Governed MLOps Workshop
Data Virtualization

Document version: June 2023

DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

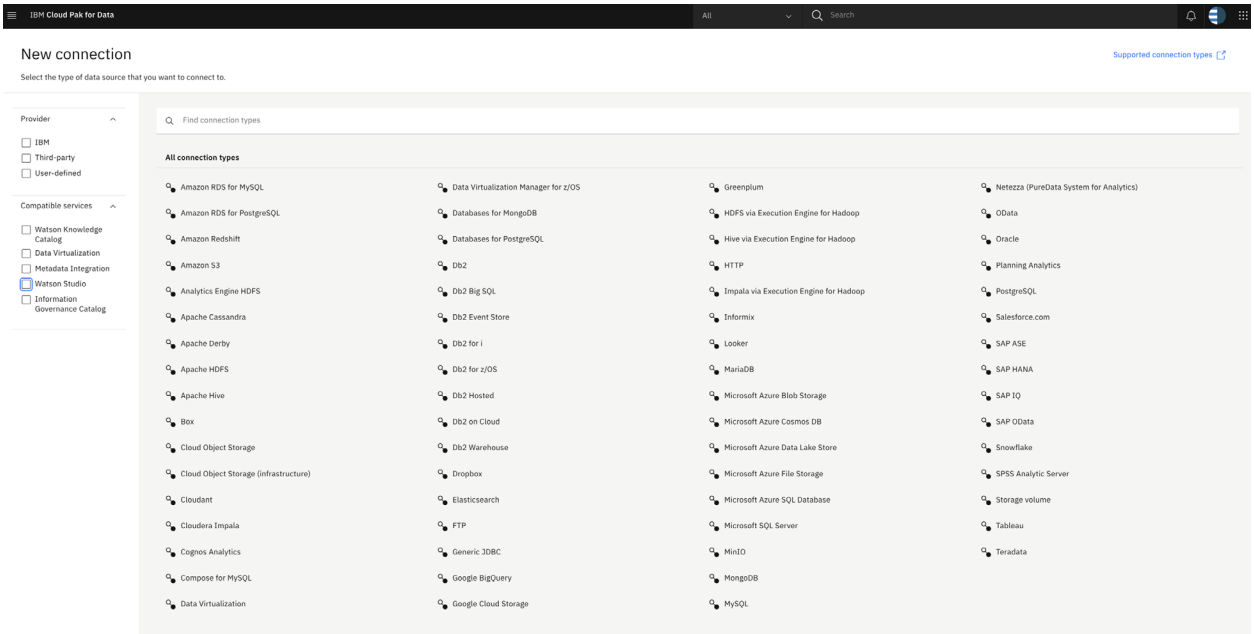
Table of Contents

Introduction

After setting up governance foundation on your Cloud Pak for Data cluster, you can proceed to establish connections to the data sources that you will leverage in training AI models for customer churn prediction.

Data Source Connections

IBM Cloud Pak for Data supports a rich set of connectors to popular data sources including 3rd party data sources, such as the ones illustrated in the figure below.



Additionally, with Cloud Pak for Data, you can create platform level connections which would be available to all services on the platform, or you can create connections specific to an individual service.

Please note that even if you create a platform level connection to a data source, you will need to review and validate that the service you intend to use supports that connection. Please review the documentation for details on what data sources are supported and by which services.

Lab objectives

In this lab, you will leverage Data Virtualization (Watson Query) service in Cloud Pak for Data to connect multiple data sources across locations and turn all this data into one logical data view. This virtual data view makes the job of getting value out of your data easy. Data Virtualization capabilities significantly increase the AI throughput of data science teams by helping data scientists efficiently access the broad set of data sources of an enterprise across a hybrid multi-cloud environment without having to copy the data.

With Data Virtualization, after creating connections to your data sources, you can quickly view all your organization's data. This virtual data view enables real-time analytics without moving data, duplication, ETLs, or additional storage requirements, so processing times are greatly accelerated. You can bring real-time insightful results to decision-making applications or analysts more quickly and dependably than methods that don't use virtualization.

In addition to connecting to data sources across hybrid cloud environments, we need to make sure that data is governed per the requirements of the enterprise. To achieve that, we will publish the data assets to a governed enterprise catalog, so that governance policies get applied to such data assets. Additionally, the catalog makes the data readily available for self-service where different data consumers can search and find the assets most suitable for their needs.

Lab data assets

For this lab, to emulate a realistic scenario that is typical of most enterprises, we assume there are three data assets that are in different formats and are available across a hybrid cloud environment:

- **Customer Personal Information:** This data captures personal information of the customers such as gender, marital status, income, age, and similar data. This data set is assumed to exist in a table in an on-prem Db2 database. Customer data which includes personal information typically includes private and sensitive information and it is common to have such data available in some on-premises data store.
- **Customer Transaction Data:** This data captures the transaction data for customers, and it can exist on-premises or in a managed database on a public cloud. In this lab, we assume this data exists in IBM Db2 managed databases on IBM public cloud.
- **Customer Churn Data:** This data captures information about whether a specific customer did churn or not. It mainly consists of a customer ID and a corresponding churn label of T (true, the customer did churn) or F (false, the customer did not churn). This dataset is typically referred to as labeled data or ground-truth which is necessary for training AI models that fall into the supervised learning category. In this lab, we will use a csv file customer_churn_labels.csv to represent such data.

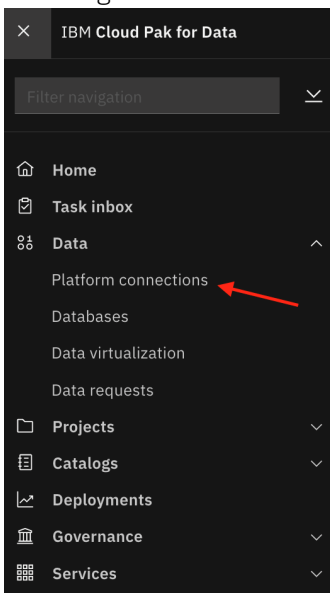
Platform Connections

The required platform connections should be created already. If not, please follow the instructions in the Appendix – Create Platform Connection.

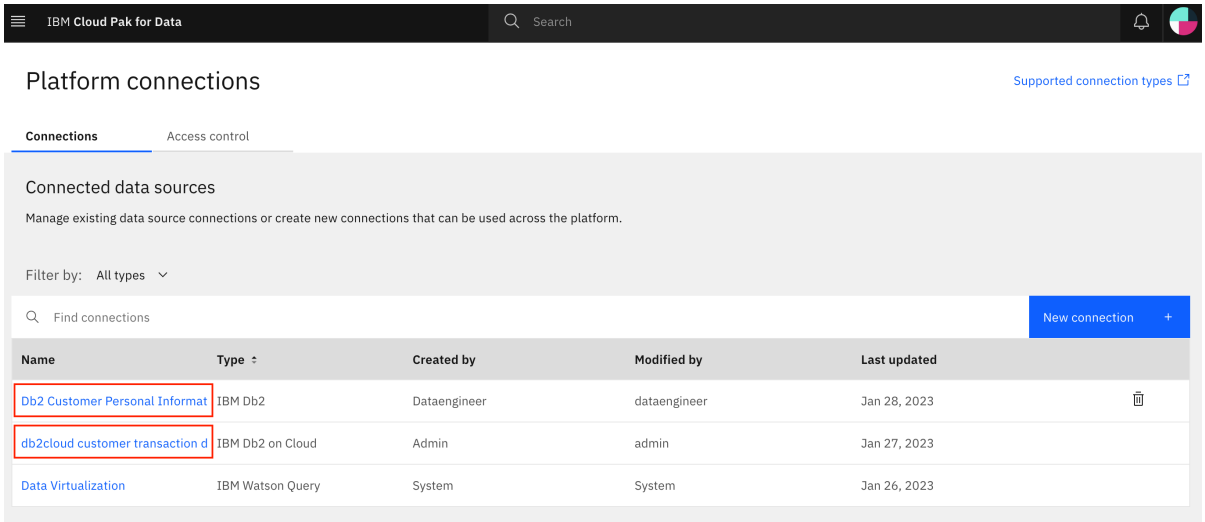
To confirm that you have the required platform connections created, execute the following:

1- Log into Cloud Pak for Data as **dataengineer** user.

2- Navigate to Platform connections by clicking on the Navigation menu (top left hamburger icon) and selecting **Data → Platform connections** (annotated with red arrow).



