

Lab Guide

Hands-on-Lab Virtualizing Db2 Warehouse data with data virtualization

Shivam R Solanki
Data Scientist
Shivam.raj.solanki@ibm.com



For decades, companies have tried to break down silos by copying data from different operational systems into central data stores for analysis, such as data marts, data warehouses, and data lakes. This is often costly and prone to error. Most struggle to manage an average of 33 unique data sources, which are diverse in structure and type, and are often trapped in data silos that are hard to find and access.

With *data virtualization*, you can query data across many systems without having to copy and replicate data, which reduces costs. It also can simplify your analytics and make them more up to date and accurate because you're querying the latest data at its source.

In this tutorial, we're going to learn how to virtualize Db2® Warehouse data with data virtualization on IBM Cloud Pak for Data to make queries across multiple data sources.

Learning objectives

In this tutorial, you will learn how to:

- [Setup the project on IBM Cloud Pak for Data](#)
- [Add datasets to IBM Cloud Pak® for Data.](#)
- [Add a data source for data virtualization.](#)
- [Virtualize the data and create a joined view.](#)
- [Assign virtualized data to a project.](#)
- [Add roles to users and perform admin tasks.](#)

Steps

Step 1. About the dataset

The dataset used for this tutorial contains information about customer churn for a telecommunications company. The data was collected in three CSV files.

1. *billing.csv*

Some of the important attributes in this file are

- Customer ID

- Contract (Month-to-month, one year, two year)
- Paperless Billing (Yes, No)
- Payment Method (Bank transfer, Credit card, Electronic check, Mailed check)
- Monthly Charges (\$)
- Total Charges (\$)
- Churn (Yes, No)

2. *Customer-service.csv*

Some of the important attributes in this file are

- Customer ID
- Gender (Male, Female)
- Senior Citizen (1, 0)
- Partner (Yes, No)
- Dependents (Yes, No)
- Tenure (1-100)

3. *products.csv*

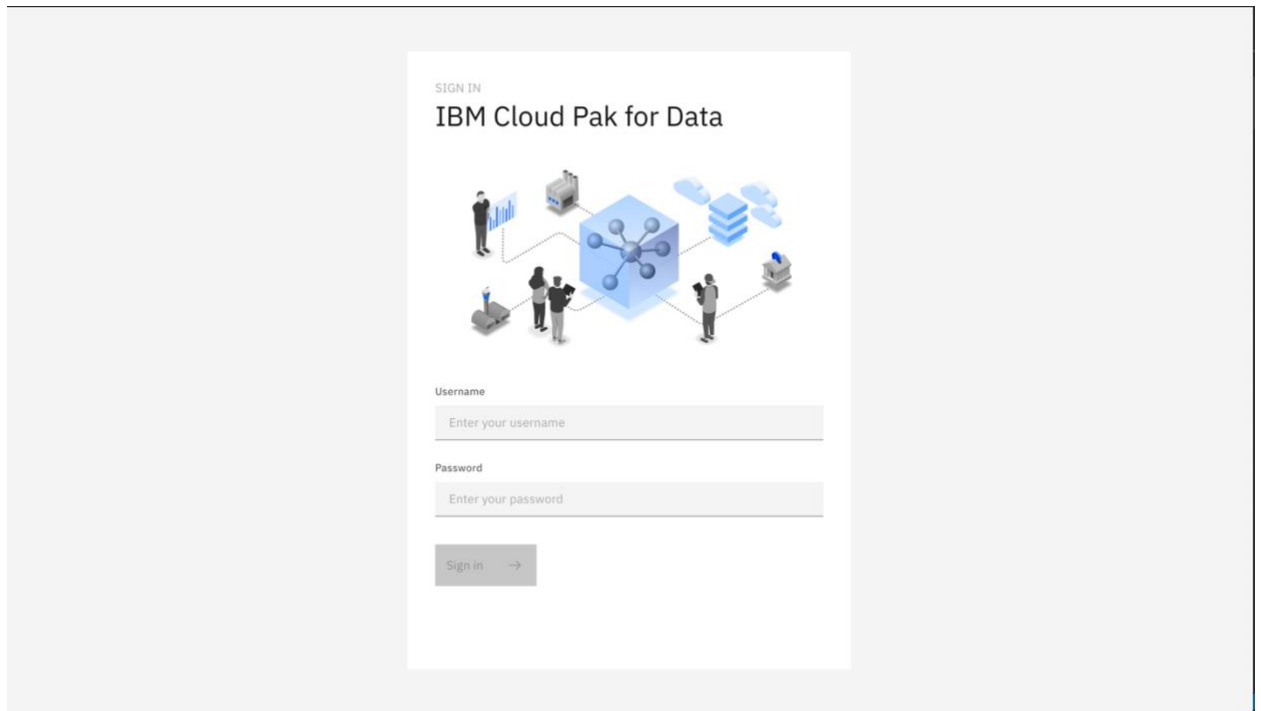
Some of the important attributes in this file are

- Customer ID
- Phone Service (Yes, No)
- Multiple Lines (Yes, No, No phone service)
- Internet Service (DSL, Fiber optic, No)
- Online Security (Yes, No, No internet service)
- Online Backup (Yes, No, No internet service)
- Device Protection (Yes, No, No internet service)
- Tech Support (Yes, No, No internet service)
- Streaming TV (Yes, No, No internet service)
- Streaming Movies (Yes, No, No internet service)

