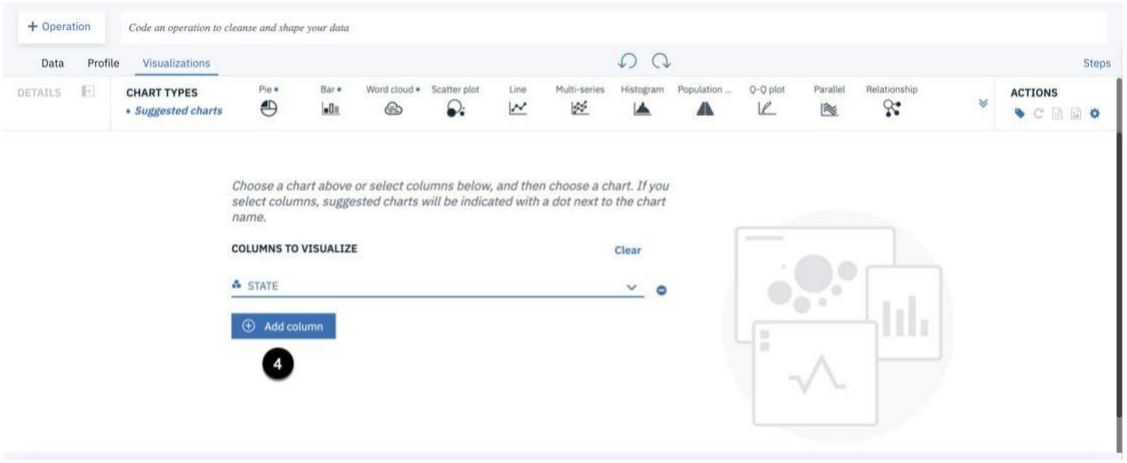
The Data Refinery has built in Visualization to quickly and easily build charts and graphs to better understand data content before shaping and to validate shaping results.

1. Click the **Visualizations** tab.

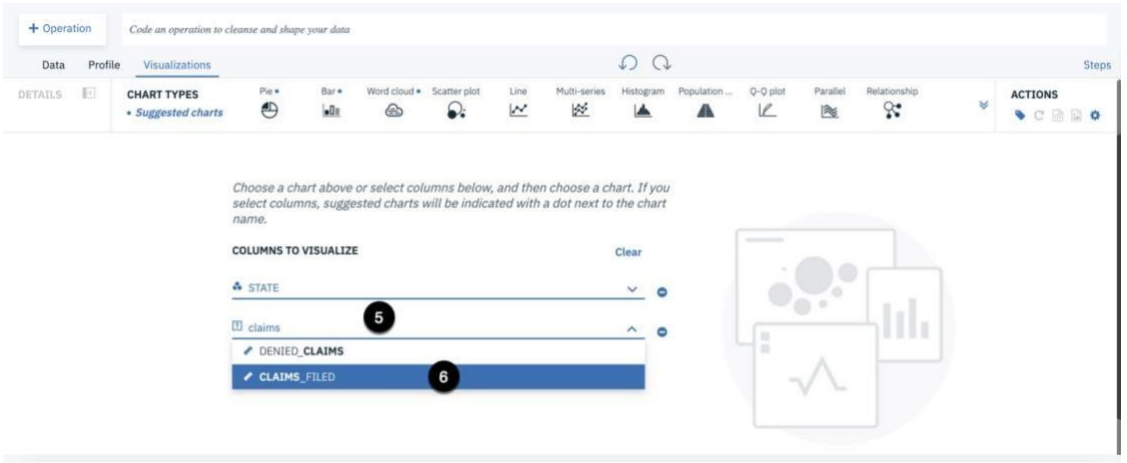


2. Click in the **COLUMNS TO VISUALIZE** area and begin typing the letters **state**.

3. Select the **state** column from the list of columns.
4. Click the **Add Column** button.

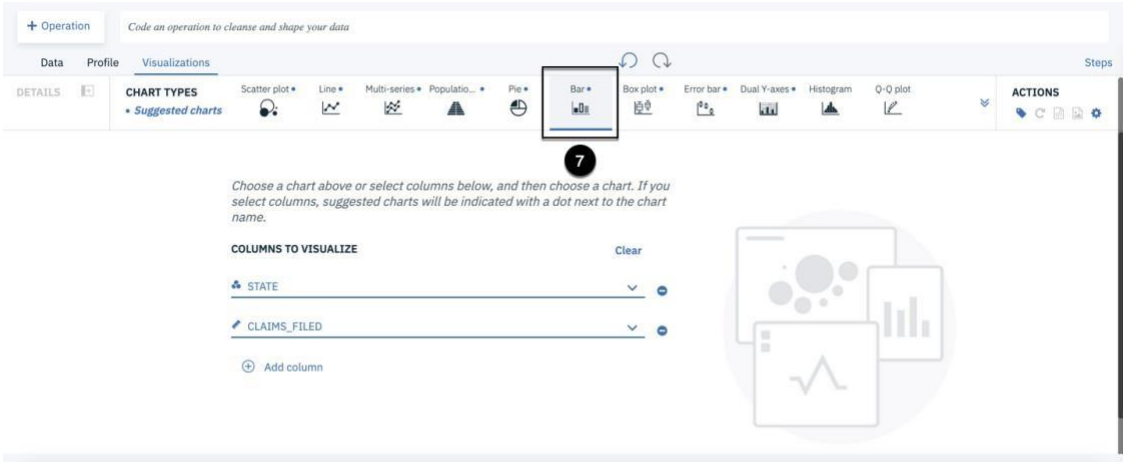


5. Click in the second line of the **COLUMNS TO VISUALIZE** area and begin typing the letters claims**.**

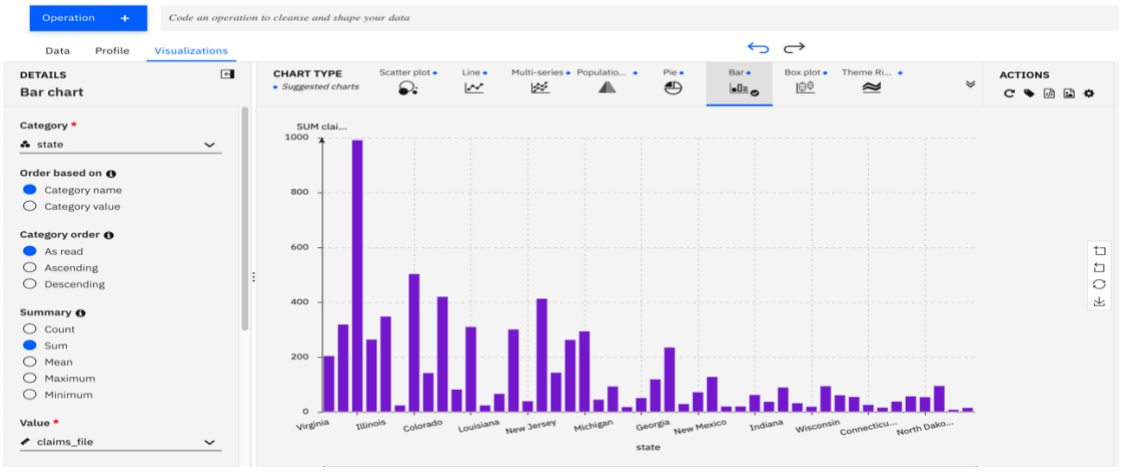


6. Select the **claims_file** column from the list of columns.

7. Click on the **Bar** chart from the **CHART TYPES** toolbar.

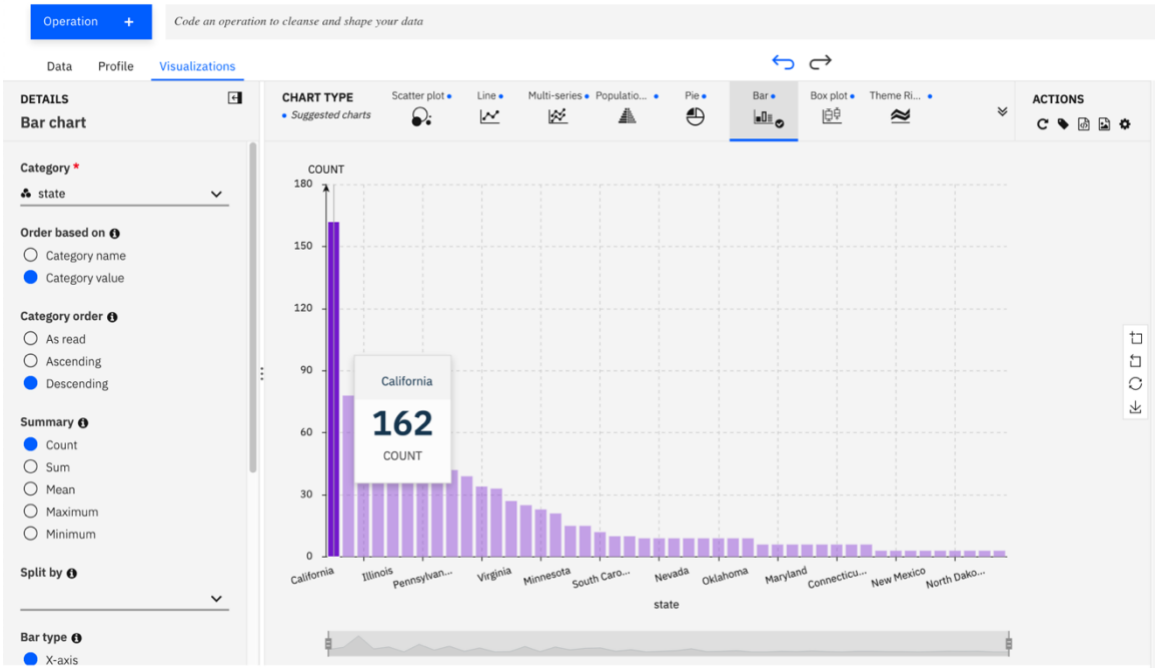


A Bar chart is displayed showing the number of Claims Filed by State.



8. Click the **Category value** radio button in the **Order based on** section.
9. Click the **Descending** radio button in the **Category order** section.
10. Click the **Count** radio button in the **Summary** section. Notice that **California** has the highest number of claims filed.