3- You should have two platform connections, the **Db2 Customer Personal Information** connection you created earlier which connects to the on-prem Db2 database and the **db2cloud customer transaction data** connection which was pre-created for you and that connects to the managed Db2 database on IBM Cloud.

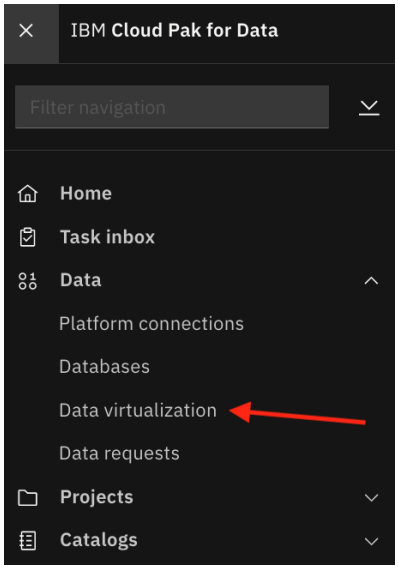


Virtualizing Data

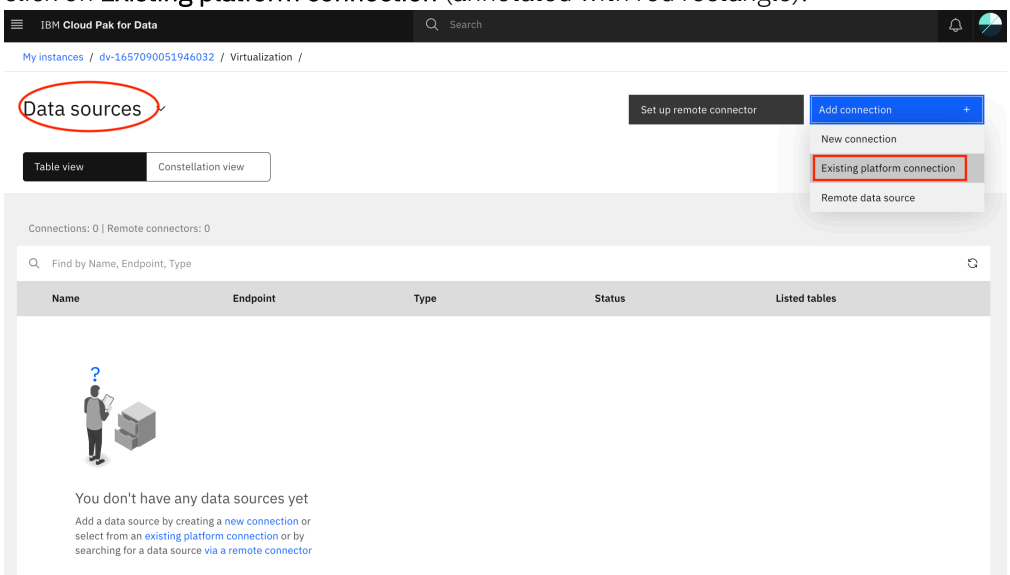
Now that you have created platform connections to two data sources, you can virtualize these data connections and then catalog them to make sure they're governed per the enterprise requirements and readily available for self-service by data science teams and business analysts.

1- Log back into Cloud Pak for Data as the **dataengineer** user.

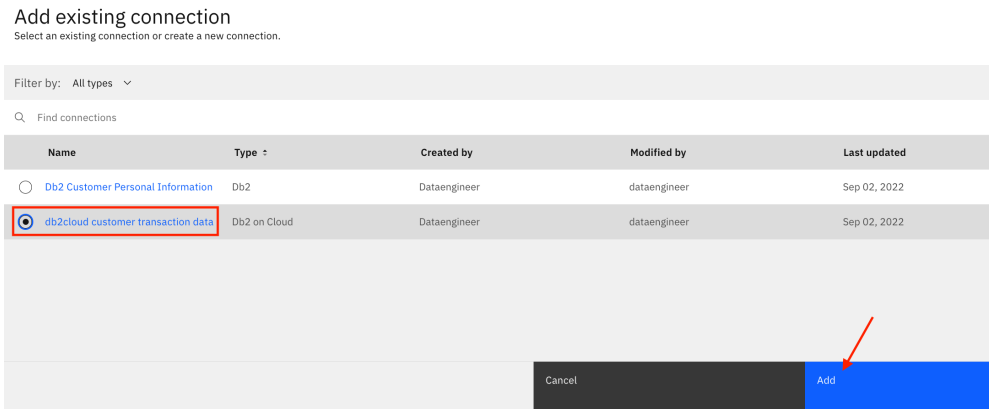
2- Navigate to the data virtualization page by clicking on the Navigation menu (top left hamburger icon) and selecting **Data → Data virtualization** (annotated with red arrow).



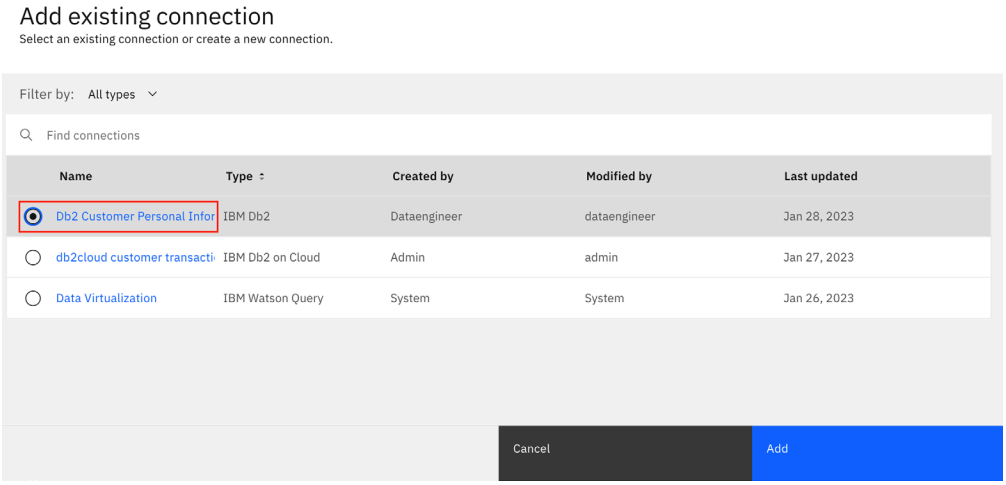
3- Click the Virtualization drop-down menu and select Data sources. On the Data virtualization **Data sources** page (annotated with red oval), click on **Add connection +** (annotated with red arrow) and click on **Existing platform connection** (annotated with red rectangle).



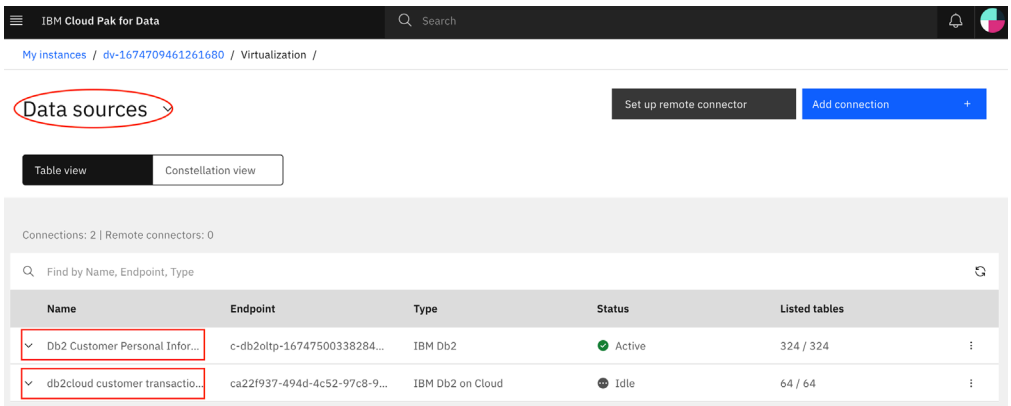
4- From the Add existing connection page, select the **db2cloud customer transaction data** connection (IBM Db2 on Cloud connection that was already created in your cluster) (annotated with red rectangle). This is where the Customer Transaction Data table was loaded. Click **Add** (annotated with red arrow). On the next page titled **Add a remote connector (optional)**, click **Skip**.