Step 2. Set up the project on IBM Cloud Pak for Data

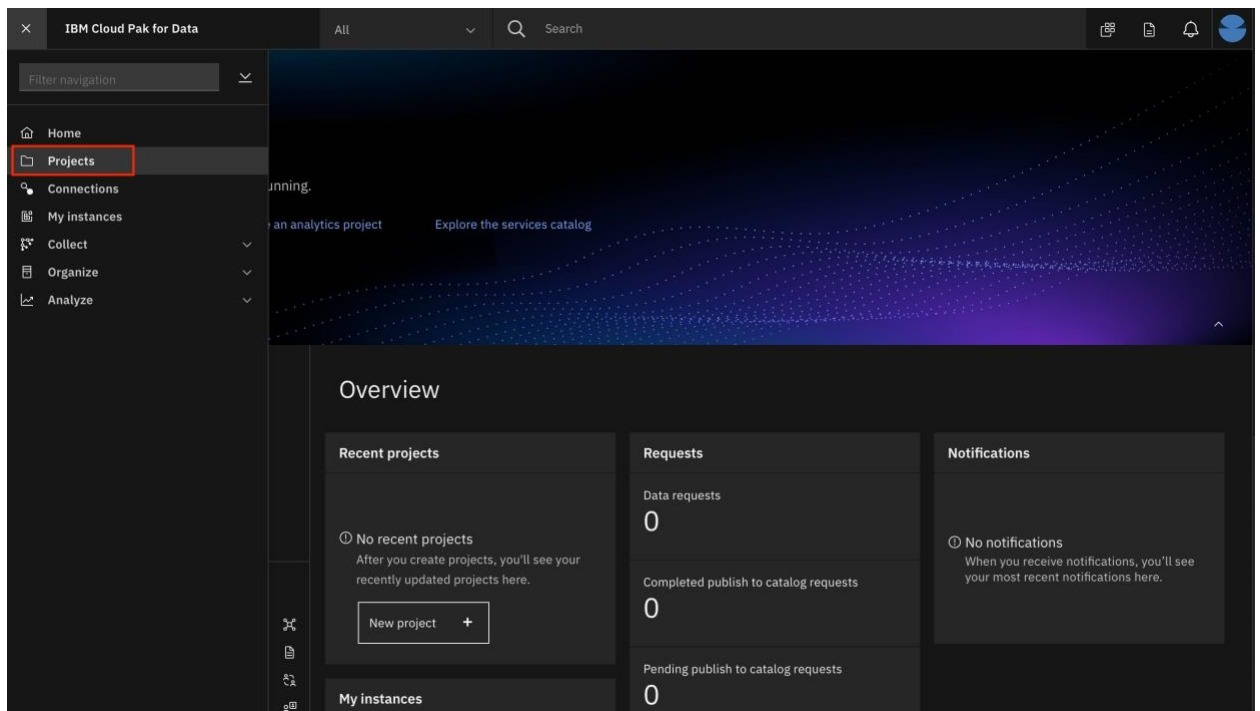
Log in to IBM Cloud Pak for Data

1. Launch a browser and navigate to the IBM Cloud Pak for Data url.

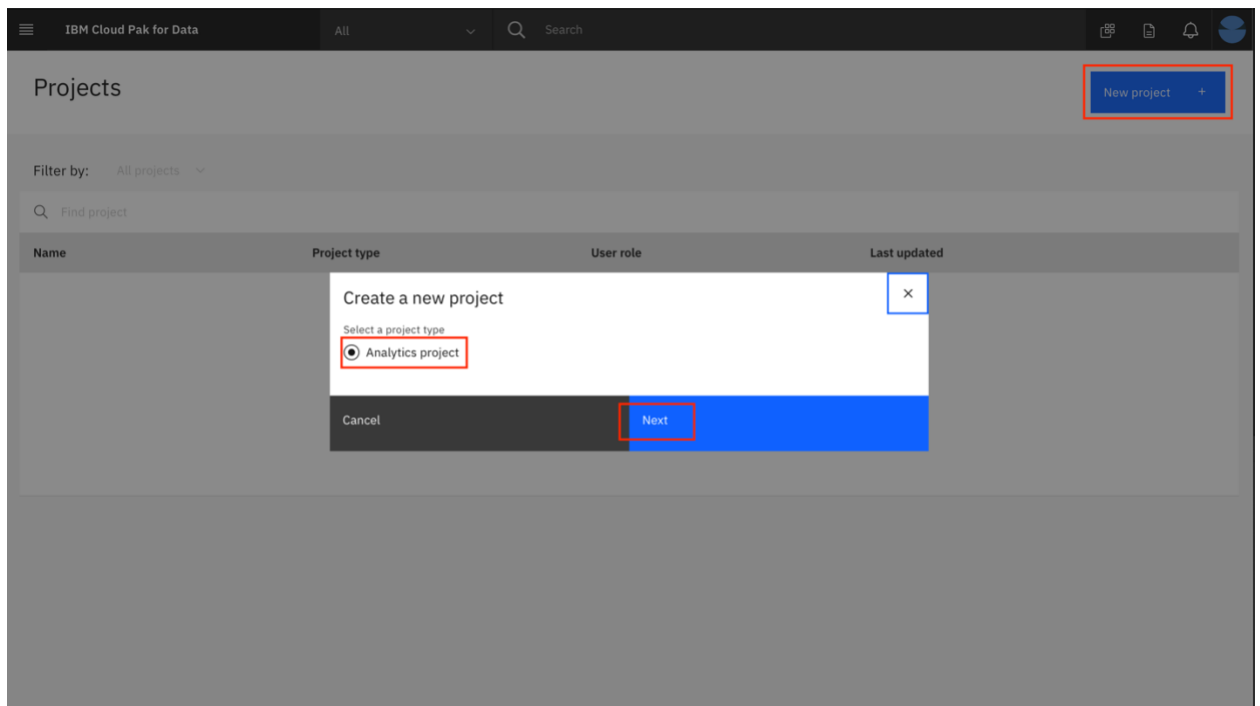


Create a new IBM Cloud Pak for Data project

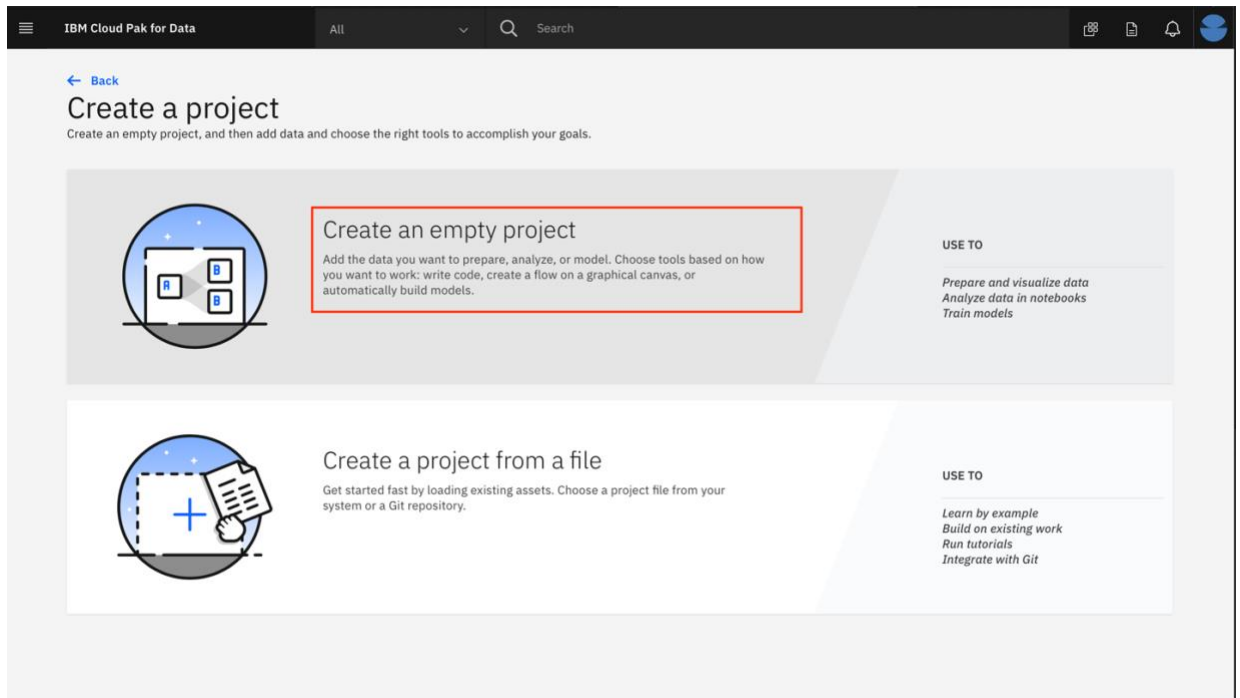
1. Go the hamburger (≡) menu and click **Projects**.



2. Click on **New project**.



3. Select **Create an empty project**.



4. Provide a name and optional description for the project and click **Create**.
(Note – For the purpose of this lab, please add your initials at the end of the project name to avoid conflict since all the participants are working in the same environment.
For ex: Telco-Churn-srs)

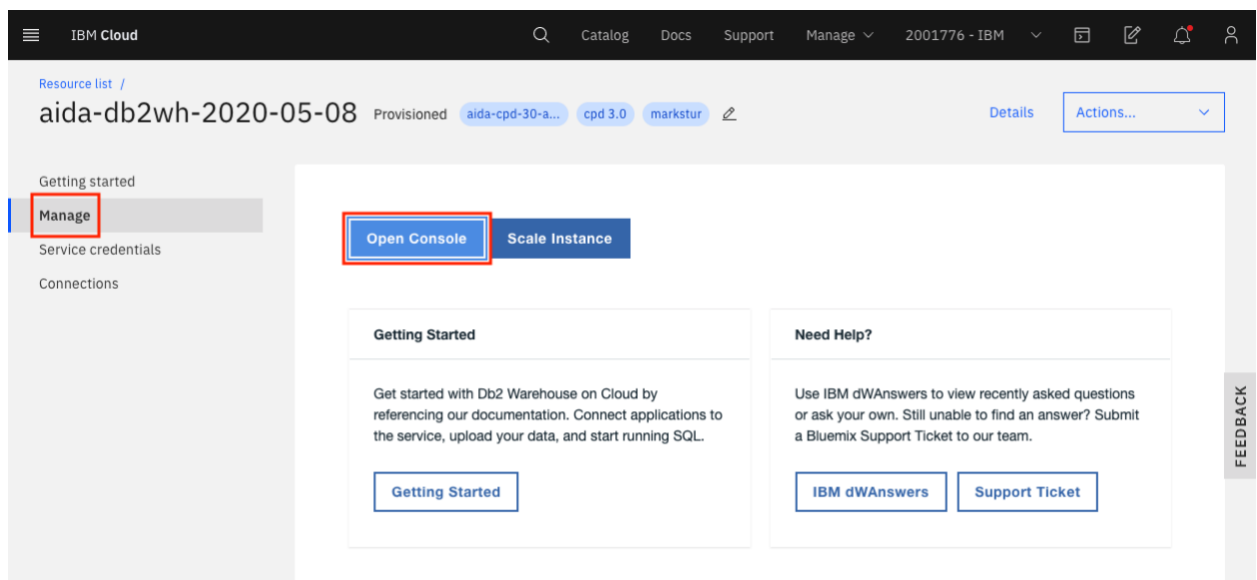
Step 3. Set up the Db2 Warehouse on IBM Cloud

We'll need a place to store our data. It is suggested to use [Db2 Warehouse on IBM Cloud](#) in order to conserve resources on the CPD cluster.

Note that IBM Cloud Pak for Data can work with any database with a JDBC connector, so Db2 warehouse is only one of many choices.

Steps below are shown for fetching connection details for Db2 Warehouse on IBM Cloud.
(Note - This step has already been done for you so you can use the credentials provided by the instructor. The screenshots shown here are for reference. You can skip to the Step 4 now)

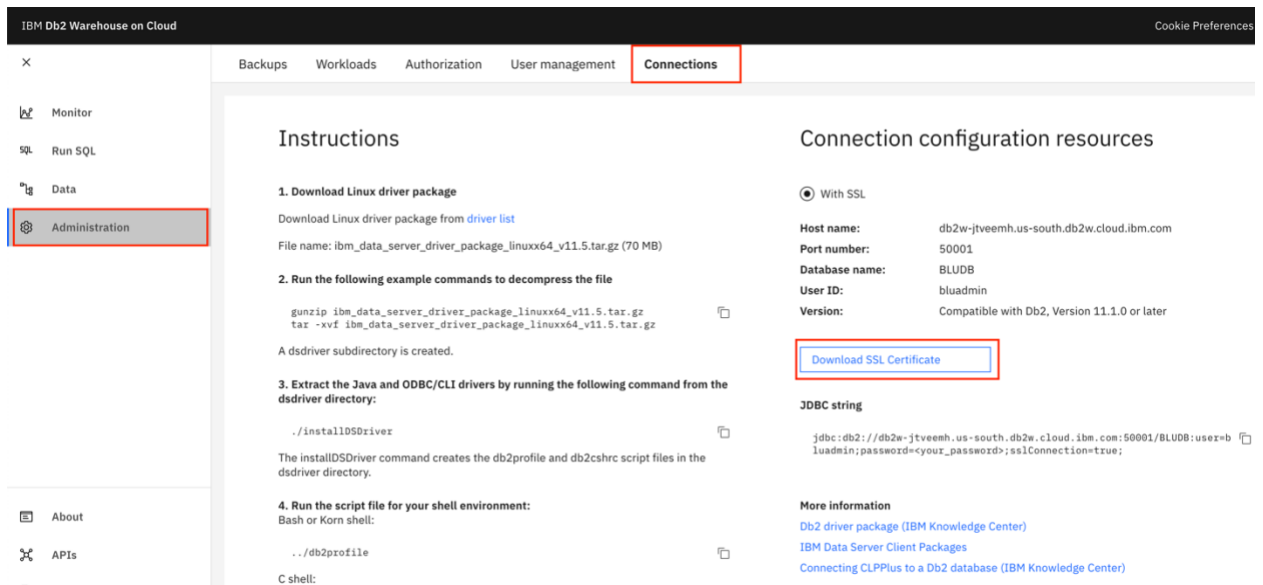
1. Go to Service Credentials and click **New credential +**. Click the **Copy to clipboard** icon and save the credentials for later.
2. Now go to Manage and click **Open Console**.



