

# Lab Guide

## Hands-on-Lab

### Watson Knowledge Catalog powered by Cloud Pak for Data

Shivam Solanki

Data Scientist

[Shivam.raj.solanki@ibm.com](mailto: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 telecommunication company, Customer churn, also known as customer attrition, is a common phenomenon because of competition and lucrative offers provided to the new customers. The telecom companies want to be equipped with monitoring churn rate in order to identify strategies for improvement.

It has been observed that with real time monitoring and targeting “to be” churning customer, it is possible to improve the customer retention rates. But for real-time monitoring, the data should be made available to the machine learning models making prediction in real time.

A significant amount of time is involved in the process of collecting and curating data in order to make it consumable. The more time required in data collection, curation and discovery, the higher the risk of churn, which are a costly outcome.

Acquiring new customer is an expensive, time-consuming process and it can cost up to 25 times more than retaining existing customers.

With a data-driven approach, the information gathering process can be expedited tremendously with immediate access to relevant information at the first possibility of churn. The use of data analytics and AI can help identify potential customer churn learning and detailed data analysis.

Using information that's available in the telecom company's enterprise Knowledge Catalog, the business can easily develop a data-driven churn analytics 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 telecom company's goal is to create a machine learning model to predict customer churn and understand the factors leading to customer. Using IBM Cloud Pak for Data, the business can easily prepare a machine learning model to assess the factors and pain points leading customer to churn. The app also needs the customer's account, billing and product information to understand the factor and build the machine learning classification model.

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 telecom 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.

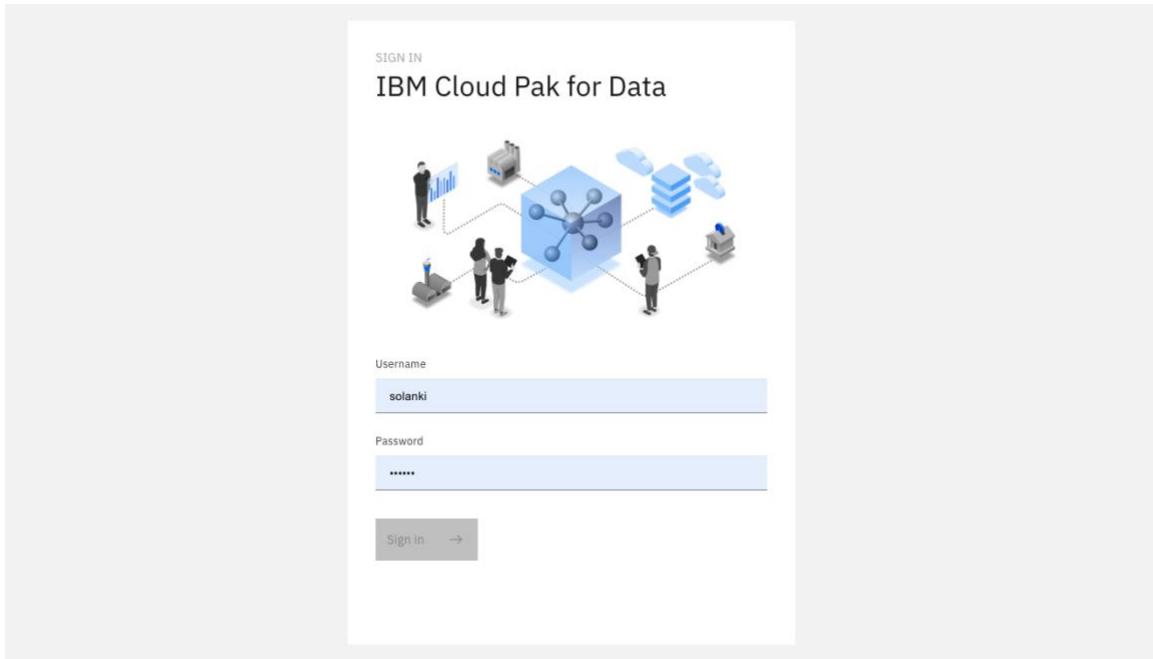
## Create an IBM Cloud Pak for Data User

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.

The image shows the IBM Cloud Pak for Data welcome screen. It features a dark-themed interface with a central banner that says "Welcome, Shivam Solanki (IBM)!" and "Use the following links to get up and running." Below the banner are several quick navigation links: "Create a service instance", "Set up an LDAP server", "Create an analytics project", "Explore the services catalog", "Create catalogs", "Explore business terms", "My instances", "My data", "Data virtualization", and "Catalogs". On the left, there's a sidebar with sections for "Quick navigation", "Resources", and "Community". A central modal window titled "Got data?" contains the message: "Your enterprise has data. Lots of data. You need to use your data to generate meaningful insights that can help you avoid problems and reach your goals. But your data is useless if you can't trust it or access it." It includes "Close" and "Next" buttons. To the right, there are sections for "Recent" (with a "No recent activity" message), "Notifications" (with a "No notifications" message), and "Pending publish to catalog requests" (showing 3 pending requests). At the bottom, there are sections for "My instances" (with a "No provisioned instances" message) and "Model Monitoring".

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 **Telco Churn** 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. On the left, there is a navigation sidebar with various options like Home, Projects, Connections, My instances, Collect, Organize, Analyze, Administer, and others. The 'Organize' section is expanded, and the 'All catalogs' option is highlighted with a red box. The main area is titled 'Welcome to IBM Cloud Pak for Data!' and shows sections for Overview, Recent projects, Requests, and Notifications. The 'Recent projects' section indicates 'No recent projects'. The 'Requests' section shows 'Data requests: 0' and 'Completed publish to catalog requests: 0'. The 'Notifications' section indicates 'No notifications'.

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

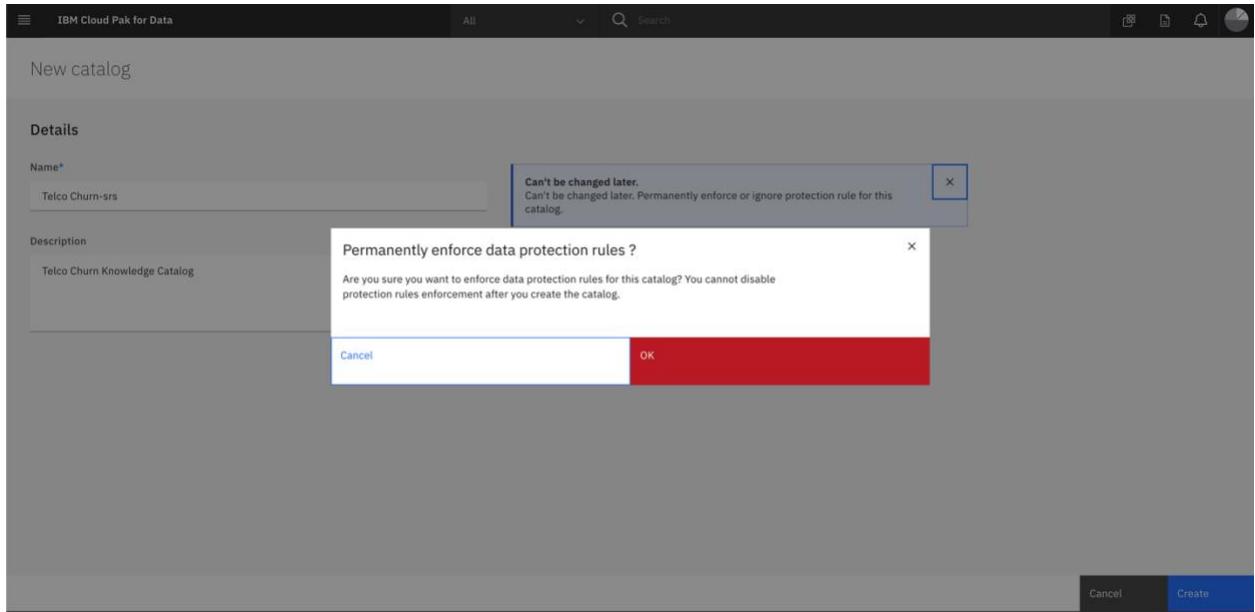
The screenshot shows a modal dialog box titled 'Your catalogs'. At the top right is a blue 'New Catalog' button. The main area contains a 'Get started' section with an info icon and the text 'Create a catalog to get started. A catalog is where you organize your assets (for example, your data, data connections, and analysis) and collaborators.' Below this is a 'Create Catalog' button.

4. Enter a Name of Telco Churn-your initials.

5. Enter a Description of Telco Churn 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 Telco Churn-initials** notification during catalog creation.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below it, the breadcrumb path shows 'Catalogs / Telco Churn-srs'. On the right side of the header, there are icons for 'Add to Catalog', a plus sign, and a refresh symbol. The main content area is titled 'Telco Churn-srs'. It has three tabs at the top: 'Browse Assets' (underlined), 'Access Control' (which is highlighted with a red box), and 'Settings'. A message box in the center says 'Now you can add assets!' with a link to get started. At the bottom of the page, there are footer links for 'Data Protection', 'Data Governance', 'Data Integration', and 'Data Quality'.

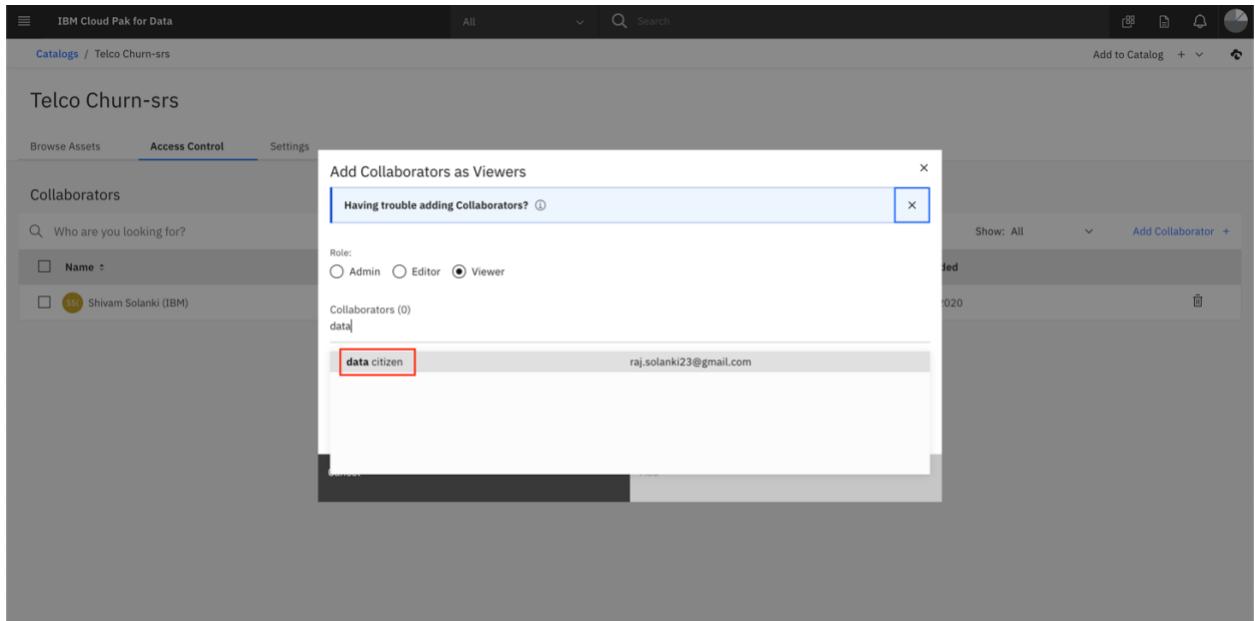
Once completed, you are brought into the newly created catalog. You will now add the **data citizen** user to the catalog as a *Viewer* so they can access the new catalog and use the data assets. You will log in as this user at the end of the lab to see how data protection rules are enforced.

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

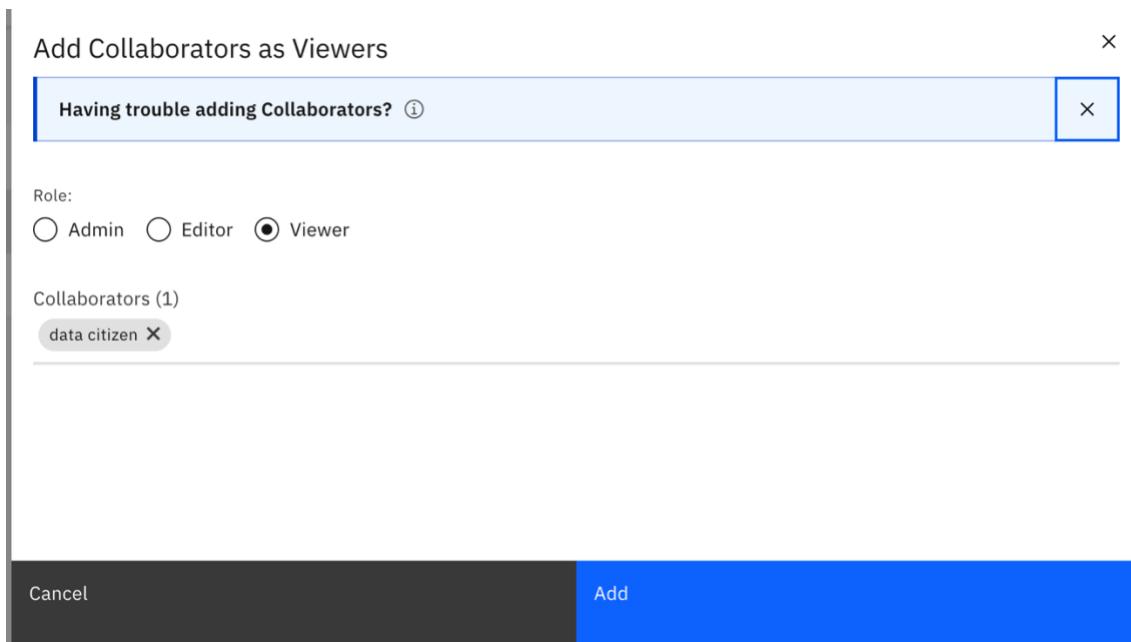
The screenshot shows the 'Access Control' tab of the catalog. At the top, there's a search bar and a 'Show: All' dropdown. On the right, there's a red box around the 'Add Collaborator' button. The main area is titled 'Collaborators' and contains a table with one row. The table columns are 'Name', 'Email', 'Role', and 'Date Added'. The single entry is 'Shivam Solanki (IBM)' with email 'Shivam.Raj.Solanki@ibm.com', role 'Admin', and date 'Sep 18, 2020'. There's also a trash icon next to the entry.

Name	Email	Role	Date Added
Shivam Solanki (IBM)	Shivam.Raj.Solanki@ibm.com	Admin	Sep 18, 2020

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



11. Type the word **data** in the search area.
12. Click on the **data citizen** user. The default role of Viewer is automatically assigned. Leave the role set to Viewer.



13. Click **Add**.
- You should now see the **data citizen** user added as a **Viewer**.

Catalogs / Telco Churn-srs

Telco Churn-srs

Browse Assets Access Control Settings

Collaborators

Who are you looking for? Show: All Add Collaborator

Name	Email	Role	Date Added
Shivam Solanki (IBM)	Shivam.Raj.Solanki@ibm.com	Admin	Sep 18, 2020
data citizen	raj.solanki23@gmail.com	Viewer	Sep 18, 2020

## Create Analytic Projects

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

The first project, which will be named **Telco Churn**, 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 Telco Churn 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 **Telco 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 Telco Churn 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 interface. On the left, a sidebar menu is open with the 'Projects' option highlighted by a red box. The main area displays a message: 'You don't have any projects right now. Create your first project now.' Below this message is a blue button labeled 'New project +'. The top navigation bar includes a search bar and filter options for 'Project type', 'User role', and 'Last updated'.

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

This screenshot shows the same interface as above, but the 'New project +' button is now highlighted with a red box. The rest of the interface remains the same, including the sidebar menu and the central message about no projects.

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

A modal dialog box titled 'Create a new project' is displayed. It contains a sub-instruction 'Select a project type' and two radio buttons: 'Analytics project' (which is selected and highlighted with a red circle) and 'Data quality project'. At the bottom of the dialog are two buttons: 'Cancel' and 'Next'. The 'Next' button is highlighted with a blue box.

5. Click the **OK** button.

6. Click on Create an empty project.

The screenshot shows the 'Create a project' page. It has two main sections: 'Create an empty project' and 'Create a project from a file'.  
The 'Create an empty project' section includes:

- A circular icon with a blue background and a white outline, containing a simplified diagram of a data flow or model structure.
- The title 'Create an empty project'.
- A brief description: '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.'
- A 'USE TO' section with three items: 'Prepare and visualize data', 'Analyze data in notebooks', and 'Train models'.

The 'Create a project from a file' section includes:

- A circular icon with a blue background and a white outline, containing a simplified diagram of a document being loaded into a system.
- The title 'Create a project from a file'.
- A brief description: 'Get started fast by loading existing assets. Choose a project file from your system or a Git repository.'
- A 'USE TO' section with three items: 'Learn by example', 'Build on existing work', 'Run tutorials', and 'Integrate with Git'.

7. Enter a Name of Telco Churn-initials and a Description of Telco Churn project for WKC.  
Click the Create button.

The screenshot shows the 'New project' dialog. It has two main sections: 'Define project details' and 'Choose project options'.  
The 'Define project details' section includes:

- A 'Name' input field containing 'Telco Churn - ss'.
- A 'Description' input field containing 'Telco Churn project for WKC'.

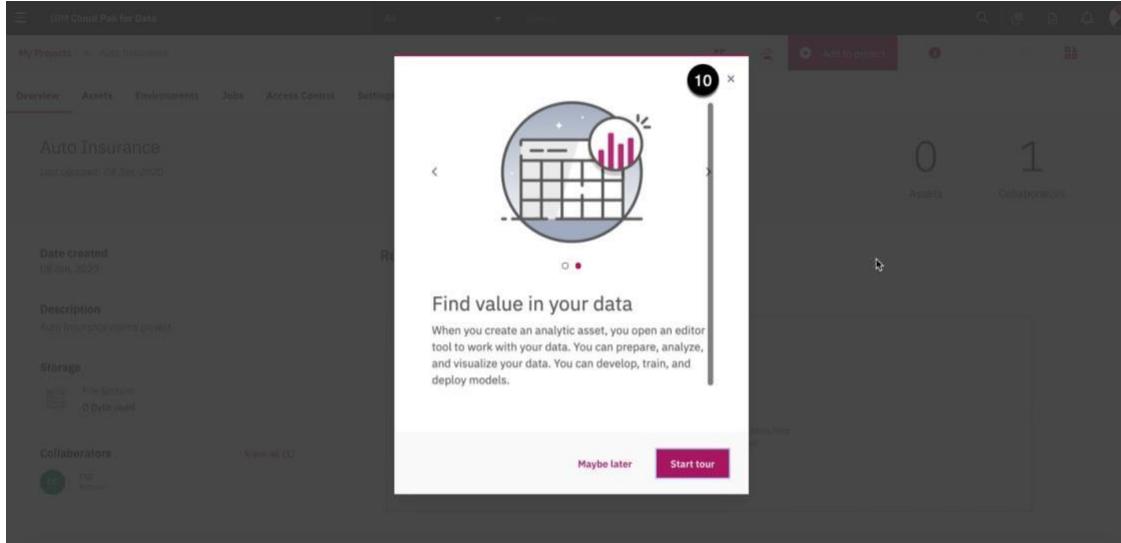
The 'Choose project options' section includes:

- A checkbox labeled 'Integrate this project with Git'.