11. Hover over the very top of the **California** bar in the chart to see the number of claims filed.



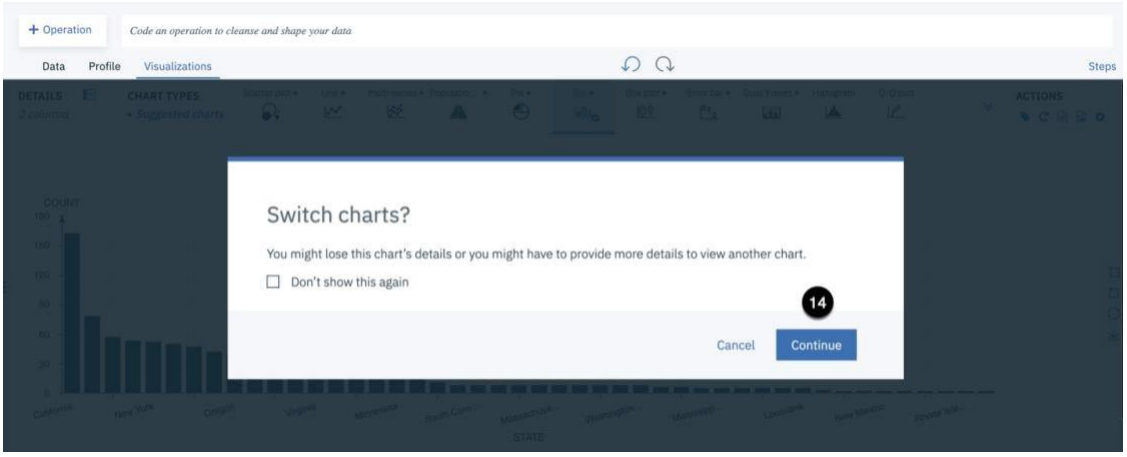
Note: The number of claims filed you see may be different than what is displayed on the screenshot because this is a multi-tenant environment with updates, deletes and inserts happening on a regular basis.

12. Click the **Details** section button to hide the Details section to gain more real estate.

    You can change the chart type, and keep the same properties, by selecting any of the suggested chart types (those that have a **dot** next to them).

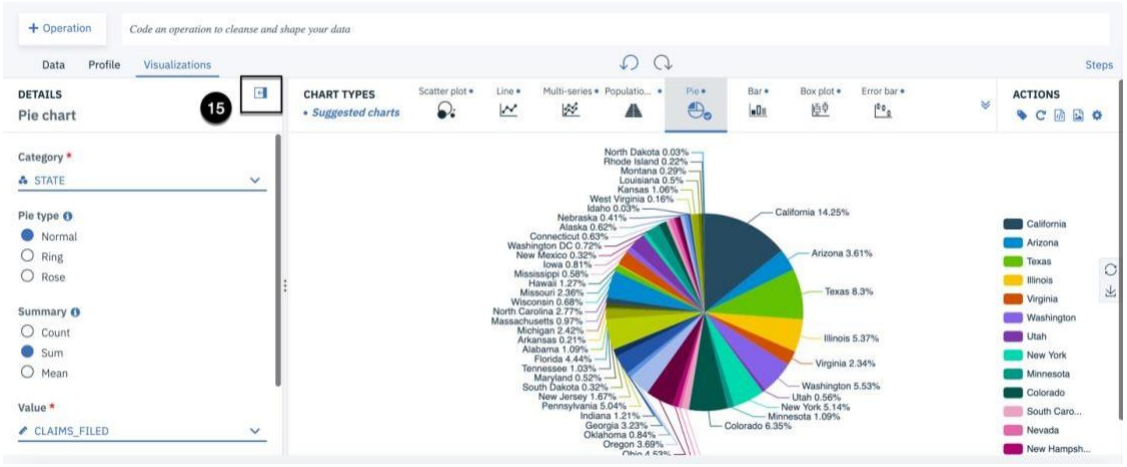
13. Click on the **Pie** chart from the **CHART TYPES** toolbar.

14. When the **Switch charts?** dialog appears, Click the **Continue** button.

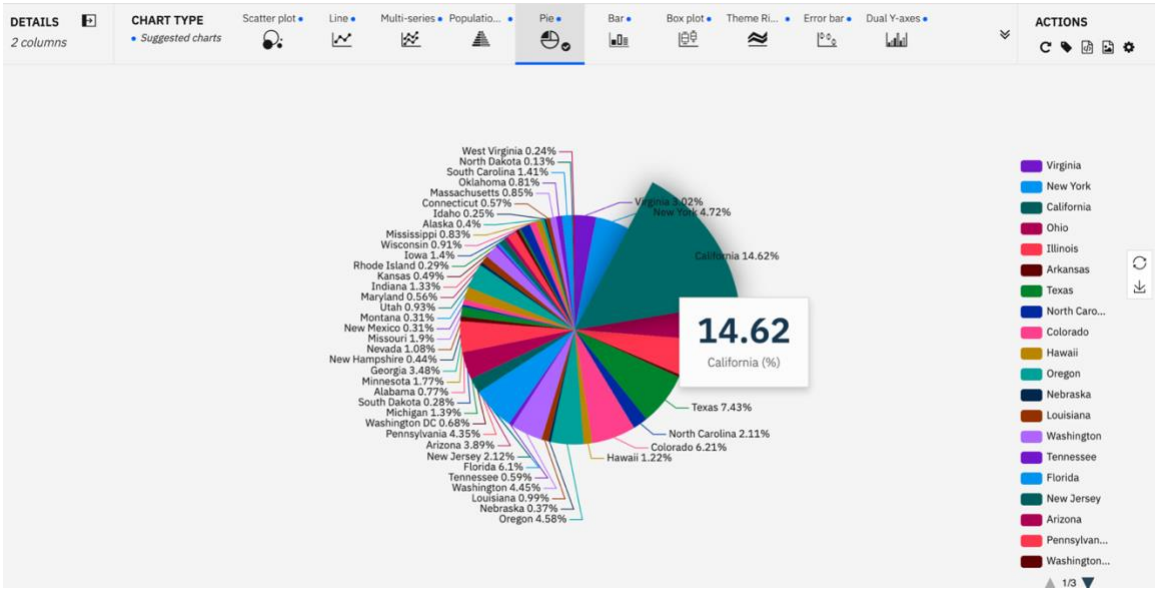


The visualization is changed to a **Pie** chart to render the result.

15. Click the **Details** section button to hide the Details section to gain more real estate.



16. Hover over the state of **California** slice in the pie chart to see the claims filed percentage compared to the other states. This further clarifies that the state of California is the leader in claims filed.



Feel free to experiment more with the **Visualizations** feature of Data Refinery on your own, using the data provided, or your own, to get more experience with the robust capabilities and see how it can assist in better understanding data before and after shaping.

Last, but surely not least, the data refinery has built in **Profiling** to allow for examination and exploration of data statistics and data type classifications to assist users in better understanding data content before and after data shaping.

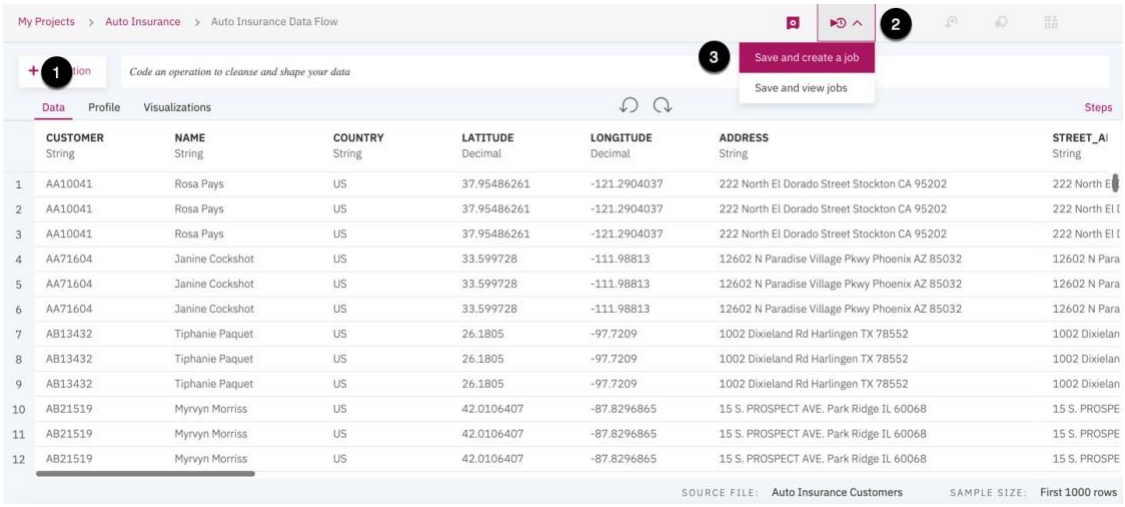
17. Click the **Profile** tab to view the profile.

Scroll to the right and down to view all the column statistics provided.

Run the Data Flow

In order to process the shaping operations, you just performed, 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 **Auto Insurance Shaped** in your **Auto Insurance** project.

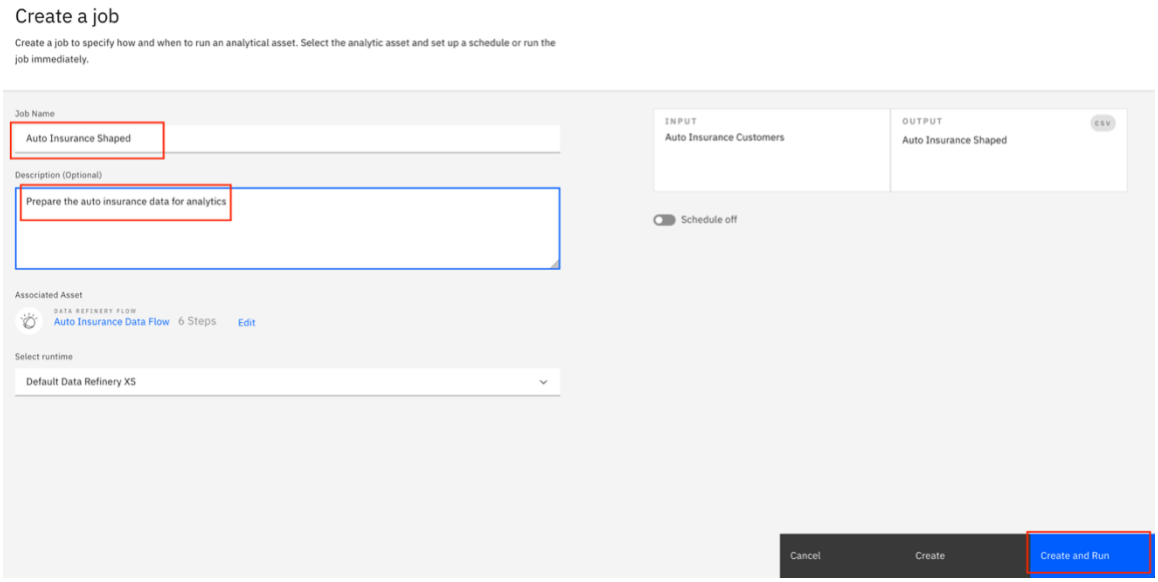
1. Click on the **Data** tab to go back to the data view.



