

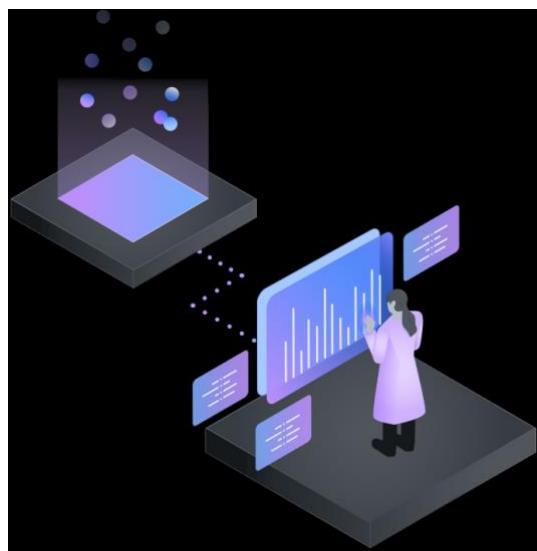
# IBM Journey to Cloud and AI

## Analytics Modernization Workshop

Featuring: IBM Cloud Pak for Data



*Deeper Dive Lab Workbook*



Lab workbook and IBM Cloud Pak for Data workshop design and environment by:  
**Burt Vialpando**, IBM Executive Analytics Architect  
**Kent Rubin**, IBM Solution Architect

September 14, 2020

## Acknowledgements

- **Duane Almeter** and **Eric Watson** for their leadership on the project
- **Daniel Kikuchi** for the CPD 3.0 cluster install environments, and other technical input and support
- **Ed Duhe, Rich Russo** and **Mitchell Odum** for providing the ESX server development platform
- **John Lucas** for product testing, workbook publishing, and project management support
- **John Van Buren** for the Organize lab designs
- **Rajesh Kartha** for Data Virtualization caching lab and DV z/OS lab
- **Benjamin Herta** for the AutoAI non-AVX work around
- **Sidney Phoon** for the last Notebook updates
- **Eric Martens** for the OpenScale lab design
- **Owais Hashmi** for OpenScale storyboarding and lab assistance
- **Rohit Gargate** for OpenScale Auto setup assistance
- **Ben Chard** for providing Analyze lab resources and assistance
- **Tom Konchan** for SPSS/IPS and Decision Optimization lab
- **Daniel Hancock** for the NPS Getting Started lab
- **David Trotter** for the z/OS work on the DV z/OS lab
- **Linda Snow** for providing assistance with complete workshop review



## Table of Contents

<b>LAB 11 COLLECT: DATA VIRTUALIZATION CACHING - DEEPER DIVE .....</b>	<b>7</b>
11.1   Lab overview.....	7
11.2   Personas represented in this lab.....	8
11.3   Logging into the CPD web client (if you have not already done so).....	9
11.4   Reviewing the dashboard: Stock Trading Analysis - Trade Co.....	10
11.5   Virtualizing the remote tables and creating a view .....	11
11.6   Creating a project to work in.....	17
11.7   Creating the Data Virtualization cache .....	26
11.8   Reviewing the Cache Management UI .....	38
11.9   Refreshing the Data Virtualization cache .....	39
11.10   Deactivating/Activating/Deleteing the Data Virtualization cache .....	40
11.11   Caching guidelines.....	41
11.12   Lab conclusion .....	41
<b>LAB 12 COLLECT: VIRTUALIZING &amp; CACHING FROM z/OS – DEEPER DIVE .....</b>	<b>41</b>
12.1   Lab overview.....	41
12.2   IBM DVM .....	41
12.3   Personas represented in this lab.....	45
12.4   Logging into the CPD web client (if you have not already done so).....	45
12.5   Reviewing the dashboard: Stock Trading Analysis - Trade Co.....	46
12.6   Adding the DV connection to DVM for z/OS.....	47
12.7   Remove existing virtual tables and views (if they exist) .....	50
12.8   Virtualizing the remote tables and creating a view .....	52
12.9   Creating the DV data cache.....	60
12.10   Lab conclusion .....	65

<b>LAB 13 ORGANIZE – DEEPER DIVE .....</b>	<b>66</b>
13.1 Lab overview .....	66
13.2 Personas represented in this lab.....	66
13.3 Logging into the CPD web client (if you have not already done so).....	67
13.4 Creating a connection to your data.....	68
13.5 Working in a project.....	73
13.6 Reviewing the business glossary.....	98
13.7 Automation Capabilities – Discovery, Classification, Term Assignment.....	101
13.8 Reviewing Classifications, Data Classes, and Reference Data .....	110
13.9 Transforming Data.....	125
13.10 Data Lineage .....	137
13.11 Lab conclusion .....	139
13.12 Additional Optional Activities .....	140
<b>LAB 14 COGNOS DASHBOARD EMBEDDED - DEEPER DIVE .....</b>	<b>145</b>
14.1 Lab overview .....	145
14.2 Persona represented in this lab.....	145
14.3 Logging into the CPD web client (if you have not already done so).....	145
14.4 Reviewing the dashboard: Monthly Metrics - Trade Co.....	146
14.5 Building the dashboard: Monthly Metrics - Trade Co.....	147
14.6 Lab conclusion .....	161
<b>LAB 15 SPSS USING NPS – DEEPER DIVE .....</b>	<b>162</b>
15.1 Lab overview .....	162
15.2 Persona represented in this lab.....	162
15.3 Logging into the CPD web client (if you have not already done so) .....	162
15.4 Working with NPS Data in SPSS .....	163
15.5 Building, evaluating, and saving the SPSS model.....	167
15.6 Creating and testing an online model deployment.....	170
15.7 Working with NPS Data in Jupyter Notebooks .....	173
15.8 Working with Batch Deployments .....	178
15.9 Lab Conclusion .....	181

<b>LAB 16 Netezza Performance Server (NPS) Getting Started.....</b>	<b>182</b>
16.1 Lab overview .....	182
16.2 Understanding Netezza Users .....	182
16.3 Connecting to the Netezza Performance Server Command Line.....	183
16.4 Checking the state of NPS .....	184
16.5 Setting up the lab database.....	184
16.6 Connecting to the Netezza system database using nzsql .....	185
16.7 Commonly used commands and SQL statements .....	187
16.8 Creating a database .....	191
16.9 Creating a table .....	193
16.10 Loading data into a table .....	195
16.11 Lab conclusion.....	196
<b>Back Page: Notices .....</b>	<b>197</b>
<b>Back Page: Trademarks and Copyrights.....</b>	<b>199</b>

[This page left intentionally blank]

---

## Lab 11 COLLECT: DATA VIRTUALIZATION CACHING - DEEPER DIVE

### 11.1 Lab overview

IBM Cloud Pak for Data (CPD) is a robust, fully integrated platform that will improve business agility and decrease the cost of creating high quality, impactful Analytics.

Data Virtualization (DV), an important component of IBM Cloud Pak for Data, helps integrate data sources across multiple types and locations and turn them into one logical data view. This virtual data lake decreases the amount of time needed to get value out of your data. With data virtualization, you can query data across many source systems without having to copy and replicate the sources into a centralized repository, which saves time and reduces cost. It simplifies your analytics and provides “current state of the business” answers because you’re querying the data at its source with all its respective policies enforced. One can create virtual objects in DV, pointing to the original data sources to be used in queries as if they were local objects. Each query executed against these virtual objects gets efficiently compiled and optimized to work directly with the source data. The query times and performance will be dictated by latency and the workload on the underlying data sources, especially while performing aggregated queries across different data sources.

With Data Virtualization Caching, it is possible to avoid trips to the original source(s) and create a temporary cache of the data. This helps immensely with query performance as the DV optimizer can choose to use the local cached copy instead of querying the remote sources repeatedly.

In this lab, you will create Data Virtualization objects that will be used within a dashboard to perform analysis followed by using the caching capability of DV to experience how query times can be drastically reduced. You will also learn how caches can be refreshed, disabled, enabled and deleted, and the best practices and guidelines to create DV Caches.

In our scenario, the Trade Co. Business Analyst (BA) initially creates a dashboard using data from a DV view, which in turn points to multiple remote data sources underneath. But rendering these dashboards and using them for analysis takes a very long time since the data has to be retrieved every time from those remote data sources. The BA works with the DV Admin, who helps create a cache for the virtualized view to help address the data retrieval performance issues observed while working with the dashboard.

## 11.2 Personas represented in this lab

The [Business Analyst](#) persona along with the [Data Engineer](#) and [Data Virtualization Administrator](#) personas will perform the exercises in this lab.

Persona (Role)	Capabilities
 Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.
 Administrator (DV/CPD)	Administrators set up and maintain the DV module within the CPD environment itself. They are responsible for granting DV access to users and administration tasks such as creating a data cache.

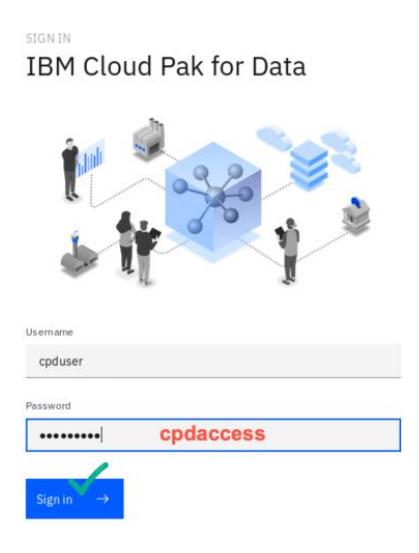
For convenience in doing this lab, instead of switching between personas, all the required privileges have been provided to the same user. The workbook will refer to the respective personas at different stages to help understand the flow of this task.

### 11.3 Logging into the CPD web client (if you have not already done so)

- \_\_1. If you are starting this lab stand-alone (without going through previous labs) do the following:
- \_\_2. Double-click the desktop icon: [Cloud Pak for Data Web Client](#).



- \_\_3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click **Sign in**.



The screenshot shows the 'SIGN IN' page for the IBM Cloud Pak for Data web client. At the top, it says 'IBM Cloud Pak for Data'. Below that is a decorative graphic showing a central blue cube with various icons like people, servers, and charts connected to it. The main form has two input fields: 'Username' containing 'cpduser' and 'Password' containing 'cpdaccess'. A green checkmark icon is positioned next to the 'Sign in' button.

SIGN IN
IBM Cloud Pak for Data

Username <input type="text" value="cpduser"/>
Password <input type="password" value="cpdaccess"/>
 <b>Sign in</b> →

## 11.4 Reviewing the dashboard: Stock Trading Analysis - Trade Co.

In an attempt to understand stock trading patterns, the business analysis for Trade Co. starts by creating a simple dashboard to find the most popular historically traded stocks. The dashboard shows the number of Shares Sold vs. the number of Daily Trades.



The Business Analyst (BA) works with the Data Engineer (DE) to get virtualized access to the different data sources required for creating the dashboard(s). The DE creates the Data Sources, Virtual Tables and finally a Virtualized View joining all the Virtual Tables and computing the basic aggregations required. The View is then shared with the BA, who can then proceed with creating the dashboard.

However, once the dashboard is initially created, the BA notices delays in rendering it. Since every request from the dashboard has to fetch the data from its original source(s), latency starts to play an important role, slowing down response. The BA works with the Data Virtualization Admin to create a cache for the View, which helps speed up query times significantly and hence rendering of the dashboard(s).

## 11.5 Virtualizing the remote tables and creating a view

The Data Virtualization process begins by adding data sources to virtualize and is typically done by the Data Engineer.



### 11.5.1 Navigate to Data virtualization

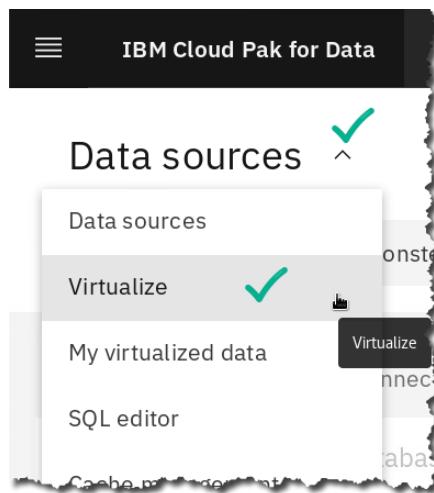
4. Start at the [Navigation Menu](#).  
Click [Collect](#) ⇒ [Data virtualization](#).



### 11.5.2 Creating virtualized tables

5. With the data sources successfully created, the next step is to virtualize the tables needed for this exercise.

From the top left go to [Data Sources](#) ⇒ [Virtualize](#).



- \_\_6. In the search bar, enter the string STOCK and click the search icon (magnifying glass).

The screenshot shows the 'Virtualize' interface with the 'Tables' tab selected. A search bar at the top contains the text 'stock' with a magnifying glass icon to its right, which has a green checkmark over it. Below the search bar, the text 'Showing matched tables: 12/554' is displayed. The main area lists several table names, with 'CUSTOMER\_TRANSACTIONS' and 'STOCK\_SYMBOLS' being the most prominent.

- \_\_7. Click the gear, then select Group tables with identical names.

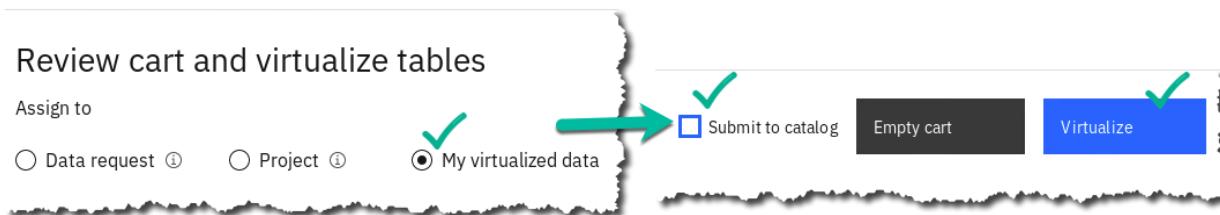


This uses the DV feature of “Schema Folding” which groups tables with the same names across data sources and shows them as a single “table” selection. (Notice how two tables called **CUSTOMER\_TRANSACTIONS** from two different databases are treated as one.)

- \_\_8. Select tables: **CUSTOMER\_TRANSACTIONS** and **STOCK\_SYMBOLS**  $\Rightarrow$  Click **Add to cart**  $\Rightarrow$  **View cart**.

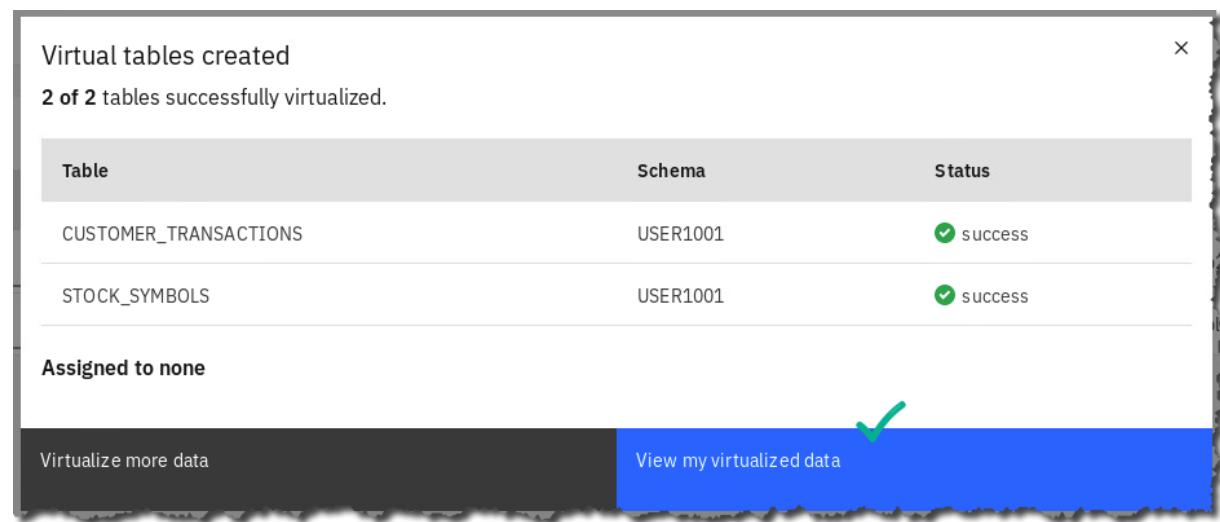
The screenshot shows a list of tables on the left with checkboxes next to them. Two checkboxes are checked: 'CUSTOMER\_TRANS...' and 'STOCK\_SYMBOLS'. A green arrow points from the 'Add to cart' button in a central overlay to a 'View cart (2)' button at the bottom right. Both the 'Add to cart' button and the 'View cart' button have green checkmarks over them.

- \_\_9. In the [Review cart and Virtualize tables](#) section, select button [My virtualized data](#)  $\Rightarrow$  Uncheck box [Submit to catalog](#)  $\Rightarrow$  Click [Virtualize](#).



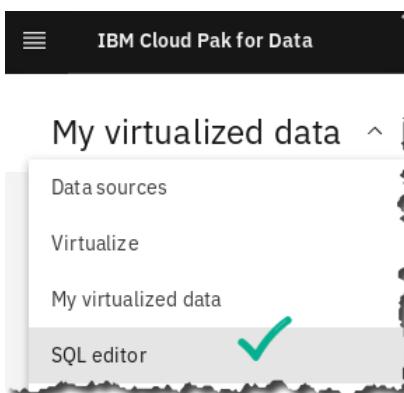
- \_\_10. Notice the **two** virtual tables you just created, which are actually comprised of three different tables in three databases in the IBM Cloud. The CUSTOMER\_TRANSACTIONS virtual table is a schema folded table pointing to **two** remote Db2 Warehouse tables and the STOCK\_SYMBOLS virtual table points to **single** table in the MySQL database.

Click [View my virtualized data](#).



- \_\_11. You will now be creating a virtualized view of these virtualized tables using the SQL editor.

Select [Menu \(My virtualized data\)](#)  $\Rightarrow$  [SQL editor](#).



\_\_12. Copy and paste the SQL below into the SQL editor.

 Data Engineer	<p>Note: You can copy and paste this SQL directly from the Unified Desktop by using the File Browser to open this file:</p> <p> ⇒ Home ⇒ Downloads ⇒ DVCaching_SQL.txt.</p> <p>Alternately, you can download this file by opening a browser tab and using this link: <a href="https://ibm.box.com/v/Workshop-V9-DVCaching-SQL">https://ibm.box.com/v/Workshop-V9-DVCaching-SQL</a></p>
---	---

```
DROP VIEW VIEW_CUST_TXN_SYMBOL_COM;
CREATE VIEW VIEW_CUST_TXN_SYMBOL_COM
AS
SELECT
"SYM"."COMPANY",
"SYM"."SYMBOL",
"CUSTID",
"TOTAL_QUANTITY",
"TXN_COUNT"
FROM
"USER1001"."STOCK_SYMBOLS" "SYM",
(SELECT
"USER1001"."CUSTOMER_TRANSACTIONS"."CUSTID" "CUSTID",
"USER1001"."CUSTOMER_TRANSACTIONS"."SYMBOL" "SYMBOL",
"USER1001"."CUSTOMER_TRANSACTIONS"."TRANSACTION_DATE"
"TRANSACTION_DATE",
SUM("USER1001"."CUSTOMER_TRANSACTIONS"."UNITS_TRADED") as
"TOTAL_QUANTITY",
COUNT(*) as "TXN_COUNT"
FROM
"USER1001"."CUSTOMER_TRANSACTIONS"
GROUP BY CUSTID,SYMBOL,TRANSACTION_DATE) "ST"
WHERE RTRIM("SYM"."SYMBOL")= RTRIM("ST"."SYMBOL");
```

13. Click **Run all** which should create the virtualized view successfully.

```
* Untitled - 1
1
2  DROP VIEW VIEW_CUST_TXN_SYMBOL_COM;
3  CREATE VIEW VIEW_CUST_TXN_SYMBOL_COM
4  AS
5  SELECT
6    "SYM"."COMPANY",
7    "SYM"."SYMBOL",
8    "CUSTID",
9    "TOTAL_QUANTITY",
10   "TXN_COUNT"
11  FROM
12  "USER1001"."STOCK_SYMBOLS" "SYM",
13  (SELECT
14    "USER1001"."CUSTOMER_TRANSACTIONS"."CUSTID" "CUSTID",
15    "USER1001"."CUSTOMER_TRANSACTIONS"."SYMBOL" "SYMBOL",
16    "USER1001"."CUSTOMER_TRANSACTIONS"."TRANSACTION_DATE" "TRANSACTION_DATE",
17    SUM("USER1001"."CUSTOMER_TRANSACTIONS"."UNITS_TRADED") as "TOTAL_QUANTITY",
18    COUNT(*) as "TXN_COUNT"
19  FROM
20  "USER1001"."CUSTOMER_TRANSACTIONS"
21  GROUP BY CUSTID,SYMBOL,TRANSACTION_DATE) "ST"
22  WHERE RTRIM("SYM"."SYMBOL") = RTRIM("ST"."SYMBOL");
23
```

**Run all**  Remember my selection

Note: The first time this script is run, the **DROP VIEW** statement will fail since the view does not exist. It will execute cleanly on subsequent runs.

Result - Jul 23, 2020 7:03:49 PM

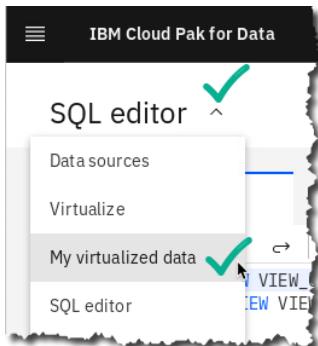
This is OK on the first run

```
Result - Jul 23, 2020 7:03:49 PM
^  ● DROP VIEWVIEW_CUST_TXN_SYMBOL_COM
Status: Failed
Error message
"USER1001.VIEW_CUST_TXN_SYMBOL_COM" is an undefined name.. SQLCODE=-204, SQLSTATE=42704, DRIVER=4.2.6.14
Learn more about this error
```

```
^  ✓ CREATEVIEWVIEW_CUST_TXN_SYMBOL_COM AS SELECT "SYM"."COMPANY", ...
Run time: 0.217 s
```

\_\_14. Preview the contents of the virtualized view to confirm the data is being correctly retrieved.

Go to [Menu \(SQL editor\)](#)  $\Rightarrow$  [My virtualized data](#).



\_\_15. On the view you just created, [VIEW\\_CUST\\_TXN\\_SYMBOL\\_COM](#), click the ellipses at the end of that line ... and then select [Preview](#).



 Data Engineer	<p>It may take a minute or two for the view to render.</p> <p>If this step fails, check to make sure that you have copied the SQL code properly into the SQL editor and run it again. Keep in mind that a successful execution of the SQL may not necessarily mean the view was actually created properly.</p>
--	--

\_\_16. The Virtualized View preview should render as below.

VIEW_CUST_TXN_SYMBOL_COM				
Created on: Jul 29, 2020 7:17:28 AM				
<a href="#">Preview</a> <a href="#">Table structure</a> <a href="#">Metadata</a>				
5 Columns				
COMPANY	SYMBOL	CUSTID	TOTAL_QUANTITY	
VAR CHAR	VARCHAR	SMALLINT	INTEGER	
Apple Inc.	AAPL	0	19	
Apple Inc.	AAPL	0	8	
Apple Inc.	AAPL	0	31	
Apple Inc.	AAPL	0	87	
Apple Inc.	AAPL	0	99	
Apple Inc.	AAPL	0	7	

## 11.6 Creating a project to work in

The Virtualized View referencing different data sources is now created. The Data Engineer will share the Virtualized View with the Business Analyst (BA) who can then work on creating the dashboard.

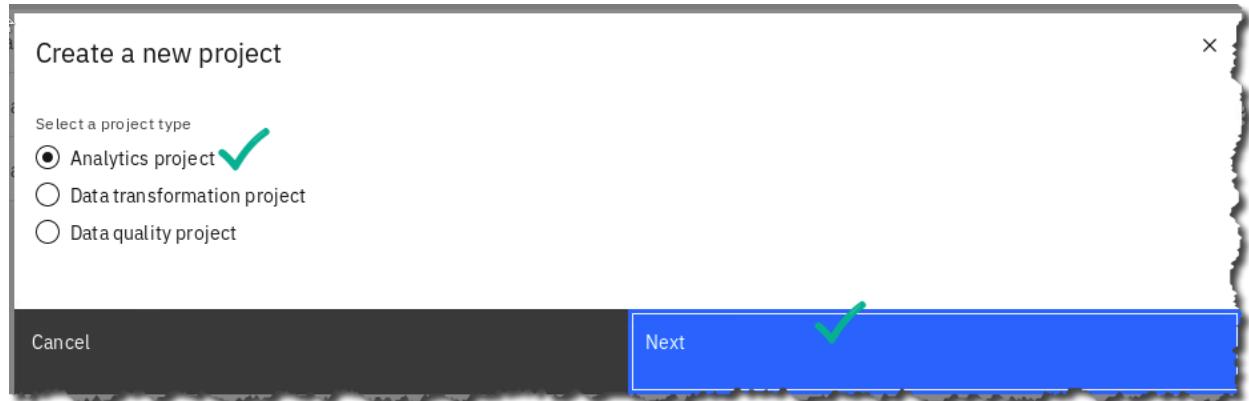


With the access to the Virtualized View in place, the BA begins to work on the dashboard by creating an Analytics Project.

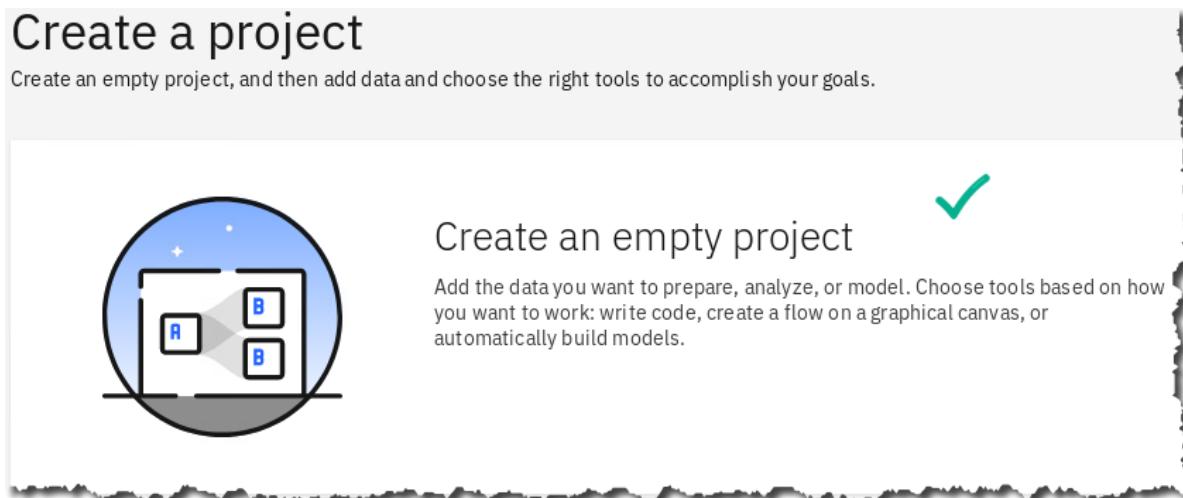
- 17. Click [Navigation Menu](#) ⇒ [Projects](#) ⇒ [New project](#).



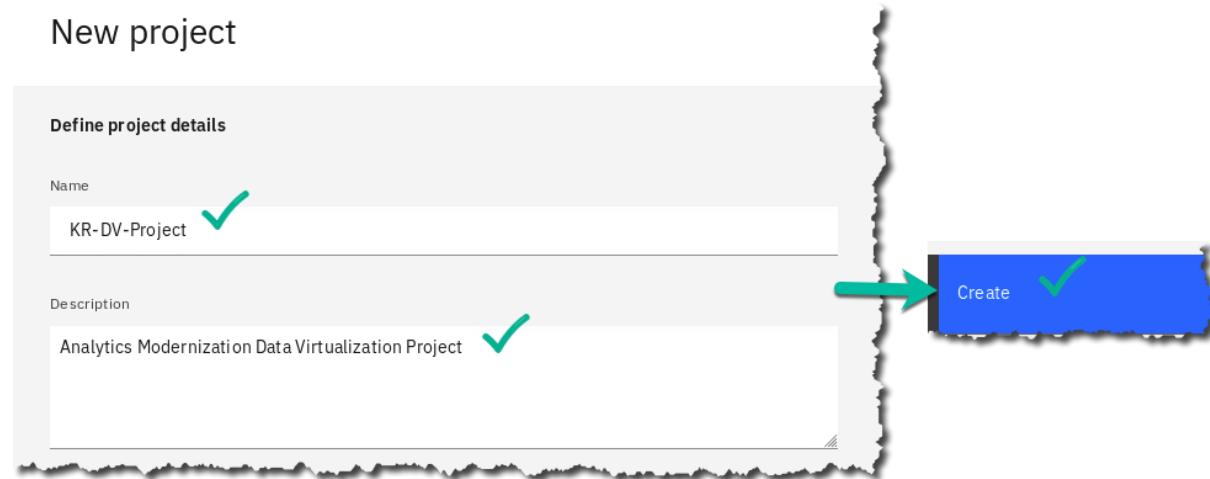
- 18. Choose [Analytics project](#) ⇒ [Next](#).



- \_\_19. Click on tile [Create an empty project](#).

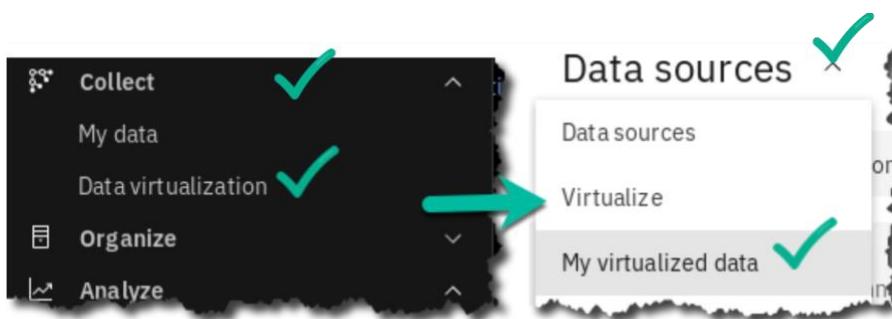


- \_\_20. Provide a name of '[your initials-DV-Project](#)' and description for the project and click [Create](#).

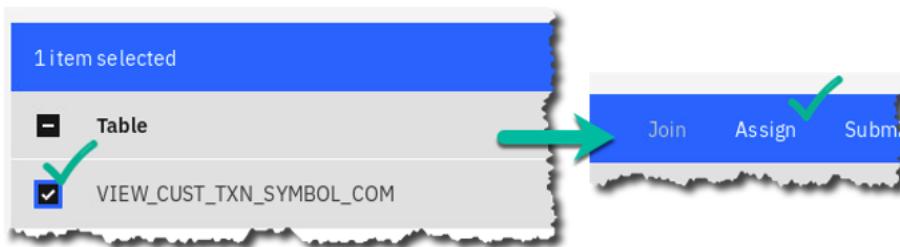


#### 11.6.1 Adding the previously created DV view to the new project

- \_\_21. Start at the [Navigation Menu](#) ⇒ Click [Collect](#) ⇒ [Data virtualization](#) ⇒ [Menu \(Data Sources\)](#) ⇒ [My virtualized data](#).



- \_\_22. Check [VIEW\\_CUST\\_TXN\\_SYMBOL\\_COM](#) ⇒ click Assign.



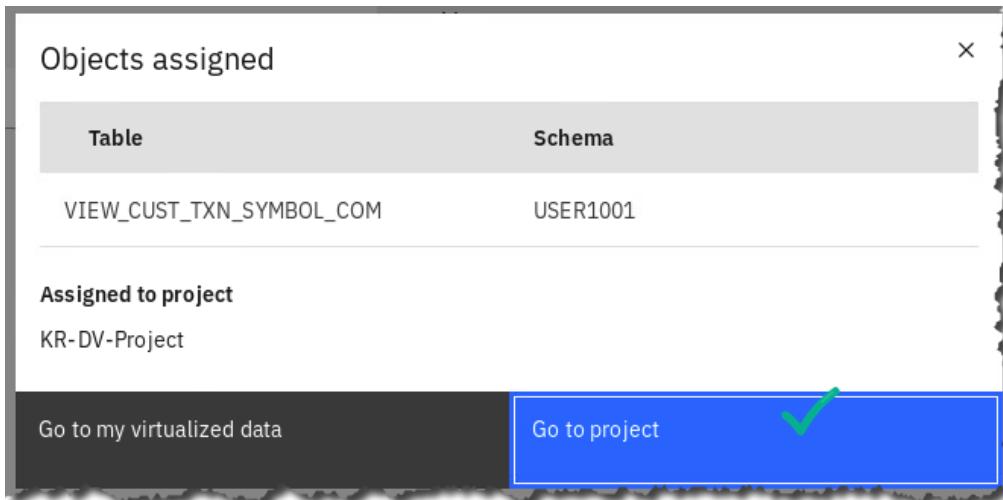
- \_\_23. Choose [your\\_initials-DV-Project](#) (whatever you have named it) ⇒ Assign.

### Assign virtual objects



- \_\_24. The DV View is now successfully assigned to the project.

Click [Go to project](#).



\_\_25. On the Project page, click the [Assets](#) tab.

Review the Data assets section where the newly assigned virtualized view should now show up, along with the Connection it needs in order to render it from the data sources. (Your connection Name will vary from what is shown below.)

My projects / KR-DV-Project

Overview Assets Environments Jobs

What assets are you looking for?

▼ Data assets

0 assets selected.

Name

01 USER1001.VIEW\_CUST\_TXN\_SYMBOL\_COM

DS15955493260020404

#### 11.6.2 Creating a dashboard within the project to use the data asset:

\_\_26. Click [+ Add to project](#) and choose [Dashboard](#) as the asset type.

Add to project +

Choose asset type

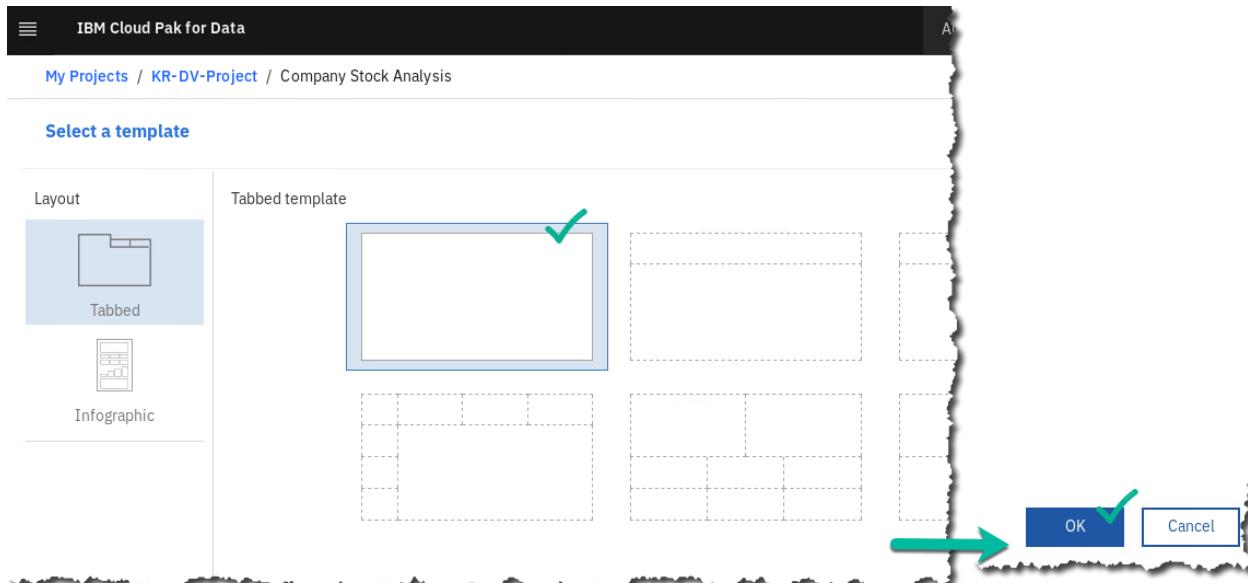
Available asset types

Data	Connection	Connect
Notebook	Dashboard	Watson
Data Refinery flow	Decision Optimizati...	

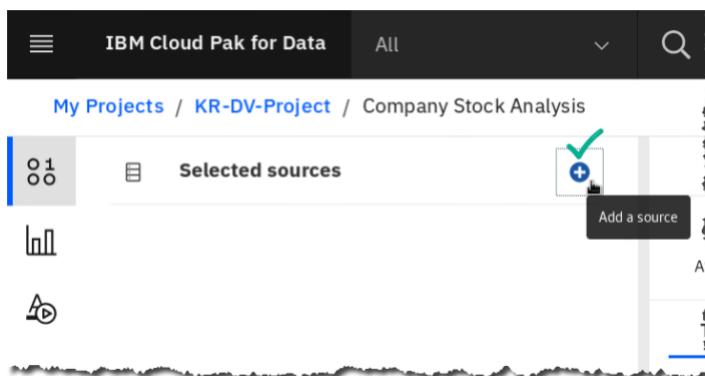
\_27. Under **Name** type: **Company Stock Analysis**  $\Rightarrow$  **Create**.



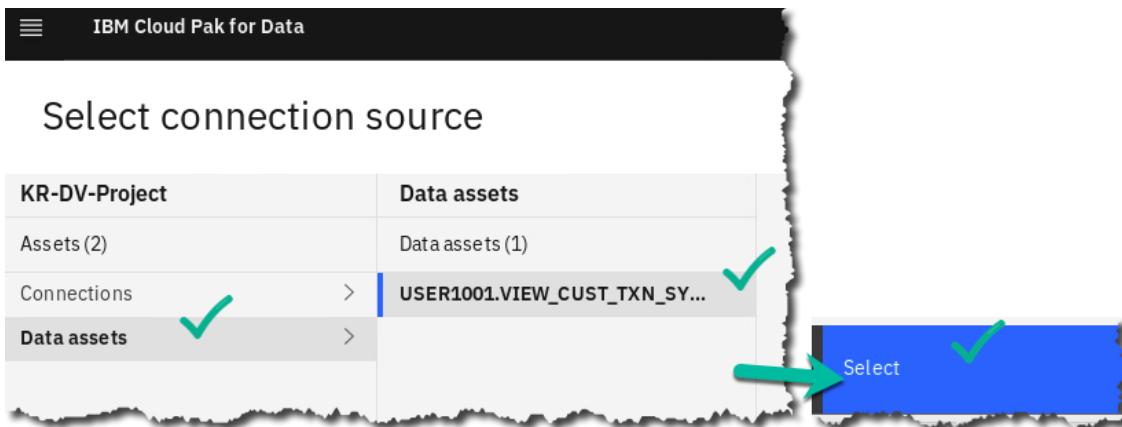
\_28. Choose the **Freeform** (default) layout  $\Rightarrow$  **OK**.



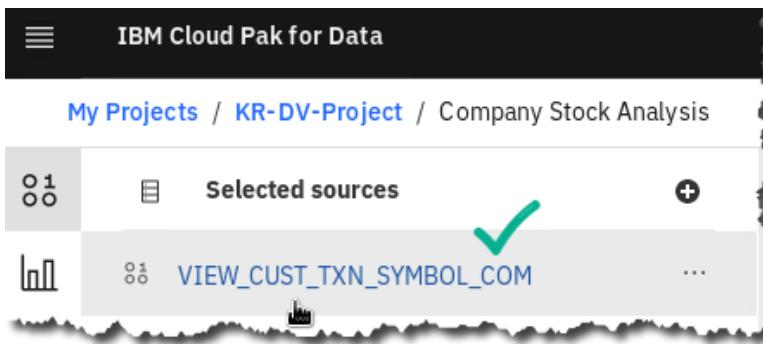
\_29. On top left of the dashboard canvas in **Selected sources** and click **+** (Add a source).



- \_\_30. Click **Data Assets** ⇒ **USER1001.VIEW\_CUST\_TXN\_SYMBOL\_COM** ⇒ **Select**.

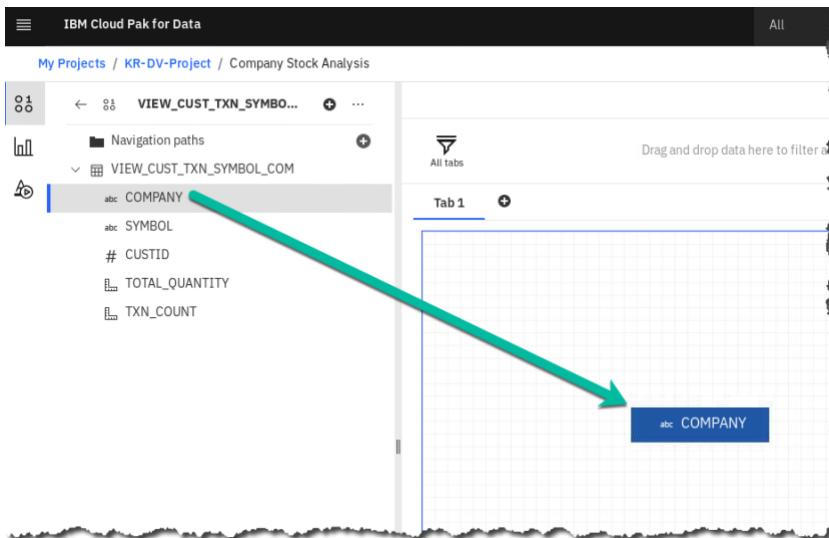


- \_\_31. Click the newly added data source to expand the view to see the columns available to be used in the dashboard.



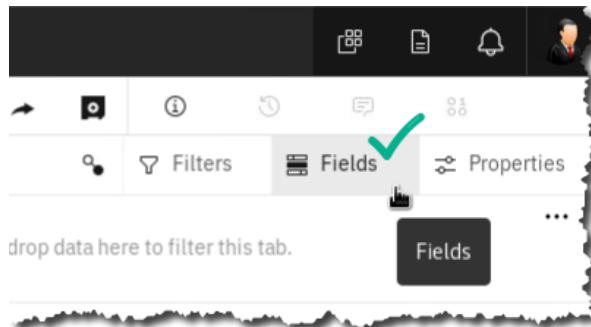
- \_\_32. Drag and drop the **COMPANY** column on the canvas.

(It may take a minute or two for this column to render as a widget – please be patient.)



\_\_33. Make sure to select the **COMPANY** widget on the canvas.

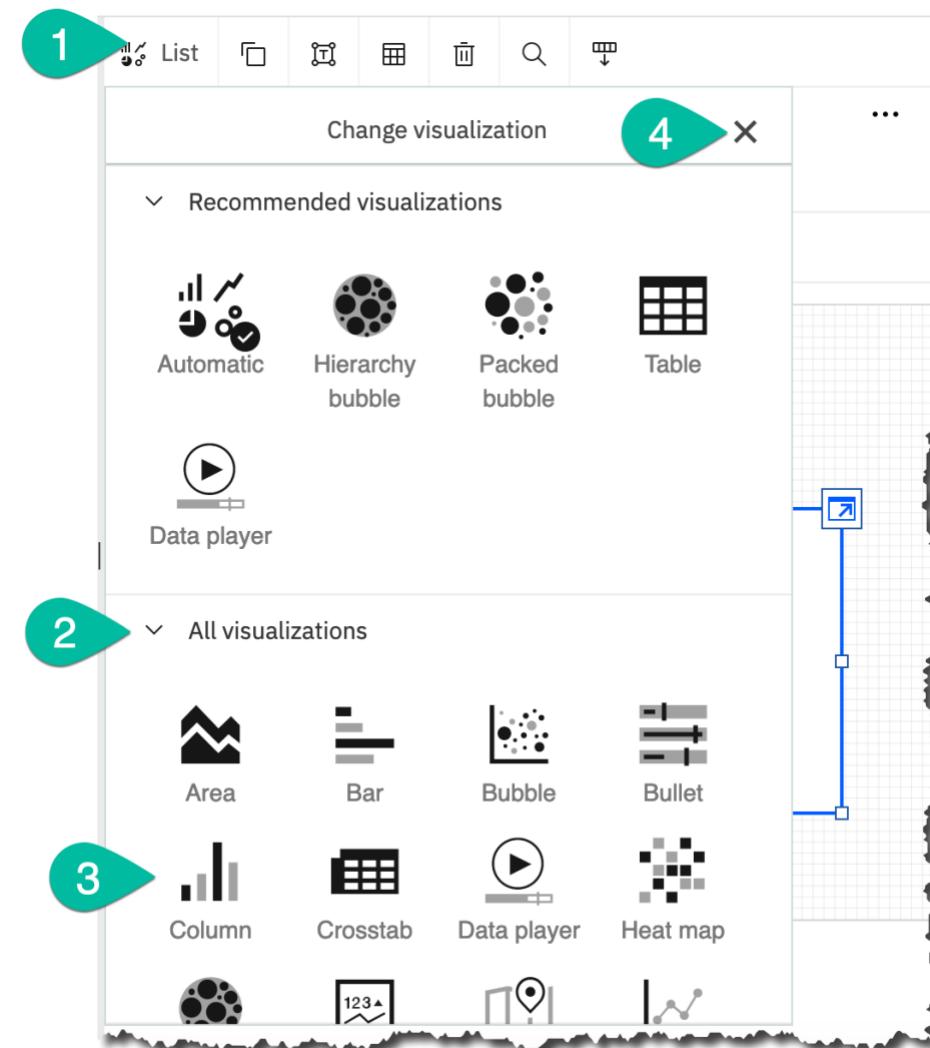
Then at the top right select the tab **Fields**.