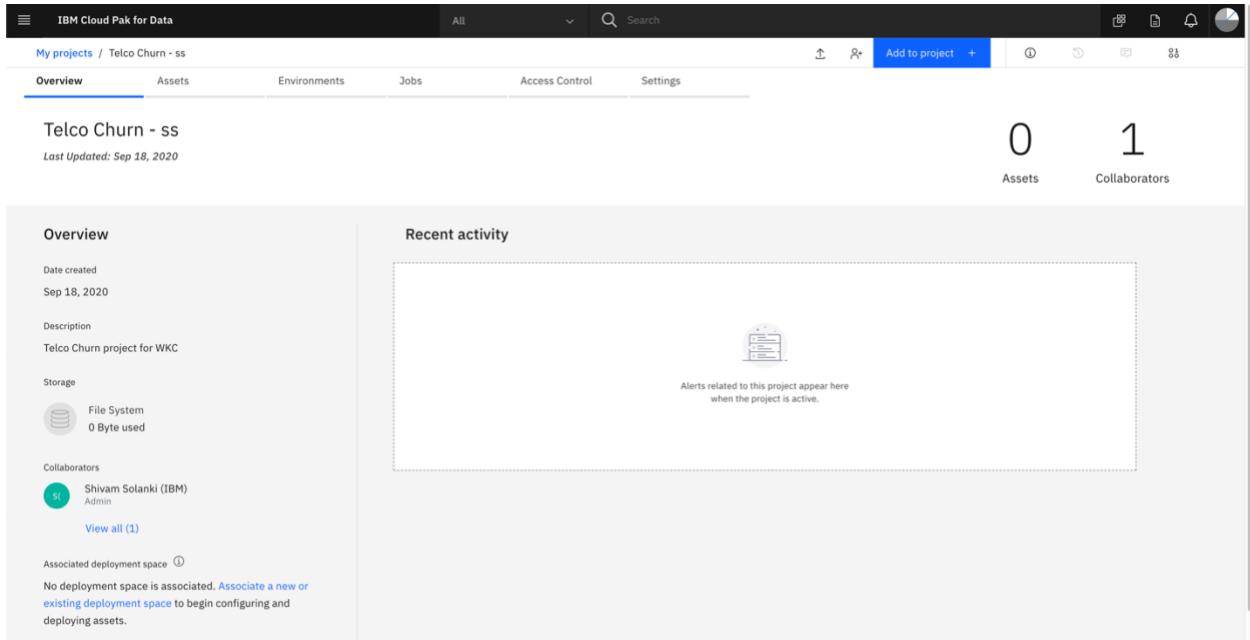
At the bottom right, there are 'Cancel' and 'Create' buttons. The 'Create' button is highlighted with a red box.

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

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



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



## Create the Telco 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 IBM Cloud Pak for Data interface. On the left, there is a navigation sidebar with the following items:

- Home
- Projects** (highlighted with a red box)
- Connections
- My instances
- Collect
- Organize
  - All catalogs
  - Information assets
  - Data and AI governance
  - Curation
  - Data quality
  - Governance workflows
  - Management
- Analyze

The main content area has columns for **Project type**, **User role**, and **Last updated**. A central message says "You don't have any projects right now. Create your first project now." with a "New project +".

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

The screenshot shows the same interface as above, but the "New project +" button in the center is highlighted with a red box.

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

A modal dialog box titled "Create a new project" is shown. It contains the instruction "Select a project type" and two radio buttons:

- Analytics project
- Data quality project

At the bottom, there are "Cancel" and "Next" buttons, with "Next" highlighted with a blue box.

5. Click the **Next** button.

6. Click on Create an empty project.

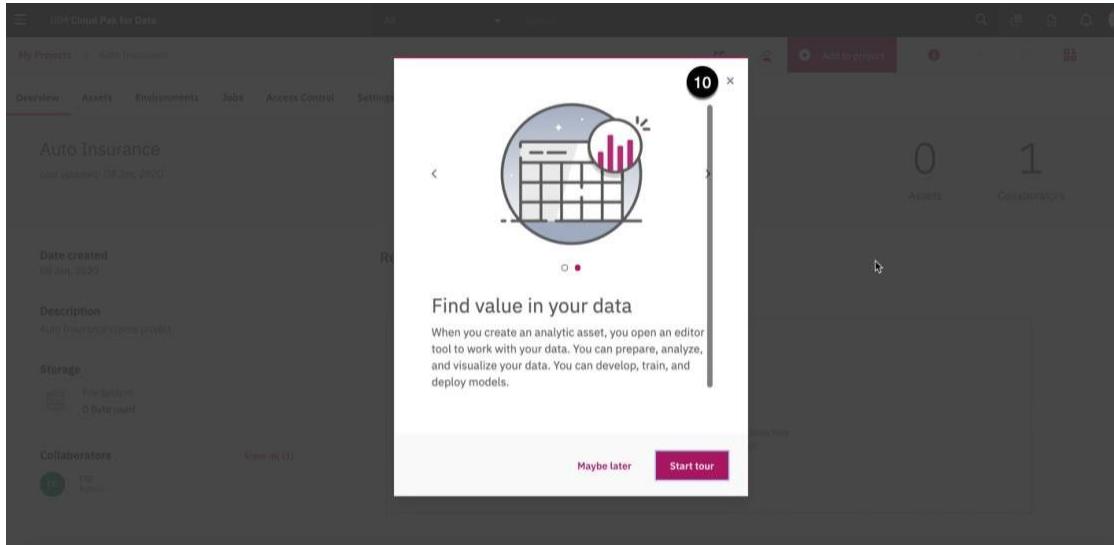
The screenshot shows the 'Create a project' page. It has two main sections: 'Create an empty project' and 'Create a project from a file'.  
**Create an empty project:**  
Icon: A circular icon showing a simplified architectural diagram with rooms and windows.  
Title: Create an empty project  
Description: 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.  
**USE TO:**  
Prepare and visualize data  
Analyze data in notebooks  
Train models  
**Create a project from a file:**  
Icon: A circular icon showing a hand holding a document with a plus sign on it.  
Title: Create a project from a file  
Description: Get started fast by loading existing assets. Choose a project file from your system or a Git repository.  
**USE TO:**  
Learn by example  
Build on existing work  
Run tutorials  
Integrate with Git

7. Enter a Name of Telco Discovery-initials, and a Description of Telco Discovery project for WKC. Click the Create button.

The screenshot shows the 'New project' creation form. It has two main sections: 'Define project details' and 'Choose project options'.  
**Define project details:**  
Name: Telco Discovery-ss  
Description: Telco Discovery project for WKC  
**Choose project options:**  
 Integrate this project with Git ⓘ  
  
At the bottom right, there are 'Cancel' and 'Create' buttons, with 'Create' being highlighted with a red box.

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.

A screenshot of the IBM Cloud Pak for Data interface showing the 'Telco Discovery-ss' project. The 'Overview' tab is selected. The project details include: Date created (Sep 18, 2020), Description (Telco Discovery project for WKC), Storage (File System, 0 Byte used), and Collaborators (Shivam Solanki (IBM) Admin). The 'Recent activity' section is empty, showing a placeholder message: 'Alerts related to this project appear here when the project is active.' Navigation icons are visible at the top right.

## 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.

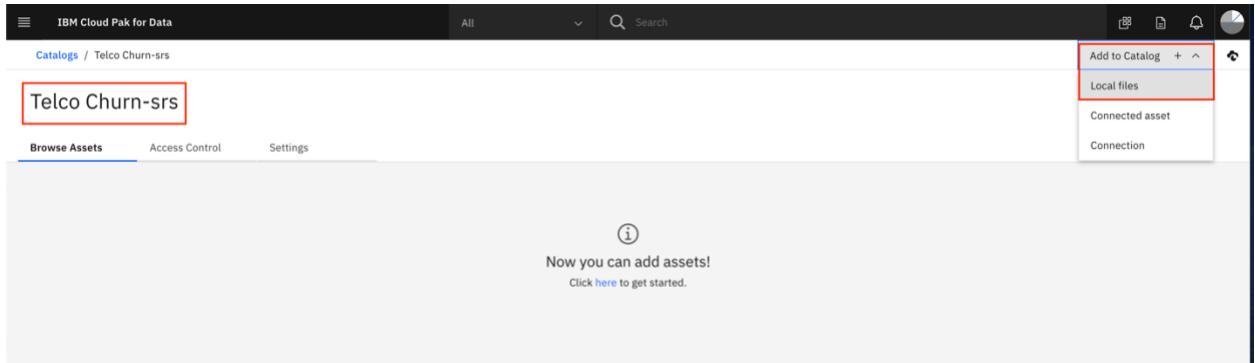
The screenshot shows the 'Telco Discovery' section of the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with tabs for Overview, Assets, Environments, Jobs, Access Control, and Settings. The 'Overview' tab is currently selected. Below the navigation bar, the title 'Telco Discovery' is displayed, along with a timestamp 'Last Updated: Aug 28, 2020'. On the right side, there's a large '0' indicating no assets found, and a 'Assets' button. The main area is currently empty.

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

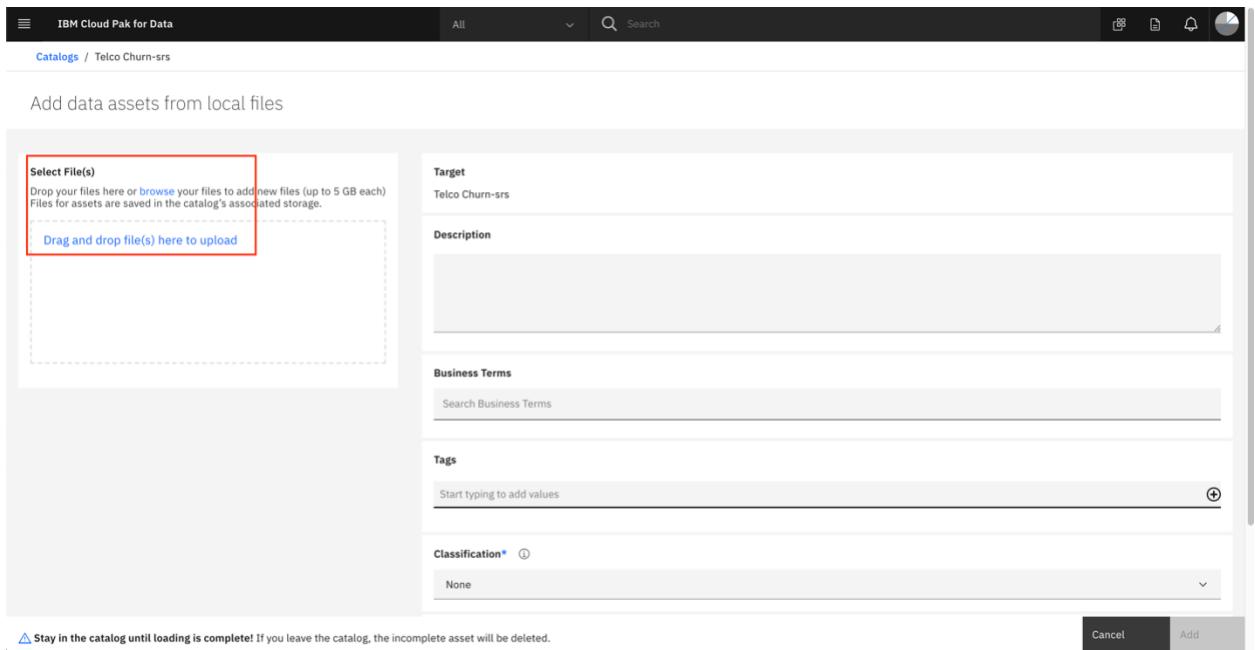
The screenshot shows the 'Organize' section of the IBM Cloud Pak for Data interface. On the left, there's a sidebar with a tree view of categories: Home, Projects, Connections, My Instances, Collect, Organize, and Analyze. The 'Organize' category is expanded, and its 'All catalogs' sub-menu item is highlighted with a red box. The main content area is titled 'Recent activity' and contains a message: 'Alerts related to this project appear here when the project is active.' There's also a small icon of a document with a checkmark.

3. Click the **Telco Churn** catalog.

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

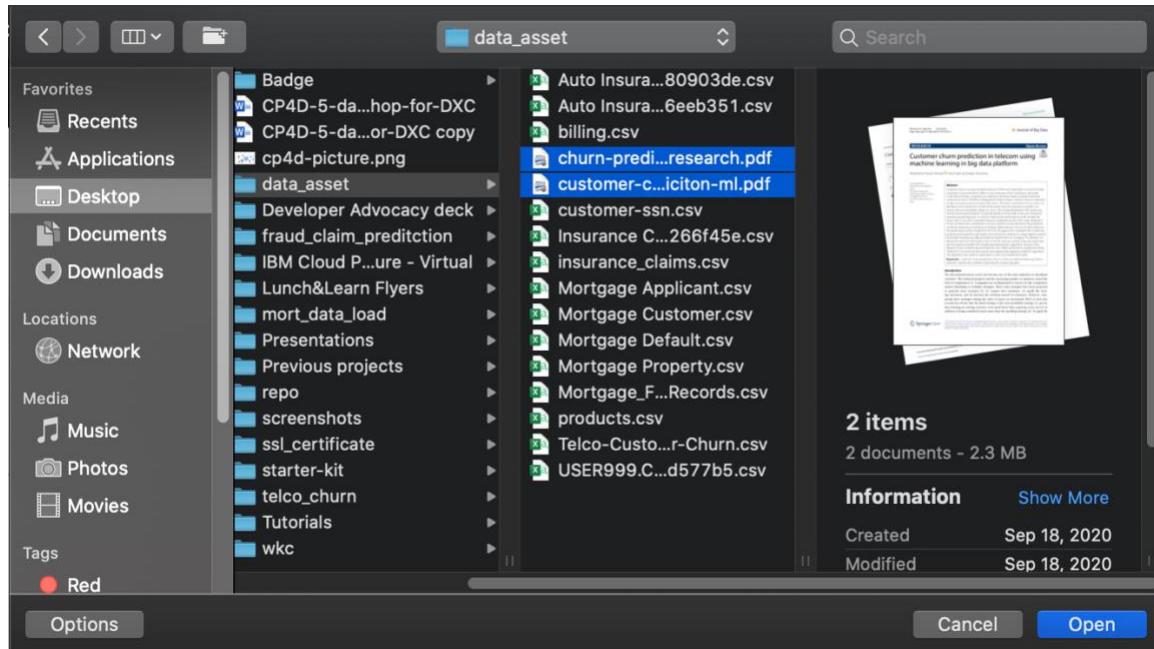


5. Click the **browse** link in the **Select File(s)** section to bring up the file selection dialog.



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

Locate the “**churn-prediction-research.pdf**” and the “**customer-churn-predictiton-ml.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.

**Business 1**

Selected Files (2)*		Edit name and format
Asset Name	Format	
churn-prediction-rese...	PDF	X
customer-churn-predi...	PDF	X

**Search B**

**Tags**

**Start typi**

**Classical**

**None**

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 “**churn-prediction-research.pdf**” file. Go to the end of the filename and remove the **.pdf** extension.

Asset Name	Format
churn-predictic...	application/pdf
-prediciton-ml	application/pdf

**Business Terms**

Search Business Terms

**Tags**

Start typing to add values

**Classification\***

None

Cancel      **Apply**

10. Click in the **Asset Name** area of the “**customer-churn-prediciton-ml.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 / Telco Churn-srs

Add data assets from local files

**Select File(s)**  
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

**Selected Files (2)\***

Asset Name	Format
churn-prediction-rese...	PDF
customer-churn-predi...	PDF

**Target**  
Telco Churn-srs

**Description**  
Telco Churn document

**Business Terms**

Search Business Terms

**Tags**

Telco Churn X

Start typing to add values

**Classification\***

None

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 **Telco Churn document**.

13. Enter a Tag of **Telco Churn** 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 Knowledge Catalog interface. On the left, there is a list of 'Selected Files (2)\*' with two entries: 'churn-prediction-rese...' and 'customer-churn-predi...'. An 'Edit name and format' button is located above this list. On the right, there are three sections: 'Business Terms' (with a search bar), 'Tags' (containing 'Telco Churn X' and 'Document'), and 'Classification\*' (with a dropdown menu). The 'Document' tag is highlighted with a red box.

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

The screenshot shows the Knowledge Catalog interface. The 'Tags' section now includes both 'Telco Churn X' and 'Document X', with 'Document' highlighted by a red box. The 'Classification\*' section is also visible. A message at the bottom left says 'Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' The 'Add' button at the bottom right is highlighted with a red box.

The screenshot shows the **Telco Churn** 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 **Telco Churn** 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.

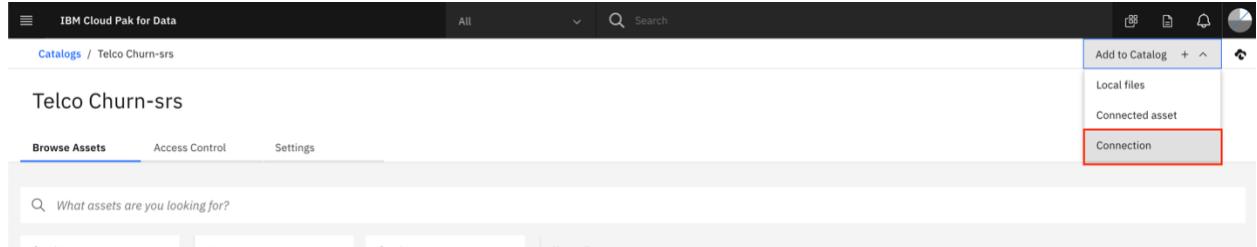
The screenshot shows the IBM Cloud Pak for Data Catalogs interface. The top navigation bar includes 'IBM Cloud Pak for Data', 'Catalogs / Telco Churn-srs', 'Telco Churn-srs', 'Add to Catalog', and various icons. Below the navigation is a search bar and a 'Browse Assets' tab. The main area displays a search interface with filters for 'Any type' (selected), 'Any source' (selected), and 'Any tag' (highlighted with a red box). The 'Any tag' dropdown shows 'Document' and 'Telco Churn'. The search results are titled 'Recently' and show two entries: 'customer-churn-predic...' and 'churn-prediction-research'. Both entries have details like owner (Shivam Solanki (IBM)), date added (Sep 18, 2020 5:18 PM), and tags (Telco Churn, Document). Below the results is a table with columns: Name, Owner, Tags, Business Terms, Type, and Date Added. Two rows are listed: 'churn-prediction-research' and 'customer-churn-predictiton-ml', both categorized as Data assets and added on Sep 18, 2020.

