

Lab Guide

Hands-on-Lab

Watson Knowledge Catalog powered by Cloud Pak for Data

Shivam Solanki

Data Scientist

Shivam.raj.solanki@ibm.com



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.

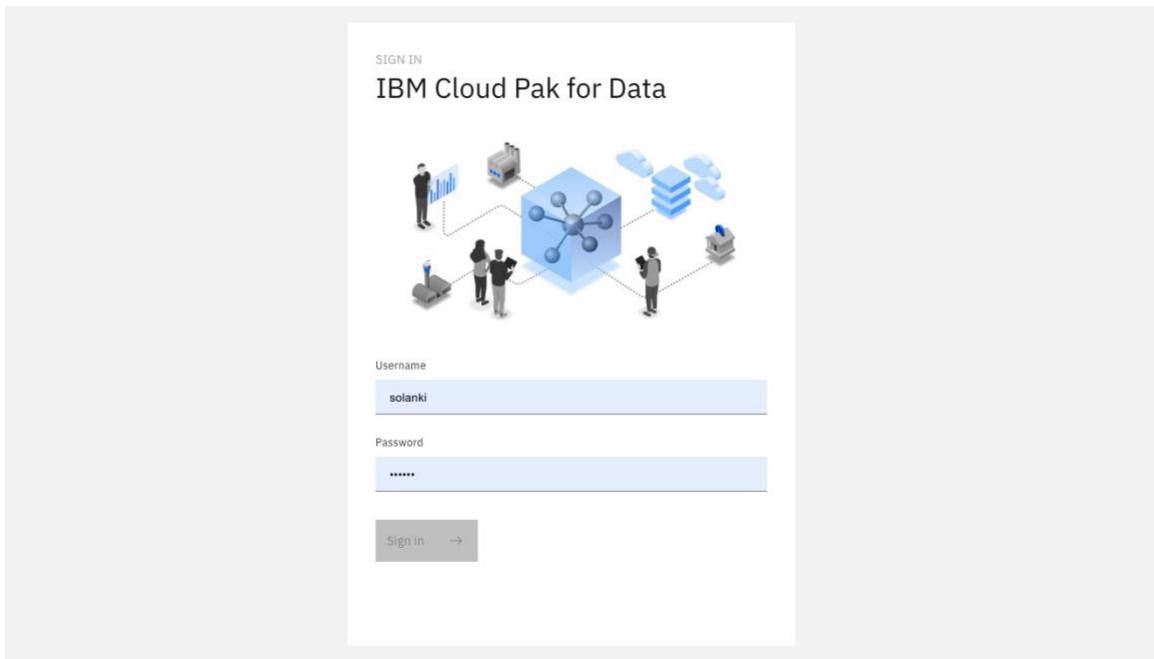
Create an IBM Cloud Pak for Data User

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.

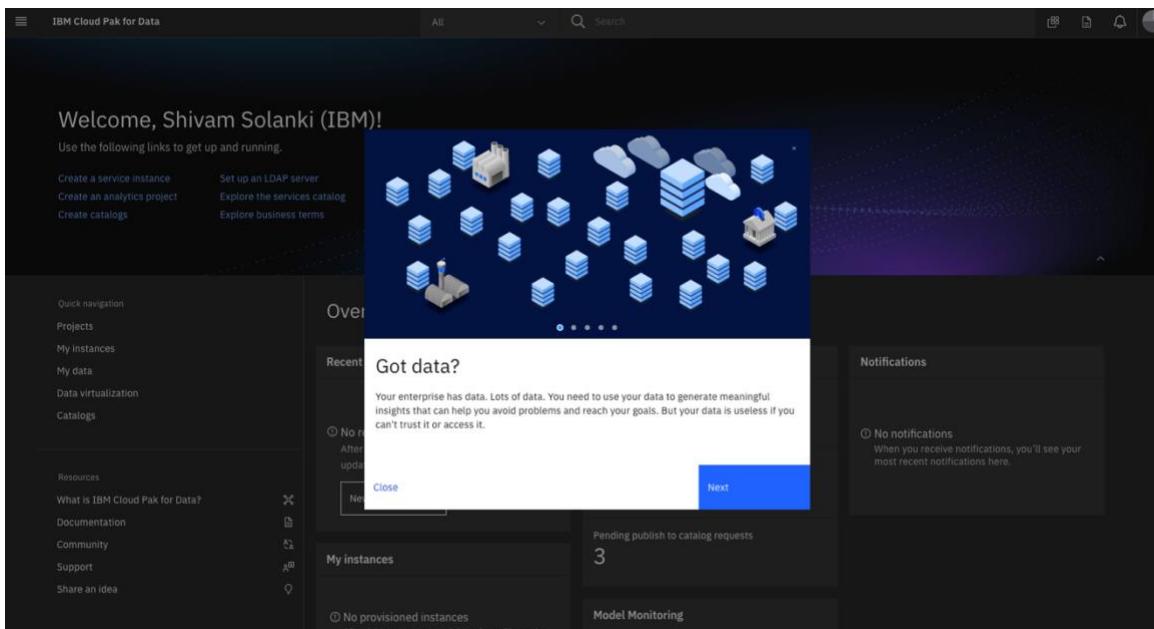
[Log in to Cloud Pak for Data](#)

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.

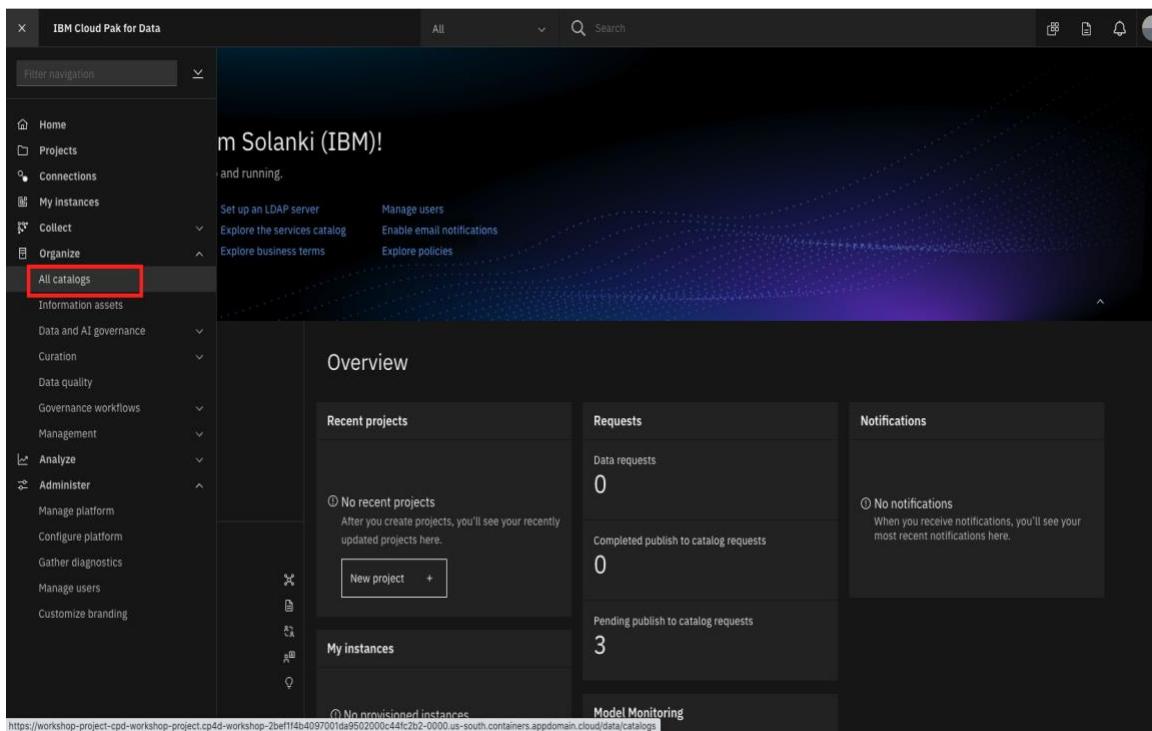
Create a Governed Catalog

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.

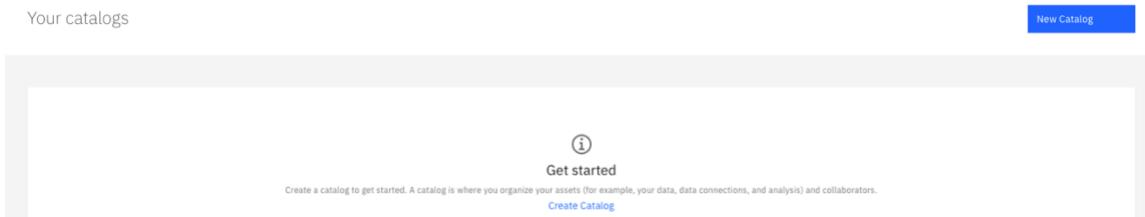
You will learn how to discover, curate and catalog data assets using an additional catalog other than the **Default Catalog** and by using some alternative methods. It will still have integration to the global business glossary and the business policies and rules to govern and protect it.

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



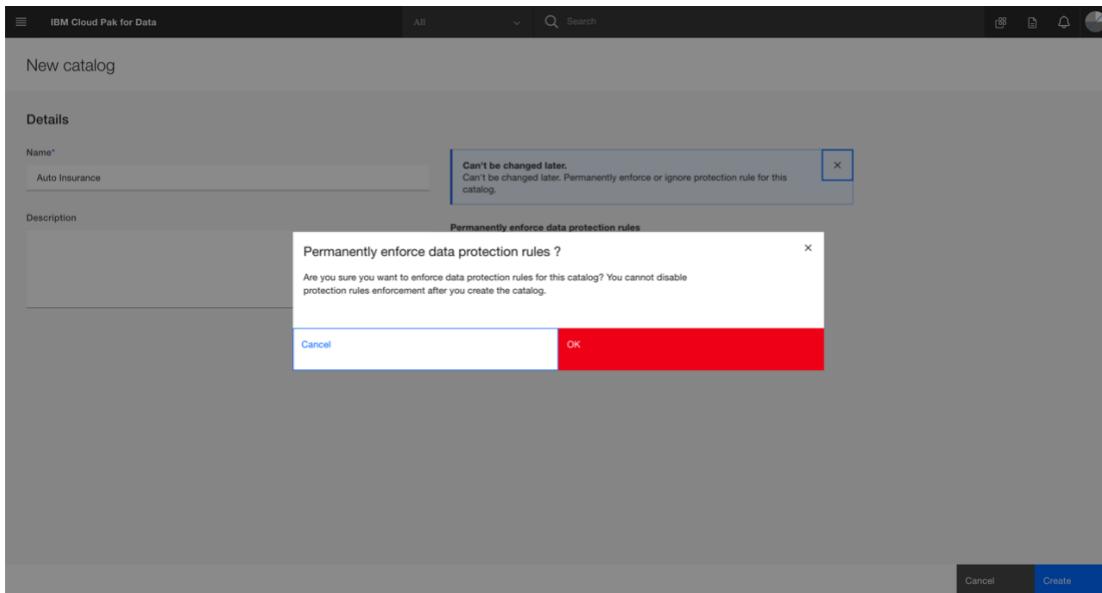
The screenshot shows the IBM Cloud Pak for Data interface. The left sidebar contains a navigation menu with several sections: Home, Projects, Connections, My instances, Collect, Organize, All catalogs (which is highlighted with a red box), Information assets, Data and AI governance, Curation, Data quality, Governance workflows, Management, Analyze, Administer, Manage platform, Configure platform, Gather diagnostics, Manage users, and Customize branding. The main content area has a dark background with a blue header bar. The header bar includes the title 'IBM Cloud Pak for Data', a search bar, and some user icons. Below the header, there's a greeting 'Hello [User Name] (IBM)!'. To the right of the greeting are links for 'Set up an LDAP server', 'Manage users', 'Explore the services catalog', 'Enable email notifications', 'Explore business terms', and 'Explore policies'. The main content area features an 'Overview' section with three tabs: 'Recent projects', 'Requests', and 'Notifications'. The 'Recent projects' tab shows 'No recent projects' and a button to 'New project'. The 'Requests' tab shows 'Data requests' at 0, 'Completed publish to catalog requests' at 0, and 'Pending publish to catalog requests' at 3. The 'Notifications' tab shows 'No notifications' and a note about receiving notifications. At the bottom of the page, there's a footer with a URL: 'https://workshop-project-cpd-workshop-project.cp4d-workshop-2bef1f4b-d097-001da9502000c44fc2b2-0000.us-south.containers.appdomain.cloud/data/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 policies** 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.

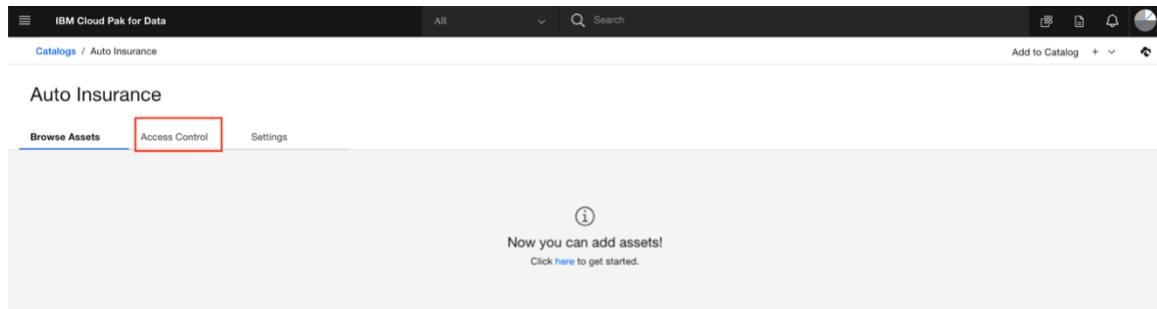


7. Click the **OK** button.

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.

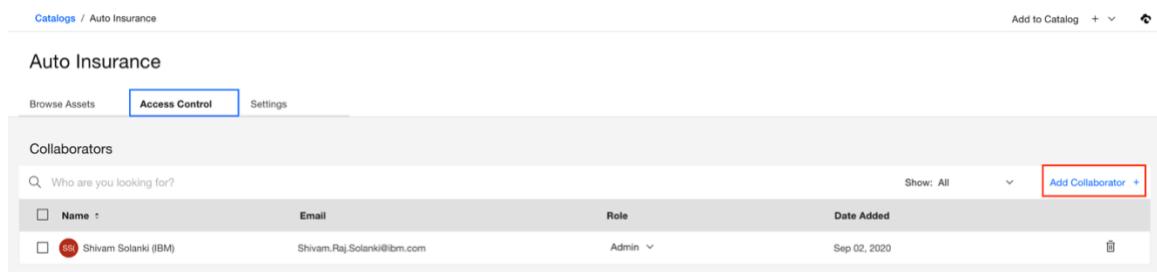
8. Click the **Create** button. You will see a **Creating Auto Insurance** notification during catalog creation.



The screenshot shows the 'Auto Insurance' catalog page. The 'Access Control' tab is highlighted with a red box. A message at the top says 'Now you can add assets!' with a link to get started. Below the message, there are tabs for 'Browse Assets', 'Access Control', and 'Settings'. The 'Access Control' tab is active.

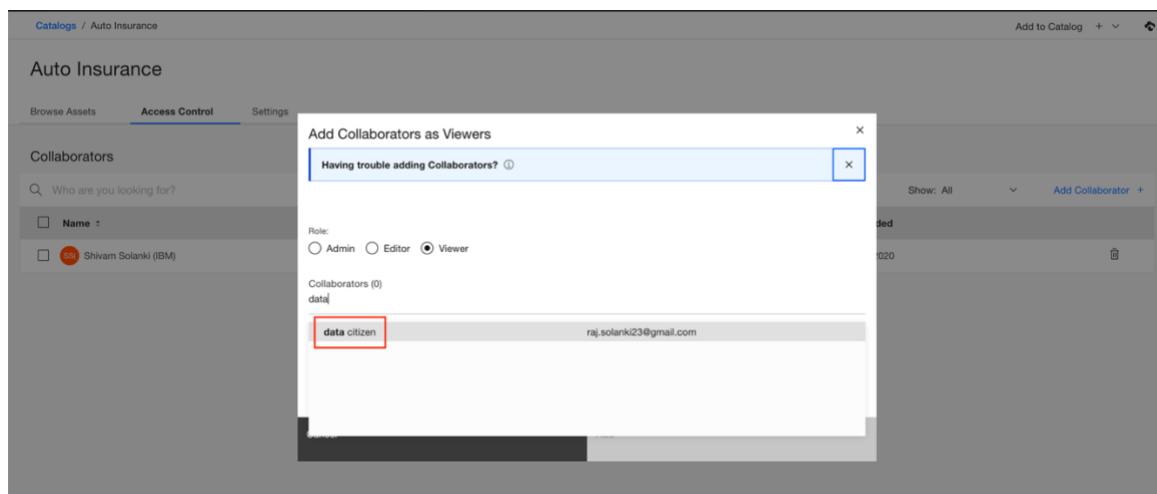
Once completed, you are brought into the newly created catalog. You will now add the **data citizen** user to the catalog as a *Viewer* so they can access the new catalog and use the data assets. You will log in as this user at the end of the lab to see how data protection rules are enforced.

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



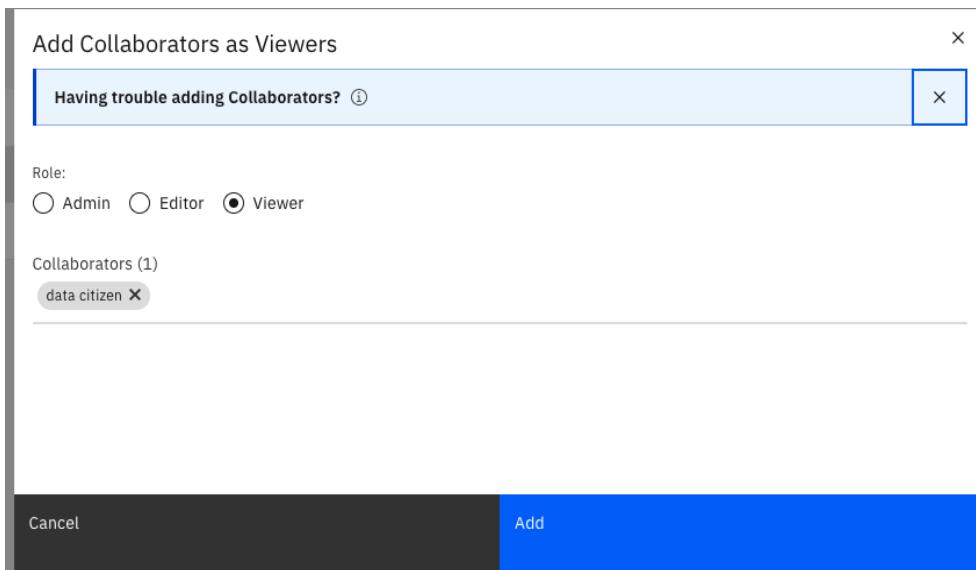
The screenshot shows the 'Access Control' tab with the 'Collaborators' section. A user named 'Shivam Solanki (IBM)' is listed with the role 'Admin' and the date 'Sep 02, 2020'. An 'Add Collaborator' button is highlighted with a red box. The 'Browse Assets' and 'Settings' tabs are also visible.