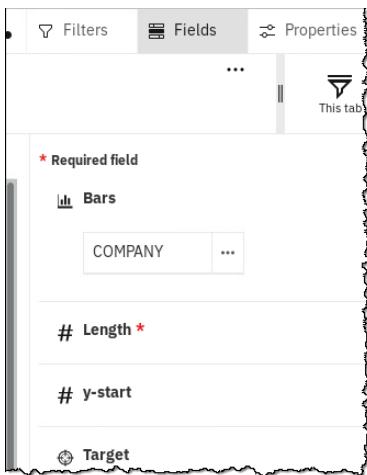
\_\_34. On the top left (above Tab 1) change visualization type from list to column by selecting:

**List box**  $\Rightarrow$  **All visualizations**  $\Rightarrow$  **Column**.

Then close the visualization select box **X**.



\_\_35. On the right of the screen, you should see this visualization formatting box.

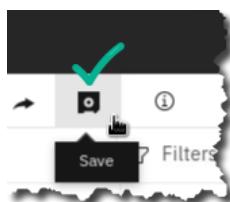


\_\_36. Format the visualization by performing these steps, in order:

1. Make sure the **COMPANY** column is used for **Bars** (if it is not, drag it there).
2. Drag and drop the **TXN\_COUNT** column or **Color**.
3. Drag and drop the **TOTAL\_QUANTITY** column for **# Length**.

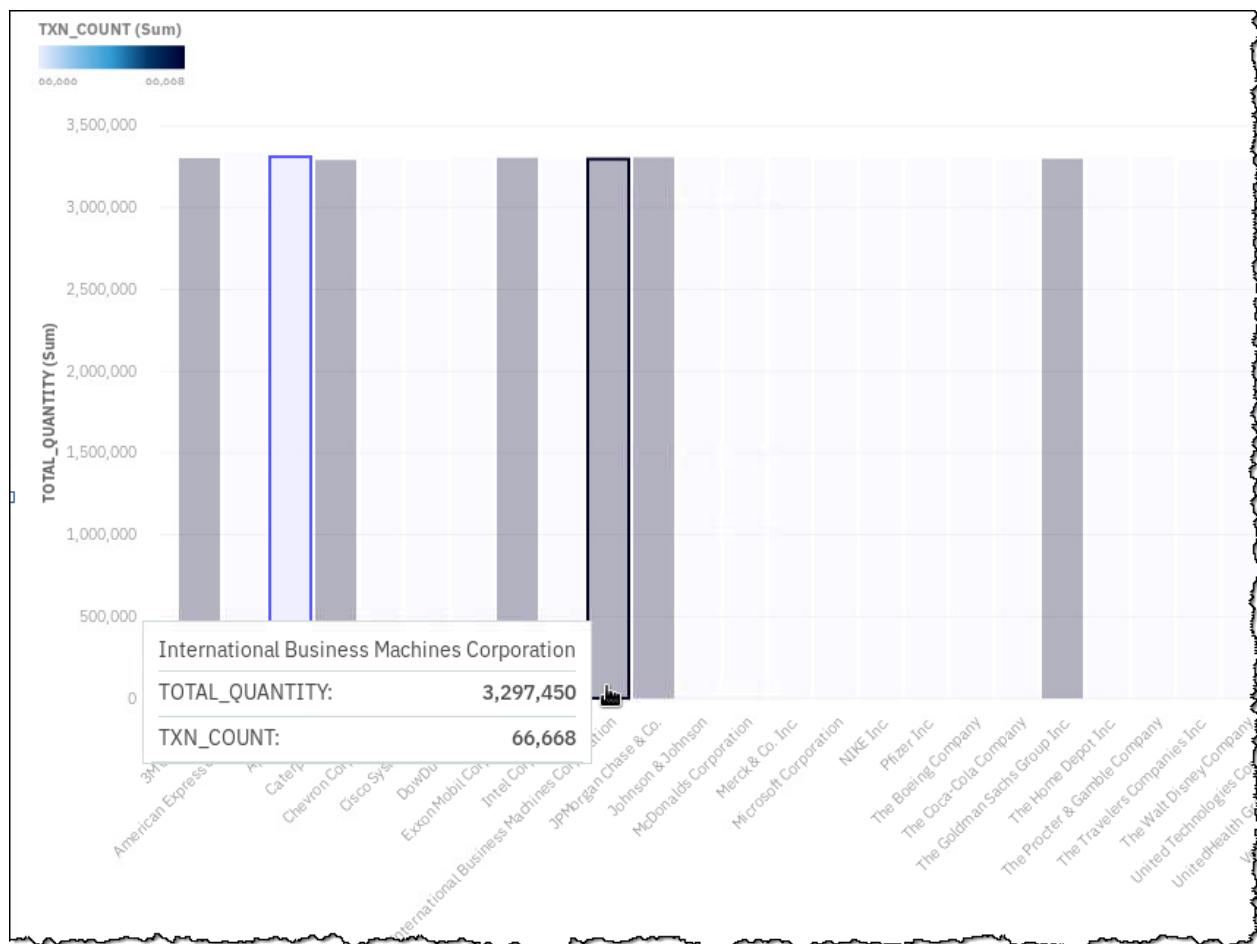
The screenshot shows the 'Fields' tab of the configuration interface. On the left, there's a sidebar with navigation paths and a list of columns: 'VIEW\_CUST\_TXN\_SYMBOL...' (highlighted), 'Navigation paths', 'VIEW\_CUST\_TXN\_SYMBOL\_COM' (expanded), 'COMPANY', 'SYMBOL', '# CUSTID', 'TOTAL\_QUANTITY', and 'TXN\_COUNT'. Three green arrows point from these columns to their corresponding fields in the main configuration box: 'COMPANY' to the 'Bars' section, 'TOTAL\_QUANTITY' to the '# Length \*' section, and 'TXN\_COUNT' to the 'Color' section.

\_\_37. Click the **Save** icon to ensure your work is saved.



- \_\_38. Perform some analysis of various companies to see the quantity traded and the transaction counts by hovering on an individual company from within the visualization.

Example: hover over the column for [International Business Machines Corporation](#).



 Business Analyst	<p>I noticed a delay while retrieving my data in this dashboard which in the long run will be unsuitable for me to use repeatedly.</p> <p>I need to reach out to my CPD Administrator and Data Engineer to find a way to make this perform better.</p>
---	--

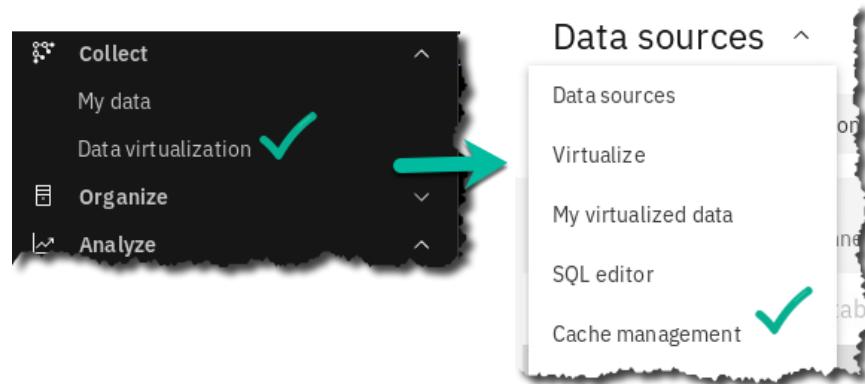
## 11.7 Creating the Data Virtualization cache

### 11.7.1 Creating the DV data cache

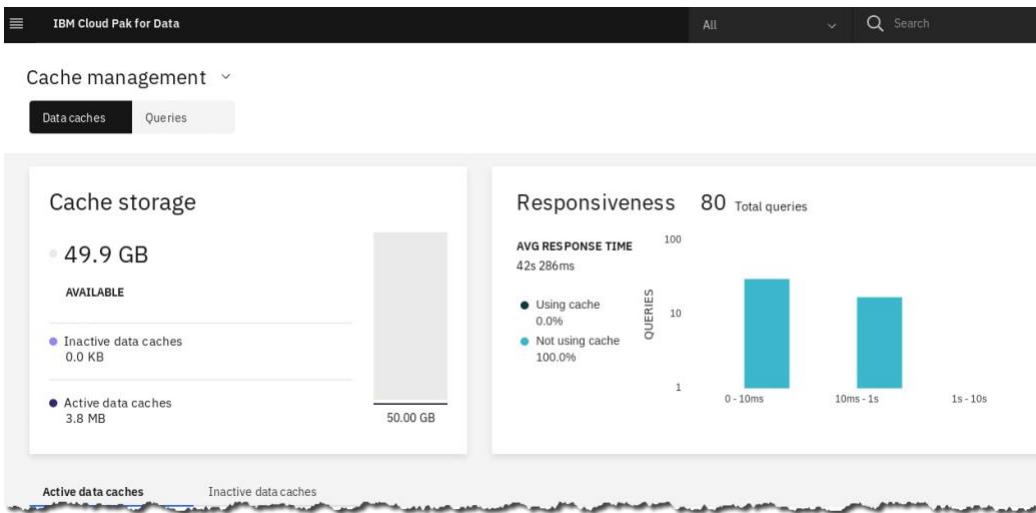


The DV Admin will first analyze the queries executed against DV and identify those generated by the dashboard. Looking at the query will help understand what type of cache needs to be created to improve performance for a better dashboard experience.

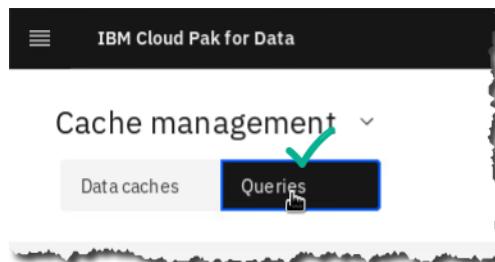
- 39. Start at the [Navigation Menu](#).
- 40. Click [Collect](#)  $\Rightarrow$  [Data virtualization](#)  $\Rightarrow$  [Menu \(Data Sources\)](#)  $\Rightarrow$  [Cache Management](#).



- 41. You will be presented with your [Cache management](#) dashboard.



\_\_42. Click tab **Queries**.



\_\_43. At the far right of the screen select **Filter**  $\Rightarrow$  Not using cache  $\Rightarrow$  Response time: 1m-10m  $\Rightarrow$  Last 24 hours  $\Rightarrow$  **Apply**.



- 44. The list of queries executed are shown with their average execution times if run more than once. Note the response time which explains the delay in working within the dashboard. (The example below is greater than 1 minute and 10 seconds)

Select the **down arrow** by first **Query ID** to see the actual SQL being run.

NOTE: If the query you chose does not contain '**SELECT VIEW\_CUST\_TXN\_SYMBOL\_COM**' at the start of the query, use the back-arrow key to return to the cache, scroll down and view another query.

Query ID	Total executions	Caches used	Avg response time with cache	Avg response time without cache
-3306915924412446352 USER1001	1	0	N/A	1 min, 10.90 sec
<pre> 1 SELECT "VIEW_CUST_TXN_SYMBOL_COMO"."COMPANY" AS "id_2065813104", COUNT("VIEW_CUST_TXN_SYMBOL_COMO"."COMPANY") AS "id1030785780", 2 SUM("VIEW_CUST_TXN_SYMBOL_COMO"."TOTAL_QUANTITY") AS "id105534131", SUM("VIEW_CUST_TXN_SYMBOL_COMO"."TXN_COUNT") AS 3 "id_1809216723" FROM "USER1001"."VIEW_CUST_TXN_SYMBOL_COM" "VIEW_CUST_TXN_SYMBOL_COMO" GROUP BY 4 "VIEW_CUST_TXN_SYMBOL_COMO"."COMPANY" ORDER BY "id_2065813104" ASC FETCH FIRST 3001 ROWS ONLY FOR FETCH ONLY 5 6 7 8 9 10 </pre>				

**View details** ✓

Click **View details**.

Cache management / Query details

Query details

Query ID: -3306915924412446352

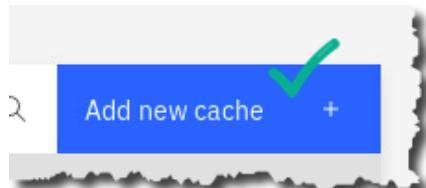
Copy this number

One could choose to create a cache from this particular query, but since the dashboards will be using all columns from the View created anyway, the DV Admin and the BA decides to create a cache for the entire virtualized view instead.

- \_\_46. Go back the main [Cache Management](#) page and click [Data caches](#).



- \_\_47. At the right of the page, click [Add new cache](#).



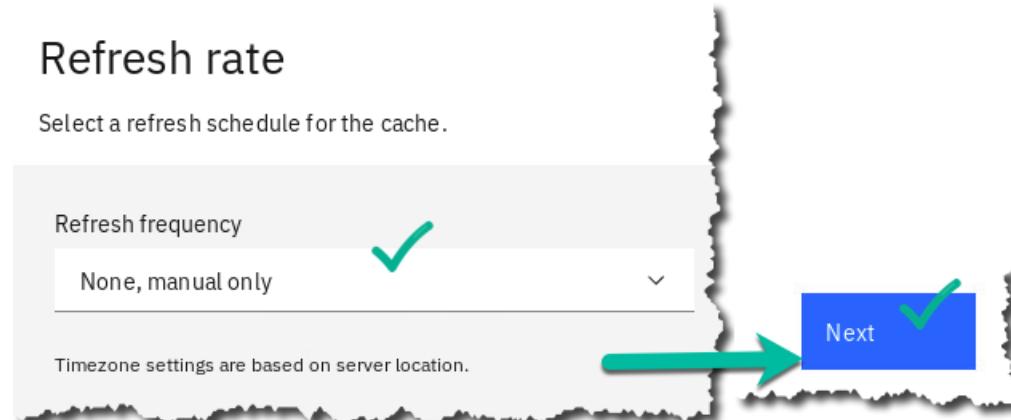
- \_\_48. This opens a SQL Editor to enter the query from which the cache will be created.  
Type in the query to select all columns from the virtualized view:

```
SELECT * FROM "USER1001"."VIEW_CUST_TXN_SYMBOL_COM"
```

Click [Next](#).



- \_\_49. In the [Refresh rate](#) choose the [default \(None, manual only\)](#) and click [Next](#).

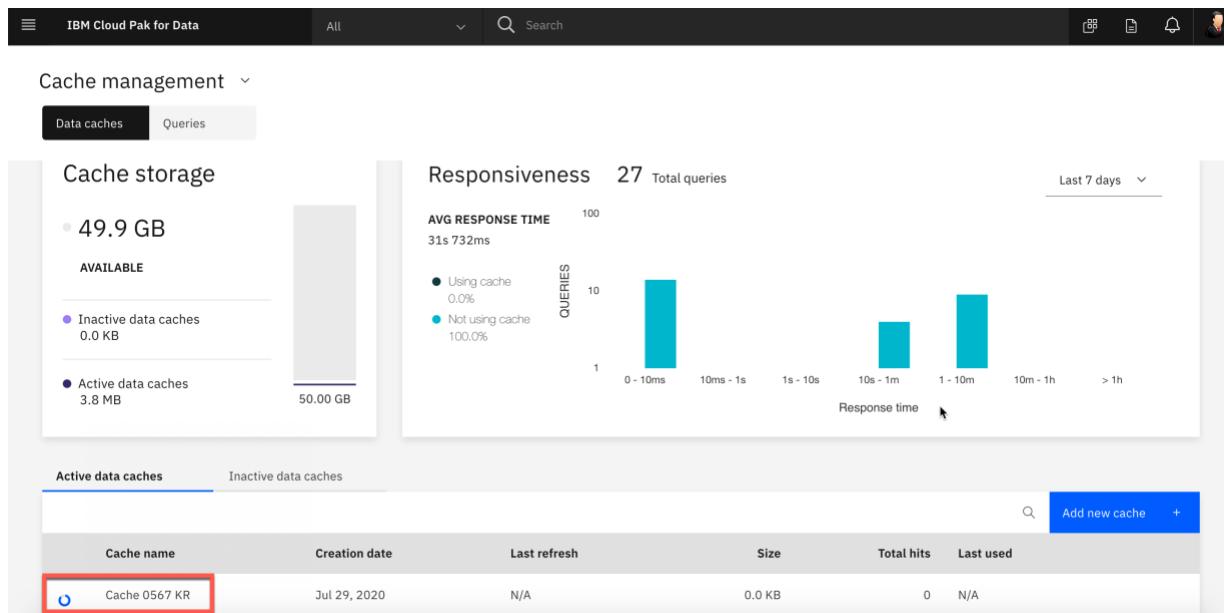


- \_\_50. Enter your **initials** at the end of the Cache name (yours may vary) then confirm the details on the final page.

Click **Create**.



The cache creation process may take some time and the main **Cache Management** page will reflect the work in progress.



- \_\_51. Once the cache creation is complete, the newly created cache shows up under the **Active data caches** along with other details and its size.

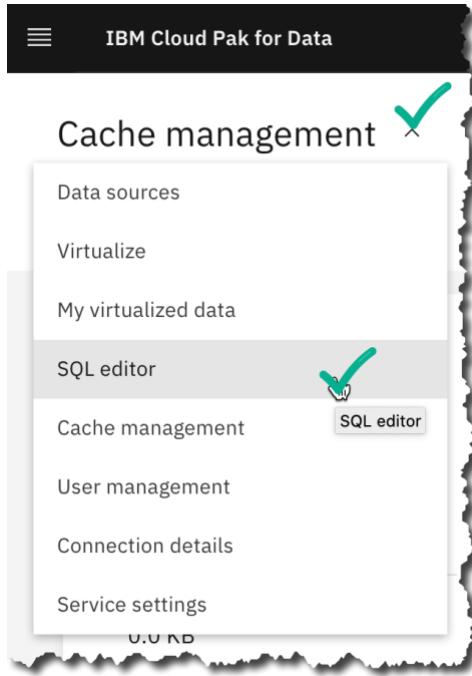
Cache name	Creation date	Last refresh	Size	Total hits	Last used
Cache 5114	Jul 08, 2020	Jul 08, 2020 9:24 PM	3.8 MB	0	N/A



Admin

**KNOWN ISSUE** – There exists a defect/issue wherein queries executed before the cache is created and activated fails to use the cache. This is because the query plan is already cached in a relational database such as Db2, so the work around is to clear the Db2 package cache and collect table statistics on the virtual tables created above. You will do that next.

\_52. Click **Menu (Cache management) ⇒ SQL Editor.**

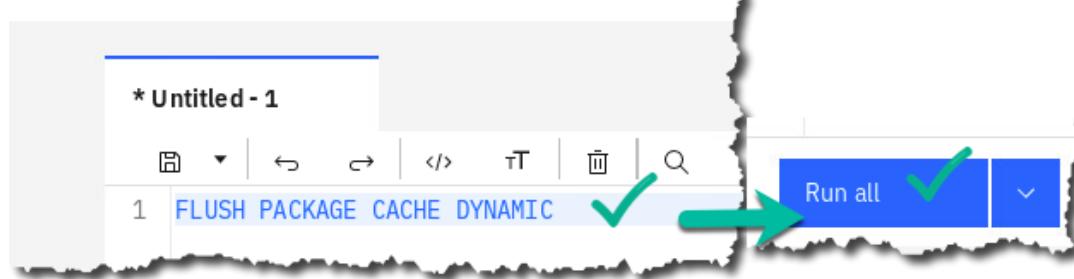


\_53. Remove the view SQL and replace it with the following:

**FLUSH PACKAGE CACHE DYNAMIC**

Click **Run all**.

**SQL editor** ▾



In a production setting, flushing the entire package may not always be a good idea as it could affect query execution for all users of the database. (We did it here to make this lab simpler.)

Instead, one could opt to selectively flush the package corresponding to particular queries. For more details on that see: <http://ibm.biz/FLUSH-PACKAGE>.

\_\_54. Next, check if the table statistics have been collected. Replace the previous SQL with:

```
SELECT TABNAME, STATS_TIME, CARD , TYPE  
FROM SYSCAT.TABLES  
WHERE TABSCHEMA='USER1001'  
AND  
TABNAME  
IN ('STOCK_SYMBOLS','CUSTOMER_TRANSACTIONS');
```

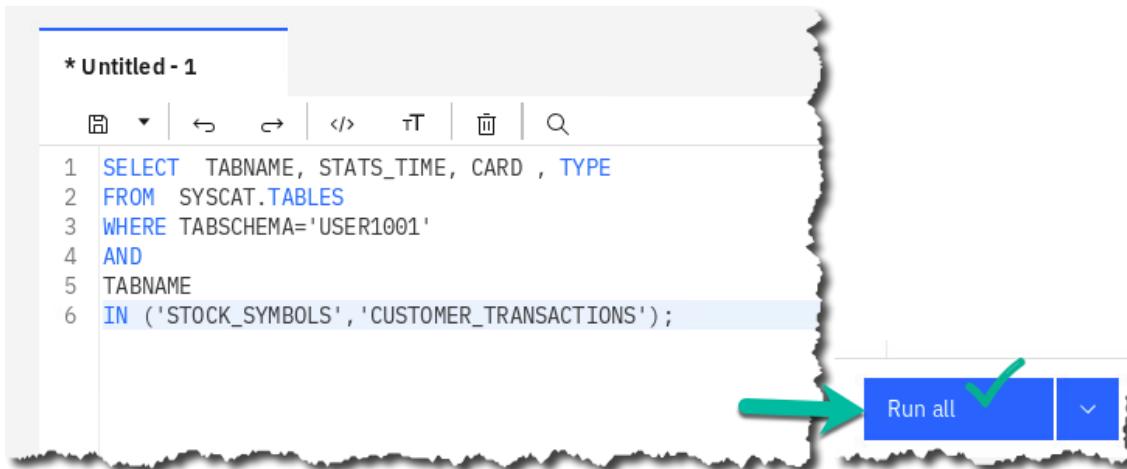
  
Data  
Engineer

Note: You can copy and paste this SQL directly from the Unified Desktop by using the File Browser to open this file:

 ⇒ Home ⇒ Downloads ⇒ DV\_SelectTabname.txt.

Alternately, you can download this file by opening a browser tab and using this link:  
<https://ibm.box.com/v/DV-SelectTabname>

\_\_55. Click [Run all](#).



If the cardinality for the virtual tables created before shows -1, this indicates that table statistics have not been collected. These statistics will be collected now.

Result - Jul 24, 2020 2:16:30 PM															
^ ② SELECT TABNAME, STATS_TIME, CARD , TYPE FROM SYSCAT.TABLES WHERE T...															
Run time: 0.093 s															
<b>Result set 1</b>															
<table border="1"> <thead> <tr> <th>TABNAME</th><th>STATS_TIME</th><th>CARD</th><th>TYPE</th></tr> </thead> <tbody> <tr> <td>CUSTOMER_TRANSACTIONS</td><td></td><td>-1</td><td>N</td></tr> <tr> <td>STOCK_SYMBOLS</td><td></td><td>-1</td><td>N</td></tr> </tbody> </table>				TABNAME	STATS_TIME	CARD	TYPE	CUSTOMER_TRANSACTIONS		-1	N	STOCK_SYMBOLS		-1	N
TABNAME	STATS_TIME	CARD	TYPE												
CUSTOMER_TRANSACTIONS		-1	N												
STOCK_SYMBOLS		-1	N												



Data  
Engineer

Note: The **SYSPROC.NNSTATS** is a procedure to collect statistics for remote tables:  
<http://ibm.biz/SYSPROC-NNSTATS>.

- 56. If your statistics are not current, then do this step. If they are current, you can skip this step.

```
CALL SYSPROC.NNSTAT(NULL, 'USER1001',
'CUSTOMER_TRANSACTIONS','","2,'/tmp/collstats1.log',?,1);
```

```
CALL SYSPROC.NNSTAT(NULL, 'USER1001',
'STOCK_SYMBOLS','","2,'/tmp/collstats1.log',?,1);
```



Data  
Engineer

Note: You can copy and paste this SQL directly from the Unified Desktop by using the File Browser to open this file:



⇒ Home ⇒ Downloads ⇒ DV\_CallSysProc.txt.

Alternately, you can download this file by opening a browser tab and using this link:  
<https://ibm.box.com/v/DV-CallSysProc> .

\_\_57. Click **Run all**.

```
* Untitled - 1
1 CALL SYSPROC.NNSTAT(NULL, 'USER1001',
2 'CUSTOMER_TRANSACTIONS', ",",2,'/tmp/collstats1.log',?,1);
3
4 CALL SYSPROC.NNSTAT(NULL, 'USER1001',
5 'STOCK_SYMBOLS', ",",2,'/tmp/collstats.log',?,1);
```

\_\_58. To confirm if the table statistics have now been collected, replace the previous SQL with the following:

```
SELECT TABNAME, STATS_TIME, CARD , TYPE
FROM SYSCAT.TABLES
WHERE TABSCHEMA='USER1001'
AND
TABNAME
IN ('STOCK_SYMBOLS','CUSTOMER_TRANSACTIONS');
```

\_\_59. Click **Run all**.

```
* Untitled - 1
1 SELECT TABNAME, STATS_TIME, CARD , TYPE
2 FROM SYSCAT.TABLES
3 WHERE TABSCHEMA='USER1001'
4 AND
5 TABNAME
6 IN ('STOCK_SYMBOLS', 'CUSTOMER_TRANSACTIONS');
```

TABNAME	STATS_TIME	CARD	TYPE
STOCK_SYMBOLS	2020-07-08 21:49:17.514...	31	N
CUSTOMER_TRANSACTIONS	2020-07-08 21:49:15.893...	2000000	

The actual rows counts for the virtual tables are shown as expected instead of -1.

With the package cache flushed and the table level statistics collected, all queries henceforth referencing the virtualized view should start using the cached copy instead of accessing the underlying data sources directly.

The DV Admin informs the BA about the cache creation and asks the BA to re-run the dashboard.





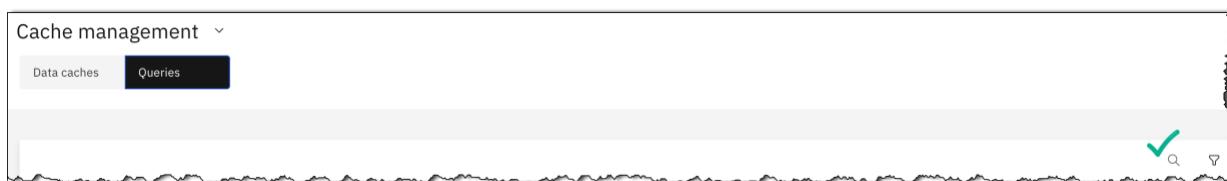
\_\_60. Return to the dashboard created previously.

Navigation Menu  $\Rightarrow$  Projects  $\Rightarrow$  your\_initials-DV-Project  $\Rightarrow$  Assets  $\Rightarrow$  Dashboards  $\Rightarrow$  Company Stock Analysis

- \_\_61. The dashboard should now render faster with much better response times while analyzing the stock data. The queries executed by the dashboard against the virtualized view now use the DV cache.
- \_\_62. Return to Data Virtualization, to Menu  $\Rightarrow$  Cache Management  $\Rightarrow$  Queries to get some insights into the queries you just ran.

*Note:* Because there is sometimes a slight delay in getting all the information about the latest queries and cache info via the UI, you can search for the query by Query ID (that you hopefully noted in the previous step.)

Click on the Search (magnifying glass) icon.



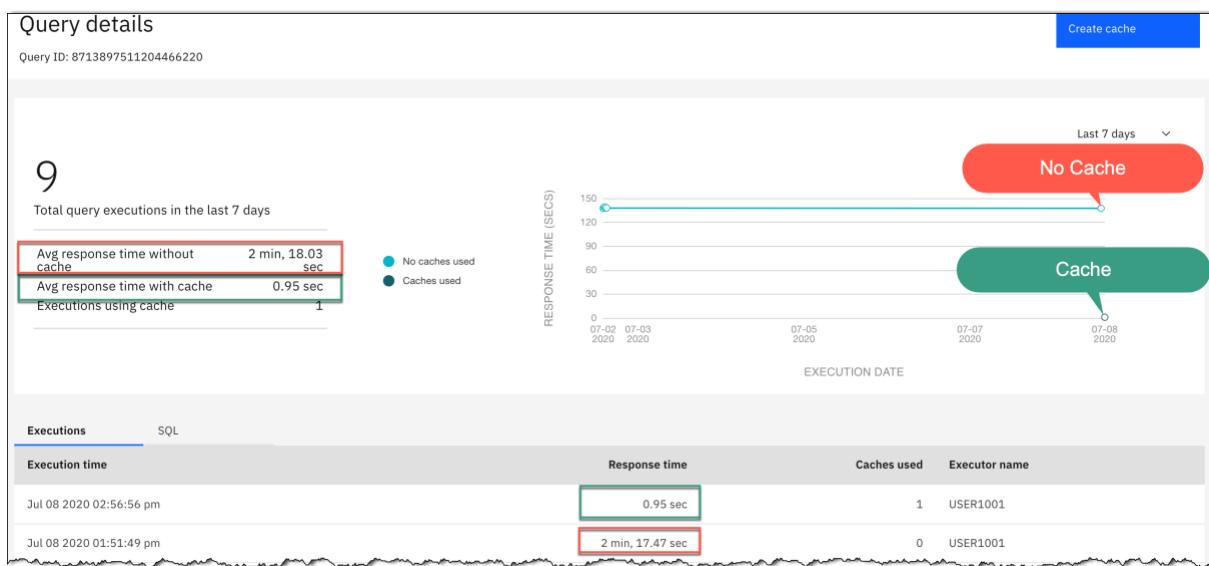
Enter the Query ID to list the query.

Query ID	Total executions	Caches used	Avg response time with cache	Avg response time without cache
8713897511204466220 USER1001	9	1	0.95 sec	2 min, 18.03 sec

- \_\_63. Expand the entry to view the query details.

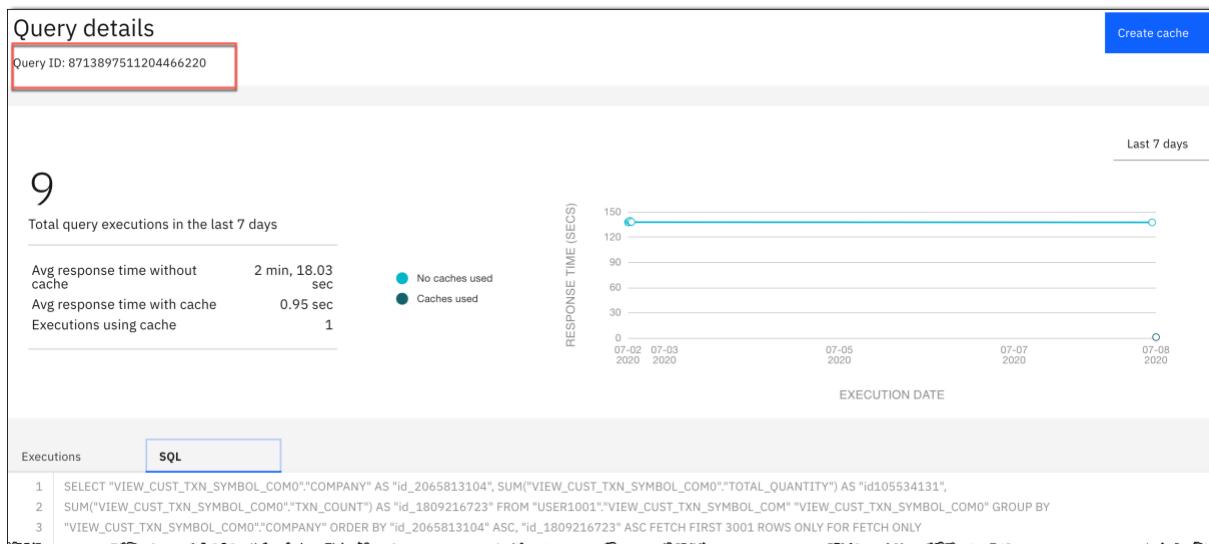
Query ID	Total executions	Caches used	Avg response time with cache	Avg response time without cache
8713897511204466220 USER1001	9	1	0.95 sec	2 min, 18.03 sec
<pre> 1 SELECT "VIEW_CUST_TXN_SYMBOL_COM0"."COMPANY" AS "id_2065813104",SUM("VIEW_CUST_TXN_SYMBOL_COM0"."TOTAL_QUANTITY") AS "id105534131", 2 SUM("VIEW_CUST_TXN_SYMBOL_COM0"."TXN_COUNT") AS "id_1809216723" FROM "USER1001"."VIEW_CUST_TXN_SYMBOL_COM" "VIEW_CUST_TXN_SYMBOL_COM0" 3 GROUP BY "VIEW_CUST_TXN_SYMBOL_COM0"."COMPANY" ORDER BY "id_2065813104" ASC, "id_1809216723" ASC FETCH FIRST 3001 ROWS ONLY FOR FETCH ONLY 4 5 6 7 8 9 10 </pre>				
<a href="#" style="color: blue;">View details</a>				

- \_\_64. Click on [View details](#) to show the number of executions and execution times.



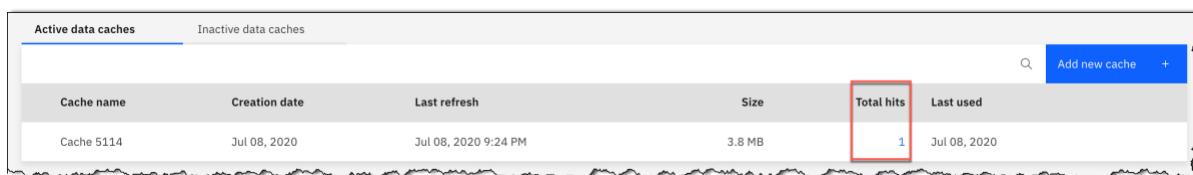
Notice in the above example (yours may vary somewhat) the query took under 1 second to run with the cache, while without cache it was 2 minutes and 18 seconds.

- \_\_65. Confirm that this is the same query that was executed in the previous steps, by confirming the Query ID and by clicking on [SQL](#).



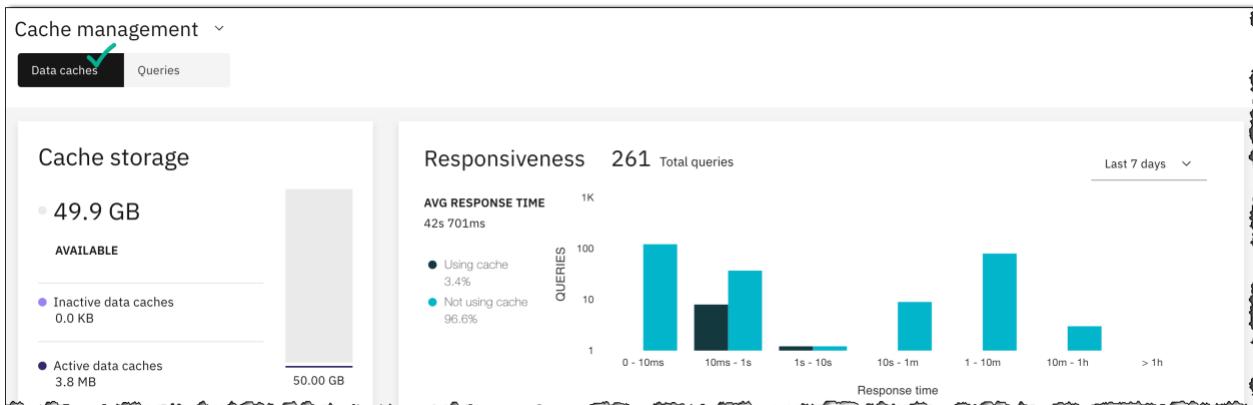
The query execution time has reduced **significantly** with the cache now being utilized instead of fetching the data from the underlying sources!

The cache hits should also start showing up under [Cache Management](#)  $\Rightarrow$  [Data caches](#).



## 11.8 Reviewing the Cache Management UI

The Cache Management UI helps the DV Admin to create, monitor and manage caches. It provides information about the active/inactive caches like size, hit counts etc. along with options to work with them.



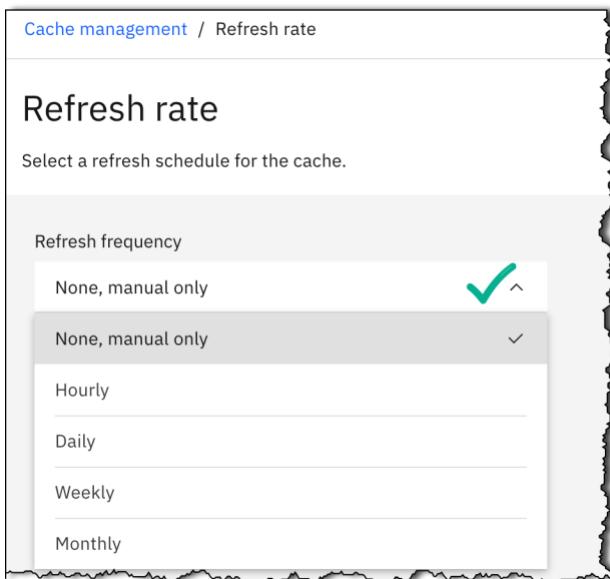
The UI also provides details about the queries executed against DV. This is particularly useful for the [DV Admin](#) in identifying slow running queries and potentially creating caches to make them run faster.

Play with some of the features of the Query section using [Filters](#).

The screenshot shows the 'Queries' section of the Cache management UI. It includes filters for 'Queries to include' (Include all selected), 'Data caches' (Select data caches checked), 'Response time' (Select range checked), 'Users' (Select users checked), and a time range ('Last 7 days' checked). A green checkmark is present above the filters and next to the 'Apply' button.

## 11.9 Refreshing the Data Virtualization cache

As data changes in the underlying data sources, it is important to refresh the cache and reflect the freshly updated data. The cache creation process provides an option to set the cache refresh rate.



One can choose from the following options in the drop down:

1. *None, manual only*: After the initial refresh, all subsequent refresh has to be manually performed by the DV Admin.
2. *Hourly*: The cache gets refreshed hourly on the set minute of the hour.
3. *Daily*: The cache gets refreshed at a specified time of day.
4. *Weekly*: The cache gets refreshed at a specified time on specific day(s) in a week.
5. *Monthly*: The cache gets refreshed at the specified time on specific day(s) and week(s) in a month.

To avoid delays and unexpected results, the best time to refresh the cache should be chosen based on the lowest DV usage/low query traffic.

## 11.10 Deactivating/Activating/Deleteing the Data Virtualization cache

The Cache Management UI provides some options to help manage caches. After a cache is created, it gets refreshed/populated by data and remains in the active state by default, to be considered by the Query Optimizer while generating the query plans.

The screenshot shows the 'Cache management' interface with the 'Data caches' tab selected. An active cache named 'Cache 5114' is listed in the table. A context menu is open over the cache row, showing options: 'View cache details', 'Edit name', 'Edit refresh rate', 'Refresh now', 'Deactivate' (which has a green checkmark), and 'Delete'.

Cache name	Creation date	Last refresh	Size	Total hits	Last used
Cache 5114	Jul 08, 2020	Jul 08, 2020 9:24 PM	3.8 MB	1	Jul 08, 2020

An active cache can be deactivated so the Query Optimizer will not consider it while generating the query plans. This could be useful if there are issues related to the cached data and needs investigation.

A cache can be deleted, which moves it into the inactive state and eventually cleaned up. The list of deleted caches can be listed using the [Deleted caches](#) option. It shows the original cache details and if needed, you could re-create them.

The screenshot shows the 'Cache management' interface with the 'Data caches' tab selected. The same active cache 'Cache 5114' is listed in the table. A context menu is open over the cache row, showing options: 'View cache details', 'Edit name', 'Edit refresh rate', 'Refresh now', 'Deactivate' (which has a green checkmark), and 'Delete'.

Cache name	Creation date	Last refresh	Size	Total hits	Last used
Cache 5114	Jul 08, 2020	Jul 08, 2020 9:24 PM	3.8 MB	1	Jul 08, 2020

The screenshot shows the 'Cache management' interface with the 'Inactive data caches' tab selected. The table lists the previously deleted cache 'Cache 5114'. A context menu is open over the cache row, showing options: 'View cache details', 'Edit name', 'Edit refresh rate', 'Refresh now', 'Deactivate' (which has a green checkmark), and 'Delete'.

Cache name	Creation date	Last refresh
Cache 5114	Jul 08, 2020	Jul 08, 2020 9:24 PM

**Deleted caches**

## 11.11 Caching guidelines

The DV cache is a very useful feature for improving DV query performance. However, there are some important points to remember:

1. The DV cache can only be created by the DV Admins.
2. A cached object has the same permissions as the underlying DV object.
3. Use queries that perform some sort of aggregation from the underlying data to create the cache. Performing aggregations help create summaries, and thus limit data cached.
4. When possible avoid duplicating data that exists in its original form in the underlying data sources.
5. Identify the most commonly used query or parts of queries and consider creating the cache from them. That way multiple users and queries can benefit from those caches.
6. Limit the size of the data to be cached. Note that the cache uses the storage allocated to the DV instance, hence it is common to all DV users. Large caches will limit the availability of space for other users/scenarios.
7. Time taken to refresh a cache will depend on the amount of data, latency and availability of the underlying data sources.
8. A cache will not be used for a query if there is no match for the statement or column(s) and the virtual object being queried.
9. The DV Cache Management UI takes couple of minutes to refresh the queries executed, cache hit counts and other relevant information.

## 11.12 Lab conclusion

Data Virtualization (DV), as part of the Collect phase, facilitates accessing data from various data sources and running queries across them. Since data movement is limited, all of the access rules and policies for creating copies remain preserved. In situations where query response times are paramount, DV provides the caching facility, which was covered in this lab.

### **\*\* End of Lab 11 - Collect: Data Virtualization Caching – Deeper Dive**

Lab by Rajesh Kartha, Edited by Burt Vialpando and Kent Rubin

---

## Lab 12 COLLECT: VIRTUALIZING & CACHING FROM Z/OS – DEEPER DIVE

### 12.1 Lab overview

Data Virtualization (DV) is an important component of IBM Cloud Pak for Data (CPD) to help integrate data sources across multiple types and locations and turn them into one logical data view. As shown in the previous DV Caching Deep Dive lab, virtualizing tables from databases hosted in the cloud and then leveraging caching helps to drastically improve query performance.

In addition to the cloud, almost all enterprises will have large amounts of data stored in on-premises data sources. Data hosted on z/OS systems can be a major portion of that.

While there exist multiple ways to access data residing in a z/OS system, using IBM Data Virtualization Manager (DVM) is one of the most popular ones. Cloud Pak for Data Virtualization (CPD-DV) has the ability to virtualize and ingest any mainframe data that is available, leveraging the Data Virtualization Manager for z/OS.

### 12.2 IBM DVM

IBM Data Virtualization Manager (DVM) for z/OS® provides virtual, integrated views of data residing on IBM Z®. It enables users and applications read/write access to IBM Z data in place, without having to move, replicate or transform the data.

IBM DVM enables data structures that were designed independently to be used together. Traditional data movement approaches can negatively impact the opportunity to benefit from data where and when it is needed. By unlocking IBM Z data using popular, industry-standard APIs, DVM for z/OS can save you time and money.

DVM for z/OS provides access to IBM Z data sources such as VSAM (Virtual Sequential Access Method), IBM Db2 for z/OS, Adaptable Database System (ADABAS), Integrated Database Management System (IDMS), IBM Information Management System (IMS), IBM System Management Facilities (SMF) and non-IBM Z data sources, all without the need for mainframe skills.

You can also simplify the development of AI applications directly from IBM Z data with CPD.

For more information on DVM for z/OS, check out this link: <http://ibm.biz/DV-Manager-ZOS>.

Also check out the IBM Demos page: <http://ibm.biz/IBM-Demos>

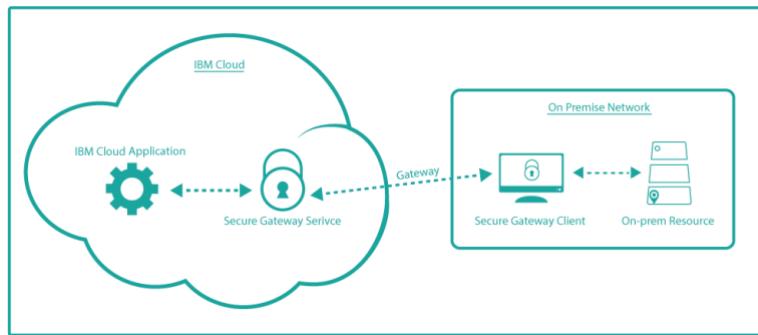
### 12.2.1 Secure Gateway

Accessing on-prem data from an external system can be challenging and there are different ways to achieve that. A simple and cost-effective way in the IBM Cloud is to use the Secure Gateway service.