Name	Owner	Tags	Business Terms	Type	Date Added
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn, Document		Data asset	Sep 18, 2020
customer-churn-predictiton-ml	Shivam Solanki (IBM)	Telco Churn, Document		Data asset	Sep 18, 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 **Document** and **Telco Churn** tags have been added to the **Tags** filter area.

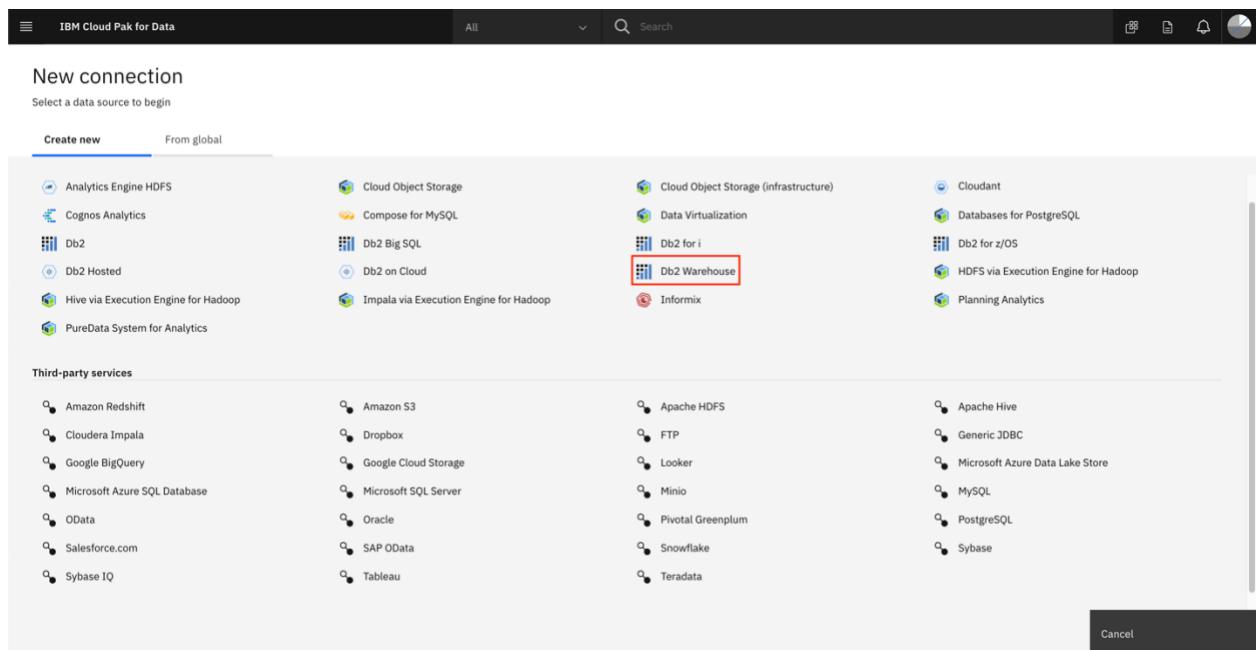
## Auto Discover Data Assets

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



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

Click on the **Db2 Warehouse** connector.



3. Enter the following parameters:

Name: **Db2-warehouse-churn-initials**

Description: **Knowledge Catalog Tutorial Db2 warehouse**

Database: **BLUDB**

Username: **bluadmin**

Hostname or IP Address field: **db2w-jtveemh.us-south.db2w.cloud.ibm.com**

Password: **GxdualWj\_f4aZq82Ah@4@KqmHAKEF**

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

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

The screenshot shows the 'New connection' configuration page in the IBM Cloud Pak for Data interface. The 'Connection overview' section includes fields for 'Name' (Db2-churn-srs) and 'Description' (Knowledge Catalog Tutorial Db2). The 'Connection Details' section includes fields for 'Database' (BLUDB), 'Port (optional)' (50001), 'Host name or IP Address' (dashdb-txn-flex-yp-dal09-168.services.dal.bluemix.net), and 'Port is SSL-enabled (optional)' (checkbox checked, note: [Deprecated] The port is configured to accept SSL connections). The 'Connection discovery' section includes a checked checkbox for 'Discover data assets'. The 'Project for discovered assets' dropdown is set to 'Telco Discovery-ss'. The 'Credentials' section includes fields for 'User name' (bluadmin) and 'Password' (redacted). The 'API Key' section is empty.

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
Db2-warehouse-churn-srs	Shivam Solanki (IBM)			Connection	Sep 18, 2020
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn Document		Data asset	Sep 18, 2020
customer-churn-predictiton...	Shivam Solanki (IBM)	Telco Churn Document		Data asset	Sep 18, 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.

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

The screenshot shows the 'Overview' tab for a connection named 'Db2-warehouse-churn-srs'. The 'Tags' section is highlighted with a red box. It contains the text 'There are no tags available for this asset.' Below the 'Tags' section, there is a 'Reviews' section with a star rating of 0 reviews. The 'Connection' section shows 'Source' as 'Db2 Warehouse' and 'type:' as 'Db2 Warehouse'. The 'Classification' section shows 'None'.

- Click in the **Tags** section and select the **Telco Churn** tag from the list of tags.

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

The screenshot shows the 'Overview' tab for the same connection. The 'Tags' section now displays a list of tags, with 'Telco Churn' selected and highlighted with a red box. The input field 'Start typing to add values' is also highlighted with a red box. The other sections remain the same as in the previous screenshot.

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

CONNECTION

## Db2-warehouse-churn-srs

**Overview**    Access    Review

**Description**  
Knowledge Catalog Tutorial Db2 warehouse

**Added:** Sep 18, 2020 5:43 PM

**Business terms**  
There are no terms available for this asset.

**Tags**  
Telco Churn X Warehouse X  
Start typing to add values +  
Cancel   **Apply**

**Reviews**  
☆☆☆☆☆ 0 reviews

**Connection**  
Source: Db2 Warehouse  
type:

(i) Connection Preview  
Click [here](#) to edit connection.

12. The discovery process has been running as a service in the background, discovering and populating data assets into the **Telco 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 **Telco Discovery** project.

Notice the number of data assets that were discovered. This discovery, when it was run, auto discovered 211 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
CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 05:43 PM
CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 05:43 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 telco 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 telco customers, you can preview the first 1000 records of the asset.

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

customer...	gender	SeniorCit...	Partner	Depende...	tenure	NATIONA...	CREDITCARD_NU...	CREDITCARD...	CREDITCARD...
8947-YRTDV	Male	0	Yes	Yes	32	120-22-7181	6011660000000870	Discover	20-Aug
3161-ONRWK	Male	0	Yes	Yes	60	582-27-7752	370000064279716	American Express	18-Oct
0114-RSRRW	Female	0	Yes	No	10	706-03-2182	4111100000003500	VISA	22-Mar
4565-NLZBV	Female	0	Yes	No	71	215-02-8472	4111950000004810	VISA	22-Jul
0031-PVLZI	Female	0	Yes	Yes	4	385-65-7820	4111160000001880	VISA	20-Sep
7206-GZCDC	Female	1	No	No	1	521-13-0320	30000004315747	Diners Club	23-Apr
6682-VCIXC	Female	0	Yes	Yes	43	747-09-7762	370000056600192	American Express	19-Sep
4791-QRGMF	Male	0	Yes	No	59	496-64-0786	370000018025389	American Express	18-Jun
6475-VHUIZ	Female	0	Yes	No	23	581-91-7471	3088080000000820	JCB	19-Jul
3910-MRQOY	Female	0	Yes	No	72	438-10-2387	4111420000002020	VISA	19-Nov
0661-WCQNQ	Male	0	Yes	No	22	409-67-9193	30000000237960	Diners Club	20-Apr
7537-RBWEA	Female	0	No	No	1	289-93-9530	370000075016917	American Express	23-Dec
4656-CAURT	Male	0	No	No	69	182-36-3850	30000003010331	Diners Club	22-Mar
0121-SNYRK	Male	0	No	No	50	650-70-0577	3088370000001050	JCB	19-Dec
1768-ZAIFU	Female	1	No	No	1	842-77-3741	4111190000002510	VISA	21-Jun
4671-LXRDQ	Male	0	No	No	2	861-80-1036	601163000001680	Discover	20-Oct
3733-LSYCE	Female	0	Yes	No	15	510-69-9056	30000001784374	Diners Club	22-Sep

The information panel on the right shows a tag of **TELCO**. This is the schema in the Db2 Warehouse instance that the table came from. Also, if you look at the columns, you will see that there are columns related to telco like tenure etc. This is the telco customers data asset that's needed.

The next several steps will demonstrate how you would publish this data asset to the **Telco Churn** catalog. However, **you will not** publish it from the project. You will catalog it from the **Telco Churn** 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.

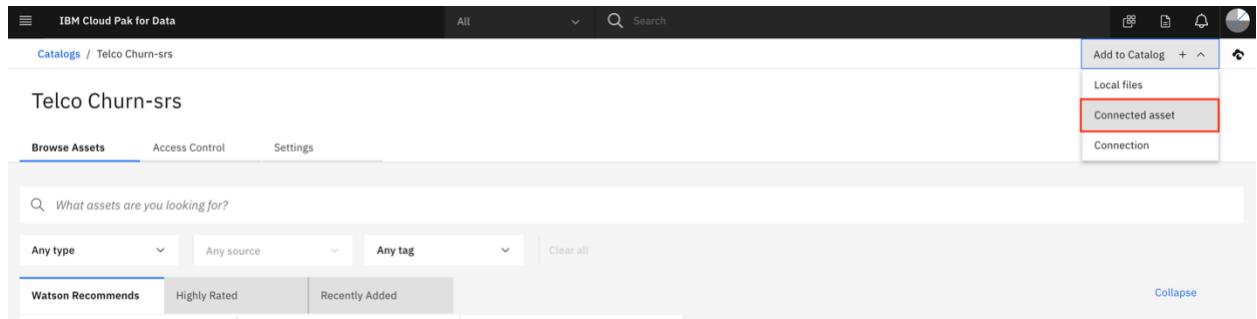
The screenshot shows the 'Assets' tab in the IBM Cloud Pak for Data interface. A search bar at the top contains the text 'custom'. Below it, a section titled 'Data assets' shows a table with two rows. The first row has a checkbox next to 'Name' which is unchecked. The second row has a checked checkbox next to 'Name' and is highlighted with a red box. The table columns are 'Name', 'Type', 'Created by', and 'Last modified'. The 'Name' column for both rows shows 'CUSTOMERS'. The 'Type' column shows 'Data Asset'. The 'Created by' column shows 'Shivam Solanki (IBM)'. The 'Last modified' column shows 'Sep 18, 2020, 05:43 PM'. At the top right of the table area, there are 'Publish' and 'Remove' buttons, with 'Publish' also highlighted with a red box.

18. Click the **Telco Discovery** link at the top of the page to go back to the Telco 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 **Telco Churn** 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; **Billing**, **Customers** and **Products** using the **Connected asset** catalog method. These tables are needed for the auto insurance claims analysis processing.

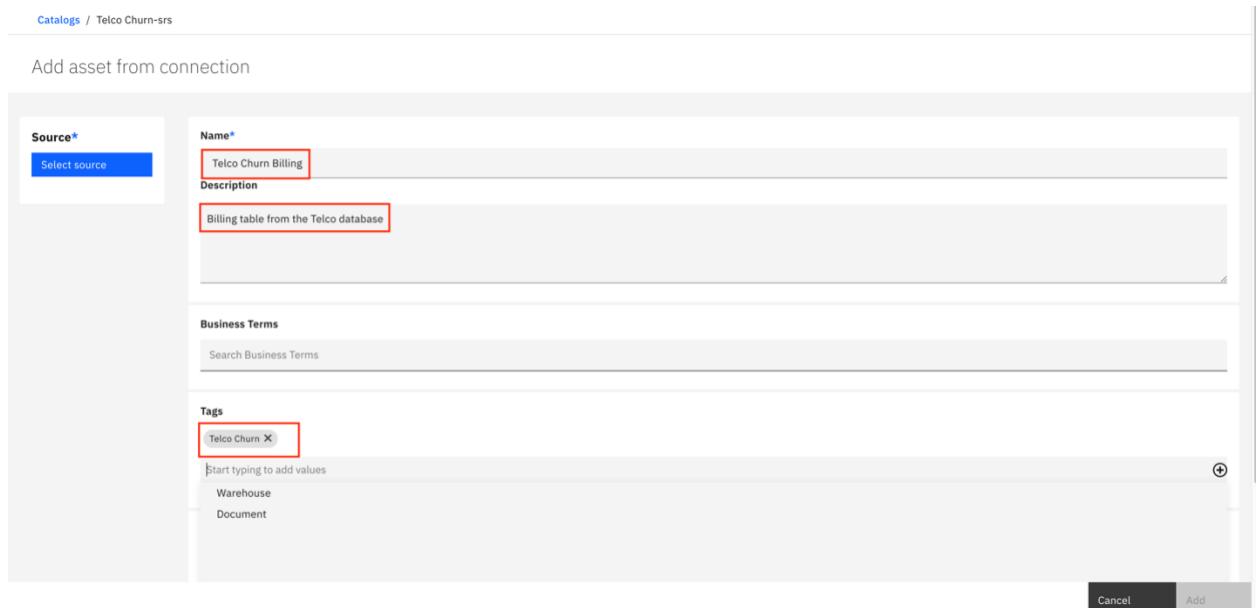
1. Click **Add to Catalog** → **Connected asset** from the Catalog menu.



The screenshot shows the IBM Cloud Pak for Data Catalog interface. The top navigation bar has 'Catalogs / Telco Churn-srs'. On the right, there's a sidebar with 'Add to Catalog' followed by a plus sign and a dropdown menu. The 'Connected asset' option is highlighted with a red box. Below the sidebar, there's a search bar and some filters like 'Any type', 'Any source', and 'Any tag'. The main area shows 'Watson Recommends', 'Highly Rated', and 'Recently Added' sections. A 'Collapse' button is at the bottom right.

2. Enter a Name of **Telco Churn Billing** and a Description of **Billing table from the Telco database**. Click in the **Tags** area and select the **Telco Churn** 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' form. It has a 'Source\*' section with a 'Select source' button. The 'Name\*' section has a field with 'Telco Churn Billing' and a 'Description' field with 'Billing table from the Telco database', both highlighted with red boxes. The 'Business Terms' section has a 'Search Business Terms' input. The 'Tags' section has a list with 'Telco Churn X' selected, and other options like 'Warehouse' and 'Document'. At the bottom right are 'Cancel' and 'Add' buttons.

3. Click the **Select Source** button to choose a Connection to add connected assets from.
4. Select the **Db2-warehouse** connection → **TELCO** schema → **BILLING** table. Click the **Select** button.

The screenshot shows the 'Catalogs' interface with the path 'Catalogs / Telco Churn-srs'. On the left, there's a tree view under 'Connections' with 'Db2-warehouse-churn-srs' selected. This node has 'Schemas (8)' listed below it, including 'AUDIT', 'DB2INST1', 'IBM\_SAILFISH', 'INSURANCE', 'NULLIDR1', 'NULLIDRA', 'QASN', and 'TELCO'. The 'TELCO' node is highlighted with a red box. Under 'TELCO', there are 'Tables (3)' listed: 'BILLING', 'CUSTOMERS', and 'PRODUCTS'. The 'BILLING' node is also highlighted with a red box. At the bottom right of the interface are two buttons: 'Cancel' and 'Select', with 'Select' being 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.
  7. Enter a Name of **Telco Churn Customers** and a Description of **Customer table from the Telco database**. Click in the **Tags** area and select the **Telco Churn** 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 → **TELCO** 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 **Telco Churn Products** and a Description of **Products table from the Telco database**. Click in the **Tags** area and select the **Telco Churn** 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 → **TELCO** schema → **PRODUCTS** table. Click the **Select** button.

15. Click the **Add** button.

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

Item Type	Item Name	Owner	Added	Tags	Reviews
Data asset	Telco Churn Products	Shivam Solanki (IBM)	Sep 18, 2020 6:07 PM	Telco...	0 reviews
Data asset	Telco Churn Customers	Shivam Solanki (IBM)	Sep 18, 2020 6:05 PM	Telco...	0 reviews
Data asset	Telco Churn Billing	Shivam Solanki (IBM)	Sep 18, 2020 5:59 PM	Telco...	0 reviews
Connection	Db2-warehouse-churn-srs	Shivam Solanki (IBM)	Sep 18, 2020 5:43 PM	Telco...	0 reviews
Data asset	churn-prediction-research	Shivam Solanki (IBM)	Sep 18, 2020 5:18 PM	Telco...	0 reviews
Data asset	customer	Shivam Solanki (IBM)	Sep 18, 2020 5:18 PM	Telco...	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.