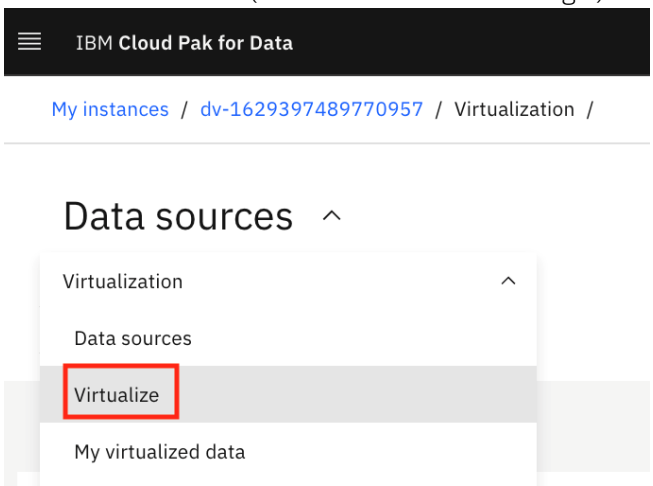
5- Repeat the process to add the **Db2 Customer Personal Information** connection (the on-prem IBM Db2 connection you created earlier) where the Customer Personal Information table exists. Click **Add** (annotated with red arrow). On the next page titled **Add a remote connector (optional)**, click **Skip**.



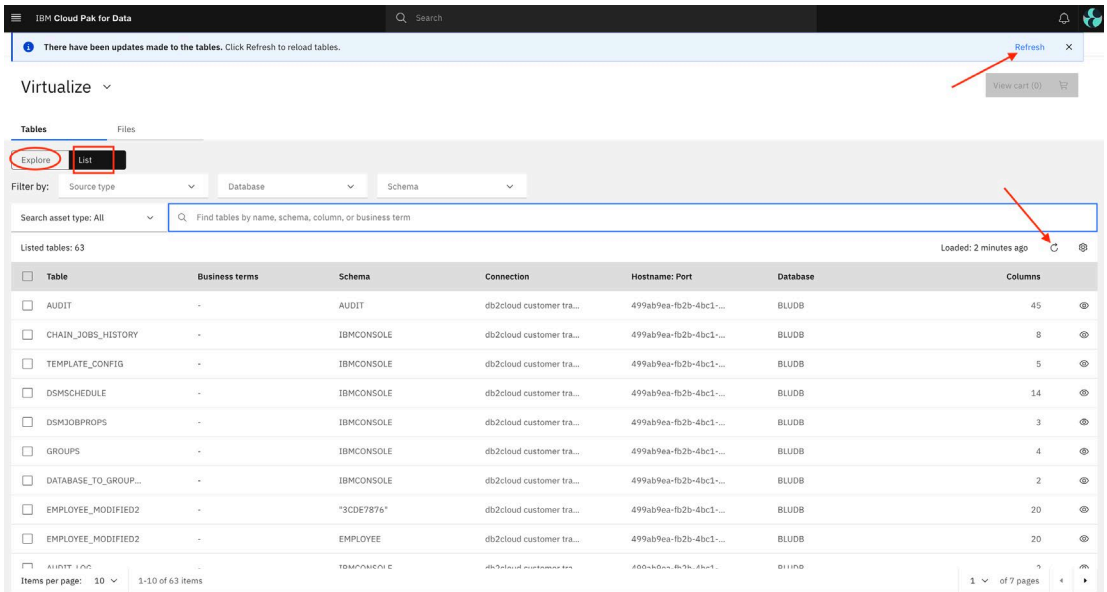
6- On the Data Sources page under Data Virtualization, you should now have both data sources added.



7- After adding the connections as sources to Data virtualization, click on the Data virtualization menu and select **Virtualize** (annotated with red rectangle).

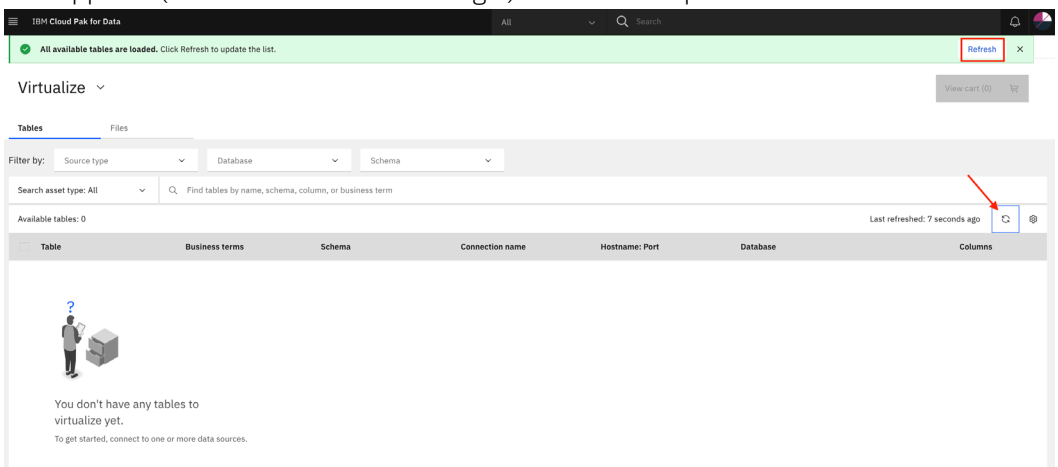


8- On the Virtualize page, you can explore the available tables from each of the defined data sources by clicking the **Explore** button (annotated with red oval). Alternatively, you can click the **List** button to list all available tables from all the defined data sources. Click the **List** button to view the list of all tables. If not all tables show from all data sources as expected, click the Refresh link or icon (annotated with red arrows).

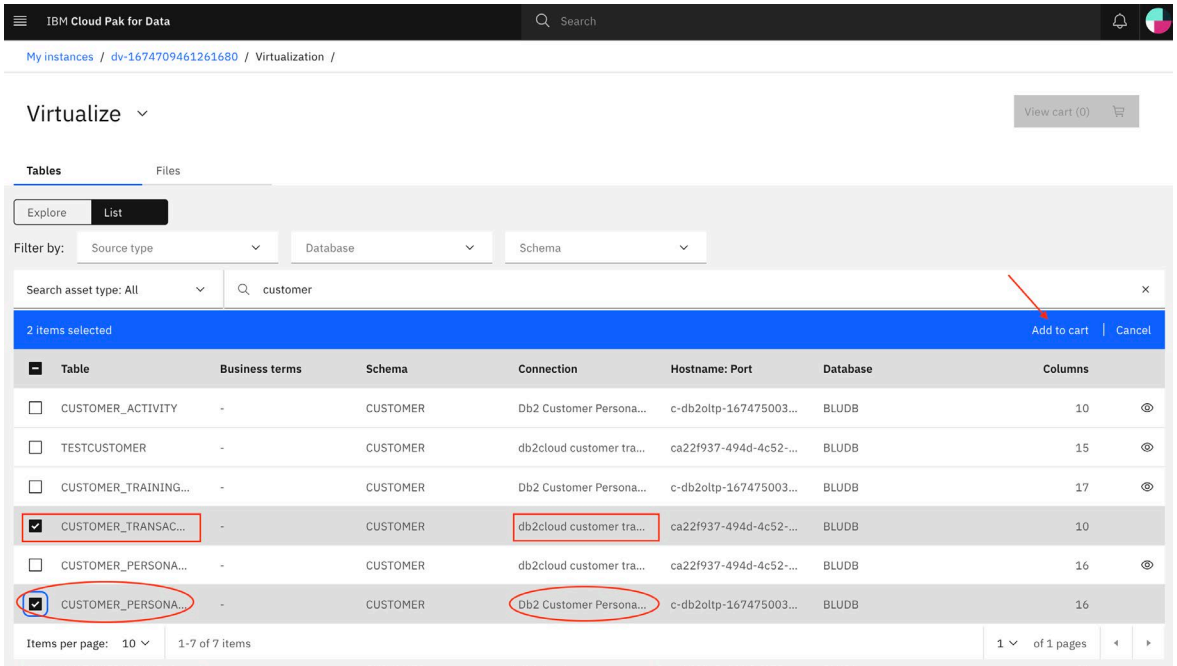


9- If you don't see all the expected data sets on the Virtualize page, make sure to refresh the page by clicking on the **refresh icon** (annotated with red arrow) and after that the **Refresh link** on the green bar

that appears (annotated with red rectangle). Now all the updated data should be visible.