IBM Secure Gateway for IBM Cloud maintains a single persistent encrypted connection between the Secure Gateway client in the on-prem network and the Secure Gateway server in the IBM Cloud. Data can be securely transmitted bidirectionally between on-prem and external resources.



The IBM Secure Gateway (SG) consists of a Gateway service, an on-prem SG Client and an on-prem Destination.



External applications can securely connect to the Gateway service, which in turn connects to the on-prem destination via the configured SG client. The external applications are unaware of the on-prem destination details. They only know the IBM Cloud Secure Gateway service endpoint and the corresponding credentials to be used. More information about IBM Cloud Secure Gateway service is available in its documentation: <http://ibm.biz/Getting-Started> and <http://ibm.biz/Client-Install>.

In this lab, you will create a DV virtualized table that spans three different data sources: one z/OS on-prem data source via IBM DVM, and two Db2 Warehouse tables hosted in the IBM Cloud. This allows querying across all of them via a single DV query and then uses the caching capability of DV to improve query times.

In our scenario, the Trade Co. Business Analyst (BA) builds on the dashboard created in the previous DV Caching Deeper Dive lab using data from a DV view, which in turn points to two remote IBM Cloud Db2 Warehouse data sources underneath. However, there is now a need for more historical analysis with the addition of a third z/OS data source that hosts transaction data from previous years. The BA works with the Data Engineer and DV Admin, who can create the necessary virtual objects (tables and views) and a data cache for the final virtualized view to help address the data retrieval performance issues.

## 12.2.2 Setup (informational only)

Information in this section is only for information purposes. All the setup steps have already been performed on your workshop cluster to facilitate this lab completion.

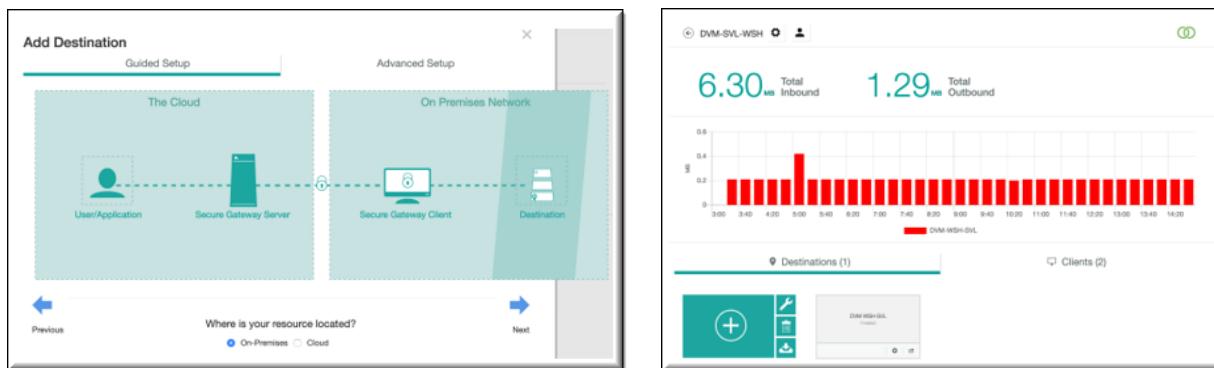
### z/OS Data:

In this lab example, a transactional process on z/OS has stored records of stock market trades into a VSAM data set. VSAM data sets arrange data in fixed format records by an index key, by a relative byte address, or by a relative record number. On the mainframe, VSAM data sets are cataloged and used frequently for easy retrieval by z/OS application and processes. DVM for z/OS has the ability to create a customized virtual table or virtual view definition over this VSAM data set.

CPD-DV, using a DVM for z/OS connection, can query the virtual table defined over the VSAM data using ANSI-standard SQL, and virtualize this stock trade data with other DV data sources.

### Access:

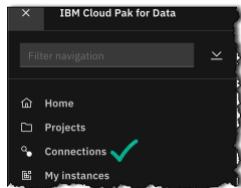
The VSAM data and the DVM for z/OS application reside on a mainframe server (z/OS) at the Washington Systems Center on IBM premises. This system is in a secured network zone and access from the CPD system is provided through a Secure Gateway client, running inside the IBM network configured to an IBM Cloud Secure Gateway service.



***Connection:***

IBM DVM for z/OS comes with its own set of JDBC drivers and CPD-DV uses these to connect and virtualize tables from the z/OS system. For this lab, these JDBC drivers have already been loaded into the Cloud Pak for Data instance and are ready to create a connection. We will show you how we did it here. ***NOTE: You do not need to complete these steps for this workshop.***

To upload the DVM JDBC driver, from the navigation menu go to [Connections](#).



Click on [Upload driver](#).



Provide a name and upload the JDBC jars along with the names of the Driver class name and the JDBC URL prefix. Click on [Upload](#).

Connections / Upload driver

### Upload driver

Enable users to connect to other data sources by uploading the JDBC drivers for the data source.

DVM z/OS	<input checked="" type="checkbox"/>
<b>JDBC driver JAR files</b>	
Upload all the required JAR files for your JDBC driver.	
Drop your JAR file here or browse for a file to upload.	
dv-jdbc-3.1.201910231912.jar	<input checked="" type="checkbox"/>
log4j-api-2.8.2.jar	<input checked="" type="checkbox"/>
log4j-core-2.8.2.jar	<input checked="" type="checkbox"/>
Driver class name ⓘ	
com.rs.jdbc.DvDriver	<input checked="" type="checkbox"/>
JDBC URL prefix ⓘ	
jdbc:rs:dv	<input checked="" type="checkbox"/>

[Cancel](#) [Upload](#)

Existing drivers

Find drivers

**RESULT**

Connection type	Date added	JAR files
DVM for z/OS	Jun 23, 2020 11:49 AM	dv-jdbc-3.1.201910231912.jar, log4j-api-2.8.2.j...

## 12.3 Personas represented in this lab

The [Business Analyst](#) persona along with the [Data Engineer](#) and [Data Virtualization Administrator](#) personas will perform the exercises in this lab.

Persona (Role)	Capabilities
 Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.
 Administrator (DV/CPD)	Administrators set up and maintain the DV module within the CPD environment itself. They are responsible for granting DV access to users and administration tasks like creating a data cache.

For convenience in doing this lab, instead of switching between personas, all the required privileges have been provided to the same user. The workbook will refer to the respective personas at different stages to help understand the flow of this task.

## 12.4 Logging into the CPD web client (if you have not already done so)

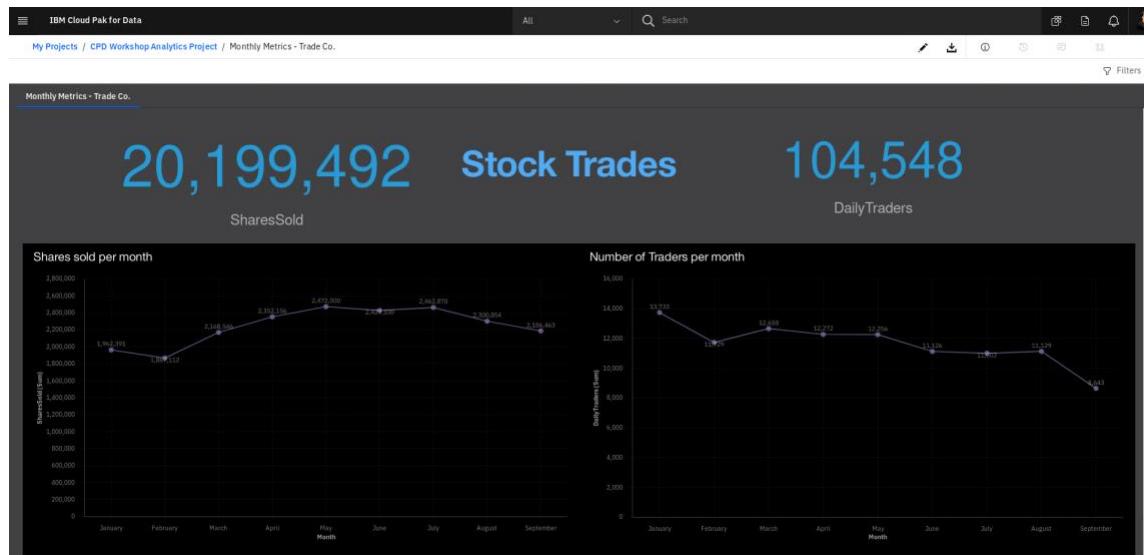
- \_\_1. If you are starting this lab stand-alone (without going through previous labs) do the following:
- \_\_2. Double-click the desktop icon: [Cloud Pak for Data Web Client](#).



- \_\_3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click [Sign in](#).

## 12.5 Reviewing the dashboard: Stock Trading Analysis - Trade Co.

In the attempt to understand stock trading patterns, the business analyst for Trade Co. starts by creating a simple dashboard to find the most popularly traded stocks historically. The dashboard shows the number of Shares Sold vs. the number of Daily Trades.



This lab follows the same narrative as that of the DV Caching Deeper Dive lab. The Business Analyst (BA) works with the Data Engineer (DE) to get virtualized access to the different data sources required for creating the dashboard(s). The DE creates the Data Sources, Virtual Tables and finally a Virtualized View joining all the Virtual Tables and computing the basic aggregations required. The View is then shared with the BA, who can then proceed with creating the dashboard.

However, once the dashboard is initially created, the BA notices delays in rendering the visualizations. Since every request from the dashboard has to fetch the data from its original source(s), latency starts to play an important role, slowing down response. The BA works with the DV Admin to create a cache for the View, which helps speed up query times significantly and hence rendering of the dashboard(s).

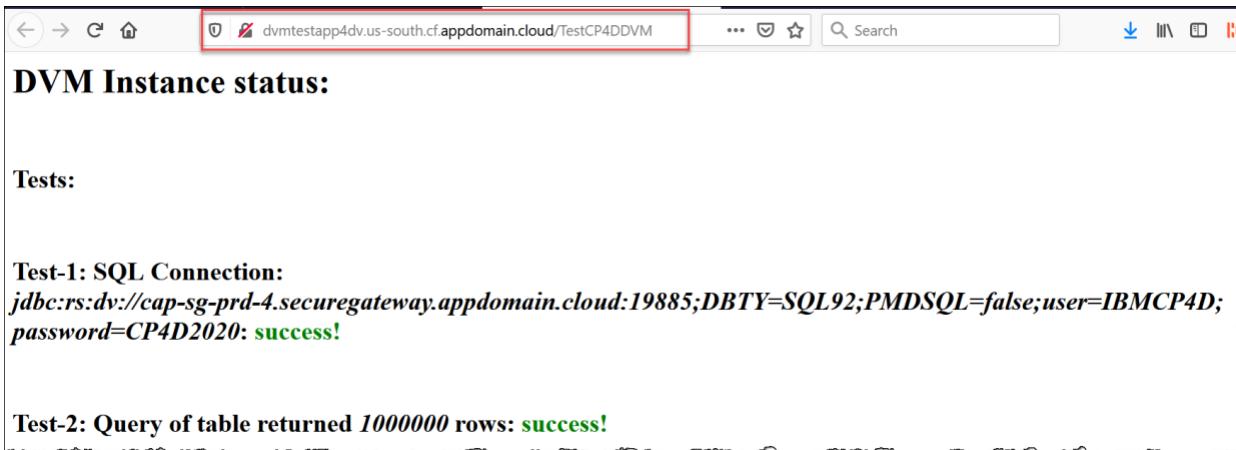
## 12.6 Adding the DV connection to DVM for z/OS

The Data Virtualization process begins by adding data sources to virtualize and is typically done by the Data Engineer.

Persona (Role)
 Data Engineer (DV/CPD)

- 4. Before proceeding with the creation of the data connection, ensure that the data source is accessible.

Click on this embedded [link](#) to confirm, or type: <https://ibm.biz/cpd-workshop-dvm>.



**DVM Instance status:**

Tests:

**Test-1: SQL Connection:**  
`jdbc:rs:dv://cap-sg-prd-4.securegateway.appdomain.cloud:19885;DBTY=SQL92;PMDSQL=false;user=IBMCMP4D;password=CP4D2020: success!`

**Test-2: Query of table returned 1000000 rows: success!**

The above link (<http://dvmtestapp4dv.us-south.cf.appdomain.cloud/TestCP4DDVM>) is a verification test running in the IBM Cloud to confirm if the z/OS and DVM systems are accessible and hence able to be queried.



If the above link times out or throws an error like this:

This page contains the following errors:

error on line 1 at column 1: Document is empty

Below is a rendering of the page up to the first error.

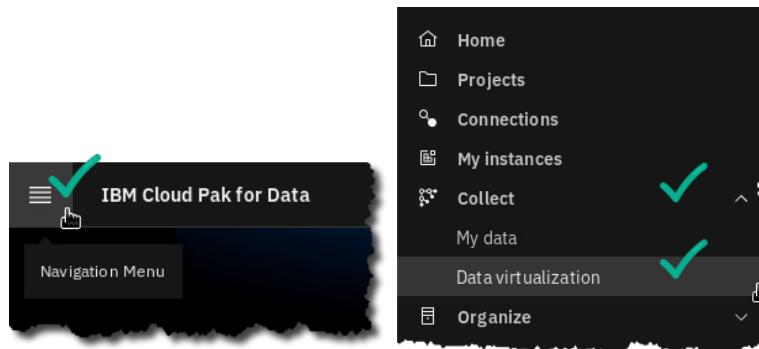
that means that the z/OS and DVM systems are not accessible for some reason.

Please contact Rajesh Kartha. Email [kartha@us.ibm.com](mailto:kartha@us.ibm.com) Slack: @Rajesh Kartha.

### 12.6.1 Navigating to Data virtualization

—5. Start at the [Navigation Menu](#).

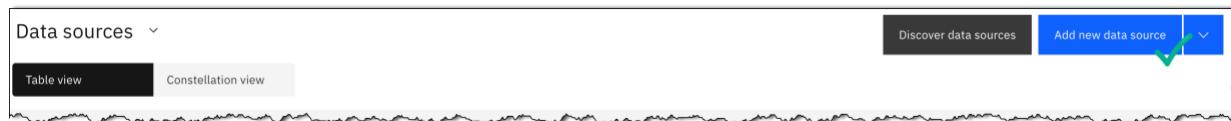
Click [Collect](#) ⇒ [Data virtualization](#).



### 12.6.2 Creating a new data connection to DVM for z/OS

—6. The JDBC drivers for connecting to DVM for z/OS have already been loaded and the system is ready to create a connection.

Click [Add new data source](#).



—7. Enter the details from the table below and select DVM for z/OS from the drop down.

Note: You can get the Connection information by downloading this file: <http://ibm.biz/DV-Z-OS>.

(If you are using the Unified Desktop, simply open a browser and type in the URL.)

Connection Name	POT-DVM
Description	DVM Connection
Connection type	DVM for z/OS
JDBC URL	jdbc:rs:dv://cap-sg-prd-4.securegateway.appdomain.cloud:19885;DBTY=SQL92;PMDSQL=false
Username	IBMCP4D
Password	CP4D2020

- \_\_8. Click on **Test connection** to test it.

The screenshot shows the 'New connection' dialog box. The 'Connection name' field contains 'POT-DVM' with a green checkmark. The 'Description (optional)' field has a placeholder 'Enter a description for the connection'. The 'Connection type' dropdown is set to 'DVM for z/OS' with a green checkmark. The 'JDBC URL' field contains 'jdbc:rs:dv://cap-sg-prd-4.securegateway.appdomain.cloud:19885;DBTY=SQL' with a green checkmark. The 'Username' field contains 'IBMCPP4D' with a green checkmark, and the 'Password' field contains a masked password with a green checkmark. At the bottom, the 'Test connection' button is highlighted with a blue border and a green checkmark, while the 'Cancel' and 'Create' buttons are dark grey.

- \_\_9. Once the **Test connection** succeeds, click on **Create** to create the connection.

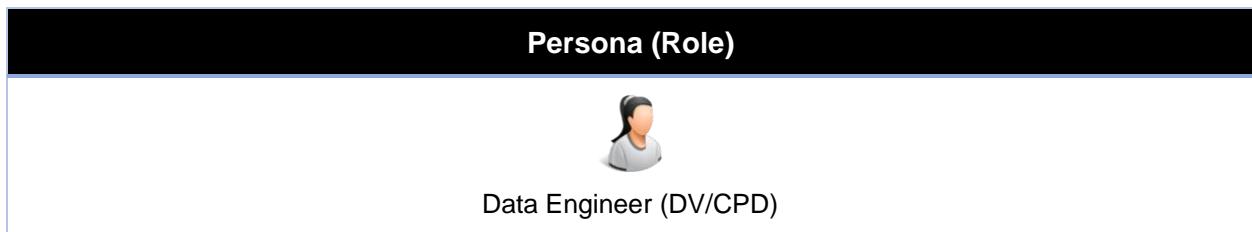
The screenshot shows the 'New connection' dialog box after testing. A green success message at the top states 'Success The test connection was successful. Click Create to save the connection information.' The rest of the form fields are identical to the previous screenshot, with the 'Test connection' button now highlighted with a blue border and a green checkmark, and the 'Create' button visible below it.

The connection is successfully created.

The screenshot shows the 'Add connection' dialog box. It lists a single connection named 'POT-DVM' under the 'Name' column. The 'Type' column shows 'DVM for z/OS', and the 'URL' column shows 'jdbc:rs:dv://cap-sg-prd-4.securegateway.appdomain.cloud:19885;DBTY=SQL92;PMDSO...'. The 'Created By' column shows 'CPD User'. The entire row for 'POT-DVM' is highlighted with a red box.

## 12.7 Remove existing virtual tables and views (if they exist)

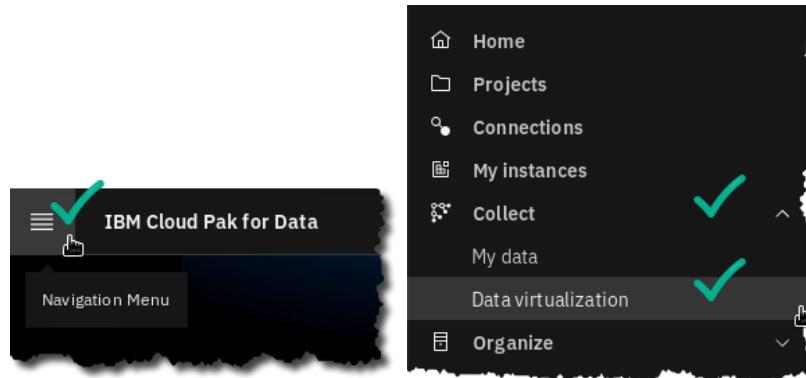
The Data Virtualization process begins by adding data sources to virtualize and is typically done by the Data Engineer.



### 12.7.1 Navigate to Data virtualization

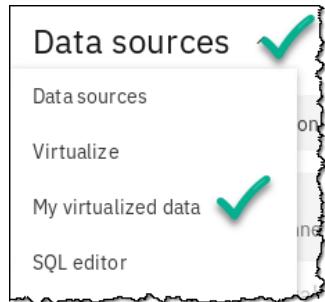
\_\_10. Start at the [Navigation Menu](#).

Click [Collect](#) ⇒ [Data virtualization](#).



### 12.7.2 Delete virtual views and tables from the previous DV Caching lab

\_\_11. From the drop-down menu (Data Sources) select [My virtualized data](#).



\_\_12. Find the virtualized table [VIEW\\_CUST\\_TXN\\_SYMBOL\\_COM](#).

(Note: If you did not do the Data Virtualization Deeper Dive lab before this one, then these assets will not exist and you can skip this and go to the next section.)

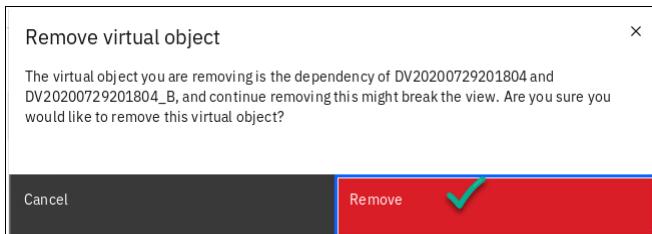
Click the [ellipses](#) on that table (at the end of the line) then [Remove](#).

Table	Schema	Created on
<input type="checkbox"/> VIEW_CUST_TXN_SYMBOL_COM	USER1001	Jul 29, 2020 3:08:54 PM
<input type="checkbox"/> STOCK_SYMBOLS	USER1001	Jul 29, 2020 7:09:30 AM
<input type="checkbox"/> CUSTOMER_TRANSACTIONS	USER1001	Jul 29, 2020 7:09:30 AM

Items per page: 10 | 1-3 of 3 items

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

\_\_13. Click [Remove](#) to confirm.



\_\_14. In the same way, remove virtual tables: [STOCK\\_SYMBOLS](#) and [CUSTOMER\\_TRANSACTIONS](#).

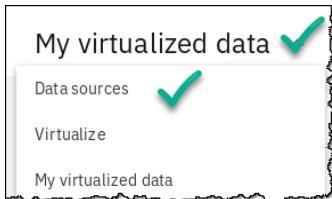
<input checked="" type="checkbox"/> STOCK_SYMBOLS	<b>Remove these as well</b>	USER1001
<input checked="" type="checkbox"/> CUSTOMER_TRANSACTIONS		USER1001

## 12.8 Virtualizing the remote tables and creating a view

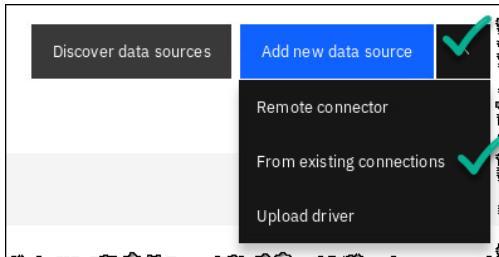
### 12.8.1 Add the DVM data source

- \_\_15. With the connection added in the section Adding the DV connection to DVM for z/OS, we will now add that as a data source to our Data Virtualization environment.

From the drop-down **Menu (My virtualized data)** select **Data sources**.



- \_\_16. On the top far right, select **Add new data source**  $\Rightarrow$  **From existing connections**



- \_\_17. Select the button for **POT-DVM**  $\Rightarrow$  **Next**.



 Data Engineer	<p>If you receive an error message that says the data source is already created, that is OK. Just move on to the next step. The data source should be available to review. Click the refresh button on your browser to make sure you can see it.</p>
--	--

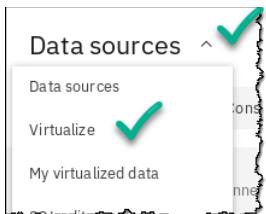
- \_\_18. You should now see this data source in your DV Data Sources screen.

Hostname: Port	Database	Type	Username
db2w-dpiyzso.us-south.db2w.cloud.ibm.com: 50000	BLUDB	Db2 Family	dvuser
cap-sg-prd-4.securegateway.appdomain.cloud: 19885	SQL92	Data Virtualization for z/OS	IBMCLOUD

### 12.8.2 Create virtualized tables

- 19. With the data sources successfully created, the next step is to virtualize the tables needed for this exercise.

From the drop-down menu (My Data sources) select Virtualize.



- 20. In the search bar, enter the string stock.

- 21. Notice there are three tables available called **CUSTOMER\_TRANSACTIONS**.

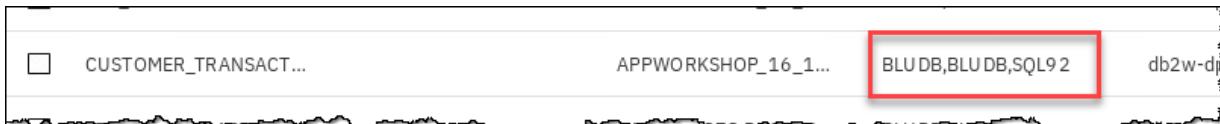
The first two tables are from two Db2 Warehouse on Cloud databases called BLUDB, and the third table is from our z/OS database called SQL92 that is available through the DVM connection you created earlier.

<input type="checkbox"/>	CUSTOMER_TRANSACTIONS	-	APPWORKSHOP_16_17	BLUDB
<input type="checkbox"/>	CUSTOMER_TRANSACTIONS	-	APPWORKSHOP_18_19	BLUDB
<input type="checkbox"/>	CUSTOMER_TRANSACTIONS	-	DVSQ	SQL92

- 22. Click the gear icon, then select **Group tables with identical names**.

- \_23. This “Table Grouping” (AKA “Schema Folding”) leverages a powerful DV feature that groups tables with the same names across data sources and presents them as a single entity. For you RDBMS people, think of it as a “union all” join across databases of different types (Db2, SQL Server, Oracle, Db2 z/OS, etc.) located anywhere in the world, all with one click of a button!

Example: Notice how three tables called **CUSTOMER\_TRANSACTIONS** from three different databases (two called BLUDB, one called SQL92) are treated as one table.



- \_24. Select tables **CUSTOMER\_TRANSACTIONS** and **STOCK\_SYMBOLS**.

Click [Add to cart](#).

Table	Business terms	Schemas	Databases/File Path	Hostname: Port	Columns	Matched columns	Grouped tables
<input type="checkbox"/> OLD_CUSTOMER_T...	APPWORKSHOP_16...	BLUDB,BLUDB	db2w-dpiyzso.us-s...	7	STOCK_PRICE	2	
<input checked="" type="checkbox"/> CUSTOMER_TRANS...	APPWORKSHOP_16...	BLUDB,BLUDB,SQL92	db2w-dpiyzso.us-s...	8	STOCK_PRICE	3	
<input type="checkbox"/> STOCK_TRANSACTI...	DASH100878,DASH...	BLUDB,BLUDB	db2w-dpiyzso.us-s...	6		2	
<input type="checkbox"/> TEST_CUST_TXN	TESTDV	BLUDB	db2w-dpiyzso.us-s...	7	STOCK_PRICE	1	
<input type="checkbox"/> TOBEDELETED	TESTDV	BLUDB	db2w-dpiyzso.us-s...	7	STOCK_PRICE	1	
<input type="checkbox"/> TESTTYPE	DVUSER,STOCKS	BLUDB,STOCKS	db2w-naxmdbf.us-...	2		2	
<input type="checkbox"/> TEST_TXN	TESTDATA	BLUDB	db2w-naxmdbf.us-...	7	STOCK_PRICE	1	
<input checked="" type="checkbox"/> STOCK_SYMBOLS	STOCKS	STOCKS	sl-us-south-1-porta...	2		1	

- \_25. Click [View cart](#).



- \_\_26. In the section [Review cart and Virtualize tables](#), notice the **Grouped tables** column which shows the number 3 to indicate the three **CUSTOMER\_TRANSACTIONS** tables residing in three different databases to be schema folded as a single virtual table.

Select the [My virtualized data](#) button.

Uncheck the [Submit to catalog](#) box.

Click [Virtualize](#).

Virtualize / Review cart and virtualize tables

Review cart and virtualize tables

Assign to

Data request  Project  My virtualized data

Submit to catalog

Empty cart

Virtualize

Table	Schema	Source schema	Databases/File Path	Hostname: Port	Grouped tables
CUSTOMER_TRAN...	USER1001	APPWORKSHOP_16_1...	BLUDB, BLUDB, SQL92	db2w-dpiyzo.us-south...	3 ✓
STOCK_SYMBOLS	USER1001	STOCKS	STOCKS	sl-us-south-1-portal.5...	1

- \_\_27. Click [View my virtualized data](#) to view the tables created.

Virtual tables created

2 of 2 tables successfully virtualized.

Table	Schema	Status
STOCK_SYMBOLS	USER1001	✓ success
CUSTOMER_TRANSACTIONS	USER1001	✓ success

Assigned to none

Virtualize more data

View my virtualized data

My virtualized data

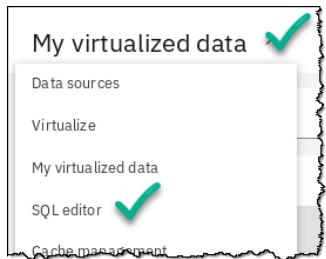
Find virtual objects

Filter by: All types

Table	Schema	Created on
CUSTOMER_TRANSACTIONS	USER1001	Jul 13, 2020 12:05:41 PM
STOCK_SYMBOLS	USER1001	Jul 13, 2020 12:05:40 PM

28. You will now be creating a virtualized view of these virtualized tables using the SQL editor.

Select menu (My virtualized data) ⇒ SQL editor.



29. Copy and paste the SQL below into the SQL editor.

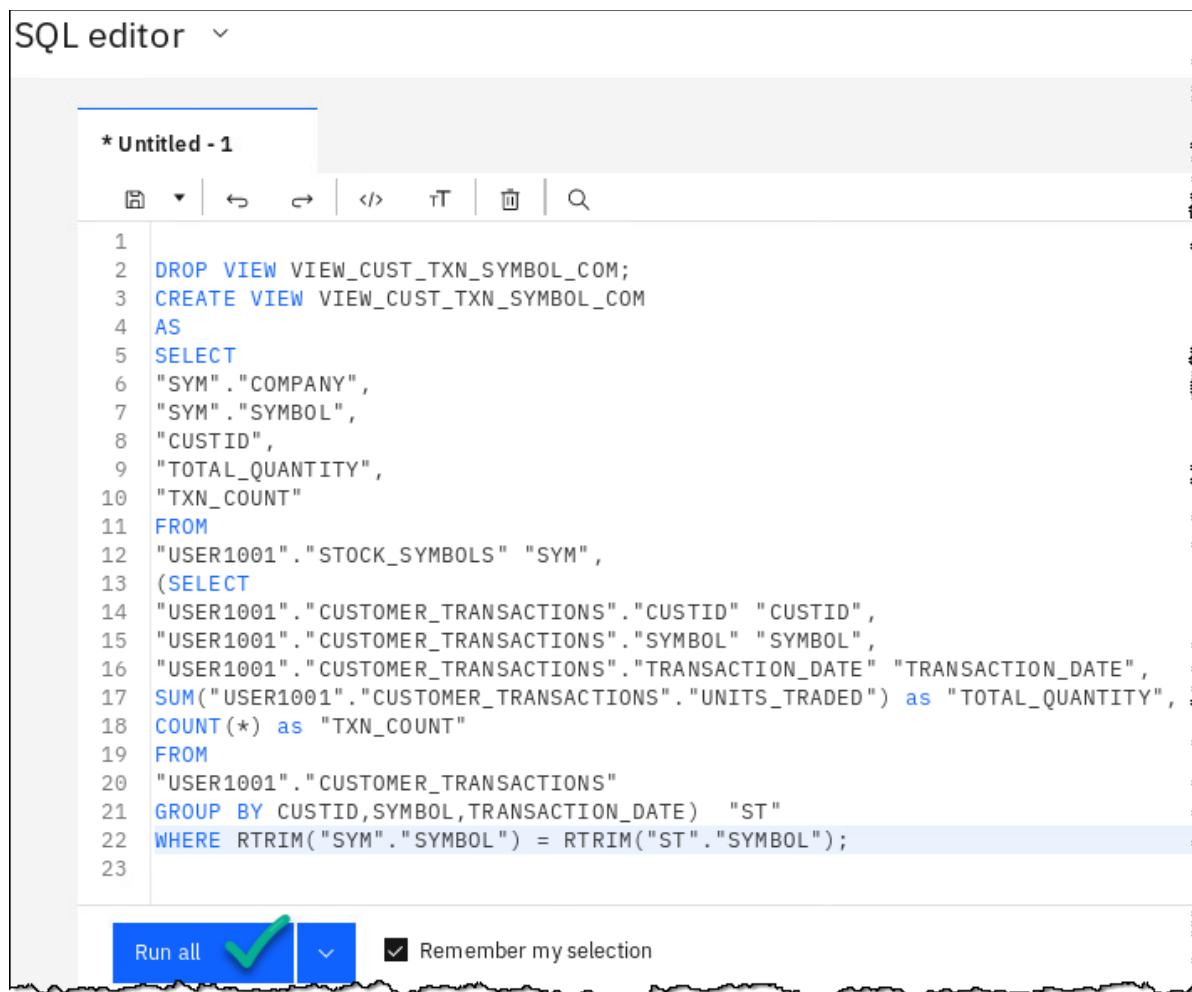
Note: You can copy and paste this SQL directly from the Unified Desktop by using the File Browser to open this file:

  ⇒ Home ⇒ Downloads ⇒ DVCache\_SQL.txt .

Alternately, you can download this file by opening a browser tab and using this link:  
<http://ibm.biz/DV-Caching-SQL>.

```
DROP VIEW VIEW_CUST_TXN_SYMBOL_COM;
CREATE VIEW VIEW_CUST_TXN_SYMBOL_COM
AS
SELECT
"SYM"."COMPANY",
"SYM"."SYMBOL",
"CUSTID",
"TOTAL_QUANTITY",
"TXN_COUNT"
FROM
"USER1001"."STOCK_SYMBOLS" "SYM",
(SELECT
"USER1001"."CUSTOMER_TRANSACTIONS"."CUSTID" "CUSTID",
"USER1001"."CUSTOMER_TRANSACTIONS"."SYMBOL" "SYMBOL",
"USER1001"."CUSTOMER_TRANSACTIONS"."TRANSACTION_DATE"
"TRANSACTION_DATE",
SUM("USER1001"."CUSTOMER_TRANSACTIONS"."UNITS_TRADED") as
"TOTAL_QUANTITY",
COUNT(*) as "TXN_COUNT"
FROM
"USER1001"."CUSTOMER_TRANSACTIONS"
GROUP BY CUSTID,SYMBOL,TRANSACTION_DATE) "ST"
WHERE RTRIM("SYM"."SYMBOL")= RTRIM("ST"."SYMBOL");
```

- \_\_30. Click **Run all** which should create the virtualized view successfully.



The screenshot shows an SQL editor window titled "SQL editor". The title bar has a dropdown arrow. Below the title bar is a toolbar with icons for file operations, search, and other functions. The main area is labeled "\* Untitled - 1". The code listed is:

```

1  DROP VIEW VIEW_CUST_TXN_SYMBOL_COM;
2  CREATE VIEW VIEW_CUST_TXN_SYMBOL_COM
3  AS
4  SELECT
5    "SYM"."COMPANY",
6    "SYM"."SYMBOL",
7    "CUSTID",
8    "TOTAL_QUANTITY",
9    "TXN_COUNT"
10   FROM
11   "USER1001"."STOCK_SYMBOLS" "SYM",
12   (SELECT
13     "USER1001"."CUSTOMER_TRANSACTIONS"."CUSTID" "CUSTID",
14     "USER1001"."CUSTOMER_TRANSACTIONS"."SYMBOL" "SYMBOL",
15     "USER1001"."CUSTOMER_TRANSACTIONS"."TRANSACTION_DATE" "TRANSACTION_DATE",
16     SUM("USER1001"."CUSTOMER_TRANSACTIONS"."UNITS_TRADED") as "TOTAL_QUANTITY",
17     COUNT(*) as "TXN_COUNT"
18   FROM
19   "USER1001"."CUSTOMER_TRANSACTIONS"
20   GROUP BY CUSTID,SYMBOL,TRANSACTION_DATE) "ST"
21   WHERE RTRIM("SYM"."SYMBOL") = RTRIM("ST"."SYMBOL");
22
23

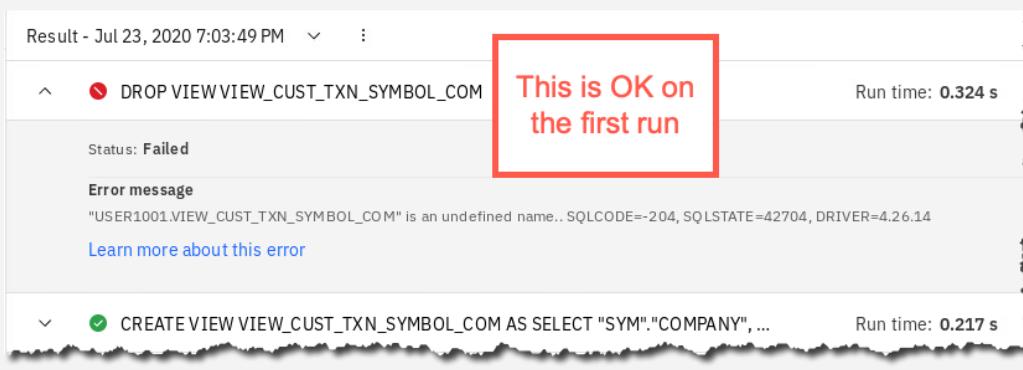
```

At the bottom of the editor, there is a toolbar with a "Run all" button containing a green checkmark, and a checkbox labeled "Remember my selection".

Note: The first time this script is run, the **DROP VIEW** statement will fail since the view does not exist. It will execute cleanly on subsequent runs.



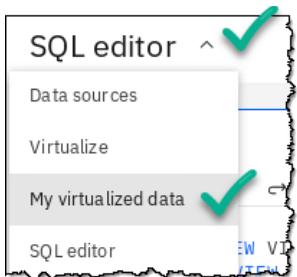
Data  
Engineer



This is OK on the first run

Result - Jul 23, 2020 7:03:49 PM		
<span style="color: red;">✖</span> DROP VIEWVIEW_CUST_TXN_SYMBOL_COM	Status: Failed	Run time: 0.324 s
Error message		"USER1001.VIEW_CUST_TXN_SYMBOL_COM" is an undefined name.. SQLCODE=-204, SQLSTATE=42704, DRIVER=4.26.14
<a href="#">Learn more about this error</a>		
<span style="color: green;">✔</span> CREATE VIEW VIEW_CUST_TXN_SYMBOL_COM AS SELECT "SYM"."COMPANY", ...		Run time: 0.217 s

- \_\_31. Select menu (SQL editor)  $\Rightarrow$  My virtualized data.



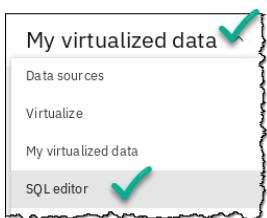
- \_\_32. The virtualized view should now be here: [VIEW\\_CUST\\_TXN\\_SYMBOL\\_COM](#).

My virtualized data		
	Schema	Created on
<input type="checkbox"/> TABLE		
<input type="checkbox"/> VIEW_CUST_TXN_SYMBOL_COM	USER1001	Jul 13, 2020 12:53:36 PM
<input type="checkbox"/> CUSTOMER_TRANSACTIONS	USER1001	Jul 13, 2020 12:53:14 PM
<input type="checkbox"/> STOCK_SYMBOLS	USER1001	Jul 13, 2020 12:53:14 PM

### 12.8.3 Check performance

- \_\_33. With the VIEW in place, it would be worthwhile to run a simple aggregated SQL query in the SQL editor to see its performance.

Select menu (My virtualized data)  $\Rightarrow$  SQL editor.



- \_\_34. Replace the existing SQL with the following by just typing it in, copy from [/Downloads/DV\\_CheckPerformance.txt](#) file or download from <http://ibm.biz/DV-Check-Performance>.

```
SELECT COMPANY, COUNT(*) as COUNT
FROM USER1001.VIEW_CUST_TXN_SYMBOL_COM
GROUP BY COMPANY;
```

(This query performs a simple COUNT across the tables based on the companies traded.)

\_\_35. Click **Run all**.

The aggregated query is run over the View which in turns gets executed across all three underlying tables in the different data sources via the schema folded virtual table created earlier.

Because we do not have a cache and it is going against all the data in every table, this will take a few minutes to complete – please be patient.

SQL editor

```
* Untitled - 1
1 SELECT COMPANY, COUNT(*) AS COUNT
2 FROM USER1001.VIEW_CUST_TXN_SYMBOL_COM
3 GROUP BY COMPANY;
```

Result - Jul 30, 2020 11:28:07 AM

COMPANY	COUNT
3M Company	60735
American Express Company	60690
Apple Inc.	91335
Caterpillar Inc.	60811
Chevron Corporation	60806

Run time: 203.692 s

\_\_36. Notice the time taken for running the query.

The time taken for the simple query execution is unacceptable. The Data Engineer reaches out to the DV Admin to create a cache

 Data Engineer	<p>Based on the time taken to execute the simple aggregated query over the View, makes it a perfect candidate to create a DV data cache from and that will be the next step.</p> <p>If this step fails, check to make sure that you have copied the SQL code properly in the SQL editor and run it again. Keep in mind that a successful run of the SQL may not mean it was actually created properly.</p>
--	--

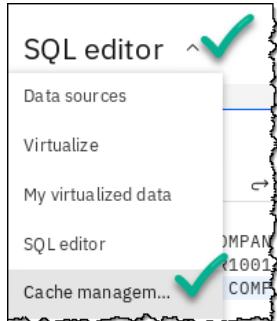
## 12.9 Creating the DV data cache

Given the performance issues observed during the SQL execution, creating a DV data cache from the View is a prudent approach.

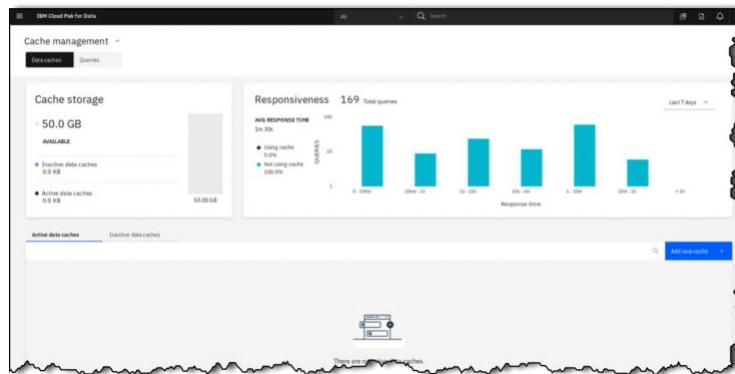


The DV Admin can analyze the queries executed against DV and identify those generated by the dashboard. Looking at the query will help understand what type of cache needs to be created to improve performance for a better dashboard experience.

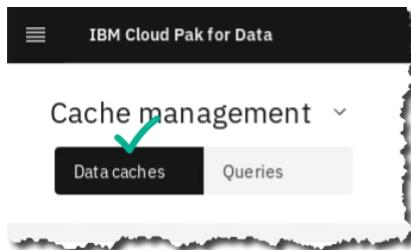
- \_37. Click [Menu \(SQL editor\)](#)  $\Rightarrow$  [Cache Management](#).



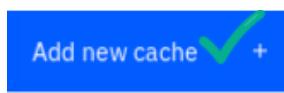
- \_38. You should be presented with your [Cache Management](#) Dashboard.



- \_39. Click on [Data caches](#).



- 40. Towards the bottom right of the screen, find and click **Add new cache +**.



- 41. This opens a SQL Editor to type in the query from which the cache will be created. Type in the query to select all columns from the virtualized view:

```
SELECT * FROM USER1001.VIEW_CUST_TXN_SYMBOL_COM
```

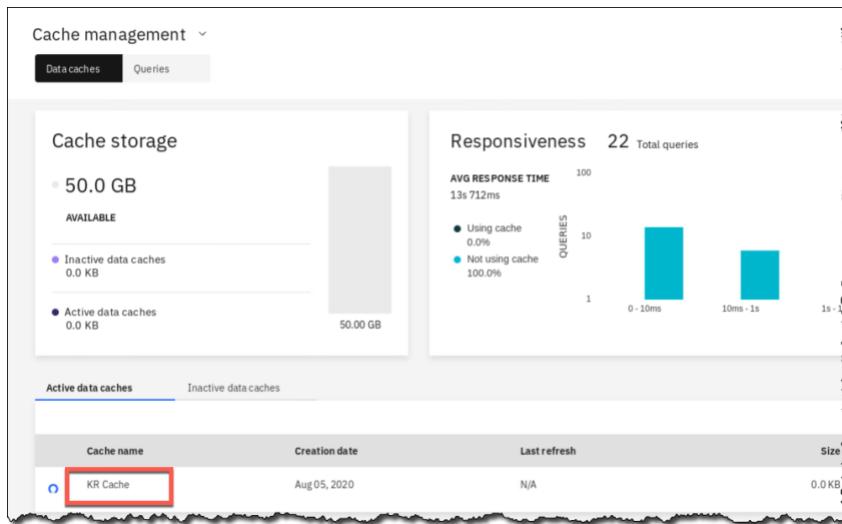
Click **Next**.

- 42. In the *Refresh rate* choose the **default (None, manual only)** and click **Next**.

- 43. Confirm the details on the final page and click **Create**.

(Note: Your cache name will vary from what is shown below.)