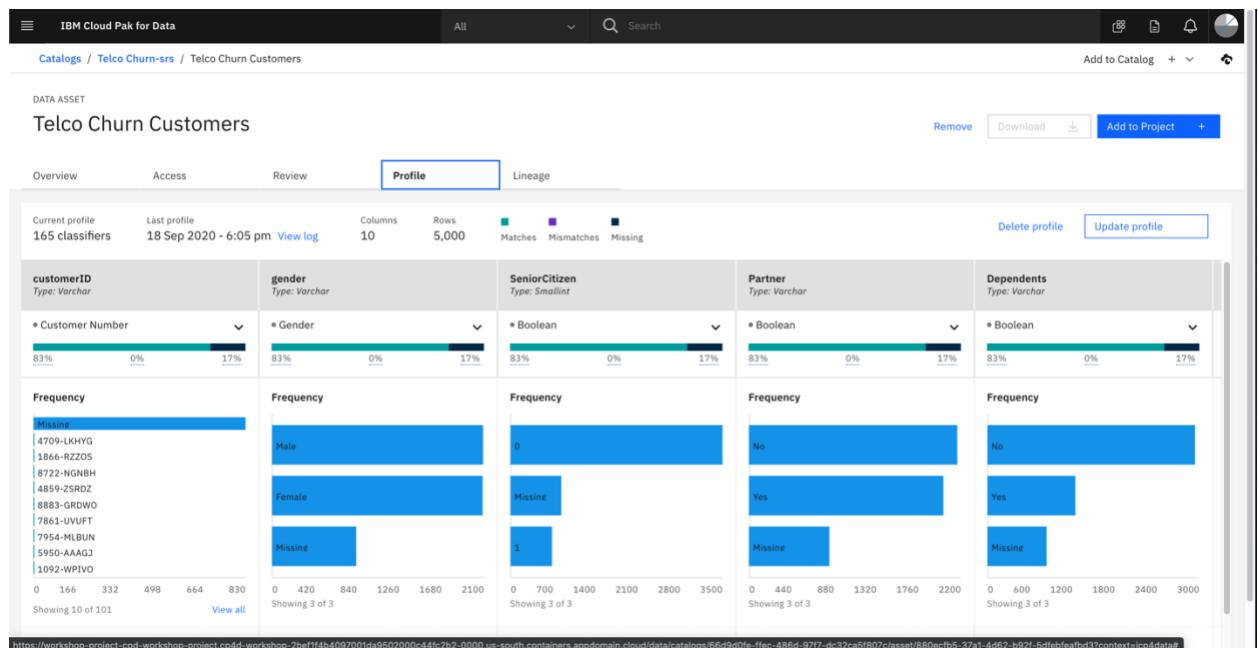
In this task, you will visit the **Profile** section of a structured data asset to examine the profiling and classification features provided. You will also visit the **Review** section to experience how you can rate and review assets to allow others to easily identify and evaluate them based on their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

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

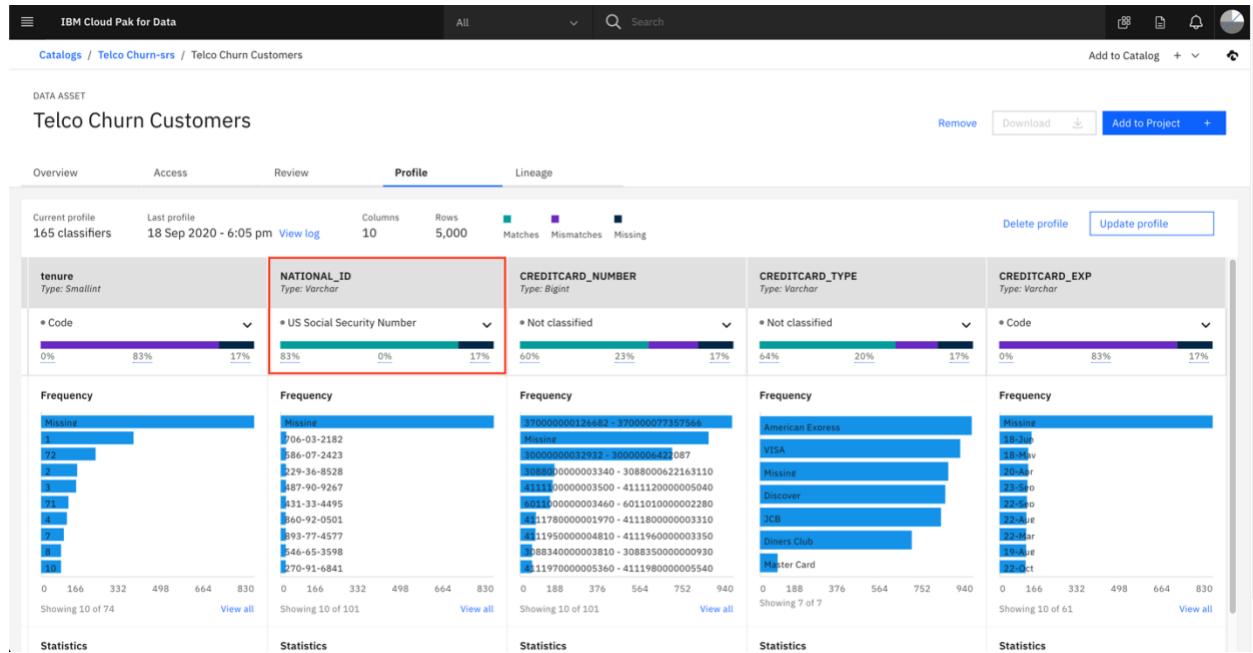
The profile should automatically appear. If not, and you are presented with a method to create or update the profile, follow the instructions to do so.



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 and CREDITCARD\_EXP columns are not. You will also notice they look like sensitive information and should be protected by data protection rules.

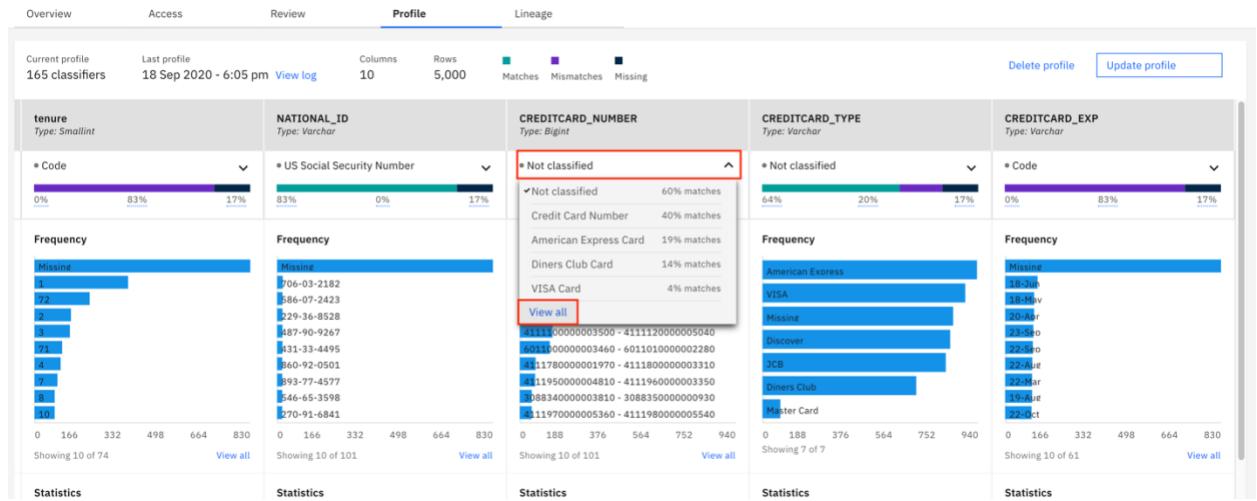
There are no data protection rules created in the glossary of this Cloud Pak for Data cluster to protect sensitive information. Even if there were, they would not be enforced and stop you from viewing the data content because you added the data asset to the catalog and are the owner. If another user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

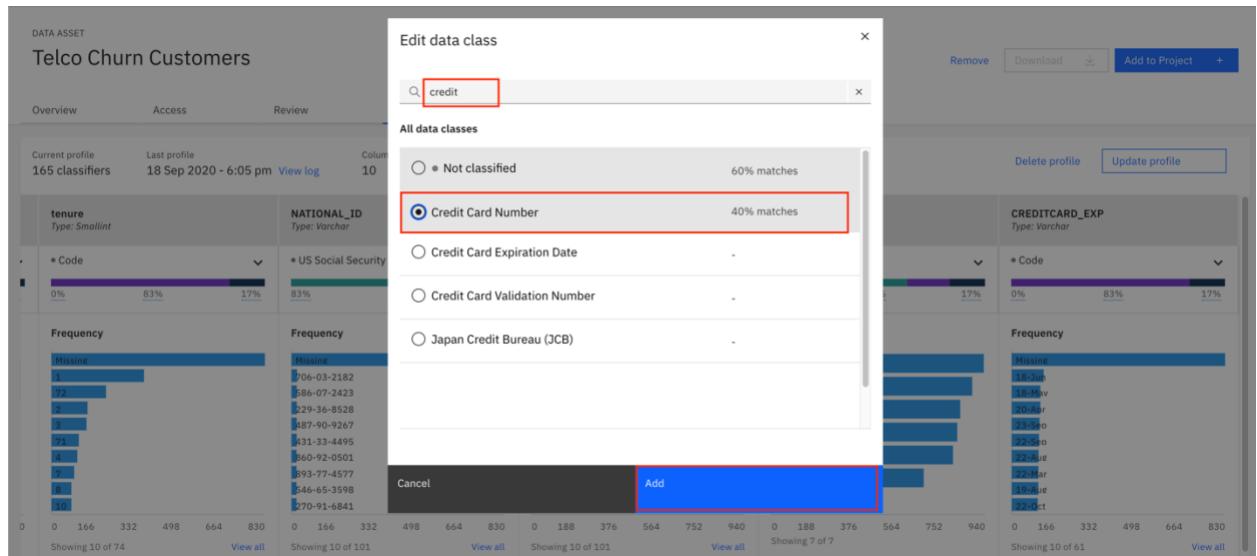
In any case, in order for any data protection rule based on a data class to be enforced, the data class assignments have to be correct. To protect the Credit Card number and

Credit Card Expiration Date of a Credit Card, the data classes for the CREDITCARD\_EXP and CREDITCARD\_NUMBER columns have to be set properly.

- Click on the down arrow next to the **Not Classified** classification of the CREDITCARD\_NUMBER 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\_EXP column.
- Click the **View All** menu item.
- Type in **credit card** in the search area.

11. Hover next to the **Credit Card Expiration Date** data class and click **Add**.

12. Click the **Close** button.

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

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

This Telco customer data is quality data that comes from the trusted telco data warehouse. However, in order to get the full value of this data it needs to be combined with the telco billing and products data.

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

Telco Churn Customers

Remove

Overview Access Review Profile Lineage

Overall rating  
0.0  
☆☆☆☆ 0 reviews

Review summary  
5 (0)  
4 (0)  
3 (0)  
2 (0)  
1 (0)

Your Review

SSK Shivam Solanki (IBM) Sep 18, 2020  
★★★★☆

This Telco customer data is quality data that comes from the trusted telco data warehouse. However, in order to get the full value of this data it needs to be combined with the telco billing and products data.

Cancel Submit

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

Catalogs / Telco Churn-srs / Telco Churn Customers Add to Catalog +

DATA ASSET

Telco Churn Customers

Remove Download Add to Project

Overview Access Review Profile Lineage

Overall rating  
4.0  
★★★★☆ 1 review

Review summary  
5 (0)  
4 (1)  
3 (0)  
2 (0)  
1 (0)

Your Review

SSK Shivam Solanki (IBM) Sep 18, 2020  
★★★★☆

This Telco customer data is quality data that comes from the trusted telco data warehouse. However, in order to get the full value of this data it needs to be combined with the telco billing and products data.

Cancel Submit

16. Click the **Telco Churn** 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 **Telco Churn Customers** data asset is the most highly rated data asset with 1 review.

### Telco Churn-srs

The screenshot shows the Watson Recommendations interface for the 'Telco Churn-srs' catalog. The 'Highly Rated' section is highlighted with a red box. The 'Telco Churn Customers' data asset is listed first, showing it was added on Sep 18, 2020 at 6:05 PM, owned by Shivam Solanki (IBM), and has a 5-star rating with 1 review. Other assets like 'churn-prediction-research' and 'customer-churn-predictiton-ml' are also listed with their details and ratings.

18. From the data asset list below, click on the **customer-churn-predictiton-ml** asset to view its properties.

The screenshot shows the properties of the 'customer-churn-predictiton-ml' data asset. The 'Overview' tab is selected, displaying the document's title, author, and basic metadata. A large preview area shows the document's content, which is a research paper titled 'Customer churn prediction in telecom using machine learning in big data platform' published in the Journal of Big Data. The 'Research' tab is also visible, providing the full text of the paper.

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 research paper on telco churn but will not be useful for our telco churn analytics project.

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

Catalogs / Telco Churn-srs / customer-churn-predictiton-ml

DATA ASSET  
customer-churn-predictiton-ml

Remove Download

Overall rating  
0.0  
☆☆☆☆☆ 0 reviews

Review summary  
5 (0)  
4 (0)  
3 (0)  
2 (0)  
1 (0)

Your Review  
Shivam Solanki (IBM) Sep 18, 2020  
★★★☆☆

Very interesting research paper on telco churn but will not be useful for our telco churn analytics project.

Cancel Submit

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

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

Catalogs / Telco Churn-srs

Add to Catalog +

Telco Churn-srs

Browse Assets Access Control Settings

What assets are you looking for?

Any type Any source Any tag

Watson Recommends Highly Rated Recently Added

Collaps

Data asset	Connection	Data asset	Data asset
Telco Churn Customers	churn-prediction-research	Db2-warehouse-churn-srs	Telco Chu
Owner: Shivam Solanki (IBM) Added: Sep 18, 2020 6:05 PM Tags: Telco...	Owner: Shivam Solanki (IBM) Added: Sep 18, 2020 5:18 PM Tags: Docu... Telco...	Owner: Shivam Solanki (IBM) Added: Sep 18, 2020 5:18 PM Tags: Docu... Telco...	Owner: Shivam Solanki (IBM) Added: Sep 18, 2020 5:59 PM Tags: Telco...
★★★★☆ 1 review	★★★☆☆ 1 review	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews

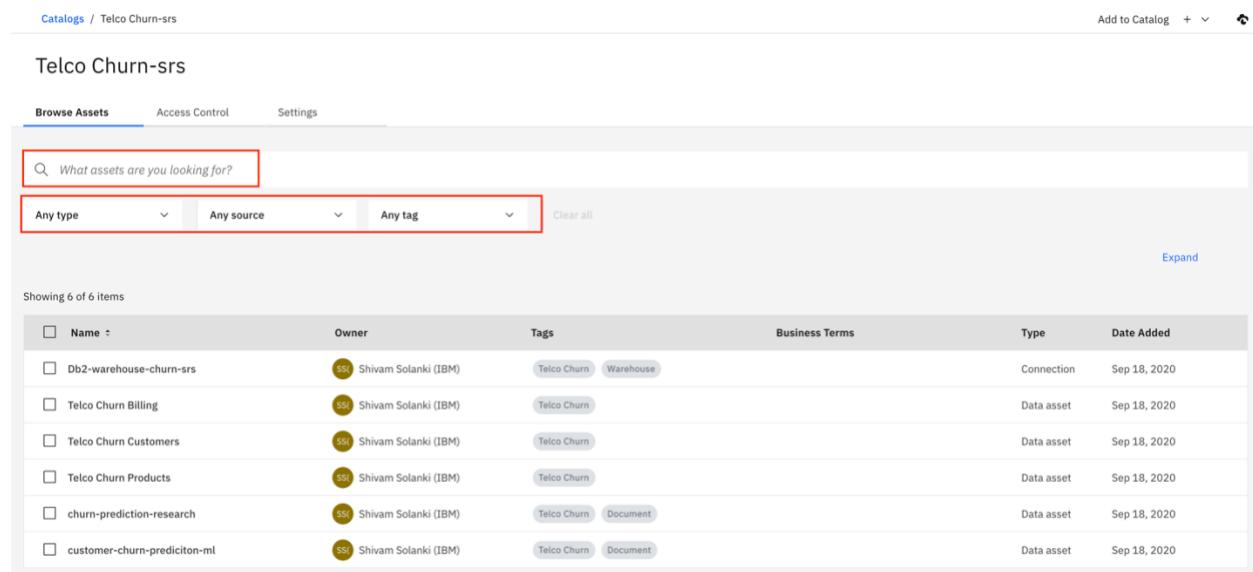
Showing 6 of 6 items

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

Notice that the **customer-churn-prediction-ml** 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 for the 'Telco Churn-srs' catalog. At the top, there's a search bar with placeholder text 'What assets are you looking for?' and three dropdown filters: 'Any type', 'Any source', and 'Any tag'. The 'Any tag' dropdown is highlighted with a red border. Below the filters, a message says 'Showing 6 of 6 items'. A table lists six data assets:

Name	Owner	Tags	Type	Date Added
Db2-warehouse-churn-srs	Shivam Solanki (IBM)	Telco Churn, Warehouse	Connection	Sep 18, 2020
Telco Churn Billing	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
Telco Churn Customers	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
Telco Churn Products	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn, Document	Data asset	Sep 18, 2020
customer-churn-prediciton-ml	Shivam Solanki (IBM)	Telco Churn, Document	Data asset	Sep 18, 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.

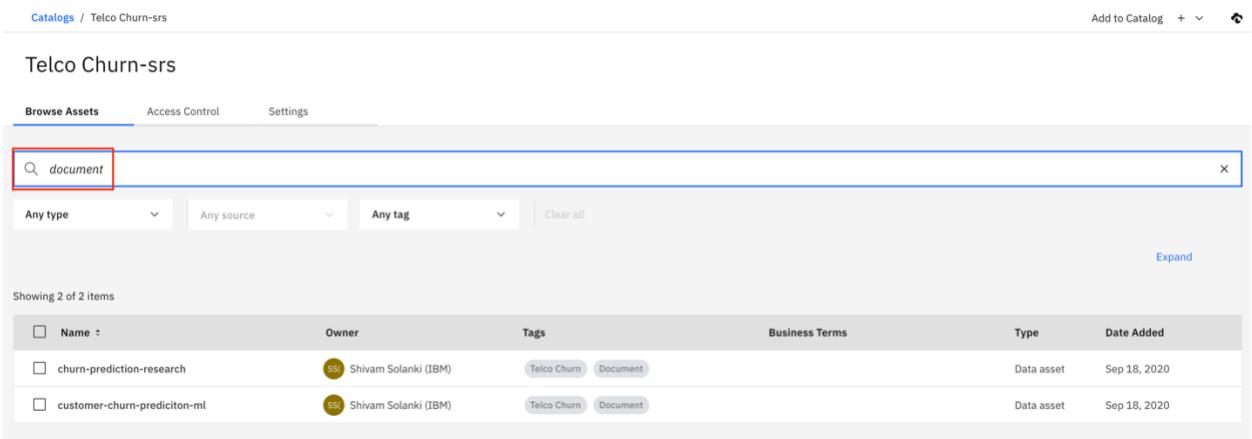
### Shop using Suggestions

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 'Telco Churn-srs' catalog page. The search bar at the top contains the text 'document'. Below the search bar, there are filters for 'Any type', 'Any source', 'Any tag', and a 'Clear all' button. A red box highlights the search bar. The results section shows 'Showing 2 of 2 items' with two data assets listed:

Name	Owner	Tags	Type	Date Added
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020
customer-churn-predictiton-ml	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020

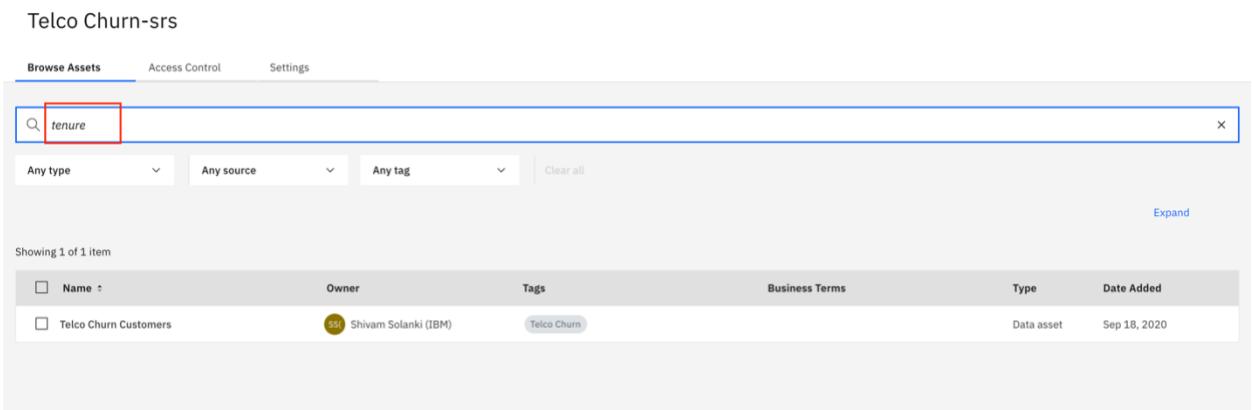
Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **document**.