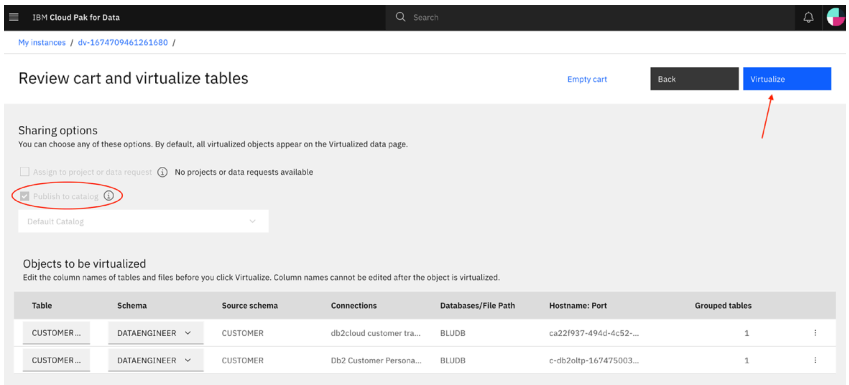


10- On the Virtualize page, select the data sets you wish to virtualize. Use the search bar and type CUSTOMER to filter the relevant data sets. Specifically, select the **CUSTOMER_TRANSACTION_DATA** table (annotated with red rectangle). Make sure you're selecting this table from the ***db2cloud customer transaction data*** connection (also annotated with red rectangle). Additionally, select the **CUSTOMER_PERSONAL_INFO** table (annotated with red oval) and make sure you're selecting this table from the ***Db2 Customer Personal Info*** connection (also annotated with red oval). Once selected, click on **Add to cart** (annotated with red arrow). Once the datasets are added to cart, click the **View cart** button.
Note that the View cart button becomes active only after assets are added to the cart.

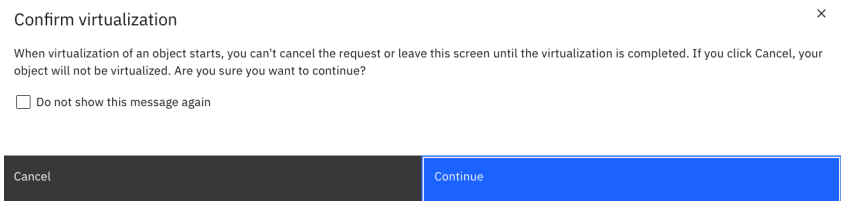


11- On the **Review cart and virtualize tables** page, review the correct tables are selected to be virtualized (annotated with red rectangle). Also note that the check box next to **Publish to catalog** is selected (annotated with red oval) so these virtualized tables will be published to the catalog where they are

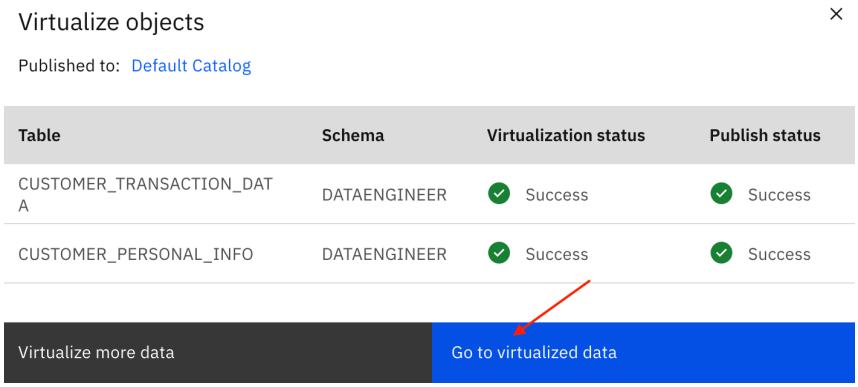
governed and made available for access by data consumers. Once you have reviewed the cart, click the **Virtualize** button (annotated with red arrow).



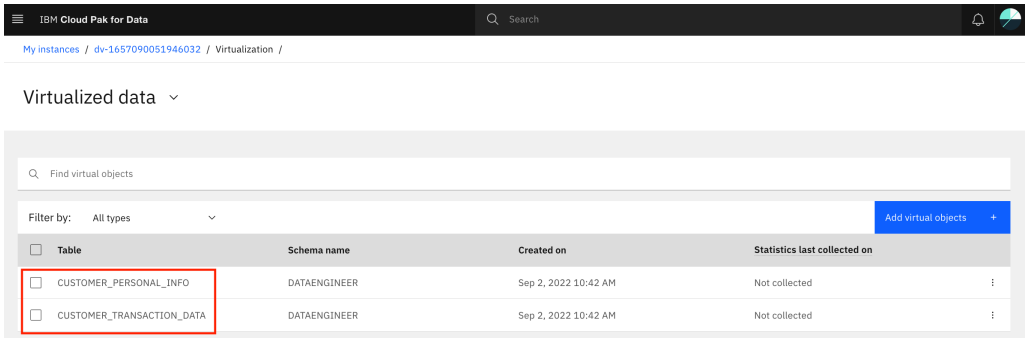
12- On the Confirm virtualization pop-up window, click **Continue**.



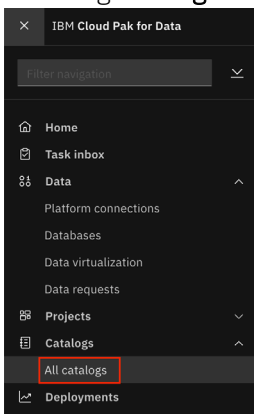
13- Wait for data virtualization to complete and then click **Go to virtualized data** (annotated with red arrow).



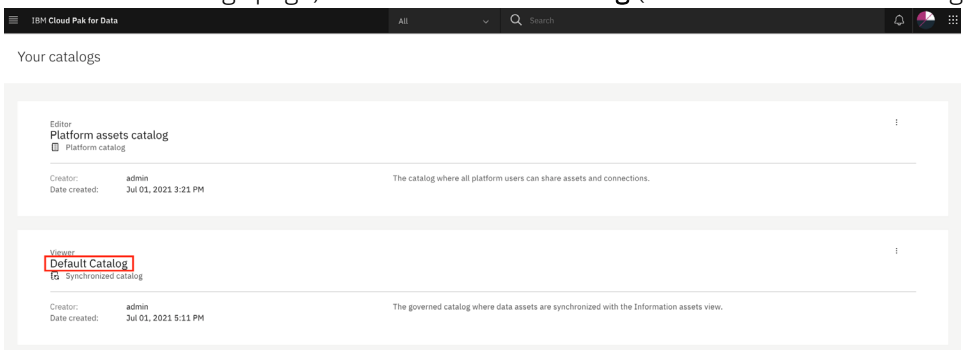
14- On the **Virtualized data** page, validate the two data sets (annotated with red rectangle) are virtualized and ready to be consumed.



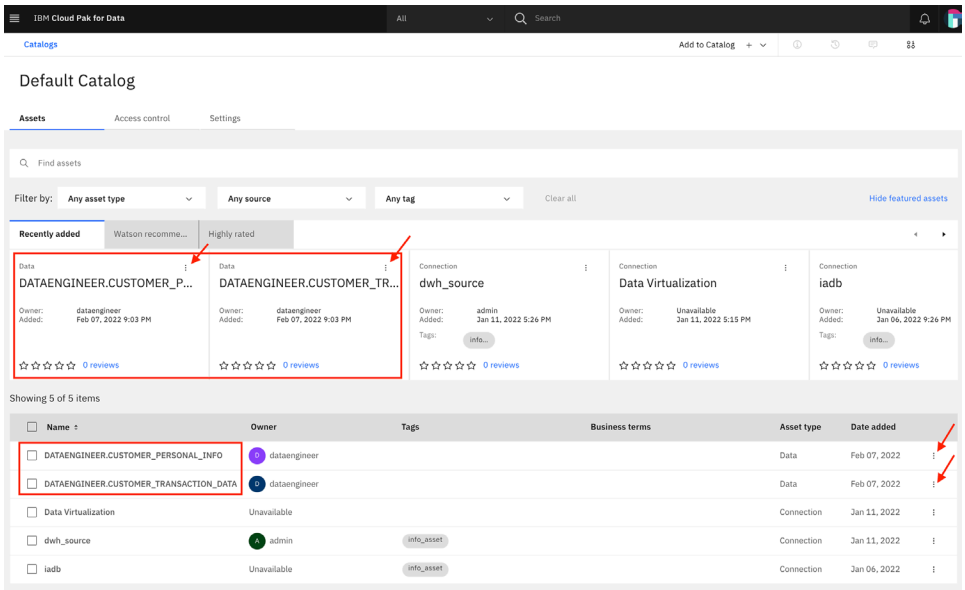
15- Next navigate to the **All catalogs** page by clicking the Navigation menu (top left hamburger icon) and selecting **Catalogs → All catalogs** (annotated with red rectangle).