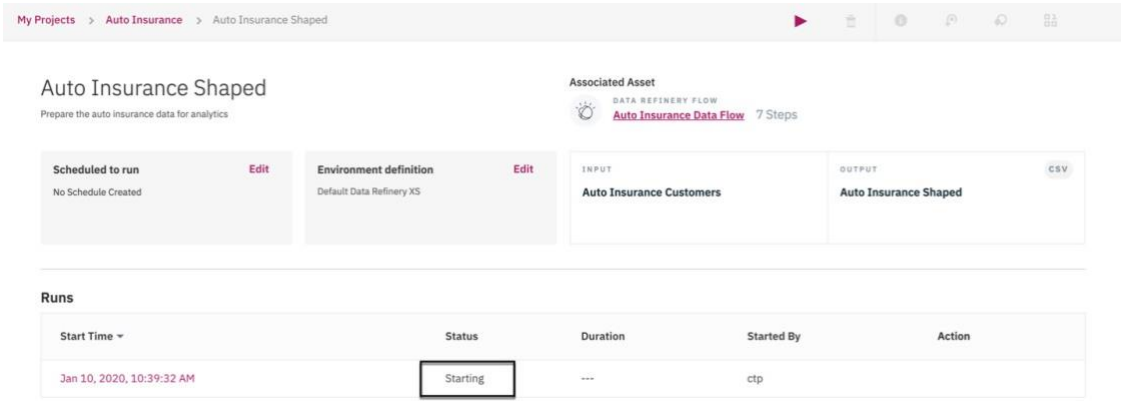
2. Click the **Jobs** button on the toolbar.

3. Select the **Save and create a job** menu item.

4. Enter a Job Name of **Auto Insurance Shaped** with the proper case, and spaces between the words.



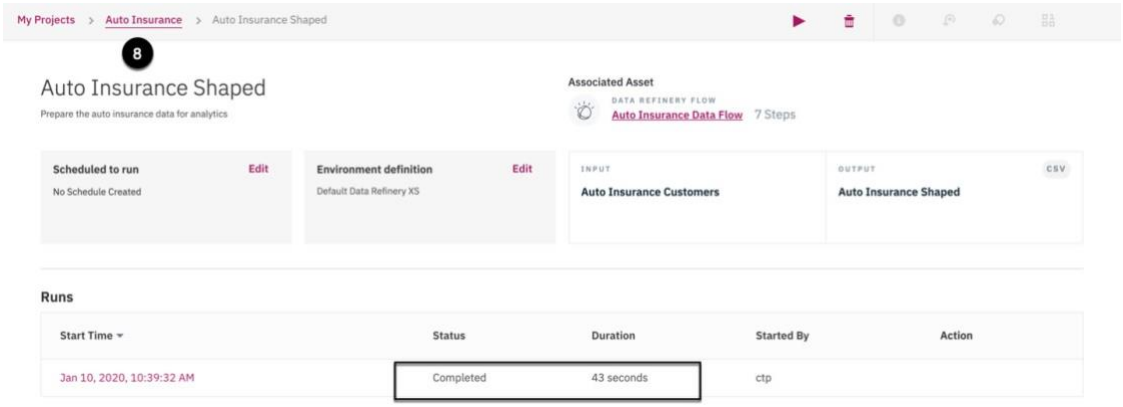
5. Copy and paste, or enter, this bolded text: **Prepare the auto insurance data for analytics** into the job **Description** field.
6. Use the **Default Data Refinery XS** runtime, it should be pre-selected.
7. Click the **Create and Run** button.



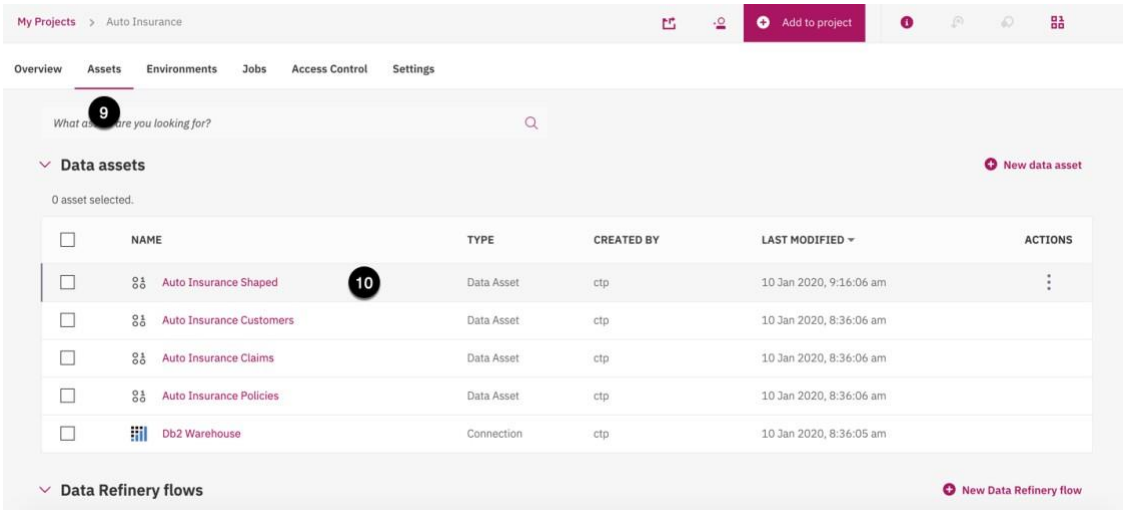
The status will change from **Queued** to **Starting** to **Running** to **Completed**.

You can use your browser's refresh function to refresh the page to see the data flow status updates. Wait until the data flow status changes to **Completed** before proceeding to the next step. It should take a minute or less to finish.

8. Click on the **Auto Insurance** project navigation link on the toolbar to get back to the sections of the project.



9. You should be taken to the **Assets** tab of the project. If not, click on the **Assets** tab to view the project assets.

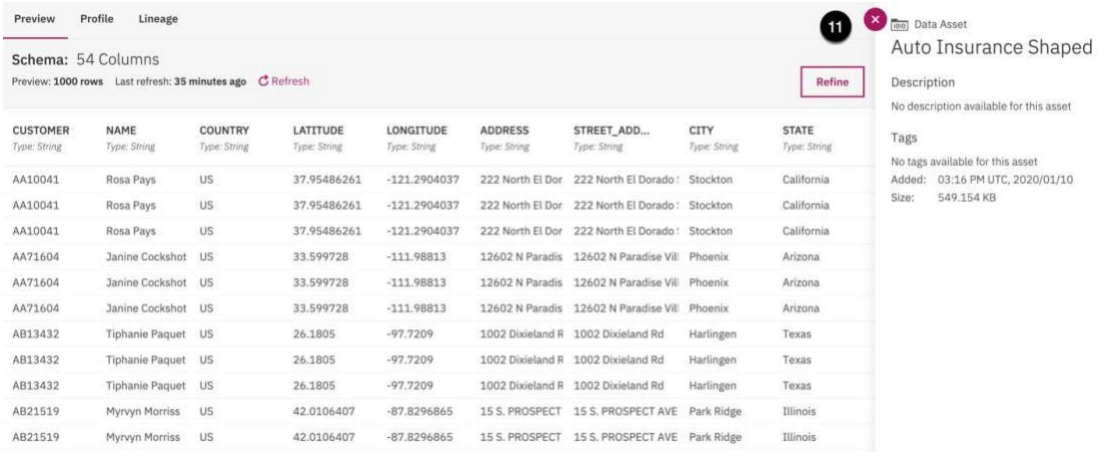


Notice that you now have a new data asset named **Auto Insurance Shaped**. This is the CSV dataset the Data Refinery generated based on your shaping recipe.

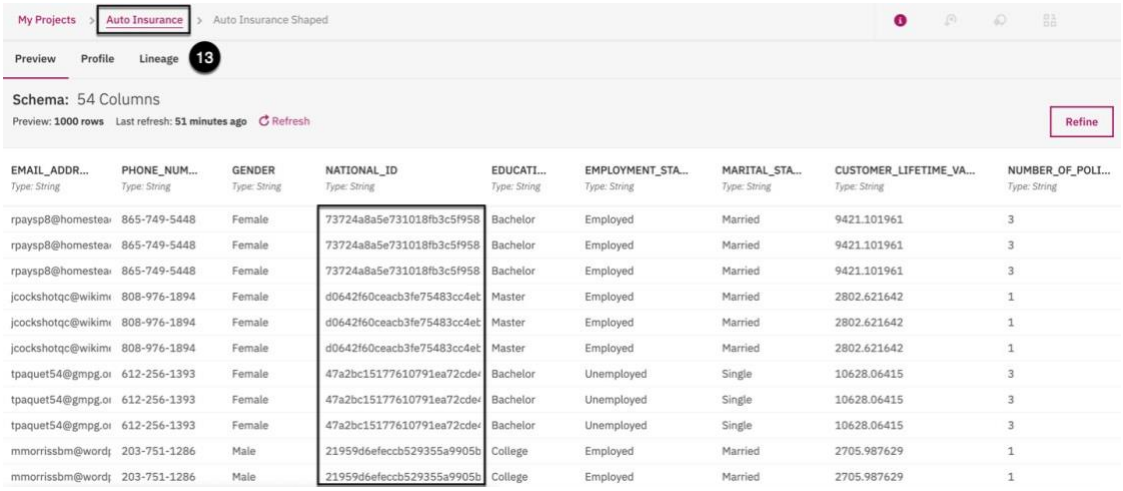
10. Hover over the **Auto Insurance Shaped** data asset and click on it to preview the data and verify the data flow results.

    Notice that the **national_id** column was anonymized and that the **claim_id_y** column has been removed.

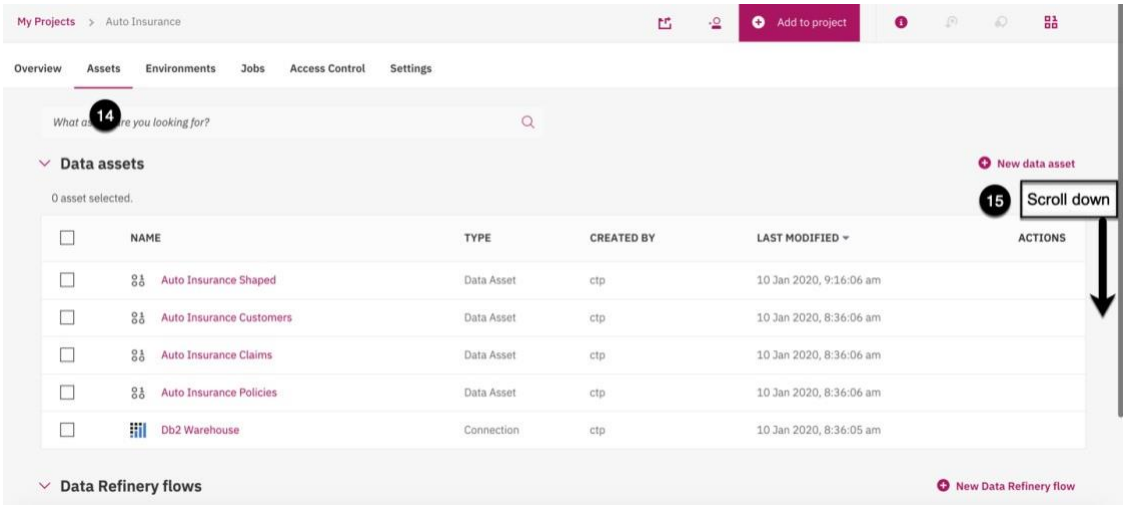
11. Click the **X** to close the preview.