- \_\_44. The cache creation process may take some time and the main [Cache Management](#) page will reflect the work in progress.



- \_\_45. Once the cache creation is complete, the newly created cache shows up under the [Active data caches](#) along with other details and its size.

The screenshot shows the 'Active data caches' table. It has columns for 'Cache name', 'Creation date', and 'Last refresh'. One row is visible: 'KR Cache' was created on Aug 05, 2020, and last refreshed on Aug 05, 2020 at 9:09 PM. The 'KR Cache' cell is highlighted with a red box.

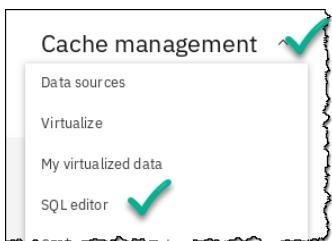
### 12.9.1 Flush the package cache and updating the statistics



Admin

**KNOWN ISSUE** – There exists a defect/issue wherein a query executed before the cache is created and activated fails to use the cache. This is because the query plan is already cached in a relational database such as Db2, so the work around is to clear the Db2 package cache and collect table statistics on the virtual tables created above. You will do that next.

- \_\_46. Click [Menu \(Cache management\)](#) ⇨ [SQL Editor](#).



\_\_47. Remove the view SQL and replace it with the following:

**FLUSH PACKAGE CACHE DYNAMIC**

\_\_48. Click [Run all](#).

```
* Untitled - 1
1 FLUSH PACKAGE CACHE DYNAMIC
```

Result - Jul 30, 2020 12:07:25 PM

FLUSH PACKAGE CACHE DYNAMIC

Status: Success | Affected rows: 0

In a production setting, flushing the entire package may not always be a good idea as it could affect query execution for all users of the database. (We did it here to make this lab simpler.)

Instead, one could opt to selectively flush the package corresponding to particular queries. For more details on that see: <http://ibm.biz/FLUSH-PACKAGE>.

\_\_49. Next, check if the table statistics have been collected. Replace the previous SQL with the following: (You can also download this SQL from <http://ibm.biz/DV-Select-Tab> ).

```
SELECT TABNAME, STATS_TIME, CARD , TYPE
FROM SYSCAT.TABLES
WHERE TABSCHEMA='USER1001'
AND
TABNAME
IN ('STOCK_SYMBOLS','CUSTOMER_TRANSACTIONS');
```

\_\_50. Click [Run all](#).

```
* Untitled - 1
1 SELECT TABNAME, STATS_TIME, CARD , TYPE
2 FROM SYSCAT.TABLES
3 WHERE TABSCHEMA='USER1001'
4 AND
5 TABNAME
6 IN ('STOCK_SYMBOLS', 'CUSTOMER_TRANSACTIONS');
```

Result - Jul 30, 2020 12:11:09 PM

SELECT TABNAME, STATS\_TIME, CARD , TYPE FR... Run time: 0.098 s

TABNAME	STATS_TIME	CARD	TYPE
CUSTOMER_TRANSACTIONS		-1	N
STOCK_SYMBOLS		-1	N

If the cardinality for the virtual tables created before shows -1, this indicates that table statistics have not been collected. These statistics will be collected now.

Note: the [SYSPROC.NNSTATS](http://ibm.biz/SYSPROC-NNSTATS) is a procedure to collect statistics for remote tables: <http://ibm.biz/SYSPROC-NNSTATS>.

\_\_51. If your card shows -1, then perform this step. If not, then you can skip this step.

Replace the previous SQL with the following:

(Or use SQL found here: <http://ibm.biz/DV-Call-SYS>)

```
CALL SYSPROC.NNSTAT(NULL, 'USER1001',
'CUSTOMER_TRANSACTIONS','','',2,'/tmp/collstats1.log',?,1);
```

```
CALL SYSPROC.NNSTAT(NULL, 'USER1001',
'STOCK_SYMBOLS','','',2,'/tmp/collstats1.log',?,1);
```

\_\_52. Click [Run all](#).

```
* Untitled - 1
1 CALL SYSPROC.NNSTAT(NULL, 'USER1001',
2 'CUSTOMER_TRANSACTIONS','','',2,'/tmp/collstats1.log',?,1);
3
4 CALL SYSPROC.NNSTAT(NULL, 'USER1001',
5 'STOCK_SYMBOLS','','',2,'/tmp/collstats1.log',?,1);
```

Status: Success | Affected rows: 0

Parameters		
Name	Type	Data type
?	OUT	VARCHAR

\_\_53. To confirm if the table statistics have now been collected, replace the previous SQL with the following (use the SQL from before):

```
SELECT TABNAME, STATS_TIME, CARD , TYPE
FROM SYSCAT.TABLES
WHERE TABSCHEMA='USER1001'
AND
TABNAME
IN ('STOCK_SYMBOLS','CUSTOMER_TRANSACTIONS');
```

\_\_54. Click [Run all](#).

TABNAME	STATS_TIME	CARD
STOCK_SYMBOLS	2020-07-30 16:14:06.95...	31
CUSTOMER_TRANSACTIONS	2020-07-30 16:14:05.22...	3000000

The actual row counts for the virtual tables are shown as expected instead of -1.

- \_\_\_55. With the package cache flushed and the table level statistics collected, all queries henceforth referencing the virtualize view should start using the cached copy instead of accessing the underlying data sources directly.

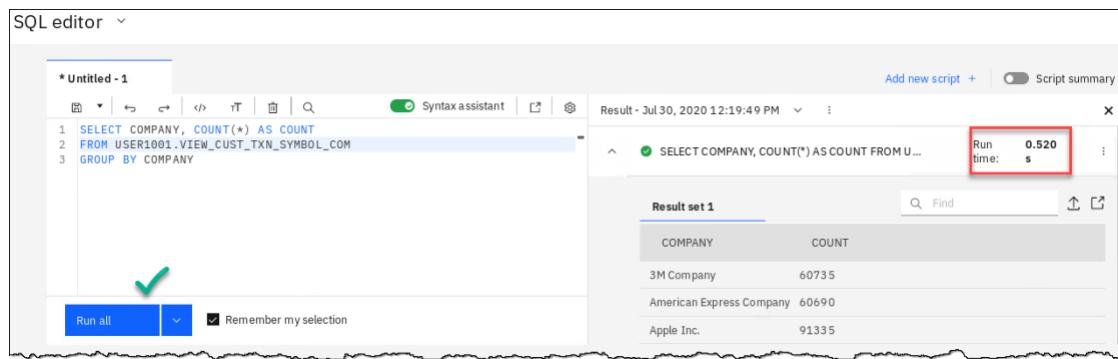
Re-run the previously execute aggregated query on the View:

Remove the previous SQL and type in this again:

```
SELECT COMPANY, COUNT(*) AS COUNT
FROM USER1001.VIEW_CUST_TXN_SYMBOL_COM
GROUP BY COMPANY;
```

This query performs a simple COUNT across the tables based on the companies traded.

- \_\_\_56. Click [Run all](#).



The screenshot shows the SQL editor interface. In the left pane, the script is displayed:

```
* Untitled - 1
1 SELECT COMPANY, COUNT(*) AS COUNT
2 FROM USER1001.VIEW_CUST_TXN_SYMBOL_COM
3 GROUP BY COMPANY
```

In the bottom left corner of the editor, there is a green checkmark icon above the "Run all" button. The "Run all" button is highlighted with a blue background and a white checkmark icon. To the right of the editor is the result pane:

Result - Jul 30, 2020 12:19:49 PM

Run time: **0.520 s**

COMPANY	COUNT
3M Company	60735
American Express Company	60690
Apple Inc.	91335

The Data Engineer informs the BA of the cache creation and shares the View and other relevant details with the BA to continue the dashboard work.

## 12.10 Lab conclusion

Data Virtualization (DV), as part of the Collect phase, facilitates accessing data from various data sources such as DVM for z/OS and performs queries across them. Since data movement is limited, all the access rules and the policies for creating copies remain preserved. In situations where query response times are paramount, DV provides the caching facility, which was covered in this lab.

### \*\* End of Lab 12 – Collect: Virtualizing & Caching from z/OS – Deeper Dive

Lab by Dave Trotter and Rajesh Kartha, Edited by Burt Vialpando and Kent Rubin - IBM

## Lab 13 ORGANIZE – DEEPER DIVE

### 13.1 Lab overview

Many organizations find it difficult to understand their own data because it originates from many sources, is dispersed across many silos, and is controlled by different teams.

This [Organize](#) lab will show you how to uncover the hidden data in your organization's data and how to build a lineage that is otherwise difficult to establish. Cloud Pak for Data helps your organization move from the manual processes required to establish relationships between data to an automated one aided by the platform's built-in machine learning capabilities.

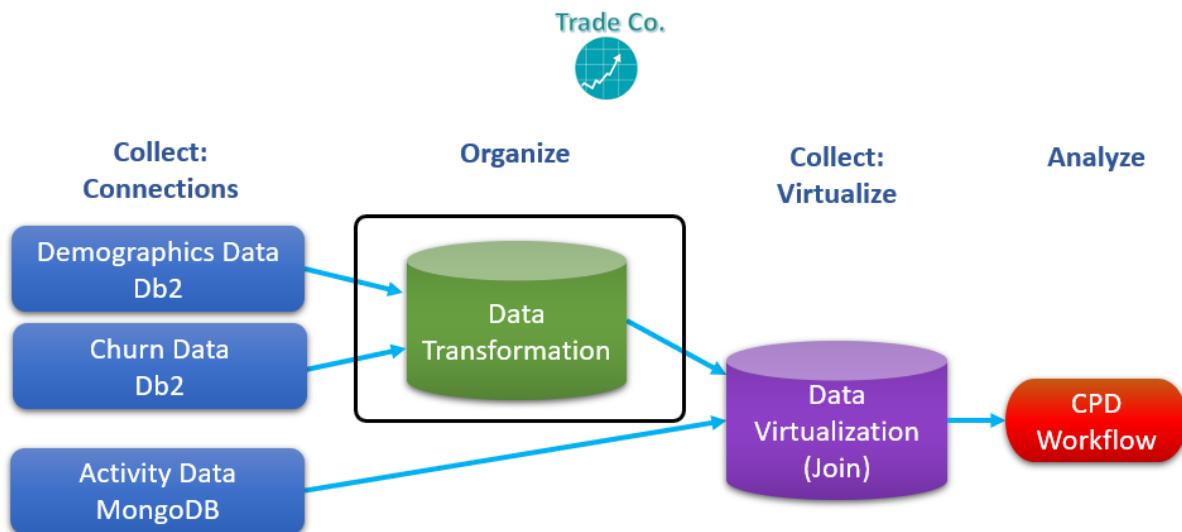
### 13.2 Personas represented in this lab

The [Data Steward](#) and [Data Engineer](#) personas are most likely to perform most of the [Organize](#) tasks shown in this lab.

Persona (Role)	Capabilities
 Data Steward	Data Stewards integrate and transform the data, and provide governance, lineage and classification of the data.
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

The [Data Steward](#) and [Data Engineer](#) personas often work closely together to prepare the data for analytics processing by other personas. For example, in this lab one of the things the [Data Steward](#) will do is to ensure Business Terms are assigned to data assets. In turn, the [Data Engineer](#) will then use that information to find the appropriate data to extract and transform to create a final table of the data sources joined together.

The Data Steward persona also works closely with the [Data Quality Analyst](#) persona.

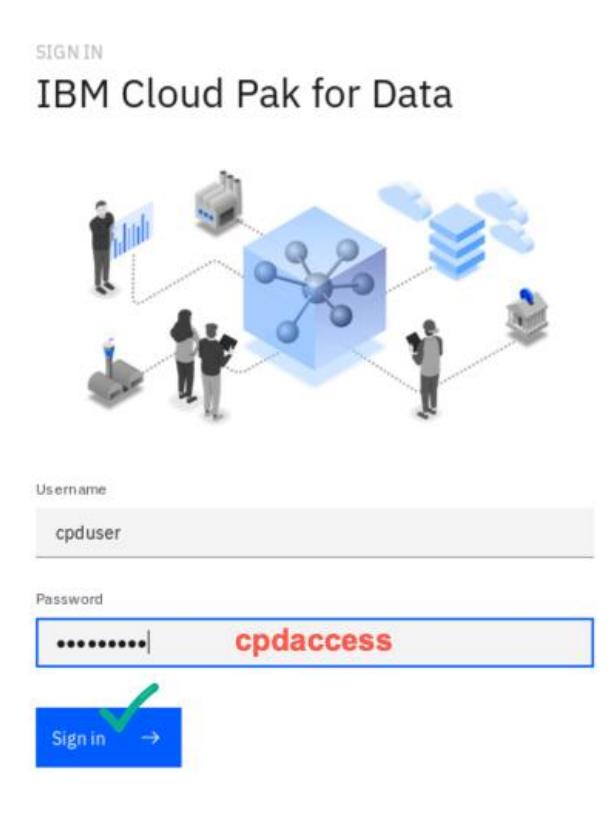


### 13.3 Logging into the CPD web client (if you have not already done so)

- 1. If you are starting this lab stand-alone (without going through previous labs) do the following:
- 2. Double-click the desktop icon: [Cloud Pak for Data Web Client](#).



- 3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click [Sign In](#).

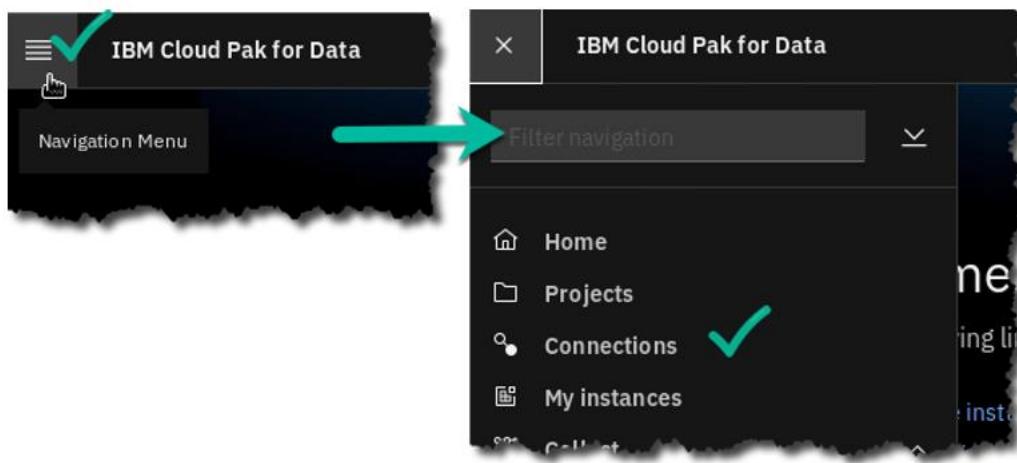


The screenshot shows the "SIGN IN" page for IBM Cloud Pak for Data. At the top, there is a decorative graphic of a central node connected to various icons representing people, databases, and servers. Below the graphic, the text "IBM Cloud Pak for Data" is centered. The sign-in form consists of two input fields: "Username" containing "cpduser" and "Password" containing "cpdaccess". A blue "Sign in" button at the bottom has a green checkmark icon and a right-pointing arrow. The entire form is set against a light gray background.

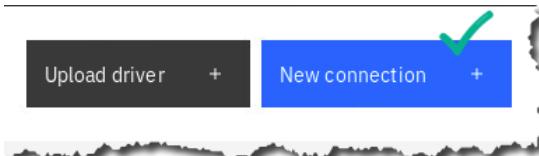
## 13.4 Creating a connection to your data

Data Connections are one of the primary ways that Cloud Pak for Data can access data both within the application and across various sources, including cloud, on-premises, application, semi-structured, etc. These connections can be created globally and then used by users within the different capabilities of the CPD platform.

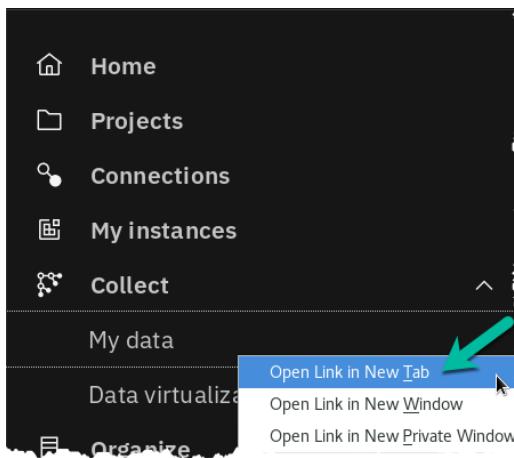
- 4. If you are starting this lab stand-alone (without going through previous labs) do the following: In the CPD web client, start at the [Navigation Menu](#) ⇨ Click [Connections](#).



- 5. Click [New connection](#).



- 6. Before creating a connection, you will need to know the credentials to the database. You can find this by clicking on [Navigation Home](#), [Collect](#), then [right click on My data](#) and [Open Link in New Tab](#).



- \_\_7. Click on the tab Databases ⇒ Click on the tile Db2 Advanced Edition Ellipse ⇒ Details.

The screenshot shows a user interface titled 'My data'. At the top, there are tabs for 'Data sets', 'Data sources', 'Data requests', and 'Databases', with 'Databases' being the active tab, indicated by a blue underline and a green checkmark. Below the tabs is a search bar with the placeholder 'Find databases' and a 'Filter by: Types' dropdown. The main area displays two database entries: 'Db2 Advanced Edition' and 'MongoDB-1'. The 'Db2 Advanced Edition' entry has a context menu open, with the 'Details' option highlighted by a green checkmark. Other options in the menu include 'Open database', 'Configure', 'Submit connection for approval', 'Manage access', and 'Delete'. The 'MongoDB-1' entry is listed below it, showing it is a 'MongoDB Enterprise' instance created on Jun 23, 2020, at 11:22 AM.



Data  
Steward

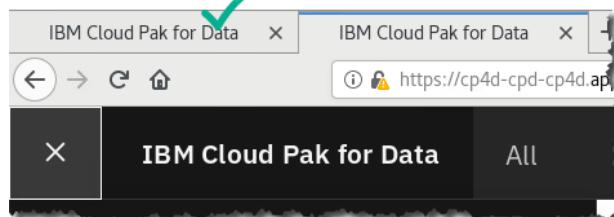
One of the advantages of using a full platform that uses containers and microservices is that you can have a host of capabilities that can be turned on, and with the correct user privilege, instances of those capabilities (like a new MongoDB or Db2 instance) can be created with the click of a button – just as you would be able to do in a Public Cloud, but behind your firewall.

- \_\_8. Page down to the bottom of the screen and you'll see the access information for our database. Copy down this information to enter in our Data Connections window.

The screenshot shows the 'Access information' section of a configuration interface. It includes fields for 'Username' (user1001), 'Password' (mh3b\*4v1\_4F?X\_WK), and 'JDBC Connection URL' (jdbc:db2://worker5.clusterw932030/BLUDB). The 'Port' field is also visible. The 'Hostname' field under 'Nodes' is highlighted with a red box.

HOSTNAME	CPU	MEMORY
worker5.clusterw9	5 cores	27 GiB

- \_\_9. Now return to the previous tab.



10. Enter the parameters to create a connection to the internal Db2 database used for this workshop. Below is a sample; Connection Name and Description can be whatever you choose, the rest of the credentials you can get from the previous step.

## New connection

The screenshot shows the 'New connection' dialog box with the following fields filled in:

- Connection name:** Db2\_Source\_Local (with a green checkmark)
- Description (optional):** Db2 Local Database (with a green checkmark)
- Connection type:** Db2 (with a green checkmark)
- Host:** worker5.clusterw9 (with a green checkmark)
- Port:** 32030 (with a green checkmark)
- Database:** BLUDB (with a green checkmark)
- Username:** user1001 (with a green checkmark)
- Password:** A masked password (with a green checkmark)

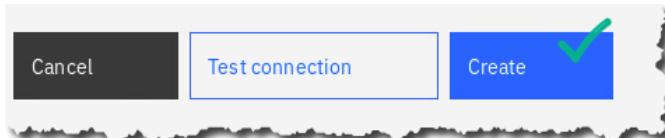
At the bottom, the 'Test connection' button is highlighted with a green checkmark, indicating a successful test.

 <b>Data Steward</b>	<p>Note: You may copy and paste the password as shown below. (if this doesn't work, then try highlighting the password and use [CTL]-c then [CTL]-v).</p> <p>Access information</p> <table border="1"> <tbody> <tr> <td>Username</td><td>user1001</td></tr> <tr> <td>Password</td><td>mh3b*4v1_4F?X_WK</td></tr> <tr> <td>JDBC Connection URL</td><td>jdbc:db2://worker5.clusterw9:32030 /BLUDB</td></tr> </tbody> </table>		Username	user1001	Password	mh3b*4v1_4F?X_WK	JDBC Connection URL	jdbc:db2://worker5.clusterw9:32030 /BLUDB
Username	user1001							
Password	mh3b*4v1_4F?X_WK							
JDBC Connection URL	jdbc:db2://worker5.clusterw9:32030 /BLUDB							
								

- \_\_11. Click **Test connection** (to verify your credentials are correct).
- \_\_12. That will return a successful message if the connection parameters are correct.



- \_\_13. Click the **Create** button to add the connection to CPD. At this point it is a 'Global Connection'. We will later add it to different projects, catalogs, etc.



## 13.5 Working in a project

Working with Projects is an important part of the Cloud Pak for Data experience. Projects allow you to organize your work into specific areas to which you can control access, and once you are happy with the results, you can publish the assets and the findings to the Catalog for general user consumption.

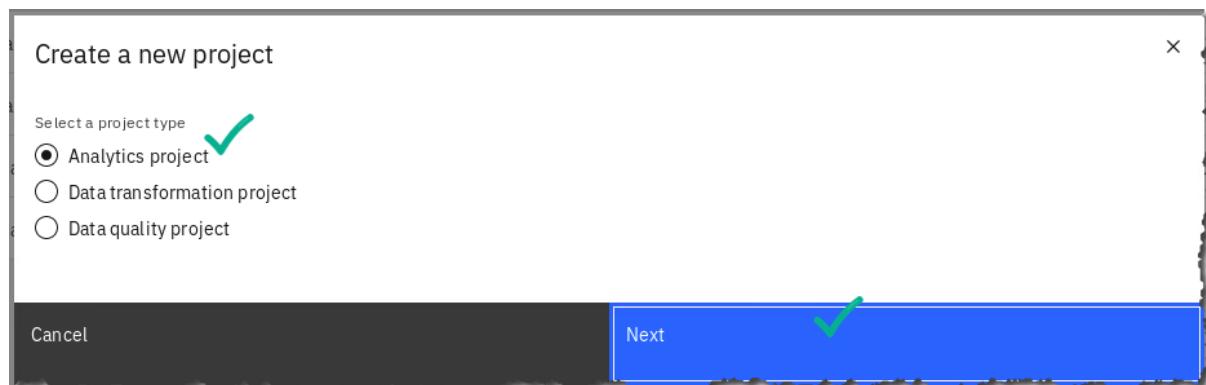
- 14. Let's start by creating a project:

In the CPD web client, click the [Navigation Menu](#) ⇒ [Projects](#) ⇒ Click [New project](#).



- 15. There are three primary types of projects you can create.

Click [Analytics project](#) ⇒ [Next](#).



- \_\_16. You can create an empty project, create a project by importing existing assets from your file system, or from a Git repository.

Click tile [Create an empty project](#).

← Back

## Create a project

Create an empty project, and then add data and choose the right tools to accomplish your goals.

Create an empty project

Add the data you want to prepare, analyze, or model. Choose the tools you want to work: write code, create a flow on a graphic interface, or automatically build models.

- \_\_17. Add the [Name](#) and [Description](#) for your project.

IBM Cloud Pak for Data

### New project

Define project details

Name

Data Analysis Project

Description

A project to analyze our data and prepare it for publishing to the catalog

Choose project options

Integrate this project with Git

 Data Steward	<p>Note: You can also integrate the project to Git, which allows for automatic saves to a Git repository – we won't be doing that today.</p>
---	--

\_\_18. Click **Create**.



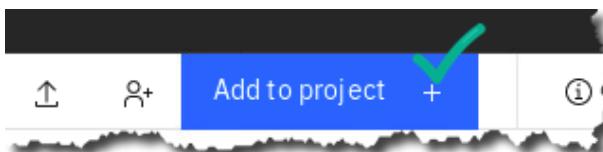
This will create a project and put you into the project overview screen.

This screen shows summary information about who is a collaborator on this project, what assets have been created or exist in this project, etc. Along the top you see tabs for the different types of things we'll do with this Project, for example, Assets, Environments, etc.

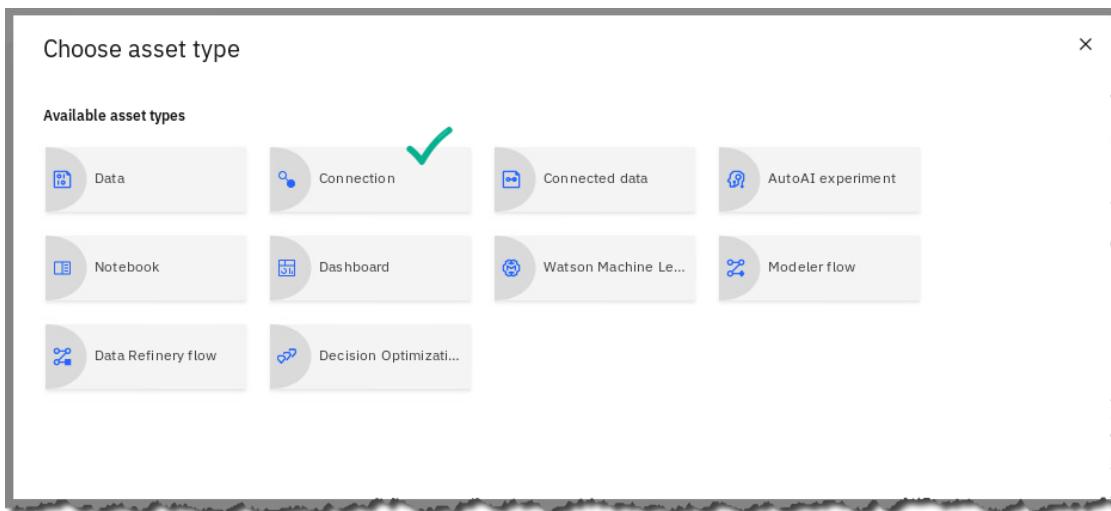
### 13.5.1 Adding a connection

Next, we'll start to bring data into our project so we can start to analyze, understand and enhance our data, to get it ready for publishing to a Catalog for public use.

Click [Add to project](#).

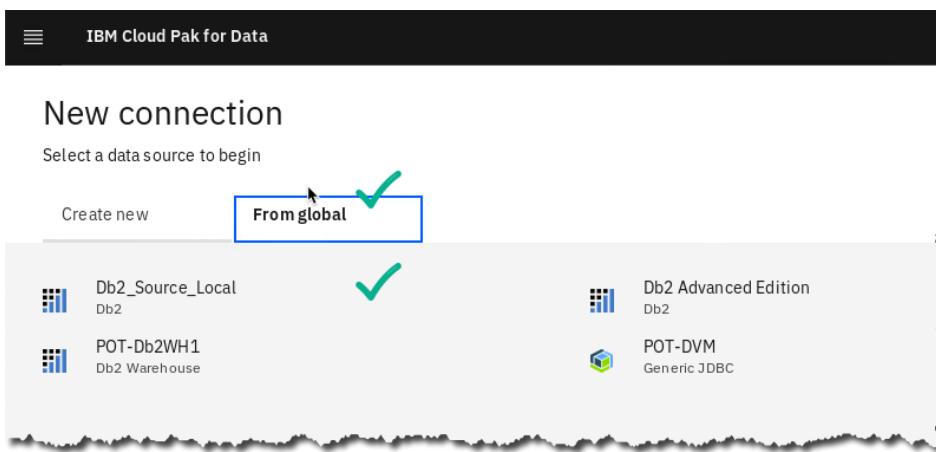


\_\_19. Choose [Connection](#), and we will add the connection we created earlier.



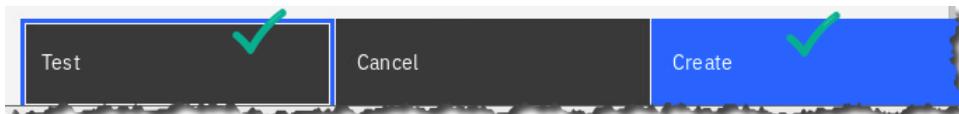
\_\_20. Here, you see a number of different choices for creating a connection. This is one of the strengths of CPD – the breadth of different types of data you can access directly from the platform. As we said, we're going to add the connection we created earlier.

Choose [From global](#)  $\Rightarrow$  [Db2\\_Source\\_Local](#).



 Data Steward	<p>Note: If this message appears, click the X</p> <div style="border: 1px solid #ccc; padding: 5px; background-color: #fff; margin-top: 10px;"> <b>⚠️</b> You can't discover assets through a connection that has personal credentials. <span style="float: right;">X</span> </div>
--	---

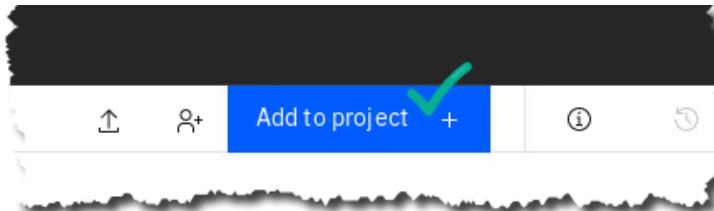
- \_\_21. Click [Test](#) (to validate we can connect to our data source), then click [Create](#).



### 13.5.2 Adding connected data

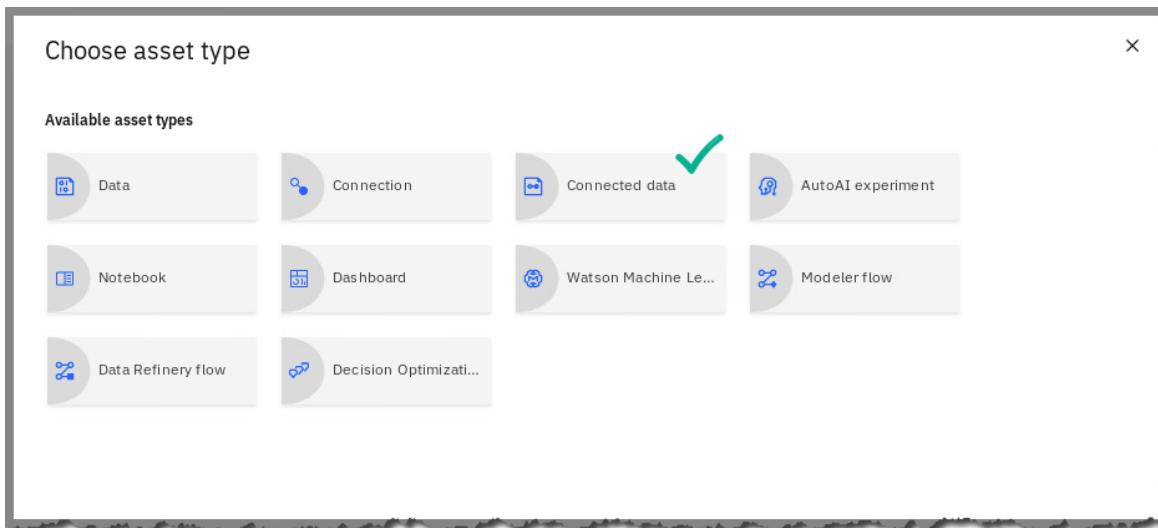
- \_\_22. You've now added your first asset to the project (a connection). Next we'll want to use that connection to bring in some data and start to analyze and understand it.

Once again, click [Add to Project](#).



- \_\_23. Notice you can also add files directly from your desktop, by either dragging or browsing your file system.

Now choose [Connected data](#).



- \_\_24. Click **Select source**.

The screenshot shows the 'New connected data asset' page. At the top, there's a navigation bar with 'My Projects / Data Analysis Project / New connected data asset'. Below it, the main title is 'New connected data asset' with the subtitle 'Add a table or file that's accessed through a connection.' On the left, there's a sidebar labeled 'Source' with a 'Select source' button highlighted by a green checkmark.

- \_\_25. Next, we'll choose our Connection, Schema, and Table(s). First, click **Db2\_Source\_Local** (or the name of your connection).

We are interested in two tables, that are in two different schemas.

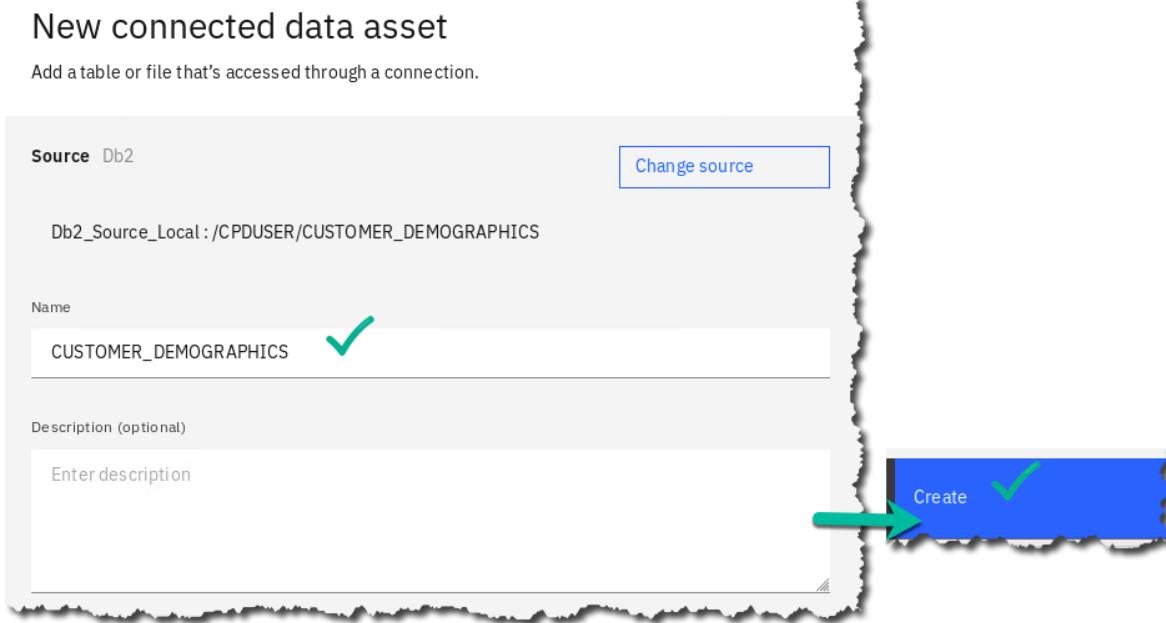
Choose the schema **CPDUSER**  $\Rightarrow$  **CUSTOMER\_DEMOGRAPHICS**  $\Rightarrow$  **Select**.

Select connection source

The screenshot shows a 'Select connection source' interface with a table titled 'Connections'. The table has three columns: 'Connections', 'Db2\_Source\_Local', and 'CPDUSER'. Under 'Connections', 'Db2\_Source\_Local' is selected and highlighted with a green checkmark. In the 'Db2\_Source\_Local' column, 'CUSTOMER\_DEMOGRAPHICS' is also highlighted with a green checkmark. A large blue button at the bottom right is labeled 'Select' with a green checkmark, and a green arrow points to it from the right.

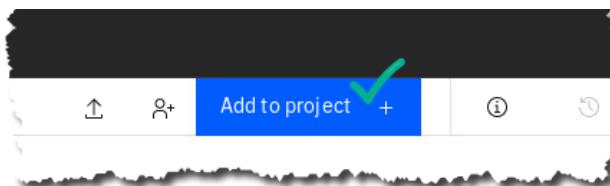
Connections	Db2_Source_Local	CPDUSER
Connections (1) <span style="color: green;">✓</span>	Schemas (7)	Tables (5)
Db2_Source_Local <span style="color: green;">✓</span>	AIOSFASTPATHICP >	BATCH_JOB_OUTPUT
	AIOSFASTPATHPROC >	CUSTOMER_CHURN
	AIOSFASTPATHPROCERROR >	CUSTOMER_DEMOCHURN
	CPDUSER <span style="color: green;">✓</span>	CUSTOMER_DEMOGRAPHICS <span style="color: green;">✓</span>
	DB2COGNOS >	JOINED_CUSTOMERS
	SOLUTIONS >	
	USER1001 >	

- \_\_26. Now we'll give it a name to refer to it by in our project. Enter CUSTOMER\_DEMOGRAPHICS for the Source name  $\Rightarrow$  [Create](#).

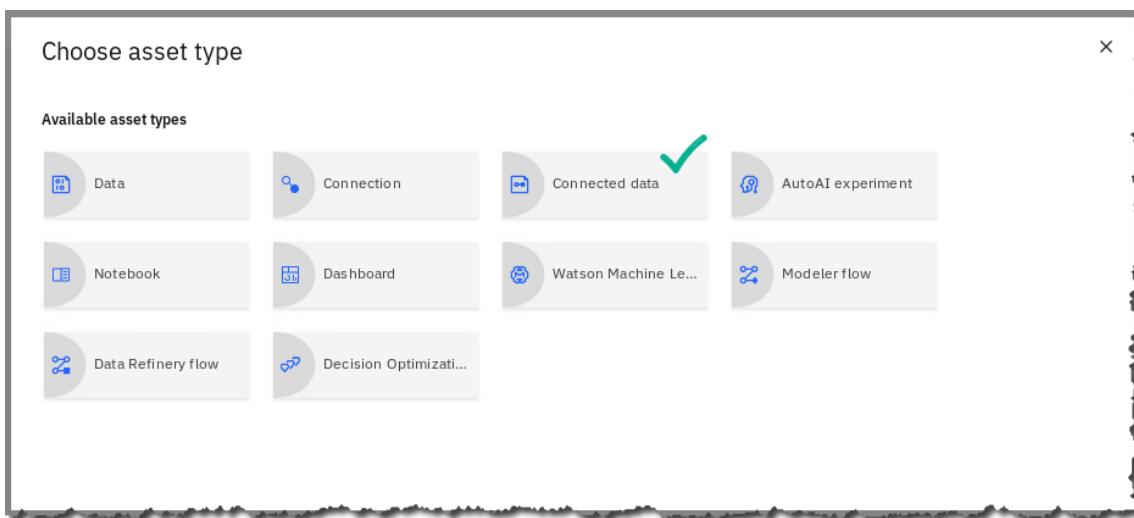


- \_\_27. You've now added your first data asset to the project! Let's add a second table from another schema.

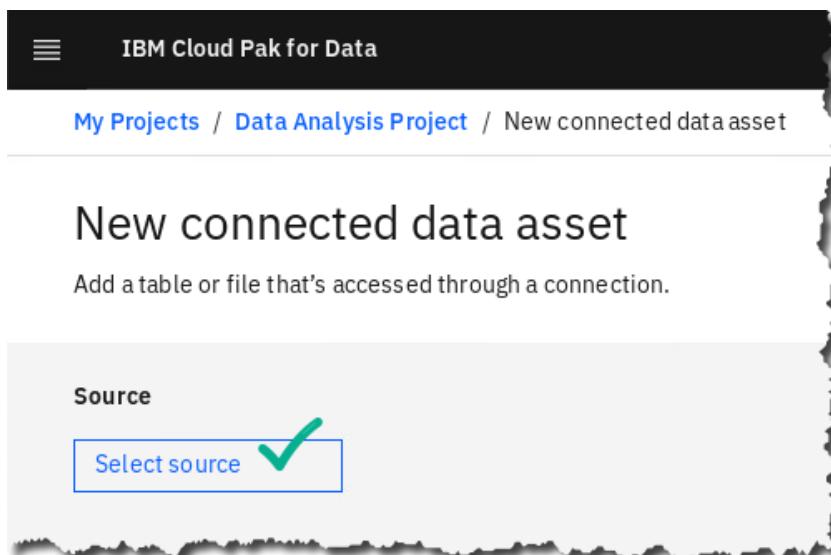
Once again, click [Add to Project](#).



- \_\_28. Choose [Connected data](#) again.

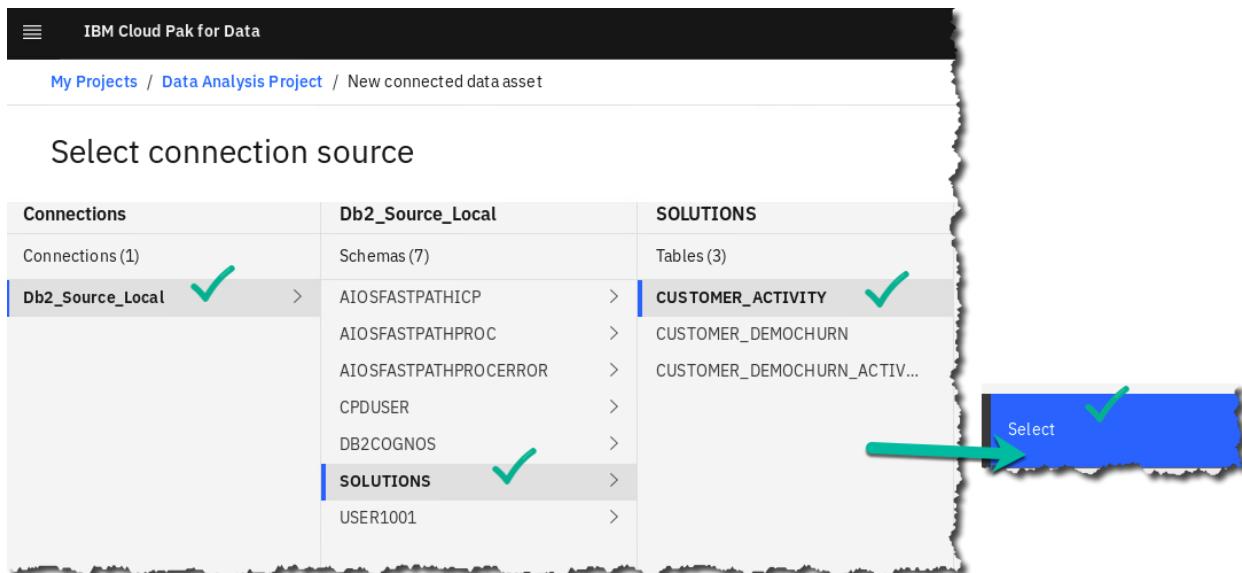


\_\_29. Click **Select source**.



The screenshot shows the 'New connected data asset' page. At the top, there's a navigation bar with 'My Projects / Data Analysis Project / New connected data asset'. Below it, the title 'New connected data asset' is displayed. A sub-instruction says 'Add a table or file that's accessed through a connection.' On the left, there's a 'Source' section with a 'Select source' button, which has a green checkmark over it.

\_\_30. Now select – Db2\_Source\_Local  $\Rightarrow$  SOLUTIONS  $\Rightarrow$  CUSTOMER\_ACTIVITY  $\Rightarrow$  then Select.



The screenshot shows the 'Select connection source' interface. It has three main sections: 'Connections', 'Db2\_Source\_Local', and 'SOLUTIONS'. Under 'Connections', 'Db2\_Source\_Local' is selected and highlighted with a green checkmark. Under 'Db2\_Source\_Local', there are several items: 'Schemas (7)', 'AIOSFASTPATHICP', 'AIOSFASTPATHPROC', 'AIOSFASTPATHPROCERROR', 'CPDUSER', 'DB2COGNOS', 'SOLUTIONS' (which also has a green checkmark), and 'USER1001'. Under 'SOLUTIONS', there are three items: 'Tables (3)', 'CUSTOMER\_ACTIVITY' (which has a green checkmark), 'CUSTOMER\_DEMOCHURN', and 'CUSTOMER\_DEMOCHURN\_ACTIV...'. A large blue arrow points from the bottom right towards a 'Select' button, which is also highlighted with a green checkmark.

- \_\_31. Name the asset CUSTOMER\_ACTIVITY ➔ Create.

Source Db2

Db2\_Source\_Local : /SOLUTIONS/CUSTOMER\_ACTIVITY

Name

CUSTOMER\_ACTIVITY ✓

Description (optional)

Enter description

Create ✓

- \_\_32. When complete, your assets list for your project should contain 3 assets; two tables and a connection.

<input type="checkbox"/>	Name	Type	Created by	Last modified
<input type="checkbox"/>	CUSTOMER_ACTIVITY ✓	Data Asset	CPD User	Jun 29, 2020, 1:07 PM
<input type="checkbox"/>	CUSTOMER_DEMOGRAPHICS ✓	Data Asset	CPD User	Jun 29, 2020, 12:58 PM
<input type="checkbox"/>	Db2_Source_Local ✓	Connection	CPD User	Jun 29, 2020, 12:48 PM

### 13.5.3 Exploring and Profiling data

One of the most powerful capabilities of Cloud Pak for Data is its ability to automatically create descriptive information, business classifications, etc., about your data. This information will then be exploited to do things such as determine where sensitive information is, decide how we should handle such information, describe who can see it and who cannot, assign business terminology to the data to make it easy for business users to locate and understand, and a whole host of other capabilities.