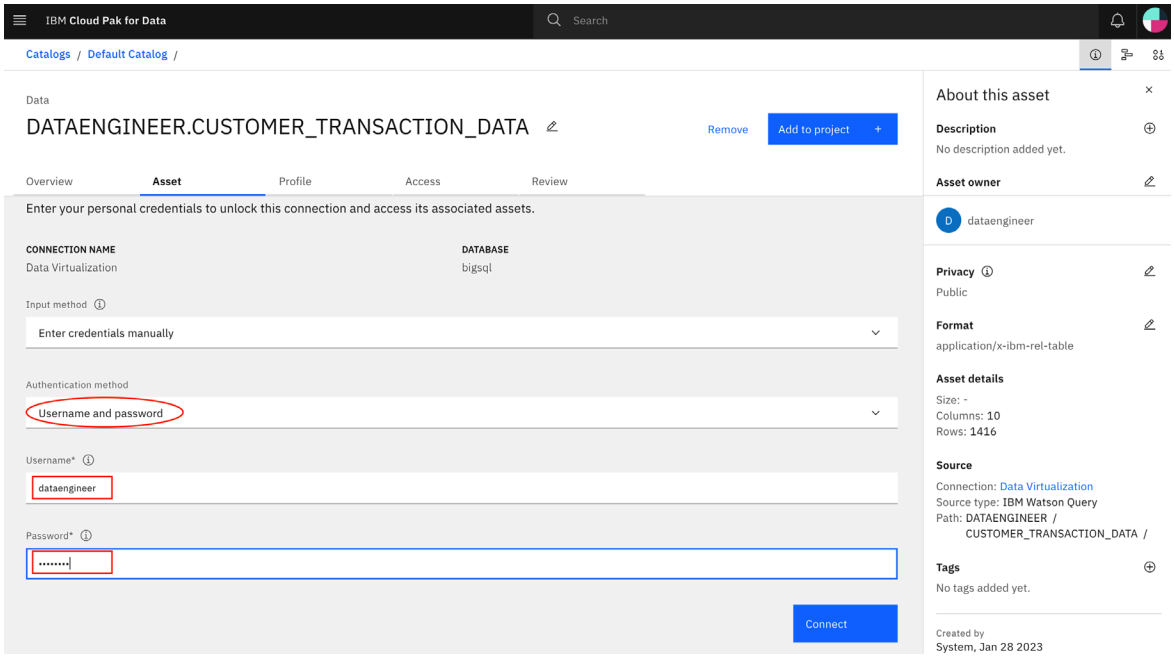
16- On the Your catalogs page, select the **Default Catalog** (annotated with red rectangle)



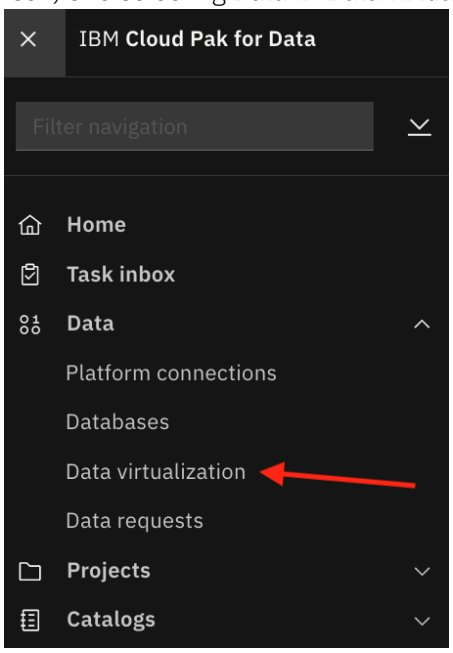
17- On the Default Catalog page, find the two data sets which have been virtualized and cataloged (annotated with red rectangle). Feel free to explore these data sets by clicking on either of them and reviewing information like Overview, Asset, Access, Review, Profile and Activities.
Click the open and close list of options (annotated with red arrow) next to either of these data sets and click **Open**.



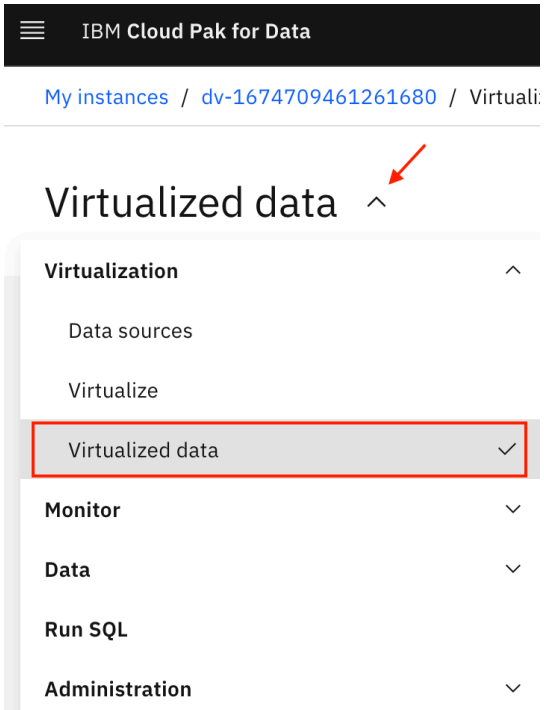
18- On the Data set page, click **Asset** tab (annotated with red oval) to preview the data. If you get a message to Unlock connection with personal credentials, select the Authentication method to be **Username and password** (annotated with red oval) and provide the credentials for the **dataengineer** user (annotated with red rectangle) and click **Connect** at the bottom right of the page.



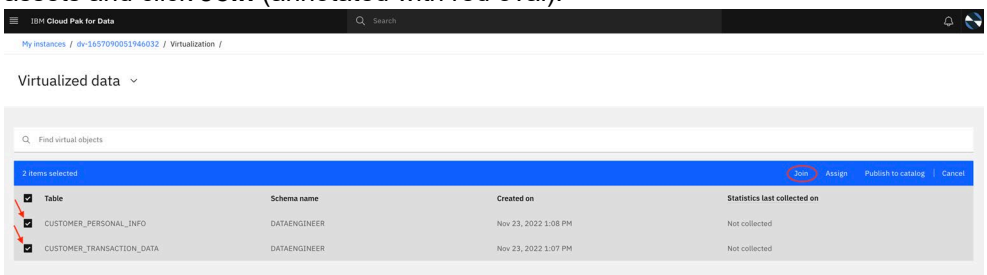
19- Navigate back to the data virtualization page by clicking on the Navigation menu (top left hamburger icon) and selecting **Data → Data virtualization** (annotated with red arrow).



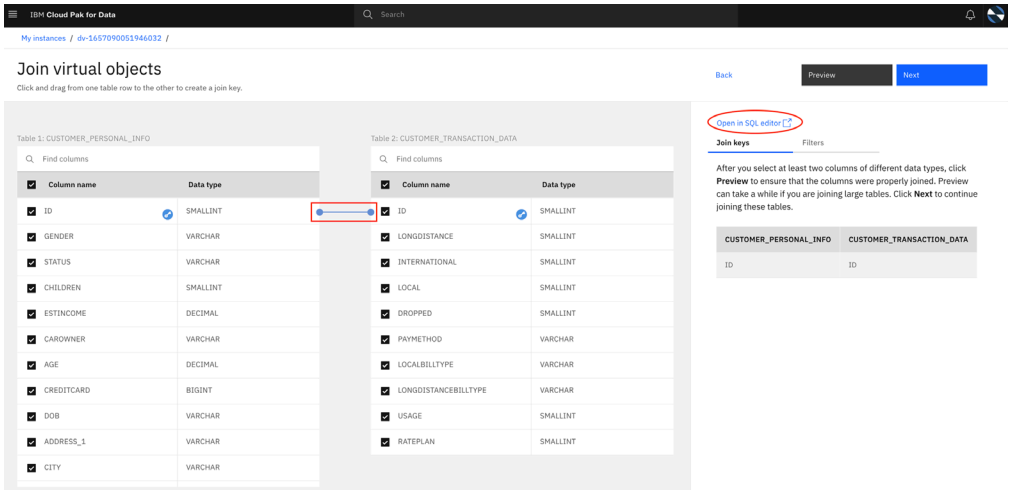
20- Click the Virtualization drop-down menu (annotated with red arrow) and select **Virtualized data** (annotated with red rectangle).