## 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 **Warehouse**.

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

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



The screenshot shows the 'Telco Churn-srs' catalog page. The search bar at the top contains the text 'tenure'. Below the search bar, there are filters for 'Any type', 'Any source', 'Any tag', and a 'Clear all' button. A red box highlights the search bar. The results section shows 'Showing 1 of 1 item' with one data asset listed:

Name	Owner	Tags	Type	Date Added
Telco Churn Customers	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020

## 5. Inside the Knowledge Catalog search area type in **tenure**.

Why is the **Telco Churn Customers** table in the result set? This is an example of the search finding an asset that has a column that contains the consecutive characters **tenure** in its name. This table has column named **tenure** 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 **research** in the search area and view the results.
- Enter the characters **ml** 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.

Telco Churn-srs

The screenshot shows the 'Browse Assets' tab selected in the top navigation bar. Below it is a search bar containing the placeholder 'What assets are you looking for?'. Underneath the search bar are three dropdown filters: 'Any type', 'Any source', and 'Any tag', each with a collapse arrow icon. To the right of these filters is a 'Clear all' button. Further down, there is an 'Expand' button enclosed in a red box. At the bottom of the screen, there is a table header with columns: Name, Owner, Tags, Business Terms, Type, and Date Added. The table below shows 6 items.

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.

Telco Churn-srs

This screenshot shows the same interface as the previous one, but the 'Any type' filter dropdown is now expanded, revealing two options: 'Connection' and 'Data asset'. The 'Data asset' option is highlighted with a red box. The rest of the interface, including the search bar, other filters, and the table at the bottom, remains the same.

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.

## Telco Churn-srs

Name	Owner	Tags	Type	Date Added
Telco Churn Billing	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
Telco Churn Customers	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
Telco Churn Products	Shivam Solanki (IBM)	Telco Churn	Data asset	Sep 18, 2020
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020
customer-churn-predictiton-ml	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020

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

Name	Owner	Tags	Type	Date Added
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020
customer-churn-predictiton-ml	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020

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

Name	Owner	Tags	Type	Date Added
churn-prediction-research	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020
customer-churn-predictiton-ml	Shivam Solanki (IBM)	Telco Churn Document	Data asset	Sep 18, 2020

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

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

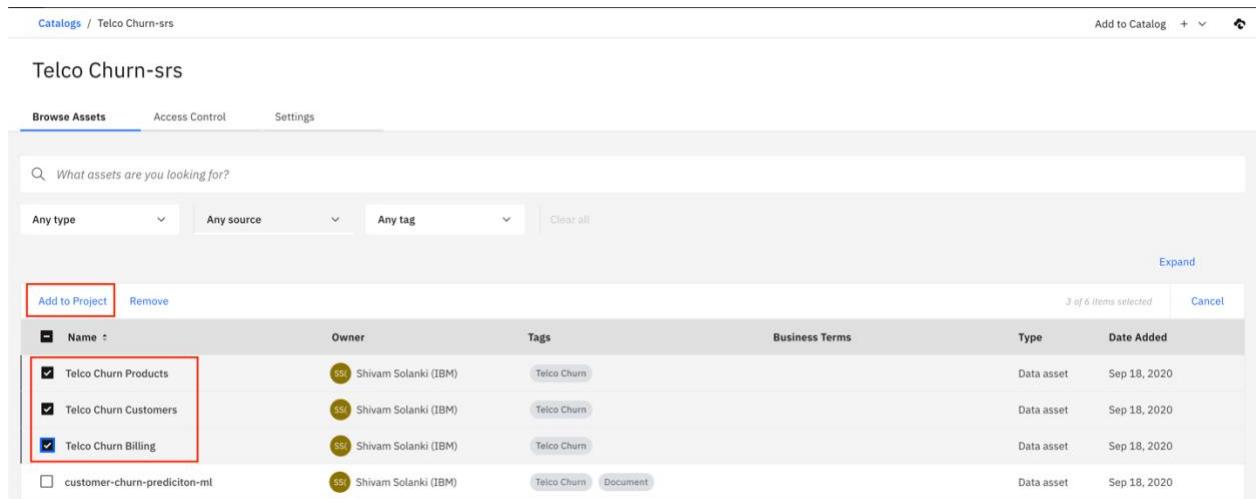
## Prepare Data for Analytics and AI

In this task, you will take the structured data you cataloged in the **Telco Churn** catalog into the **Telco Churn** 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 **Telco Churn** 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 **Telco Churn** project from the **Telco Churn** catalog.

1. Click the **check box** next to the **Telco Churn Products** data asset.



The screenshot shows the Catalogs interface with the path 'Catalogs / Telco Churn-srs'. The 'Browse Assets' tab is selected. A search bar at the top says 'What assets are you looking for?'. Below it are filters for 'Any type', 'Any source', 'Any tag', and a 'Clear all' button. An 'Expand' button is located in the top right of the asset list. The asset list table has columns: Name, Owner, Tags, Business Terms, Type, and Date Added. Three items are selected, indicated by checked checkboxes in the 'Name' column:

Name	Owner	Tags	Business Terms	Type	Date Added
<input checked="" type="checkbox"/> Telco Churn Products	Shivam Solanki (IBM)	Telco Churn		Data asset	Sep 18, 2020
<input checked="" type="checkbox"/> Telco Churn Customers	Shivam Solanki (IBM)	Telco Churn		Data asset	Sep 18, 2020
<input checked="" type="checkbox"/> Telco Churn Billing	Shivam Solanki (IBM)	Telco Churn		Data asset	Sep 18, 2020
<input type="checkbox"/> customer-churn-prediction-ml	Shivam Solanki (IBM)	Telco Churn Document		Data asset	Sep 18, 2020

2. Click the **check box** next to the **Telco Churn Customers** data asset.
3. Click the **check box** next to the **Telco Churn Billing** data asset.

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

**Target\***

Telco Churn - ss

**Selected assets (3)**

Asset Name	Catalog	Connection
Telco Churn Billing	Telco Churn-srs	Db2-warehouse-churn-srs
Telco Churn Customers	Telco Churn-srs	Db2-warehouse-churn-srs
Telco Churn Products	Telco Churn-srs	Db2-warehouse-churn-srs

**Connections to be added (1)**

Db2-warehouse-churn-srs Telco Churn-srs Knowledge Catalog Tutorial Db2 warehouse
--

**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.

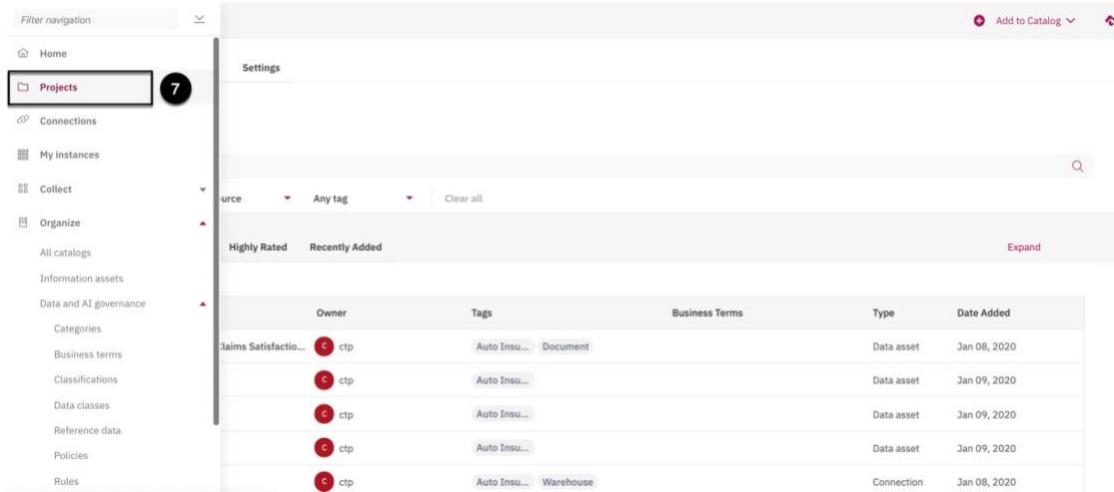
- Select the **Telco Churn** project from the list of Target projects.
- Click the **Add** button.

**Catalogs** > Auto Insurance

**Add to Catalog**

You are brought back into the **Telco Churn** 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.

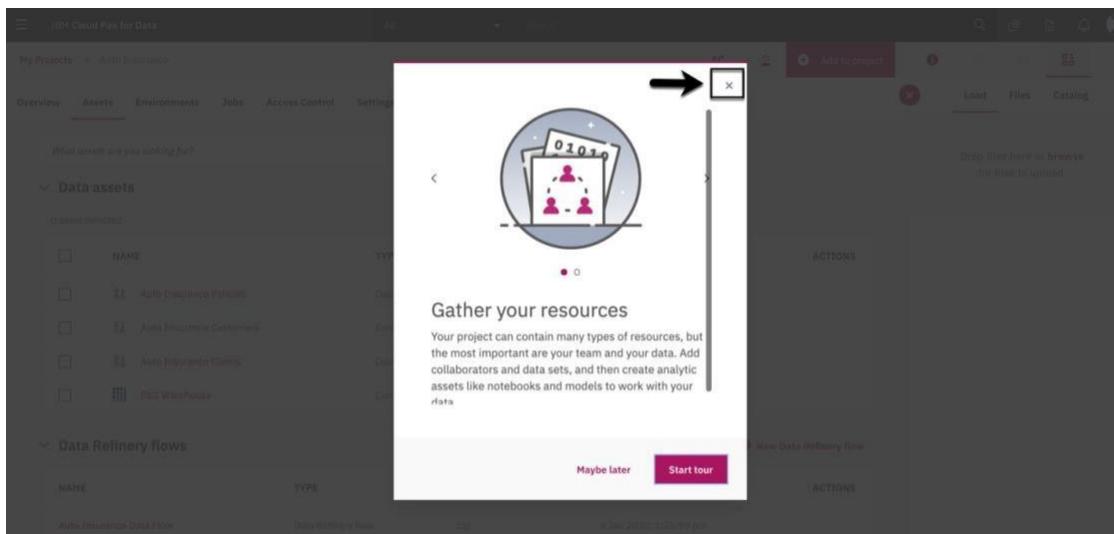
7. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.



The screenshot shows the navigation menu on the left side of the interface. The 'Projects' option is highlighted with a red box and a large number '7' indicating it is the current selection. Other options include 'Home', '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 'Settings' panel titled 'Claims Satisfaction' with a search bar and a table listing data assets and connections. The table has columns for Owner, Tags, Business Terms, Type, and Date Added. The data listed includes 'Auto Insurance Policies' (Owner: ctp, Tags: Auto Insu..., Document, Type: Data asset, Date Added: Jan 08, 2020), 'Auto Insurance Customers' (Owner: ctp, Tags: Auto Insu..., Type: Data asset, Date Added: Jan 09, 2020), 'Auto Insurance Claims' (Owner: ctp, Tags: Auto Insu..., Type: Data asset, Date Added: Jan 09, 2020), 'Auto Insurance Policies' (Owner: ctp, Tags: Auto Insu..., Type: Data asset, Date Added: Jan 09, 2020), and 'Auto Insurance Warehouse' (Owner: ctp, Tags: Auto Insu..., Warehouse, Type: Connection, Date Added: Jan 08, 2020).

8. Click the **Projects** menu.

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



The screenshot shows the 'Auto Insurance' project page. On the left, there are sections for 'Overview', 'Assets' (selected), 'Environments', 'Jobs', 'Access Control', and 'Settings'. Under 'Assets', there are sections for 'Data assets' and 'Data Refinery flows'. The 'Data assets' section lists 'Auto Insurance Policies', 'Auto Insurance Customers', 'Auto Insurance Claims', and 'Auto Insurance Warehouse'. The 'Data Refinery flows' section lists 'Auto Insurance Data Flow' and 'Data Refinery Flow'. A central modal dialog box titled 'Gather your resources' is displayed, containing text about adding collaborators and data sets, and creating analytic assets like notebooks and models. It includes a 'Start tour' button and 'Maybe later' link. At the bottom of the page, there are 'Load', 'Files', and 'Catalog' buttons.

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

## Refine the Data

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

The screenshot shows the 'IBM Cloud Pak for Data' interface. At the top, there's a navigation bar with 'My projects / Telco Churn - ss'. Below it is a horizontal menu with tabs: 'Overview' (disabled), 'Assets' (highlighted with a red box), 'Environments', 'Jobs', 'Access Control', and 'Settings'. A search bar is positioned above the main content area. The main area is titled 'Data assets' with a subtitle '0 assets selected.' Below this is a table listing four assets:

<input type="checkbox"/>	Name	Type	Created by	Last modified
<input type="checkbox"/>	Telco Churn Customers	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM
<input type="checkbox"/>	Telco Churn Billing	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM
<input type="checkbox"/>	Telco Churn Products	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM

To the right of the table, a context menu is open for the first asset ('Telco Churn Customers'). The menu items are: 'Publish to Catalog', 'Refine' (highlighted with a red box), 'Download', 'Promote', and 'Remove'.

2. Select the **ellipses...** to the right of the **Telco Churn Customers** data asset to view the data asset action menu.
3. Select the **Refine** menu item.

You are brought into the **data preparation** component of the analytic project to begin shaping the **Telco Churn 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.

My Projects > Auto Insurance > Auto Insurance Customers > Refine data

**Operation** Code an operation to cleanse and shape your data

**Data** Profile Visualizations

**Get perspective on your data**

1 Cleanse and shape your data on the Data tab. Validate your data and find anomalies on the Profile tab. Get insights into your data on the Visualizations tab.

2 3 4 5 6 1 of 4 Next

	COUNTRY	LATITUDE	LONGITUDE
1	String US	41.75113981	-88.0127658
2	String US	39.2781	-120.1203
3	String US	47.76121	-122.3464
4	String US	33.88081187	-118.0288381
5	String US	34.02091598	-84.31698227
6	String US	41.77028751	-88.20481022
7	String AR47B49 Janine McCreath	40.9581	-74.0747
8	String AS97690 Milicent Caveau	40.8591	-73.9694
9	String AW77988 Agnes Woodfield	38.9974	-105.0672
10	String AY40674 Jamal Duddle	35.080383	-81.70435
11	String AZ34845 Dov Gabriely	42.339208	-71.134786
12	String BA75404 Dee dee Mugglesone	42.43478788	-81.43361611
13	String BB882067 Ellery Clorenshaw	39.58050569	-105.1353397
14	String BC66536 Joelyn Pilgram	33.98205235	-118.2491117
15	String BD76386 Marie Flinders	41.9138533	-88.31240528

SOURCE FILE: Auto Insurance Customers SAMPLE SIZE: First 328 rows

**Steps**

0 STEPS

**Data Source** Auto Insurance Customers

**DATA REFINERY FLOW DETAILS**

LOCATION Auto Insurance

DATA REFINERY FLOW NAME Auto Insurance Customers\_flow

Enter a description of the Data Refinery flow

**STEPS** 0

**DATA REFINERY FLOW OUTPUT**

LOCATION Auto Insurance/Data assets

DATA SET NAME Auto Insurance Customers s...

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

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

My Projects / Telco Churn - ss / Telco Churn Customers / Refine data

**Operation** + Code an operation to cleanse and shape your data

**Data** Profile Visualizations

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	NATIONALITY	CREDITCARDDEBT
1	0114-RSRRW	Female	0	Yes	No	10	115-27-7498	XXXXXXXXXX
2	7206-GZCDC	Female	1	No	No	1	793-37-6798	XXXXXXXXXX
3	4791-QRGMF	Male	0	Yes	No	59	908-13-5300	XXXXXXXXXX
4	3838-OZURD	Male	0	Yes	No	66	014-12-2539	XXXXXXXXXX
5	1371-DWPAZ	Female	0	Yes	Yes	0	452-09-8933	XXXXXXXXXX
6	2017-CCLBLH	Female	0	No	No	8	403-59-2982	XXXXXXXXXX
7	9415-DPEWS	Female	0	No	No	18	028-98-3917	XXXXXXXXXX
8	5624-RYAMH	Female	0	No	No	9	463-48-0895	XXXXXXXXXX
9	9272-LSVYH	Male	0	No	No	10	894-07-4968	XXXXXXXXXX
10	3249-ZPQRG	Male	0	No	No	4	636-30-4575	XXXXXXXXXX
11	3084-DOWLE	Female	0	Yes	No	72	473-07-9503	XXXXXXXXXX

**Steps**

0

**Information**

**Details** Edit

**LOCATION** Telco Churn - ss

**DATA REFINERY FLOW NAME** Telco Churn Customers \_...

Enter a description of the Data Refinery flow

**STEPS** 0

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

**DATA REFINERY FLOW DETAILS**

**LOCATION**  
Telco Churn - ss

Data Refinery Flow Name \*  
**Telco Churn Data Flow**

Description  
**Prepare the telco churn data for analytics**

Cancel **Apply**

**STEPS**  
0

**DATA REFINERY FLOW OUTPUT**

Location [Edit Output](#)  
Telco Churn - ss/Data assets

Data Set Name  
**Telco Churn Customers \_sh...**

Enter a description of the resulting data set.  
**✓ If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.**

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

**DATA REFINERY FLOW DETAILS**

**LOCATION**  
Telco Churn - ss

**DATA REFINERY FLOW NAME**  
**Telco Churn Data Flow**

Prepare the telco churn data for analytics

**STEPS**  
0

**DATA REFINERY FLOW OUTPUT**

Edit output

Location \*  
Telco Churn - ss/Data assets

Data set name \*  
**Telco Churn Shaped**

Description  
**Telco Churn Customer, Products and Billing data combined and refined**

File format  
CSV

The first line of the file contains column headers

7. Copy and paste, or enter, this bolded text **Prepare the telco churn 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.

The screenshot shows the 'Data Refinery Flow Output' dialog box. On the left, under 'DATA REFINERY FLOW DETAILS', there is a 'LOCATION' section with 'Telco Churn - ss'. Below it is a 'DATA REFINERY FLOW NAME' section with 'Telco Churn Data Flow' and a description 'Prepare the telco churn data for analytics'. Under 'STEPS', it says '0'. On the right, under 'DATA REFINERY FLOW OUTPUT', there is a 'Location' field set to 'Telco Churn - ss/Data assets', a 'Data Set Name' field set to 'Telco Churn Shaped', and a checkbox with the text 'If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.' A blue 'Edit Output' button is visible above the location field. At the bottom of the dialog, there is a note 'Review the Data Refinery flow details and the Data Refinery flow output details.' and a red 'Done' button.

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 **Telco Churn** project.

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

My Projects / Telco Churn - ss / Telco Churn Customers / Refine data

**DATA REFINERY FLOW DETAILS**

**LOCATION**  
Telco Churn - ss

**DATA REFINERY FLOW NAME**  
**Telco Churn Data Flow**  
Prepare the telco churn data for analytics

**STEPS**  
0

**DATA REFINERY FLOW OUTPUT**

[Edit Output](#)

**Location**  
Telco Churn - ss/Data assets

**Data Set Name**  
**Telco Churn Shaped**

Telco Churn Customer, Products and Billing data combined and refined

If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.

Review the Data Refinery flow details and the Data Refinery flow output details.

**Done**

11. Copy and paste, or enter, this bolded text **Telco Churn Customer, Products and Billing 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.