Get SSL certificate for Db2 Warehouse on IBM Cloud

You will need an SSL cert for IBM Cloud Pak for Data to use the IBM Cloud Db2 Warehouse instance.

1. In the Db2 Warehouse console, from the upper-left hamburger (☰) menu, click **Administration** → **Connections**, then **Download SSL Certificate**.



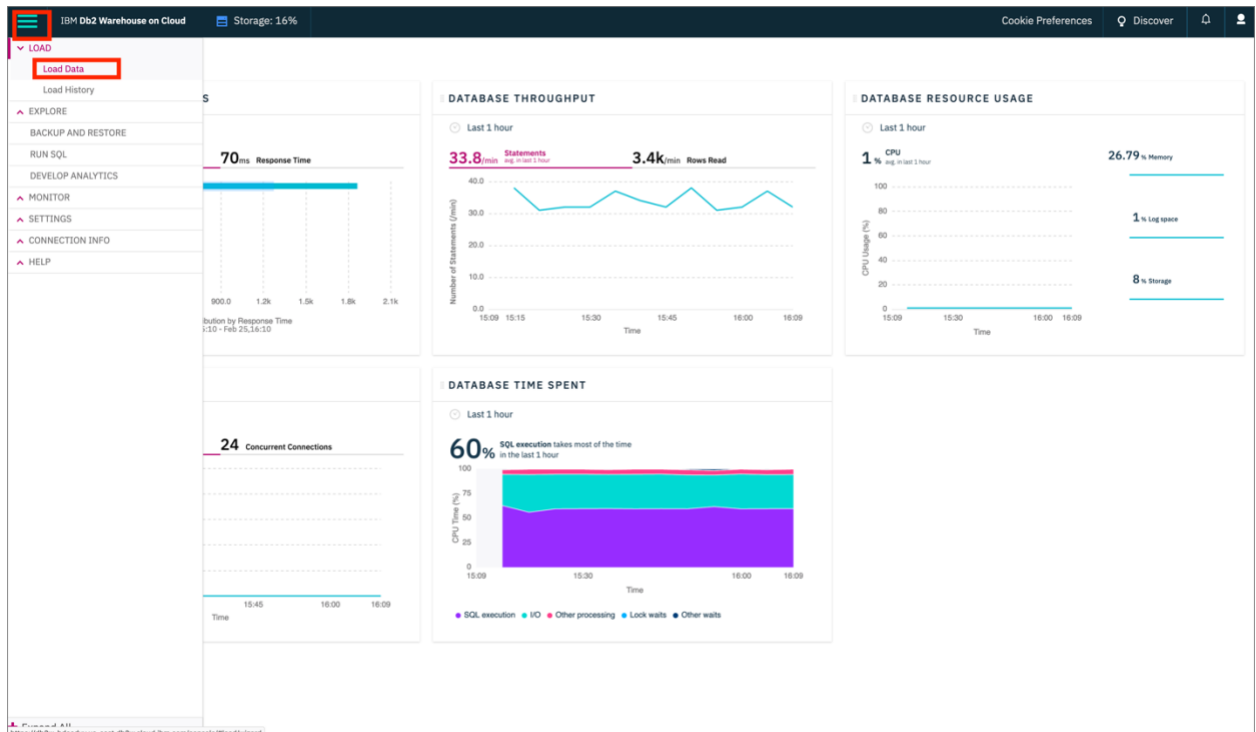
2. You will need to convert the SSL certificate from .crt to a .pem file using [OpenSSL](#). Run the following command:

```
bash
openssl x509 -in DigiCertGlobalRootCA.crt -out DigiCertGlobalRootCA.pem -outform PEM -inform DER
```

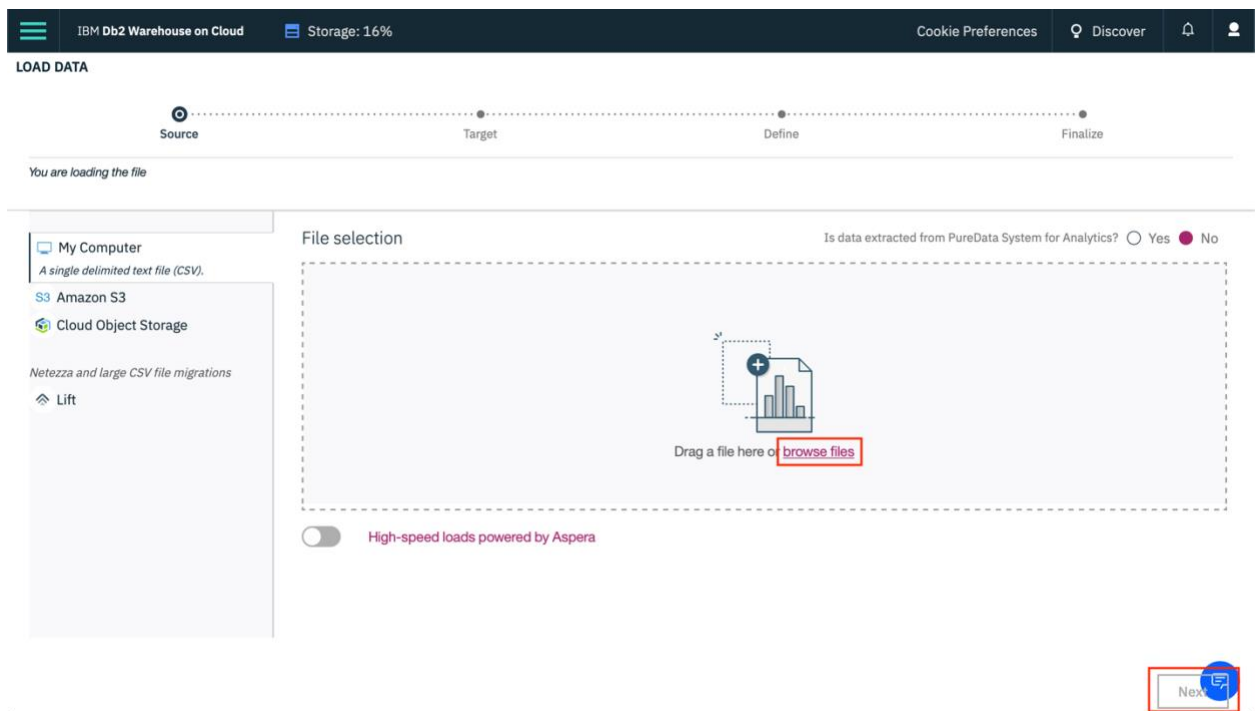
Seed the Db2 Warehouse on IBM Cloud

Steps below are shown for uploading data to Db2 Warehouse on IBM Cloud. (Note - The 3 data files *claims.csv*, *customer.csv* and *policies.csv* have been uploaded to Db2 Warehouse on IBM Cloud for you so you can skip to the Step 4 now)

1. From the upper-left hamburger (☰) menu, click **LOAD** → **Load data**.



2. Click **browse files** and select the [billing.csv](#) file after downloading it from the link, then click **Next**.



- Choose Schema **TELCO** and click **+ New table**. Under Create a new Table, provide BILLING as the name of the table and click **Create**, then **Next**.

Load Data | Load History | Tables | Views | Aliases | MQTs | Schemas | Sequences | Application objects

Source (selected) | Target | Define | Finalize

You are loading the file *billing.csv*

Select a load target

Schema: TELCO (selected)

Table: No entries found.

Create a new table: BILLING (entered), Create (button)

Back | Next (button)

- Accept the defaults and click **Next**. On the next screen, click **Begin Load**.

Load Data | Load History | Tables | Views | Aliases | MQTs | Schemas | Sequences | Application objects

Source (selected) | Target (selected) | Define | Finalize

You are loading the file *billing.csv* into **TELCO.BILLING**

Code page (character encoding): 1208 (UTF-8) | Separator: | Header in first row: | Time & date format: | Detect data types: |

	CUSTOMERID VARCHAR(10)	CONTRACT VARCHAR(14)	PAPERLESSBILLING VARCHAR(3)	PAYMENTMETHOD VARCHAR(25)	MONTHLYCHARGES DECIMAL(7,2)	TOTALCHARGES DECIMAL(8,2)
1	7590-VHVEG	Month-to-month	Yes	Electronic check	29.85	29.85
2	5575-GNVDE	One year	No	Mailed check	56.95	1889.5
3	3668-QPYBK	Month-to-month	Yes	Mailed check	53.85	108.15
4	7795-CFOCW	One year	No	Bank transfer (automatic)	42.3	1840.75
5	9237-HQITU	Month-to-month	Yes	Electronic check	70.7	151.65
6	9305-CDSKC	Month-to-month	Yes	Electronic check	99.65	820.5
7	1452-KIOVK	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4
8	6713-OKOMC	Month-to-month	No	Mailed check	29.75	301.9
9	7892-POOKP	Month-to-month	Yes	Electronic check	104.8	3046.05
10	6388-TABGU	One year	No	Bank transfer (automatic)	56.15	3487.95

Back | Next (button)