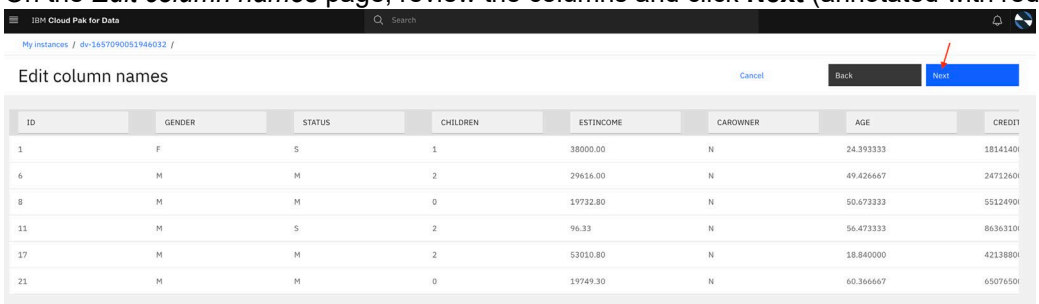
21- Data Vritualizataion also enables you to create virtualized views against the original data sources. To illustrate, check the selection squares (annotated with red arrows) next to the two virtualized data assets and click **Join** (annotated with red oval).



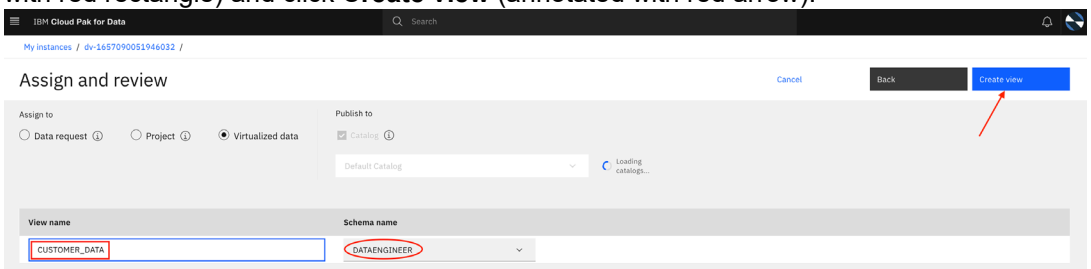
22- On the *Join virtual objects* page, drag and drop a connection (annotated with red rectangle) from the ID field on CUSTOMER_PERSONAL_INFO table to the ID field on CUSTOMER_TRANSACTION_DATA table to create a joined view of these tables using the ID key. For this simple join, you can do so using the UI. To create more complex virtualized joins or views, you can leverage the SQL editor by clicking the **Open in SQL editor** link (annotated with red oval). Click **Next**.



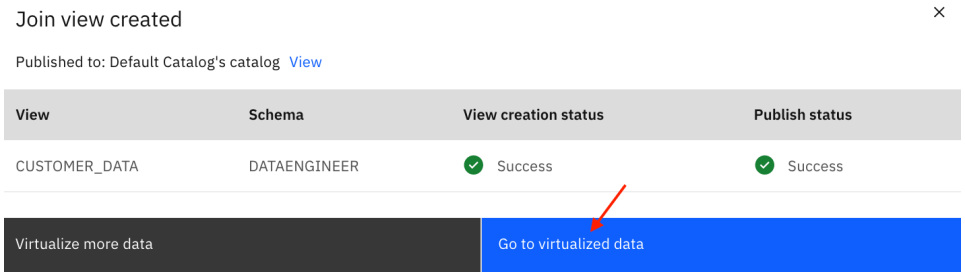
23- On the *Edit column names* page, review the columns and click **Next** (annotated with red arrow).



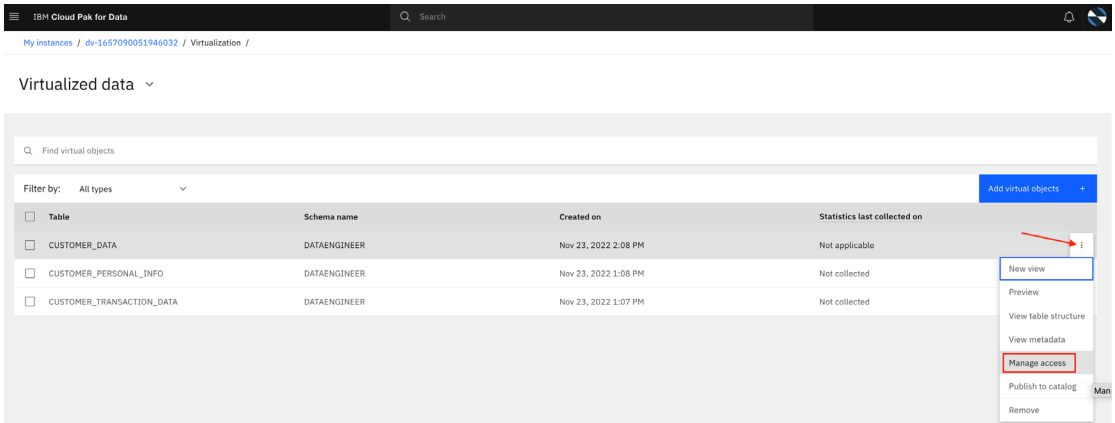
24- On the *Assign and review* page, provide a name for the virtual view, CUSTOMER_DATA (annotated with red rectangle) and click **Create view** (annotated with red arrow).



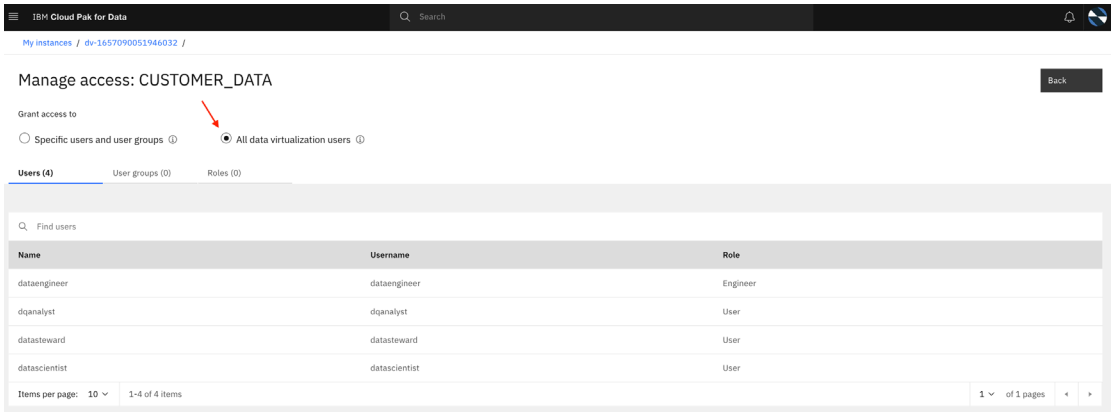
25- On the *Join view created* pop-up, click on **Go to virtualized data** (annotated with red arrow).



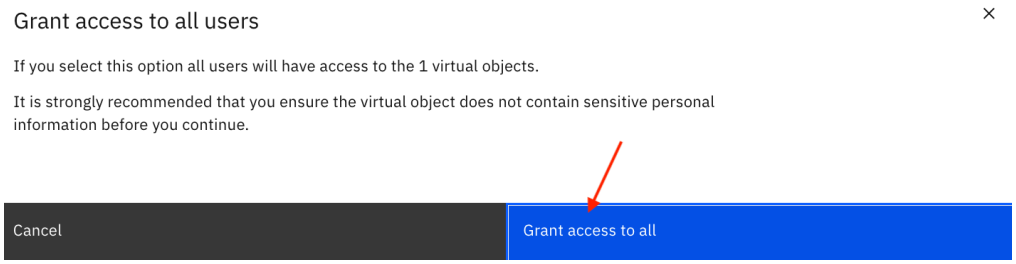
26- On the *Virtualized data* page, click the actions menu (3 vertical dots annotated with red arrow) to the right of CUSTOMER_DATA table, and select **Manage access** (annotated with red rectangle).



27- Select the radio button next to **All data virtualization users** (annotated with red arrow) to provide access to this virtualized table to all users who have access to data virtualization.