My Projects / Telco Churn - ss / Telco Churn Customers / Refine data

**DATA REFINERY FLOW DETAILS**

**LOCATION**  
Telco Churn - ss

**DATA REFINERY FLOW NAME**  
**Telco Churn Data Flow**  
Prepare the telco churn data for analytics

**STEPS**  
0

**DATA REFINERY FLOW OUTPUT**

[Edit Output](#)

**Location**  
Telco Churn - ss/Data assets

**Data Set Name**  
**Telco Churn Shaped**

Telco Churn Customer, Products and Billing data combined and refined

If the data set already exists, overwrite the data in the existing data set with the Data Refinery flow output.

Review the Data Refinery flow details and the Data Refinery flow output details.

**Done**

13. Click the **Done** button.

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

The screenshot shows the Cloud Pak for Data Workshop interface. On the left, there's a data preview table with columns: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, NATIONALITY, and CREDITCARD. The table contains 17 rows of sample data. On the right, there's a 'Details' panel with sections for 'Information' (containing 'Edit' and 'LOCATION Telco Churn - ss'), 'Data Refinery Flow Name' (set to 'Telco Churn Data Flow'), and 'Steps' (set to 0). Below these are sections for 'Data Refinery Flow Output' (Location: Telco Churn - ss/Data assets) and 'Data Set Name' (set to 'Telco Churn Shaped'). A red box highlights the close button ('X') in the top right corner of the Details panel.

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

## Combine Data

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

The screenshot shows the 'Operation' menu interface. On the left, there's a sidebar with a search bar and sections for 'CLEANSE', 'ORGANIZE', and 'NATURAL LANGUAGE'. Under 'CLEANSE', several operations like 'Convert column value to missing' and 'Remove duplicates' are listed. Under 'ORGANIZE', 'Join' is highlighted with a red box. On the right, a main panel displays a table of data with columns: customerID, gender, SeniorCitizen, Partner, and Dependents. At the bottom right of the main panel, it says 'SOURCE'.

customerID	gender	SeniorCitizen	Partner	Dependents
0114-RSRRW	Female	0	Yes	No
7206-GZCDC	Female	1	No	No
4791-QRGMF	Male	0	Yes	No
3838-OZURD	Male	0	Yes	No
1371-DWPAZ	Female	0	Yes	Yes
2017-CCBLH	Female	0	No	No
9415-DPEWS	Female	0	No	No
5624-RYAMH	Female	0	No	No
9272-LSVYH	Male	0	No	No
3249-ZPQRG	Male	0	No	No
3084-DOWLE	Female	0	Yes	No
1084-MNSMJ	Female	0	Yes	Yes
1131-SUEKT	Male	0	Yes	Yes
9432-RUVSL	Female	0	No	No
8060-HIWJJ	Male	0	No	No
9089-UOWJG	Female	0	Yes	Yes
9885-AIBVB	Male	0	Yes	No

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

The screenshot shows the 'Join' operation configuration screen. It has a dropdown for 'Join method' with 'Inner join' selected (highlighted with a red box). Below it, a note says 'Returns only the rows in each data set that match rows in the other data set. Returns one row in the original data set for each matching row in the joining data set.' A note below that says 'The default suffix for each data set will be used to differentiate any duplicate column names in the resulting data set.' On the left, there's a 'Source' section with a 'Telco Churn Cu...' dataset and a 'Data set to join' section with a '+ Add Data Set' button. The main panel shows a table of data with columns: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, and NATION. The 'Dependents' column from the source table is renamed 'Dependents' and the 'SeniorCitizen' column is renamed 'SeniorCitizen'.

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	NATIO
0114-RSRRW	Female	0	Yes	No	10	115-2'
7206-GZCDC	Female	1	No	No	1	793-3'
4791-QRGMF	Male	0	Yes	No	59	908-1'
3838-OZURD	Male	0	Yes	No	66	014-1'
1371-DWPAZ	Female	0	Yes	Yes	0	452-0'
2017-CCBLH	Female	0	No	No	8	403-5'
9415-DPEWS	Female	0	No	No	18	028-9'
5624-RYAMH	Female	0	No	No	9	463-4'
9272-LSVYH	Male	0	No	No	10	894-0'
3249-ZPQRG	Male	0	No	No	4	636-3'
3084-DOWLE	Female	0	Yes	No	72	473-0'
1084-MNSMJ	Female	0	Yes	Yes	51	495-7'
1131-SUEKT	Male	0	Yes	Yes	61	558-6'
9432-RUVSL	Female	0	No	No	48	424-7'
8060-HIWJJ	Male	0	No	No	64	254-9'
9089-UOWJG	Female	0	Yes	Yes	12	123-3'

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

4. Click the **+ Add Data Set** button in the **Data set to join** section.
5. Click on the **Data assets** section.
6. Select the **Telco Churn Billing** data asset.
7. Click the **Apply** button.
8. Scroll down in the Join properties area until you see the JOIN KEYS section. Click in the **Telco Churn 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 configuration screen. On the left, there's a description of what a join does: 'Returns only the rows in each data set that match rows in the other data set. Returns one row in the original data set for each matching row in the joining data set.' Below this, under 'Source' and 'Data set to join', two datasets are listed: 'Telco Churn Cu...' and 'Telco Churn Bill...'. Each has a suffix input field ('\*Suffix \_X' and '\_Y'). In the center, the 'JOIN KEYS' section is highlighted with a red box. It contains a table with two rows: 'Telco Churn Cust...' and 'Telco Churn Billing'. Under 'customerID' in both rows, the 'customerID' column is selected. At the bottom right of the main panel, there's a 'SOURCE FILE: Telco' label. At the very bottom, there are 'Cancel' and 'Next' buttons, with 'Next' also highlighted with a red box.

9. Click in the **Telco Churn Billing** 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 Unchecking columns excludes them from the join result.

Operation X | Code an operation to cleanse and shape your data

Join

<input checked="" type="checkbox"/> SeniorCitizen	customerID String
<input checked="" type="checkbox"/> Partner	gender String
<input checked="" type="checkbox"/> Dependents	SeniorCitizen String
<input checked="" type="checkbox"/> tenure	Partner String
<input checked="" type="checkbox"/> NATIONAL_ID	Dependents String
<input type="checkbox"/> CREDITCARD_NUMBER	tenure String
<input type="checkbox"/> CREDITCARD_TYPE	NATIONAL. String
<input type="checkbox"/> CREDITCARD_EXP	
<input checked="" type="checkbox"/> Contract	
<input checked="" type="checkbox"/> PaperlessBilling	
<input checked="" type="checkbox"/> PaymentMethod	
<input checked="" type="checkbox"/> MonthlyCharges	
<input checked="" type="checkbox"/> TotalCharges	
<input checked="" type="checkbox"/> Churn	

Back | Apply | SOURCE FILE: Telco Churn Customers | SAM

## 12. Click the **Apply** button.

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

Operation + | Code an operation to cleanse and shape your data

Data Profile Visualizations | Steps

	NATIONAL... String	Contract String	PaperlessBilling String	PaymentMethod String	MonthlyCharges Decimal	TotalCharges Decimal	Churn String
1	863-36-5468	Month-to-month	No	Mailed check	19.4	415.4	No
2	171-22-9346	Month-to-month	Yes	Electronic check	89.75	5496.9	No
3	278-47-2963	One year	No	Bank transfer (automatic)	75.7	3876.2	No
4	740-36-9151	Two year	No	Bank transfer (automatic)	115.15	8349.45	No
5	182-07-4726	Month-to-month	No	Electronic check	24.35	41.85	Yes
6	091-04-0989	Month-to-month	No	Mailed check	33.6	33.6	No
7	981-42-7607	Month-to-month	Yes	Bank transfer (automatic)	79.9	343.95	Yes
8	794-25-2312	One year	No	Bank transfer (automatic)	20.15	405.6	No
9	641-23-7039	One year	No	Mailed check	20.05	96.8	No
10	266-04-3726	Two year	Yes	Credit card (automatic)	116.8	8456.75	No
11	712-59-7172	Month-to-month	Yes	Electronic check	90.05	368.1	Yes
12	914-03-3091	Two year	Yes	Electronic check	108.2	7840.6	No
13	825-90-0828	Two year	No	Bank transfer (automatic)	56.75	1304.85	No
14	678-73-1895	One year	Yes	Credit card (automatic)	24.5	1497.9	No
15	857-56-4800	Two year	Yes	Credit card (automatic)	79.2	4590.35	No
16	975-15-9861	Month-to-month	No	Bank transfer (automatic)	87.65	2766.4	No
17	250-61-9589	One year	Yes	Bank transfer (automatic)	75.35	2243.9	No

SOURCE FILE: Telco Churn Customers | SAMPLE SIZE: First 140 rows

1 Steps

Data Source: Telco Churn Customers

Join: JUST ADDED  
inner-joined data from Telco Churn Billing based on columns customerID, customerID

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' interface with the 'Join' operation highlighted. The main pane displays a sample of data from the 'Telco Churn Customers' file, showing columns like NationalID, Contract, PaperlessBilling, PaymentMethod, and MonthlyCharges. The sidebar lists various operations under categories like 'CLEANSE', 'ORGANIZE', and 'NATURAL LANGUAGE'. The right panel shows the 'Steps' section with one step added: 'Data Source: Telco Churn Customers' and 'Join: JUST ADDED'.

NATIONALID	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges
863-36-5468	Month-to-month	No	Mailed check	19.4
171-22-9346	Month-to-month	Yes	Electronic check	89.75
278-47-2963	One year	No	Bank transfer (automatic)	75.7
740-36-9151	Two year	No	Bank transfer (automatic)	115.15
182-07-4726	Month-to-month	No	Electronic check	24.35
091-04-0989	Month-to-month	No	Mailed check	33.6
981-42-7607	Month-to-month	Yes	Bank transfer (automatic)	79.9
794-25-2312	One year	No	Bank transfer (automatic)	20.15
641-23-7039	One year	No	Mailed check	20.05
266-04-3726	Two year	Yes	Credit card (automatic)	116.8
712-59-7172	Month-to-month	Yes	Electronic check	90.05
914-03-3091	Two year	Yes	Electronic check	108.2
825-90-0828	Two year	No	Bank transfer (automatic)	56.75
678-73-1895	One year	Yes	Credit card (automatic)	24.5
857-56-4800	Two year	Yes	Credit card (automatic)	79.2
975-15-9861	Month-to-month	No	Bank transfer (automatic)	87.65
250-61-9589	One year	Yes	Bank transfer (automatic)	75.35

SOURCE FILE: Telco Churn Customers SAMPLE SIZE: First 140 rows

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 'Inner join' method is selected in the dropdown. The 'Source' section lists 'Telco Churn Cu...' and the 'Data set to join' section lists 'Add Data Set'. The right panel shows the 'Steps' section with one step added: 'Data Source: Telco Churn Customers' and 'Join: JUST ADDED'.

Combine data from two data sets based on a comparison of the values in specified key columns.

Inner join

Returns only the rows in each data set that match rows in the other data set. Returns one row in the original data set for each matching row in the joining data set.

The default suffix for each data set will be used to differentiate any duplicate column names in the resulting data set.

Source: Telco Churn Cu... Data set to join: Add Data Set \*Suffix: \_X \_Y

NATIONALID	Contract	PAPERLESSBILLING	PAYMENTMETHOD	MONTHLYCHARGES
863-36-5468	Month-to-month	No	Mailed check	19.4
171-22-9346	Month-to-month	Yes	Electronic check	89.75
278-47-2963	One year	No	Bank transfer (automatic)	75.7
740-36-9151	Two year	No	Bank transfer (automatic)	115.15
182-07-4726	Month-to-month	No	Electronic check	24.35
091-04-0989	Month-to-month	No	Mailed check	33.6
981-42-7607	Month-to-month	Yes	Bank transfer (automatic)	79.9
794-25-2312	One year	No	Bank transfer (automatic)	20.15
641-23-7039	One year	No	Mailed check	20.05
266-04-3726	Two year	Yes	Credit card (automatic)	116.8
712-59-7172	Month-to-month	Yes	Electronic check	90.05
914-03-3091	Two year	Yes	Electronic check	108.2
825-90-0828	Two year	No	Bank transfer (automatic)	56.75
678-73-1895	One year	Yes	Credit card (automatic)	24.5
857-56-4800	Two year	Yes	Credit card (automatic)	79.2
975-15-9861	Month-to-month	No	Bank transfer (automatic)	87.65

SOURCE FILE: Telco Churn Customers SAMPLE SIZE: First 140 rows

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

17. Click on the **Data assets** section.
18. Select the **Telco Churn Products** 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 **Telco Churn Customers JOIN KEYS** column selection list on the left and select the **customerID** column as the join key column.

The screenshot shows the Data Workshop interface for performing a join operation. On the left, there's a 'Join' configuration pane with two data sources: 'Telco Churn Cu...' and 'Telco Churn Pro...'. The 'JOIN KEYS' section is highlighted with a red box, showing the selection of 'customerID' from the 'Telco Churn Customers' side. The main table displays 140 rows of data with columns like 'NATIONA...', 'Contract', 'PaperlessBilling', 'PaymentMethod', and 'MonthlyCharges'. On the right, the 'Steps' panel shows a single step: 'Data Source: Telco Churn Customers'.

21. Click in the **Telco Churn Products** JOIN KEYS column selection list on the right and select the **customerID** column as the join key column.
22. Click the **Next** button.
23. Click the **Apply** button.

The screenshot shows the Data Shaper interface with the following details:

- Operation:** Join
- Left Table:** Telco Churn Customers (140 rows)
- Right Table:** Telco Churn Billing (140 rows)
- Join Type:** inner-joined data from Telco Churn Billing based on columns customerID, customerID
- Selected Columns (Left):** Dependents, tenure, NATIONAL\_ID, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport.
- Selected Columns (Right):** NATIONA..., Contract, PaperlessBilling, PaymentMethod, MonthlyCharges.
- Toolbar Buttons:** Back, Apply (highlighted with a red box), Save, Undo, Redo, Refresh, Help, Steps.
- Steps Panel:** Shows 1 Step: Data Source - Telco Churn Customers, and 1 Join: JUST ADDED (inner-joined data from Telco Churn Billing based on columns customerID, customerID).

The join should complete successfully. **Scroll** to the right to see that the Telco Churn Products table columns are now appended at the end of the Telco Churn Customers and Billing 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 Shaper interface after saving the data flow, with the following details:

- Operation:** +
- Left Table:** Telco Churn Customers (43 rows)
- Right Table:** Telco Churn Billing (43 rows)
- Join Type:** inner-joined data from Telco Churn Billing based on columns customerID, customerID
- Selected Columns (Left):** InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
- Selected Columns (Right):** NATIONA..., Contract, PaperlessBilling, PaymentMethod, MonthlyCharges.
- Toolbar Buttons:** Back, Apply, Save (highlighted with a red box), Undo, Redo, Refresh, Help, Steps.
- Steps Panel:** Shows 2 Steps: Data Source - Telco Churn Customers, and 2 Joins: JUST ADDED (inner-joined data from Telco Churn Billing based on columns customerID, customerID) and JUST ADDED (inner-joined data from Telco Churn Products based on columns customerID, customerID).

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.

[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治理 (Data Governance) interface. On the left, there's a navigation bar with tabs: Operation (+), Profile, and Visualizations. Below this is a table with columns: Partner String, Dependents String, tenure String, NATIONAL\_ID String, Contract String, PaperlessBilling String, PaymentMethod String, and Monthly! Decimal. A context menu is open over the NATIONAL\_ID column, with the 'Substitute' option highlighted and surrounded by a red box. To the right of the table, a sidebar titled 'Steps' lists two steps: 'Data Source' (Telco Churn Customers) and 'Join' (inner-joined data from Telco Churn). At the bottom, it says 'SOURCE FILE: Telco Churn Customers' and 'SAMPLE SIZE: First 43 rows'.

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.

This screenshot shows the same data治理 interface after the 'Substitute' operation has been applied. The NATIONAL\_ID column now contains random strings like '7a5d8c5b9a1cc77...', '847389ca91b5349...', etc. The column action menu is no longer visible. The sidebar 'Steps' now includes a third step: 'Substitute' (JUST ADDED), which is described as 'Substituted random strings for the data in NATIONAL\_ID'. The bottom status bar remains the same: 'SOURCE FILE: Telco Churn Customers' and 'SAMPLE SIZE: First 43 rows'.

## Sort Data

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

The screenshot shows the Data Studio interface with a table of customer data. The table has columns: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, NATIONAL..., and Contract. A context menu is open over the first row of the 'customerID' column, with the 'Sort ascending' option highlighted and enclosed in a red box. To the right of the table, a sidebar titled 'Steps' shows three steps: 'Data Source' (Telco Churn Customers), 'Join' (inner-joined data from Telco Churn Billing based on columns customerID, customerID), and 'Substitute' (JUST ADDED). At the bottom of the table, it says 'SOURCE FILE: Telco Churn Customers' and 'SAMPLE SIZE: First 43 rows'.

	customerID String	gender String	SeniorCitizen String	Partner String	Dependents String	tenure String	NATIONAL... String	Contract String
1	9093-FPDLG	Remove	0	No	No	11	7a5d8c5b9a1cc77...	Month-to-montl
2	6838-HVLXG	Remove duplicates	0	No	No	12	847389ca91b5349...	Month-to-montl
3	2277-DJJDL	Remove empty rows	1	Yes	No	60	111d68fe41eb97d...	Month-to-montl
4	8922-NPKB3	<b>Sort ascending</b>	0	Yes	Yes	42	f76c40f76d6b8712...	Two year
5	6963-EZQEE	Sort descending	1	Yes	No	70	c261b9a7c29da41...	Two year
6	1564-HJUVY	Substitute	0	No	No	4	c201afb8b73343bf...	Month-to-montl
7	2672-HUYVI	CONVERT COLU...>	0	No	No	6	cceb9fec0a6f3e76...	Month-to-montl
8	7508-MYB0G	TEXT >	0	Yes	No	14	b6ba43ceee0009df...	Month-to-montl
9	0148-DCDOS		0	No	No	25	ed608491faef943...	Month-to-montl
10	8347-GDTMP	View All	0	Yes	No	64	75f6a4041ac07f3f...	Two year
11	0428-IKYCP	Male	0	Yes	No	22	52fcb378c34ea071...	Month-to-montl
12	4003-OCTMP	Female	0	Yes	No	31	cff0834c184bc136...	One year
13	7225-IILWY	Male	0	Yes	Yes	68	9f7b9bcf2b855450...	Two year
14	2668-TZSPS	Male	0	No	No	1	531hb307d8b05ea...	Month-to-montl
15	8869-TORSS	Female	0	No	No	48	5c94940cf0f9b06...	One year
16	6505-OZNPQ	Female	0	No	No	6	44b4efda78fb08d9c...	Two year
17	0953-LGOVU	Male	0	Yes	Yes	12	daa02ca9f96e3f39...	Month-to-montl

SOURCE FILE: Telco Churn Customers    SAMPLE SIZE: First 43 rows

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.