12. Click on the **Auto Insurance** project navigation link on the toolbar to go back to the project home page. You should be taken to the **Assets** tab of the project. If not, click on the **Assets** tab to view the project assets.

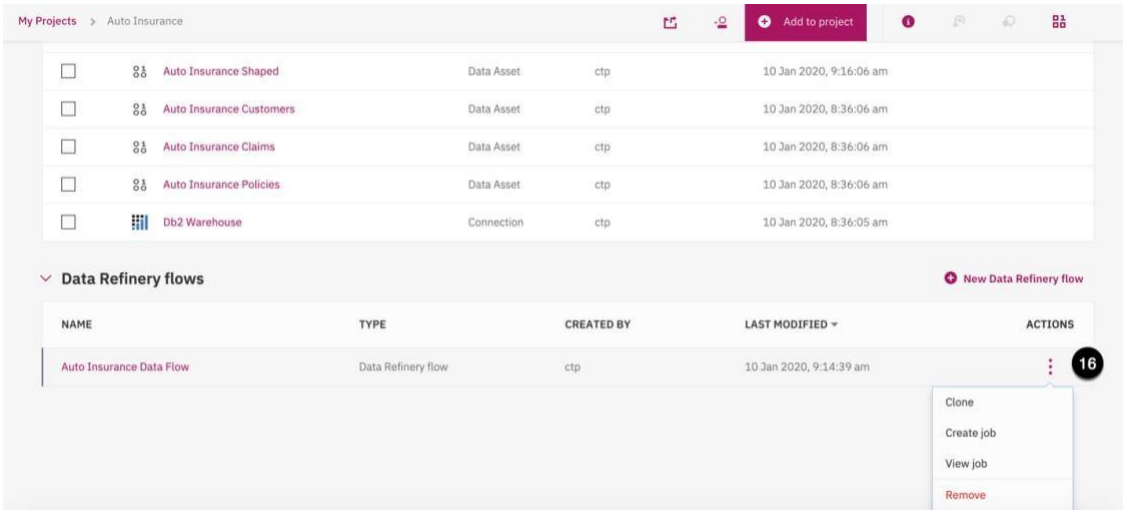


13. Scroll down to the **Data Refinery flows** section of the **Assets** tab.



Once a data flow is saved, it is placed in the **Data Refinery flows** section of the project. You should see your saved **Auto Insurance Data Flow**.

14. Select the **ellipses…** to the far right under the data flow **ACTIONS** column.



Note: You may have to scroll down in the UI to see the entire menu depending on the browser you are using.

You can perform the following actions from the data flow actions menu:

- **Clone** - Creates a copy of the data flow. The flow is added to the Data Refinery flows list as “original-name copy 1”.
- **Create job** - Opens the job creation dialog to create a new job.
- **View job** - Opens the job page for data flow.
- **Remove** - Deletes the data flow from your project.

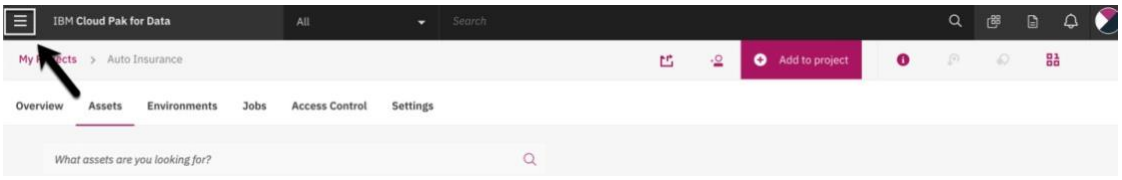
You can also click on the **data flow** to open it in the data flow shaper to modify it and view the data flow’s result set and recipe steps.

Protect Sensitive Information

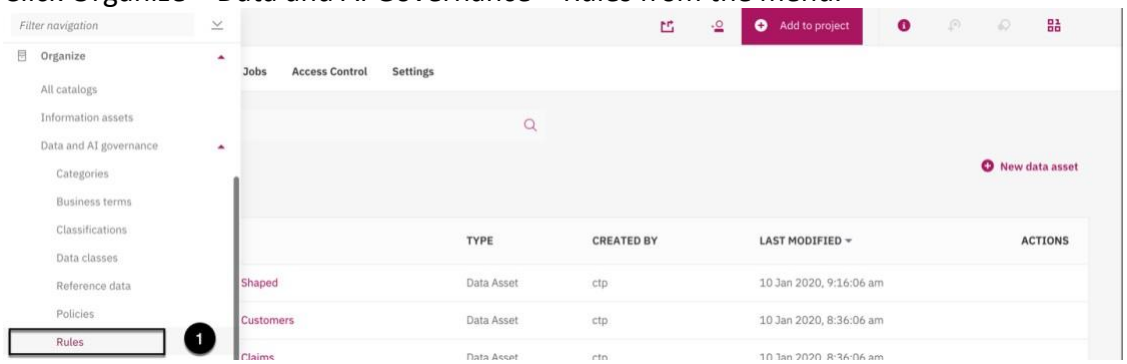
In this section you will learn how to protect sensitive information by creating **Data Protection Rules**. You will create data protection rules to obfuscate (i.e. Mask) **US Social Security Numbers** and redact **Credit Card Information** and then validate that they are being enforced.

Create Data Protection Rules

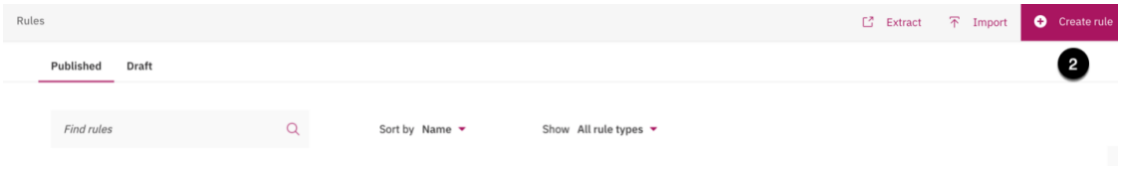
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.



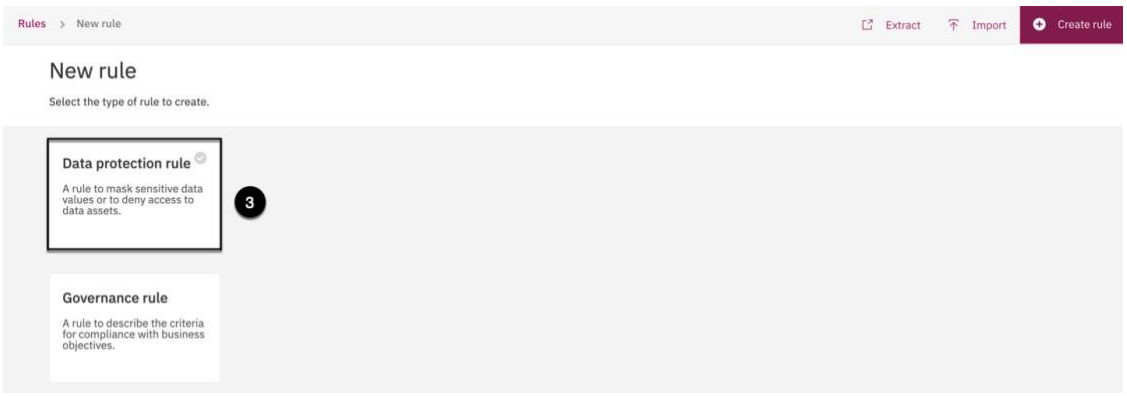
Click Organize > Data and AI Governance > Rules from the menu.



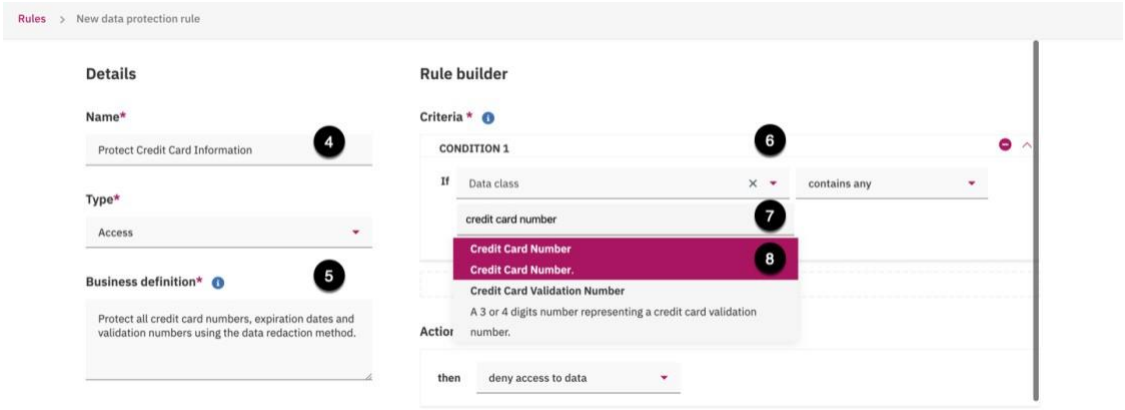
2. Click the **Create rule** button.



3. Click the **Data protection rule** type.



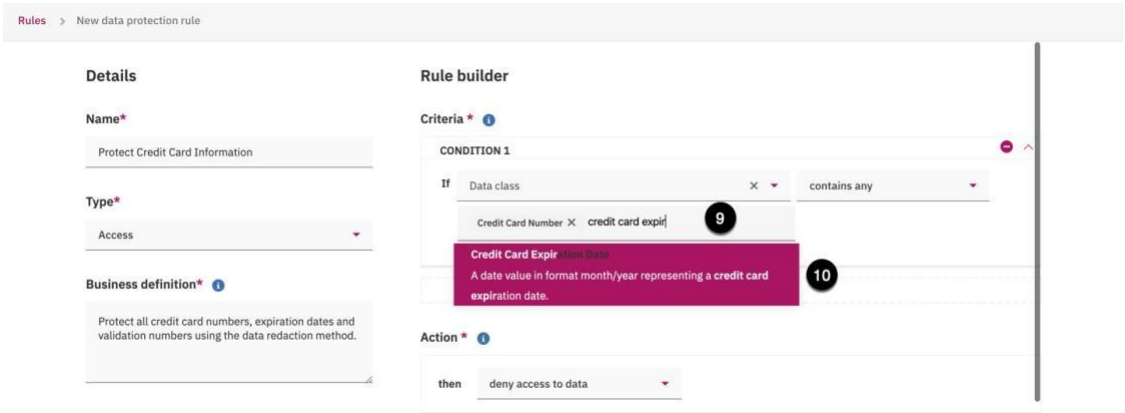
4. Enter a Name of **Protect Credit Card Information**.