28- On the confirmation pop-up window, click **Grant access to all** (annotated with red arrow).



29- Click **Back**. Repeat steps 26-28 to grant access to the other two virtualized tables, **CUSTOMER_PERSONAL_INFO** and **CUSTOMER_TRANSACTION_DATA**.

Summary

In this exercise, you have created the foundations for a governed data fabric by virtualizing your data sources and publishing the virtualized data assets to your catalog which enforces your organization's governance and compliance requirements. Data will be available through the catalog for your organization's consumers to search, find, and leverage in their business intelligence and AI applications.

Appendix

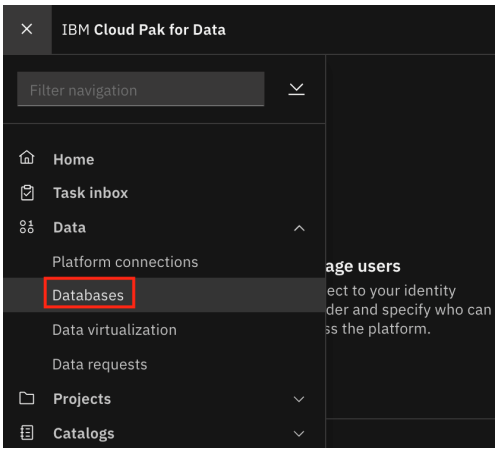
Create Platform Connection

To start with, you will step through the creation of a platform connection to Db2 database running on the same Cloud Pak for Data cluster. This Db2 will include relevant data assets that will be consumed throughout this workshop.

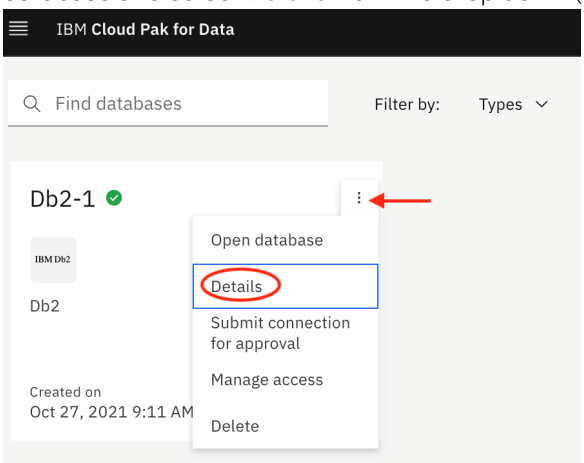
To setup that platform connection to Db2, you need to find the deployment ID for your Db2 instance, by executing the following steps:

1- Navigate to your Cloud Pak for Data url and log in as **admin** user.

2- Navigate to databases by clicking on the Navigation menu (top left hamburger icon) and selecting **Data ➔ Databases** (annotated with red rectangle).

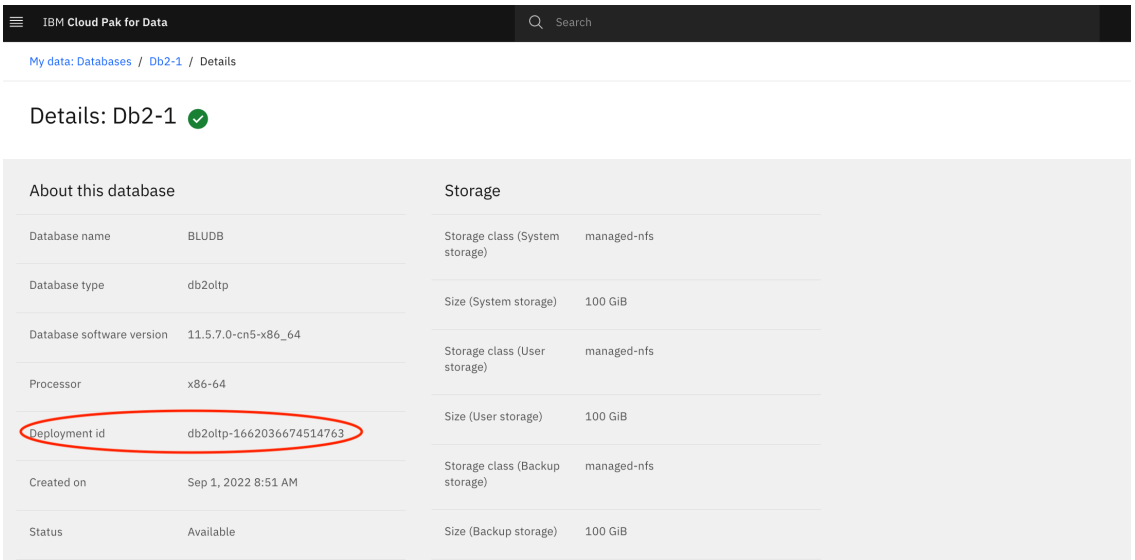


3- Click the **open and close list of options** menu (annotated with red arrow) next to your Db2-1 database and select **Details** from the drop down (annotated with red oval)



4- On the database details page, find the Deployment id. In the example below, the Deployment id is db2oltp-1662036674514763 (annotated with red oval). Your Db2 instance will most likely have

different values.

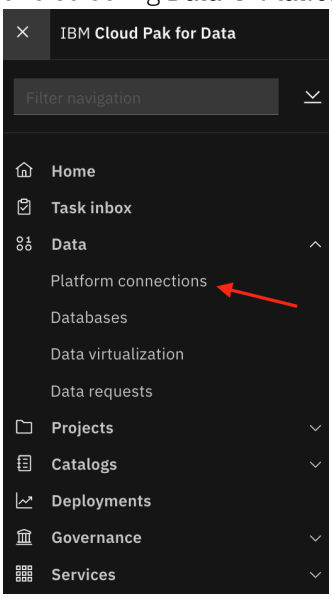


5- Capture the Db2 deployment ID as you will need it later. For the example referenced here, the values would be:

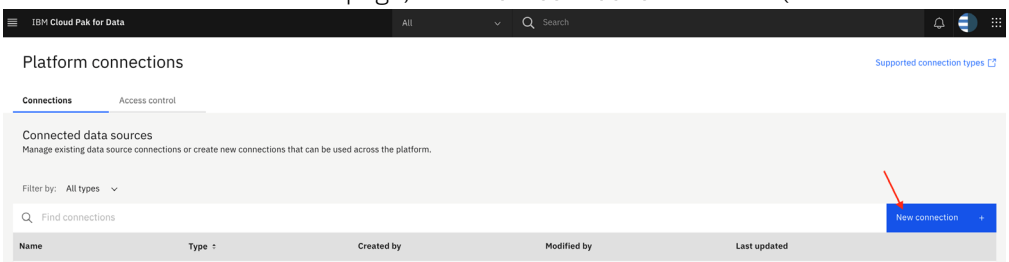
```
"Deployment ID": db2oltp-1662036674514763
```

6- Next, you will create platform connections to access the data assets referenced earlier which are needed for the churn prediction project. Platform connections are available to be consumed by all services of the platform provided the services support the data source type.

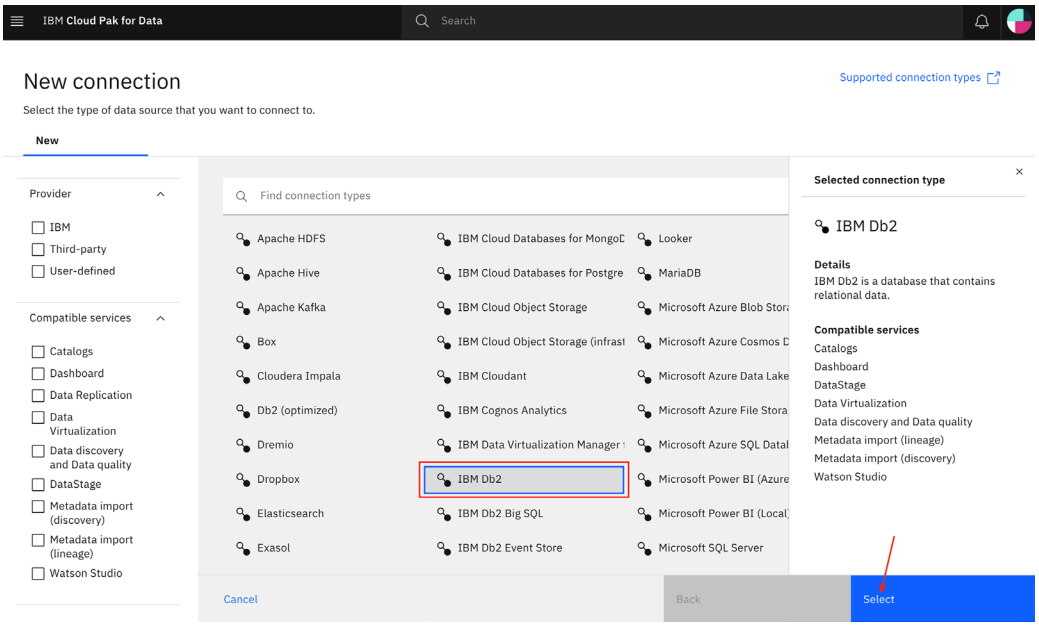
7- Navigate to Platform connections by clicking on the Navigation menu (top left hamburger icon) and selecting **Data → Platform connections** (annotated with red arrow).