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



The screenshot shows a modal dialog titled 'Add Collaborators as Viewers'. It contains a message 'Having trouble adding Collaborators?' and a 'Role' section with radio buttons for 'Admin', 'Editor', and 'Viewer', where 'Viewer' is selected. Below that is a 'Collaborators (0)' section with a search bar and a list containing 'data citizen'. The 'data citizen' entry is highlighted with a red box. The background shows the 'Access Control' tab of the 'Auto Insurance' catalog.

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**.

The screenshot shows the "Auto Insurance" catalog page. The "Access Control" tab is selected. The "Collaborators" section displays a table with two rows. The first row is for "Shivam Solanki (IBM)" with email "Shivam.Raj.Solanki@ibm.com" and role "Admin". The second row is for "data citizen" with email "raj.solanki23@gmail.com" and role "Viewer". The "data citizen" row is highlighted with a red box. The "Add Collaborator" button is visible at the top right of the table.

Name	Email	Role	Date Added
Shivam Solanki (IBM)	Shivam.Raj.Solanki@ibm.com	Admin	Sep 02, 2020
data citizen	raj.solanki23@gmail.com	Viewer	Sep 02, 2020

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

Create Analytic Projects

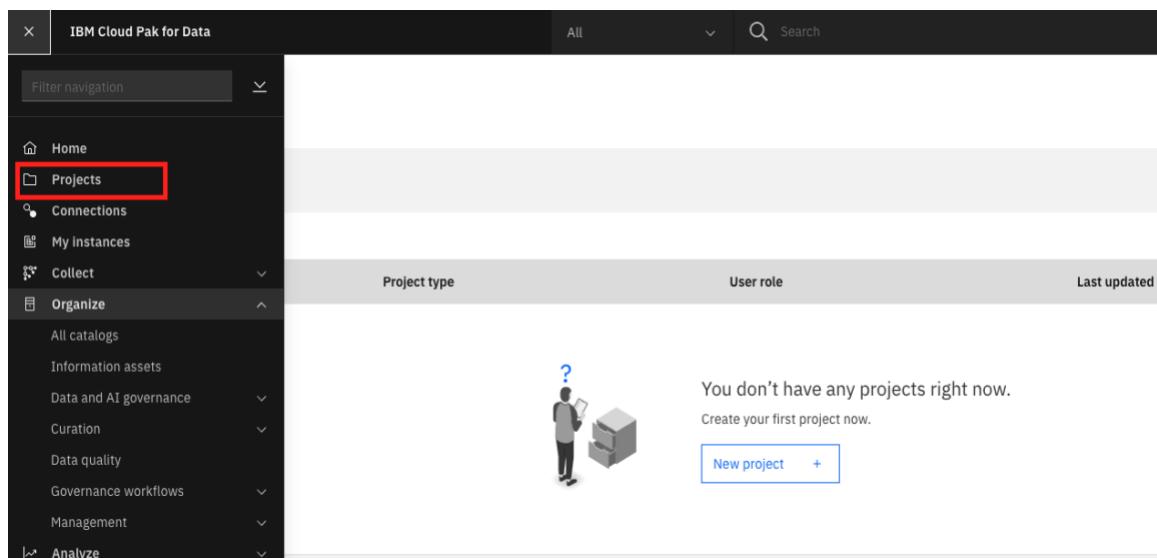
In this task, you will create two Cloud Pak for Data analytics projects.

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.

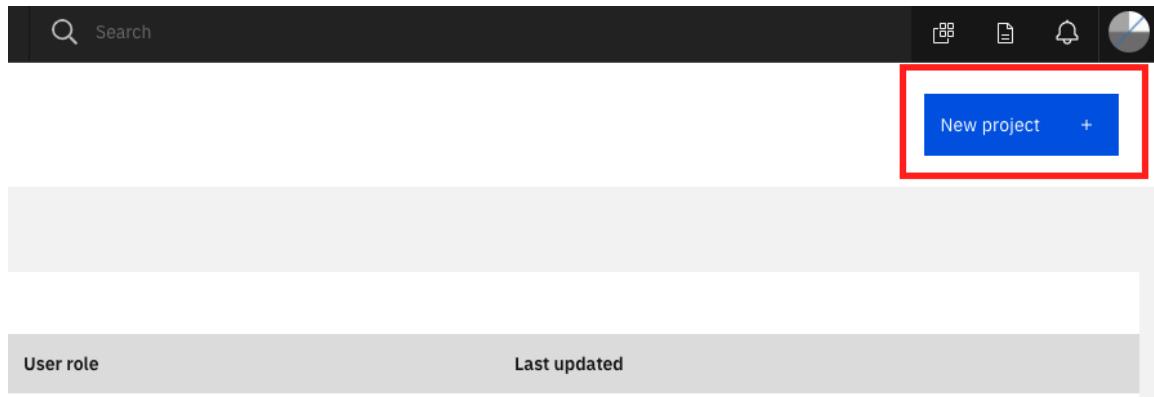
Create the Auto Insurance Project

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

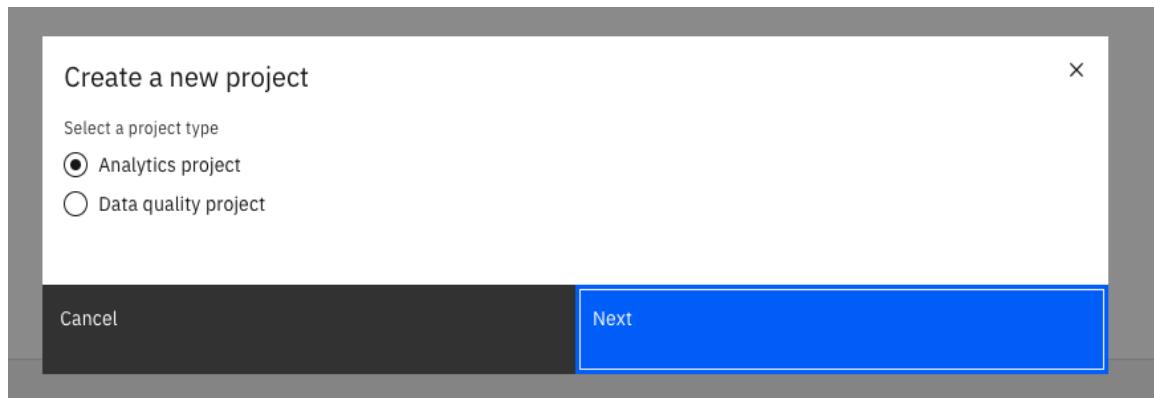


The screenshot shows the IBM Cloud Pak for Data web interface. The top navigation bar has a search bar and a 'Search' button. The main menu on the left includes 'Home', 'Projects' (which is highlighted with a red box), 'Connections', 'My instances', 'Collect', 'Organize', 'All catalogs', 'Information assets', 'Data and AI governance', 'Curation', 'Data quality', 'Governance workflows', 'Management', and 'Analyze'. The right side of the screen displays a table header for 'Project type', 'User role', and 'Last updated'. Below the table, there is a message: 'You don't have any projects right now. Create your first project now.' with a 'New project +' button. A small icon of a person standing next to a question mark is also present.

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



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



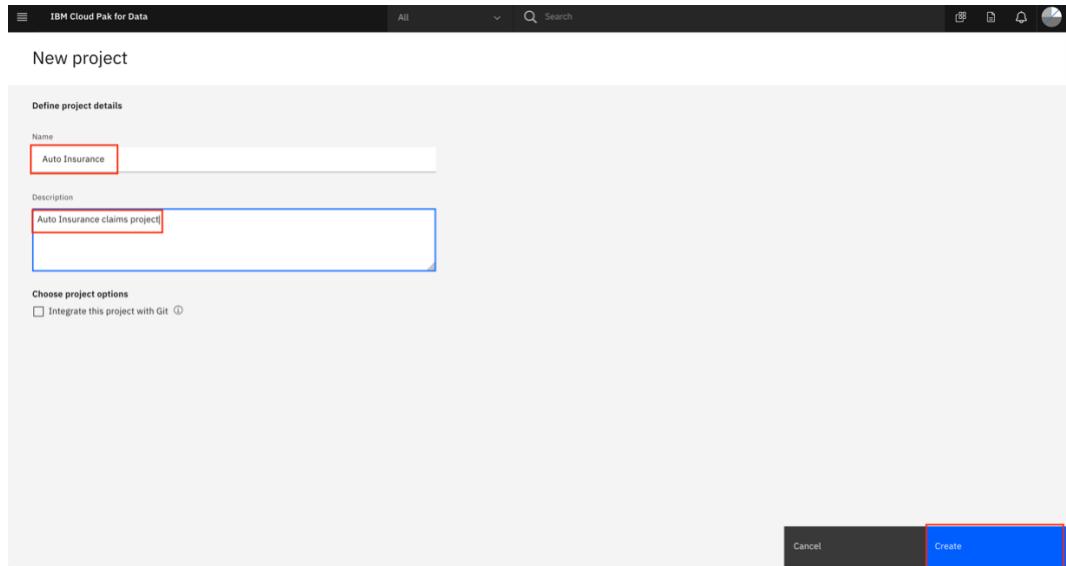
5. Click the **OK** button.

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

A screenshot of the "Create a project" interface. It shows two main options: "Create an empty project" and "Create a project from a file".

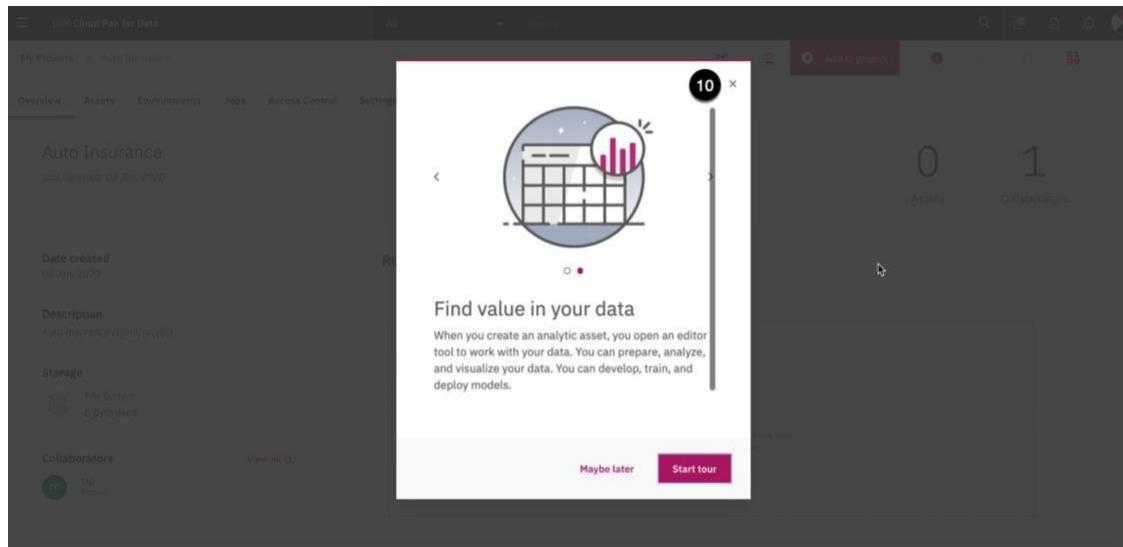
- Create an empty project:** This section features a circular icon with a blue gradient and a white outline, containing a small diagram of a flowchart or data preparation interface. The text says "Create an empty project" and "Add the data you want to prepare, analyze, or model. Choose tools based on how you want to work: write code, create a flow on a graphical canvas, or automatically build models." To the right, under "USE TO", are three items: "Prepare and visualize data", "Analyze data in notebooks", and "Train models".
- Create a project from a file:** This section features a circular icon with a blue gradient and a white outline, containing a hand holding a document with a plus sign on it. The text says "Create a project from a file" and "Get started fast by loading existing assets. Choose a project file from your system or a Git repository." To the right, under "USE TO", are four items: "Learn by example", "Build on existing work", "Run tutorials", and "Integrate with Git".

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.

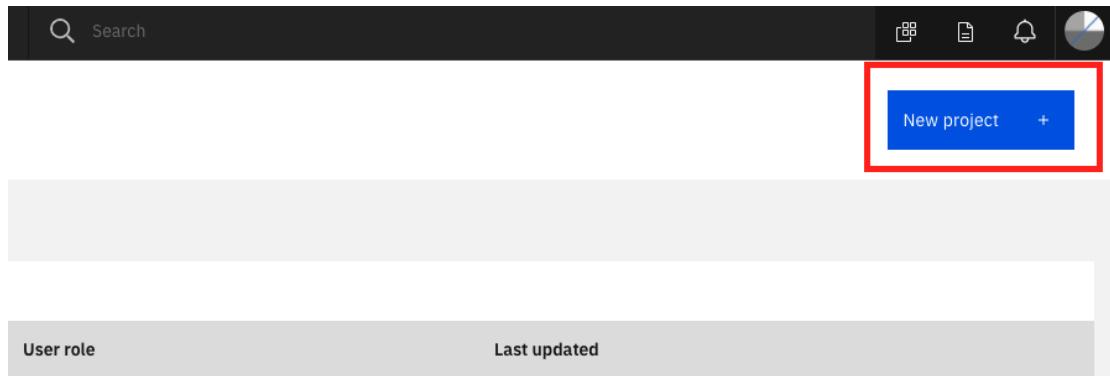
The screenshot shows the 'Auto Insurance' project details. At the top right, there are counts for 'Assets' (0) and 'Collaborators' (1). Below this, the 'Overview' section displays basic project information: Date created (Sep 02, 2020), Description (Auto Insurance claims project), Storage (File System, 0 Byte used), and Collaborators (Shivam Solanki (IBM), Admin). A note indicates no deployment space is associated. The 'Recent activity' section is currently empty.

Create the Auto Discovery Project

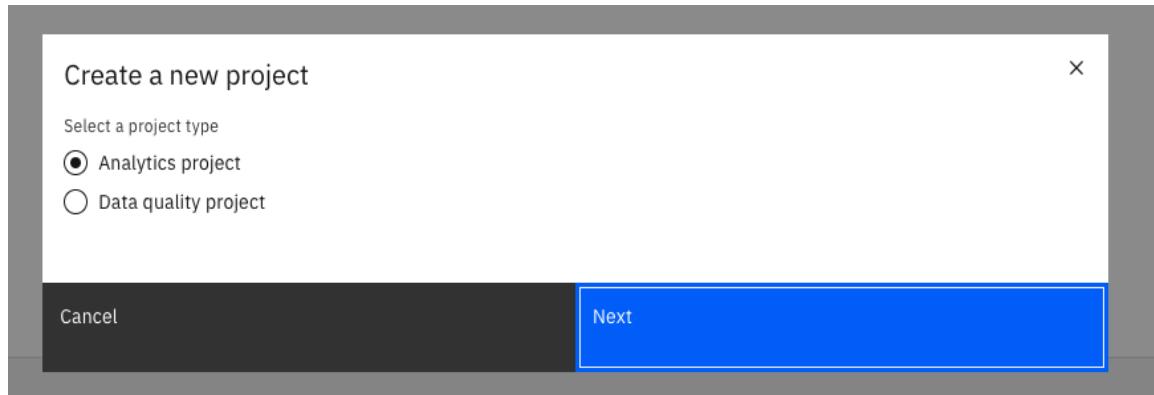
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.
2. Click the **Projects** menu.

The screenshot shows the navigation menu on the left. The 'Projects' option is highlighted with a red box. The main area displays a message: 'You don't have any projects right now. Create your first project now.' with a 'New project +' button.