5. Copy and Paste the following bolded text into the **Business definition**:

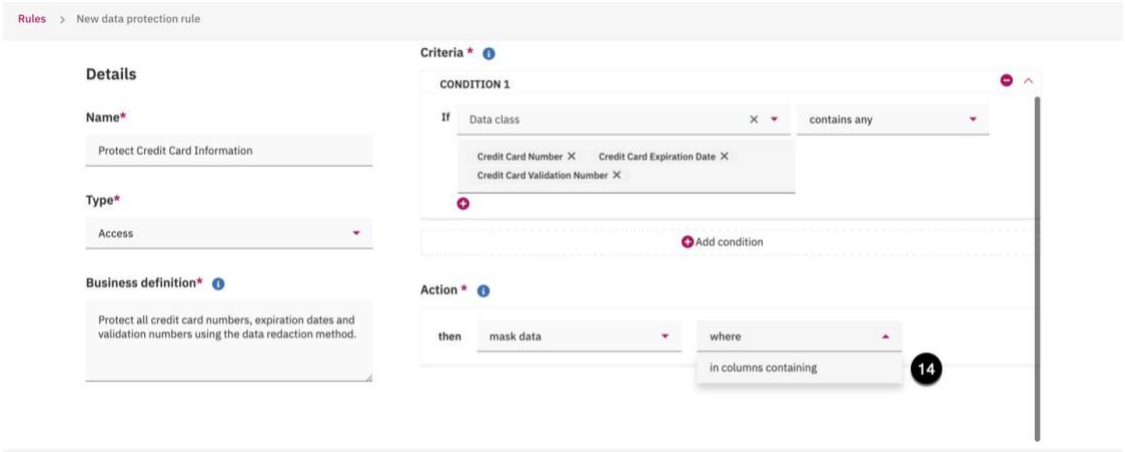
   Protect all credit card numbers, expiration dates and validation numbers using the data redaction method

6. In the Condition 1 area, for the **If** statement, select **Data Class**.
7. In the *Search for a data class* area type **credit card number**
8. Select the **Credit Card Number** data class from the list.
9. In the *Search for a data class* area type **credit card valid**.



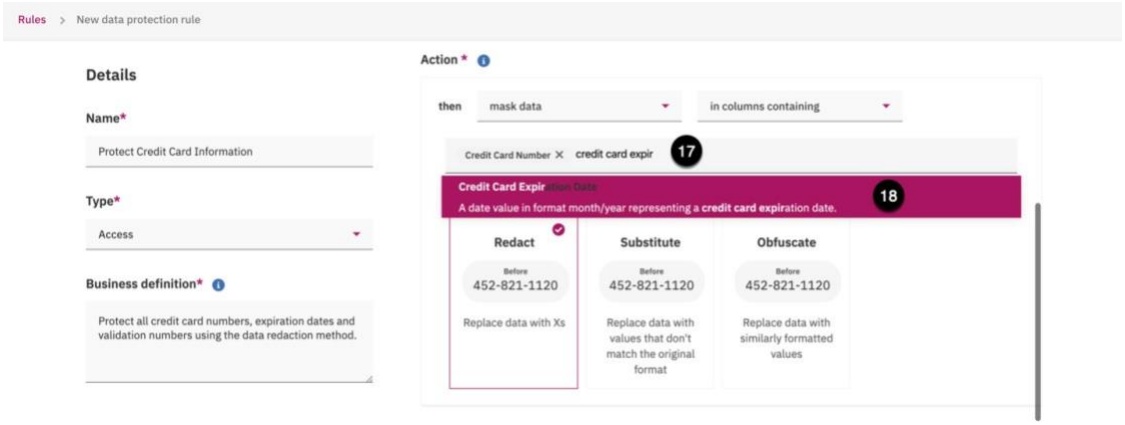
10. Select the **Credit Card Validation** data class from the list.
11. In the Action area, for the **then** clause, select **mask data**.
12. In the Action area, for the **where** clause, select **in columns containing**.
13. In the *Search for a data class* area type **credit card number**

14. Select the **Credit Card Number** data class from the list.

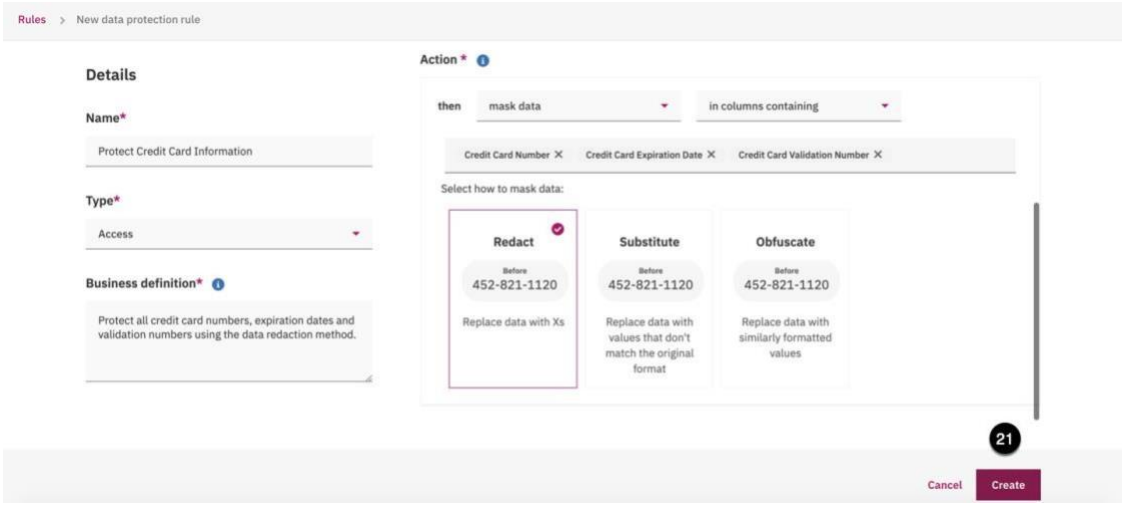


15. In the *Search for a data class* area type **credit card valid**.

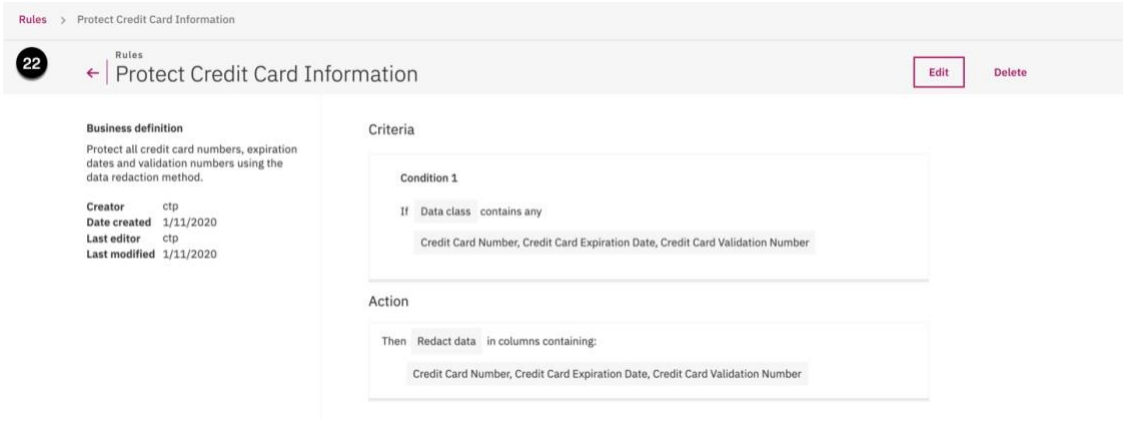
16. Select the **Credit Card Validation Number** data class from the list.



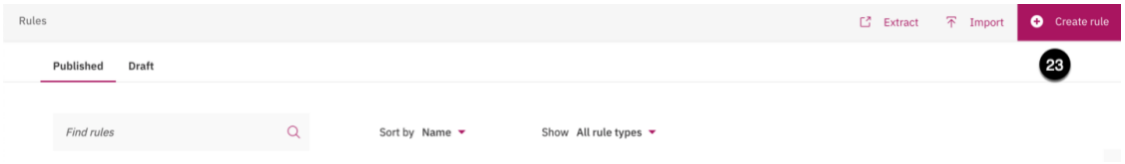
17. Click **Create**.



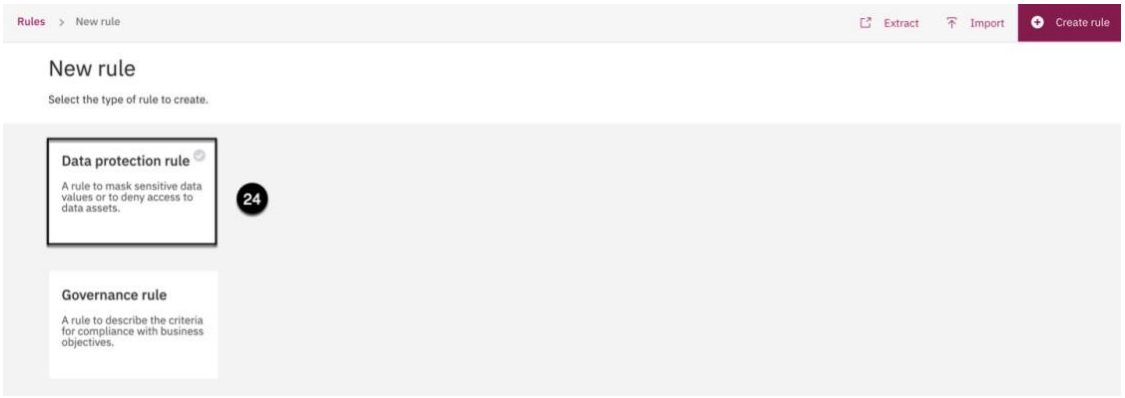
18. Click the **Rules** bread crumb to go back to the rules section.