## 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.

The screenshot shows the IBM Cloud Pak for Data interface. The top navigation bar includes 'IBM Cloud Pak for Data', 'My Projects / Telco Churn - ss / Telco Churn Customers / Refine data', and a search bar. The main area has tabs for 'Operation' (selected), 'Profile', and 'Visualizations'. Below these is a table with 17 rows of data from 'Telco Churn Customers'. The table columns are: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, NATIONAL\_ID, and Contract. The right side of the interface features a sidebar titled 'Steps' which details the data flow operations:

- Data Source:** Telco Churn Customers
- Join:** inner-joined data from Telco Churn Billing based on columns customerID,customerID
- Substitute:** Substituted random strings for the data in NATIONAL\_ID
- Sort ascending:** JUST ADDED

At the bottom of the sidebar, it says 'Sorted rows by customerID'.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	NATIONAL_ID	Contract
1	0148-DDOS	Male	0	No	No	25	ed608491faef943...	Month-to-month
2	0428-ICKYCP	Male	0	Yes	No	22	52fc5378c34e071...	Month-to-month
3	0484-FFVBJ	Male	0	No	No	32	69c6fa24db773bfd...	One year
4	0953-LGOVU	Male	0	Yes	Yes	12	daa02ca9f96e3139...	Month-to-month
5	1303-SRDOK	Female	0	Yes	Yes	55	a9a54046d2e3a74...	Two year
6	1564-HJUVY	Male	0	No	No	4	c201afb8b73343bf...	Month-to-month
7	1936-CZAKF	Male	0	Yes	No	49	a846bb210a92ab35...	Two year
8	2181-TIDSV	Male	0	Yes	Yes	68	4dd9f1721d20f8f0...	Two year
9	2277-DJJDL	Male	1	Yes	No	60	111d68fe41eb97d...	Month-to-month
10	2576-HXMPA	Female	0	No	No	1	b2f1289ae765c05...	Month-to-month
11	2640-PMGFL	Male	0	No	Yes	27	3b247d8132c09fb...	Month-to-month
12	2668-TZSPS	Male	0	No	No	1	531bb307d8b05ea...	Month-to-month
13	2672-HUVVI	Female	0	No	No	6	ccceb9fec0a6f3e76...	Month-to-month
14	2674-MIAHT	Female	0	No	No	4	2a5f27c8552ef9e9...	Month-to-month
15	2960-NKRSO	Male	0	No	No	24	8714b36855479d5...	Month-to-month
16	3043-TYBNO	Male	0	No	No	3	0008ad22b2a29386...	Month-to-month
17	3237-AJGEH	Female	0	Yes	Yes	3	a46fd0a549d5f055...	Month-to-month

SOURCE FILE: Telco Churn Customers SAMPLE SIZE: First 43 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 **Telco Churn Shaped** with the proper case, and spaces between the words.

Job Name  
Telco Churn

Description (Optional)  
Prepare the telco churn data for analytics

Associated Asset  
DATA REFINERY FLOW  
Telco Churn Data Flow 4 Steps Edit

Select runtime  
Default Data Refinery XS

Create and Run

5. Copy and paste, or enter, this bolded text: **Prepare the telco churn 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.

My Projects / Telco Churn - ss / Telco Churn

Telco Churn

Prepare the telco churn data for analytics

Associated Asset  
DATA REFINERY FLOW  
Telco Churn Data Flow 4 Steps

Scheduled to run  
No Schedule Created

Environment definition  
Default Data Refinery XS

INPUT  
Telco Churn Customers

OUTPUT  
Telco Churn Shaped

Runs (1)

Start Time	Status	Duration	Started By	Action
Sep 18, 2020 7:21:42 PM	Running	---	Shivam Solanki (IBM)	:

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 **Telco Churn** project navigation link on the toolbar to get back to the sections of the project.

The screenshot shows the 'Telco Churn' project page. At the top, there's a header with tabs for 'Telco Churn Customers' and 'Telco Churn - ss'. Below the header, the project name 'Telco Churn' is displayed. A sub-header says 'Prepare the telco churn data for analytics'. To the right, it shows an 'Associated Asset' section for a 'DATA REFINERY FLOW' named 'Telco Churn Data Flow' with 4 steps. Below this are sections for 'Scheduled to run' (No Schedule Created), 'Environment definition' (Default Data Refinery XS), 'INPUT' (Telco Churn Customers), and 'OUTPUT' (Telco Churn Shaped). A CSV download button is also present. Under the 'Runs (1)' section, a table lists one run: 'Start Time' (Sep 18, 2020 7:21:42 PM), 'Status' (Completed), 'Duration' (2 minutes 34 seconds), 'Started By' (Shivam Solanki (IBM)), and an 'Action' column with a three-dot menu.

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

The screenshot shows the 'Assets' tab in the 'IBM Cloud Pak for Data' interface. The top navigation bar includes 'My projects / Telco Churn - ss', a search bar, and buttons for 'Add to project' and '+'. Below the navigation, there are tabs for 'Overview', 'Assets' (which is selected and highlighted with a red box), 'Environments', 'Jobs', 'Access Control', and 'Settings'. A search bar is present. The main area displays a list of 'Data assets' with the following details:

Name	Type	Created by	Last modified
Telco Churn Shaped	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 07:24 PM
Telco Churn Customers	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM
Telco Churn Billing	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM
Telco Churn Products	Data Asset	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM
Db2-warehouse-churn-srs	Connection	Shivam Solanki (IBM)	Sep 18, 2020, 06:55 PM

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

10. Hover over the **Telco Churn 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.

11. Click the X to close the preview.

My Projects / Telco Churn - ss / Telco Churn Shaped

Preview Profile Lineage

Schema: 22 Columns  
Preview: 1000 rows

Last refresh: just now Refine

customer... String	gender String	SeniorCit... String	Partner String	Depende... String	tenure String	NATIONA... String	Contract String	PaperlessBi... String	PaymentMe... String	MonthlyCha... String
0002-ORFBO	Female	0	Yes	Yes	9	ca0fb3f5f1e981t	One year	Yes	Mailed check	65.60
0003-MKNFE	Male	0	No	No	9	fdb83f6cd267faf	Month-to-month	No	Mailed check	59.90
0004-TLHLJ	Male	0	No	No	4	8d75b8f384896l	Month-to-month	Yes	Electronic check	73.90
0011-IGKFF	Male	1	Yes	No	13	6aedb9f54fc4e0l	Month-to-month	Yes	Electronic check	98.00
0013-EXCHZ	Female	1	Yes	No	3	64720d472fffb6l	Month-to-month	Yes	Mailed check	83.90
0013-MHZWF	Female	0	No	Yes	9	2c8f4b215b098l	Month-to-month	Yes	Credit card (autom)	69.40
0013-SMEOE	Female	1	Yes	No	71	d3a7386f47668t	Two year	Yes	Bank transfer (auto	109.70
0014-BMAQU	Male	0	Yes	No	63	a51fb1141a9b0t	Two year	Yes	Credit card (autom)	84.65
0015-UOCQJ	Female	1	No	No	7	0325c10bf15b6cl	Month-to-month	Yes	Electronic check	48.20
0016-QLJIS	Female	0	Yes	Yes	65	c8b929250c079t	Two year	Yes	Mailed check	90.45
0017-DINOC	Male	0	No	No	54	5d883fa8b178bf	Two year	No	Credit card (autom)	45.20
0017-IUDMW	Female	0	Yes	Yes	72	d400513048ee5t	Two year	Yes	Credit card (autom)	116.80
0018-NYROU	Female	0	Yes	No	5	de8a5e4502ad	Month-to-month	Yes	Electronic check	68.95
0019-EFAEP	Female	0	No	No	72	32d210dedc614t	Two year	Yes	Bank transfer (auto	101.30
0019-GFNTW	Female	0	No	No	56	fa0dbe32efddc6t	Two year	No	Bank transfer (auto	45.05
0020-INWCK	Female	0	Yes	Yes	71	2d08197f582bd	Two year	Yes	Credit card (autom)	95.75
0020-JDNXP	Female	0	Yes	Yes	34	9a39f969ebbd3t	One year	No	Mailed check	61.25



12. Click on the **Telco Churn** 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.

IBM Cloud Pak for Data

All Search

My Projects / Telco Churn - ss / Telco Churn Shaped

Preview Profile Lineage

Schema: 22 Columns  
Preview: 1000 rows

Last refresh: 1 minute ago Refine

customer... String	gender String	SeniorCit... String	Partner String	Depende... String	tenure String	NATIONA... String	Contract String	PaperlessBi... String	PaymentMe... String	MonthlyCha... String	TotalChar... String	Churn String	PhoneSe... String
0002-ORFBO	Female	0	Yes	Yes	9	ca0fb3f5f1e981t	One year	Yes	Mailed check	65.60	593.30	No	Yes
0003-MKNFE	Male	0	No	No	9	fdb83f6cd267faf	Month-to-month	No	Mailed check	59.90	542.40	No	Yes
0004-TLHLJ	Male	0	No	No	4	8d75b8f384896l	Month-to-month	Yes	Electronic check	73.90	280.85	Yes	Yes
0011-IGKFF	Male	1	Yes	No	13	6aedb9f54fc4e0l	Month-to-month	Yes	Electronic check	98.00	1237.85	Yes	Yes
0013-EXCHZ	Female	1	Yes	No	3	64720d472fffb6l	Month-to-month	Yes	Mailed check	83.90	267.40	Yes	Yes
0013-MHZWF	Female	0	No	Yes	9	2c8f4b215b098l	Month-to-month	Yes	Credit card (autom)	69.40	571.45	No	Yes
0013-SMEOE	Female	1	Yes	No	71	d3a7386f47668t	Two year	Yes	Bank transfer (auto	109.70	7904.25	No	Yes
0014-BMAQU	Male	0	Yes	No	63	a51fb1141a9b0t	Two year	Yes	Credit card (autom)	84.65	5377.80	No	Yes
0015-UOCQJ	Female	1	No	No	7	0325c10bf15b6cl	Month-to-month	Yes	Electronic check	48.20	340.35	No	Yes
0016-QLJIS	Female	0	Yes	Yes	65	c8b929250c079t	Two year	Yes	Mailed check	90.45	5957.90	No	Yes
0017-DINOC	Male	0	No	No	54	5d883fa8b178bf	Two year	No	Credit card (autom)	45.20	2460.55	No	No
0017-IUDMW	Female	0	Yes	Yes	72	d400513048ee5t	Two year	Yes	Credit card (autom)	116.80	8456.75	No	Yes
0018-NYROU	Female	0	Yes	No	5	de8a5e4502ad	Month-to-month	Yes	Electronic check	68.95	351.50	No	Yes
0019-EFAEP	Female	0	No	No	72	32d210dedc614t	Two year	Yes	Bank transfer (auto	101.30	7261.25	No	Yes
0019-GFNTW	Female	0	No	No	56	fa0dbe32efddc6t	Two year	No	Bank transfer (auto	45.05	2560.10	No	No
0020-INWCK	Female	0	Yes	Yes	71	2d08197f582bd	Two year	Yes	Credit card (autom)	95.75	6849.40	No	Yes

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

The screenshot shows the 'Assets' tab selected in the navigation bar. The main area displays a list of data assets, including 'Telco Churn Shaped', 'Telco Churn Customers', 'Telco Churn Billing', 'Telco Churn Products', and 'Db2-warehouse-churn-srs'. Below this, the 'Data Refinery flows' section is shown, containing a single entry: 'Telco Churn Data Flow'. A red box highlights this section.

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

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

The screenshot shows the 'Data Refinery flows' section with the 'Telco Churn Data Flow' entry. To the right of this entry is a small ellipsis (...). A red box highlights the context menu that appears when this ellipsis is clicked, listing options: 'Clone', 'Create job', 'View job', and '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.

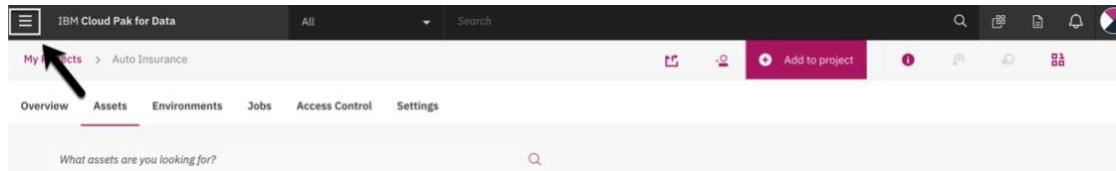
You can also click on the **data flow** to open it in the data flow shaper to modify it and view the data flow’s result set and recipe steps.

## Protect Sensitive Information

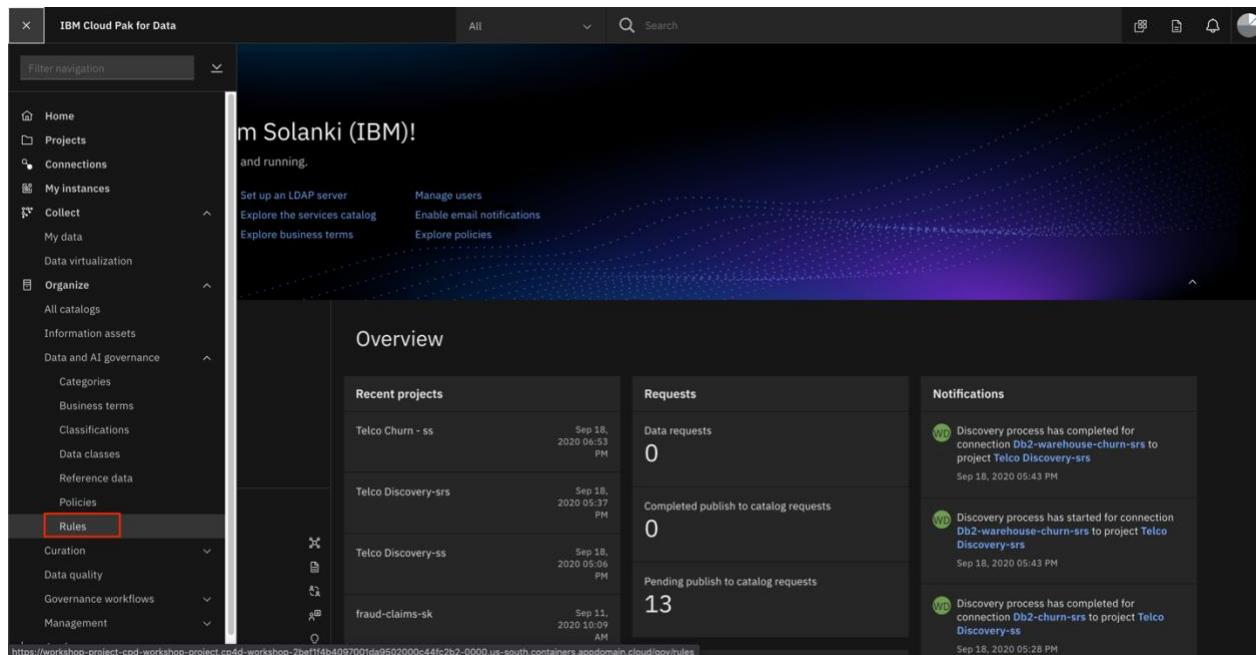
In this section you will learn how to protect sensitive information by creating **Data Protection Rules**. You will create data protection rules to obfuscate (i.e. Mask) **US Social Security Numbers** and redact **Credit Card Information** and then validate that they are being enforced.

### Create Data Protection Rules

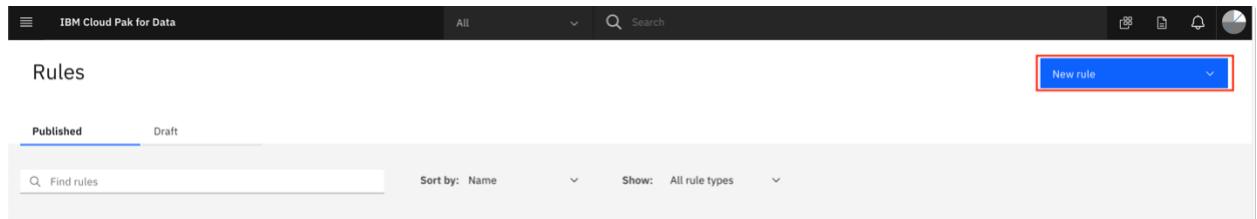
1. Click the **IBM Cloud Pak for Data** navigation menu in the top left corner.



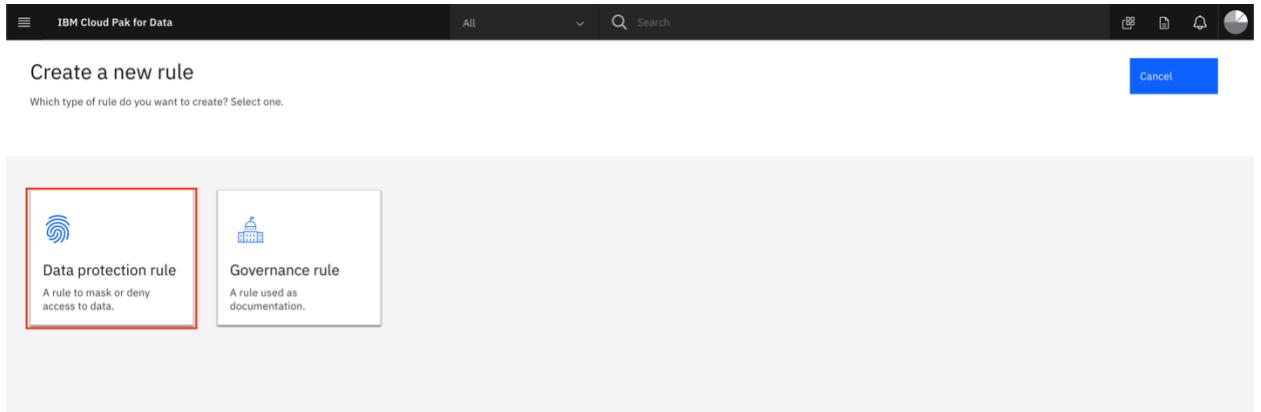
Click Organize > Data and AI Governance > Rules from the menu.



2. Click the **New rule** dropdown button and select Create rule.



3. Click the Data protection rule type.



4. Enter a Name of Protect Credit Card Information - initials.

A screenshot of the 'New data protection rule' configuration screen. On the left, the 'Rule Details' section shows 'Name\*' set to 'Protect Credit Card Information - srs' and 'Type\*' set to 'Access'. On the right, the 'Rule builder' section shows 'Criteria \*' with 'Condition 1' set to 'If Data class contains any cred'. Below that, 'Action' is set to 'Credit Card Number'. A blue bar highlights the 'Credit Card Number' action.

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 search **credit card exp.**

**Rule Details**

**Name\***: Protect Credit Card Information - srs

**Type\***: Access

**Business definition\***: Protect all credit card numbers, expiration dates and validation numbers using the data redaction method

**Rule builder**

**Criteria \***: Condition 1

If Data class contains any

- Credit Card Number X cred
- Credit Card Expiration Date**: A date value in format month/year representing a credit card expiration date.
- Credit Card Number
- Credit Card Validation Number

Action

the

A 3 or 4 digits number representing a credit card validation number.