- 33. Next we'll start to get an understanding of our data; click on the **CUSTOMER\_DEMOGRAPHICS** table.

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 the navigation bar, the 'Assets' tab is selected. A search bar is present above a list of assets. The asset list includes:

- CUSTOMER\_ACTIVITY**: Data Asset, CPD User, Jun 29, 2020, 1:07 PM
- CUSTOMER\_DEMOGRAPHICS**: Data Asset, CPD User, Jun 29, 2020, 12:58 PM (marked with a green checkmark)
- Db2\_Source\_Local**: Connection, CPD User, Jun 29, 2020, 12:48 PM

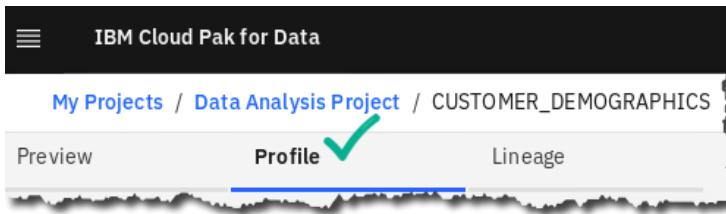
- 34. This will bring up a preview window showing you a sample of our data, and a number of options to work with it. Take a minute to familiarize yourself with this screen.

The screenshot shows the 'Preview' tab for the CUSTOMER\_DEMOGRAPHICS table. It displays a schema with 18 columns and a preview of 1000 rows. The columns are:

ID	GENDER	STATUS	CHILDREN	ESTINCO...	HOMEOW...	AGE	TAXID	CREDITC...	DOB	ADDRESS...	ADDRESS...	CITY	STA
0	F	S	1	38000	N	24	147889187	6549061697939	1947-11-11	159 HUTTON ST		ABSECON	NJ
1	M	M	2	29616	N	49	113772166	6436360484417	1992-03-17	31 WOODLAND R		SAINT LOUIS	MO
2	M	M	0	19732.8	N	51	132420919	4849378808118	1907-09-08	1910 COCHRAN I		KEARNY	NJ
3	M	S	2	96.33	N	56	700548452	2926742654852	1980-04-29	187 HAYES MILL		RUSTON	LA
4	F	M	2	52004.8	N	25	141013706	4132500804622	1979-01-16	RR 1 BOX 57B		MONTGOMERY	AL
5	M	M	2	53010.8	N	19	163371244	2231773884473	1992-12-06	7850 45TH AVE N		CHESTER	MA
6	M	M	1	75004.5	N	65	182544864	7349439804241	1911-04-24	RR 1 BOX 47		NEW CASTLE	PA
7	M	M	0	19749.3	N	60	206227068	5553618912566	1912-11-23	515 KENSINGTO		ISSAQAH	WA
8	M	S	1	57626.9	Y	44	131099071	9119007527242	1916-04-26	6077 STATE ROU		SHAVERTOWN	PA
9	M	M	2	20078	N	33	119762649	6813572896826	1977-02-09	188 W OLYMPIC I		EL PASO	TX
10	F	M	2	47902	N	26	817366094	2046608099384	1905-02-01	21579 LARAMIE		PHILADELPHIA	PA
11	M	M	1	7545.96	Y	17	451541224	4045479553572	1924-10-01	1 PLAINVILLE CI		HAVERHILL	MA
12	F	S	0	78851.3	N	48	124158559	8979358234254	1933-06-14	4716 SW VIOLA C		TENAFLY	NJ
13	F	S	1	17540.7	Y	63	163930462	9050922700714	1942-12-12	1 HAUN RD		PHAROAH	OK
14	F	M	0	83891.9	Y	61	165912006	6325263828540	1947-01-25	4220 BARSTOW		ARENA	ND
15	F	M	2	28220.8	N	39	235315405	4562971044212	1946-02-20	10167 OAK HOLL		STRASSTOWN	PA
16	F	S	0	28589.1	N	16	730825728	8111200911782	1916-10-05	D18 CALLE 5		HOUSTON	TX
17	F	M	2	5237.63	N	49	908144755	1349244456282	1938-10-02	3579 N 47TH AVE		KENOSHA	WI

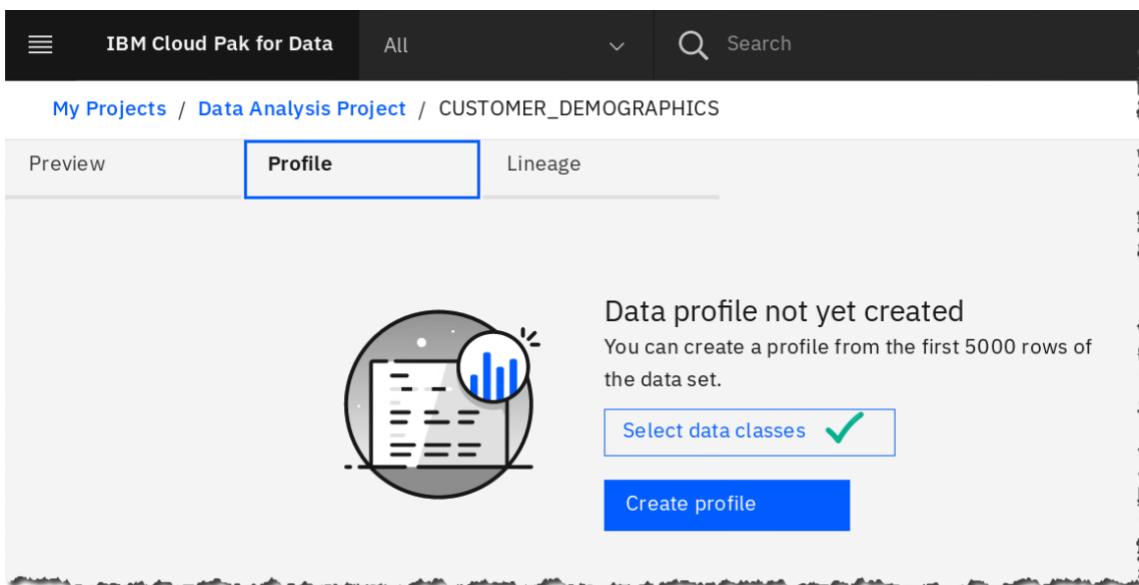
- \_\_35. The first thing we'll do is profile the data; profiling is a common practice of inspecting the data to understand basic characteristics of your columns such as frequency distributions, formats, completeness, etc. In Cloud Pak for Data, you will see that it goes a lot further than that...

- \_\_36. Click the tab **Profile**.



This will open the profile launch window.

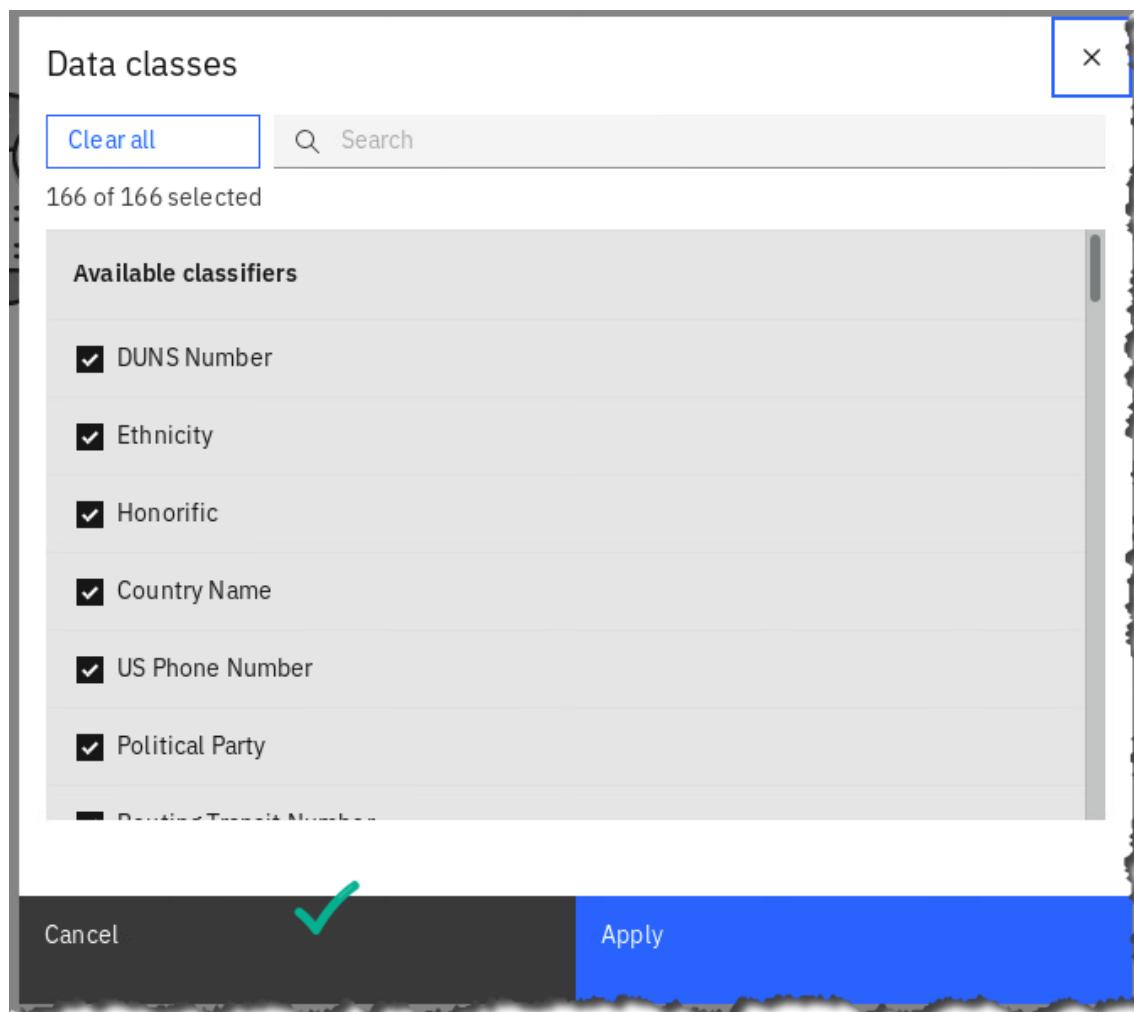
- \_\_37. First, click **Select Data Classes**.



- \_\_38. This will show a list of all “out of the box’ data classes that can be assigned automatically during profiling to the different data columns. (Sometimes this does not populate – if so, [Cancel](#) and reselect.)

This window allows you to de-select certain classes if they aren’t pertinent to your business or to this profiling exercise. For now, we will leave them all active. In addition, custom data classes can be defined for your business and added to this repository.

Click [Cancel](#) (to return to the profile launch screen).



39. Now click **Create Profile**.

My Projects / Data Analysis Project / CUSTOMER\_DEMOGRAPHICS

Preview      **Profile**      Lineage

Data profile not yet created  
You can create a profile from the first 5000 rows of the data set.

Select data classes

Create profile ✓

Note: The profiling process will launch. This will take a few moments to complete so feel free to leave this screen and return, but it should only take between 2-3 minutes.

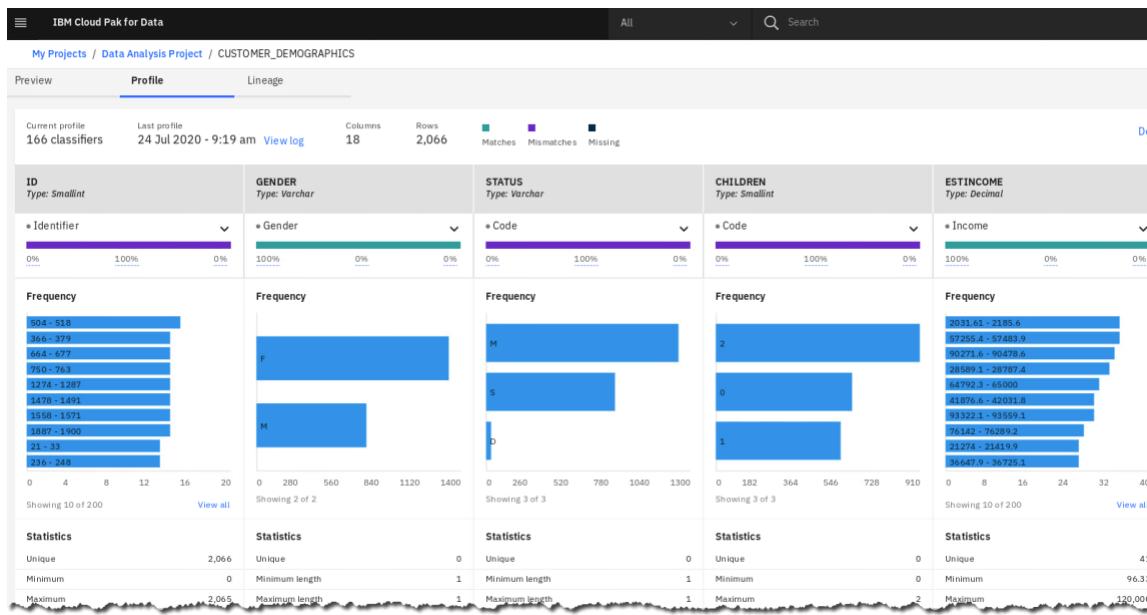
Updating data profile  
Feel free to continue working or stay here and refresh the page occasionally to get an update on the profile's status. When the profile's ready, you'll be able to view it.

Click **refresh** on the browser to see if it finished.

IBM Cloud Pak for Data

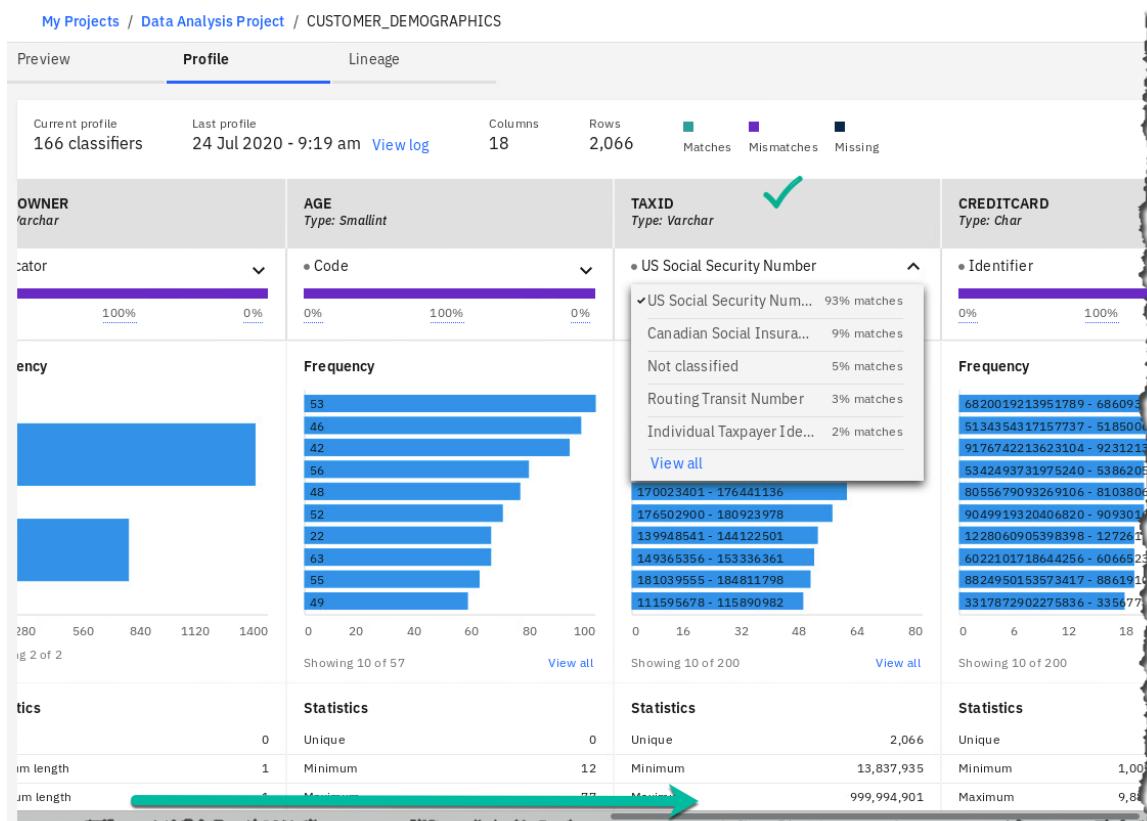
IBM Cloud Pa

—40. When it is complete, you will see:



—41. This is a bird's eye view of the results that have been generated by the profile run. Note that you can immediately determine frequencies and pay attention to the data classes that have been identified. Here we are looking at Gender, which was automatically determined during profiling based on the column metadata and the data itself.

Scroll right to see other columns, like TAXID...



- \_\_42. Click back to the section [Preview](#).

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there is a navigation bar with the text "IBM Cloud Pak for Data" and "All". Below the navigation bar, the path "My Projects / Data Analysis Project / CUSTOMER\_DEMOGRAPHICS" is displayed. A horizontal menu bar below the path contains three items: "Preview" (which has a green checkmark icon and is highlighted with a blue border), "Profile", and "Lineage".

- \_\_43. Click on the down arrow for **TAXID** to see how Profiling found the identifier: [US Social Security Number](#).

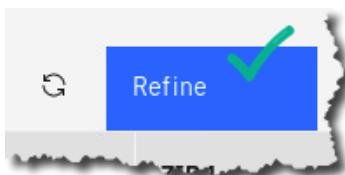
The screenshot shows a data preview table with four columns: AGE, TAXID, CREDITC..., and DOB. The TAXID column is currently active, indicated by a dropdown menu that is open over the first row. The dropdown menu shows the value "US So..." followed by a truncated version "US Social Se..." and a full version "Canadian So...". Other options in the dropdown include "Not classifi..." and "Routing Tra...". The other columns show numerical values (e.g., 24, 49, 51, 56, 25) and date/timestamp values (e.g., 1941-01-01, 1992-01-01, 1907-01-01, 1986-01-01, 1979-01-01). A green checkmark is placed on the dropdown menu and its truncated version.

AGE Smallint	TAXID String	CREDITC... Char	DOB Date
Code	US So...	Identifi...	Date
24	US Social Se...	6549061697939	1941-01-01
49	Canadian So...	6436360484417	1992-01-01
51	Not classifi...	4849378808118	1907-01-01
56		2926742654852	1986-01-01
25	Routing Tra...	4132500804622	1979-01-01
10	162271244	2231773884473	1992-01-01

### 13.5.4 Refining visualizations

The next steps will allow us to use visualization techniques to further get an understanding of our data; visualization is a powerful way to get quick insights and make decisions about what we want to do next.

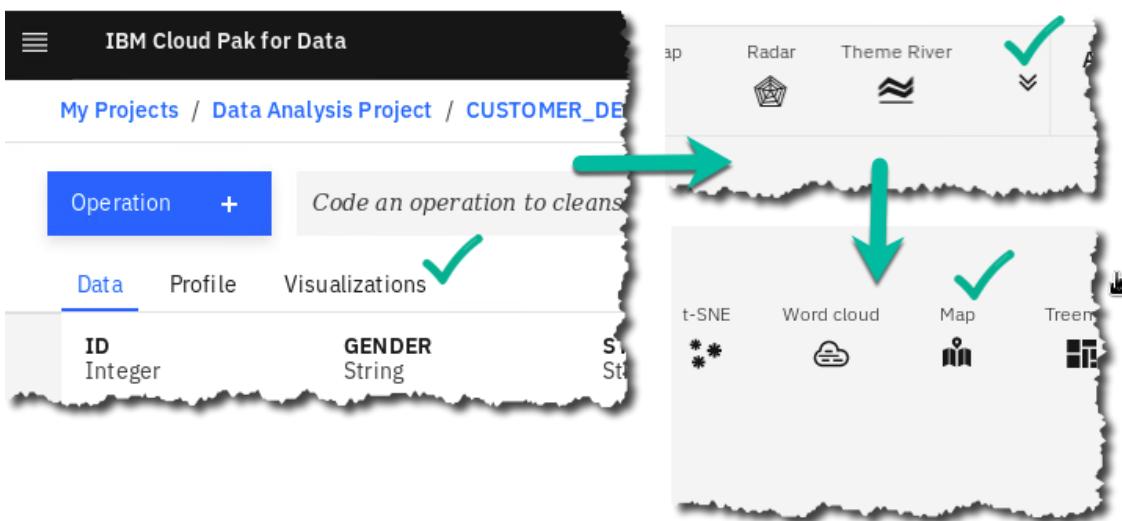
- \_44. Click [Refine](#).



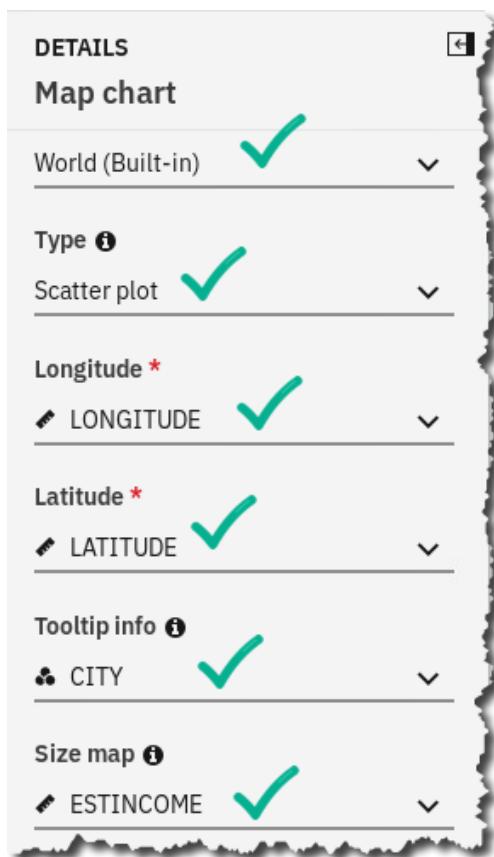
 Data Steward	<p>Note: If you receive this error, back up in the browser and refresh to fix.</p> 
---	---

- \_45. Click [Visualizations](#) ⇒ [Map](#).

NOTE: You may need to click the double drop down arrow to expand the list of visualizations to find [Map](#).

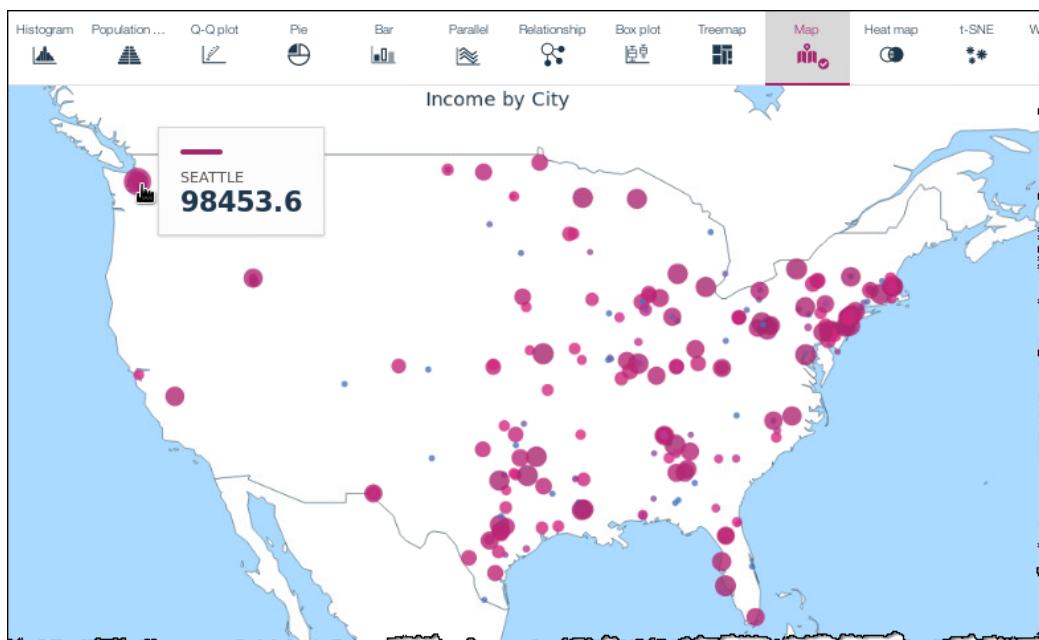


\_\_\_46. Fill in the **Details** as shown below.



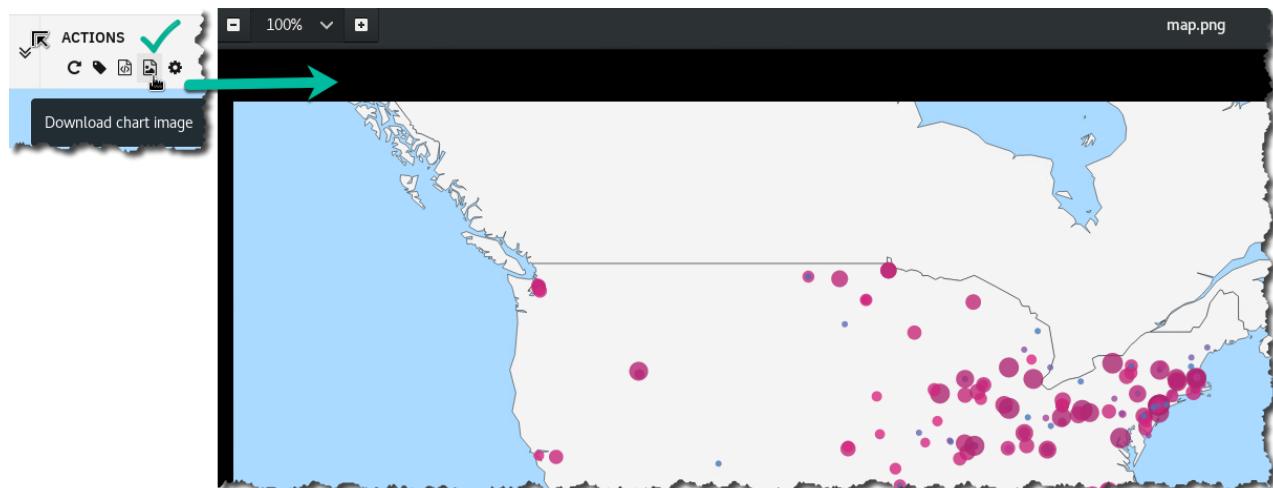
\_\_\_47. Use your mouse's wheel to **zoom in** and center the United States on the visualization.

Hover over any circle on the map, which indicates the Estimated Income by that City.

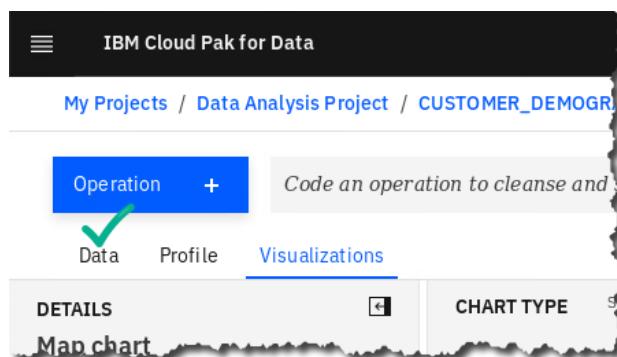


- 48. If the Data Scientist or Business Analyst want to keep a visualization, they can download it by clicking on the image icon shown.

Then, they can either save it, or review it with [Open with Image Viewer](#) (depending on browser and OS).



- 49. Close the Image viewer (if you have it open) and click on the [Data](#) tab to return to your data view (you may see a popup asking if you want to leave the page after clicking [Data](#). If so, choose yes).

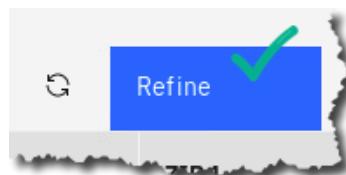


### 13.5.5 Data refinery flows

You should now be back at the data grid view page. (See below)

Data Refinery flows are a powerful way to make adjustments to your data, in a visual, self-service environment. As you choose steps, it will create a pipeline process; each step is reversible and editable as well so that you may back up in case of error.

- 50. Click [Refine](#) if you are not already in the Data Refinery Flow Editor.



- \_\_51. The screen will refresh with a slightly different look; you'll still have your data grid, but you'll also see some new options for 'Operation', 'Steps', etc.

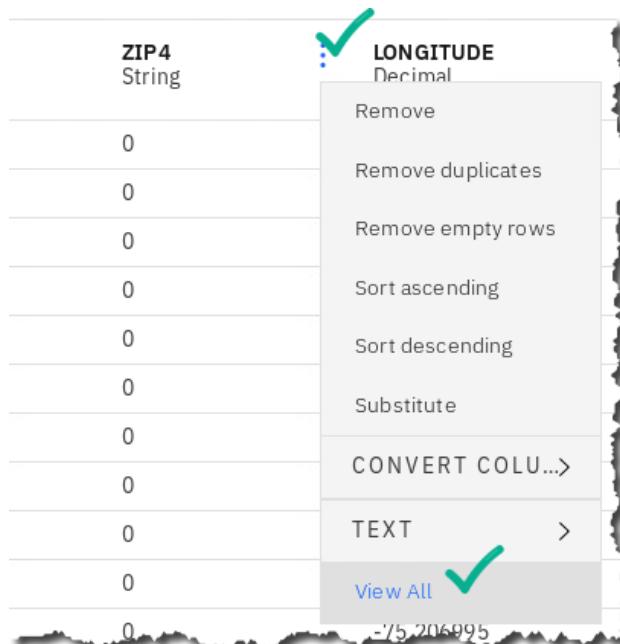
	ID Integer	GENDER String	STATUS String	CHILDREN Integer	ESTINCOME Decimal	HOMEOWNER String	AGE Integer
1	0	F	S	1	38000	N	24
2	1	M	M	2	29616	N	49
3	2	M	M	0	19732.8	N	51
4	3	M	S	2	96.33	N	56
5	4	F	M	2	52004.8	N	25
6	5	M	M	2	53010.8	N	19
7	6	M	M	1	75004.5	N	65
8	7	M	M	0	19749.3	N	60
9	8	M	S	1	57626.9	Y	44
10	9	M	M	2	20078	N	33
11	10	F	M	2	47902	N	26
12	11	M	M	1	7545.96	Y	17

- \_\_52. You have been working in what is known as the Data Refinery. This module gives you many different capabilities for transforming, or "shaping" your data, using an easy to understand spreadsheet and menu-driven function paradigm.

The first thing we want to do is scroll over to our two zip code fields, ZIP and ZIP4. Note that the zip field sometimes contains only 4 digits instead of 5. This commonly happens when the field at some point was defined as numeric and the leading digit is a zero – it automatically gets dropped. Many east coast regions have zip codes that start with zero – we'll want to pad the field to get the 0 back. Secondly, we'll want to pad all ZIP4 fields to have 4 zeros if there is a zero in the field.

- \_\_53. Click on the **ellipses** next to the ZIP4 field. This will bring up a floating menu with a number of actions, or what we call operations, that you can use to transform our zip field. Some commonly used functions are called out directly, but we want to pad, so...

Click on **View All**.



*Tip: Use your browser zoom in/out feature to see the View All option.*

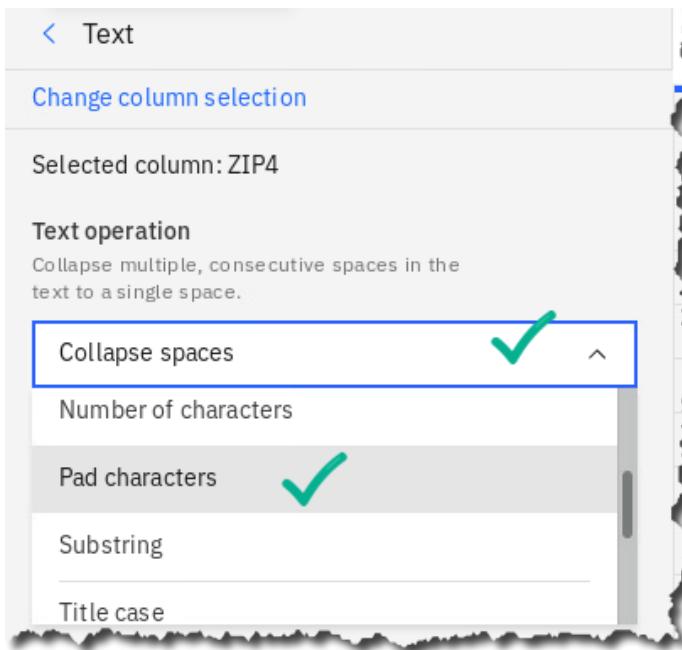
- \_\_54. Now we see all of the shaping operations by category on the left, and the column we're concentrating on, ZIP4.

We're making a text transformation to pad missing numbers with zeros, so click **Text**.



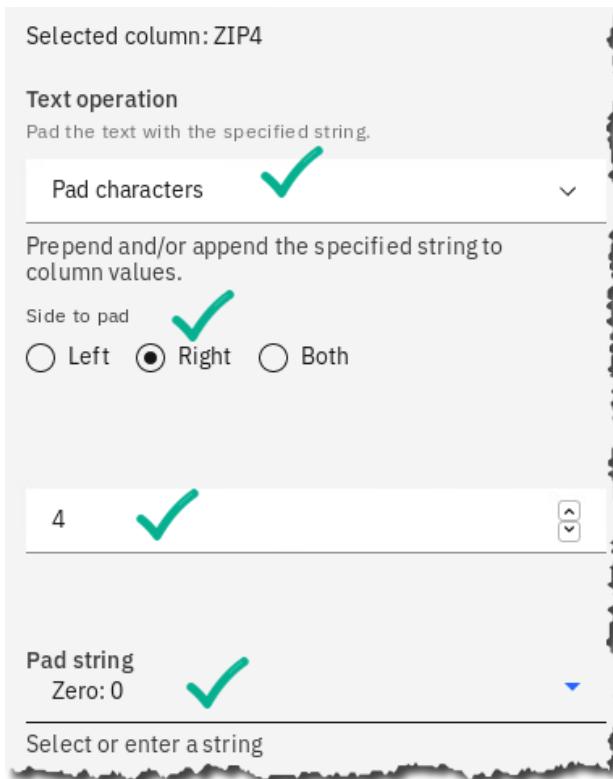
—55. Next, we choose the operation – click the dropdown below [Text Operation](#).

Choose [Pad Characters](#).



—56. This brings up a window where you put in the parameters for padding characters. This function takes 3 parameters – [Side to pad: \(Right\)](#), what the length of the resulting field should be [\(4\)](#), and what [pad string to use: \(0\)](#).

Here are the parameters you should choose:



- 57. After entering these parameters, click the **Apply** button. This will take you back to the data grid, and our field ZIP4 will now be padded if necessary.



- 58. We're now going to do a similar thing to the ZIP field. We want to pad the zip field with a zero, if necessary, to make the field 5 characters. *Don't look at the screenshot below!* Try it on your own first – but if you're struggling, the screenshots below should help.

The screenshot shows the 'Text' operation configuration for the 'ZIP' column:

- Selected column: ZIP**
- Text operation**: Pad the text with the specified string.
  - Pad characters**:
  - Side to pad**:  Left
  - Right
  - Both
- Value**: 5
- Pad string**: Zero: 0
- String**: Select or enter a string

On the right side, the 'Text' operation panel shows the following settings:

- Text**:
- CLEANSE**
- Text**:
- Change column selection**
- Selected column: ZIP**
- Text operation**: Collapse multiple, consecutive spaces in the text to a single space.
  - Collapse spaces**:
  - Number of characters**
  - Pad characters**:
  - Substring**
  - Title case**

A large green arrow points from the 'Text' operation panel down to the 'Pad string' input field. Another green arrow points from the 'Pad string' input field to the 'Apply' button.

- \_\_\_59. When you click **Apply** you're returned to the data grid. Scroll over to ZIP and verify the leading zeros are in place:

BAJADERO	PR	00616 ✓	0
JERSEY CITY	NJ	07305	

- \_\_\_60. What's also great about refinery is that as you add operations it keeps a list of those operations, in easy to understand English, not code – and you can delete a step at any point by clicking the garbage can on that step – the data will revert back to its state prior to the step.

2 Steps

Data Source

CUSTOMER\_DEMOGRAPHICS

Text     

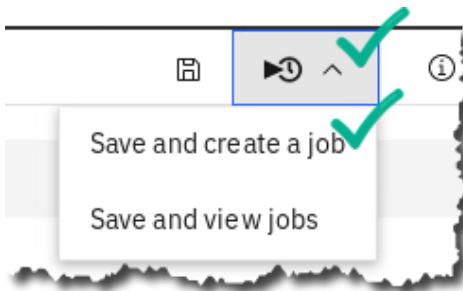
Padded text in ZIP4 with 0 on the right for total length of 4 into ZIP4

Text      JUST ADDED

Padded text in ZIP with 0 on the left for total length of 5 into ZIP

- \_\_\_61. Now that we've built a pipeline with two steps, we want to run it so we can persist the data somewhere. To do that, we save the pipeline as a "job".

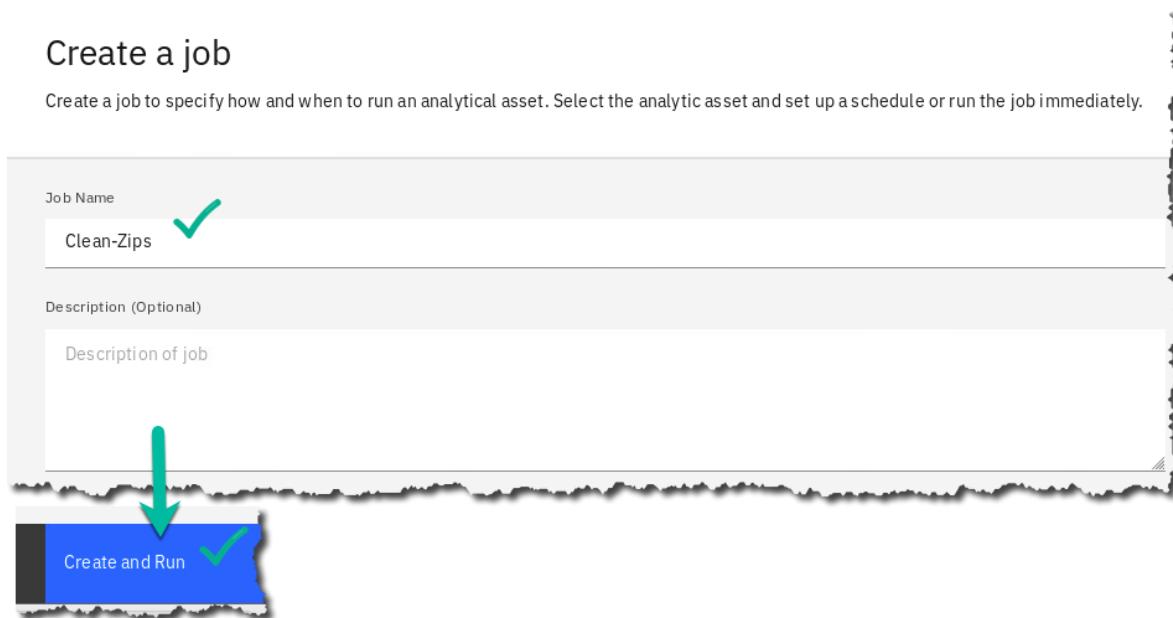
Click the **down arrow** icon next to the clock and play arrow icons, then click **Save and Create a Job**.



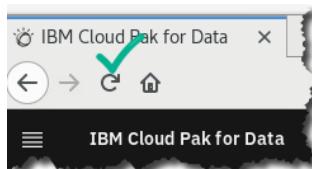
- \_\_62. Next we need to name the job and then run it.  
Enter the name **Clean-Zips**  $\Rightarrow$  **Create and Run**.

### Create a job

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



- \_\_63. The job takes about a minute to complete. You can see it running on the screen.  
Hit the **refresh** icon to see the status more quickly.

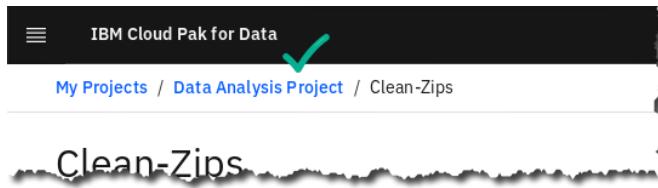


 Data Steward	<p>Note: Due to an issue, the first attempt to run the job may fail; if so, just re-run and it should be successful.</p>
--	--

- \_\_64. When **Completed**, you can check the logs of the job if you wanted to. This can be especially useful if the job is complex. This job is only a simple two step job so we will not review it now.



- \_\_65. Click on the breadcrumb trail to return back to the project.



\_\_66. A new output was created from the Refine job.

Click on it: [CUSTOMER\\_DEMOGRAPHICS\\_shaped.csv](#).

IBM Cloud Pak for Data

My projects / Data Analysis Project

Overview Assets Environments Jobs Access Control Settings

What assets are you looking for?

**Data assets**

0 assets selected.

<input type="checkbox"/> Name	Type
<input type="checkbox"/> CSV CUSTOMER_DEMOGRAPHICS_shaped.csv ✓	Data Asset
<input type="checkbox"/> Ø CUSTOMER_DEMOGRAPHICS	Data Asset
<input type="checkbox"/> Ø CUSTOMER_ACTIVITY	Data Asset

\_\_67. Scroll over to find the ZIP and ZIP4 columns. Notice the data is padded with four zeros when there was previously only one zero in ZIP4 and the ZIP field is left padded with a 0.

	ZIP4 String	LONGITU... String
	0000	-90.285918
	0000	-92.629521
	0000	
	0000	
	0000	-106.34411
	0000	-75.206995
	1057	
	0000	

 Data Steward	<p>The power of <a href="#">Refine</a> can be used by more than just the Data Steward because it is launched from a Project, which enables many more personas the ability to shape and refine data assets that the Data Steward may have given them access to. This gives anyone the ability to explore and shape data on their own with self-service capabilities.</p> <p>If you need to make a permanent change to the data, for example, to write it back to a database from where it came, that is where <a href="#">Transform</a> comes in. We will explore Transform later in this lab.</p>
---	---

## 13.6 Reviewing the business glossary

Business Glossaries have been around in various incarnations for about ten years. They have enjoyed varying degrees of success. The problem with glossaries, and with data governance in general, is that it's too costly and manually-intensive to maintain and enhance these glossaries, and the associated metadata, etc., over a long period of time. Today you'll see how Cloud Pak for Data's Automation capabilities finally make well-maintained, value-added governance solutions a reality.

A **business glossary** consists of **categories** and **terms**.

**Categories** provide the logical structure for the glossary so that you can browse and understand the relationships among terms and categories in the glossary. Categories can be organized in a hierarchy based on their meaning and relationships to one another.

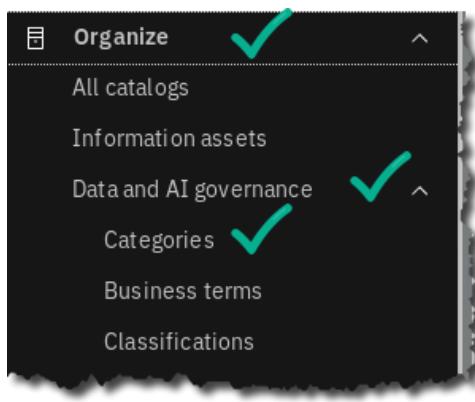
A **Business Term** is a word or phrase that describes a characteristic of the enterprise. Terms are the fundamental building blocks of the glossary. Each Business Term has a parent Category, but it can also be referenced by other Categories. When you create a Business Term, you need to provide a meaningful name. Business Terms can be assigned to other Business Terms, and to other asset types as well.

**Business Terms** can be associated with a variety of different data assets in your organization, from tables and columns and files and fields to policies, ETL processes, data models, etc. This makes it easy for business users to find what they are looking for using terminology they understand.

Let's explore some of the categories and terms that are pre-loaded into our environment.

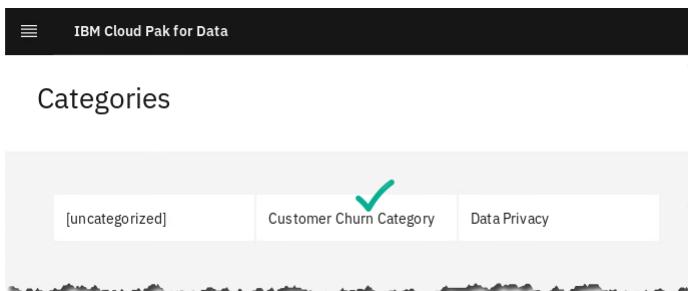
- 68. Start at the [Navigation Menu](#).

Click [Organize](#)  $\Rightarrow$  [Data and AI governance](#)  $\Rightarrow$  [Categories](#).



\_\_69. Review a category already created for you to review.

Click [Customer Churn Category](#).