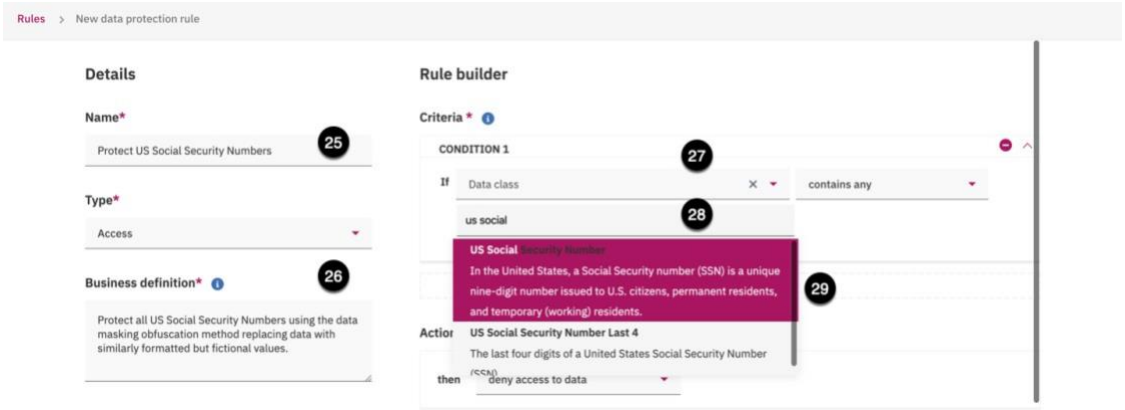
19. Click the **Create rule** button.



20. Click the Data protection rule type.



21. Enter a Name of **Protect US Social Security Numbers**.



22. Copy and Paste the following bolded text into the **Business definition**:

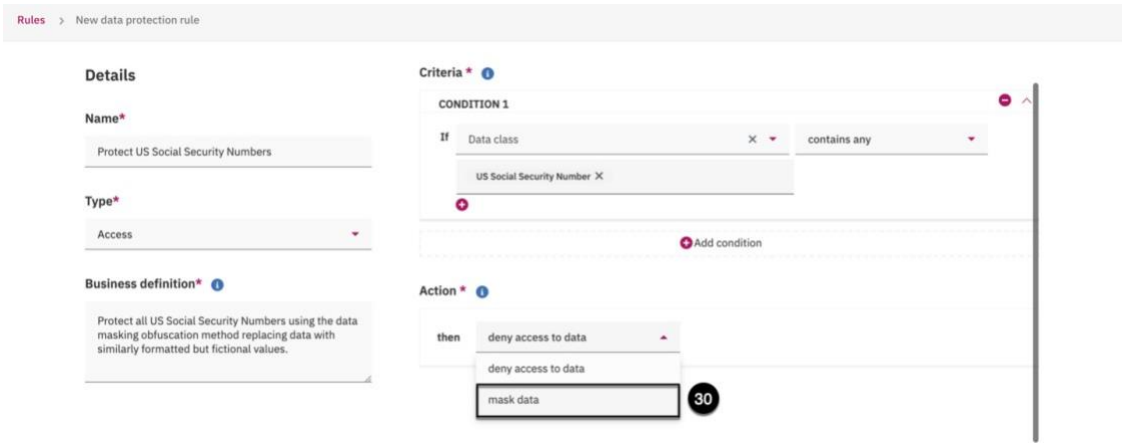
    Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values

23. In the Condition 1 area, for the **If** statement, select **Data Class**.

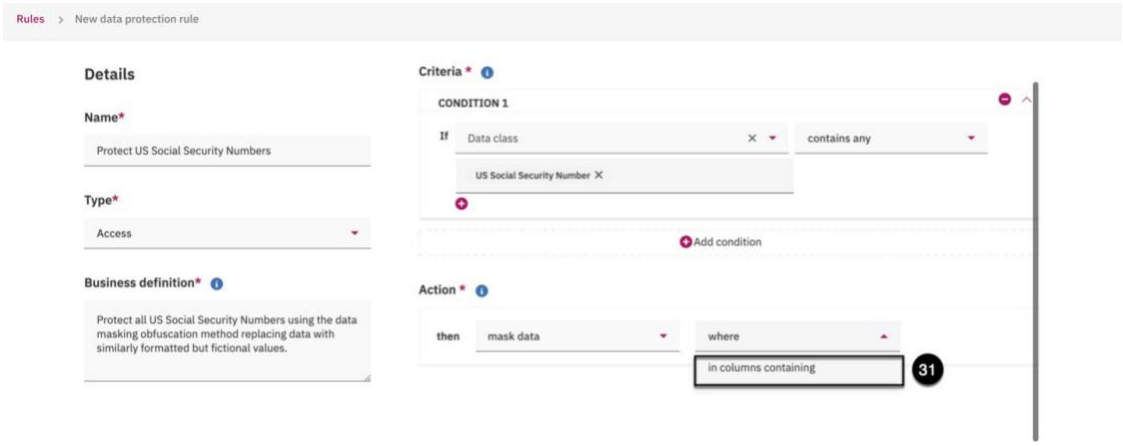
24. In the *Search for a data class* area type **us social**

25. Select the **US Social Security Number** data class from the list.

26. In the Action area, for the **then** clause, select **mask data**.

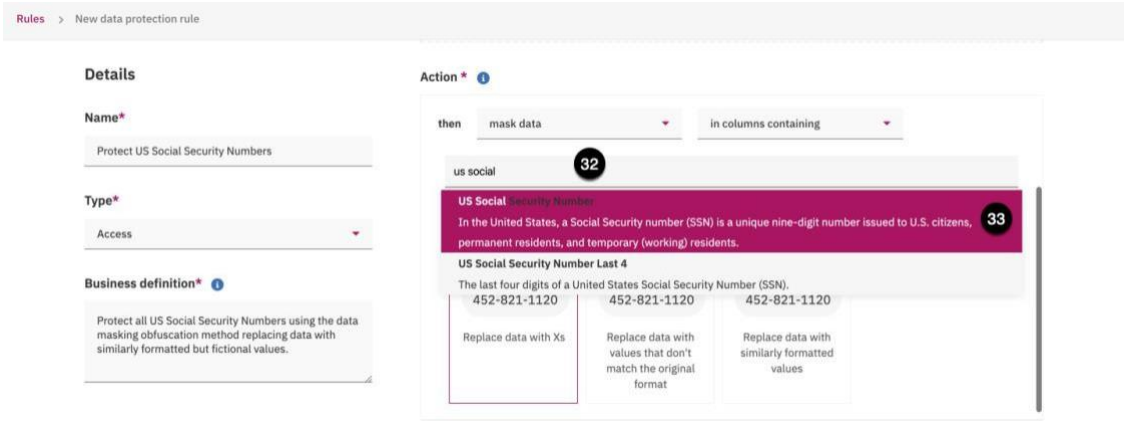


27. In the Action area, for the **where** clause, select **in columns containing**.

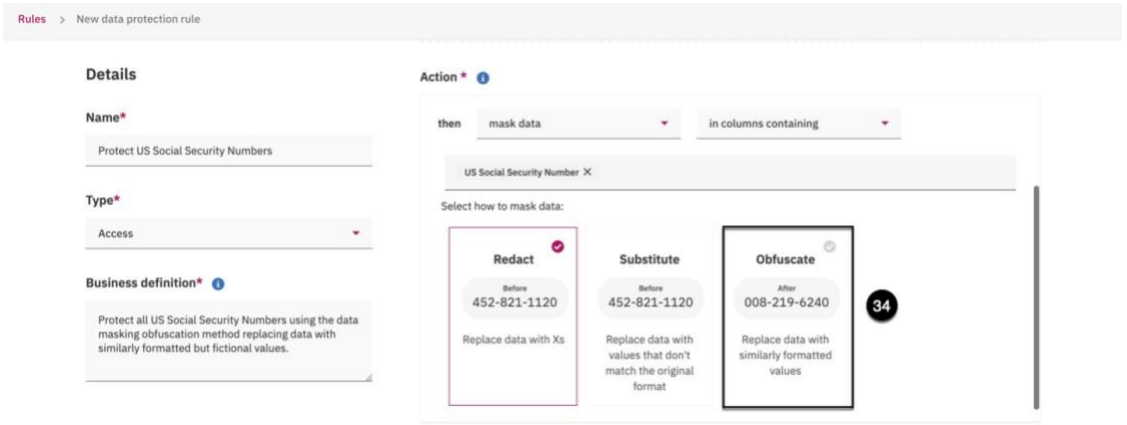


28. In the *Search for a data class* area type **us social**

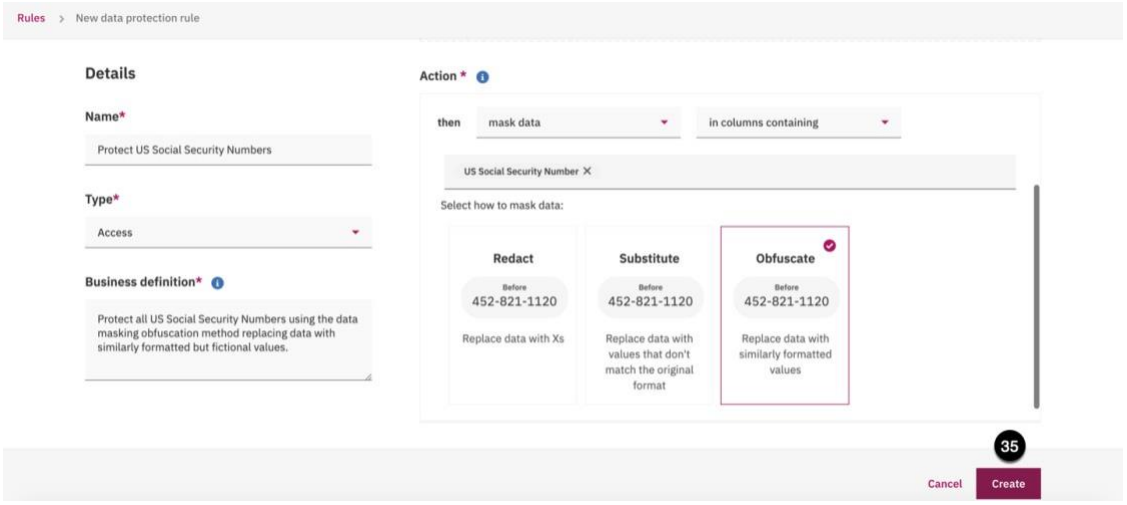
29. Select the **US Social Security Number** data class from the list.