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



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

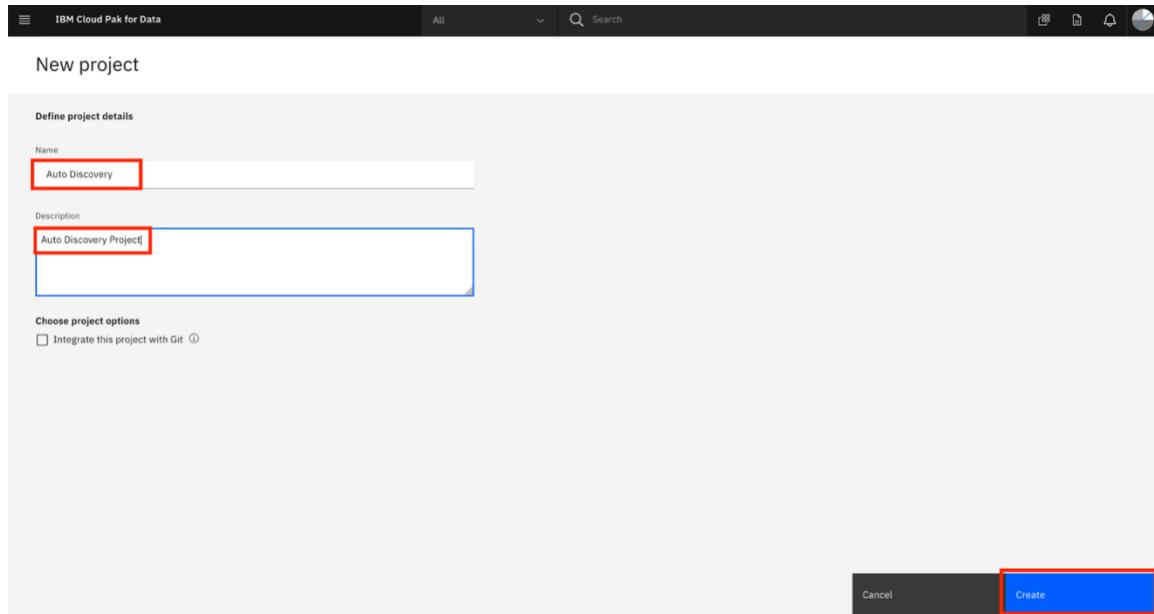


5. Click the **Next** button.

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

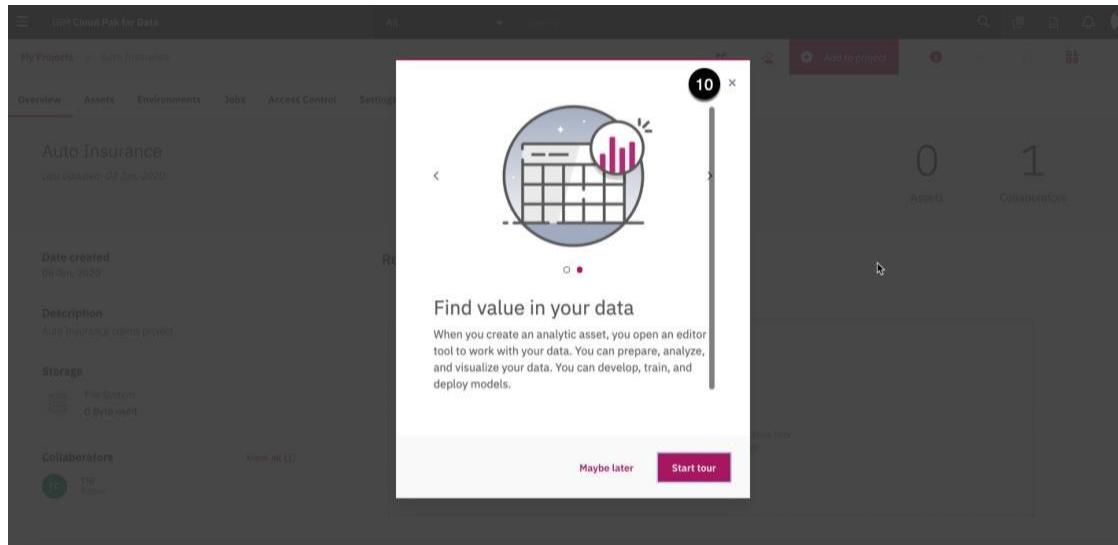
The screenshot shows the "Create a project" page. At the top, there is a back arrow and the text "Create a project". Below this, a sub-header says "Create an empty project" with the sub-instruction "Create an empty project, and then add data and choose the right tools to accomplish your goals.". To the left of this text is a circular icon containing a stylized representation of a data flow or code editor. To the right is a "USE TO" section with three items: "Prepare and visualize data", "Analyze data in notebooks", and "Train models". Below this section is another sub-header "Create a project from a file" with the sub-instruction "Get started fast by loading existing assets. Choose a project file from your system or a Git repository.". To the left of this text is a circular icon containing a hand holding a document with a plus sign on it. To the right is another "USE TO" section with four items: "Learn by example", "Build on existing work", "Run tutorials", and "Integrate with Git".

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.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with the title "IBM Cloud Pak for Data". Below it, a sub-navigation bar shows "My projects / Auto Discovery". The main content area has tabs for "Overview", "Assets", "Environments", "Jobs", "Access Control", and "Settings", with "Overview" being the active tab.

Auto Discovery

Last Updated: Sep 02, 2020

Overview

- Date created: Sep 02, 2020
- Description: Auto Discovery Project
- Storage:
File System 0 Byte used
- Collaborators:
Shivam Solanki (IBM) Admin
- 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.

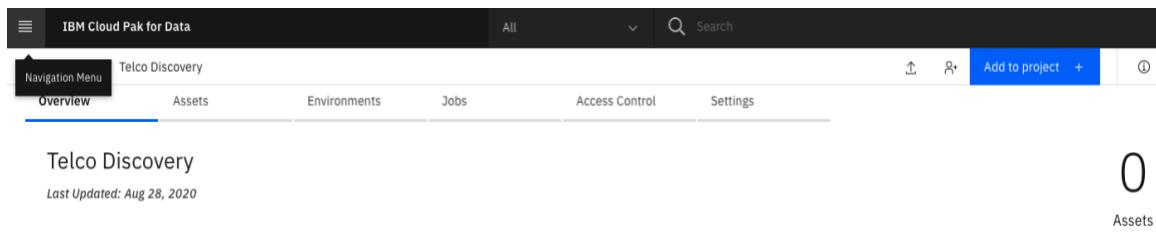
0 Assets 1 Collaborators

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.



The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a dark header bar with the "IBM Cloud Pak for Data" logo on the left, a search bar with a magnifying glass icon and the word "Search" on the right, and a dropdown menu with "All" selected. Below the header is a navigation menu with tabs: "Navigation Menu" (highlighted), "Overview" (selected), "Assets", "Environments", "Jobs", "Access Control", and "Settings". To the right of the tabs are icons for "Add to project" (blue button) and a help icon. The main content area is titled "Telco Discovery" and displays the message "Last Updated: Aug 28, 2020". On the right side of this title, there is a large "0" and the word "Assets".

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

The screenshot shows the IBM Cloud Pak for Data interface. On the left, there is a navigation sidebar with various sections like Home, Projects, Connections, My instances, Collect, Organize, Analyze, and Administer. Under the Organize section, the 'All catalogs' option is highlighted with a red box. The main content area is titled 'Recent activity' and contains a placeholder message: 'Alerts related to this project appear here when the project is active.' There is also a small icon of a document with a plus sign.

- Click the **Auto Insurance** catalog.

The screenshot shows the 'Your catalogs' page. It displays a card for the 'Auto Insurance' catalog, which was created by Shivam Solanki (IBM) on Sep 02, 2020 at 8:50 PM. The card title is 'Auto Insurance Knowledge Catalog'. A blue 'New Catalog' button is located in the top right corner of the page.

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

The screenshot shows the 'Auto Insurance' catalog page. At the top right, there is a 'Add to Catalog' button with a dropdown menu. The 'Local files' option is highlighted with a red box. Below the catalog card, there is a message: 'Now you can add assets! Click [here](#) to get started.' The bottom of the page has tabs for 'Browse Assets', 'Access Control', and 'Settings'.

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

Add data assets from local files

Drop your files here or [browse](#) your files to add new files (up to 5 GB each)
Files for assets are saved in the catalog's associated storage.

Drag and drop file(s) here to upload

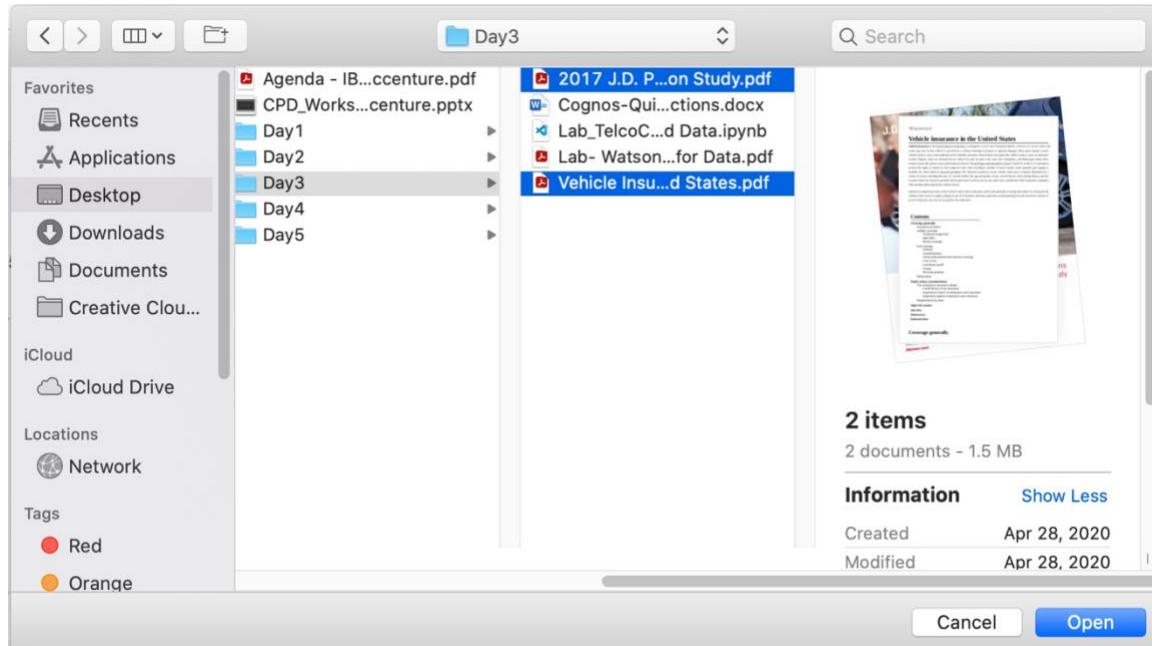
Target
Auto Insurance

Description

Business Terms
Search Business Terms

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.

Add data assets from local files

Selected Files (2)*

6 Edit name and format

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

ASSET NAME	FORMAT	
Vehicle Insurance Doc United States.pdf	PDF	(X)
2017 J.D. Power U.S. Auto Claims Satisfaction Survey.pdf	PDF	(X)

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.

Asset Name	Format
satisfaction Stud	application/pdf
c United States	application/pdf

Cancel Apply

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.

Catalogs / Auto Insurance
Files for assets are saved in the catalog's associated storage.

Drag and drop file(s) here to upload

Selected Files (2)*

Asset Name	Format
2017 J.D. Power U.S. ... PDF	X
Vehicle Insurance Doc... PDF	X

Description
Auto Insurance document

Business Terms
Search Business Terms

Tags
Auto Insurance X

Start typing to add values

Classification* ⓘ

None

Privacy
Public Private

All catalog members can find and use the asset.

Members

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

Cancel Add

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.

The screenshot shows the 'Selected Files' interface with two files listed: '2017 J.D. Power U.S. ...' and 'Vehicle Insurance Doc...'. On the right side, there is a 'Business Terms' section with a search bar and a 'Tags' section. The 'Tags' section contains 'Auto Insurance X' and 'Document'. The 'Document' tag is highlighted with a red box.

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

The screenshot shows the asset catalog creation dialog. It displays 'Selected Files (2)*' and 'Edit name and format' buttons. Below is a table with 'Asset Name' and 'Format' columns. To the right, there are sections for 'Business Terms' and 'Tags', which contain 'Auto Insurance X' and 'Document X'. A note at the bottom says 'Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' The 'Add' button is highlighted with a red box.

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.

Catalogs / Auto Insurance

Add to Catalog + 

Auto Insurance

Browse Assets Access Control Settings

What assets are you looking for?

Any type Any source Any tag

Watson Recommends Highly Rated Recently Added

Clear all

Auto Insurance Document

Data asset

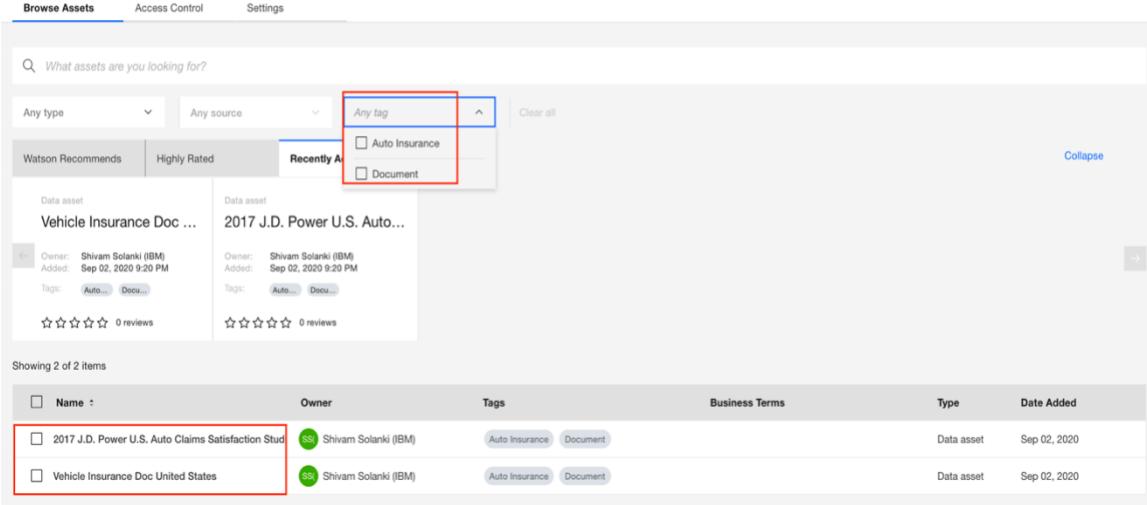
Vehicle Insurance Doc ... 2017 J.D. Power U.S. Auto...

Owner: Shivam Solanki (IBM) Owner: Shivam Solanki (IBM)
Added: Sep 02, 2020 9:20 PM Added: Sep 02, 2020 9:20 PM
Tags: Auto... Docu... Tags: Auto... Docu...
0 reviews 0 reviews

Collapseshow

Showing 2 of 2 items

<input type="checkbox"/> Name :	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/> 2017 J.D. Power U.S. Auto Claims Satisfaction Stud	Shivam Solanki (IBM)	Auto Insurance Document		Data asset	Sep 02, 2020
<input type="checkbox"/> Vehicle Insurance Doc United States	Shivam Solanki (IBM)	Auto Insurance Document		Data asset	Sep 02, 2020



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.

The screenshot shows the 'Auto Insurance' catalog page. At the top right, there is a 'Catalogs / Auto Insurance' breadcrumb, an 'Add to Catalog' button with a plus icon, and a 'Connection' button with a gear icon. A large number '1' is overlaid on the 'Connection' button. Below the header, there are tabs for 'Browse Assets', 'Access Control', and 'Settings'. The main area is titled 'Auto Insurance' and contains a search bar with placeholder text 'Q. What assets are you looking for?'. Below the search bar are filter options: 'Filter' (set to 'Any type'), 'Any tag', 'Modified on' (set to 'Modified by'), 'Modified by' (set to 'Modified by'), and 'Clear all'. The 'Connection' button is highlighted with a red box and a large number '1'.

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.

The screenshot shows the 'New connection' dialog with the 'From global' tab selected. The 'Create new' tab is also visible. The interface is divided into two main sections: 'IBM services' and 'Third-party services'.
IBM services:

- Analytics Engine HDFS
- Cognos Analytics
- Db2** (highlighted with a red box)
- IBM Db2 database
- Hive via Execution Engine for Hadoop
- PureData System for Analytics

Third-party services:

- Amazon Redshift
- Cloudera Impala
- Google BigQuery
- Microsoft Azure SQL Database
- OData
- Salesforce.com
- Sybase IQ
- Amazon S3
- Dropbox
- Google Cloud Storage
- Microsoft SQL Server
- Oracle
- SAP OData
- Tableau
- Apache HDFS
- FTP
- Looker
- Minio
- Pivotal Greenplum
- Snowflake
- Teradata
- Apache Hive
- Generic JDBC
- Microsoft Azure Data Lake Store
- MySQL
- PostgreSQL
- Sybase

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: **YmY4ZjlyMTg2YTFI**

Check the **Discover data assets** check box under **Connection discovery**

Select the **Auto Discovery** project from the “**Project for discovered assets**” selection list.

New connection (Db2 - Db2)

Enter information for the selected data source

Connection overview

Connection Details

Connection discovery

Project for discovered assets *

Credentials

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.

Name	Owner	Tags	Business Terms	Type	Date Added
2017 J.D. Power U.S. Auto Claims Satisfaction Stud	Shivam Solanki (IBM)	Auto Insurance, Document		Data asset	Sep 02, 2020
Db2	Shivam Solanki (IBM)			Connection	Sep 02, 2020
Vehicle Insurance Doc United States	Shivam Solanki (IBM)	Auto Insurance, Document		Data asset	Sep 02, 2020

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.

The screenshot shows a web-based interface for managing data connections. At the top, there's a breadcrumb navigation: Catalogs / Auto Insurance / Db2. Below the header, the title is CONNECTION and the specific asset name is Db2. There are two main tabs: Overview (which is selected) and Access. The Overview tab contains several sections: Description (Knowledge Catalog Tutorial Db2), Added: Sep 02, 2020 9:53 PM, Business terms (no terms available), Tags (no tags available), and Reviews (0 reviews). To the right of the Overview tab, there's a Connection Preview section with a link to edit the connection. At the bottom of the Overview tab, it says Source type: Db2. On the far right of the header, there are Remove and Download buttons.

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.

CONNECTION
Db2

Overview Access Review

Description Knowledge Catalog Tutorial Db2

Added: Sep 02, 2020 9:53 PM

Business terms There are no terms available for this asset.

Tags

Start typing to add values +

Auto Insurance Document

Source type: Db2

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.

CONNECTION
Db2

Remove Download Add to Project +

Overview Access Review

Description Knowledge Catalog Tutorial Db2

Added: Sep 02, 2020 9:53 PM

Business terms There are no terms available for this asset.

Tags

Auto Insurance Db2

Start typing to add values +

Cancel Apply

Reviews 0 reviews

Connection

Source type: Db2

Classification

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.

The screenshot shows the 'My projects' section of the IBM Cloud Pak for Data interface. The 'Auto Discovery' project is selected. The top navigation bar includes 'All', 'Search', 'Add to project', and various icons. Below the navigation is a header with tabs: 'Overview' (selected), 'Assets' (highlighted with a red box), 'Environments', 'Jobs', 'Access Control', and 'Settings'. The 'Assets' tab shows a count of 156 assets and 1 collaborator. The 'Recent activity' section displays two entries from today: 'Discovery process has completed for connection Db2 to project Auto Discovery' at 9:55 PM and 'Discovery process has started for connection Db2 to project Auto Discovery' at 9:53 PM. The 'Overview' panel on the left shows details like date created (Sep 02, 2020), description (Auto Discovery Project), storage (File System, 0 Byte used), and collaborators (Shivam Solanki (IBM) Admin). It also indicates no deployment space is associated.

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.