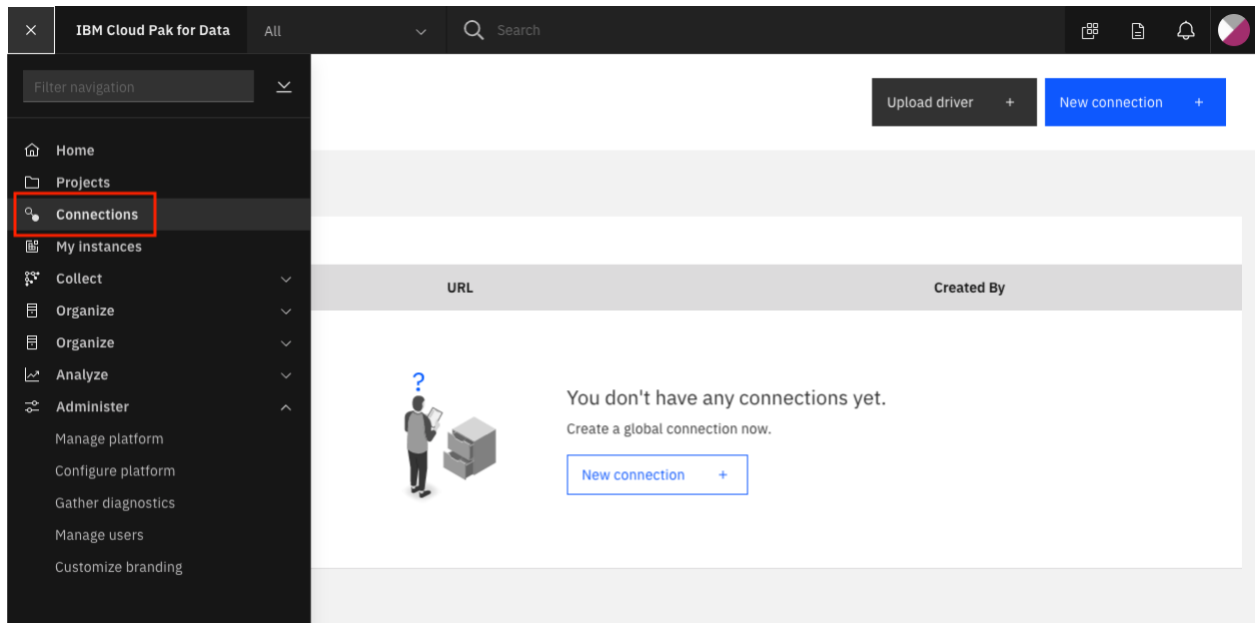
- Repeat for the [products.csv](#) file, naming the table `PRODUCTS` and the [customer-service.csv](#) table `CUSTOMERS`.

Now the Db2 warehouse has been set up on IBM Cloud.

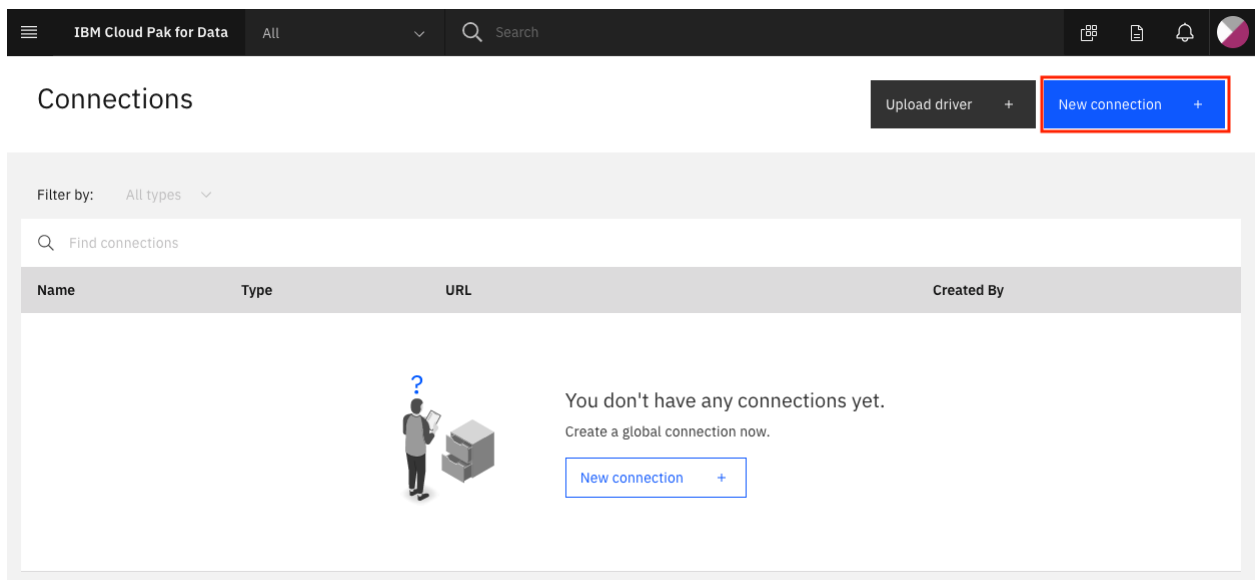
Step 4. Add a new data source connection

Add the new data source

1. To add a new data source, go to the hamburger (≡) menu and click on the **Connections** option.



2. At the overview, click **New connection +**.



Enter connection details for Db2 Warehouse on IBM Cloud

1. Start by giving your new connection a name, and select **Db2 Warehouse on Cloud** as your connection type. More fields should appear. Fill in the new fields with the credentials for your Db2 Warehouse connection. Click the checkbox for **Use SSL**.

(Note – Please add your initials at the end of the connection name with a dash. For example, db2-warehouse-srs in order to avoid conflict)

2. Enter the following service credentials to create a connection to the Db2 warehouse on IBM Cloud.

"hostname": "db2w-jtveemh.us-south.db2w.cloud.ibm.com",
"password": "GxdualWj_f4aZq82Ah@4@KqmHAKeF",
"port": 50001,
"db": "BLUDB",
"username": "bluadmin",

3. Download the [SSL certificate](#) and upload it to create connection.

IBM Cloud Pak for Data

Success The test connection was successful. Click Create to save the connection information.

New connection

Connection name: db2-warehouse-srs

Description (optional): DB2 Warehouse on IBM Cloud

Connection type: Db2 Warehouse on Cloud

Host: db2w-jtveemh.us-south.db2w.cloud.ibm.com

Port: 50001

Database: BLUDB

Username: bluadmin

Password: [masked]

Connection parameters: sslConnection=true;sslTrustStoreLocation=/user-home/_global_/security/customer-tr

JDBC URL: jdbc:db2://db2w-jtveemh.us-south.db2w.cloud.ibm.com:50001/BLUDB:sslConnection=...

☒ Use SSL

Verify SSL certificate

SSL certificate (optional): Upload your SSL certificate file from your local file system.

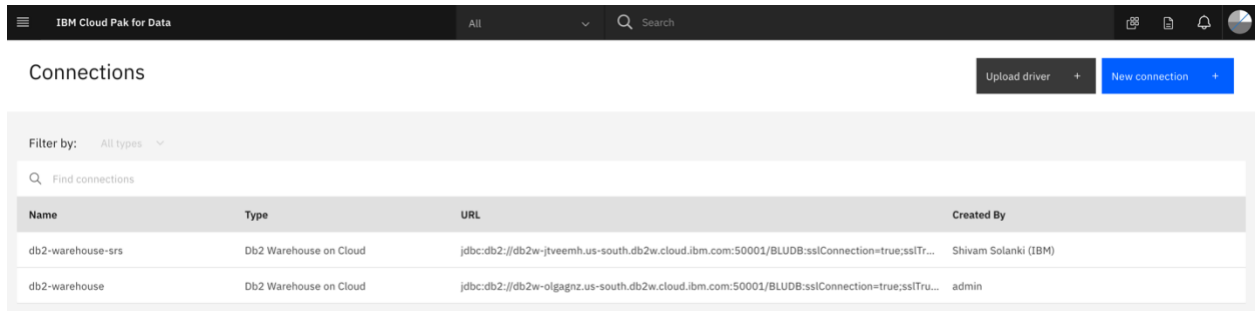
Drag your SSL certificate files here or browse for the files to upload

DigCertGlobalRootCA.pem

Certificate server host name (optional): Enter server host name

Cancel Test connection Create

The new connection will be listed in the overview.



The screenshot shows the 'Connections' page in IBM Cloud Pak for Data. At the top, there's a header with 'IBM Cloud Pak for Data', a dropdown menu set to 'All', and a search bar. Below the header, there are buttons for 'Upload driver' and 'New connection'. A filter section shows 'Filter by: All types' and a search input 'Find connections'. The main content is a table with columns: Name, Type, URL, and Created By. Two connections are listed: 'db2-warehouse-srs' and 'db2-warehouse', both of type 'Db2 Warehouse on Cloud'. The first connection was created by 'Shivam Solanki (IBM)' and the second by 'admin'.

Name	Type	URL	Created By
db2-warehouse-srs	Db2 Warehouse on Cloud	jdbc:db2://db2w-ivteemh.us-south.db2w.cloud.ibm.com:50001/BLUDB:sslConnection=true;sslTru...	Shivam Solanki (IBM)
db2-warehouse	Db2 Warehouse on Cloud	jdbc:db2://db2w-olgagnz.us-south.db2w.cloud.ibm.com:50001/BLUDB:sslConnection=true;sslTru...	admin