8- On the Platform connections page, click **New connection +** button (annotated with red arrow).



9- Select IBM Db2 connection type (annotated with red rectangle) and click **Select** (annotated with red arrow).



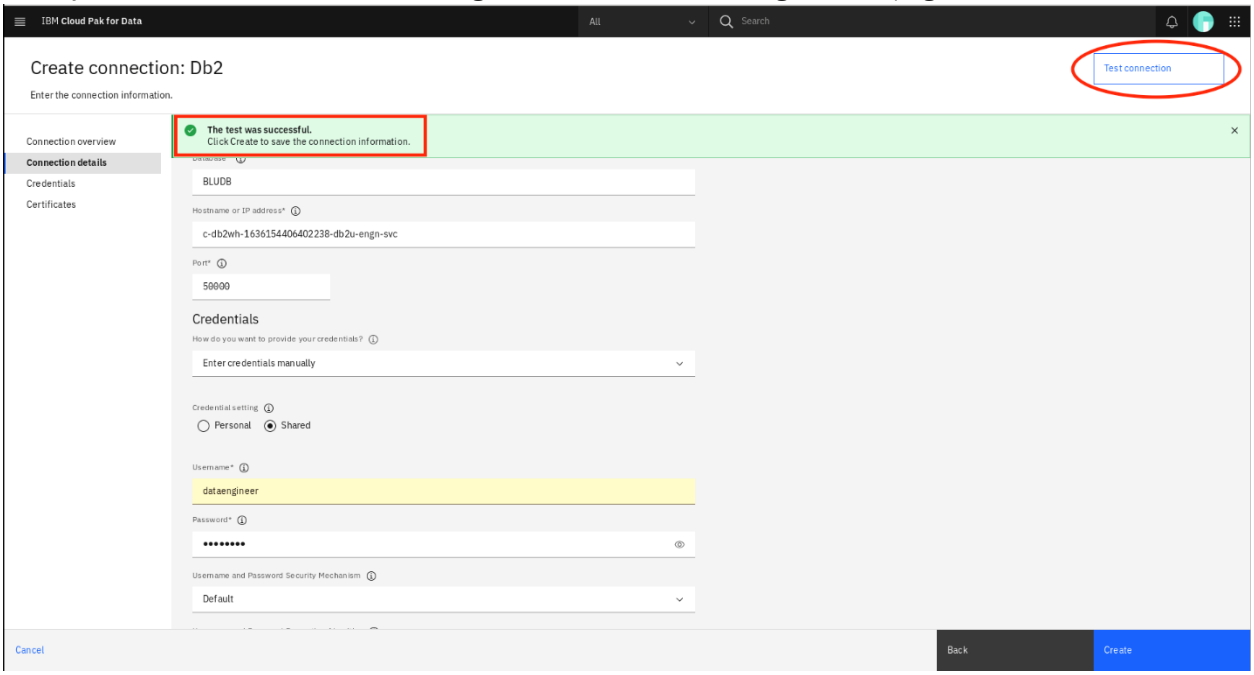
10- Provide a Name <Db2 Customer Personal Information> and an optional Description for the connection and provide the required connection details to access the Db2 instance which were obtained earlier. The username and password credentials should be the credentials for the user who has access to that Db2 instance; in this case, it is the **admin**.

```
"Database": BLUDB,

"Hostname or IP address": c-<YOUR-DEPLOYMENT-ID>-db2u-engn-svc

"port": 50000

"username": admin

"password": your_admin_password
```

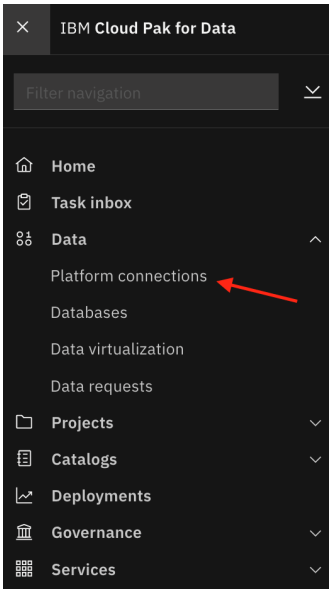
The Hostname or IP address is the Deployment ID you captured earlier, with the “c-“ prefix and “-db2u-engn-svc” suffix.

Make sure the box next to Port is SSL-enabled on the bottom left of the page is unchecked. Then click Test connection (annotated with red oval). You should see a message in green on top of the page confirming that “The test was successful” (annotated with red rectangle).

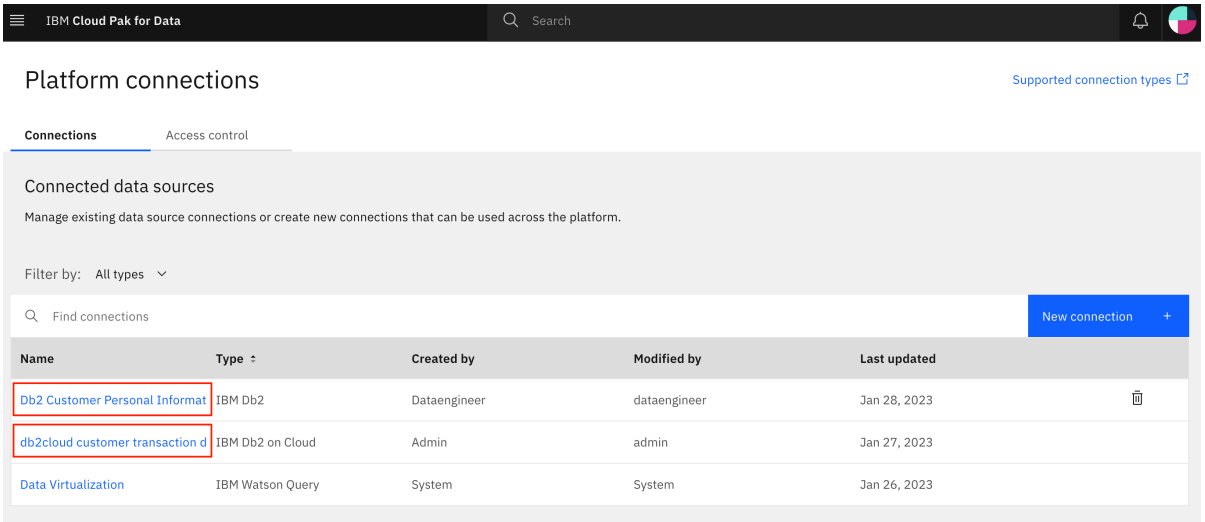
Once you see the successful test message, click **Create** (bottom right of the page).



11- Navigate back to the Platform connections page by clicking on the Navigation menu (top left hamburger icon) and selecting **Data → Platform connections** (annotated with red arrow).



12- You should have two platform connections, the **Db2 Customer Personal Information** connection you just created which connects to the on-prem Db2 database and the **db2cloud customer transaction data** connection which was pre-created for you and that connects to the managed Db2 database on IBM Cloud.


IBM Cloud Pak for Data
Search
Platform connections
Supported connection types
Connections
Access control
Connected data sources
Manage existing data source connections or create new connections that can be used across the platform.
Filter by: All types
Find connections
New connection
Name
Type
Created by
Modified by
Last updated
Db2 Customer Personal Informat
IBM Db2
Dataengineer
dataengineer
Jan 28, 2023
db2cloud customer transaction d
IBM Db2 on Cloud
Admin
admin
Jan 27, 2023
Data Virtualization
IBM Watson Query
System
System
Jan 26, 2023