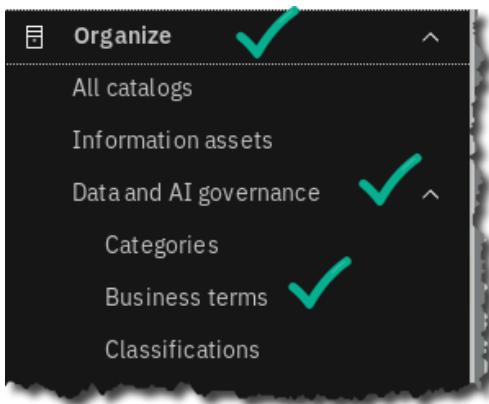


\_\_70. Review the Business Terms and Policy for the Category.

Governance artifacts		
Name	Description	Type
Days Since Last Trade	Number of days since the customer executed a trade on our platform	Business term
Employee ID	A number that identifies our employees for HR purposes.	Business term
Gender	Data indicating gender, if specified	Business term
Home Owner	Flag indicating whether a customer owns a home	Business term
Income	Annual household income of the customer	Business term
Marital Status	Status indicating whether a customer is married, single, etc.	Business term
Net Gains and Net Losses are Mutually Exclusive	A customer can only have a value in Net Gains or Net Losses	Policy

Notice that on this page you could create a new Category, but do not do so. We will only be working with this one category.

\_\_71. Click [Navigation Menu](#) ⇒ [Organize](#) ⇒ [Data and AI governance](#) ⇒ [Business terms](#).



\_\_72. Here you can edit or add new Business terms that are either in Published or Draft mode.

Click on one to review it in more detail.

The screenshot shows the 'Business terms' page in the IBM Cloud Pak for Data interface. The 'Published' tab is selected, indicated by a blue underline. There are two entries listed:

- Days Since Last Trade** ✓  
Number of days since the customer executed a trade on our platform  
[Customer Churn Category](#)  
Last modified: Jun 22, 2020
- Gender**  
Customer Churn Category  
[Customer Churn Category](#)



Data Steward

You can create your own Glossary with Categories and Business Terms manually or import them from a file. In addition, you can import Industry Models from IBM for industries such as finance, banking, healthcare, and insurance, and import them into CPD.

See the services screen then [Industry solutions](#).

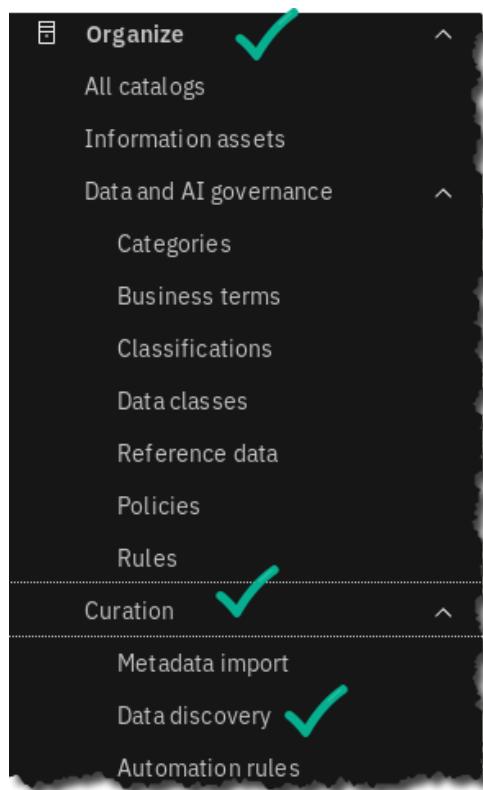
## 13.7 Automation Capabilities – Discovery, Classification, Term Assignment

As mentioned in the business glossary section, the key to successful data governance lies in the use of automation and machine learning techniques to make decisions, create relationships, and generally enhance the value, timeliness and accuracy of your data governance activities. What you'll be doing next is known as Auto Discovery. This is a process that will crawl the data you point it at, and do several things:

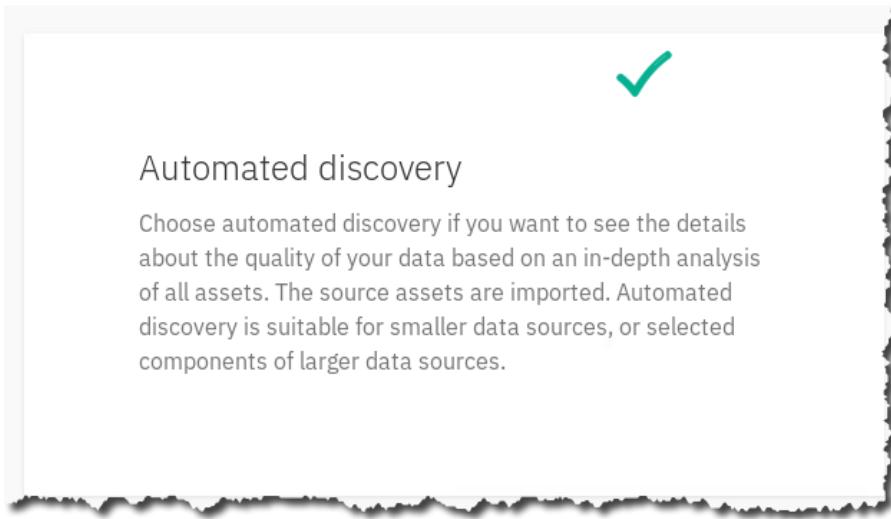
- \_\_a. *Classify the data* – this is the process that will assign business classifications to your data elements.
- \_\_b. *Score the data* – this will allow you to quickly understand how clean your data is.
- \_\_c. *Assign Terms to your data* – this is where CPD will look at the data you're pointing to, take the business glossary terms, and based on both the data in the columns and the metadata and classifications already determined, assign a confidence rating to specific terms that it matches to the columns. You can set a threshold for when it automatically makes the connection between terms and columns, or you can use a workflow-driven review process to inspect the confidences and determine whether to assign or not.

\_\_73. Let's get started. Go to the [Navigation Menu](#).

Click [Organize](#) ⇒ [Curation](#) ⇒ [Data discovery](#).



- 74. This will bring up a screen showing any past results and allow you to create a new discovery job. Click [Automated discovery](#).



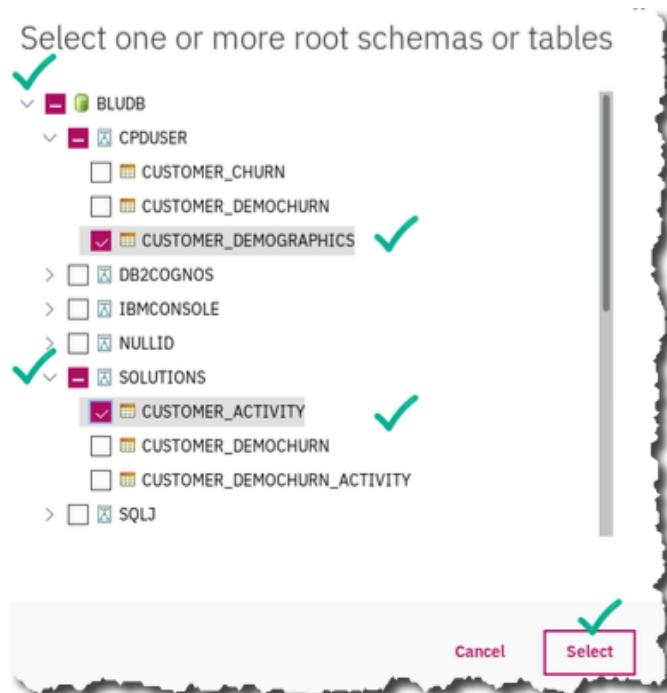
- 75. Next, we'll choose the connection that gets us to our data, and the options we want for this discovery run:

For **Connection** choose the connection we created earlier, i.e., [Db2 Advanced Edition](#), then click [Browse](#).

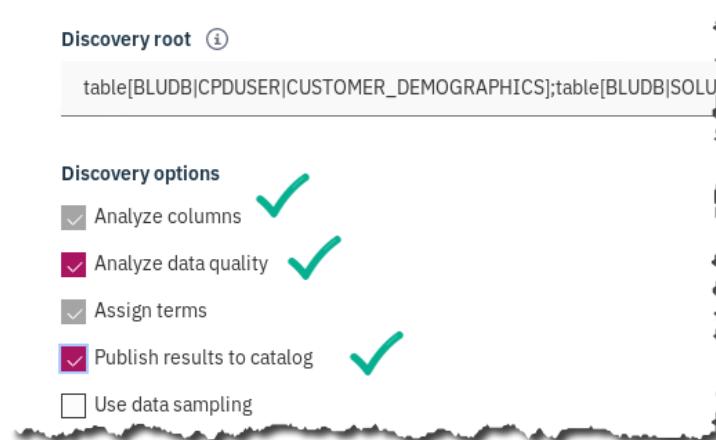
A screenshot of a web page titled "Automated discovery job". The page has a dark header bar with the text "IBM Cloud Pak for Data" and "All". The main content area includes a "Connection \*" dropdown menu where "Db2 Advanced Edition" is selected, indicated by a green checkmark. Below it is a "Discovery root" input field containing the placeholder text "Example: schema[db\_name/schema\_name];table[db\_name/schema\_name/table\_name]" and a "Browse" button, which is also highlighted with a green checkmark. Under "Discovery options", there are three checkboxes: "Analyze columns", "Analyze data quality", and "Assign terms", all of which are empty. The entire screenshot is framed by a thick black border.

- \_\_76. This will bring up a screen where we can choose which tables, schemas, or an entire database for which we want to run Discovery. We're going to choose two tables that will be our main tables for later transformation.

Click to expand BLUDB then CPDUSER  $\Rightarrow$  CUSTOMER\_DEMOGRAPHICS, and SOLUTIONS  $\Rightarrow$  CUSTOMER\_ACTIVITY. Then Select .



- \_\_77. Next, click the checkboxes for Analyze Columns, Analyze Data Quality and Publish to Catalog. (Assign terms will automatically be checked.)



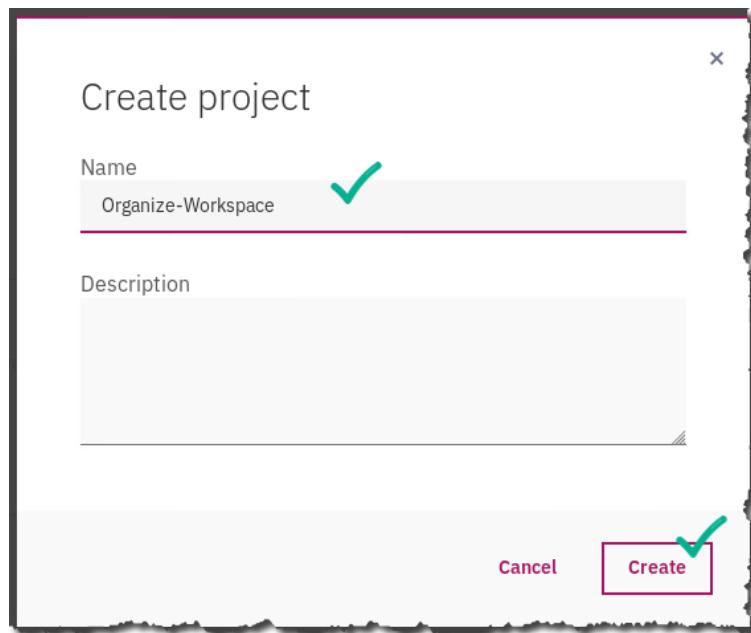
- 78. Lastly, you need to choose where to place the results for further analysis or inspection – these are called *Projects*.

Scroll down to find the **Select a project** dropdown.

Choose **Add a project**.

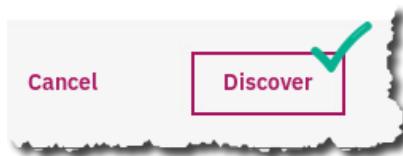


- 79. This will bring up a screen where you can name the Project **Organize-Workspace** and create it.



\_\_80. Once the project is created you'll be taken back to the Discovery screen.

Now click **Discover**.



\_\_81. This will bring up a screen where the status of the discovery job is shown; we see the **Import** portion is running; this imports the metadata from the tables into our workspace.

\_\_82. Click the **refresh** button on the right to update the status.



Number of tables 0	
Tables	Status
	Phase Import <b>Running</b>
	Start June 30, 2020, 1:00 PM

The process will take a minute or two to complete – when it's complete you'll see the blue **Running** turn to green **Finished**.

Discovered assets							
Number of schemas 2		Number of tables 2					
Asset name	Asset type	Tables	Status	Phase	Import	Analyze	Phase
CUSTOMER_DEMOGRAPHICS	Table	1	Running	Phase	Import <b>Finished</b>	Analyze <b>Running</b>	Phase
				Start	June 30, 2020, 1:00 PM	Start	June 30, 2020, 1:00 PM
				End	June 30, 2020, 1:00 PM	Done	0% Successful 0% Cancelled 0% Failed 0%
CUSTOMER_ACTIVITY	Table	1	Running	Phase	Import <b>Finished</b>	Analyze <b>Running</b>	Phase
				Start	June 30, 2020, 1:00 PM	Start	June 30, 2020, 1:00 PM
				End	June 30, 2020, 1:00 PM	Done	0% Successful 0% Cancelled 0% Failed 0%

\_\_83. We can now take a look at what the discovery process found.

When finished, click the eye icon next to the **CUSTOMER\_DEMOGRAPHICS** table entry.

Discovered assets							
Number of schemas 2		Number of tables 2					
Asset name	Asset type	Tables	Status	Actions			
CUSTOMER_DEMOGRAPHICS	Table	1	Running	Phase	Import <b>Finished</b>	Analyze <b>Running</b>	Actions
				Start	June 30, 2020, 1:00 PM	Start	June 30, 2020, 1:00 PM
				End	June 30, 2020, 1:00 PM	Done	0% Successful 0% Cancelled 0% Failed 0%
CUSTOMER_ACTIVITY	Table	1	Running	Phase	Import <b>Finished</b>	Analyze <b>Running</b>	Actions
				Start	June 30, 2020, 1:00 PM	Start	June 30, 2020, 1:00 PM
				End	June 30, 2020, 1:00 PM	Done	0% Successful 0% Cancelled 0% Failed 0%

—84. This takes us into the view of our demographics table.

Click the > next to the table name to expand the columns.

<input type="checkbox"/> Name	Quality score
<input type="checkbox"/> CUSTOMER_DEMOGRAPHICS	98%

—85. You will then see that the discovery process has automatically determined data classes for many of the columns, as well as assigned Business Terms if the confidence of the term being a match is > 80%. This is incredibly powerful – in the past this sort of work was done completely manually, and often the results were error-ridden. Research suggests this automation eliminates 85%-90% of the time it takes to do this manually.

Discovered data assets					
<input type="checkbox"/>	Name	Quality score	Data class	Assigned terms	Last analyzed
<input type="checkbox"/>	CUSTOMER_DEMOGRAPHICS	98%	—	—	Jun 30, 2023
	ADDRESS_1	98%	Text 100% ✓	—	Jun 30, 2023
	ADDRESS_2	100%	NoClassDetected 100% ✓	—	Jun 30, 2023
	AGE	99%	Code 100% ✓	—	Jun 30, 2023
	CHILDREN	100%	Code 100% ✓	—	Jun 30, 2023
	CITY	95%	City 95% ✓	—	Jun 30, 2023
	CREDITCARD	100%	Identifier 100% ✓	—	Jun 30, 2023
	DOB	100%	Date of Birth 100% ✓	—	Jun 30, 2023
	ESTINCOME	100%	NoClassDetected 100% ✓	—	Jun 30, 2023
	GENDER	100%	Gender 100% ✓	Gender 100% ✗ ✓	Jun 30, 2023
	HOMEOWNER	100%	Indicator 100% ✓	Home Owner 100% ✗ ✓	Jun 30, 2023
	ID	100%	Credit Card Validation ... 100% ✓	—	Jun 30, 2023
	LATITUDE	100%	Latitude 100% ✓	—	Jun 30, 2023
	LONGITUDE	100%	Longitude 100% ✓	—	Jun 30, 2023

Note: Your Quality score may show numbers slightly different than above.

- \_\_86. So, what can we do with this information? Since we decided to automatically publish it to the catalog, we can look at some of the assets and see what we can find. First, let's check the information that's been collected in the workspace from the discovery.

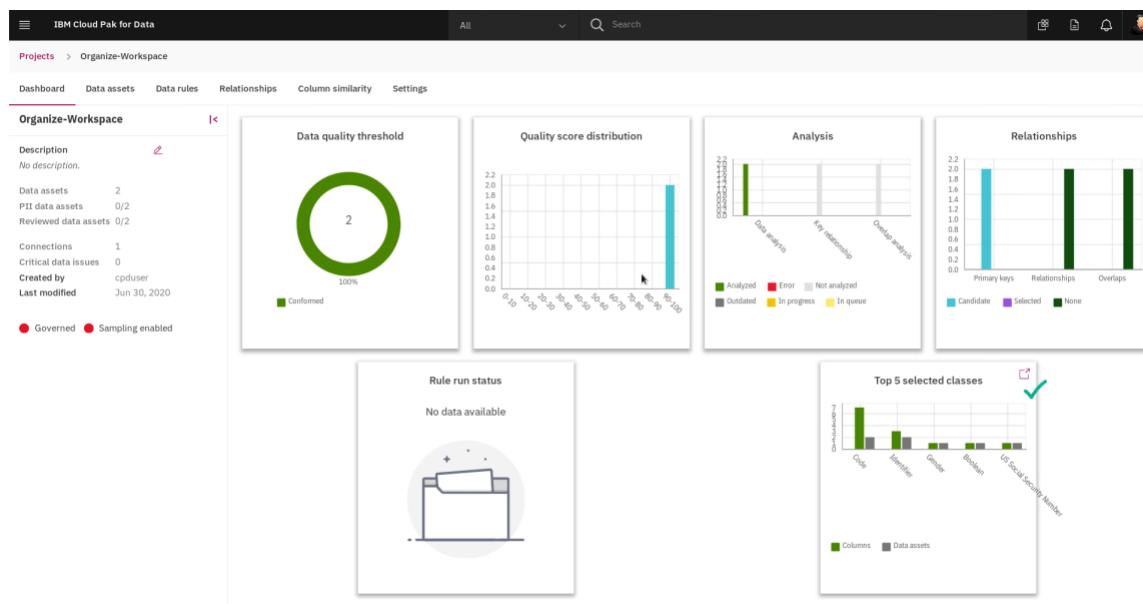
Click the [back arrow](#) to return to our main discovery listing.

The screenshot shows the 'Discover' section of the IBM Cloud Pak for Data interface. At the top, there is a navigation bar with a back arrow icon and the text 'IBM Cloud Pak for Data'. Below the navigation bar, the project name 'CUSTOMER\_DEMOGRAPHICS' is displayed. The main area is titled 'Discover' and contains several sections: 'General information', 'Discover options', and 'Source asset import'. A green checkmark is placed next to the 'Project' field, which is set to 'Organize-Workspace'.

- \_\_87. Next, click the link to jump into the [Organize-Workspace](#) project.

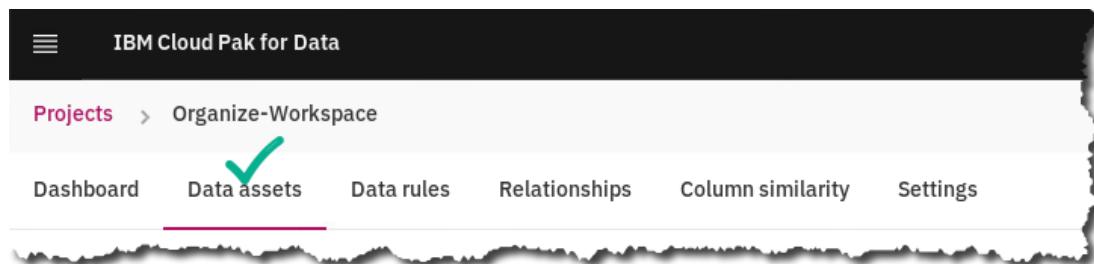
The screenshot shows the 'Organize-Workspace' project dashboard. It includes sections for 'General information' (Start: July 24, 2020, 12:31 PM; Assets included in the discovery: 10), 'Discover options' (Project: Organize-Workspace, checked), 'Discovery options used' (Column analysis, Term assignment, Data quality analysis, Publish), and 'Source asset import' (All assets). A green checkmark is placed next to the 'Project' field.

- \_\_88. We're now shown a dashboard with some highlighted information about our discovery. If you scroll down, you'll see the top data classes selected from the data that we ran discovery against, and also the quality score distribution across our data sets. The distribution shows that the data sets are pretty clean, scoring 90-100% on quality scores.



\_\_89. Next, let's look at some of the other characteristics of our data elements that have been discovered.

Click the link that says **Data assets**.



\_\_90. Next, click on the **CUSTOMER\_DEMOGRAPHICS** tile.

A screenshot of the 'Data assets' page. At the top, there is a search bar and filter options ('Sort by: Default', 'Filter by: All data assets'). Below this, two data asset tiles are displayed. The first tile is for 'CUSTOMER\_ACTIVITY' and the second is for 'CUSTOMER\_DEMOGRAPHICS'. The 'CUSTOMER\_DEMOGRAPHICS' tile has a green checkmark icon to its right. Both tiles show the following details:

- Progress: 99% (for CUSTOMER\_ACTIVITY) and 97% (for CUSTOMER\_DEMOGRAPHICS)
- Type: B.SOLUTIONS (for CUSTOMER\_ACTIVITY) and B.CPDUSER (for CUSTOMER\_DEMOGRAPHICS)
- Database connection: jdbc:db2://worker5.clusterw9:32030/BLUDB: BLUD
- Last imported: Jun 30, 2020, 13:00
- Last data analysis: Jun 30, 2020, 13:02
- Last published: Jun 30, 2020, 13:02
- Threshold: 80%

- 91. We are now looking at all the columns in the table, and different information about each column; Data Class, Business Term assigned, even the prevailing format and whether the field is unique. You can click on a column name to get even more detailed information.

The screenshot shows the IBM Cloud Pak for Data interface. On the left, there's a sidebar with navigation links: Columns, Governance, Data quality, Data classes, Data types, Rules, Primary keys, and Foreign keys. Below these are sections for 'CUSTOMER\_DEMOGRAPHICS' (Description, Data quality score, Columns, Rows, Reviewed, Threshold, Analysis status, Last analysis) and a 'Find a column' search bar. The main area is titled 'Columns (18)' and lists 18 columns with their details: ID (Credit Card Validation Number, Format 9999), GENDER (Gender, Format Gender 100%, A), STATUS (Code, Format A), CHILDREN (Code, Format 9), ESTINCOME (Income, Format 99999.99), HOMEOWNER (Indicator, Format Home Owner 100%, A), AGE (Code, Format 99), TAXID (US Social Security Number, Format 9999999999999999), CREDITCARD (Identifier, Format 9999999999999999), DOB (Date of Birth, Format NATIVE/DEFAULT), and ADDRESS\_1 (Text, Format NA).

 Data Steward	<p>A <a href="#">Data Dictionary</a> contains a Business Glossary (Categories and Business Terms) as well as Information Governance Policies and Rules to ensure data compliance with business objectives, as well as data security.</p> <p>In this lab we have a beginning sample of these items, but in reality, a Data Dictionary for any organization is quite large and can and should be updated frequently as new data sources, regulations, and other criteria require it.</p>
---	--

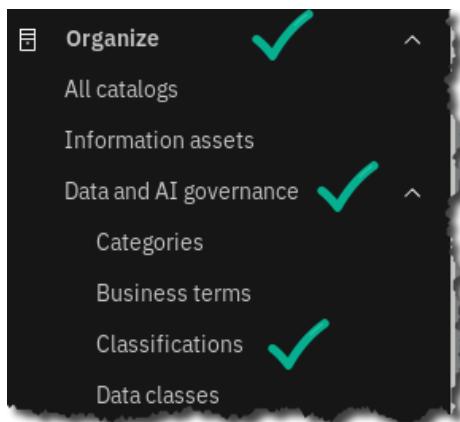
## 13.8 Reviewing Classifications, Data Classes, and Reference Data

### 13.8.1 Classifications

A [classification](#) describes the sensitivity level of data. In catalogs, a classification describes the sensitivity of a whole data asset to help catalog members understand the asset. You can use classifications to describe Business Terms, Data Classes, Reference Data Sets, and Governance Rules. In data protection Rules, you can include Classifications in conditions to identify the type of data to restrict.

\_\_92. Start at the [Navigation Menu](#).

Click [Organize](#) ⇒ [Data and AI governance](#) ⇒ [Classifications](#).



\_\_93. Scroll to find Classification [Confidential](#) and click on it.

A screenshot of the 'Classifications' page in the IBM Cloud Pak for Data interface. The title bar says 'IBM Cloud Pak for Data'. The main area shows a table with one row. The row contains the classification name 'Confidential' with a checkmark, a detailed description of what confidential data is, a category link to '[uncategorized]', and a note that it was last modified on Jun 18, 2020. The table has columns for 'Published' (which is selected) and 'Draft'. There are also filters for 'Find classifications', 'Sort by: Name', 'Show: All', and an 'Edit' button.

- \_\_94. The Classification is described here. You could also add the primary Category here, but there is no need to do so now.

The screenshot shows the 'Classifications' page in the IBM Cloud Pak for Data interface. A classification named 'Confidential' is selected and marked as 'Published'. The 'Description' field is expanded and highlighted with a red box, containing the following text: 'Confidential data is data that if compromised in some form, is likely to result in significant and/or long-term harm to the institution and/or individuals whose data it is.' Below this, there is a link 'Access to confidential information is...'. The 'Primary category' section is also visible.

### 13.8.2 Data classes

**Data classes** describe the contents of data in a column of a relational or structured data set. Data classes are assigned to columns when profiling a structured data asset and shown on the **Profile** page in a Catalog or Project.

Watson Knowledge Catalog provides a predefined set of Data Classes. Some Data Classes are categorized into groups, for example:

- If you select **Date**, it also includes **Date of Birth**.
- If you select **Driver's License**, it also includes all dependent driver licenses listed below.

- \_\_95. Start at the [Navigation Menu](#).

Click [Organize](#)  $\Rightarrow$  [Data and AI governance](#)  $\Rightarrow$  [Data classes](#).

The screenshot shows the navigation menu of the Watson Knowledge Catalog. The path 'Organize > Data and AI governance > Data classes' is highlighted with green checkmarks, indicating the steps to reach the Data classes section.

\_\_96. Scroll to find **Account Number** and click on it.

The screenshot shows the 'Data classes' list in the IBM Cloud Pak for Data interface. The 'Account Number' data class is selected, highlighted with a green checkmark icon. The description 'A value representing an Account Number.' is visible, along with examples '123456' and the primary category '[uncategorized]'. Other data classes like 'Address Line 3' are also listed.

\_\_97. The Data Class is described here. You could also add the primary Category here.

The screenshot shows the detailed view of the 'Account Number' data class. The 'Description' field is highlighted with a red box and contains the text 'A value representing an Account Number.'. The 'Examples' field shows '123456'. The 'Primary category' section shows '[uncategorized]' with a green checkmark icon. The 'Secondary categories' section indicates 'No secondary categories added yet.' The 'Data matching' section has a toggle switch set to 'On'.

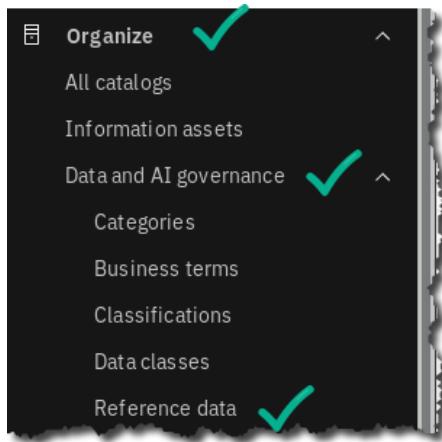
### 13.8.3 Reference Data Set

Reference Data Sets define lists of permissible values that are allowed for use within a data field and may be referenced by Business Terms, Policies, Rules and Data Classes in Watson Knowledge Catalog.

You can capture, manage, and socialize reference data — setting it up once and re-using the reference data in other places.

- \_\_98. Start at the [Navigation Menu](#).

Click [Organize](#) ⇒ [Data and AI governance](#) ⇒ [Reference data](#).



- \_\_99. Scroll to find Reference data set [State and Province Codes](#) and click on it.

\_\_100. Scroll down to review the Reference data set rows (the data).

	Code	Value	Description
AA	Armed Forces (the) Americas	Click to add description	
AB	Alberta	Click to add description	
AE	Armed Forces Europe	Click to add description	
AK	Alaska	Click to add description	
AL	Alabama	Click to add description	

\_\_101. Click section **Related content** and review how you can relate Rules or Data classes to it.

This reference data doesn't have any related content yet.  
Types of artifacts that can be related to reference data include:

Data classes
 Rules

### 13.8.4 Data Protection & Masking

Another challenging requirement in Data Governance is the ability to ensure that only the correct users or roles see the correct data. Often, this can get very granular, such that a particular role needs to view a set of columns in a table, but other columns in the same table are off limits. Like with Business Glossaries, the challenge is in how to ensure data protection when there is an exploding volume of sources – enter automation.

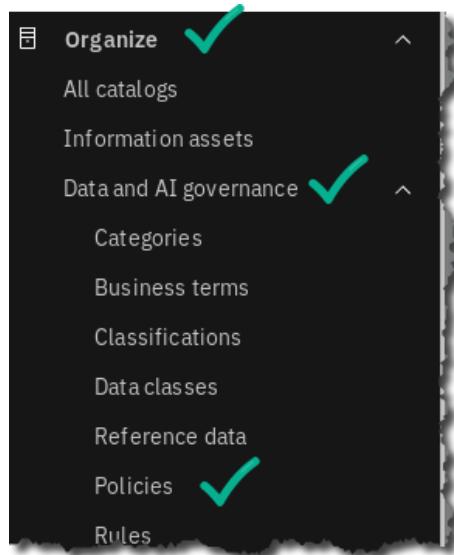
Data Protection Rules are a unique and incredibly powerful capability in Cloud Pak for Data. They allow you to set up masking or redaction capabilities, on a columnar or tabular basis once, and any data that subsequently gets added to the catalog, if it has the same characteristics as the rule states, will automatically be masked or redacted. This is another benefit of the automated data discovery and classification process.

It works like this:

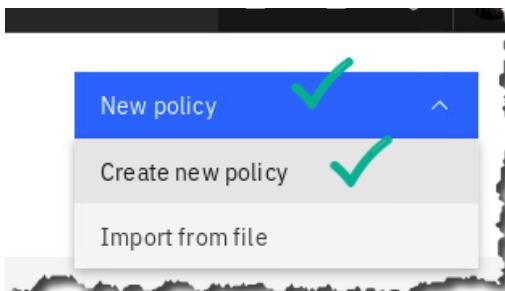
- You build a data protection rule stating that any column in the data catalog that is classified as ‘Date of Birth’ must be redacted from view when the user is a developer.
- You run data discovery on a new data source that you want to integrate into your catalog, data lake, etc.
- Data discovery automatically classifies any field it determines through ML and fuzzy logic to be DOB information with the data class “Date of Birth”.
- You associate this data protection rule with a more general policy stating that certain data elements must be redacted.
- Immediately this data protection rule will be executed any time a developer tries to view a Date of Birth column.

Let's see how this works in practice. The first thing we will do is create a policy around masking PII data.

\_\_102. Click [Navigation menu](#)  $\Rightarrow$  [Organize](#)  $\Rightarrow$  [Data and AI Governance](#)  $\Rightarrow$  [Policies](#).



\_\_103. We'll create a new policy, so click New policy  $\Rightarrow$  Create new policy.



\_\_104. Now fill in the following information (Policy name = Mask PII Data) and click Save as draft.

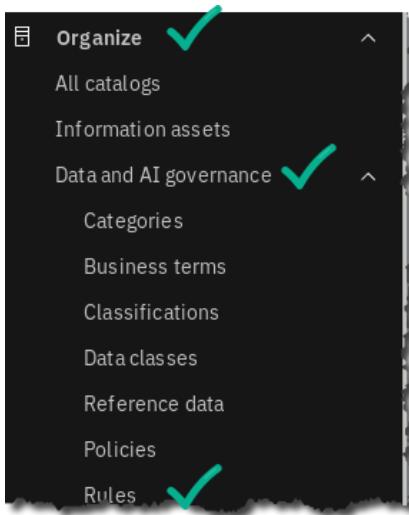
The screenshot shows a "Create new Policy" dialog box. The "Policy name" field contains "Mask PII Data" with a green checkmark. The "Primary category" section shows "[uncategorized]" with a "Change" button. The "Description (optional)" section has a placeholder "Type description here". At the bottom, there are "Cancel" and "Save as draft" buttons, with "Save as draft" being highlighted by a green checkmark.

We save this as a draft because Cloud Pak For Data has a built-in workflow capability that ensures that all governance artifacts, rules, policies, etc. need to be reviewed before being published.

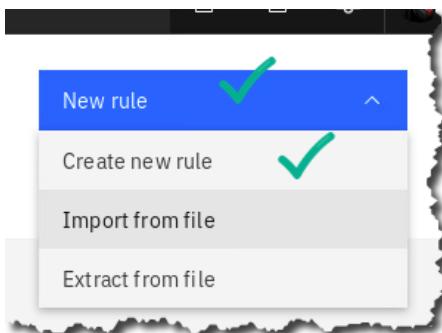
\_\_105. Publish the policy – click Publish – Publish.



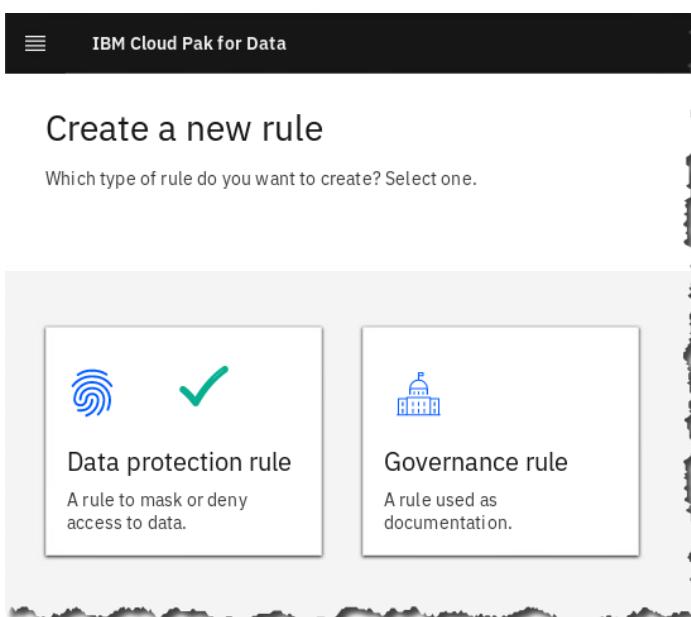
\_\_106. Next, we'll create our data protection rule, and then link it to our policy.  
Click [Navigation menu](#) ⇒ [Organize](#) ⇒ [Data and AI Governance](#) ⇒ [Rules](#).



\_\_107. Click [Create new rule](#).



\_\_108. Choose [Data protection rule](#).



109. Now we'll enter the criteria for our rule.

Fill out the form as shown here:

Name: Mask DOB for Developer user

Type: Access

Business definition: A rule to ensure that PII masks date of birth.

If Data class contains any 'Date of Birth'

and (+)

User name contains any Developer

Action

then mask data in columns containing Date of Birth

Redact

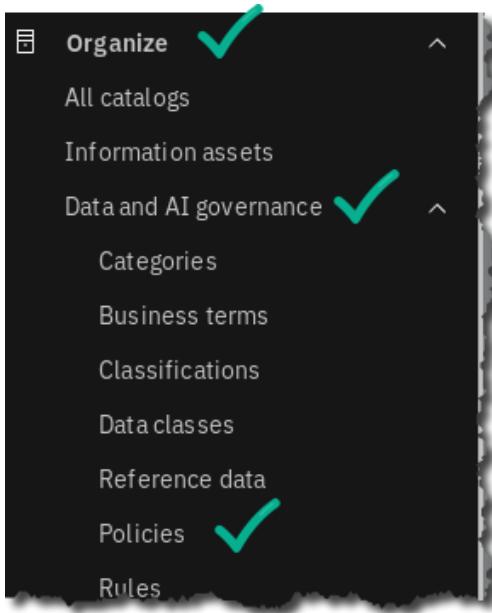
and click Create rule.

The screenshot shows the 'New data protection rule' interface in IBM Cloud Pak for Data. It consists of three main panels:

- Rule Details:** Contains fields for Name (Mask DOB for Developers), Type (Access), and Business definition (A rule to ensure that PII masks date of birth).
- Conditions:** Shows the logical expression "If Data class contains any 'Date of Birth' AND User name contains any 'Developer'".
- Action:** Shows the selection "then mask data in columns containing Date of Birth" and the choice of "Redact" (Before 452-821-1120) and "Replace data with Xs".

\_\_110. Next, we want to associate this rule to the Policy we created earlier. For data protection rules to be enforced, they need to be associated with one or more policies; this ensures that there is a business-documented reason (the policy) for having the rule.

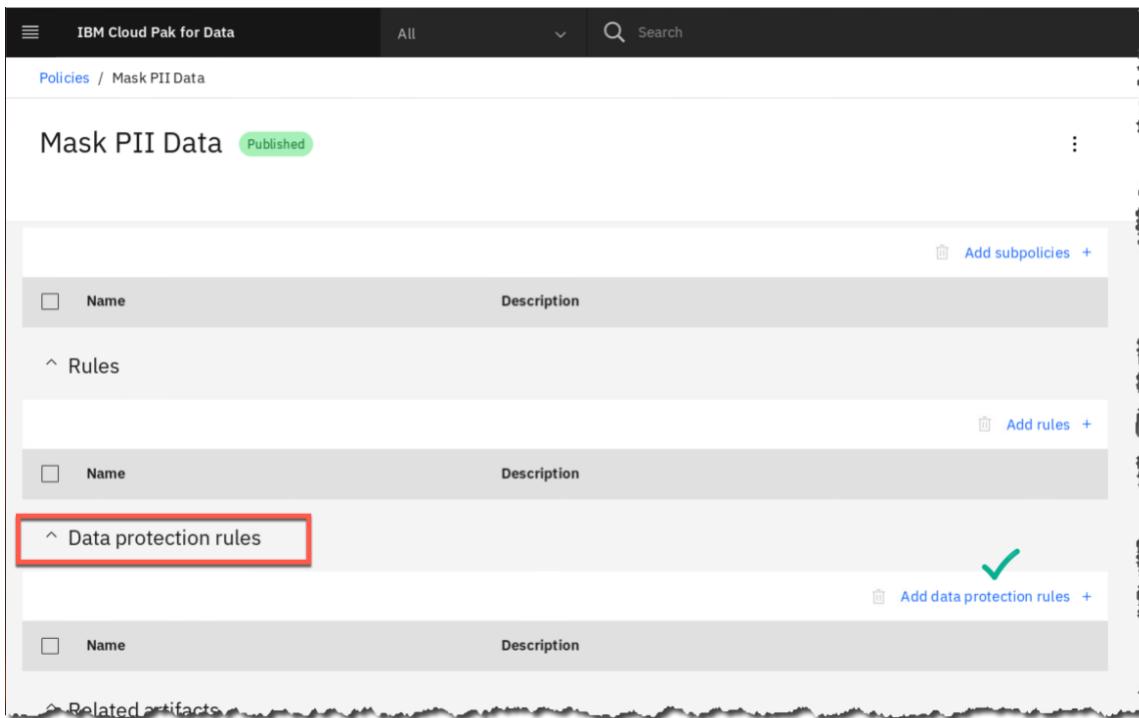
Click **Navigation menu** ⇒ **Organize** ⇒ **Data and AI Governance** ⇒ **Policies**.



\_\_111. Click to open the policy called **Mask PII Data**.

A screenshot of the 'Policies' page in IBM Cloud Pak for Data. The 'Published' tab is selected. A search bar shows 'Find policies'. The 'Sort by:' dropdown is set to 'Name'. Two policies are listed: 'Data Privacy' and 'Mask PII Data'. 'Data Privacy' is described as a 'Company-wide data privacy policy for securing private data'. 'Mask PII Data' is described as 'Mask PII Data'. Both policies have a blue folder icon and were last modified on Jun 30, 2020. A green checkmark is placed next to 'Mask PII Data'.

- \_\_112. Next, scroll down until you see the section of the page called **Data Protection Rules** and click **Add Data Protection Rule**.

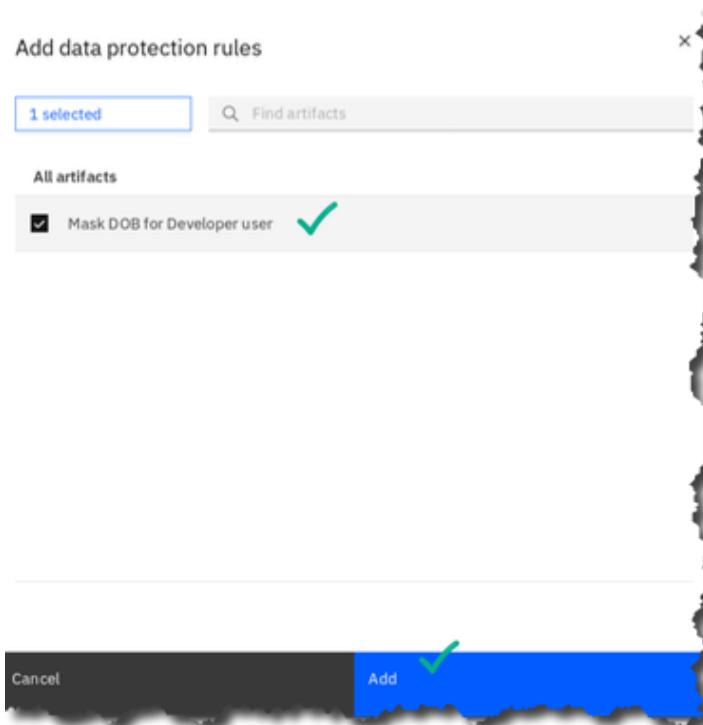


The screenshot shows the 'Mask PII Data' policy page in the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and a 'Published' status indicator. Below the navigation, the policy name 'Mask PII Data' is displayed. Under the policy name, there are sections for 'Rules' and 'Data protection rules'. The 'Data protection rules' section is highlighted with a red box. A green checkmark icon is visible in the top right corner of the 'Add data protection rules' button.

- \_\_113. Click the checkbox next to our Data Protection Rule.

Then click **Add**.

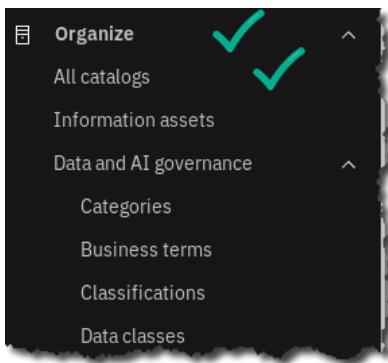
Note: There's no need to hit save after adding the Rule; it's automatically saved, and you will get a small indicator of that in the upper right of the screen.



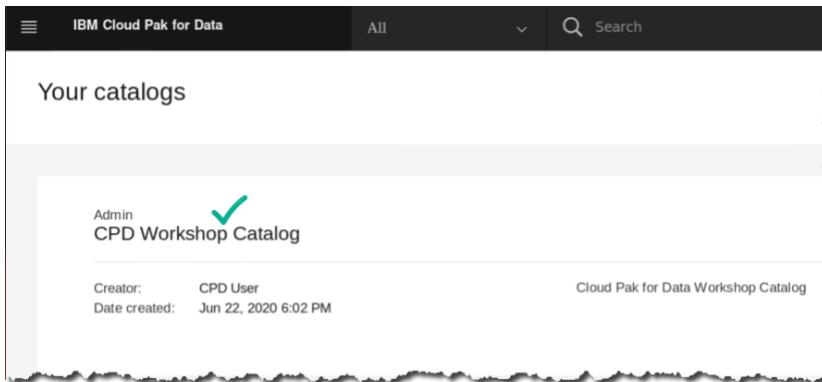
The screenshot shows the 'Add data protection rules' dialog box. It has a header 'Add data protection rules' and a sub-header '1 selected'. Below this, there's a search bar 'Find artifacts' and a list titled 'All artifacts'. A single item, 'Mask DOB for Developer user', is listed with a checked checkbox and a green checkmark. At the bottom of the dialog, there are 'Cancel' and 'Add' buttons. The 'Add' button is highlighted with a blue background and a green checkmark icon.

\_\_114. The next thing we'll want to do is add a collaborator to our catalog – catalogs can have collaborators with various roles; we will add our 'developer' user, and then we'll log in as that user to see the masking in action.