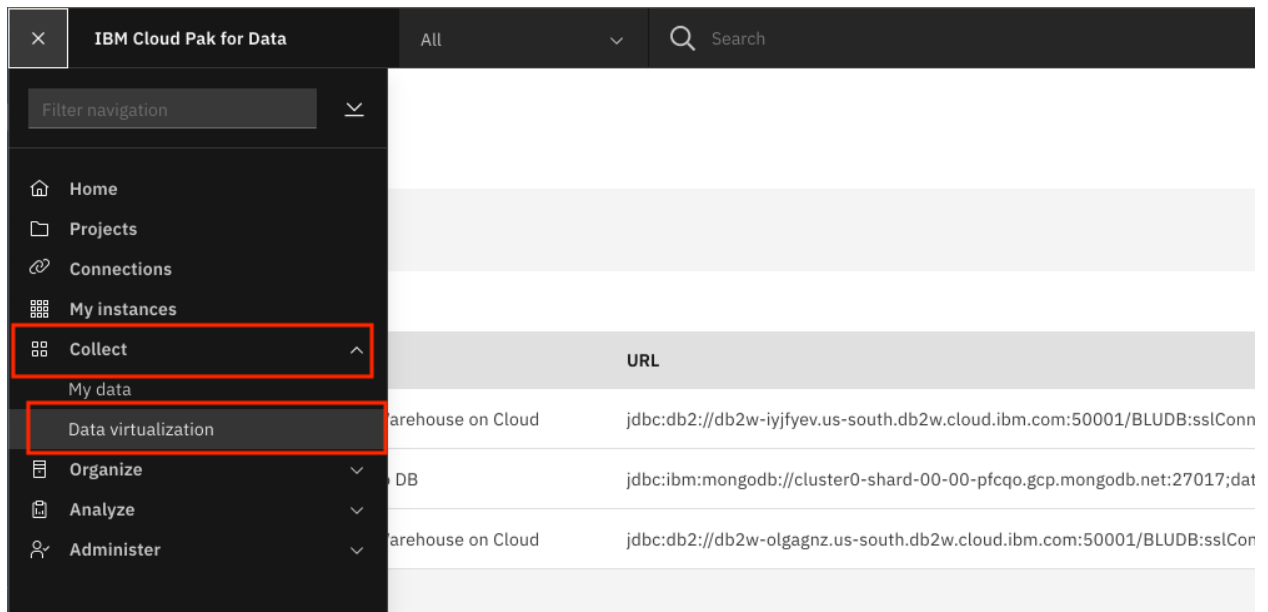
Step 5. Virtualize Db2 data with data virtualization

NOTE: This section requires Admin user access to the IBM Cloud Pak for Data cluster.

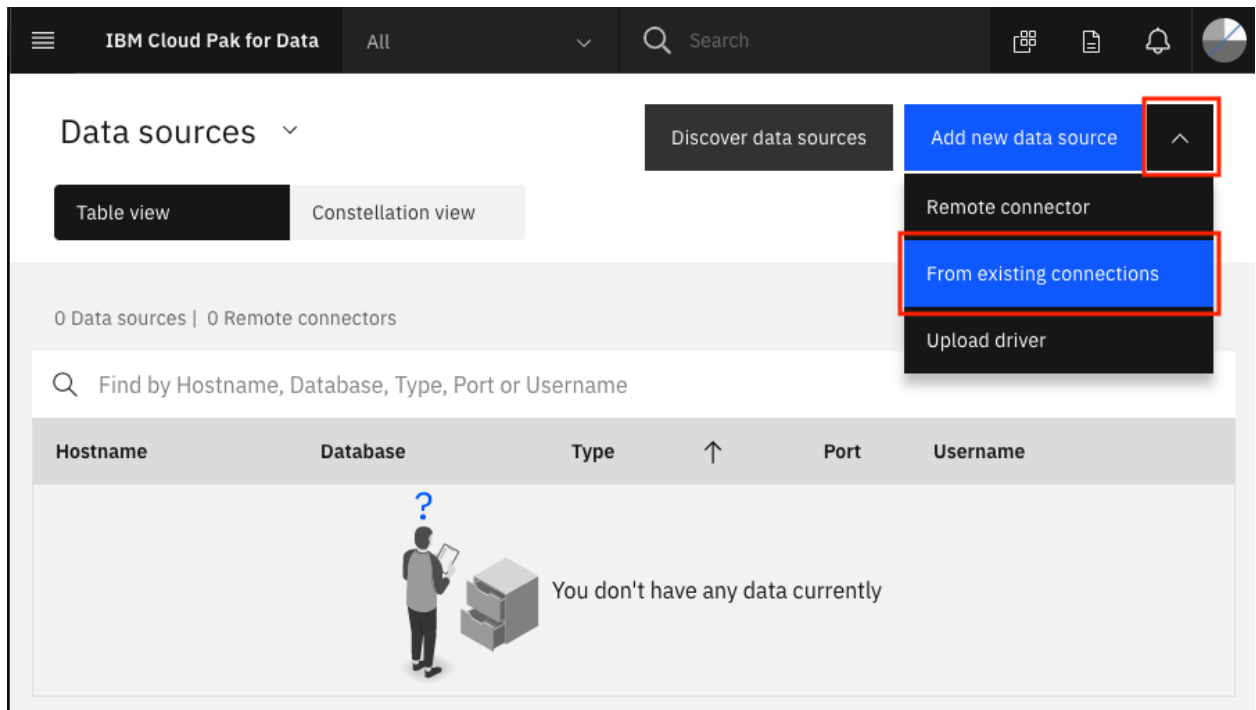
For this section, we'll use the data virtualization tool to import the data from Db2 Warehouse, which is now exposed as a connection in IBM Cloud Pak for Data.

Add a data source to Data Virtualization

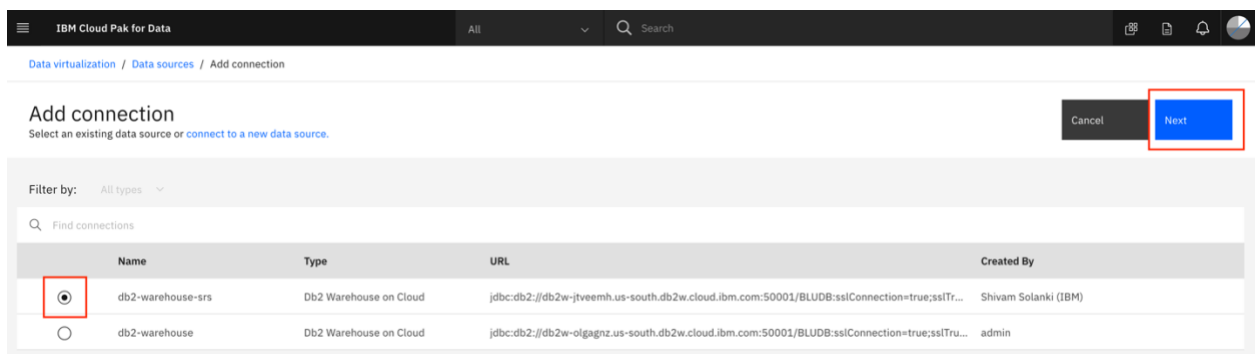
1. To launch the data virtualization tool, go the hamburger (☰) menu and click **Collect** and then **Data Virtualization**.



2. At the empty overview, click the drop-down next to Add new data source and select **From existing connections**



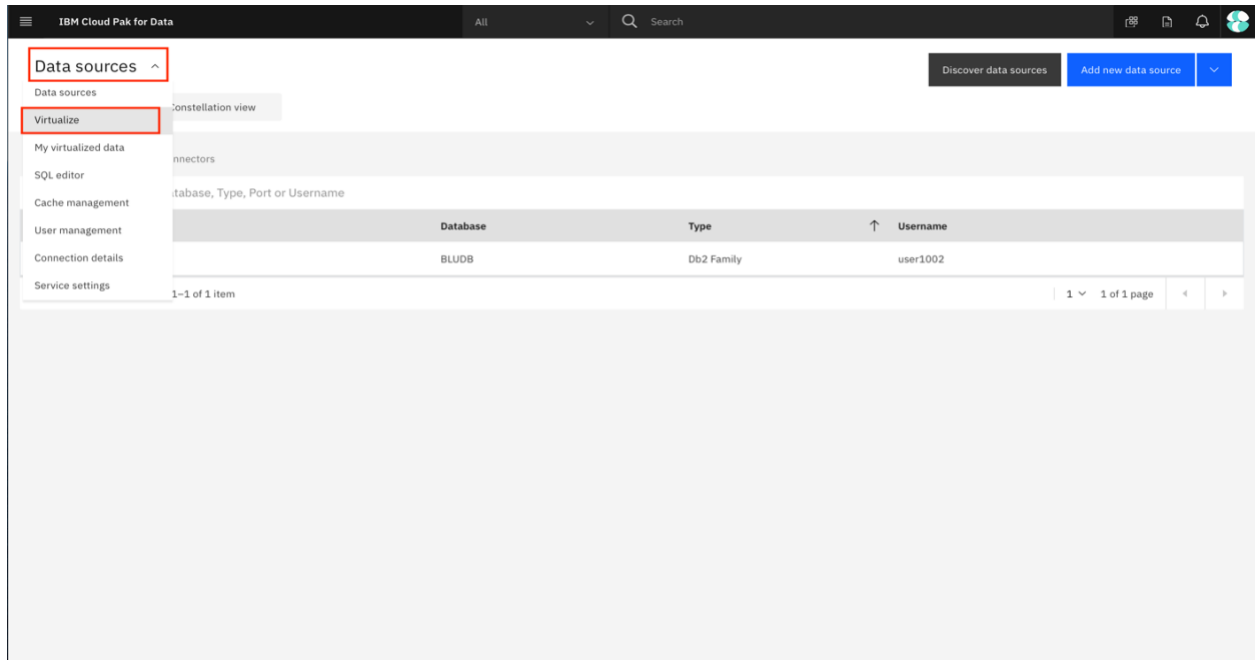
3. Select the data source we made in the previous step and click **Next**.