Name	Type	Created by	Last modified
CUSTOMER_OFFERS	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER_ATTRITION	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER_ACTIVITY	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM

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.

insured_occupant	marital_status	customer_lifetime_value	no_of_policies	no_of_closed_complaints	no_of_communications	no_of_complaints	no_of_open_complaints
craft-repair	Single	9209.41167	7	2	8	3	0
machine-op-inspect	Divorced	4428.038374	8	1	8	0	0
sales	Married	7228.993348	3	0	0	8	0
armed-forces	Single	2321.88367	1	1	4	9	0
sales	Married	3341.66335	3	0	7	5	5
tech-support	Married	8417.159538	7	0	2	3	0
prof-specialty	Single	10083.48688	2	0	0	2	0
tech-support	Single	2750.540853	1	0	2	2	0
other-service	Married	4665.245047	3	1	9	3	1
priv-house-serv	Divorced	4170.572119	1	0	1	4	0
exec-managerial	Divorced	2858.287371	1	0	8	7	3
exec-managerial	Married	5075.420953	3	2	9	0	0
protective-serv	Divorced	5935.572179	8	0	4	1	0
armed-forces	Married	3193.817363	1	0	0	5	0
machine-op-inspect	Married	7605.637524	1	0	2	8	1
transport-moving	Single	5926.729379	5	0	2	4	5
machine-op-inspect	Single	4880.713626	1	1	0	7	0
marketing	Married	1326.037705	9	n	7	4	n

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.

The screenshot shows the 'Assets' tab selected in the navigation bar. A search bar at the top contains the text 'custom'. In the main list, a checkbox is checked next to the 'CUSTOMERS' data asset. A red box highlights the 'Publish' button in the top right corner of the list header, and another red box highlights the checked checkbox for 'CUSTOMERS'.

Name	Type	Created by	Last modified
CUSTOMER_OFFERS	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER_ATTRITION	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER_ACTIVITY	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMER	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM
CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 02, 2020, 9:54 PM

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.

The screenshot shows the 'Catalogs / Auto Insurance' page. In the top right corner, there is a dropdown menu with the title 'Add to Catalog'. The 'Connected asset' option is highlighted with a red box. Below the menu, the main content area is titled 'Auto Insurance' and contains tabs for 'Browse Assets', 'Access Control', and 'Settings'. A search bar at the bottom asks 'What assets are you looking for?'. The overall interface is clean and modern, typical of a cloud-based data management tool.

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.

The screenshot shows the 'Add asset from connection' dialog box. It has several sections: 'Source' (Db2), 'Name*' (Auto Insurance Claims), 'Description' (All U.S. auto insurance claims), 'Business Terms' (Search Business Terms), 'Tags' (Auto Insurance selected), 'Classification*' (None), and 'Privacy'. At the bottom right, there are 'Cancel' and 'Add' buttons, with 'Add' being highlighted with a red box. The 'Tags' section is particularly relevant to the task at hand.

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.

The screenshot shows the 'Catalogs / Auto Insurance' interface. On the left, under 'Connections', there is one entry: 'Db2'. This is highlighted with a red box. To its right, under 'Schemas(10)', there is a list of schemas. One schema, 'AUTO_INSURANCE', is highlighted with a red box. Under 'Tables(3)', there are three tables: 'CLAIMS', 'CUSTOMERS', and 'POLICIES'. The 'CLAIMS' table is also highlighted with a red box. At the bottom right of the interface, there are two buttons: 'Cancel' and 'Select'. The 'Select' button is highlighted with a red box.

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.

The screenshot shows the 'Catalogs / Auto Insurance' interface. At the top right, there is a dropdown menu with the options 'Add to Catalog', '+', and a refresh icon. Below this, there is a list of categories: 'Local files', 'Connected asset' (which is highlighted with a red box), and 'Connection'. At the bottom of the interface, there is a search bar with the placeholder text 'Q What assets are you looking for?'.

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.

Connection	Data asset	Data asset	Data asset	Data asset	Data asset
Db2	Vehicle Insurance Doc Unit...	Auto Insurance Policies	Auto Insurance Customers	Auto Insurance Claims	2017 J.D. I
Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 10:28 PM	Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 10:27 PM	Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 10:25 PM	Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 9:53 PM	Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 9:20 PM	Owner: Shivam Solanki (IBM) Added: Sep 02, 2020 9:15 PM
Tags: Auto...	Tags: Auto...	Tags: Auto...	Tags: Db2	Tags: Auto..., Docu...	Tags: Auto...
☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews

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.

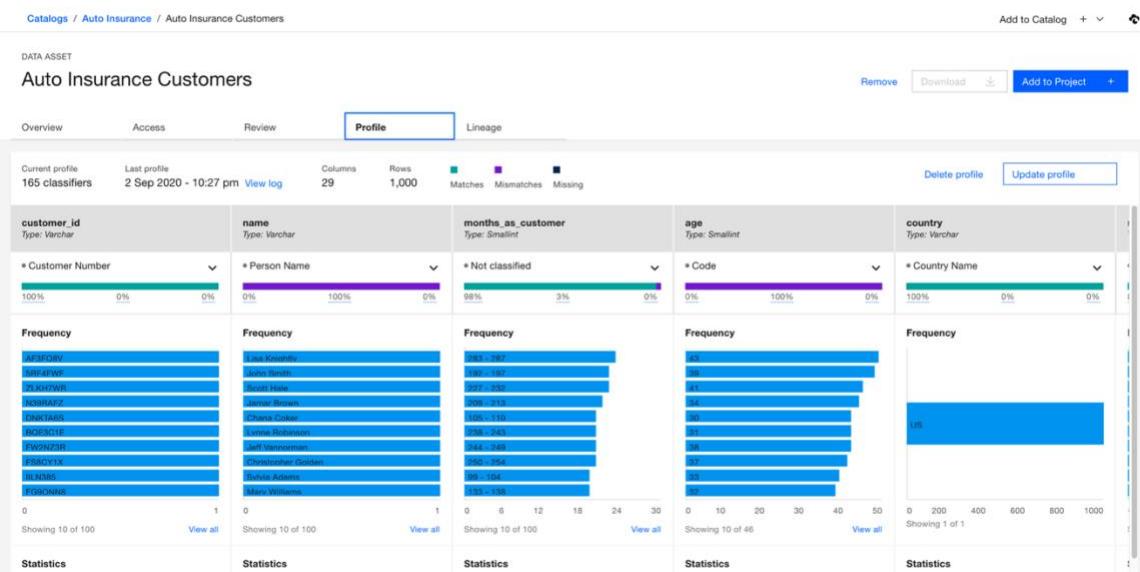
In this task, you will visit the **Profile** section of a structured data asset to examine the profiling and classification features provided. You will also visit the **Review** section to experience how you can rate and review assets to allow others to easily identify and evaluate them based on their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

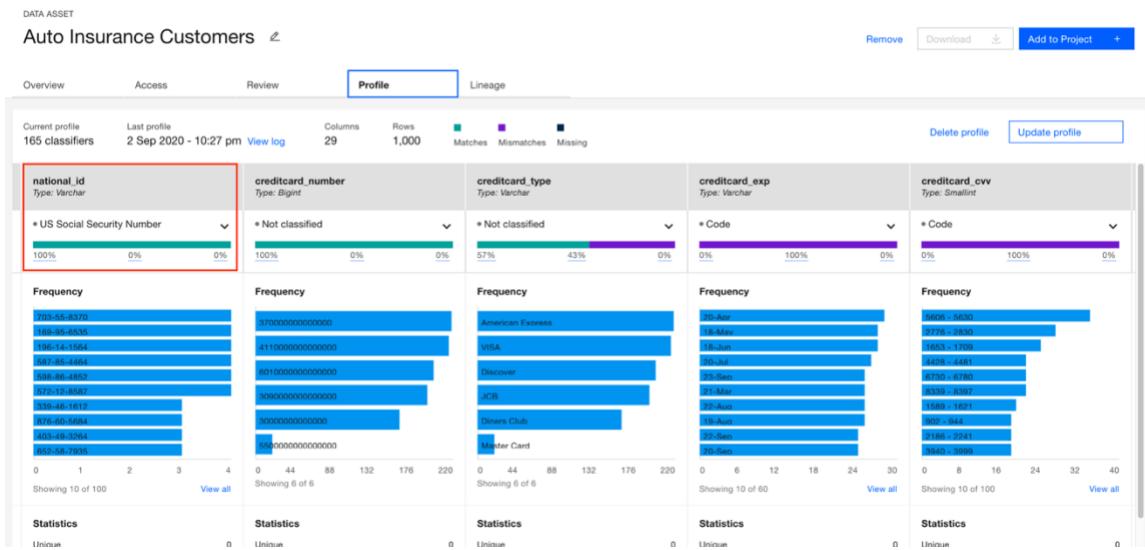
The profile should automatically appear. If not, and you are presented with a method to create or update the profile, follow the instructions to do so.



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.



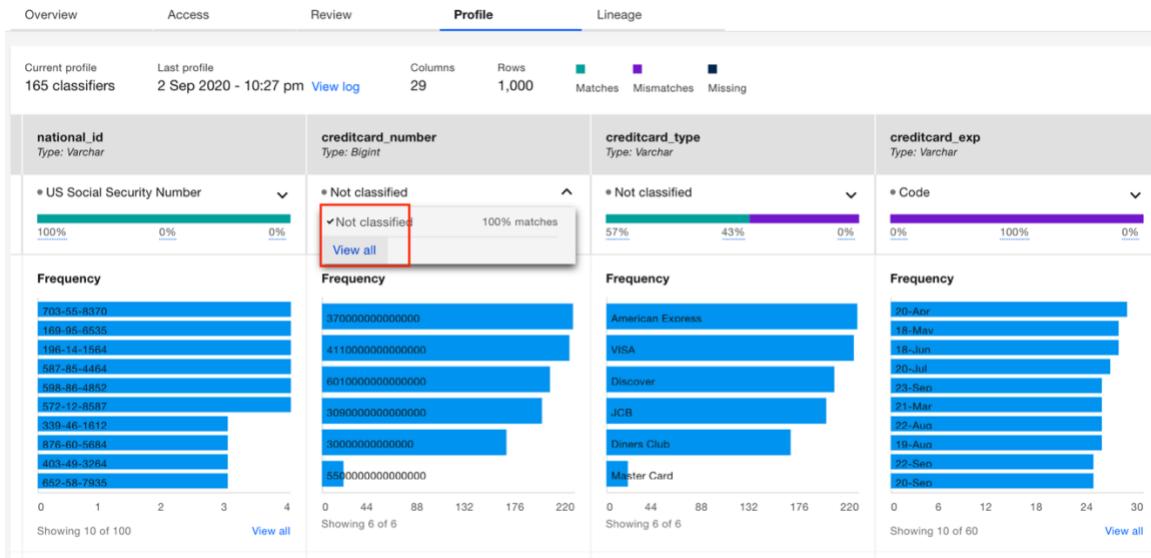
Note: You will notice that the national_id column is classified correctly but the creditcard_number, creditcard_exp, creditcard_cvv and creditcard_cvv columns are not. You will also notice they look like sensitive information and should be protected by data protection rules.

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.

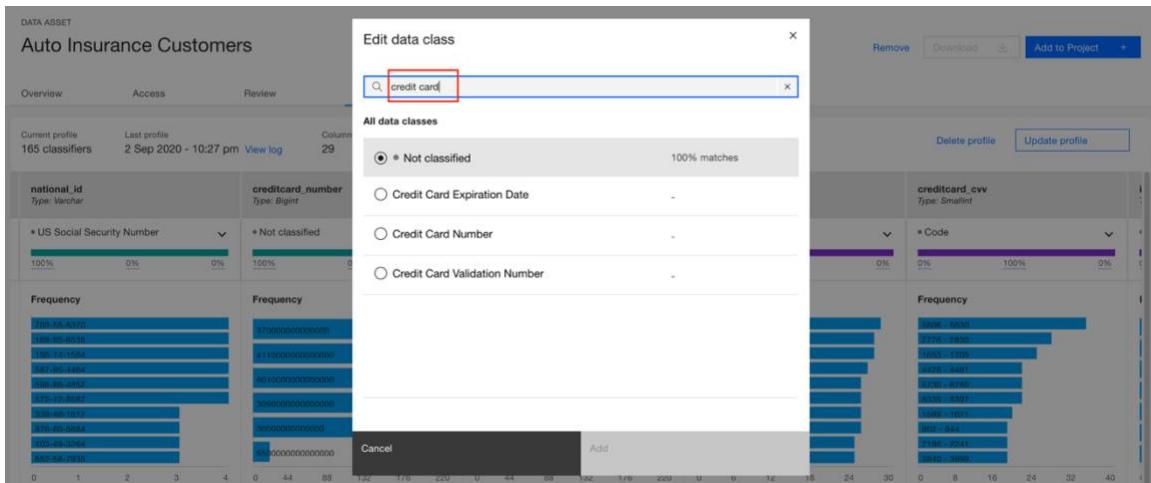
Time permitting, you will get a chance to experience how data protection works at the end of the lab so you can see it in action.

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.

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



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



- Click the **Close** button.
- Click on the down arrow next to the **Not classified** classification of the CREDITCARD_CVV column.
- Click the **View All** menu item.
- Type in **credit card** in the search area.
- Hover next to the **Credit Card Validation Number** data class and click **Use**.
- 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.

The screenshot shows the 'Review' tab of the 'Auto Insurance Customers' asset. On the left, there's a sidebar with 'Overall rating' (0.0) and a 5-star rating scale. Below that is a 'Review summary' table showing counts for each star rating. The main area is titled 'Your Review' and contains a review by 'Shivam Solanki (IBM)' dated 'Sep 02, 2020'. The review text is highlighted with a red box. At the bottom right of the review area is a 'Submit' button, which is also highlighted with a red box.

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

The screenshot shows the main details page for the 'Auto Insurance Customers' asset. It includes a navigation bar, a 'DATA ASSET' header, and a sidebar with 'Overall Rating' (4.0) and a 5-star rating scale. The main content area shows a 'My Review' section with a review by 'ctp' dated 'Jan 09, 2020', which is also highlighted with a red box. At the top right, there are 'Remove', 'Download', and 'Add to Project' buttons.

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.

The screenshot shows the 'Auto Insurance' catalog page. At the top, there are filters for 'Any type', 'Any source', 'Any tag', and a search bar. Below the filters, there are two tabs: 'Watson Recommends' and 'Highly Rated'. The 'Highly Rated' tab is selected, indicated by a red circle with the number 20. Under this tab, the 'Auto Insurance Customers' data asset is listed first, with a red arrow pointing to it. It has a 5-star rating and 1 review. Other assets listed include 'Auto Insurance Claims', 'Vehicle Insurance Doc Uni...', 'Db2 Warehouse', and '2017 J.D. P'. Below this section, there is a table with columns for Name, Owner, Tags, Business Terms, Type, and Date Added. The '2017 J.D. Power U.S. Auto Claims Satisfaction Study' asset is selected, indicated by a red circle with the number 21. The table row for this asset shows it was added by 'ctp' on 'Jan 09, 2020' and is a 'Data asset'.

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 screenshot shows the '2017 J.D. Power U.S. Auto Claims Satisfaction Study' data asset details. At the top, there are tabs for 'Overview', 'Access', 'Review', and 'Lineage'. The 'Overview' tab is selected, indicated by a red circle with the number 23. On the left, there are sections for 'Description' (Auto Insurance document), 'Business Terms' (None), 'Tags' (Auto Insurance, Document), 'Reviews' (0 reviews), and 'Classification' (None). On the right, there is a large preview window showing a blue car accident scene with the 'J.D. POWER' logo. The preview window has controls for zooming and orientation, and a red arrow labeled 22 points to the bottom right corner of the preview area.

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.

The screenshot shows the 'Review' tab of a data asset page. A large callout box highlights the review area. An arrow labeled '25' points to the star rating input field, which is set to three stars. Another arrow labeled '24' points to the text area containing the review description. A third arrow labeled '26' points to the 'Submit' button at the bottom right of the review form.

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

The screenshot shows the same data asset page after the review was submitted. The 'Overall Rating' section now shows 3.0 with one review. The review summary bar indicates one review at the 3-star level. The review itself is visible in the 'My Review' section.

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

The screenshot shows the 'Auto Insurance' catalog page. A callout box labeled '28' highlights the 'Highly Rated' section. Another callout box labeled '29' highlights the 'Collapse' button in the top right corner of the asset card for the 'Vehicle Insurance Doc Uni...' asset.

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

The screenshot shows the Knowledge Catalog interface. At the top, there's a navigation bar with 'Catalogs' and 'Auto Insurance'. On the right, there are buttons for 'Add to Catalog' and a user icon. Below the navigation is a search bar with placeholder text 'What assets are you looking for?' and a search button. Underneath the search bar are filter options: 'Any type', 'Any source', 'Any tag', and 'Clear all'. To the right of these filters is a 'Organized' button. Below the filters, there are three tabs: 'Watson Recommends' (which is selected), 'Highly Rated', and 'Recently Added'. To the right of these tabs is a 'Suggested' section with a 'Expand' button. The main area shows a table of assets with 6 items. The columns are: Name, Owner, Tags, Business Terms, Type, and Date Added. The assets listed are:

Name	Owner	Tags	Business Terms	Type	Date Added
2017 J.D. Power U.S. Auto Claims Sat...	ctp	Auto In..., Docum...		Data asset	Jan 09, 2020
Auto Insurance Claims	ctp	Auto In...		Data asset	Jan 09, 2020
Auto Insurance Customers	ctp	Auto In...		Data asset	Jan 09, 2020
Auto Insurance Policies	ctp	Auto In...		Data asset	Jan 09, 2020
Db2 Warehouse	ctp	Auto In..., Wareho...		Connection	Jan 09, 2020

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**.

The screenshot shows the 'Auto Insurance' catalog page. At the top, there are tabs for 'Browse Assets', 'Access Control', and 'Settings'. Below the tabs is a search bar containing the text 'document' (marked with a circled '1'). To the right of the search bar is a clear button (marked with a circled '2'). The main area displays a table of 'Showing 2 of 2 items' with columns: Name, Owner, Tags, Business Terms, Type, and Date Added. Two entries are listed:

Name	Owner	Tags	Business Terms	Type	Date Added
2017 J.D. Power U.S. Auto Claims Satisf...	Ricardo Buglio	Auto In... Docum...		Data asset	Aug 14, 2019
Vehicle Insurance Doc United States	Ricardo Buglio	Auto In... Docum...		Data asset	Aug 14, 2019

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.