Go to [Navigation Menu](#)  $\Rightarrow$  [Organize](#)  $\Rightarrow$  [All catalogs](#).



\_\_115. Click on [CPD Workshop Catalog](#).

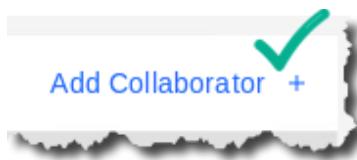


 Data Steward	Note – This is the catalog where our data assets were automatically published after running Auto Discovery.
--	---

\_\_116. This will bring you into the screen that shows the assets in the catalog, and tabs for different types of information. Click the tab **Access Control**.

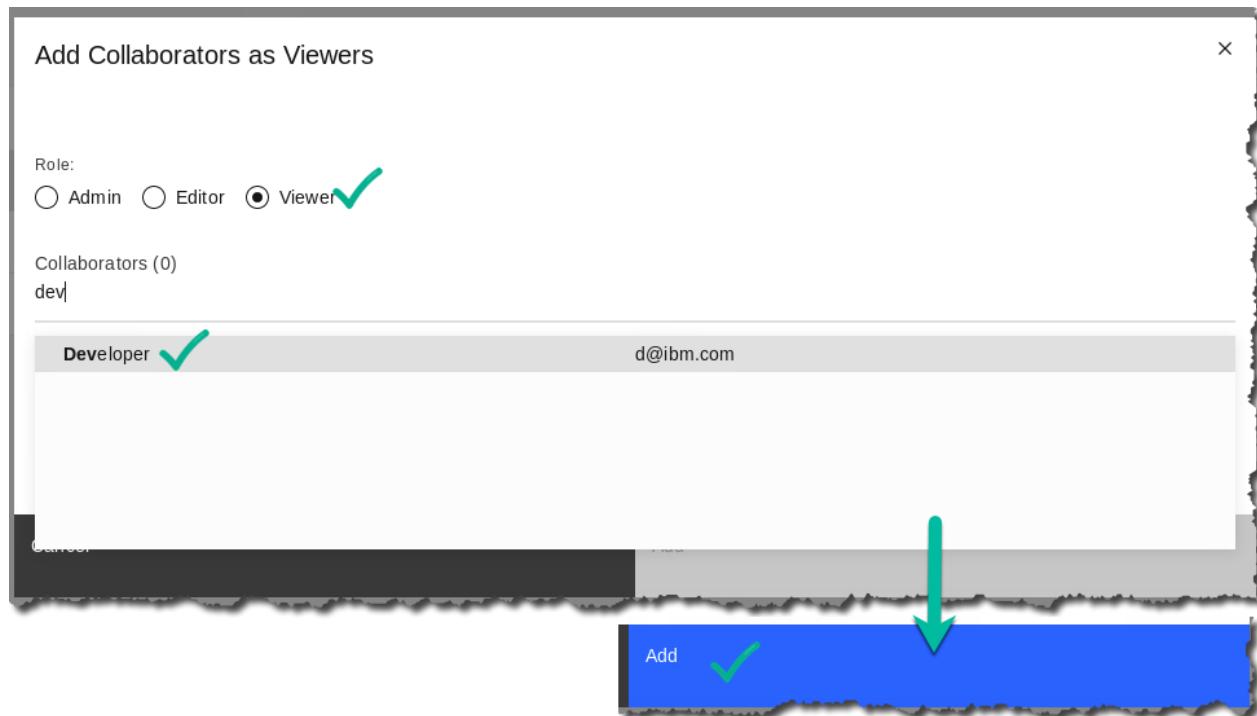
The screenshot shows the IBM Cloud Pak for Data interface with the "Catalogs / CPD Workshop Catalog" path. The "Access Control" tab is highlighted with a green checkmark. A search bar and filter options for type, source, and tag are visible. Below, three asset cards are displayed under "Watson Recommends": "Db2 Advanced Edition" (Connection), "Customer Demographics" (Data asset), and "Customer Activity" (Data asset). Each card shows owner information (CPD User), add date (Jun 22, 2020), and review counts (0 reviews for Db2 and Customer Demographics, 1 review for Customer Activity).

\_\_117. Next, click the link on the right side of the screen that says **Add Collaborator**.



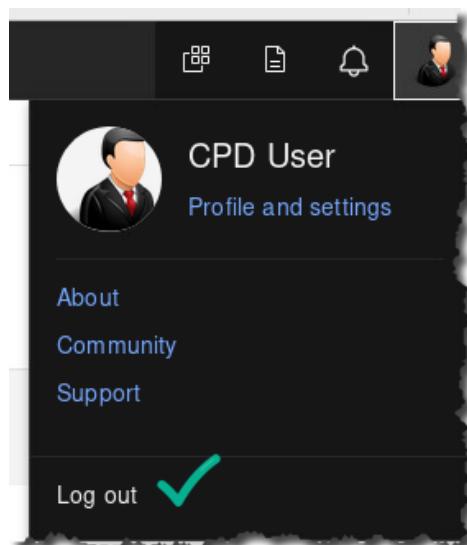
\_\_118. Next, start typing the word **Developer** and you will get a drop down of all the users in the system with that in their name.

Choose **Developer** from the list  $\Rightarrow$  click **Add**. This will add the Developer as a Viewer (default).

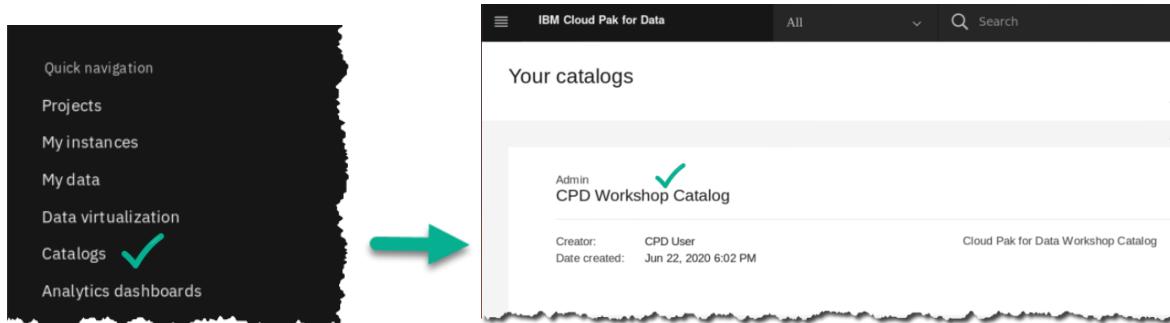


\_\_119. We've now added the user as a collaborator to this project. Let's log in as that user.

First log out CPD User by clicking the **User Icon** in the upper right corner  $\Rightarrow$  **Log out**.



- \_\_120. Then log back in as [developer](#) (password should be saved in the browser, but it is [cpdaccess](#)).
- \_\_121. When the welcome window comes up, click [Explore Catalogs](#), and on the next screen open the catalog [CPD Workshop Catalog](#).



- \_\_122. Next, we'll choose the table [CPDUSER.CUSTOMER\\_DEMOGRAPHICS](#).

- \_\_123. When the Table preview opens, you will see that the [DOB](#) field is all 'X's and there is a padlock icon over the column.

CREDITC...	DOB
Identifier	Date ...
29267426548528	XXXXXXX
55536189125665	XXXXXXX
89793582342548	XXXXXXX
81112009117829	XXXXXXX
88255529919040	XXXXXXX
12815532810902	XXXXXXX
90714118732757	XXXXXXX

- \_\_124. Now log out and then log back in as user [cpduser](#) with password of [cpdaccess](#).

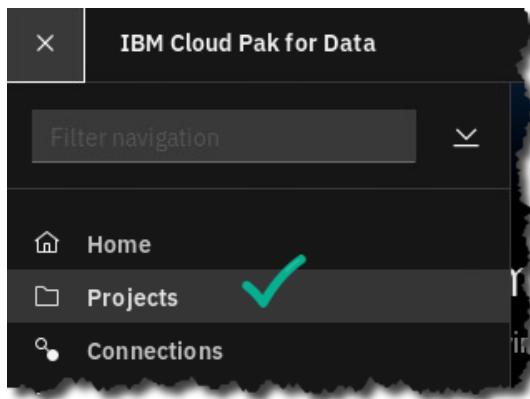
## 13.9 Transforming Data

One of the basic capabilities necessary in a good information architecture is the ability to implement sophisticated transformation and data manipulation. This capability is traditionally known as Extract/Transform/Load, or ETL. In Cloud Pak for Data, because of the strong governance capabilities, the ETL process automatically creates a documented data flow that enables data lineage. Data lineage is the ability to see exactly what happens to data as it moves through your organization's systems. This information is critical in today's world of data privacy regulation, where the ability to document your organization's data processing steps is a requirement for compliance.

### 13.9.1 Reviewing an existing job

- \_125. ETL processes, or 'jobs', are developed in a special type of project called a [Transform Project](#). So, let's open our transform project and get started.

Click [Navigation Menu](#) ⇒ [Projects](#).



- \_126. Open the project [CPD\\_Workshop\\_Transform\\_Project](#).

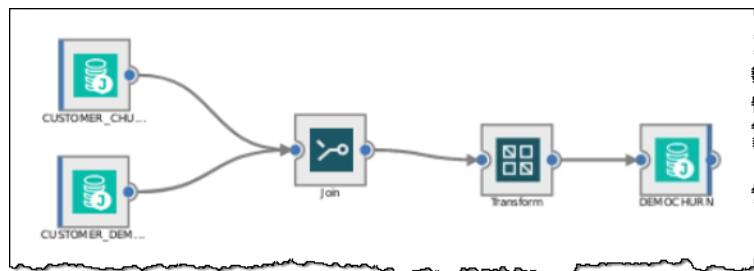
Name	Project type
Data Analysis Project	Analytics
CPD Workshop Analytics Project	Analytics
KR-DV-Project	Analytics
CPD_Workshop_Transform_Project	Data transform

- \_\_127. When you open the project, you are shown a list of any jobs that exist. There are also tabs along the top of the screen for the different types of artifacts that are typically used in developing jobs – Connections, Table Definitions, Parameter Sets, and Jobs

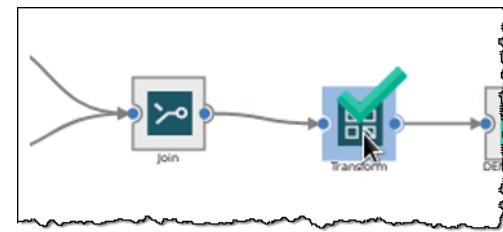
Let's inspect the job that's there. Click on the name to open it:

The screenshot shows a software interface for managing projects. At the top, it says "Project: CPD\_Workshop\_Transform\_Project". Below that is a toolbar with "Connections", "View", "Sort by", a search bar ("Search (1 shown)"), and dropdown menus for "All types" and "Name". A large list area is titled "Name ▲" and contains one item: "Transform\_Job\_Demo\_Churn" with a green checkmark next to it.

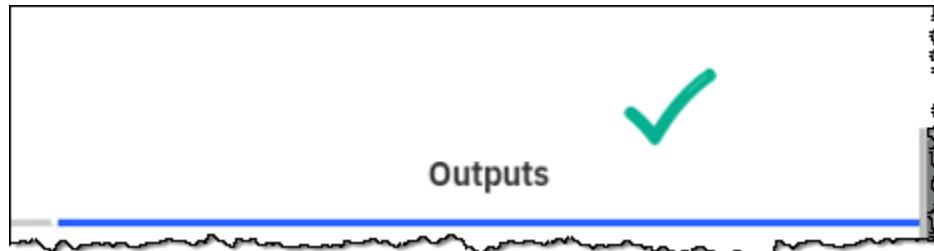
- \_\_128. When we open the job, we see a graphical data flow moving left to right. This job actually joins two tables, then does some work to transform some of the columns, and finally writes out the joined data to a new table.



- \_\_129. Double click on the **Transform** stage – this opens up a window that shows our transformation logic.



- \_\_130. Click on **Outputs**.



\_\_131. Now we see a grid with Columns for derivation, column name, etc.

Scroll all the way to the bottom and we see a row with a derivation, for the column AGE\_GROUP.



	LINK_52.ZIP4	ZIP4
	Link_52.LONGITUDE	LONGITUDE
	Link_52.LATITUDE	LATITUDE
	If Link_52.AGE < 18 THEN "Child" ELSE IF Link_52.AGE < 30 THEN "Young adult" ELSE IF Link_52.AGE < 65 THEN "Adult" ELSE "Senior"	AGE_GROUP 

Runtime column propagation

\_\_132. Double click to see the full derivation; in this case we're creating a column to hold a value that describes the age group of each individual.

Derivation Builder - AGE\_GROUP VARCHAR(11)



```
If Link_52.AGE < 18 THEN "Child" ELSE IF Link_52.AGE < 30 THEN "Young adult" ELSE IF Link_52.AGE < 65 THEN "Adult" ELSE "Senior"
```

\_\_133. Cancel out of the stages until you're back to the job flow diagram.

Then click the **x** in the upper right corner to close the job.

### 13.9.2 Creating a job

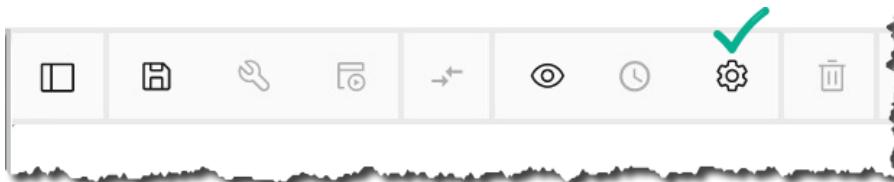
We will next create our own Transformation job.

\_\_134. Click Jobs Tab, then **Create** ⇒ **Parallel job**.

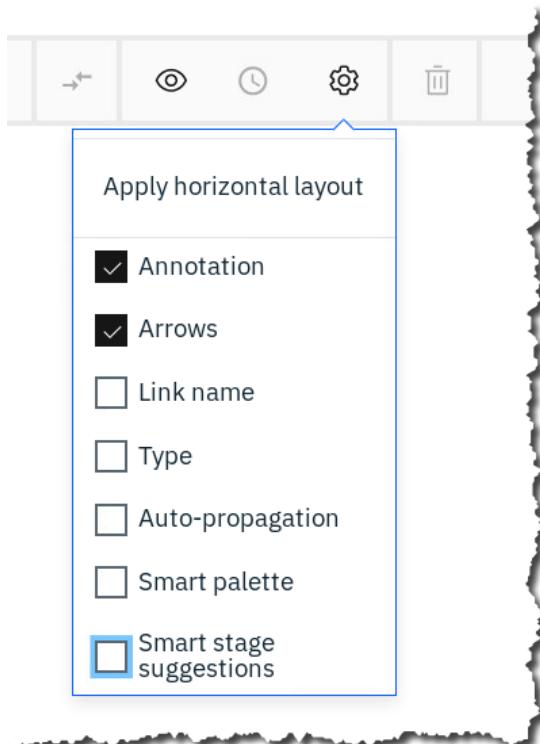


\_\_135. This created a blank canvas we can use to develop our job. Before we begin, first ensure the settings are correct.

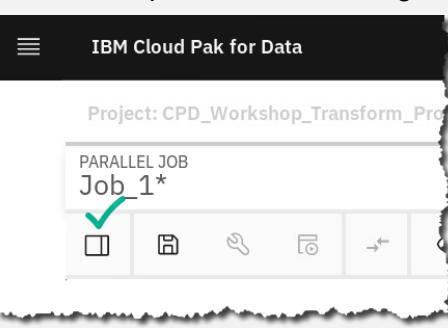
Click the gear icon for [Settings](#).



\_\_136. Then ensure the list shows the same checked and unchecked options as below:



\_\_137. On the left you see all of the different stages available for connecting to data, transforming it, routing it, etc.

 Data Steward	<p>Note: If the palette isn't showing, click the leftmost icon on the button bar:</p>  A screenshot of the IBM Cloud Pak for Data interface. On the left is a sidebar with a "Data Steward" icon. The main area shows a "Project: CPD_Workshop_Transform_Pro" with a "PARALLEL JOB" named "Job_1*". Below the project name is a toolbar with several icons. The first icon in the toolbar, which is a window icon, has a green checkmark above it, indicating it was clicked to open the palette.
---	--

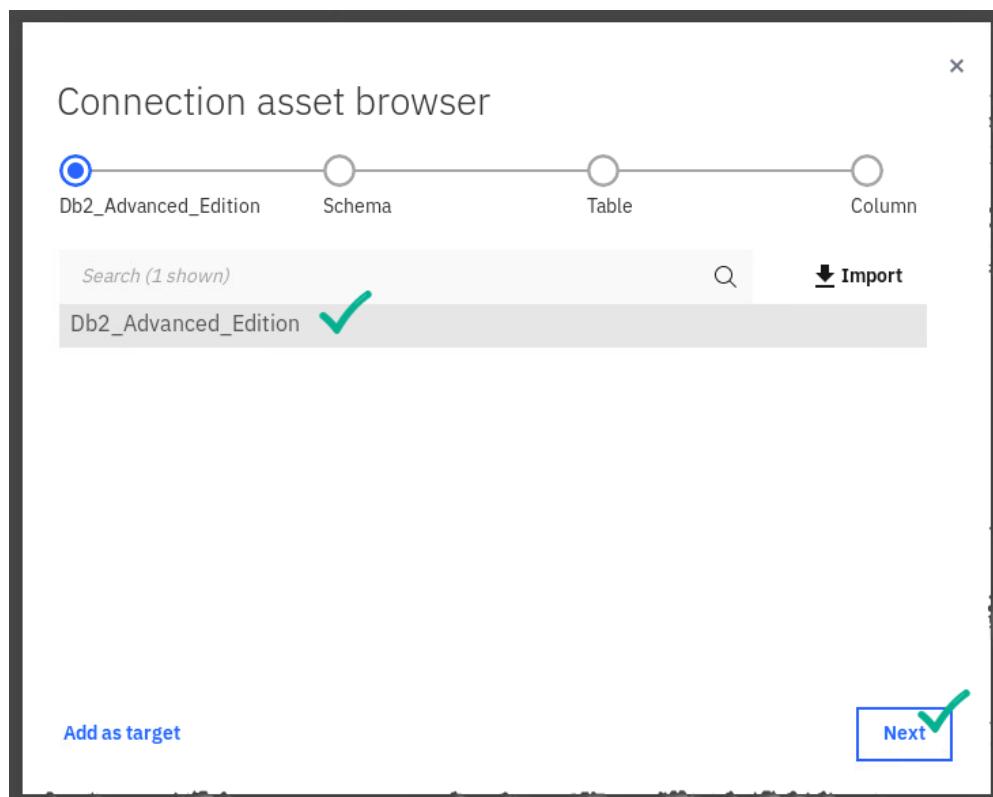
—138. We're going to create a job that joins our two tables, **CUSTOMER\_DEMOGRAPHICS** and **CUSTOMER\_ACTIVITY**. Let's start by dropping two stages to represent these tables.

On the palette, find the **Connection** icon. It should be the first one on the palette.

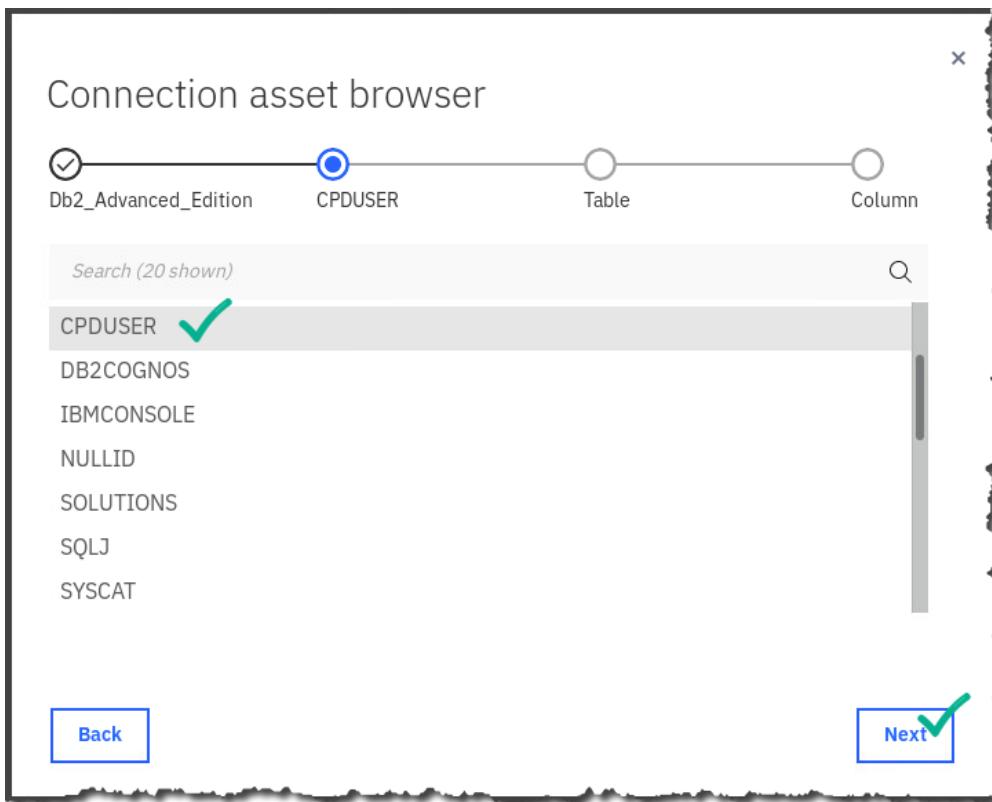
Left click the icon, drag it to the canvas, and left click again to release it. This will open up a window where you'll be guided to fill out the connection information for the database we'll be extracting from.



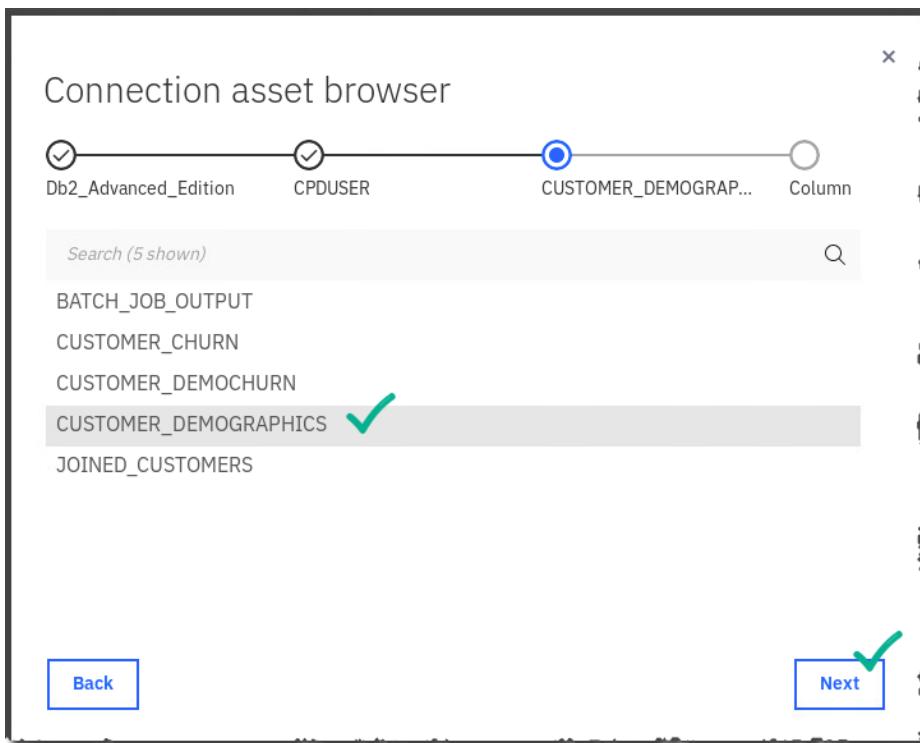
—139. The Connection asset browser appears. Click **Db2 Advanced Edition** to select our Db2 Advanced Edition connection, choose it from the list and then click **Next**.



\_\_140. On the next window, choose Schema [CPDUSER](#), and click [Next](#).



\_\_141. Next, it's time to choose our table – we're going to choose [CUSTOMER\\_DEMOGRAPHICS](#) and then click [Next](#).



\_\_142. Now we see a list of the columns in our table.

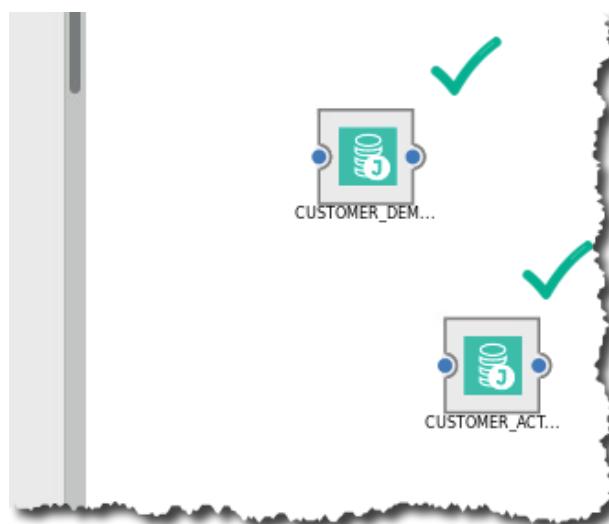
Click [Add to job](#).

The screenshot shows a modal window titled "Connection asset browser". At the top, there are four circular checkboxes: the first three are checked (Db2\_Advanced\_Edition, CPDUSER, CUSTOMER\_DEMOGRAP...), and the fourth is uncheckable (Column). Below the checkboxes is a table with the following data:

	Name	Type	Length
<input checked="" type="checkbox"/>	ID	SMALLINT	0
<input checked="" type="checkbox"/>	GENDER	VARCHAR	1
<input checked="" type="checkbox"/>	STATUS	VARCHAR	1
<input checked="" type="checkbox"/>	CHILDREN	SMALLINT	0
<input checked="" type="checkbox"/>	ESTINCOME	DECIMAL	9

At the bottom left is a "Back" button, and at the bottom right is an "Add to job" button, which is highlighted with a blue box and a green checkmark.

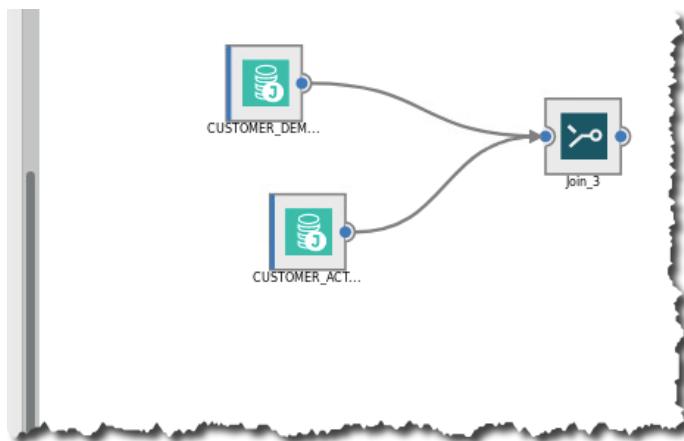
\_\_143. Now, we'll do the same process for our second table. Repeat the same steps you just did, but this time, change the schema you choose to 'SOLUTIONS' and the table name to '['CUSTOMER\\_ACTIVITY'](#)'. When finished, your canvas should look like this:



\_\_144. Single click the **CUSTOMER\_DEMOGRAPHICS** icon on the canvas.

\_\_145. Navigate to the palette on the left and do the following:

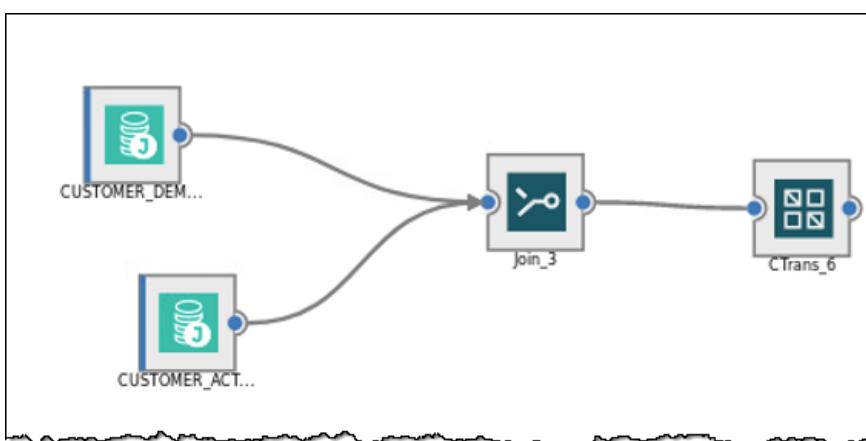
- Find the **Join** icon on the palette (in the Stages section.).
- Click it once.
- Hold your mouse button down and drag the stage to the canvas.
- This should draw a link from **CUSTOMER\_DEMOGRAPHICS** to the join stage. [If you do not see the link being drawn, remove the Join and make sure that you click **CUSTOMER\_DEMOGRAPHIC** first.]
- Next, single click on the **CUSTOMER\_ACTIVITY** stage.
- Click the small blue button on the stage, then drag a link to the join stage.
- You should see the join between **CUSTOMER\_DEMOGRAPHIC** and **CUSTOMER\_ACTIVITY**.



\_\_146. If you do not get it right it the first time, delete the **Join** from canvas and try again.

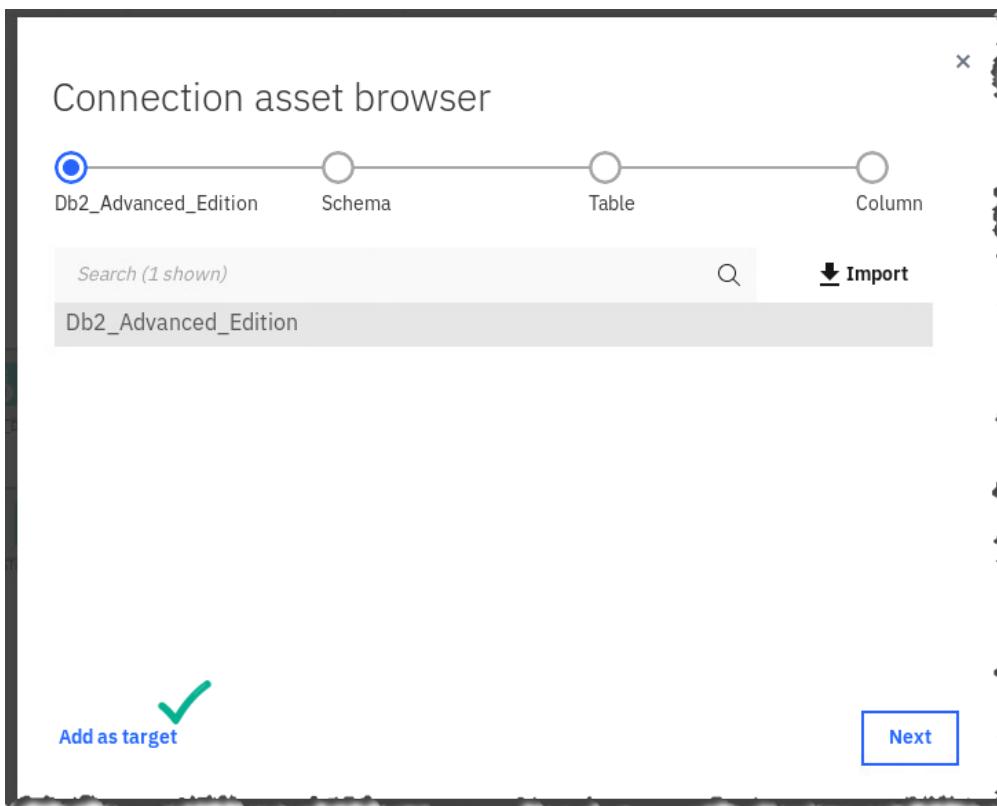
\_\_147. Single click **Join** on the canvas to select it. (Make sure you select it before going to next step.)

\_\_148. Next, go to the palette, select the **transformer** stage, and drag it onto the screen.

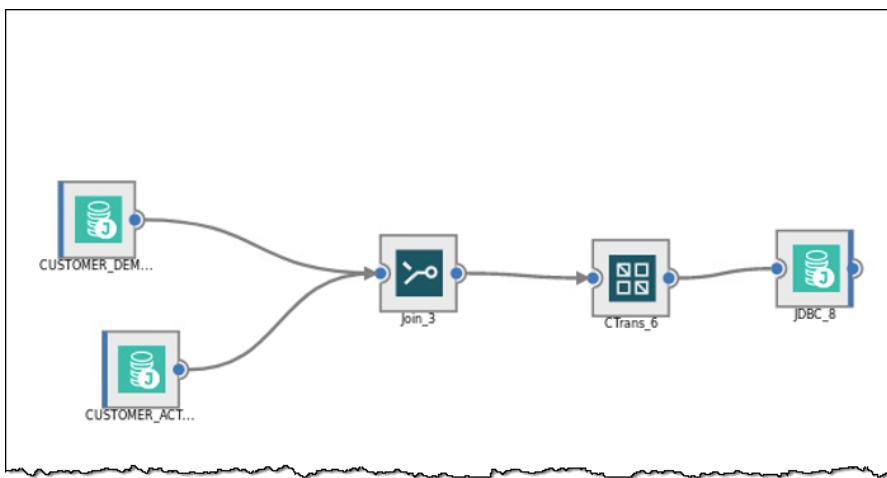


\_\_149. Lastly, we'll add our target stage. Single click on the Transform stage. Then select the **Connector** icon at the top of the palette, and drag it onto the canvas. This will pop up the connections configuration window.

\_\_150. Click the link at the bottom that says **Add as Target**. This should immediately link the target stage with the Transformer stage, but if not, draw the link from the transformer stage to the target.



You should see the following:



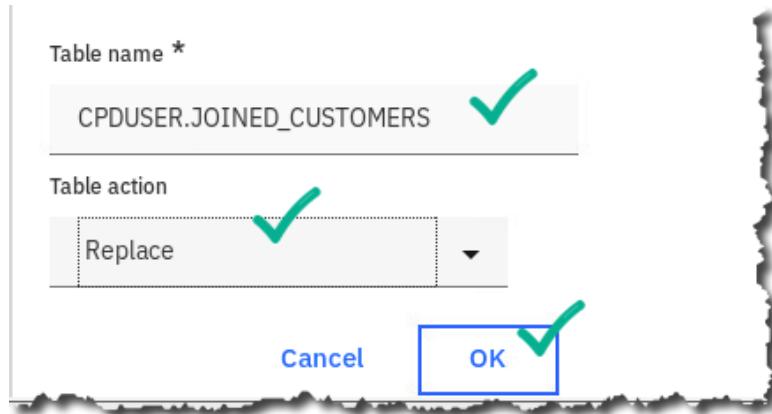
\_\_151. Next, double click the **last connection icon**.

This will bring up the property sheet for our target table. We'll want to put in a table name, and we'll also want to create the table as part of the job run if it is not there.

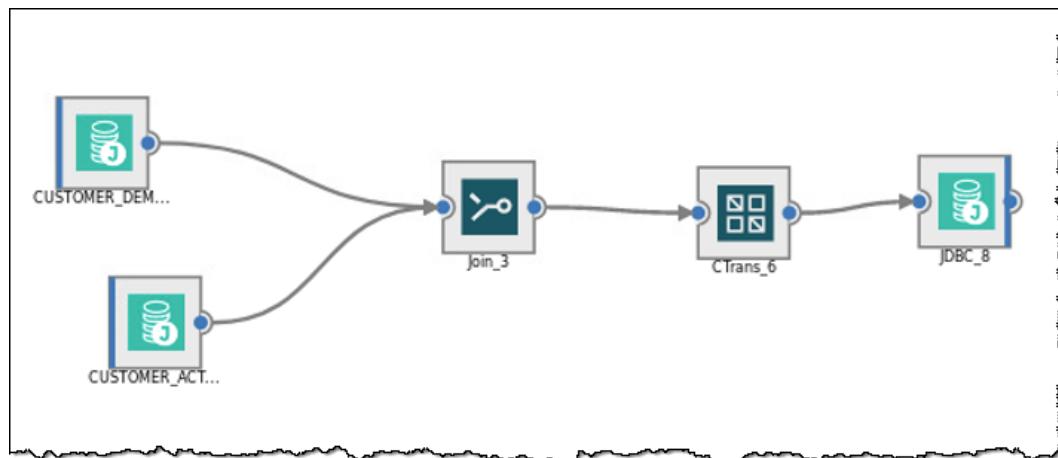
\_152. First, under the **Table name** field, enter **CPDUSER.JOINED\_CUSTOMERS**, and then under **Table action** choose **Replace**.

This will attempt to drop and re-create the table, but if no table of the same name exists, it will just create the table.

Click **OK** to save the properties.



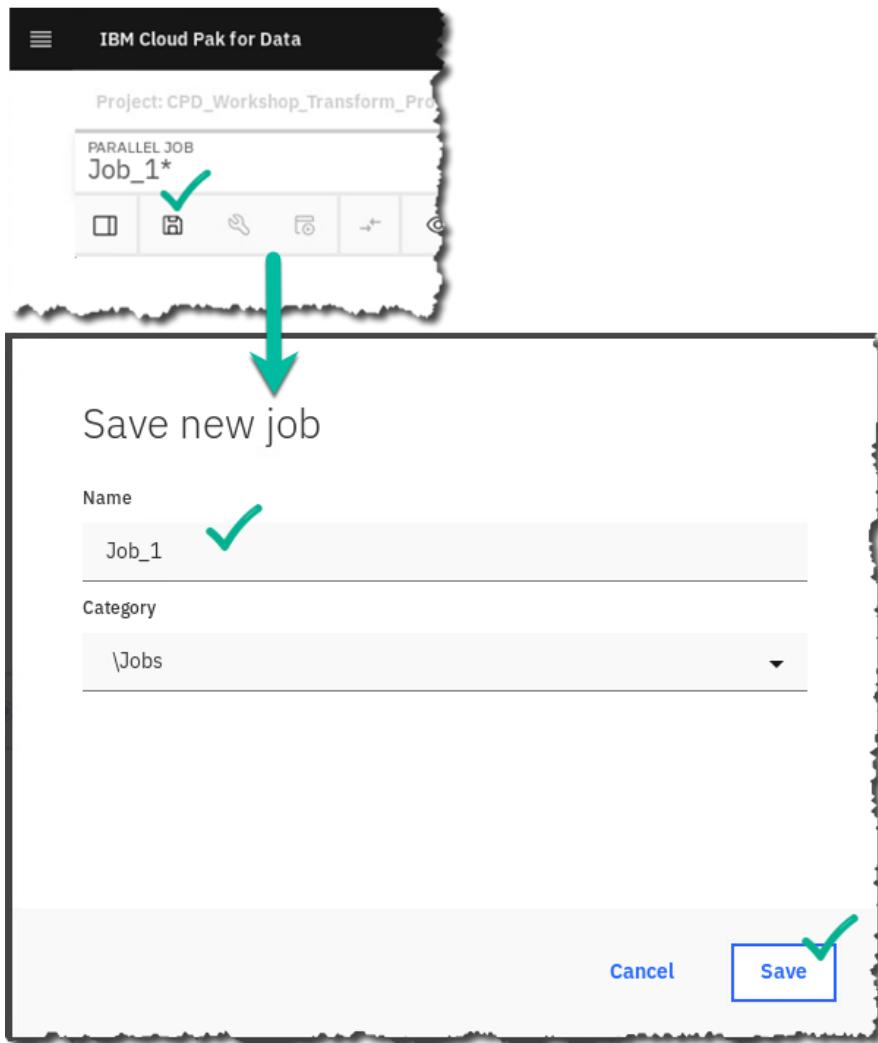
\_153. Your job should now look like this:



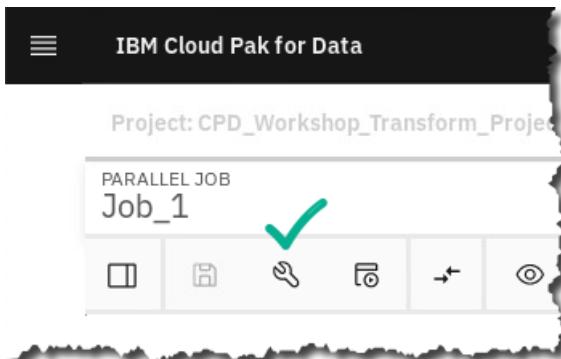
\_\_154. Next, we'll save, compile and then run the job.

First, click the **diskette** icon to save the job.

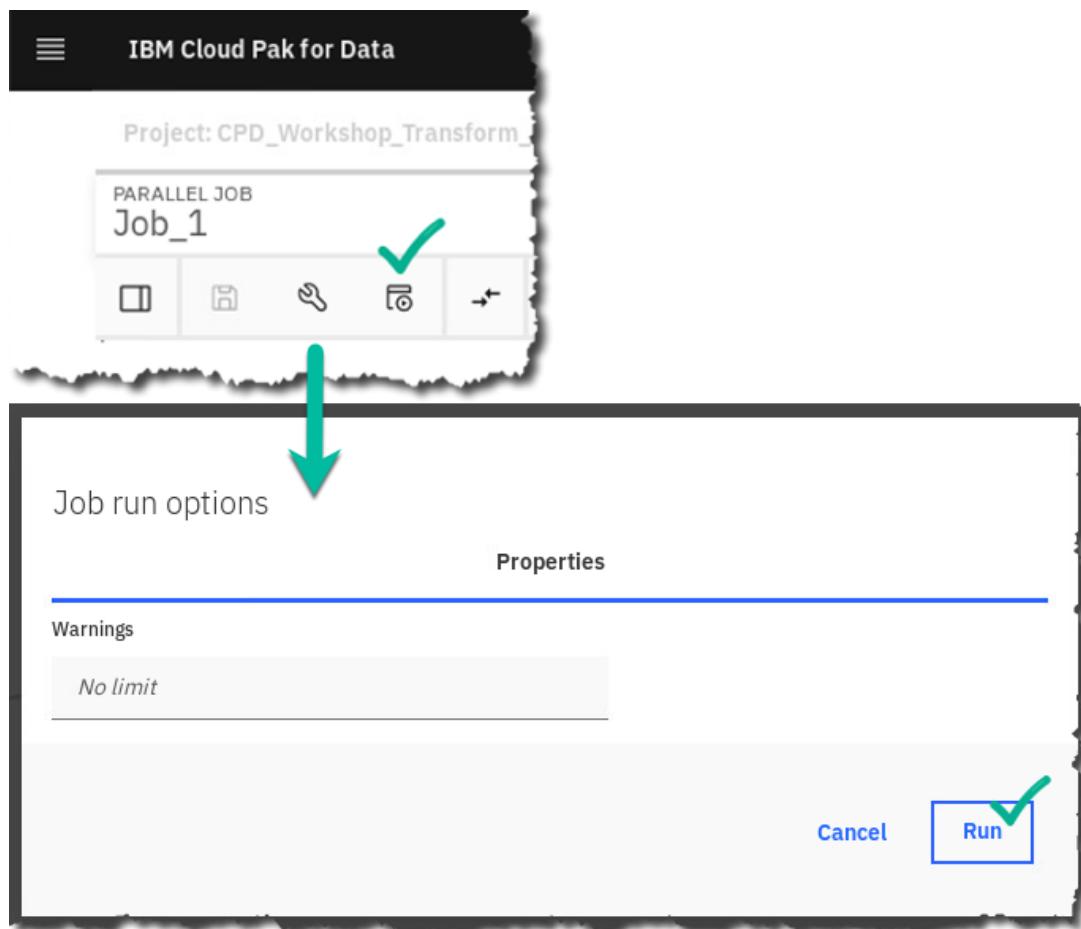
Give it a name (or take the default **Job\_1**) and click **Save**.



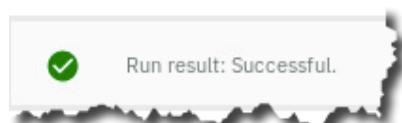
\_\_155. Next, we'll compile the job by clicking the **wrench** icon.



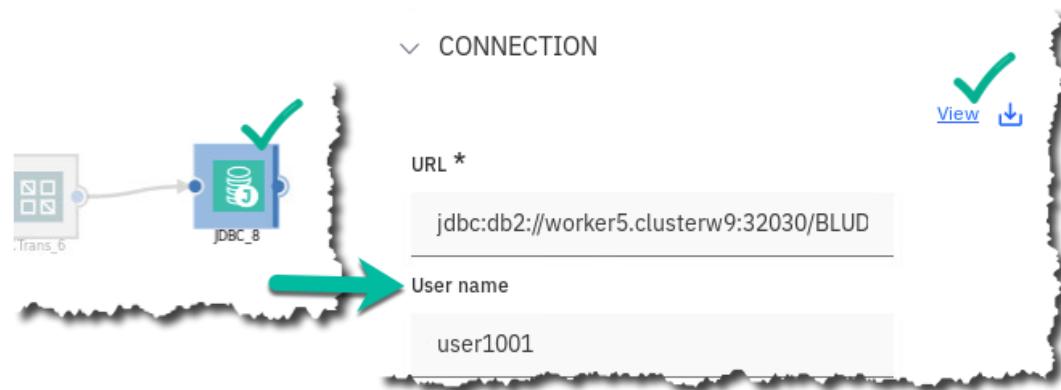
\_\_156. Begin to run the job by clicking the [run](#) icon, then [Run](#) the job.