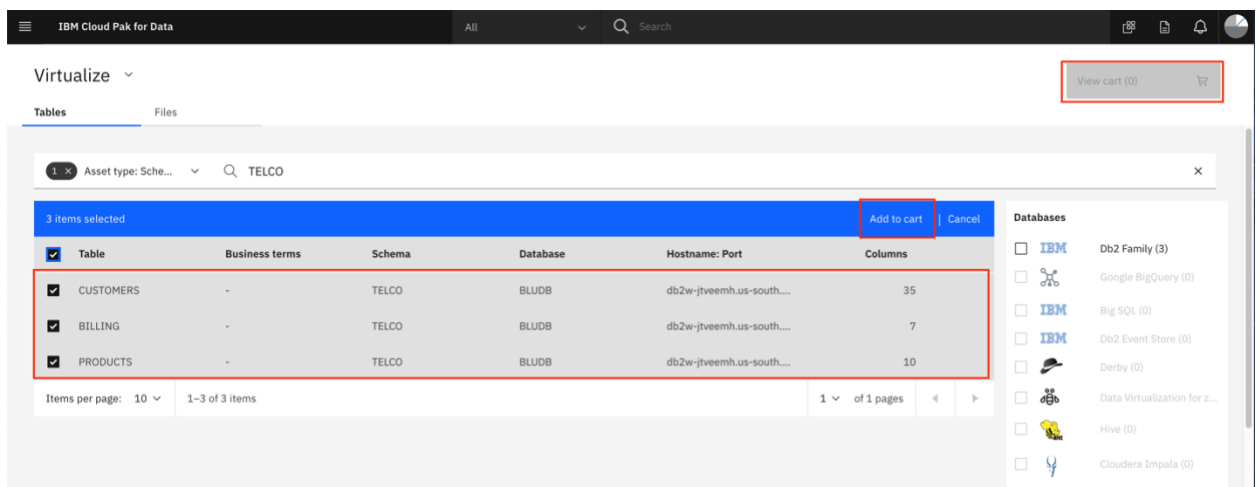
Start virtualizing data

Because we now have access to the Db2 Warehouse data, we can virtualize the data to our IBM Cloud Pak for Data project.

1. Click on the **Data Sources** drop-down and choose **Virtualize**.



2. Several tables will appear (many are created as sample data when a Db2 Warehouse instance is provisioned) in the table. Find the tables that was created earlier, the previous instructions suggested naming them CUSTOMERS, PRODUCTS, and BILLING. Or you can search them by filtering on Schema = TELCO. Once selected, click **Add to cart** and then **View Cart**.



3. The next panel prompts the user to choose which project to assign the data to. Choose **My virtualized data** and *uncheck* the box that says Submit to catalog. Rename your tables to avoid conflict with the existing virtualized tables from another project and/or user. Click **Virtualize** to start the process.

Review cart and virtualize tables

Assign to

☐ Data request ☐ Project ☒ My virtualized data

Submit to catalog Empty cart Virtualize

Objects to be virtualized

Table	Schema	Source schema	Databases/File Path	Hostname: Port	Grouped tables
CUSTOMERS-TELCO-SRS	USER1003	TELCO	BLUDB	db2w-jvveemh.us-south.db2w.clo...	1
BILLING-TELCO-SRS	USER1003	TELCO	BLUDB	db2w-jvveemh.us-south.db2w.clo...	1
PRODUCTS-TELCO-SRS	USER1003	TELCO	BLUDB	db2w-jvveemh.us-south.db2w.clo...	1

- You'll be notified that the virtual tables have been created. Let's see the new virtualized data from the data virtualization tool by clicking **View my virtualized data**.

Review cart and virtualize tables

Assign to

☐ Data request ☐ Project ☒ My virtualized data

Submit to catalog Empty cart Virtualize

Objects to be virtualized

Table	Schema	Status
BILLING-TELCO-SRS	USER1003	success
PRODUCTS-TELCO-SRS	USER1003	success
CUSTOMERS-TELCO-SRS	USER1003	success

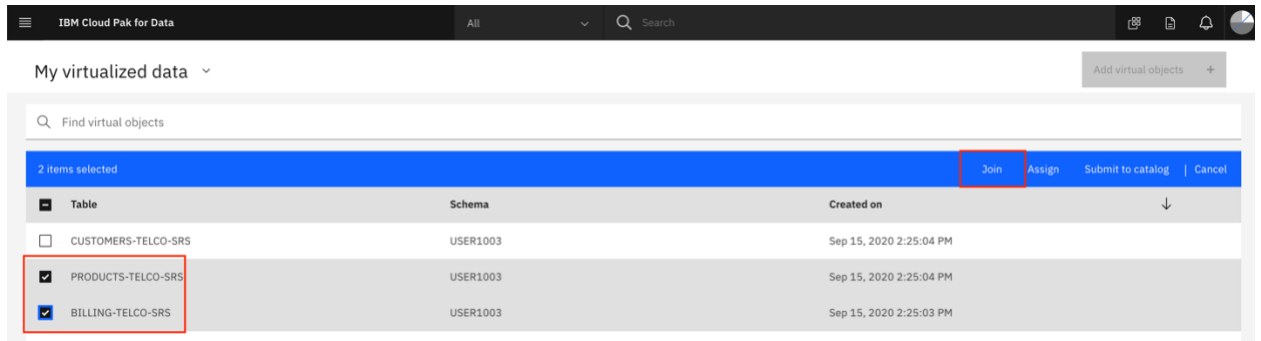
Assigned to none

Virtualize more data View my virtualized data

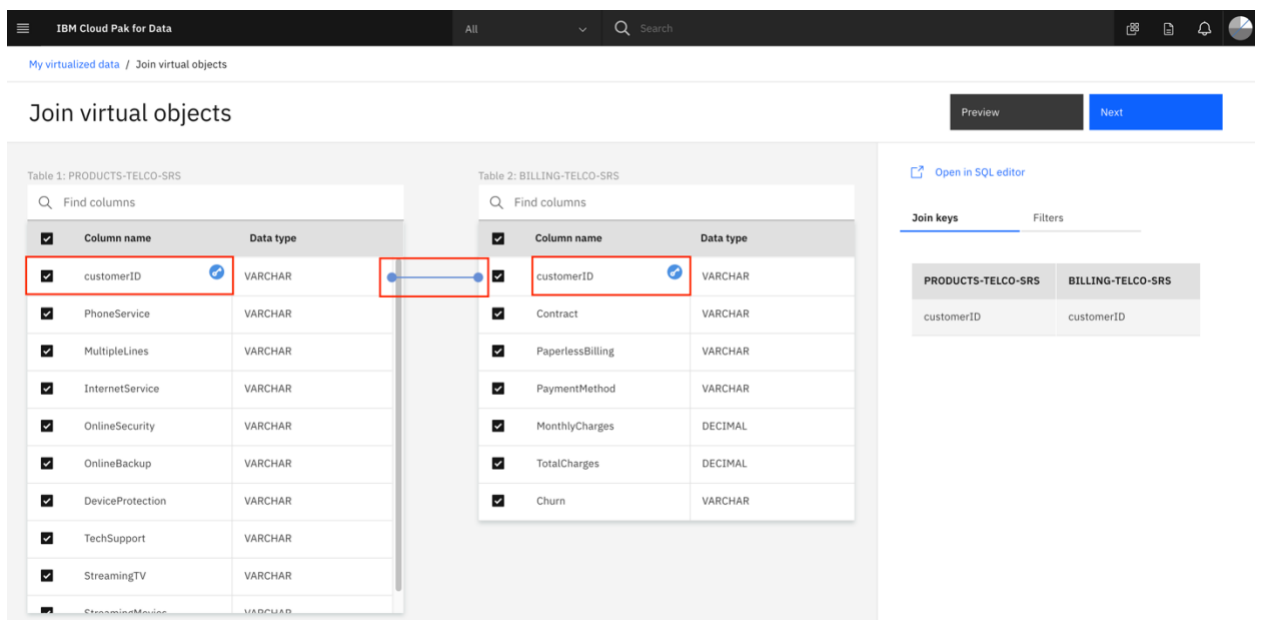
Join the virtualized data

Now we are going to *join* the tables we created so we have a merged set of data. It will be easier to do it here rather than in a notebook where we would have to write code to handle three different datasets.

- Click on any two tables (BILING-TELCO and PRODUCTS-TELCO, for instance), then click the **Join** button.



- To join the tables, we need to pick a key that is common to both datasets. Here we choose to map **customer_id** from the first table to **customer_id** on the second table. Do this by clicking one and dragging it to the other. When the line is drawn, click **Next**.



- Next, you have a chance to edit column names, but we will keep them as-is. Click **Next**.

IBM Cloud Pak for Data

All

Search

My virtualized data / Join virtual objects / Edit column names

Edit column names

Next

Preview load complete

customerID	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtectio	TechSupport	StreamingT
1174-FGIFN	Yes	Yes	Fiber optic	No	No	No	Yes	No
6620-HVDUJ	Yes	No	DSL	Yes	Yes	No	Yes	No
4701-MLJPN	Yes	No	DSL	No	No	Yes	Yes	No
1032-MAELW	Yes	Yes	DSL	Yes	Yes	Yes	Yes	No
7641-TQFHN	Yes	Yes	Fiber optic	No	No	No	Yes	Yes
1552-TKMXS	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet s

- In the next panel, we'll give our joined data a unique name such as BILLINGPRODUCTS-SRS (to be consistent with SQL standards, pick an uppercase name). Under **Assign to**, choose **My virtualized data** and uncheck the box that says Submit to catalog. Click **Create view** to start the process.