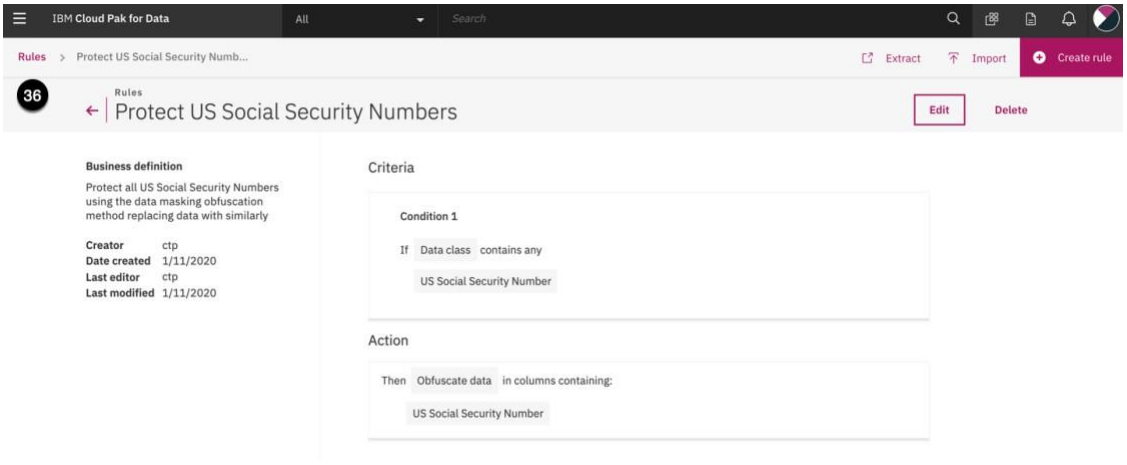
30. Click the **Obfuscate** masking method.



31. Click **Create**.

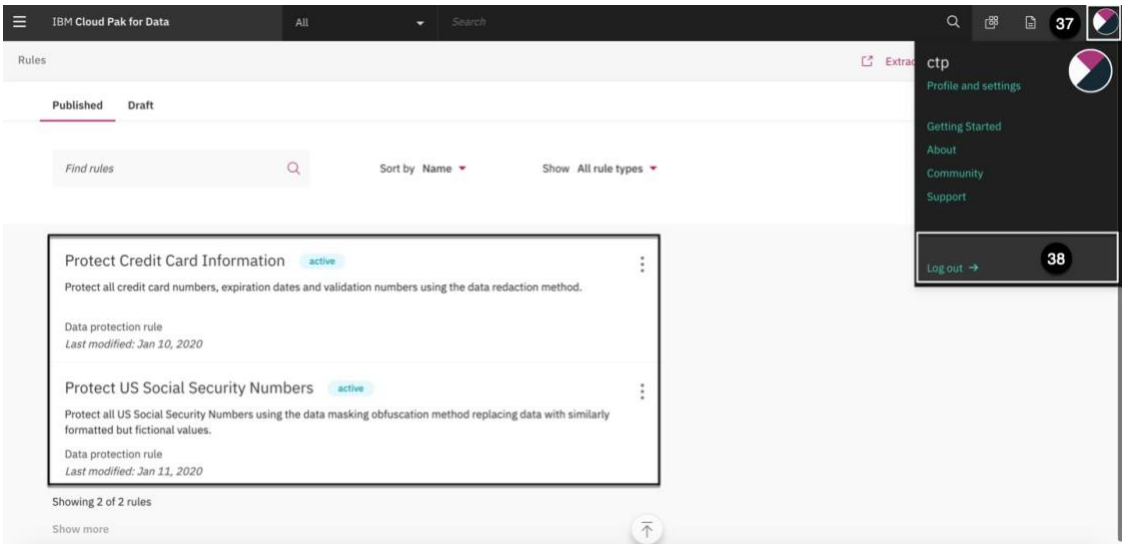


32. Click the **Rules** breadcrumb from the menu.



You should see the two data protection rules in the published tab. If not, refresh the page using your browser's refresh method.

33. Click on the **Profile and settings** icon in the top right corner.

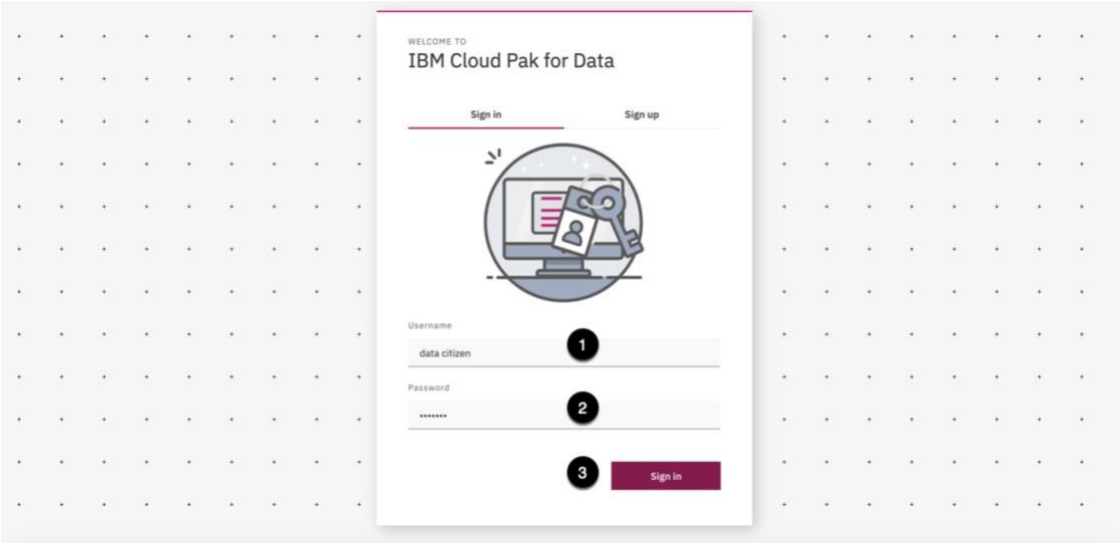


34. Click **Log out**.

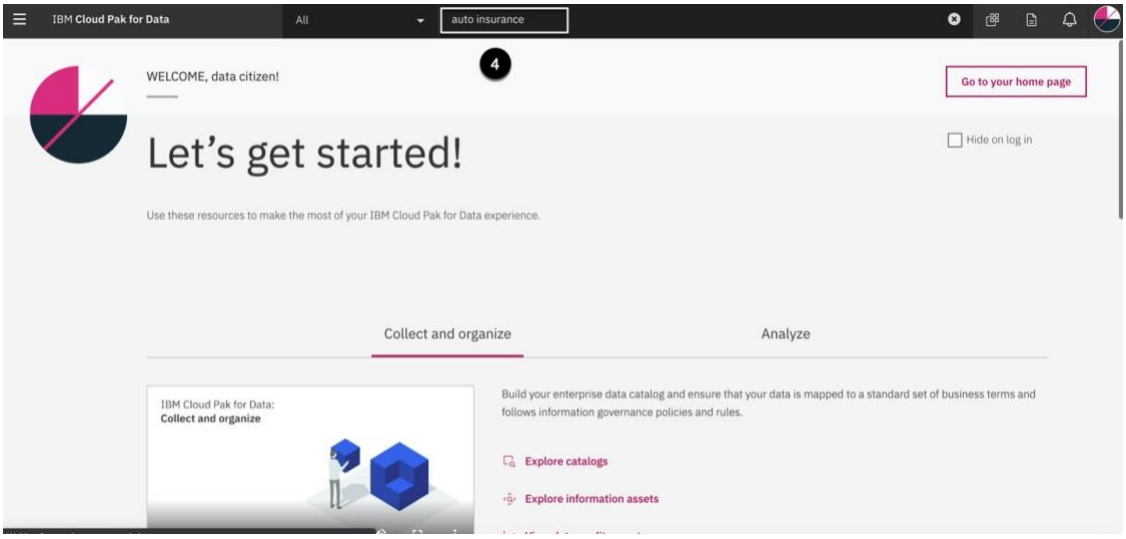
Validate Data Protection Rules

You will now log in as a data citizen to validate that you can search and find the data you are looking for and that the data protection rules are being enforced as defined.

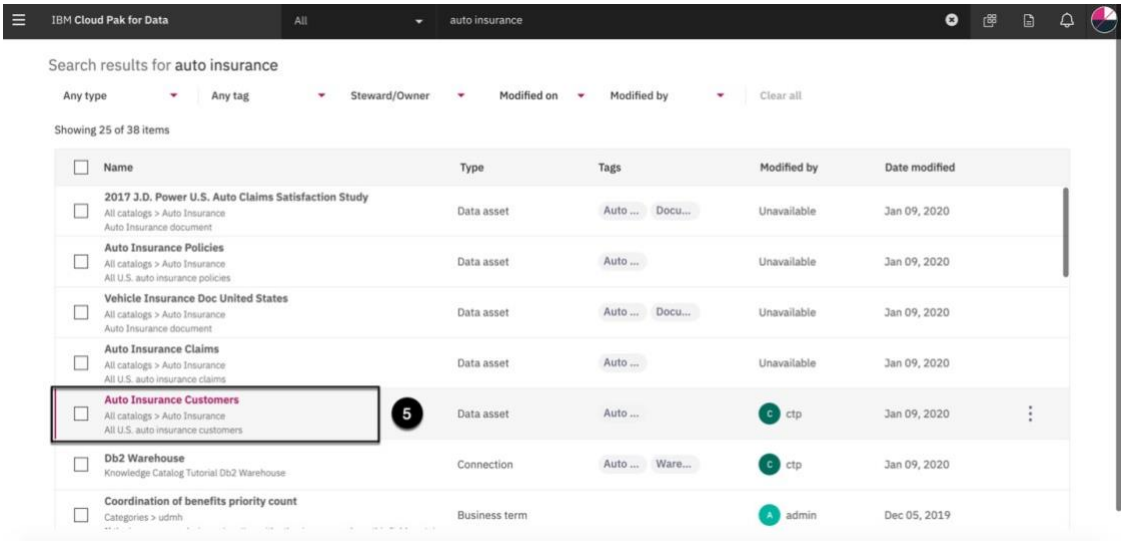
1. Enter **data citizen** as the Username.