10. Select the **Credit Card Expiration Date** 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.

New data protection rule

**Rule Details**

**Name\***: Protect Credit Card Information - srs

**Type\***: Access

**Business definition\***: Protect all credit card numbers, expiration dates and validation numbers using the data redaction method

**Rule builder**

**Criteria \***: Condition 1

If Data class contains any

- Credit Card Number X Credit Card Expiration Date X

Add new condition +

**Action \***: Action 1

then mask data

where in columns containing

- Credit Card Number X

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

16. Select the **Credit Card Expiration Date** data class from the list.

The screenshot shows the 'Rule builder' interface. On the left, under 'Rule Details', there is a 'Name\*' field containing 'Protect Credit Card Information - srs'. Below it, a 'Type\*' field is set to 'Access'. Under 'Business definition\*', there is a note: 'Protect all credit card numbers, expiration dates and validation numbers using the data redaction method'. On the right, under 'Criteria \*', a condition for 'Data class' is defined, selecting 'Credit Card Number' and 'Credit Card Expiration Date'. Under 'Action \*', the action is 'mask data', targeting 'Credit Card Number' and 'Credit Card Expiration Date'. The 'Redact' option is selected, showing a preview of 'Before 452-821-1120' being replaced by 'Replace data with Xs'. Other options like 'Substitute' and 'Obfuscate' are also shown.

17. Click **Create**.

The screenshot shows the 'New data protection rule' dialog. It has fields for 'Name\*' (Protect Credit Card Information - srs), 'Type\*' (Access), and 'Business definition\*' (Protect all credit card numbers, expiration dates and validation numbers using the data redaction method). The 'Action \*' section is identical to the one in the previous screenshot, with 'Redact' selected and a preview of the masked data. The 'Create rule' button is highlighted with a red box.

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

The screenshot shows the 'Criteria' and 'Action' sections of a rule configuration. In the Criteria section, there is a condition: 'If Data class contains any Credit Card Number, Credit Card Expiration Date'. In the Action section, there is an action: 'Then Redact data in columns containing Credit Card Number, Credit Card Expiration Date'. At the top right, there are 'Edit rule' and 'Delete rule' buttons. The 'Rules' breadcrumb is highlighted with a red box at the top left.

19. Click the **New rule** dropdown and select **Create new rule** button.

The screenshot shows the 'Rules' section of the IBM Watson Knowledge Catalog. A dropdown menu is open, with 'New rule' highlighted with a red box. Other options in the menu include 'Create new rule', 'Import from file', and 'Extract from file'.

20. Click the Data protection rule type.

The screenshot shows the 'Create a new rule' dialog. It asks 'Which type of rule do you want to create? Select one.' Two options are shown: 'Data protection rule' (selected) and 'Governance rule'. The 'Data protection rule' option is highlighted with a red box.

21. Enter a Name of Protect US Social Security Numbers - initials.

New data protection rule

**Rule Details**

Name\*  
Protect US Social Security Numbers - srs

Type\*  
Access

Business definition\* ⓘ  
Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values

**Rule builder**

Criteria \* ⓘ

Condition 1

If Data class contains any Social

Canadian Social Insurance Number (SIN)  
A social insurance number (SIN) is a number issued in Canada to administer various government programs.

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.

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 **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**.

New data protection rule

**Rule Details**

Name\*  
Protect US Social Security Numbers - srs

Type\*  
Access

Business definition\* ⓘ  
Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values

**Rule builder**

Criteria \* ⓘ

Condition 1

If Data class contains any US Social Security Number

Add new condition ⓘ

Action \* ⓘ

then

- deny access to data
- mask data

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

The screenshot shows the 'Rule builder' interface in the IBM Watson Knowledge Catalog. On the left, under 'Rule Details', there is a 'Name' field set to 'Protect US Social Security Numbers - srs'. Below it, a 'Type' dropdown is set to 'Access'. Under 'Business definition', there is a note: 'Protect all US Social Security Numbers using the data masking obfuscation method replacing data with similarly formatted but fictional values'. On the right, under 'Rule builder', the 'Criteria' section has a condition 'Data class' set to 'US Social Security Number'. The 'Action' section shows 'mask data' and 'where in columns containing'.

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

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

The screenshot shows the 'Rule builder' interface in the IBM Watson Knowledge Catalog. The 'Action' section shows 'mask data' and 'where in columns containing'. A search bar above the list is set to 'US Social Security Number'. Below the search bar, a list of data classes is shown, with 'us social' highlighted. A detailed description of 'US Social Security Number' is visible, stating: '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.' Three options for masking are listed: 'US Social Security Number Last 4' (the last four digits of a United States Social Security Number (SSN)), '452-821-1120'; 'Replace data with Xs'; 'Replace data with values that don't match the original format'; and 'Replace data with similarly formatted values'.

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

New data protection rule

Rule Details

Name\* Protect US Social Security Numbers - srs

Type\* Access

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

Action \* then mask data in columns containing

Select how to mask data:

- Redact** Before 452-821-1120 Replace data with Xs
- Substitute** Before 452-821-1120 Replace data with values that don't match the original format
- Obfuscate** Before 452-821-1120 Replace data with similarly formatted values

31. Click **Create**.

IBM Watson Knowledge Catalog

New data protection rule

Create rule

Cancel

Rule Details

Name\* Protect US Social Security Numbers - srs

Type\* Access

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

Action \* then mask data in columns containing

Select how to mask data:

- Redact** Before 452-821-1120 Replace data with Xs
- Substitute** Before 452-821-1120 Replace data with values that don't match the original format
- Obfuscate** Before 452-821-1120 Replace data with similarly formatted values

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

The screenshot shows the 'Rules' page with a single rule named 'Protect US Social Security Numbers - srs'. The rule has one condition and one action. The condition states 'If Data class contains any US Social Security Number'. The action states 'Then Obfuscate data in columns containing US Social Security Number'. There are 'Edit rule' and 'Delete rule' buttons at the top right.

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.

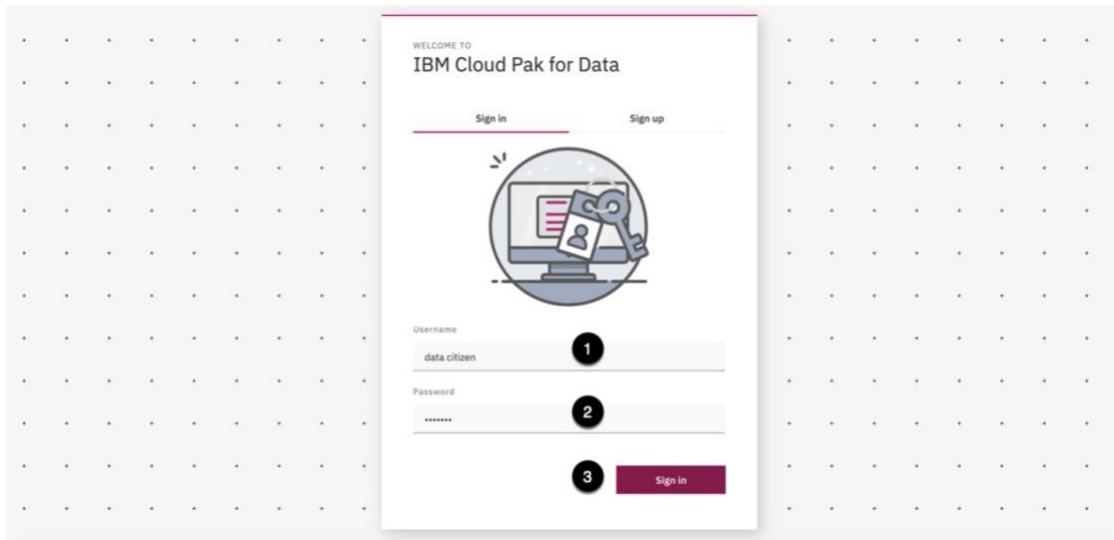
The screenshot shows the IBM Watson Knowledge Catalog interface. In the top right corner, there is a user profile icon for 'Shivam Solanki...' with options for 'Profile and settings', 'About', and 'Log out'. A red box highlights the 'Log out' button. Below the header, the 'Published' tab is selected in the 'Rules' section, showing two active rules: 'Protect US Social Security Numbers - srs' and 'Protect Credit Card Information - srs'. Both rules are described as active and last modified on Sep 21, 2020.

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 **telco churn** in the search area and press the enter key.

A screenshot of the IBM Cloud Pak for Data overview page. The top navigation bar includes a "Sign in" button, a search bar containing "telco churn" (which is highlighted with a red box), and other icons. The main area starts with a "Welcome, data citizen!" message and a "Use the following links to get up and running." section. It features four main cards: "Recent projects" (No recent projects), "Requests" (Data requests: 0), "Notifications" (No notifications), and "Overview" (My instances, My data, Data virtualization, Catalogs). On the left, there's a sidebar with "Quick navigation" (Projects, My instances, My data, Data virtualization, Catalogs) and "Resources" (What is IBM Cloud Pak for Data?, Documentation).

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

The screenshot shows the search results for "telco churn". The results table has columns: Name, Type, Tags, Modified by, and Modified on. There are six items listed:

- customer-churn-predicton-ml**: Data asset, Telco Churn, Document, System, Sep 18, 2020.
- Db2-warehouse-churn-srs**: Connection, Telco Churn, Warehouse, Shivam Solanki (IBM), Sep 18, 2020.
- Telco Churn Billing**: Data asset, Telco Churn, System, Sep 18, 2020.
- Telco Churn Customers**: Data asset, Telco Churn, Shivam Solanki (IBM), Sep 18, 2020. This item is highlighted with a red box.
- churn-prediction-research**: Data asset, Telco Churn, Document, System, Sep 18, 2020.
- Telco Churn Products**: Data asset, Telco Churn, System, Sep 18, 2020.

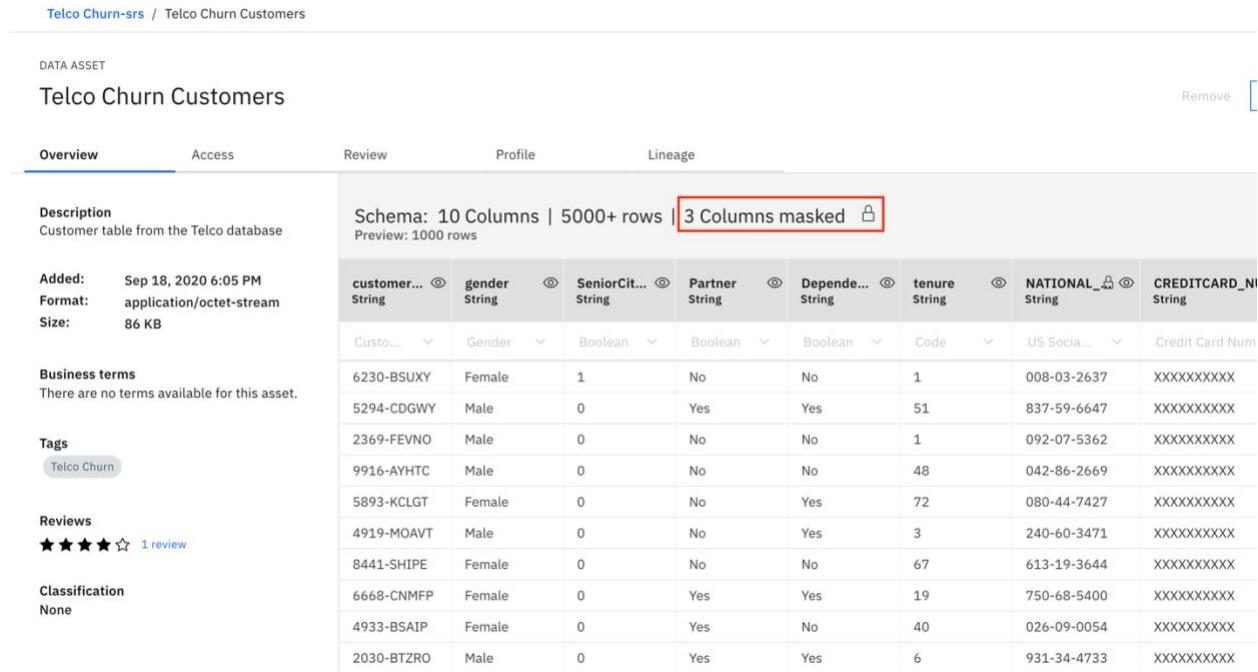
You should immediately see the message, **“Data Masking in progress”**, with a spinning progress wheel. It will take a minute to load so be patient and let it finish.

The screenshot shows the details for the Telco Churn Customers data asset. The schema is listed as 10 columns, 5000+ rows, and 3 columns masked. A message box says "Data masking in progress" and "This asset is being masked by the data enforcement rule: Protect Credit Card Information - srs. 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.

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

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

6. Click the **Refresh** button under the lock icon in the middle of the page. If that does not work, **refresh** the page using your browser's refresh method.
7. Click on the lock icon next to "3 columns masked"



Schema: 10 Columns   5000+ rows   3 Columns masked										
Preview: 1000 rows										
customer...	gender	SeniorCit...	Partner	Depende...	tenure	NATIONAL_	CREDITCARD_N			
String	String	String	String	String	String	String	String			
Customer ID	Gender	Senior Citizen	Partner Count	Dependents	Tenure	National ID	Credit Card Number			
6230-BSUXY	Female	1	No	No	1	008-03-2637	XXXXXXXXXX			
5294-CDGKY	Male	0	Yes	Yes	51	837-59-6647	XXXXXXXXXX			
2369-FEVNO	Male	0	No	No	1	092-07-5362	XXXXXXXXXX			
9916-AYHTC	Male	0	No	No	48	042-86-2669	XXXXXXXXXX			
5893-KCLGT	Female	0	No	Yes	72	080-44-7427	XXXXXXXXXX			
4919-MOAVT	Male	0	No	Yes	3	240-60-3471	XXXXXXXXXX			
8441-SHPE	Female	0	No	No	67	613-19-3644	XXXXXXXXXX			
6668-CNMFP	Female	0	Yes	Yes	19	750-68-5400	XXXXXXXXXX			
4933-BSAIP	Female	0	Yes	No	40	026-09-0054	XXXXXXXXXX			
2030-BTZRO	Male	0	Yes	Yes	6	931-34-4733	XXXXXXXXXX			

**Notice** that 2 columns are redacted, and one is obfuscated.

8. Scroll to the right until you see the **national\_id** column.

Telco Churn-srs / Telco Churn Customers

DATA ASSET

**Telco Churn Customers**

Remove Download Add to Project

Overview Access Review Profile Lineage

**Description**  
Customer table from the Telco database

**Added:** Sep 18, 2020 6:05 PM  
**Format:** application/octet-stream  
**Size:** 86 KB

**Business terms**  
There are no terms available for this asset.

**Tags**  
Telco Churn

**Reviews**  
★ ★ ★ ★ ☆ 1 review

**Classification**  
None

Schema: 10 Columns | 5000+ rows | 3 Columns masked

Last refresh: 1 minute ago

Preview: 1000 rows

NATIONAL\_ID CREDITCARD\_NUMBER CREDITCARD\_EXP

rf...	gender	SeniorCit...	Partner	Depende...	tenure	NATIONAL_ID	CREDITCARD_NUMB...	CREDITCARD_...	CREDITCARD_E...
	String	String	String	String	String	String	String	String	String
UXY	Female	1	No	No	1	008-03-2637	XXXXXXXXXX	Discover	XXXXXXX
IGWY	Male	0	Yes	Yes	51	837-59-6647	XXXXXXXXXX	Diners Club	XXXXXXX
VNO	Male	0	No	No	1	092-07-5362	XXXXXXXXXX	American Express	XXXXXXX
HTC	Male	0	No	No	48	042-86-2669	XXXXXXXXXX	Master Card	XXXXXXX
LGT	Female	0	No	Yes	72	080-44-7427	XXXXXXXXXX	VISA	XXXXXXX
JAVT	Male	0	No	Yes	3	240-60-3471	XXXXXXXXXX	JCB	XXXXXXX
IPE	Female	0	No	No	67	613-19-3644	XXXXXXXXXX	Discover	XXXXXXX
IMFP	Female	0	Yes	Yes	19	750-68-5400	XXXXXXXXXX	American Express	XXXXXXX
AIP	Female	0	Yes	No	40	026-09-0054	XXXXXXXXXX	Diners Club	XXXXXXX
ZRO	Male	0	Yes	Yes	6	931-34-4733	XXXXXXXXXX	American Express	XXXXXXX
CBS	Male	0	No	No	15	447-76-3056	XXXXXXXXXX	Discover	XXXXXXX
QVO	Male	0	No	No	1	901-25-2161	XXXXXXXXXX	VISA	XXXXXXX

Notice that the **NATIONAL\_ID**, **CREDITCARD\_NUMBER**, AND **CREDITCARD\_EXP** columns have a lock icon next to their name indicating that the data is being protected.

## 9. Click on the **lock** icon next to the **NATIONAL\_ID** column.

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

Telco Churn-srs / Telco Churn Customers

DATA ASSET

**Telco Churn Customers**

Remove Download

Profile and settings

About Community Support Log out

Overview Access Review Profile Lineage

**Description**  
Customer table from the Telco database

**Added:** Sep 18, 2020 6:05 PM  
**Format:** application/octet-stream  
**Size:** 86 KB

**Business terms**  
There are no terms available for this asset.

**Tags**  
Telco Churn

**Reviews**  
★ ★ ★ ★ ☆ 1 review

**Classification**  
None

Schema: 10 Columns | 5000+ rows | 3 Columns masked

Preview: 1000 rows

NATIONAL\_ID CREDITCARD\_NUMBER CREDITCARD\_EXP

rf...	gender	SeniorCit...	Partner	Depende...	tenure	NATIONAL_ID	CREDITCARD_NUMB...	CREDITCARD_...	CREDITCARD_E...
	String	String	String	String	String	String	String	String	String
UXY	Female	1	No	No	1	008-03-2637	XXXXXXXXXX	Discover	XXXXXXX
IGWY	Male	0	Yes	Yes	51	837-59-6647	XXXXXXXXXX	Diners Club	XXXXXXX
VNO	Male	0	No	No	1	092-07-5362	XXXXXXXXXX	American Express	XXXXXXX
HTC	Male	0	No	No	48	042-86-2669	XXXXXXXXXX	Master Card	XXXXXXX
LGT	Female	0	No	Yes	72	080-44-7427	XXXXXXXXXX	VISA	XXXXXXX
JAVT	Male	0	No	Yes	3	240-60-3471	XXXXXXXXXX	JCB	XXXXXXX
IPE	Female	0	No	No	67	613-19-3644	XXXXXXXXXX	Discover	XXXXXXX
IMFP	Female	0	Yes	Yes	19	750-68-5400	XXXXXXXXXX	American Express	XXXXXXX
AIP	Female	0	Yes	No	40	026-09-0054	XXXXXXXXXX	Diners Club	XXXXXXX
ZRO	Male	0	Yes	Yes	6	931-34-4733	XXXXXXXXXX	American Express	XXXXXXX
CBS	Male	0	No	No	15	447-76-3056	XXXXXXXXXX	Discover	XXXXXXX

## 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.