IBM Cloud Pak for Data

All

Search

My virtualized data / Join virtual objects / Edit column names / Assign and review

Assign and review

Submit to catalog

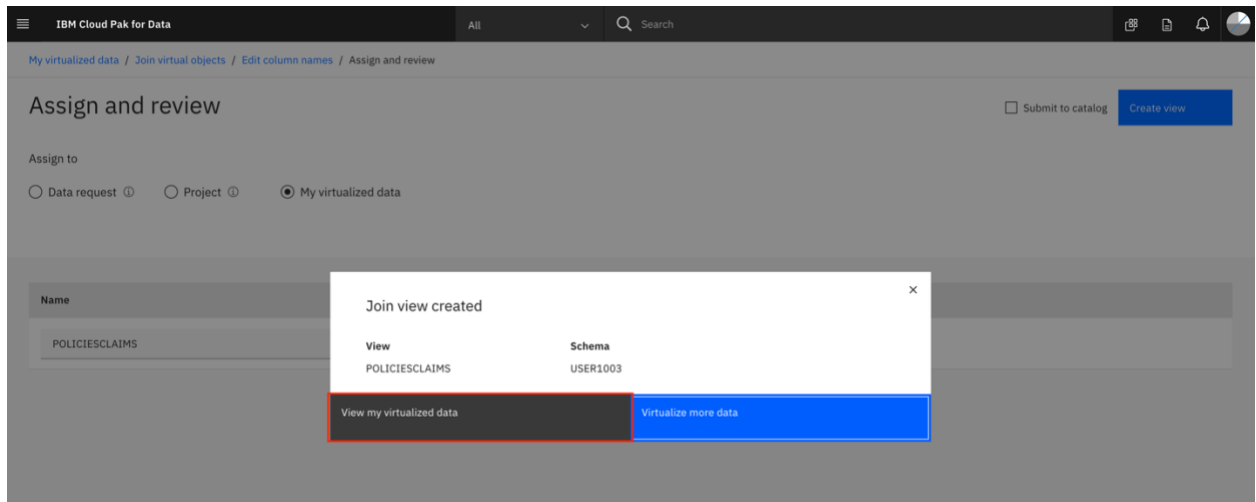
Create view

Assign to

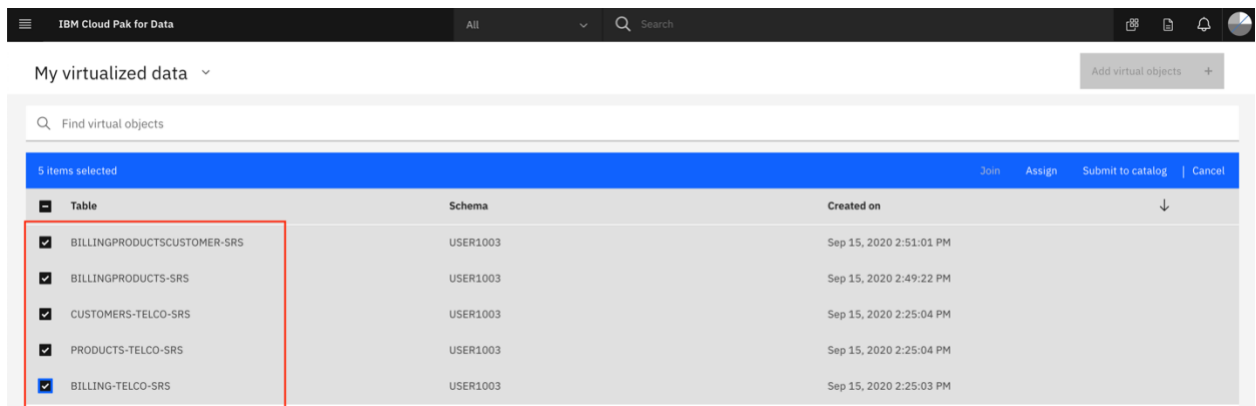
☐ Data request ⓘ ☐ Project ⓘ ☒ My virtualized data

Name	Schema
BILLINGPRODUCTS-SRS	USER1003

5. You'll be notified that the join has succeeded. Click on **View my virtualized data** to go back and see all your virtualized data.



6. **IMPORTANT** Now join the new joined view (BILLINGPRODUCTS-SRS) and the last virtualized table (CUSTOMERS-TELCO-SRS) to create a new joined view that has all three tables; let's call it BILLINGPRODUCTSCUSTOMER-SRS. Switching back to the **My virtualized data** screen should show all three virtualized tables and two joined tables. Do not go to the next section until this step is performed.



Grant access to the virtualized data

For other users to have access to the data you just virtualized, you need to grant it. Follow these steps to make your virtualized data visible to them:

1. Go to **Data Virtualization** from the hamburger (≡) menu. Click on **Menu** → **My virtualized data**.
2. Click on the virtualized data you've created, then click the three vertical dots to the right, and choose **Manage access**.

My virtualized data

Find virtual objects

5 items selected

Table	Schema	Created on
<input checked="" type="checkbox"/> BILLINGPRODUCTSCUSTOMER-SRS	USER1003	Sep 15, 2020 2:51:01 PM
<input checked="" type="checkbox"/> BILLINGPRODUCTS-SRS	USER1003	Sep 15, 2020 2:49:22 PM
<input checked="" type="checkbox"/> CUSTOMERS-TELCO-SRS	USER1003	Sep 15, 2020 2:25:04 PM
<input checked="" type="checkbox"/> PRODUCTS-TELCO-SRS	USER1003	Sep 15, 2020 2:25:04 PM
<input checked="" type="checkbox"/> BILLING-TELCO-SRS	USER1003	Sep 15, 2020 2:25:03 PM
<input type="checkbox"/> clairespolicescustomMN	USER999	Sep 9, 2020 10:46:32 AM
<input type="checkbox"/> Claim_policyMN	USER999	Sep 9, 2020 10:42:44 AM

Join Assign Submit to catalog Cancel

New view
Preview
View table structure
View metadata
Manage access
Submit to catalog
Remove

3. Click the **Specific users** radio button, then **Add user +**.

My virtualized data / Manage access

Manage Access: BILLINGPRODUCTSCUSTOMER-SRS

Grant access to

☒ **Specific users** ☐ All data virtualization users

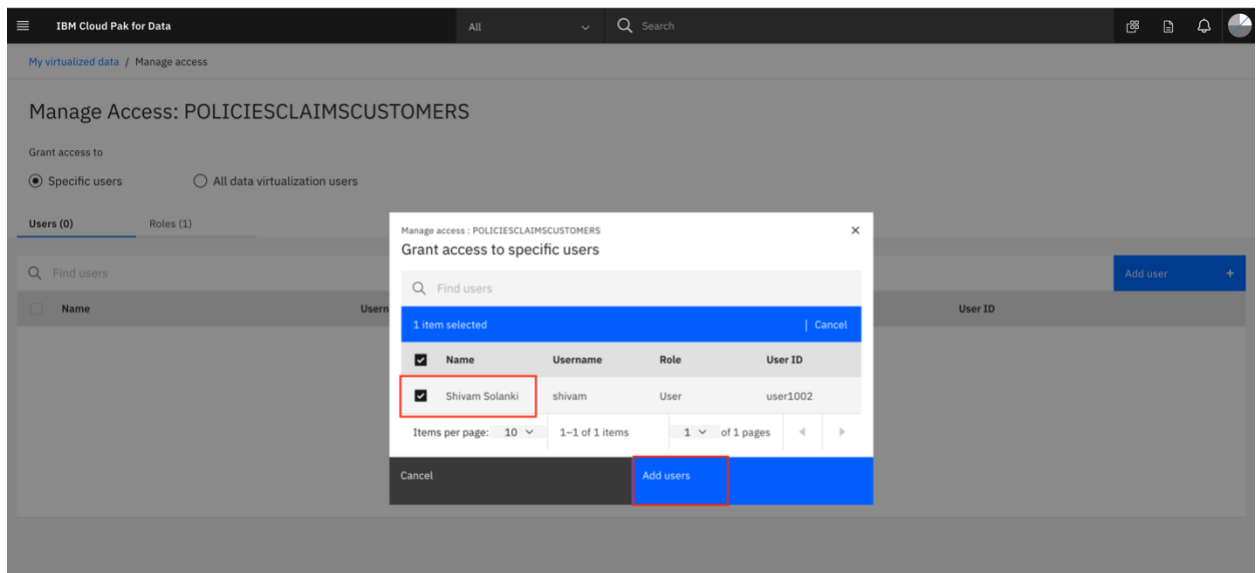
Users (0) Roles (1)

Find users

Add user +

Name	Username	Role	User ID
<p>No users have been granted access to this virtual object.</p> <p>Data Virtualization Admins can grant access to virtual objects to users with the Engineer or User roles only.</p>			

4. Select the users you wish to grant access to and click **Add users**.



Repeat the above steps for the remaining tables and views.

Assign the Engineer role to the users

IBM Cloud Pak for Data users that need to use data virtualization functions must be assigned specific roles based on their job descriptions. These roles are Admin, Engineer, User, and Steward. You can learn more about these roles on the [IBM Cloud Pak for Data product hub](#).

Let's assign the Engineer role to some users:

1. From the hamburger (☰) menu, choose the **Data Virtualization** option, then click **My virtualized data** → **User management**.

The screenshot shows the 'User management' page in the IBM Cloud Pak for Data console. The page has a dark header with the IBM Cloud Pak for Data logo, a search bar, and navigation icons. Below the header, the 'User management' title is followed by a dropdown arrow. Two tabs, 'Users' and 'Roles', are visible, with 'Users' being the active tab. The main content area is titled 'Users' and includes a subtitle: 'View which users have access to the Data Virtualization service. Grant access to more users and define their role.' There is a 'Filter by:' dropdown set to 'All roles' and a search icon. A blue 'Add users +' button is located in the top right corner. Below this is a table with four columns: 'Name', 'Username', 'Role', and 'User ID'. The table contains six rows of user data.

Name	Username	Role	User ID
admin	admin	Admin	user999
Omid Meh	omid	Admin	user1001
Scott D'Angelo	scott	Admin	user1002
Javier Torres	javier	Admin	user1003
Samaya Madhavan	samaya	Admin	user1015
mark sturdevant	mark	Admin	user1016

2. Click on **Add users +** and update the role of the user to Engineer.

This screenshot shows the same 'User management' page as the previous one, but with a modal dialog open. The modal is titled 'Grant access to new users' and has a subtitle: 'Specify which users can access the Data Virtualization service and define their role.' It contains a search bar and a table with three columns: 'Name', 'Username', and 'Role'. The table has one row with the following data: 'Mark Data Engineer', 'mark_de', and 'User'. The 'Role' cell has a dropdown menu open, showing the following options: 'Admin', 'Engineer', 'Steward', and 'User'. The 'Engineer' option is highlighted. At the bottom of the modal, there are 'Cancel' and 'Add' buttons.