The screenshot shows the 'Auto Insurance' catalog page. The search bar now contains 'claim' (marked with a circled '5'). To the right of the search bar is a clear button (marked with a circled '6'). The main area displays a table of 'Showing 3 of 3 items' with columns: Name, Owner, Tags, Business Terms, Type, and Date Added. Three entries are listed, with the last one ('Auto Insurance Policies') highlighted:

Name	Owner	Tags	Business Terms	Type	Date Added
2017 J.D. Power U.S. Auto Claims Satisf...	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018
Auto Insurance Claims	Rick Buglio	Auto In...		Data asset	Sep 05, 2019
Auto Insurance Policies	Rick Buglio	Auto In...		Data asset	Sep 05, 2019

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.

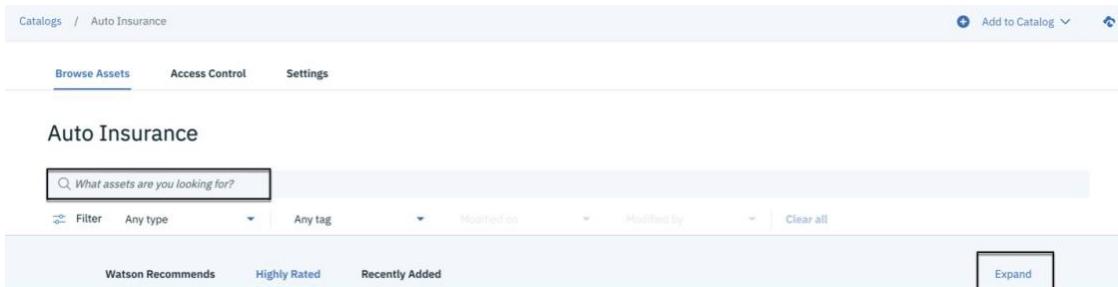
Using the same search method, type in the following searches in the **search area**. Clear the search area after each search to get the correct results:

- 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.

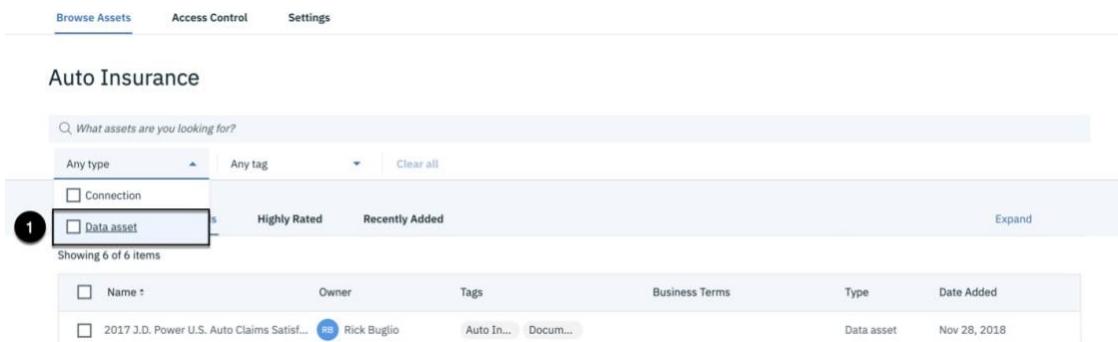
Time permitting, you can experiment on your own and type in different criteria in the **search area** to get more experience with how search works.

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.

Browse Assets Access Control Settings

Auto Insurance

What assets are you looking for?

Type Any tag

Showing 5 of 5 items

<input type="checkbox"/>	Name	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/>	2017 J.D. Power U.S. Auto Claims Satisf...	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018
<input type="checkbox"/>	Auto Insurance Claims	Rick Buglio	Auto In...		Data asset	Sep 05, 2019
<input type="checkbox"/>	Auto Insurance Customers	Rick Buglio	Auto In...		Data asset	Sep 05, 2019
<input type="checkbox"/>	Auto Insurance Policies	Rick Buglio	Auto In...		Data asset	Sep 05, 2019
<input type="checkbox"/>	Vehicle Insurance Doc United States	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018

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 catalogued are **Connections, Models, Notebooks and Dashboards**.

Browse Assets Access Control Settings

Auto Insurance

What assets are you looking for?

Type Any tag

Showing 5 of 5 items

<input type="checkbox"/>	Name	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/>	Auto Insurance					
<input type="checkbox"/>	Document					
<input type="checkbox"/>	2017 J.D. Power U.S. Auto Claims Satisf...	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018

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

Browse Assets Access Control Settings

Auto Insurance

What assets are you looking for?

Type Tag

Showing 2 of 2 items

<input type="checkbox"/>	Name	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/>	2017 J.D. Power U.S. Auto Claims Satisf...	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018
<input type="checkbox"/>	Vehicle Insurance Doc United States	Rick Buglio	Auto In... Docum...		Data asset	Nov 28, 2018

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

- 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.

The screenshot shows the 'Auto Insurance' catalog page. At the top, there are tabs for 'Browse Assets', 'Access Control', and 'Settings'. Below the tabs, there's a search bar and filters for 'Any type', 'Any source', 'Any tag', and a 'Clear all' button. A 'Watson Recommends' section is visible above the main list. The main area displays a table of data assets:

	Name	Owner	Tags	Business Terms	Type	Date Added
1	2017 J.D. Power U.S. Auto Claims Satisfaction Survey	ctp	Auto Insur...	Document	Data asset	Jan 09, 2020
2	Auto Insurance Claims	ctp	Auto Insur...		Data asset	Jan 09, 2020
3	Auto Insurance Customers	ctp	Auto Insur...		Data asset	Jan 09, 2020
4	Auto Insurance Policies	ctp	Auto Insur...		Data asset	Jan 09, 2020
	Db2 Warehouse	ctp	Auto Insur...	Warehouse	Connection	Jan 09, 2020

A modal dialog box is open over the table, titled 'Add to Project'. It contains a list of selected assets: 'Auto Insurance Claims', 'Auto Insurance Customers', and 'Auto Insurance Policies'. There are buttons for 'Add to Project' (highlighted with a red box), 'Remove', 'Expand', and 'Cancel'. The number '4' is displayed in a circle at the top left of the modal.

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.

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

Catalogs / Auto Insurance

Add to Project

Target*

Auto Discovery

Selected assets (3)

Asset Name	Catalog	Connection
Auto Insurance Claims	Auto Insurance	Db2
Auto Insurance Customers	Auto Insurance	Db2
Auto Insurance Policies	Auto Insurance	Db2

Connections to be added (1)

Db2	Auto Insurance Knowledge Catalog Tutorial Db2
-----	---

Cancel Add

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.

- Select the **Auto Insurance** project from the list of Target projects.
- Click the **Add** button.

IBM Cloud Pak for Data

All Search

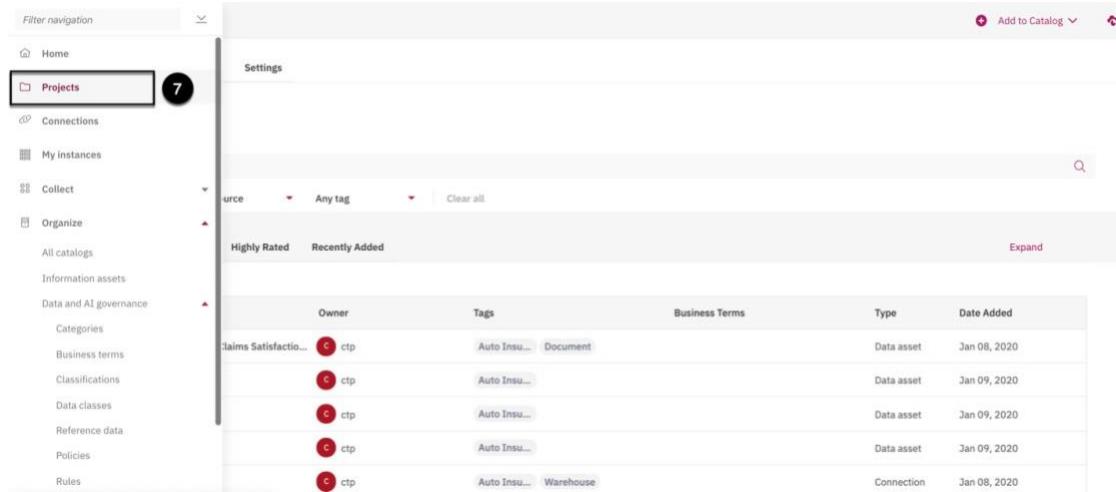
Catalogs > Auto Insurance

Add to Catalog

Browse Assets Access Control Settings

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.



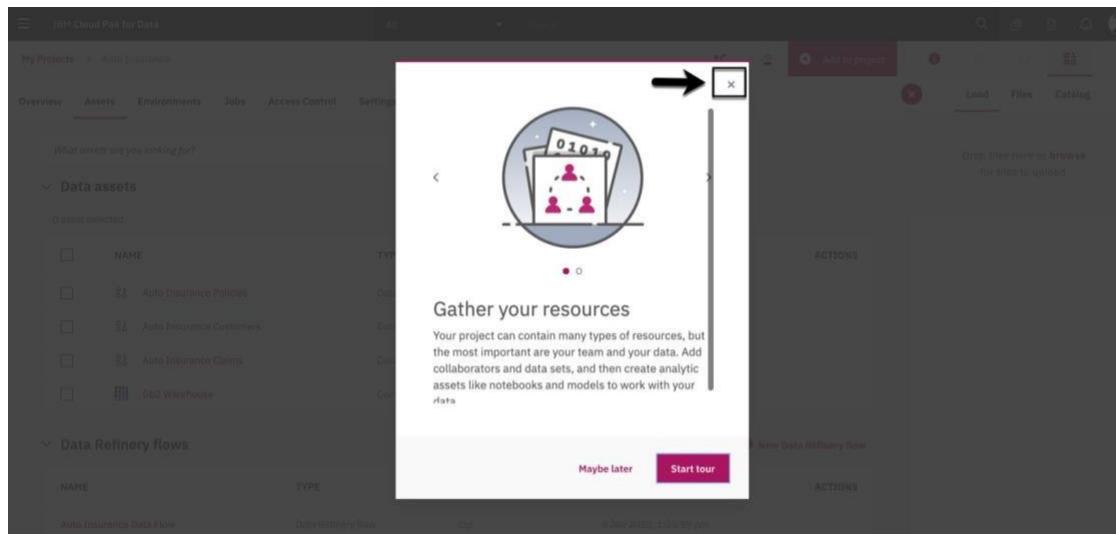
The screenshot shows the navigation menu on the left side of the interface. The 'Projects' option is highlighted with a red box and a black number '7' indicating it is the next step. Other options include 'Connections', 'My instances', 'Collect', 'Organize', 'All catalogs', 'Information assets', 'Data and AI governance', 'Categories', 'Business terms', 'Classifications', 'Data classes', 'Reference data', 'Policies', and 'Rules'. To the right of the menu is a search bar and a table titled 'Settings' showing various data assets and connections.

8. Click the **Projects** menu.



The screenshot shows the 'Projects' list page. The 'Auto Insurance' project is selected and highlighted with a red box and a black number '8'. The table lists two projects: 'Auto Insurance' and 'Auto Discovery'. The 'Auto Insurance' project is categorized under 'Analytics' and has an 'Admin' user role, last updated on 8 Jan 2020, 4:05 PM. The 'Auto Discovery' project is also under 'Analytics' and has an 'Admin' user role, last updated on 9 Jan 2020, 9:07 AM.

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



The screenshot shows the 'Auto Insurance' project dashboard. On the left, there are sections for 'Data assets' (listing 'Auto Insurance POHOS', 'Auto Insurance Customers', 'Auto Insurance Claims', and 'Auto Warehouse') and 'Data Refinery flows' (listing 'Auto Insurance Data Flow'). A central 'Gather your resources' section features a circular icon with three people and a stack of documents, with the text: 'Your project can contain many types of resources, but the most important are your team and your data. Add collaborators and data sets, and then create analytic assets like notebooks and models to work with your data.' At the bottom right of this section are 'Maybe later' and 'Start tour' buttons. A large 'X' button is located in the top right corner of the dialog box.

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

Refine the Data

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

The screenshot shows the 'Assets' tab selected in the navigation bar. Below it is a table listing four data assets: 'Auto Insurance Customers', 'Auto Insurance Policies', 'Auto Insurance Claims', and 'Db2 Warehouse'. The 'Auto Insurance Customers' row has a context menu open, with the 'Refine' option highlighted.

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

The screenshot shows the 'Data Refinery Flow Details' section. It includes fields for 'LOCATION' (Auto Insurance), 'DATA REFINERY FLOW NAME' (Auto Insurance Customers_flow), and 'DATA SET NAME' (Auto Insurance Customers_s...). On the left, there's a table titled 'Get perspective on your data' with 15 rows of sample data from the 'Auto Insurance Customers' source file. The first few rows show columns for COUNTRY, LATITUDE, and LONGITUDE.

	COUNTRY	LATITUDE	LONGITUDE
1	US	41.75113981	-88.0127658
2	US	39.2781	-120.1203
3	US	47.76121	-122.3464
4	US	33.88081187	-118.0288831
5	US	34.02091598	-84.31698227
6	US	41.77025751	-88.20481022
7	AR47849	Janine McCreathe	40.9581
8	AS97690	Milicent Caveau	40.891
9	AW77988	Agnes Woodfield	38.9974
10	AY40674	Jamail Duddle	35.080383
11	AZ34845	Dov Gabriely	42.339208
12	BA75404	Dee dee Mugglesstone	42.43478788
13	BB82067	Elleray Glorershaw	39.58050569
14	BC66536	Joelyn Pilgram	33.98205235
15	BD87486	Marie Flinders	41.9138533

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

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

The screenshot shows the Data Refinery interface. On the left, there's a table titled 'Auto Insurance Customers' with columns: CUSTOMER, NAME, COUNTRY, LATITUDE, and LONGITUDE. The table contains 11 rows of data. On the right, the 'DATA REFINERY FLOW DETAILS' panel is open, showing 'LOCATION: Auto Insurance' and 'DATA REFINERY FLOW NAME: Auto Insurance Customers_flow'. A circled number '4' is located in the top right corner of the interface.

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

The screenshot shows the 'DATA REFINERY FLOW DETAILS' panel. The 'DATA REFINERY FLOW NAME' field is highlighted and has a circled number '5' above it. The field contains the text 'Auto Insurance Customers...'. Below the field is a note: 'Enter a description of the Data Refinery flow'. The 'LOCATION' field is set to 'Auto Insurance'. The 'DATA SET NAME' field is also visible on the right side of the screen.

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

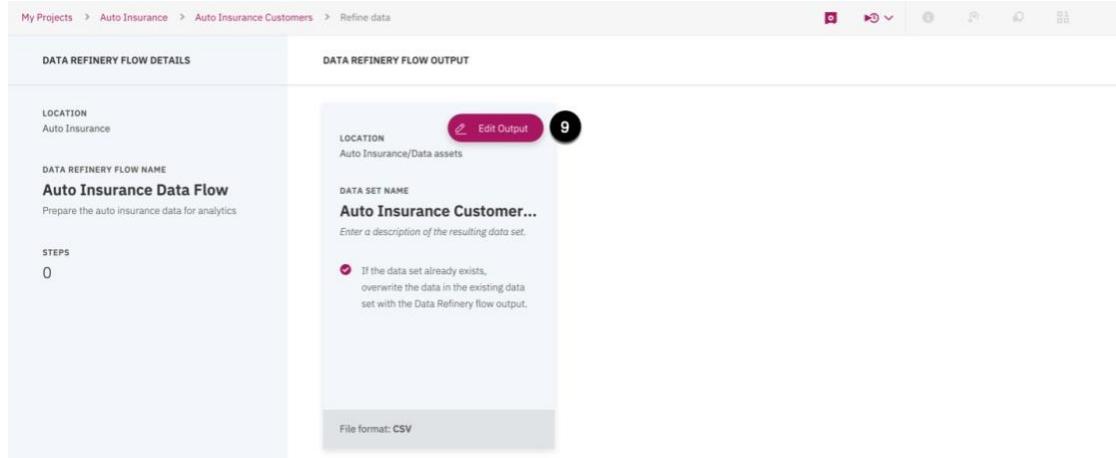
The screenshot shows the 'DATA REFINERY FLOW DETAILS' panel. The 'DATA REFINERY FLOW NAME' field now contains 'Auto Insurance Data Flow' (circled with '6'). The 'DESCRIPTION' field below it contains the text 'Prepare the auto insurance data for analytics' (circled with '7'). At the bottom, there are 'Cancel' and 'Apply' buttons, with 'Apply' highlighted and circled with '8'. The 'LOCATION' field is still set to 'Auto Insurance'. The 'DATA SET NAME' field and other settings are visible on the right.

- 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.

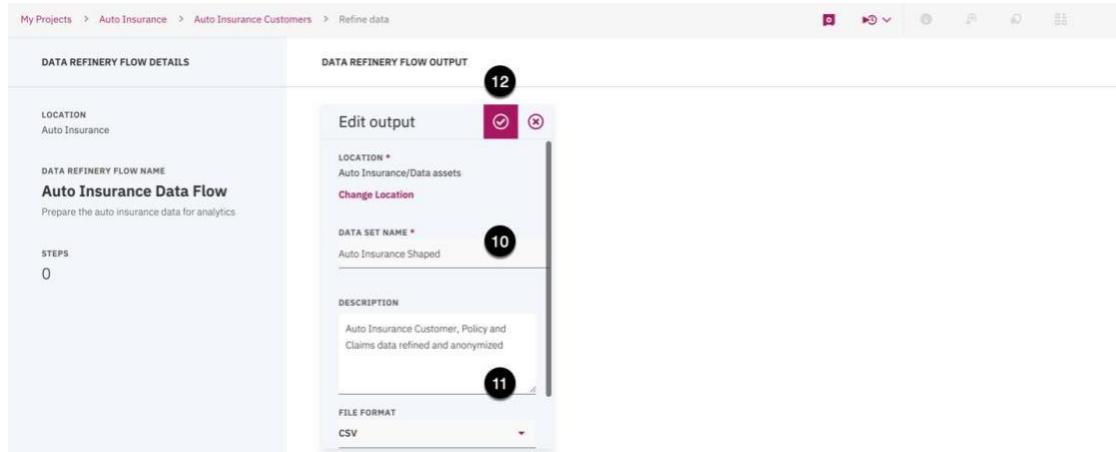
9. Click the **Edit Output** button to change the DATA SET NAME.



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.