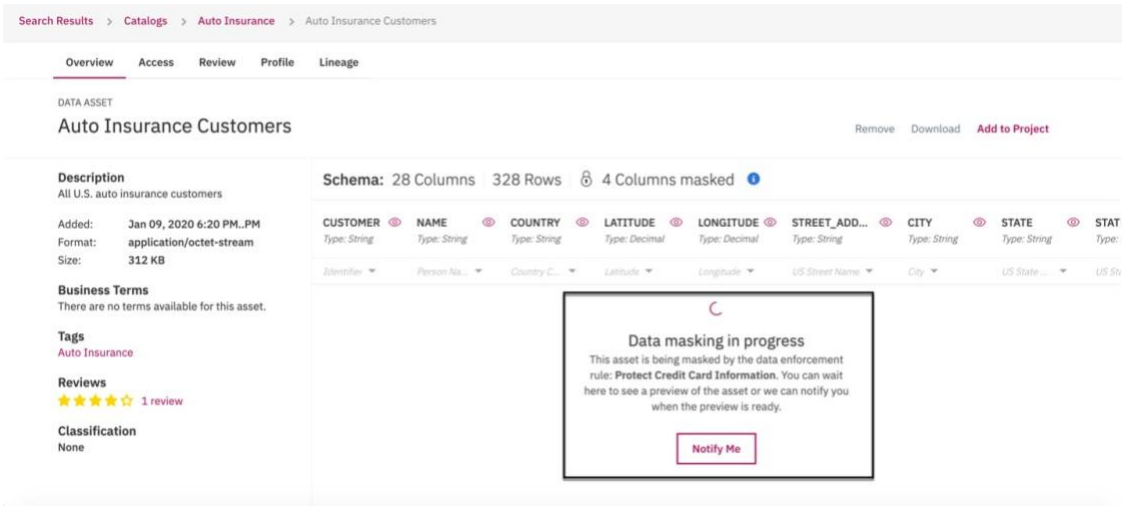
2. Enter **citizen** as the Password.
3. Click the **Sign in** button.
4. Enter **auto insurance** in the search area and press the enter key.



5. Click on the **Auto Insurance Customers** data asset from the Auto Insurance catalog.



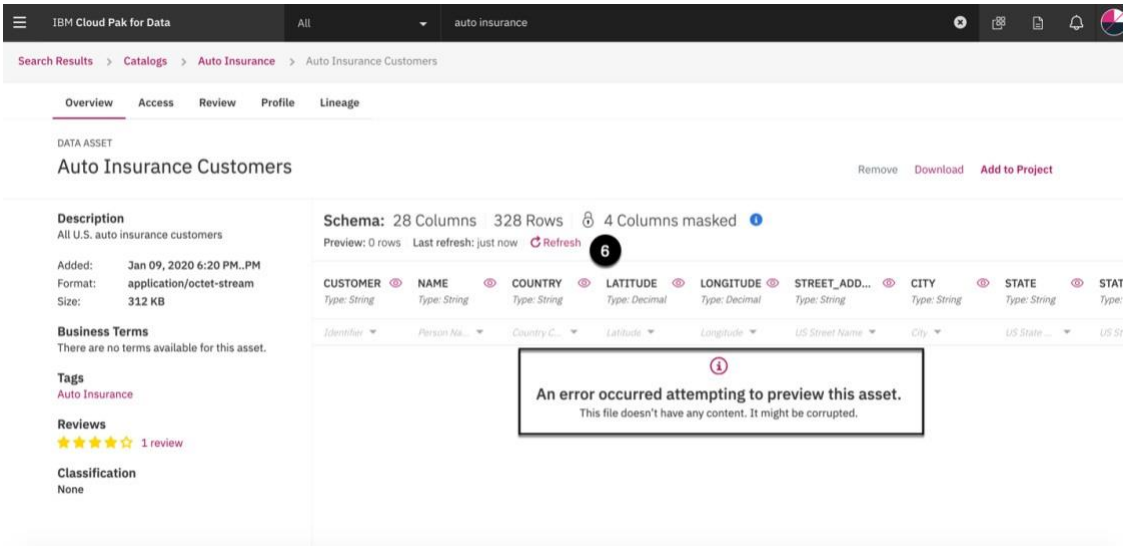
You should immediately see the message, "**Data Masking in progress",** with a spinning progress wheel. It will take a minute to load so be patient and let it finish.



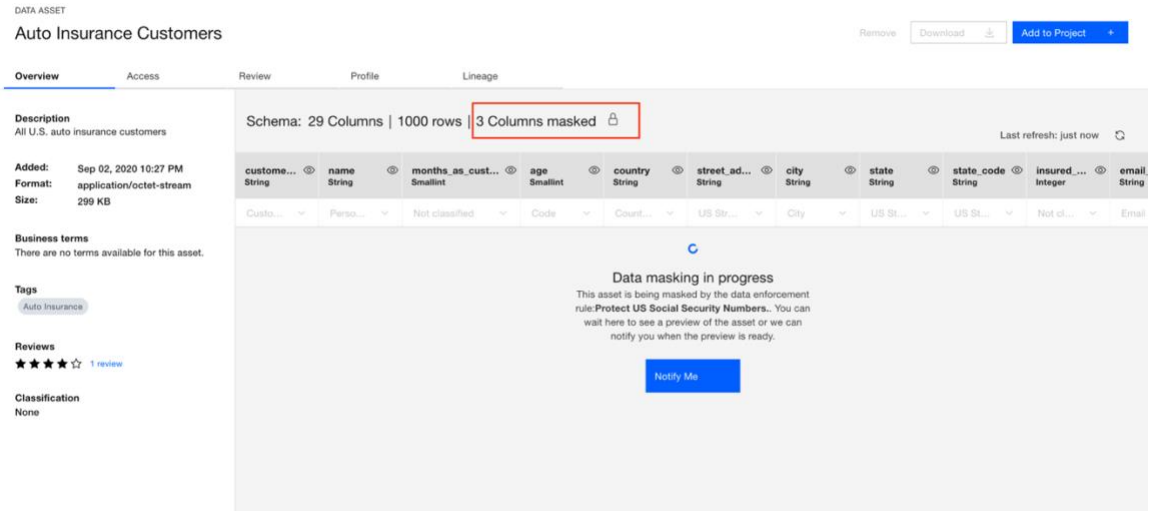
Note: The data citizen user does not own the data asset so the data protection rules will be enforced and only see the protected version of the data as defined by the data protection rules defined that are based on the data classes of the data. That is why you did the additional work to classify the additional credit card expiration date and validation columns in the data profile of the Auto Insurance Customers table.

If you see the error above, don't be alarmed. It's a known timing issue that is being addressed; the page just needs to be refreshed.

6. Click the **Refresh** button under the lock icon in the middle of the page. If that does not work, **refresh** the page using your browser's refresh method.

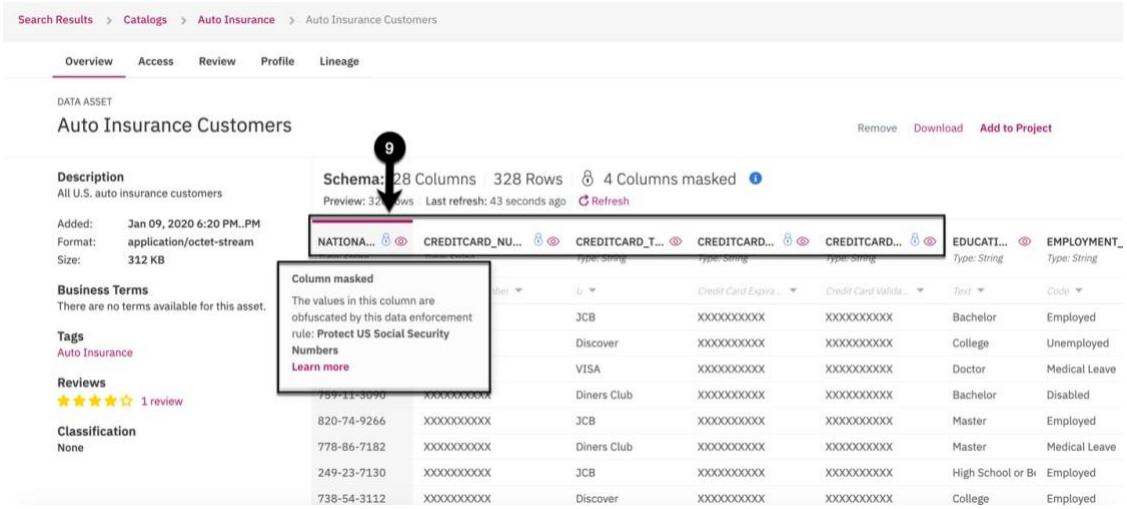


7. Hover over the **i** information icon next to the "4 columns masked" label with a lock icon in front of it.



Notice that 3 columns are redacted, and one is obfuscated.

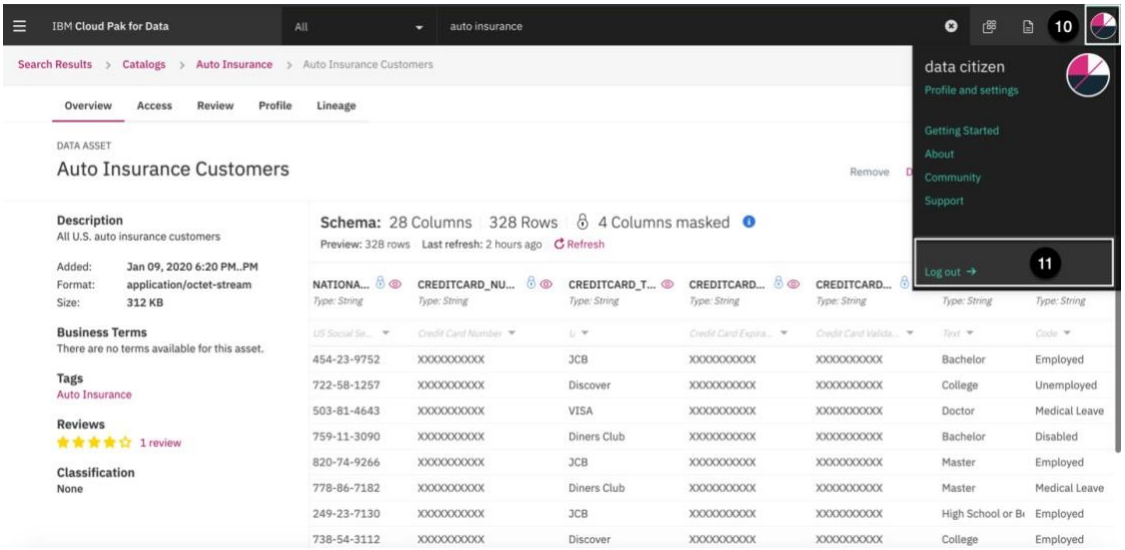
8. Scroll to the right until you see the **national_id** column.



Notice that the **national_id**, **creditcard_number**, and **creditcard_cvv** columns have a lock icon next to their name indicating that the data is being protected.

9. Hover over the **lock** icon next to the **national_id** column.

   Hover over the **lock** icon on the other columns being protected as well.



10. Click on the **Profile and settings** icon in the top right corner.

11. Click **Log out**.

Summary

You completed the IBM Watson Knowledge Catalog tutorial.

You explored: Creating a Governed Knowledge Catalog, Discovering and Cataloging Data Assets, Understanding and Socializing Data Assets, Shopping for Data, Preparing Data for Analytics and AI and Protecting Sensitive Information.

To further your education on Cloud Pak for Data and Watson Knowledge Catalog and many other IBM products and solutions, visit the IBM Demos website.