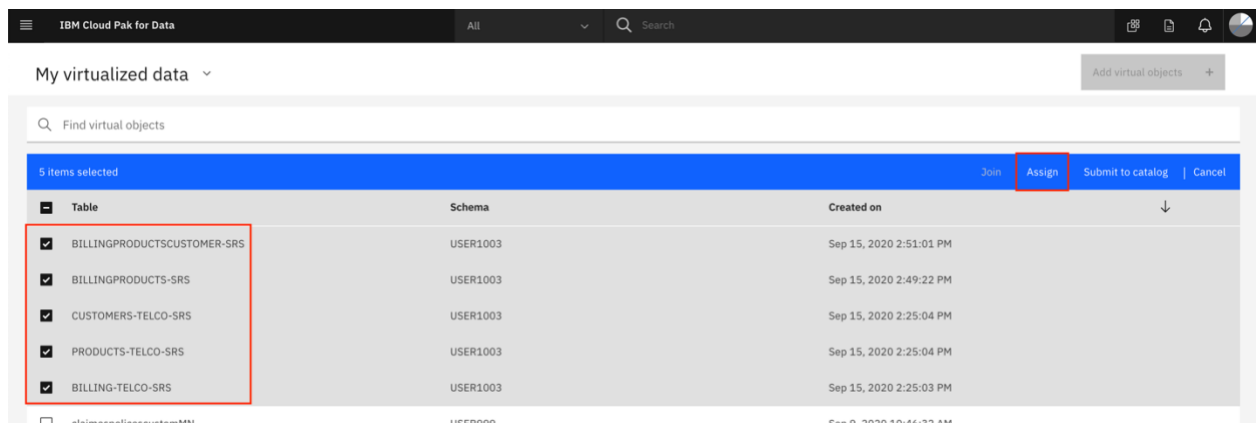
Name	Username	Role
Mark Data Engineer	mark_de	User

Step 6. Users assign virtualized data

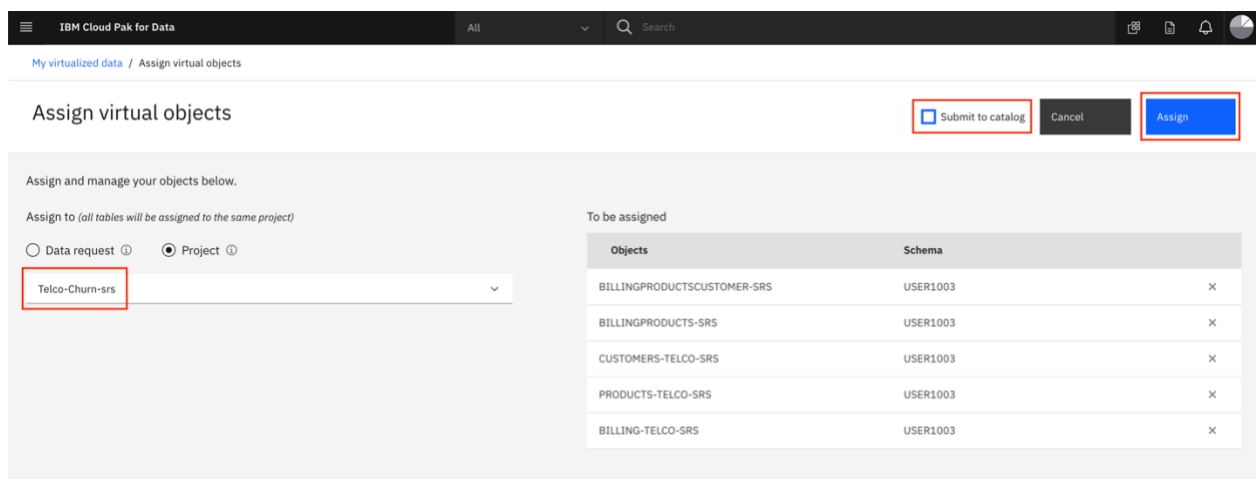
Now let's look at how a user who has access to virtualized data can assign the data to their project — how to add the virtualized data as asset to a project.

Assign the data to your project

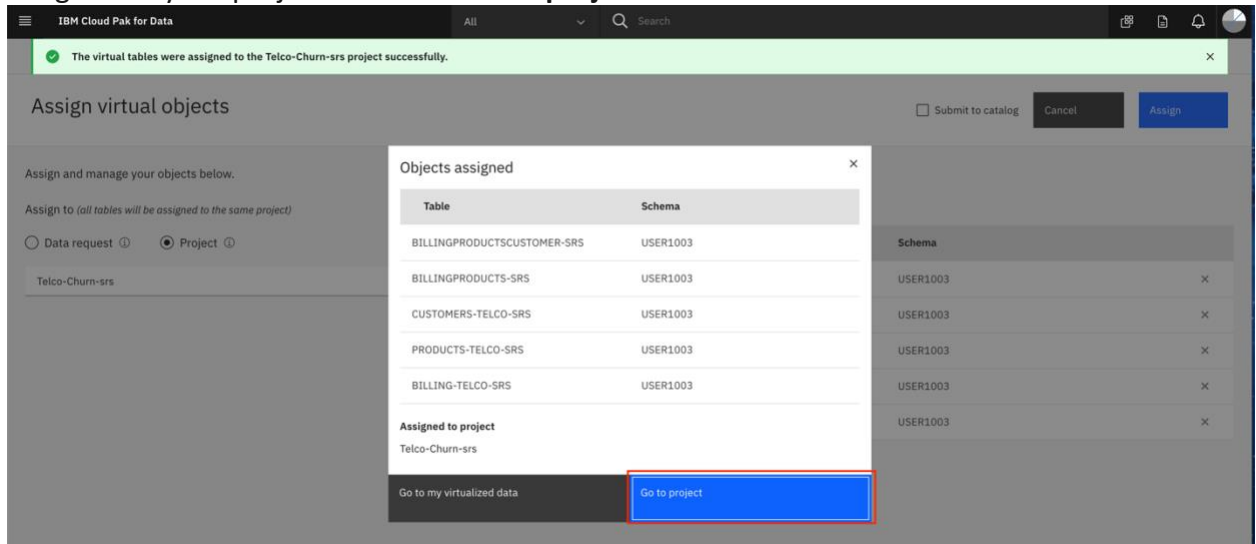
1. From the hamburger (☰) menu, click on **Collect** → **Data Virtualization**. You will be brought to the My virtualized data section. Here you should see the data you can access (or that the administrator has assigned to you). Select the checkbox next to our original tables (BILLING-TELCO-SRS, PRODUCTS-TELCO-SRS, CUSTOMERS-TELCO-SRS) and the joined tables (BILLINGPRODUCTS-SRS, BILLINGPRODUCTSCUSTOMER-SRS), and click the **Assign** button to import them into your project.




2. On the Assign virtual objects screen, choose the project to assign the data. If there is a **Submit to catalog** checkbox on the top right, uncheck it and click the **Assign** button to add the data to your project.

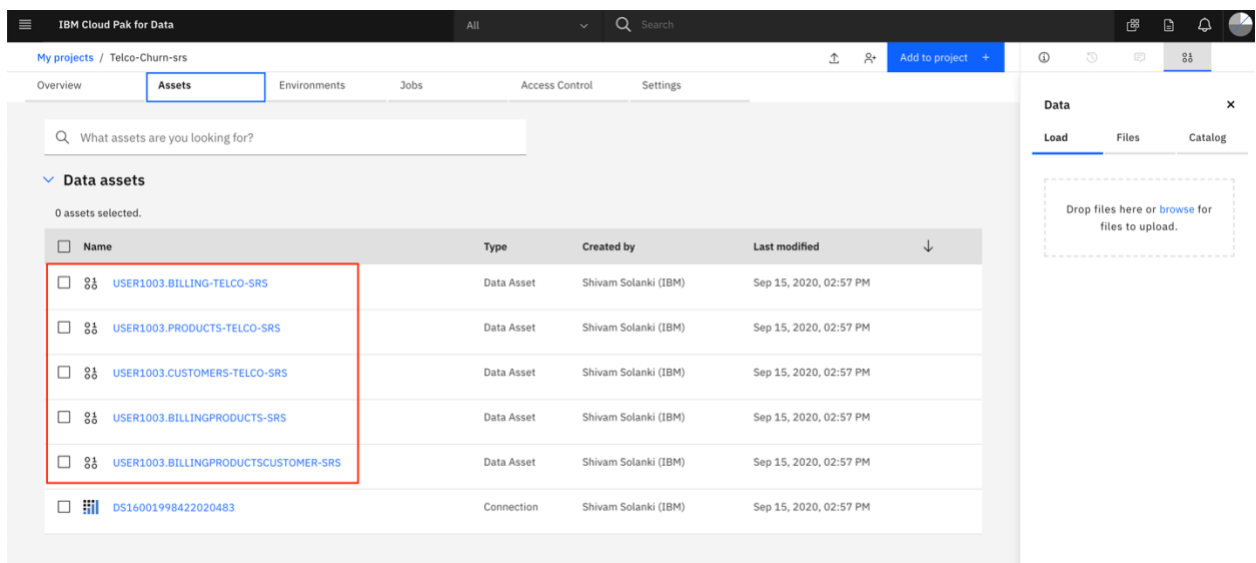


3. In the pop-up panel, you will receive a confirmation that the objects have been assigned to your project. Click the **Go to project** button.



Alternatively, close the modal and go to your projects by clicking on the hamburger () menu, then choosing **Projects**.

4. On the project page, click on the **Assets** tab to display the virtualized tables and joined tables that are now in your project.



Summary

This lab tutorial explained how to virtualize Db2 Warehouse data with data virtualization on IBM Cloud Pak for Data to make queries across multiple data sources. To continue the series and learn more about IBM Cloud Pak for Data, take a look at the next tutorial, [Data visualization with data refinery](#).