The screenshot shows the Data Refinery interface with two main sections: 'DATA REFINERY FLOW DETAILS' on the left and 'DATA REFINERY FLOW OUTPUT' on the right. In the 'Details' section, the 'LOCATION' is set to 'Auto Insurance'. The 'DATA REFINERY FLOW NAME' is 'Auto Insurance Data Flow', described as 'Prepare the auto insurance data for analytics'. The 'STEPS' count is 0. In the 'Output' section, the 'LOCATION' is 'Auto Insurance/Data assets', the 'DATA SET NAME' is 'Auto Insurance Shaped', described as 'Auto Insurance Customer, Policy and Claims data refined and anonymized'. A note indicates that if the data set already exists, it will be overwritten. The file format is specified as CSV. A 'Done' button is visible at the bottom right.

13. Click the **Done** button.

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

The screenshot shows the Data Refinery interface with the 'Edit' tab selected in the Details panel. The 'Data' tab is active, displaying a table of customer data with columns: CUSTOMER (String), NAME (String), COUNTRY (String), LATITUDE (String), and LONGITUDE (String). The table lists 15 rows of data. The 'Steps' section shows 0 steps. The 'Data Source' is 'Auto Insurance Customers'. The 'DATA REFINERY FLOW DETAILS' section includes the 'LOCATION' (Auto Insurance), 'DATA REFINERY FLOW NAME' (Auto Insurance Data Flow), and a description ('Prepare the auto insurance data for analytics'). The 'STEPS' section shows 0 steps. The 'DATA REFINERY FLOW OUTPUT' section includes the 'LOCATION' (Auto Insurance/Data assets) and 'DATA SET NAME' (Auto Insurance Shaped).

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

Combine Data

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

The screenshot shows the 'Operation' menu with the 'Join' option selected. The right side of the screen displays a preview of a data table with columns: CUSTOMER, NAME, COUNTRY, LATITUDE, LONGITUDE, and STREET_ADDRESS. The data consists of 10 rows of customer information from the 'Auto Insurance Customers' file. At the bottom, it says 'SOURCE FILE: Auto Insurance Customers' and 'SAMPLE SIZE: First 1000 rows'.

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

The screenshot shows the 'Join' configuration screen. The 'Inner join' method is selected. In the 'Data set to join' section, the 'Auto Insurance ...' asset is listed under 'Source'. There are fields for 'Prefix' and 'Suffix' with values '_X' and '_Y' respectively. The right side shows the same data preview as the previous screen, with the 'SOURCE FILE: Auto Insurance Customers' and 'SAMPLE SIZE: First 1000 rows' details at the bottom.

- Select the **Inner join** method from the join method list.
- Click the **+ Add Data Set** button in the **Data set to join** section.
- Click on the **Data assets** section.
- Select the **Auto Insurance Claims** data asset.
- 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.

The screenshot shows the 'Join' operation interface in Data Workshop. The 'JOIN KEYS' section is open, showing two dropdown menus: one for the source dataset ('Auto Insurance ...') and one for the target dataset ('Auto Insurance ...'). Both dropdowns have a suffix of '-X'. The 'JOIN KEYS' dropdown has a suffix of '-Y'. The 'CUSTOMER' entry is selected in both dropdowns. The 'Next' button at the bottom is highlighted with a large black circle containing the number 10.

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.

The screenshot shows the 'Join' operation interface in Data Workshop. On the left, a list of columns is shown with several checkboxes. The 'EDUCATION' checkbox is checked. The 'Back' button at the bottom is highlighted with a large black circle containing the number 12.

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.

The screenshot shows the 'Policies Columns' shaper in the Cloud Pak for Data Workshop. The table contains 12 rows of data from the 'Auto Insurance Customers' source. A callout box with the text 'Scroll to the Right' points to the right side of the table, where new columns from the joined 'Auto Insurance Policies' table are visible, such as 'POLICY_TYPE', 'POLICY', 'RENEW_OFFER', and 'SALES'. The 'Steps' panel on the right shows a single 'Join' step with the description: 'inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER'.

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.

The screenshot shows the 'Operation' menu in the Cloud Pak for Data Workshop. The 'Join' operation is highlighted with a callout box labeled '14'. Other operations listed include Replace missing values, Replace substring, ORGANIZE, Aggregate, Concatenate, Conditional replace, Sample, Split column, NATURAL LANGUAGE, Remove stop words, and Tokenize. The main workspace shows the same 'Policies Columns' shaper and 'Steps' panel as in the previous screenshot.

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

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

The screenshot shows the 'Join' operation configuration. The 'Join' method is set to 'Inner join' (marked with a circled '15'). The 'Data set to join' section contains two entries: 'Auto Insurance ...' with suffixes '*Suffix -x' and 'Auto Insurance ...' with suffixes '*Suffix -y'. A blue button labeled '+ Add Data Set' is visible. The right side of the screen displays a preview of the joined data, showing columns like POLICY_ID, COVERAGE, EFFECTIVE_TO..., and POLICY_TYPE. The preview table has 10 rows. The 'Steps' panel on the right shows one step: 'Data Source' set to 'Auto Insurance Customers'.

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.

The screenshot shows the 'Join' operation configuration with the 'Inner join' method selected. The 'Data set to join' section now includes 'Auto Insurance ...' with suffixes '*Suffix -x' and 'Auto Insurance ...' with suffixes '*Suffix -y'. A blue button labeled '+ Add Data Set' is visible. A modal window titled 'JOIN KEYS' is open, showing a list of columns from both data sources: 'Auto Insurance Cus...' and 'Auto Insurance Clai...'. The 'POLICY_ID' column is selected in the 'Auto Insurance Cus...' list. A blue button labeled '+ Add Join Key' is visible at the bottom of the modal. The right side of the screen displays a preview of the joined data, showing columns like LATITUDE, STREET_ADDRESS, CITY, and STATE. The preview table has 10 rows. The 'Steps' panel on the right shows one step: 'Data Source' set to 'Auto Insurance Customers'.

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 screenshot shows the Data Shaper interface with a 'Join' step selected. On the left, a list of columns from the 'Auto Insurance Customers' table is shown with checkboxes next to them. A column named 'NATIONAL_ID' has a checkbox checked and is highlighted with a black circle containing the number '23'. To the right is a table of 'Policies' data with columns: POLICY_ID, COVERAGE, EFFECTIVE_TO..., and POLICY_TYPE. The 'Steps' panel on the right shows two steps: 'Data Source' (Auto Insurance Customers) and 'Join' (inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER). The 'Join' step is labeled 'JUST ADDED'. At the bottom, there are 'Back' and 'Apply' buttons, and status information: SOURCE FILE: Auto Insurance Customers and SAMPLE SIZE: First 463 rows.

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.

The screenshot shows the Data Flow interface with a 'Claims Columns' step selected. On the left is a table of 'Claims' data with columns: CLAIM_ID, FIRST_NOTICE..., RESPONSE, CLAIM_REASON, and INCIDENT_SUMMARY. The 'Steps' panel on the right shows two steps: 'Data Source' (Auto Insurance Customers) and 'Join' (inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER). The 'Join' step is labeled 'JUST ADDED'. At the bottom, there is a 'Save' button with a black circle containing the number '24'. A callout box points to the 'Save' button with the text 'Scroll to the right' and an arrow pointing to the right. Status information at the bottom includes SOURCE FILE: Auto Insurance Customers and SAMPLE SIZE: First 143 rows.

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

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

The screenshot shows the Data Workshop interface with the 'Operation' tab selected. On the left, a sidebar lists 'FREQUENTLY USED' operations: Calculate, Convert column type, Filter, Math, Remove, Rename, Sort ascending, Sort descending, Substitute, Text, and CLEANSE. The 'Remove' option is highlighted with a red box and a step number '2'. The main area displays a table of data with columns: CLAIM_ID, FIRST_NOTICE..., RESPONSE, and CLAIM_REASON. To the right, the 'Steps' panel shows two steps: 'Data Source' (Auto Insurance Customers) and 'Join' (inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER). Below the table, it says 'SOURCE FILE: Auto Insurance Customers' and 'SAMPLE SIZE: First 143 rows'.

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

The screenshot shows the Data Workshop interface with the 'Column selection' step selected. On the left, a sidebar lists 'To begin, select a column.' with options: claim, claim_id_x, claim_id_y, and claims_file. The 'claim' option is highlighted with a red box and a step number '2'. The main area displays a table of data with columns: customer_id, name, months_as_cus..., and age. To the right, the 'Steps' panel shows two steps: 'Data Source' (Auto Insurance Customers) and 'Join' (inner-joined data from Auto Insurance Claims based on columns customer_id,customer_id). Below the table, it says 'SOURCE FILE: Auto Insurance Customers' and 'SAMPLE SIZE: First 1000 rows'.

- Click the **Next** button.
- Click the **Apply** button.
- Go to the toolbar and Click the **Save** button to save the data flow.
- Then remove **Claim_id_x** as well.

7 Code an operation to cleanse and shape your data

8

Steps

3 STEPS
Auto Insurance Customers
Join
inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER
Join
inner-joined data from Auto Insurance Claims based on columns POLICY_ID,POLICY_ID
Rename column
Renamed column CUSTOMER_x to CUSTOMER

SOURCE FILE: Auto Insurance Customers **SAMPLE SIZE:** First 1000 rows

Create New Data

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

1 Code an operation to cleanse and shape your data

2

Steps

4 STEPS
Auto Insurance Customers
Join
inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER
Join
inner-joined data from Auto Insurance Claims based on columns POLICY_ID,POLICY_ID
Rename column
Renamed column CUSTOMER_x to CUSTOMER

SOURCE FILE: Auto Insurance Customers **SAMPLE SIZE:** First 1000 rows

2. Scroll down to the ORGANIZE section and click the **Concatenate** operation.

3. Select the **street_address** column from the column selection list.

3

4

Steps

4 STEPS
Auto Insurance Customers
Join
inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER
Join
inner-joined data from Auto Insurance Claims based on columns POLICY_ID,POLICY_ID
Rename column
Renamed column CUSTOMER_x to CUSTOMER

SOURCE FILE: Auto Insurance Customers **SAMPLE SIZE:** First 1000 rows

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

Concatenate

Selected column: STREET_ADDRESS

Select the columns to concatenate with STREET_ADDRESS in the order you want.

1. STREET_ADDRESS

2. CITY

3. STATE

4. ZIP_CODE

5. EMAIL_ADDRESS

6. PHONE_NUMBER

STREET_ADDRESS
String
1001 W 75th Street
1001 W 75th Street
1001 W 75th Street
100 Northstar Dr
100 Northstar Dr
100 Northstar Dr
18325 Aurora Ave N
18325 Aurora Ave N
18325 Aurora Ave N
16610 Valley View Avenue
16610 Valley View Avenue

CITY
String
Woodridge
Woodridge
Woodridge
Truckee
Truckee
Truckee
Truckee
Shoreline
Shoreline
Shoreline
La Mirada
La Mirada

SOURCE FILE: Auto Insurance Customers SAMPLE SIZE: First 1000 rows

4 STEPS

Auto Insurance Customers

Join

inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER

Rename column

Renamed column CUSTOMER_x to CUSTOMER

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

Concatenate

Selected column: STREET_ADDRESS

Select the columns to concatenate with STREET_ADDRESS in the order you want.

1. STREET_ADDRESS

2. CITY

3. STATE

4. STATE_CODE

5. ZIP_CODE

6. EMAIL_ADDRESS

7. PHONE_NUMBER

STREET_ADDRESS
String
1001 W 75th Street
1001 W 75th Street
1001 W 75th Street
100 Northstar Dr
100 Northstar Dr
100 Northstar Dr
100 Northstar Dr
18325 Aurora Ave N
18325 Aurora Ave N
18325 Aurora Ave N
16610 Valley View Avenue
16610 Valley View Avenue

CITY
String
Woodridge
Woodridge
Woodridge
Truckee
Truckee
Truckee
Truckee
Shoreline
Shoreline
Shoreline
La Mirada
La Mirada

SOURCE FILE: Auto Insurance Customers SAMPLE SIZE: First 1000 rows

4 STEPS

Auto Insurance Customers

Join

inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER

Rename column

Renamed column CUSTOMER_x to CUSTOMER

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

STREET_ADDRESS	CITY	STATE_CODE
1001 W 75th Street	Woodridge	IL
1001 W 75th Street	Woodridge	IL
1001 W 75th Street	Woodridge	IL
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
18325 Aurora Ave N	Shoreline	WA
18325 Aurora Ave N	Shoreline	WA
18325 Aurora Ave N	Shoreline	WA
16610 Valley View Avenue	La Mirada	CA
16610 Valley View Avenue	La Mirada	CA

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

STREET_ADDRESS	CITY	STATE_CODE
1001 W 75th Street	Woodridge	IL
1001 W 75th Street	Woodridge	IL
1001 W 75th Street	Woodridge	IL
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
100 Northstar Dr	Truckee	CA
18325 Aurora Ave N	Shoreline	WA
18325 Aurora Ave N	Shoreline	WA
18325 Aurora Ave N	Shoreline	WA
16610 Valley View Avenue	La Mirada	CA
16610 Valley View Avenue	La Mirada	CA

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.

The screenshot shows the Data Workshop interface with the 'Concatenate' operation selected. The 'Steps' panel on the right lists four steps: 'Auto Insurance Customers' (Join), 'inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER', 'Join', and 'inner-joined data from Auto Insurance Claims based on columns POLICY_ID,POLICY_ID'. The main area displays a preview of the concatenated 'ADDRESS' column, which is the result of joining three columns: 'STREET_ADDRESS', 'CITY', and 'STATE_CODE'. The preview shows rows such as '1001 W 75th Street Woodridge IL' and '18325 Aurora Ave N Shoreline WA'. The 'Name of the concatenated column' field is set to 'ADDRESS'.

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.

The screenshot shows the Data Workshop interface with the 'Data' tab selected. The main area displays a table with columns: NAME, COUNTRY, LATITUDE, LONGITUDE, and ADDRESS. The ADDRESS column contains concatenated values such as '1001 W 75th Street Woodridge IL 60517'. The 'Steps' panel on the right shows five steps: 'Auto Insurance Customers' (Join), 'inner-joined data from Auto Insurance Policies based on columns CUSTOMER,CUSTOMER', 'Join', 'inner-joined data from Auto Insurance Claims based on columns POLICY_ID,POLICY_ID', and 'Rename column' (Renamed column CUSTOMER_x to CUSTOMER).

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.

The screenshot shows a data preview table with columns: PHONE_NUMBER, GENDER, NATIONAL_ID, EDUCATION, EMPLOYMENT_..., and MARITAL_STAT.... The NATIONAL_ID column contains sensitive SSN values. A vertical scroll bar on the right side of the table is highlighted with a red box and the text 'Scroll to the right' with an arrow pointing to it. A callout bubble labeled '1' points to the ellipsis menu icon in the header of the NATIONAL_ID column. Another callout bubble labeled '2' points to the 'Substitute' option in the context 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.

The screenshot shows the same data preview table after the **Substitute** operation has been applied. The NATIONAL_ID column now contains anonymized values like 64ee8c6d0d363e3362..., 64ee8c6d0d363e3362..., etc. The rest of the table and the operations sidebar remain the same.

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.

The screenshot shows the Data Refinery interface with the 'Visualizations' tab selected. A context menu is open over the 'CUSTOMER' column, listing options like 'Remove', 'Remove duplicates', 'Remove empty rows', 'Sort ascending' (which is highlighted with a black circle), 'Sort descending', 'Substitute', 'CONVERT COLUMN...', 'TEXT', and 'View All'. The 'Steps' button in the top right corner is also highlighted with a black circle. The toolbar at the top includes a save button (circled 3) and other standard icons.

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.

The screenshot shows the Data Refinery interface with the 'Visualizations' tab selected. A context menu is open over the 'STATE' column in the 'COLUMNS TO VISUALIZE' area, listing 'state' (highlighted with a black circle) and 'STATE_CODE'. The 'CHART TYPES' section is visible above, showing various chart icons. The 'Actions' button in the top right corner of the visualization panel is also highlighted with a black circle.

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.

The screenshot shows the Cloud Pak for Data Workshop interface. The top navigation bar includes 'Operation', 'Data', 'Profile', 'Visualizations' (which is the active tab), and 'Steps'. Below the tabs are 'CHART TYPES' with 'Suggested charts' listed. The main area is titled 'Choose a chart above or select columns below, and then choose a chart. If you select columns, suggested charts will be indicated with a dot next to the chart name.' A 'COLUMNS TO VISUALIZE' section contains a dropdown menu with 'STATE' selected. Below it is a blue button labeled 'Add column' with a circled number 4. To the right, there are three small preview icons representing different chart types: a bubble chart, a bar chart, and a line chart.

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

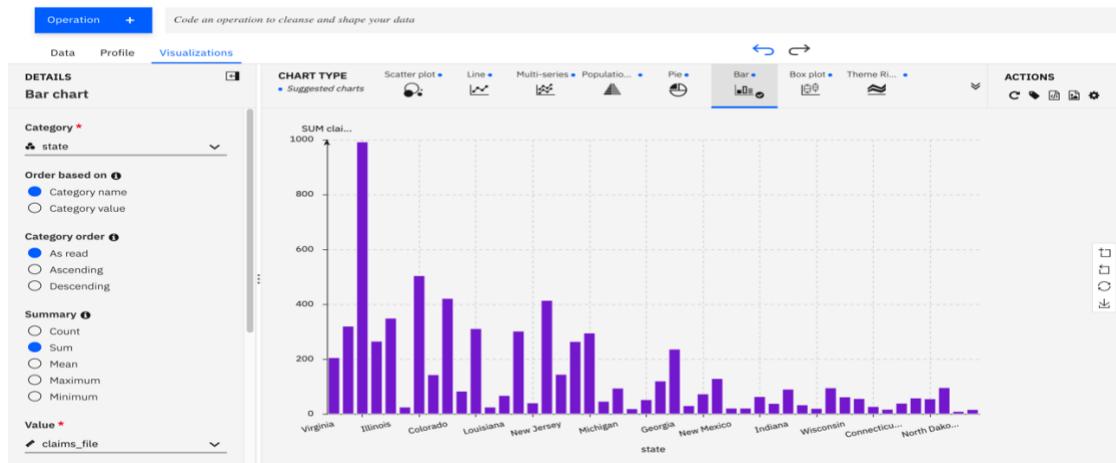
The screenshot shows the same interface as the previous one, but the 'COLUMNS TO VISUALIZE' section now lists three columns: 'STATE', 'claims', 'DENIED CLAIMS', and 'CLAIMS FILED'. The 'claims' entry is highlighted with a circled number 5, and the 'CLAIMS FILED' entry is highlighted with a circled number 6. The rest of the interface remains the same, including the 'CHART TYPES' section and the preview icons.

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

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

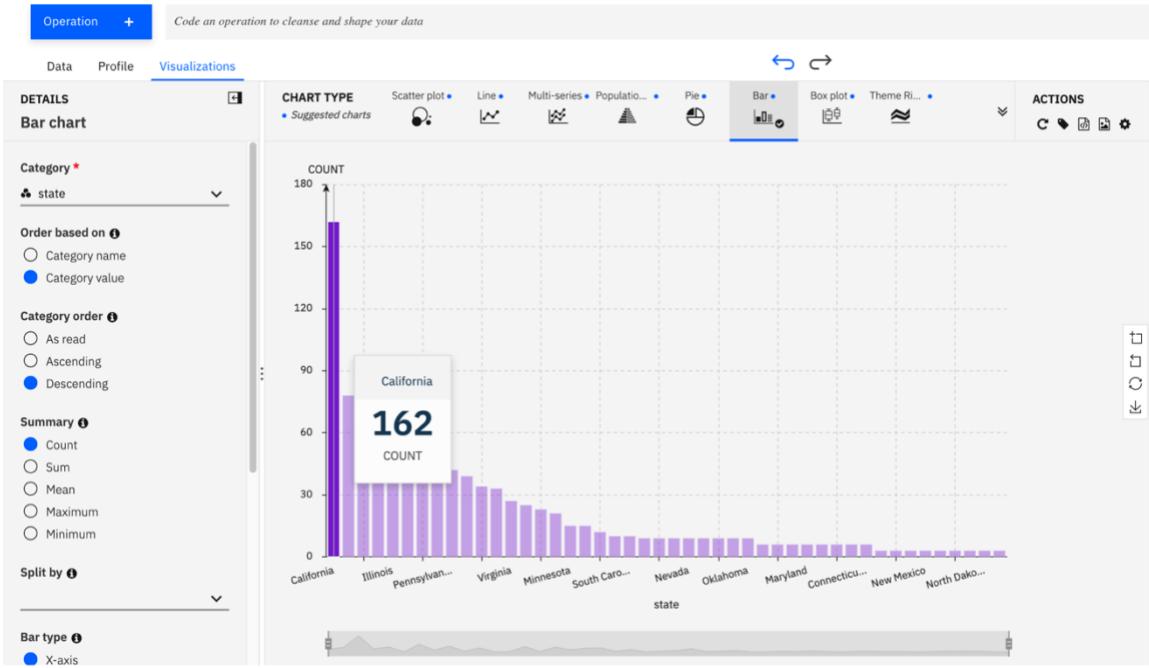
The screenshot shows the Cloud Pak for Data Workshop interface. The top navigation bar includes 'Operation', 'Data', 'Profile', and 'Visualizations'. The 'Visualizations' tab is active. The 'CHART TYPES' toolbar below the navigation bar has several options: Scatter plot, Line, Multi-series, Population, Pie, Bar (which is highlighted with a large number '7'), Box plot, Error bar, Dual Y-axes, Histogram, and Q-Q plot. To the right of the chart types is a section titled 'ACTIONS' with various icons. Below the toolbar, there is a note: 'Choose a chart above or select columns below, and then choose a chart. If you select columns, suggested charts will be indicated with a dot next to the chart name.' Under 'COLUMNS TO VISUALIZE', there are two entries: 'STATE' and 'CLAIMS_FILED'. A placeholder 'Add column' is also present. On the right side of the interface, there is a decorative graphic featuring three overlapping rectangular boxes with data visualization icons (bar chart, pie chart, line graph).

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.



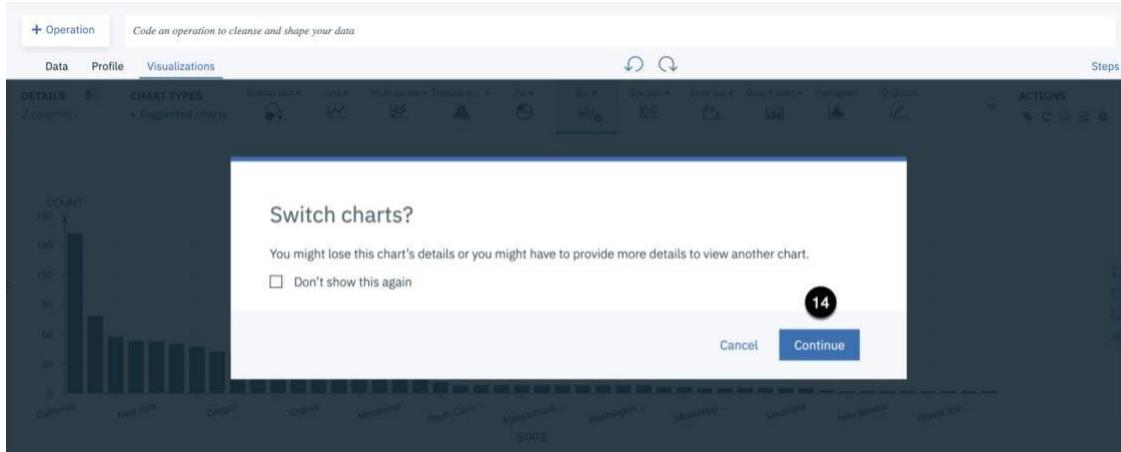
Note: The number of claims filed you see may be different than what is displayed on the screenshot because this is a multi-tenant environment with updates, deletes and inserts happening on a regular basis.

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).

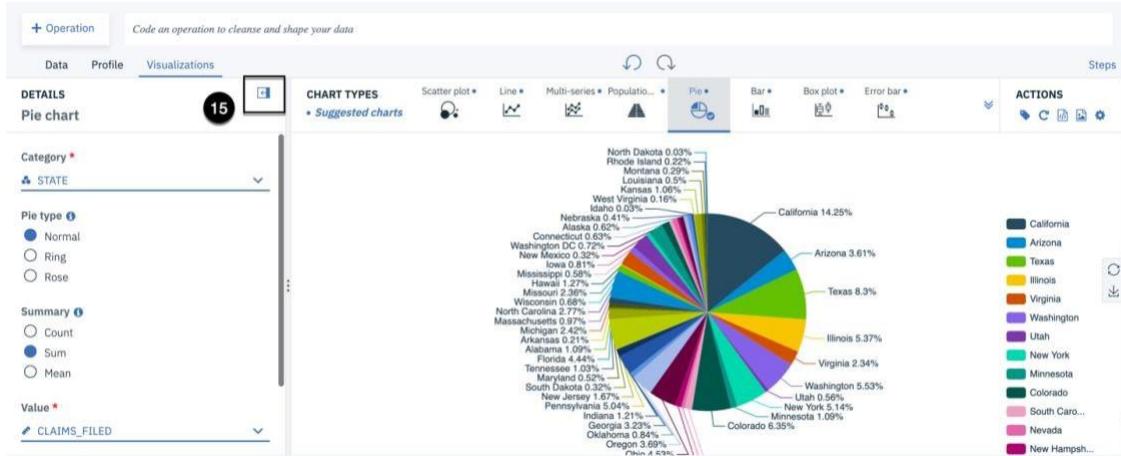
13. Click on the **Pie** chart from the **CHART TYPES** toolbar.

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

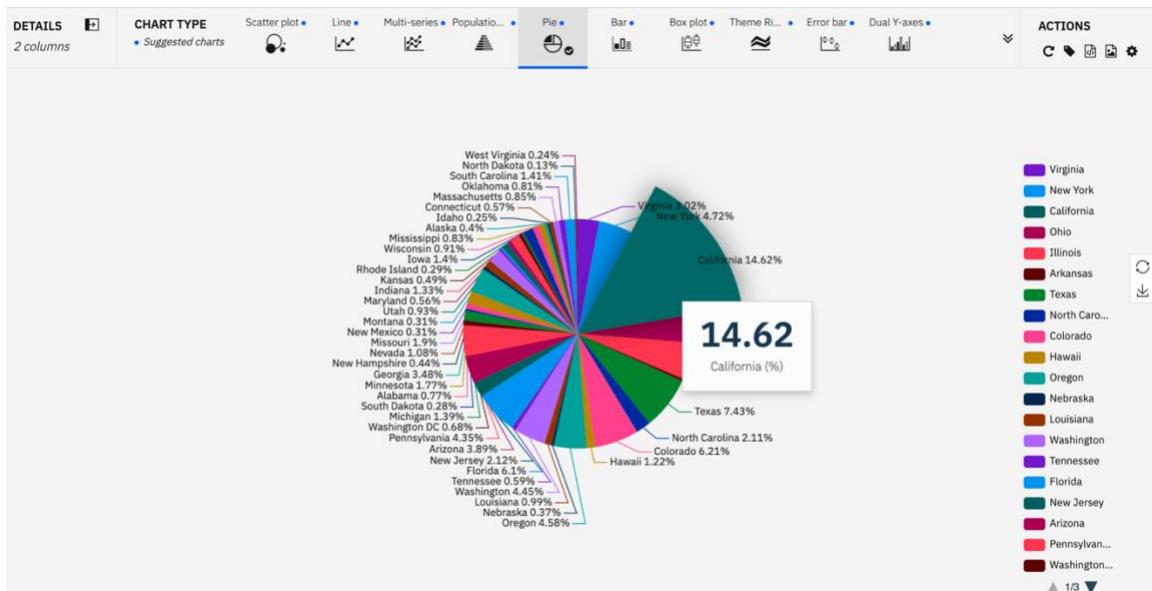


The visualization is changed to a **Pie** chart to render the result.

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.

My Projects > Auto Insurance > Auto Insurance Data Flow

+ 1 operation Code an operation to cleanse and shape your data

3 Save and create a job

Save and view jobs

Steps

CUSTOMER	NAME	COUNTRY	LATITUDE	LONGITUDE	ADDRESS	STREET_AI
1 AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dorado Street Stockton CA 95202	222 North El
2 AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dorado Street Stockton CA 95202	222 North El
3 AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dorado Street Stockton CA 95202	222 North El
4 AA71604	Janine Cockshot	US	33.599728	-111.98813	12602 N Paradise Village Pkwy Phoenix AZ 85032	12602 N Para
5 AA71604	Janine Cockshot	US	33.599728	-111.98813	12602 N Paradise Village Pkwy Phoenix AZ 85032	12602 N Para
6 AA71604	Janine Cockshot	US	33.599728	-111.98813	12602 N Paradise Village Pkwy Phoenix AZ 85032	12602 N Para
7 AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland Rd Harlingen TX 78552	1002 Dixielan
8 AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland Rd Harlingen TX 78552	1002 Dixielan
9 AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland Rd Harlingen TX 78552	1002 Dixielan
10 AB21519	Myrvyn Morriss	US	42.0106407	-87.8296865	15 S. PROSPECT AVE. Park Ridge IL 60068	15 S. PROSPE
11 AB21519	Myrvyn Morriss	US	42.0106407	-87.8296865	15 S. PROSPECT AVE. Park Ridge IL 60068	15 S. PROSPE
12 AB21519	Myrvyn Morriss	US	42.0106407	-87.8296865	15 S. PROSPECT AVE. Park Ridge IL 60068	15 S. PROSPE

SOURCE FILE: Auto Insurance Customers SAMPLE SIZE: First 1000 rows

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.

Create a job

Create a job to specify how and when to run an analytical asset. Select the analytic asset and set up a schedule or run the job immediately.

Job Name

Auto Insurance Shaped

Description (Optional)

Prepare the auto insurance data for analytics

INPUT

Auto Insurance Customers

OUTPUT

CSV

Schedule off

Associated Asset

DATA REFINERY FLOW

Auto Insurance Data Flow 6 Steps Edit

Select runtime

Default Data Refinery XS

Create and Run

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 screenshot shows the 'Runs' section of the project details page. It displays a single run entry:

Start Time	Status	Duration	Started By	Action
Jan 10, 2020, 10:39:32 AM	Starting	---	ctp	

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.

The screenshot shows the 'Runs' section of the project details page. It displays a single run entry that has completed:

Start Time	Status	Duration	Started By	Action
Jan 10, 2020, 10:39:32 AM	Completed	43 seconds	ctp	

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

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	Auto Insurance Shaped	Data Asset	ctp	10 Jan 2020, 9:16:06 am	⋮
<input type="checkbox"/>	Auto Insurance Customers	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
<input type="checkbox"/>	Auto Insurance Claims	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
<input type="checkbox"/>	Auto Insurance Policies	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
<input type="checkbox"/>	Db2 Warehouse	Connection	ctp	10 Jan 2020, 8:36:05 am	

▼ Data Refinery flows + New Data Refinery flow

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.

CUSTOMER	NAME	COUNTRY	LATITUDE	LONGITUDE	ADDRESS	STREET_ADD...	CITY	STATE
AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dor	222 North El Dorado :	Stockton	California
AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dor	222 North El Dorado :	Stockton	California
AA10041	Rosa Pays	US	37.95486261	-121.2904037	222 North El Dor	222 North El Dorado :	Stockton	California
AA71604	Janine Cockshot	US	33.599728	-111.98813	12602 N Paradis	12602 N Paradise Vill	Phoenix	Arizona
AA71604	Janine Cockshot	US	33.599728	-111.98813	12602 N Paradis	12602 N Paradise Vill	Phoenix	Arizona
AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland R	1002 Dixieland Rd	Harlingen	Texas
AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland R	1002 Dixieland Rd	Harlingen	Texas
AB13432	Tiphanie Paquet	US	26.1805	-97.7209	1002 Dixieland R	1002 Dixieland Rd	Harlingen	Texas
AB21519	Myrvyn Morris	US	42.0106407	-87.8296865	15 S. PROSPECT	15 S. PROSPECT AVE	Park Ridge	Illinois
AB21519	Myrvyn Morris	US	42.0106407	-87.8296865	15 S. PROSPECT	15 S. PROSPECT AVE	Park Ridge	Illinois

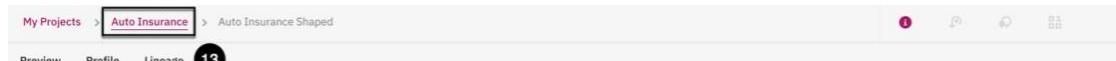
Refine X Data Asset

Description
No description available for this asset

Tags
No tags available for this asset

Added: 03:16 PM UTC, 2020/01/10
Size: 549.154 KB

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.



My Projects > **Auto Insurance** > Auto Insurance Shaped

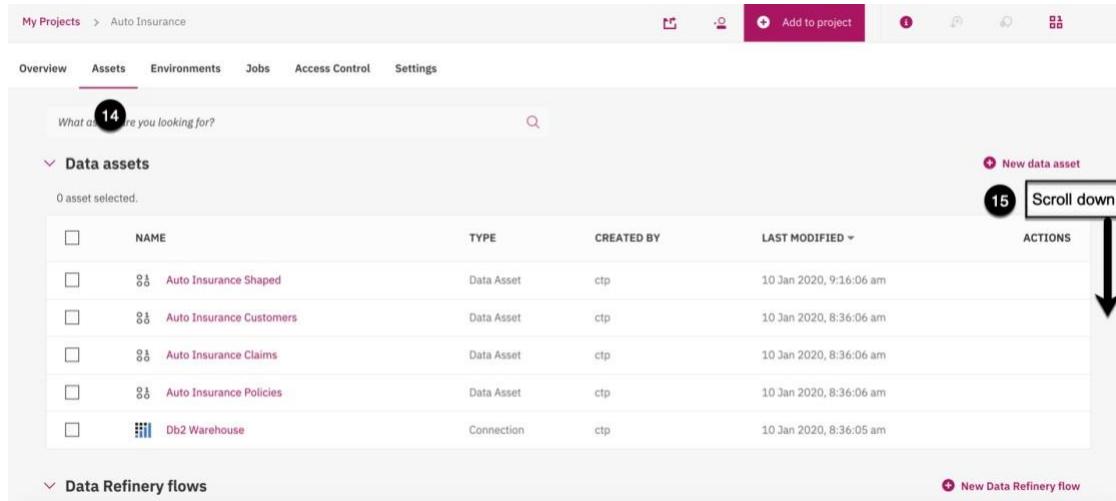
Preview Profile Lineage **13**

Schema: 54 Columns
Preview: 1000 rows Last refresh: 51 minutes ago

Refine

EMAIL_ADDR...	PHONE_NUM...	GENDER	NATIONAL_ID	EDUCATI...	EMPLOYMENT_STA...	MARITAL_STA...	CUSTOMER_LIFETIME_VA...	NUMBER_OF_POLI...
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
rpaysp8@homeste...	865-749-5448	Female	73724a8a5e731018fb3c5f958	Bachelor	Employed	Married	9421.101961	3
rpaysp8@homeste...	865-749-5448	Female	73724a8a5e731018fb3c5f958	Bachelor	Employed	Married	9421.101961	3
rpaysp8@homeste...	865-749-5448	Female	73724a8a5e731018fb3c5f958	Bachelor	Employed	Married	9421.101961	3
jcockshotqc@wikim...	808-976-1894	Female	d0642f60ceacb3fe75483cc4et	Master	Employed	Married	2802.621642	1
jcockshotqc@wikim...	808-976-1894	Female	d0642f60ceacb3fe75483cc4et	Master	Employed	Married	2802.621642	1
jcockshotqc@wikim...	808-976-1894	Female	d0642f60ceacb3fe75483cc4et	Master	Employed	Married	2802.621642	1
tpaquet54@gmpg.o...	612-256-1393	Female	47a2bc15177610791ea72cdet	Bachelor	Unemployed	Single	10628.06415	3
tpaquet54@gmpg.o...	612-256-1393	Female	47a2bc15177610791ea72cdet	Bachelor	Unemployed	Single	10628.06415	3
tpaquet54@gmpg.o...	612-256-1393	Female	47a2bc15177610791ea72cdet	Bachelor	Unemployed	Single	10628.06415	3
mmorrisbm@wordf...	203-751-1286	Male	21959d6fecccb529355a9905b	College	Employed	Married	2705.987629	1
mmorrisbm@wordf...	203-751-1286	Male	21959d6fecccb529355a9905b	College	Employed	Married	2705.987629	1

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



My Projects > Auto Insurance

Add to project

Overview Assets Environments Jobs Access Control Settings

What are you looking for?

14 Data assets

0 asset selected.

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
Auto Insurance Shaped	Data Asset	ctp	10 Jan 2020, 9:16:06 am	
Auto Insurance Customers	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
Auto Insurance Claims	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
Auto Insurance Policies	Data Asset	ctp	10 Jan 2020, 8:36:06 am	
Db2 Warehouse	Connection	ctp	10 Jan 2020, 8:36:05 am	

15 Scroll down

16 New Data Refinery flow

17 Data Refinery flows

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.

The screenshot shows a list of data assets and a data refinery flow. The 'Data Refinery flows' section contains one item:

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
Auto Insurance Data Flow	Data Refinery flow	ctp	10 Jan 2020, 9:14:39 am	⋮ 16 Clone Create job View job Remove

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.

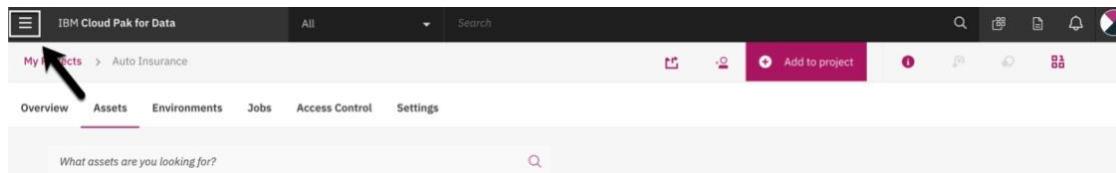
You can also click on the **data flow** to open it in the data flow shaper to modify it and view the data flow’s result set and recipe steps.

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.

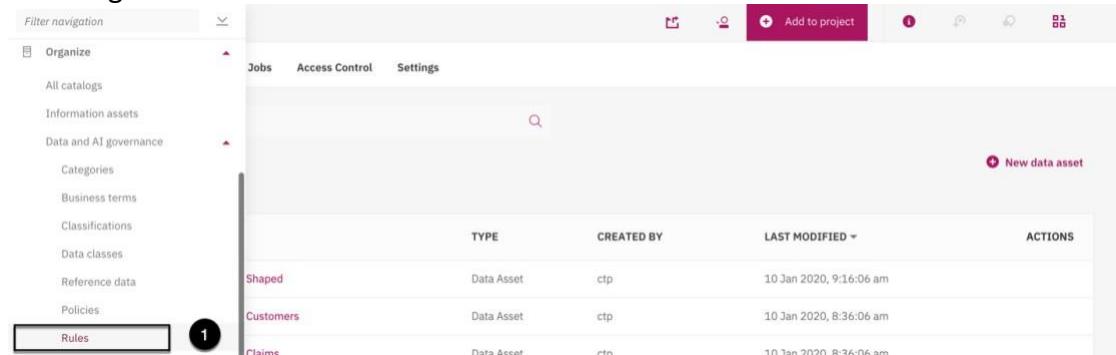
Create Data Protection Rules

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



The screenshot shows the top navigation bar of the IBM Cloud Pak for Data interface. The 'My Projects' icon is highlighted with a black arrow. The bar includes sections for 'All', 'Search', and various project management icons. Below the bar, there are tabs for 'Overview', 'Assets' (which is selected), 'Environments', 'Jobs', 'Access Control', and 'Settings'. A search bar at the bottom asks 'What assets are you looking for?'.

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



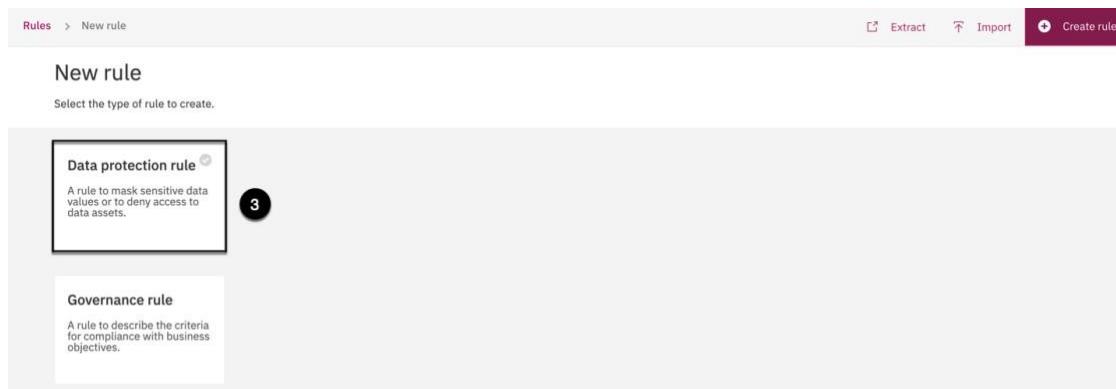
The screenshot shows the 'Organize' menu expanded under 'Data and AI governance'. The 'Rules' option is highlighted with a circled number 1. To the right, a list of existing rules is displayed in a table format with columns for TYPE, CREATED BY, LAST MODIFIED, and ACTIONS. The rules listed are 'Shaped', 'Customers', and 'Claims'.

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



The screenshot shows the 'Rules' page with a 'Published' tab selected. The 'Create rule' button is highlighted with a circled number 2. Other buttons include 'Extract', 'Import', and another 'Create rule' button.

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



The screenshot shows the 'New rule' creation page. It has a 'Rules' breadcrumb and a 'Create rule' button. The 'Data protection rule' option is highlighted with a circled number 3. A tooltip for 'Data protection rule' states: 'A rule to mask sensitive data values or to deny access to data assets.' Below it, a 'Governance rule' option is also present.

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.

The screenshot shows the 'Criteria' section of the rule configuration. It includes a condition where 'Data class' contains 'Credit Card Number' and 'Credit Card Expiration Date'. Below this, there is an 'Action' section with 'mask data' selected under 'then' and 'in columns containing' selected under 'where'. A circled number '14' is located in the bottom right corner of the action area.

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.

The screenshot shows the 'Action' section with 'mask data' selected under 'then' and 'in columns containing' selected under 'where'. A specific data class, 'Credit Card Expiration Date', is highlighted with a pink border. A circled number '17' is located above the highlighted item, and a circled number '18' is located to its right.

17. Click **Create**.

The screenshot shows the final step of creating the rule. The 'Action' section is identical to the previous screenshot, with 'Credit Card Expiration Date' selected. A circled number '21' is located in the bottom right corner of the page.

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

The screenshot shows the 'Protect Credit Card Information' rule configuration. The 'Business definition' section states: 'Protect all credit card numbers, expiration dates and validation numbers using the data redaction method.' It lists the creator as 'ctp', date created as '1/11/2020', last editor as 'ctp', and last modified as '1/11/2020'. The 'Criteria' section contains 'Condition 1': 'If Data class contains any Credit Card Number, Credit Card Expiration Date, Credit Card Validation Number'. The 'Action' section specifies: 'Then Redact data in columns containing: Credit Card Number, Credit Card Expiration Date, Credit Card Validation Number'. At the top right are 'Edit' and 'Delete' buttons.

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

The screenshot shows the 'Rules' dashboard with the 'Create rule' button highlighted. The dashboard includes tabs for 'Published' and 'Draft', search and sort functions, and buttons for 'Extract', 'Import', and 'Create rule'.

20. Click the Data protection rule type.

The screenshot shows the 'New rule' creation interface. The 'Data protection rule' option is selected and highlighted, with a description: 'A rule to mask sensitive data values or to deny access to data assets.' The 'Governance rule' option is also shown with its description: 'A rule to describe the criteria for compliance with business objectives.'

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

The screenshot shows the 'Details' and 'Rule builder' sections for creating a new data protection rule. In the 'Details' section, the 'Name*' field is filled with 'Protect US Social Security Numbers' (marked with circle 25). The 'Type*' field is set to 'Access' (marked with circle 26). In the 'Business definition*' section, the text states: 'Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values.' (marked with circle 26). In the 'Rule builder' section, 'CONDITION 1' is defined with 'If Data class contains any us social' (marked with circles 27 and 28). A tooltip for 'us social' defines it as: 'US Social Security Number. In the United States, a Social Security number (SSN) is a unique nine-digit number issued to U.S. citizens, permanent residents, and temporary (working) residents.' (marked with circle 29). The 'Action' section is set to 'US Social Security Number Last 4' (marked with circle 29).

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**.

The screenshot shows the 'New data protection rule' configuration page. In the 'Details' section, the name is 'Protect US Social Security Numbers' and the type is 'Access'. In the 'Criteria' section under 'CONDITION 1', the condition is set to 'Data class' with the value 'US Social Security Number' and the operator 'contains any'. In the 'Action' section, the 'then' clause is set to 'mask data', which is highlighted with a red box and labeled '30'.

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

The screenshot shows the 'New data protection rule' configuration page. In the 'Details' section, the name is 'Protect US Social Security Numbers' and the type is 'Access'. In the 'Criteria' section under 'CONDITION 1', the condition is set to 'Data class' with the value 'US Social Security Number' and the operator 'contains any'. In the 'Action' section, the 'then' clause is set to 'mask data' and the 'where' clause is set to 'in columns containing', which is highlighted with a red box and labeled '31'.

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

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

The screenshot shows the 'Details' section with 'Name*' set to 'Protect US Social Security Numbers' and 'Type*' set to 'Access'. In the 'Action *' section, 'then mask data' is selected under 'in columns containing' 'us social'. A callout bubble labeled 32 points to the 'us social' option. Another callout bubble labeled 33 points to the 'US Social Security Number' data class listed in the dropdown.

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

The screenshot shows the 'Details' section with 'Name*' set to 'Protect US Social Security Numbers' and 'Type*' set to 'Access'. In the 'Action *' section, 'then mask data' is selected under 'in columns containing' 'US Social Security Number X'. A callout bubble labeled 34 points to the 'Obfuscate' method, which is highlighted with a pink border. The other methods, 'Redact' and 'Substitute', are also shown with their respective before and after examples.

31. Click **Create**.

The screenshot shows the 'Details' section with 'Name*' set to 'Protect US Social Security Numbers' and 'Type*' set to 'Access'. In the 'Action *' section, 'then mask data' is selected under 'in columns containing' 'US Social Security Number X'. The 'Obfuscate' method is selected and highlighted with a pink border. The 'Create' button is visible at the bottom right, with a callout bubble labeled 35 pointing to it.

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

The screenshot shows the 'Rules' section of the IBM Cloud Pak for Data interface. A specific rule named 'Protect US Social Security Numbers' is selected. The rule's details are displayed in a card:

- Business definition:** Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values.
- Criteria:** Condition 1: If Data class contains any US Social Security Number.
- Action:** Then Obfuscate data in columns containing: US Social Security Number.

Metadata for the rule:

- Creator: ctp
- Date created: 1/11/2020
- Last editor: ctp
- Last modified: 1/11/2020

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.

The screenshot shows the user profile and settings menu. The sidebar on the right includes:

- Profile picture: ctp
- Profile and settings
- Getting Started
- About
- Community
- Support

The main content area displays the two data protection rules:

- Protect Credit Card Information** (active): Protect all credit card numbers, expiration dates and validation numbers using the data redaction method. Last modified: Jan 10, 2020.
- Protect US Social Security Numbers** (active): Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values. Last modified: Jan 11, 2020.

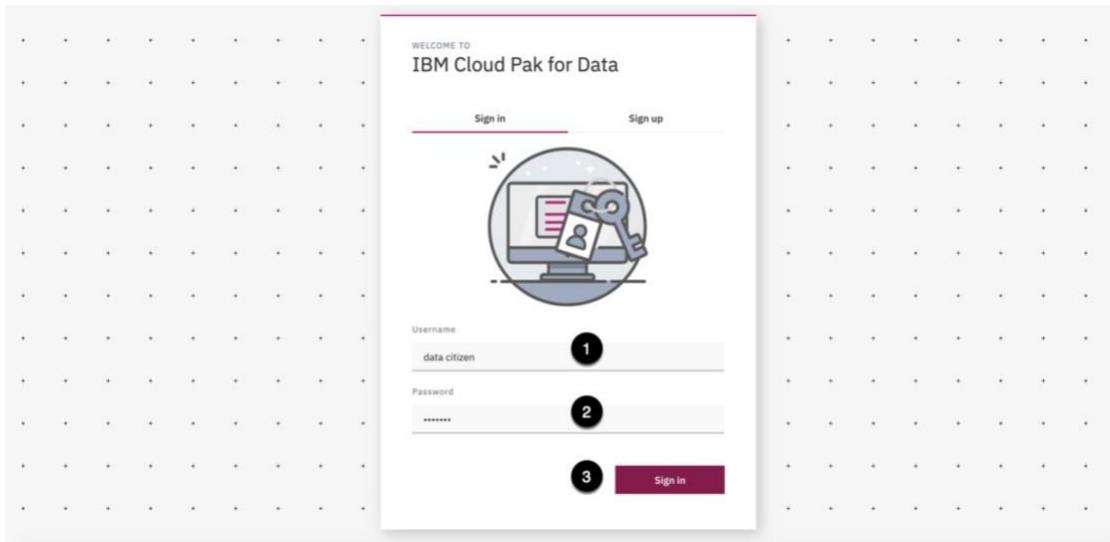
At the bottom right of the sidebar, there is a 'Log out' button.

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.

A screenshot of the IBM Cloud Pak for Data home page. The top navigation bar includes a menu icon, the title "IBM Cloud Pak for Data", a search bar with the placeholder "All", and a search input field containing "auto insurance". On the far right are icons for user profile, notifications, and help. The main content area starts with a welcome message "WELCOME, data citizen!" and a large "Let's get started!" button. Below this is a section titled "Collect and organize" with a sub-section for "IBM Cloud Pak for Data: Collect and organize". It shows an illustration of a person interacting with blue 3D cubes. To the right, there is a brief description of the service and links to "Explore catalogs" and "Explore information assets". Another tab, "Analyze", is visible above this section. At the bottom right of the main content area is a link "Go to your home page" and a checkbox "Hide on log in".

- Click on the Auto Insurance Customers data asset from the Auto Insurance catalog.

The screenshot shows the search results for 'auto insurance'. The 'Auto Insurance Customers' data asset is highlighted with a red box and a number '5' indicating it is the selected item. The table includes columns for Name, Type, Tags, Modified by, and Date modified. Other items listed include '2017 J.D. Power U.S. Auto Claims Satisfaction Study', 'Auto Insurance Policies', 'Vehicle Insurance Doc United States', 'Auto Insurance Claims', 'Db2 Warehouse', and 'Coordination of benefits priority count'.

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.

The screenshot shows the details of the 'Auto Insurance Customers' data asset. The 'Overview' tab is selected. The 'Description' section states 'All U.S. auto insurance customers'. The 'Schema' section shows 28 columns, 328 rows, and 4 columns masked. The 'Tags' section lists 'Auto Insurance'. The 'Reviews' section shows 1 review. The 'Classification' section shows 'None'. A modal window titled 'Data masking in progress' is displayed, stating 'This asset is being masked by the data enforcement rule: Protect Credit Card Information. You can wait here to see a preview of the asset or we can notify you when the preview is ready.' A 'Notify Me' button is present at the bottom of the modal.

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.

- 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.

The screenshot shows the 'Auto Insurance Customers' data asset details page. On the right, the schema table has a row labeled '4 Columns masked' with a lock icon. A modal window titled 'An error occurred attempting to preview this asset.' displays the message 'This file doesn't have any content. It might be corrupted.'

CUSTOMER	NAME	COUNTRY	LATITUDE	LONGITUDE	STREET_ADDRESS	CITY	STATE	STAT
Type: String	Type: String	Type: String	Type: Decimal	Type: Decimal	Type: String	Type: String	Type: String	Type: String

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

The screenshot shows the same data asset page after masking. Three columns in the schema table are highlighted with a red box. A modal window titled 'Data masking in progress' provides details about the masking rule and a 'Notify Me' button.

customer_id	name	months_as_customer	age	country	street_address	city	state	state_code	insured_employees	email
String	String	Smallint	Smallint	String	String	String	String	String	Integer	String

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

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

Description
All U.S. auto insurance customers

Added: Jan 09, 2020 6:20 PM...
Format: application/octet-stream
Size: 312 KB

Business Terms
There are no terms available for this asset.

Tags
Auto Insurance

Reviews
★★★★★ 1 review

Classification
None

Schema: 28 Columns 328 Rows 4 Columns masked

NATIONAL...	CREDITCARD_NU...	CREDITCARD_T...	CREDITCARD...	CREDITCARD...	EDUCATI...	EMPLOYMENT_...
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
JCB	XXXXXXXXXX	XXXXXXXXXX	Bachelor	Employed		
Discover	XXXXXXXXXX	XXXXXXXXXX	College	Unemployed		
VISA	XXXXXXXXXX	XXXXXXXXXX	Doctor	Medical Leave		
Diners Club	XXXXXXXXXX	XXXXXXXXXX	Bachelor	Disabled		
820-74-9266	XXXXXXXXXX	JCB	Master	Employed		
778-86-7182	XXXXXXXXXX	Diners Club	Master	Medical Leave		
249-23-7130	XXXXXXXXXX	JCB	XXXXXXXXXX	High School or Br	Employed	
738-54-3112	XXXXXXXXXX	Discover	XXXXXXXXXX	College	Employed	

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.

Description
All U.S. auto insurance customers

Added: Jan 09, 2020 6:20 PM...
Format: application/octet-stream
Size: 312 KB

Business Terms
There are no terms available for this asset.

Tags
Auto Insurance

Reviews
★★★★★ 1 review

Classification
None

Schema: 28 Columns 328 Rows 4 Columns masked

NATIONAL...	CREDITCARD_NU...	CREDITCARD_T...	CREDITCARD...	CREDITCARD...	EDUCATI...	EMPLOYMENT_...
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
JCB	XXXXXXXXXX	XXXXXXXXXX	Bachelor	Employed		
Discover	XXXXXXXXXX	XXXXXXXXXX	College	Unemployed		
VISA	XXXXXXXXXX	XXXXXXXXXX	Doctor	Medical Leave		
Diners Club	XXXXXXXXXX	XXXXXXXXXX	Bachelor	Disabled		
820-74-9266	XXXXXXXXXX	JCB	Master	Employed		
778-86-7182	XXXXXXXXXX	Diners Club	Master	Medical Leave		
249-23-7130	XXXXXXXXXX	JCB	XXXXXXXXXX	High School or Br	Employed	
738-54-3112	XXXXXXXXXX	Discover	XXXXXXXXXX	College	Employed	

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.