\_\_157. When the job is done you'll see a green check mark and a run result.  
(Note: A *Warning message* may appear that may be safely ignored.)



Lastly, if you'd like to see the result, double-click on the target stage (the rightmost stage) and then click [View](#) link in the property sheet. This will show us our joined result.

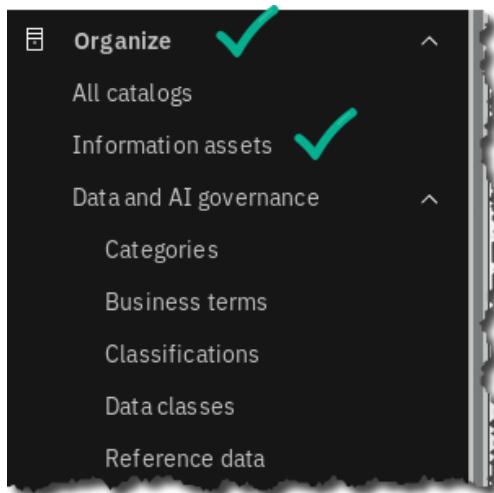


## 13.10 Data Lineage

The last thing we will explore is [Data Lineage](#). One of the main benefits of a well governed Information Architecture is the ability to describe and document how data flows through your organization, and what transformations are executed on the data. This allows organizations to better comply with regulations, understand the impact of changing different components of their Information Architecture, and a host of other benefits.

- 158. The job we have just built and run automatically documents our data flow and allows us to easily visualize what happened during this job run. Let's go see what this looks like:

Navigate to [Information Assets](#).



- 159. This will bring up a screen that shows you recently accessed assets, and a search bar.

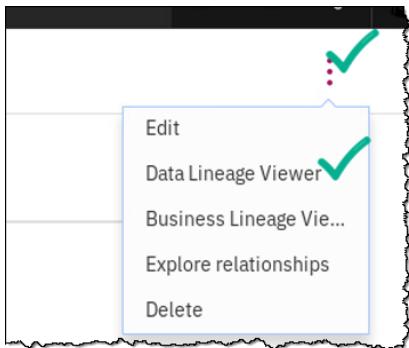
Click on the table [CUSTOMER\\_DEMOGRAPHICS](#). If the table is not there, search for it by typing [DEMO](#) into the search bar.

A screenshot of the 'Explore information assets' interface. At the top, there is a search bar with dropdowns for 'All asset types' and 'Find assets'. Below the search bar, the title 'Explore information assets' is displayed. A section titled 'My recently viewed assets' contains a table with two columns: 'Name' and 'Description'. The first row in the table is 'CUSTOMER\_DEMOGRAPHICS', which is highlighted with a green checkmark. The 'Name' column shows the table name, and the 'Description' column shows the connection details: 'dbc:db2://worker5.clusterw9:32030/BLUDB >> BLUDB >> CCPDUSER'.

Name	Description
CUSTOMER_DEMOGRAPHICS	dbc:db2://worker5.clusterw9:32030/BLUDB >> BLUDB >> CCPDUSER

\_\_160. Once chosen, a window showing table details is displayed.

Click the **ellipses** (3 vertical dots) in the upper right corner, and then click **Data Lineage Viewer**.



\_\_161. Click **Run Lineage**.



\_\_162. You will then be brought to a visualization of the flow of the transform job we created.

Notice that our source table, **CUSTOMER\_DEMOGRAPHICS**, is actually a source for two separate jobs, and that those jobs write the data to two separate target tables.

You can further explore and drill into these steps to see more detail. This is an incredibly powerful capability, which is automatically available by virtue of the Transform design process.

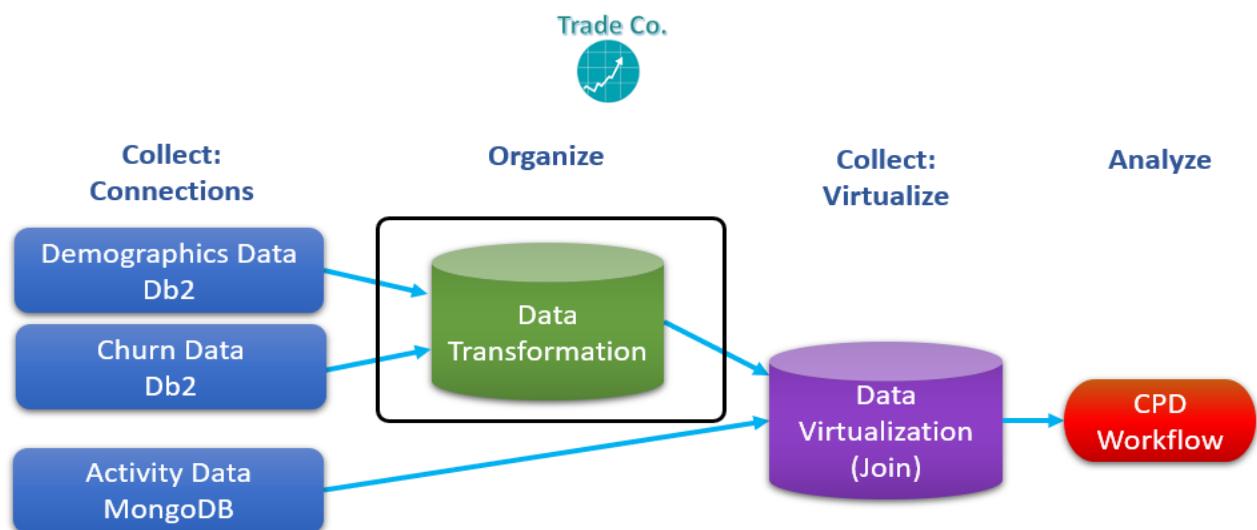
## 13.11 Lab conclusion

We have seen the value in creating a [Data Dictionary](#) by creating a [glossary](#) of [categories](#) and [terms](#) to make data searchable so that data scientists, data engineers and business analysts can shop for data. The CPD platform reduces the time-consuming data organization task, making it easier for the consumers of the data to access and analyze the trusted data they need.

This lab has shown you how to automatically [discover](#) data sources, automatically classify those sources with business classifications using CPD machine learning methods, and automatically assign business terms to those data sources.

This lab also showed you how to create [data protection rules](#), and join and move data to an analytics Advanced Edition in a [Data Transformation](#) step for further use by the data consumers in the CPD workflow.

**The steps covered here could normally take many weeks, months, and sometimes even years, to complete using traditional manual methods. Cloud Pak for Data automates these operations so that you can accelerate the time to value of your organization's analytics projects.**



### \*\* End of Lab 13 – Organize – Deeper Dive

Lab by John Van Buren, Burt Vialpando and Kent Rubin

## 13.12 Additional Optional Activities

There are many other capabilities within the ‘Organize’ section of Cloud Pak for Data. Here we will highlight one of these.

### Using Data Virtualization sources in Organize

One of the strengths of an integrated platform is that you can use the various capabilities interchangeably, reuse previously designed components, and in general create a data fabric more quickly than with individual tools.

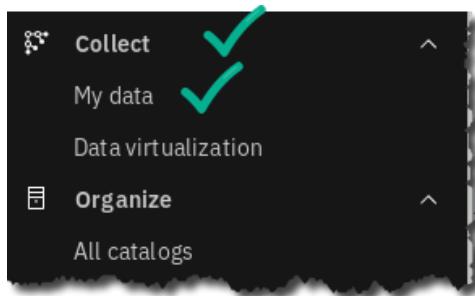
In an earlier lab, we created a virtualized view that was a join of a Db2 warehouse table and a MongoDB document store. What if we wanted to run auto discovery on that view to understand the quality of the data in those two sources when joined together? Or what if we wanted to automatically document the attributes of the view with Business Terms from our Business Glossary? Let’s go ahead and do it!

First, we need to make sure we’ve got our virtualized view created.

**NOTE:** *This view was created during the execution of Lab 3, document: LC03-CPD-3.0.1-Collect-Connections.docx*

- 163. You can verify it is there by doing the following:

From the main navigation menu, choose [Collect→My data](#).



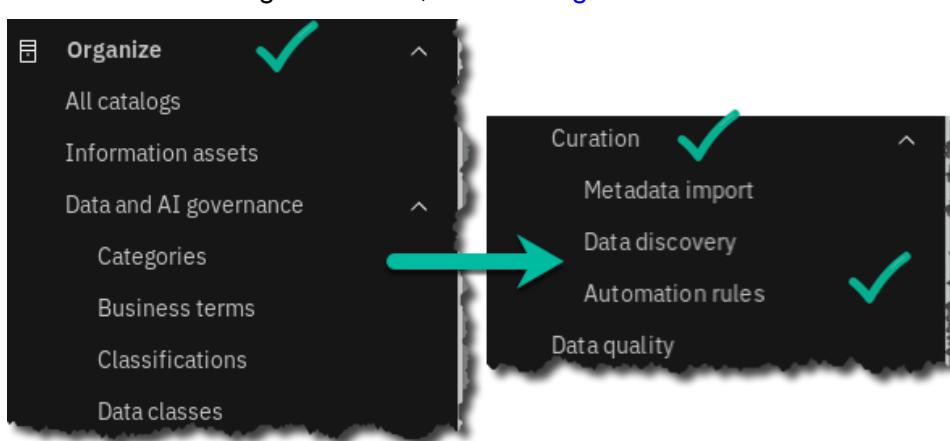
- 164. This will bring up a list of virtualized tables and views. Look for the one with the name [USER1001.CUSTOMER\\_DEMOCHURN](#); this is a view that represents the join of our DEMOGRAPHICS table, in Db2 with our ACTIVITY table, in MongoDB:

4	CPD Workshop Analytics Project	customer_demochurn_activity_analyze.csv	text/csv
5	CPD Workshop Analytics Project	USER1001.CUSTOMER_DEMOCHURN	application/octet-stream

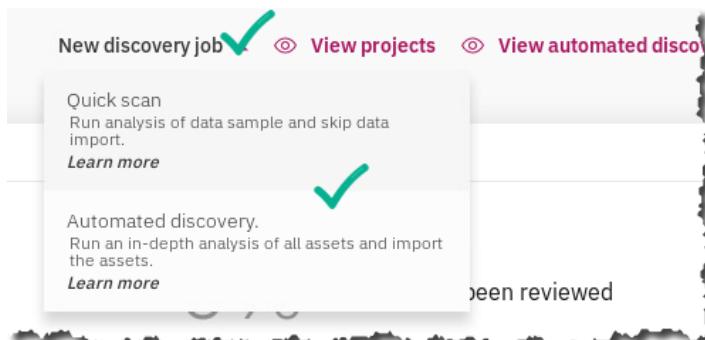
- 165. If that asset is there, you can proceed to the next step.

 Data Steward	<p>Note: If it's not there, please go back and run the Collect and Virtualize lab from our CORE.</p>
---	--

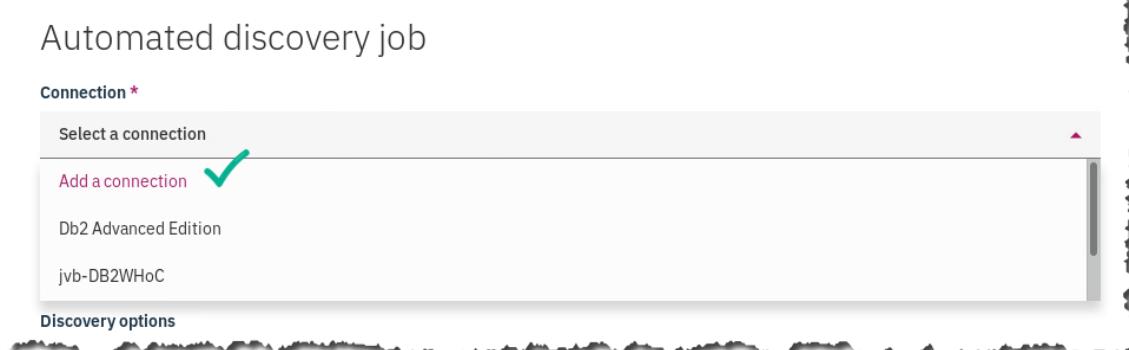
- \_\_166. Once you've established that your virtualized view is there, we want to simply run some of the actions against it the way we would against any other table.
- \_\_167. From the main navigation menu, choose [Organize](#) ⇒ [Curation](#) ⇒ [Data discovery](#).



- \_\_168. Here, we'll want to run an automated discovery just like before – choose [new discovery job](#) → [automated discovery](#)



- \_\_169. Next click the '[Select a Connection](#)' dropdown, and choose '[Add a connection](#)'



 Data Steward	<p>Note: if the connection 'DV-VIRTUAL-SOURCES' appears in the list, then you can skip adding a connection. Simply choose 'DV-VIRTUAL-SOURCES' from the list.</p>
---	---

- \_\_170. The screen will change to a list of possible connections, choose ‘DV-VIRTUAL-SOURCES’ and then click ‘Next’:

Name	Type	URL	Created By
DV-VIRTUAL-SOURCES	DB2	jdbc:db2://dv-server.cp4d.svc.cluster.local:32051/BIGSQL	CPD User

- \_\_171. This will send you back to the discovery window – click the **Browse** button to take a look at the data available –

- \_\_172. This will bring up a tree view of the database (BigSQL) and the member schemas. Navigate down until you come to ‘USER1001’ and choose our virtual view, **CUSTOMER\_DEMOCHURN\_ACTIVITY**, and click Select

- 173. Next, click the checkboxes for ‘Analyze columns’ and ‘Analyze data quality’, and then drop down the list of projects and choose ‘ORGANIZE-WORKSHOP’:

**Discovery options**

- Analyze columns ✓
- Analyze data quality ✓
- Assign terms
- Publish results to catalog
- Use data sampling

Set the maximum number of records that you want to include in your data set sample:  
Example: 2000

Select the method that you want to use to create your sample:

- Use the first x number of rows (where x = maximum number of records allowed)
- Use every Nth value (up to maximum number of records)  
Nth interval  
Example: 1000
- Use a random sampling  
Seed  
Example: 1234 Percentage  
Example: 10

Project \* ⓘ  
ORGANIZE-WORKSHOP

- 174. Click the ‘Discover’ button. This will start the discovery process and you will see a status window you can click the refresh icon periodically to see when it is finished. :

Discovery job 1594995076963 Running

**General information**

Start July 17, 2020, 10:11 AM Started by cpduser

Assets included in the discovery 1

**Discover options**

Project ORGANIZE-WORKSHOP Connection DV-VIRTUAL-SOURCES

Discovery options used: Column analysis, Term assignment, Data quality analysis

Source asset import All assets Analysis All assets

**Discovered assets**

Number of schemas 1 Number of tables 1

Asset name	Asset type	Tables	Status	Actions
CUSTOMER_DEMOCHURN_ACTIVITY	Table	1	Phase Import <span>Finished</span> Start July 17, 2020, 10:11 AM End July 17, 2020, 10:11 AM	<span>refresh</span> <span>info</span>

- 175. It will first show a status of phase Analyze as ‘running’:

**Discovered assets**

Number of schemas 1 Number of tables 1

Asset name	Asset type	Tables	Status	Phase	Import	Analyze	Actions
CUSTOMER_DEMOCHURN_ACTIVITY	Table	1	Phase Analyze <span>Running</span> Start July 17, 2020, 10:11 AM Done 0% Successful 100% Cancelled 0% Failed 0%	Import	<span>Finished</span>	<span>Running</span>	<span>refresh</span> <span>info</span>

- 176. Then, after about 60 seconds, click refresh again and the phase Analyze should change to 'Finished' Once you see that it has finished, you can click the [Eye icon](#) to see the results:

Discovered assets		Number of schemas 1			
Asset name	Asset type	Tables	Status	Phase	Actions
CUSTOMER_DEMOCHURN_ACTIVITY	Table	1	Phase Import <span style="background-color: #2e7131; color: white; padding: 2px 5px;">Finished</span> Phase Analyze <span style="background-color: #2e7131; color: white; padding: 2px 5px;">Finished</span>	Start July 17, 2020, 10:11 AM End July 17, 2020, 10:11 AM Done 100% Successful 100% Cancelled 0% Failed 0%	

- 177. This will bring up a screen showing the quality scores, data class choices made automatically, and business term assignments made automatically:

CUSTOMER_DEMOCHURN_ACTIVITY					
<input checked="" type="checkbox"/>	CUSTOMER_DEMOCHURN_ACTIVITY	97%	—	—	Jul 17, 2020, 10:13 AM
	ADDRESS_1	75%	US Street Name 76% ▾	—	Jul 17, 2020, 10:13 AM
	ADDRESS_2	100%	NoClassDetected 100% ▾	—	Jul 17, 2020, 10:13 AM
	AGE	99%	Code 100% ▾	—	Jul 17, 2020, 10:13 AM
	AGE_GROUP	100%	Code 100% ▾	—	Jul 17, 2020, 10:13 AM
	CHILDREN	100%	Code 100% ▾	—	Jul 17, 2020, 10:13 AM
	CHURNRISK	81%	City 81% ▾	—	Jul 17, 2020, 10:13 AM
	CITY	95%	City 95% ▾	—	Jul 17, 2020, 10:13 AM
	CREDITCARD	100%	Identifier 100% ▾	—	Jul 17, 2020, 10:13 AM
	DAYSSINCELASTLOGIN	100%	Code 100% ▾	Days Since Last Trade 57% ▾	Jul 17, 2020, 10:13 AM
	DAYSSINCELASTTRADE	100%	Code 100% ▾	Days Since Last Trade 100% ✘ ✓	Jul 17, 2020, 10:13 AM
	DOB	100%	Date of Birth 100% ▾	—	Jul 17, 2020, 10:13 AM
	ESTINCOME	100%	Income 100% ▾	✓	Jul 17, 2020, 10:13 AM
	GENDF^	—	Gender 100% ✘	—	Jul 17, 2020, 10:13 AM

In conclusion: what we have accomplished here is that we can understand the quality, possible data classifications, and business terms of a virtual table, in this case containing both structured Db2 data and semi-structured MongoDB data, as we can for any normal table.

The integrated nature of Cloud Pak for Data is what makes it possible to do so many of your tasks from one place.

## Lab 14 COGNOS DASHBOARD EMBEDDED - DEEPER DIVE

### 14.1 Lab overview

When embarking on machine learning projects, many organizations engage Business Analysts to help gain insight into their data. This persona can use tools like Analytics Dashboards (part of the CPD offering) or the more capable cartridge from which this service is derived: Cognos Analytics.

In this lab, you will use the Analytics Dashboards to build visualizations to help the organization understand why their customer use is declining.

In our scenario, the Trade Co. Business Analyst used this service to provide the company executives and the data scientists the information they needed to understand their problem as well as to access the effectiveness of the solution.



### 14.2 Persona represented in this lab

The [Business Analyst](#) persona is likely to perform the exercises in this lab, and that is to create visualizations to make sense of the problems the organization is facing.

Persona (Role)	Capabilities
 Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.

### 14.3 Logging into the CPD web client (if you have not already done so)

- \_\_1. If you are starting this lab stand-alone (without going through previous labs) do the following:
- \_\_2. Click the desktop icon: [Cloud Pak for Data Web Client](#).

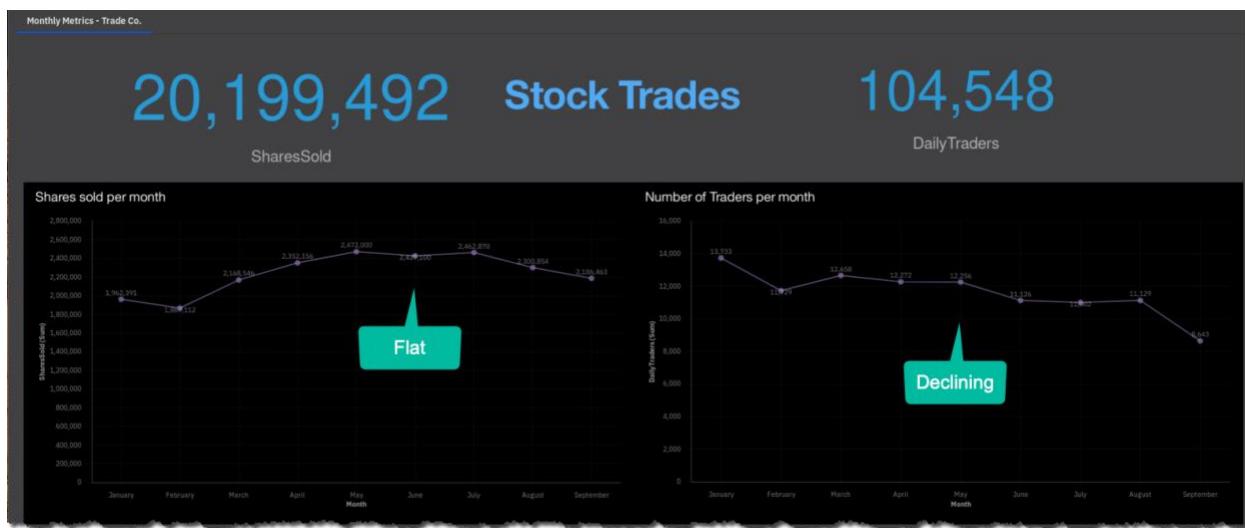


- \_\_3. The CPD web client GUI displays as shown. Use `cpduser` and `cpduser` for the *Username* and *Password* and click **Sign In**.



## 14.4 Reviewing the dashboard: Monthly Metrics - Trade Co.

In the dashboard **Monthly Metrics – Trade Co.**, the business analyst for Trading Co. has painted a picture of how the business was initially doing. The dashboard shows a relatively flat **shares sold per month**, and a declining **number of traders per month**, which the business analyst provides to the Trade Co. executives.



## 14.5 Building the dashboard: Monthly Metrics - Trade Co.

The business analyst began by analyzing current trends of customer visits and daily trades in the Trade Co. Stock Trader application. He requested the data engineers to provide a file with historical totals of visits and trades for the past year ([TraderData.csv](#)) which was deposited into the project where everyone on the team could collaborate.

Build the dashboard with that data to see the trends the business analyst discovered.

### 14.5.1 Starting the dashboard

- \_\_1. From the top left Navigation Menu  $\Rightarrow$  [Projects](#)  $\Rightarrow$  CPD Workshop Analytics Project.
- \_\_2. Under [Assets](#) click [Data Assets](#).

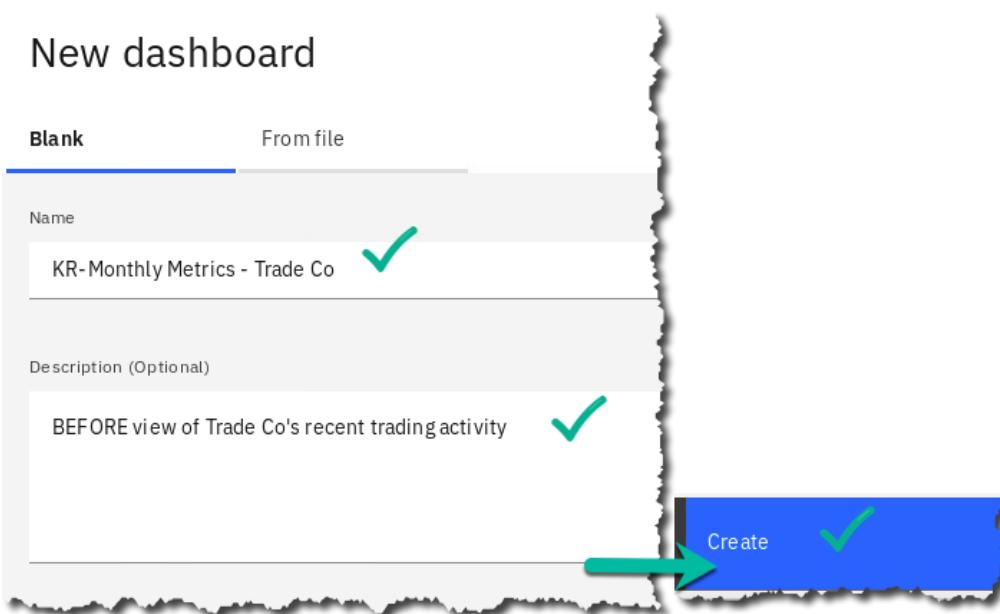
Notice [TraderData.csv](#) available in our project.

Name	Type
CSV TraderData.csv	Data Asset

- \_\_3. Click [Add to project +](#)  $\Rightarrow$  Dashboard.

Available asset types
Data
Connection
Connected data
AutoAI experiment
Notebook
Dashboard
Watson Machine Le...
Modeler flow
Data Refinery flow
Decision Optimizati...

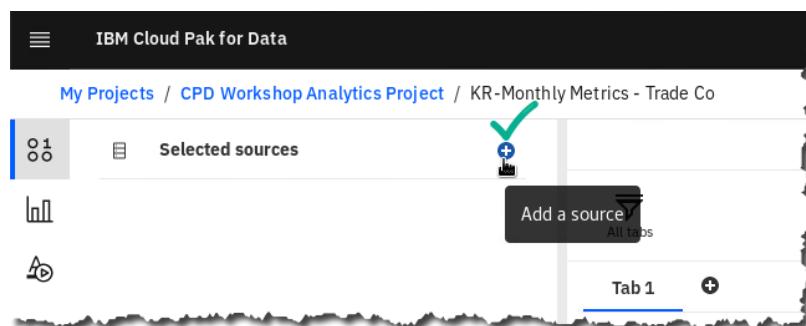
- \_\_4. In the **Blank** tab, fill in the **Name** and **Description** as shown below, then click **Create**  
 - for the name, use **initials-Monthly Metrics – Trade Co.**



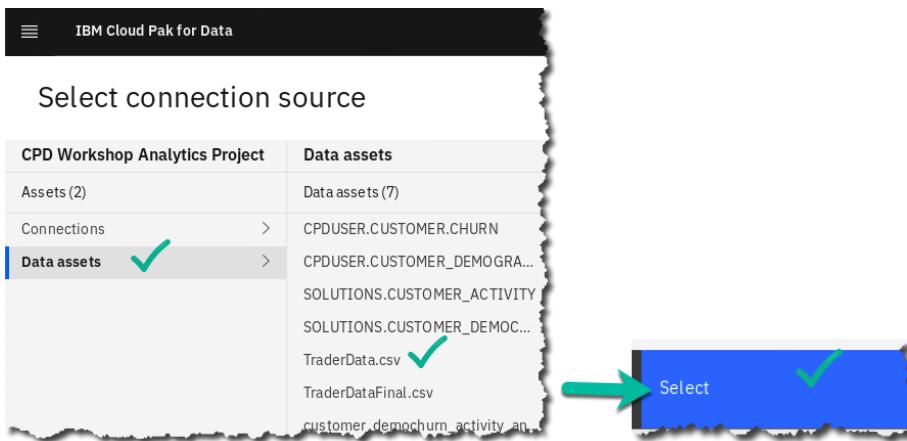
- \_\_5. You are now presented with a choice of canvas templates. Select the one that looks as shown => Click **OK**.



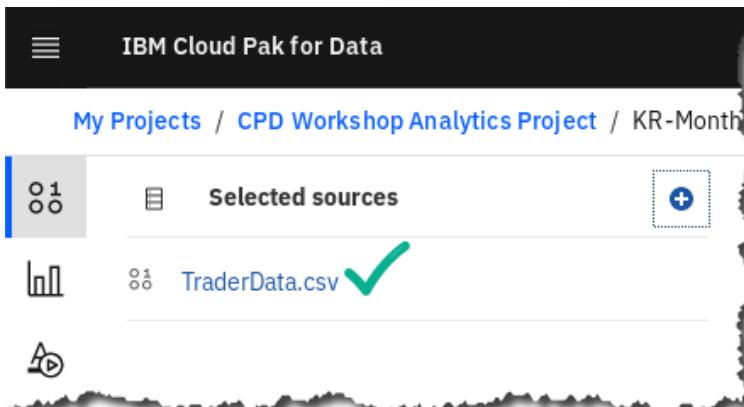
- \_\_6. From the **Selected sources** area near the top left of your screen, click **+**.



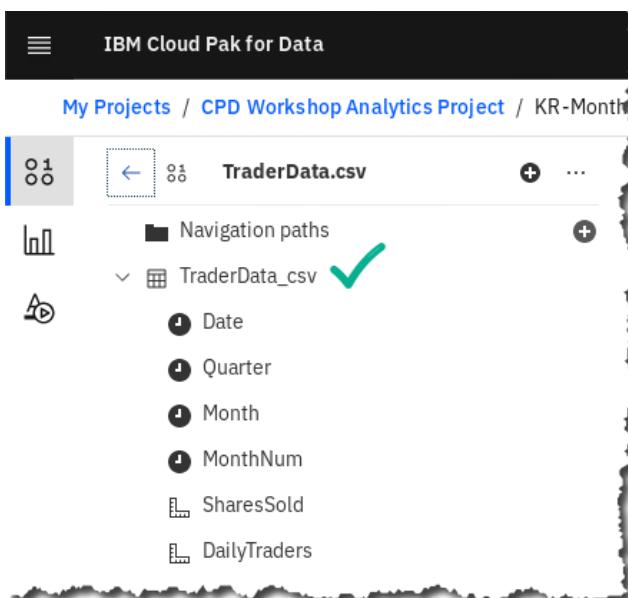
- \_\_7. Expand arrow next to **Data assets** and choose: **TraderData.csv**  $\Rightarrow$  **Select**.



- \_\_8. Click on **TraderData.csv** once it is in the Selected sources. This will allow you to expand it.

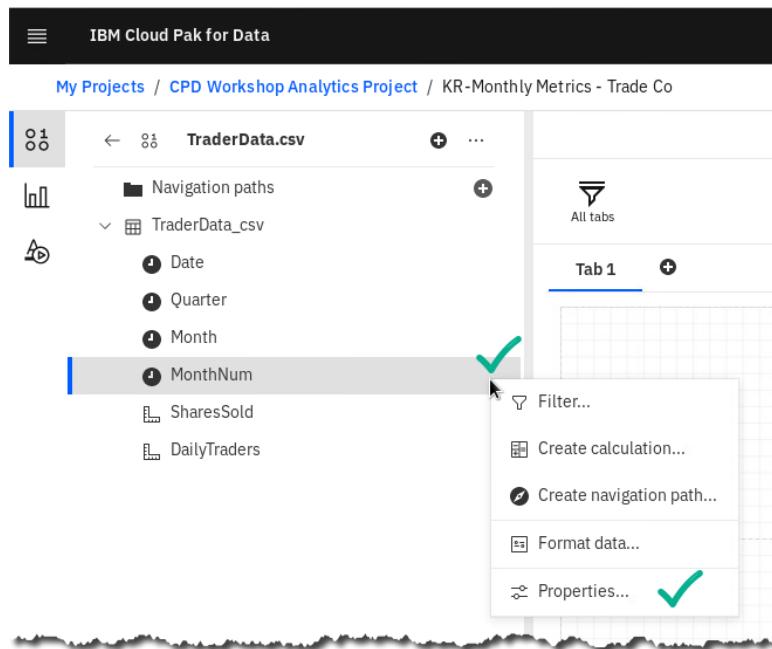


- \_\_9. Now expand **TraderData.csv**.

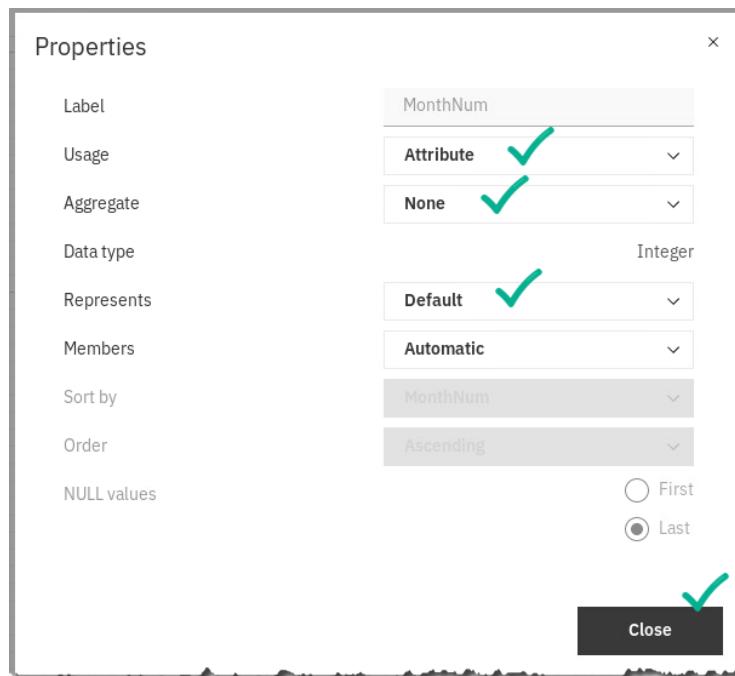


- \_\_10. Next, ensure the properties of the column data are what you want to be represented correctly within the dashboard.

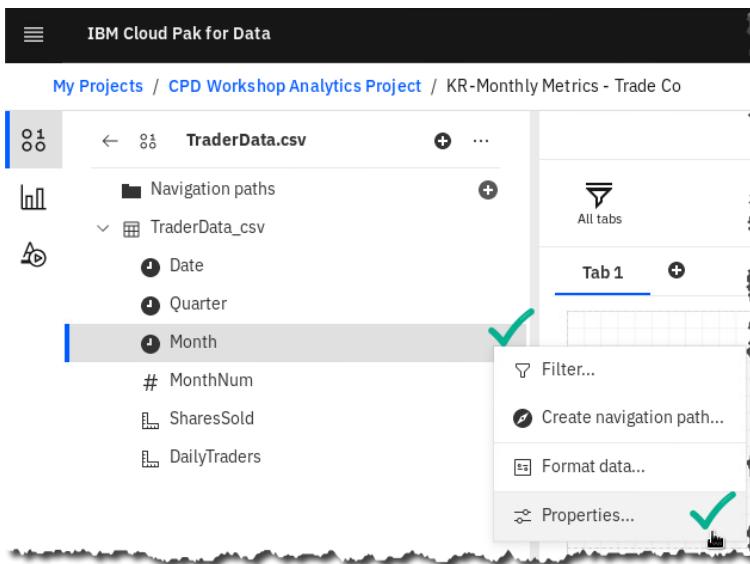
Click the MonthNum ellipsis, then select Properties from the flyout menu.



- \_\_11. MonthNum is an attribute on which we sort our months. Change MonthNum usage to be Attribute and Aggregate to be None, Represents Default, then click Close.



\_\_12. Select the **Properties** of Month by selecting the ellipses next to Month.

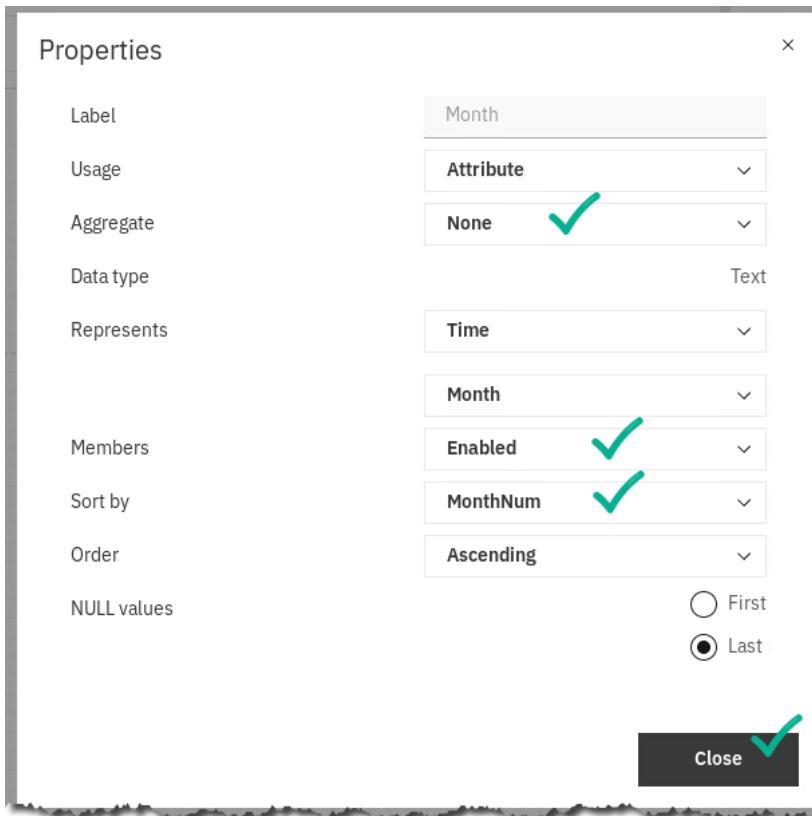


\_\_13. Change **Aggregate** to **None**.

Change **Members** to **Enabled**.

Select **Sort by – MonthNum**.

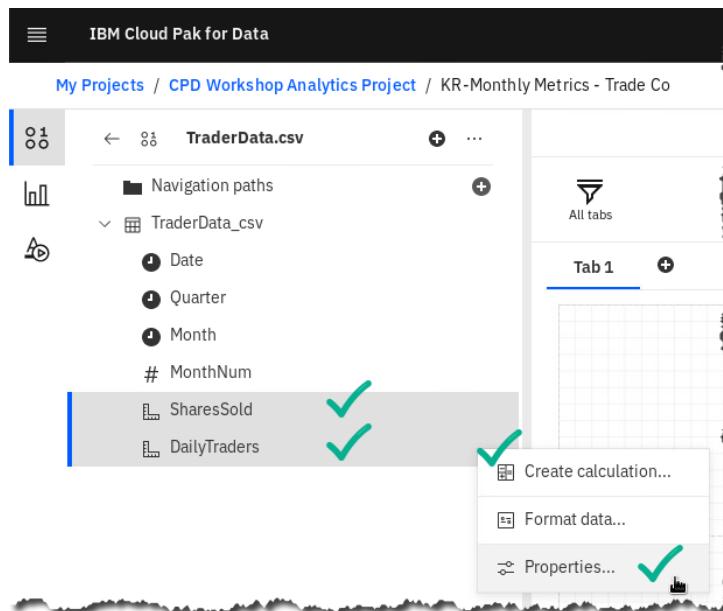
Click **Close**.



\_\_14. Next, set the aggregation of our measures for totaling.

Hold the **shift key down** and click **SharesSold** and **DailyTraders** to select both.

Click the **ellipsis** and then click **Properties**.



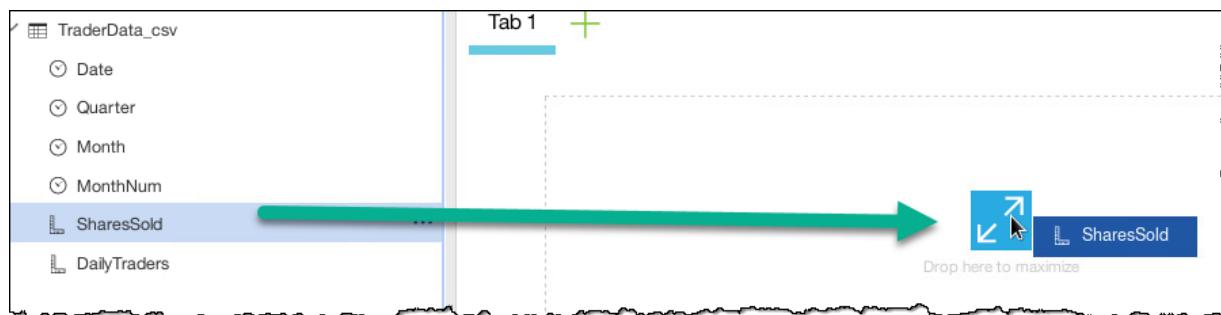
\_\_15. Make sure *Aggregate* is **Total** in the drop-down menu.

Click **Close**.

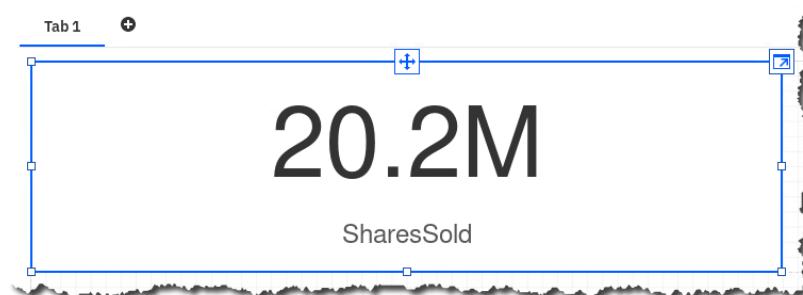


### 14.5.2 Building the first visualizations

- 16. Drag **SharesSold** to the top left box, hovering over the *Drop here to maximize* area when it turns blue, then releasing the mouse. This gives us a total of **Shares sold** over all time.



- 17. If you did not drop at the right place it won't fill the template area. You can still adjust the guide to match with the outline of the left box. It should look as below.



- 18. Drag **DailyTraders** to the top right box, hovering over the *Drop here to maximize* area when it turns blue as well. This maximizes this metric in this box. This gives us a total of trades over all time.

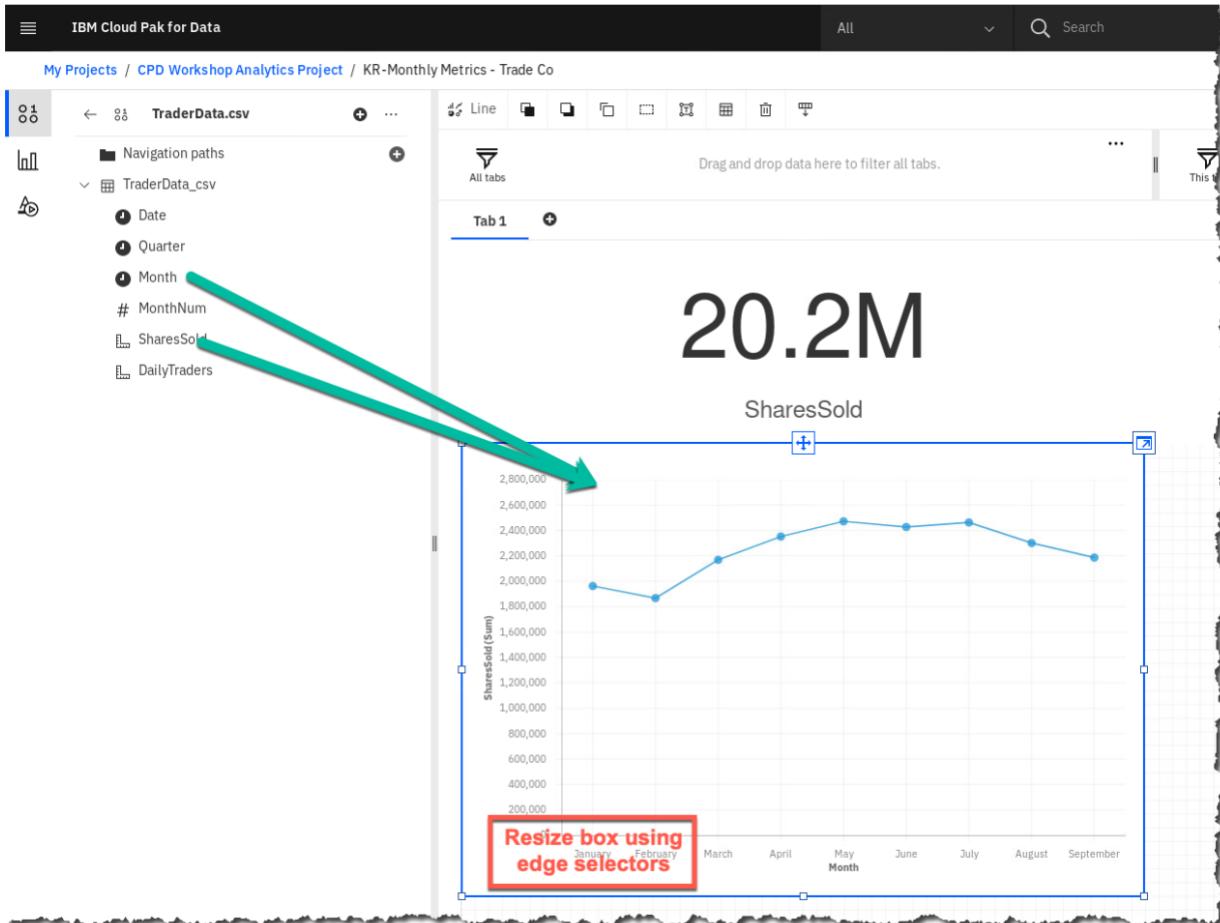


- 19. After you complete both top boxes, the dashboard should display as shown:

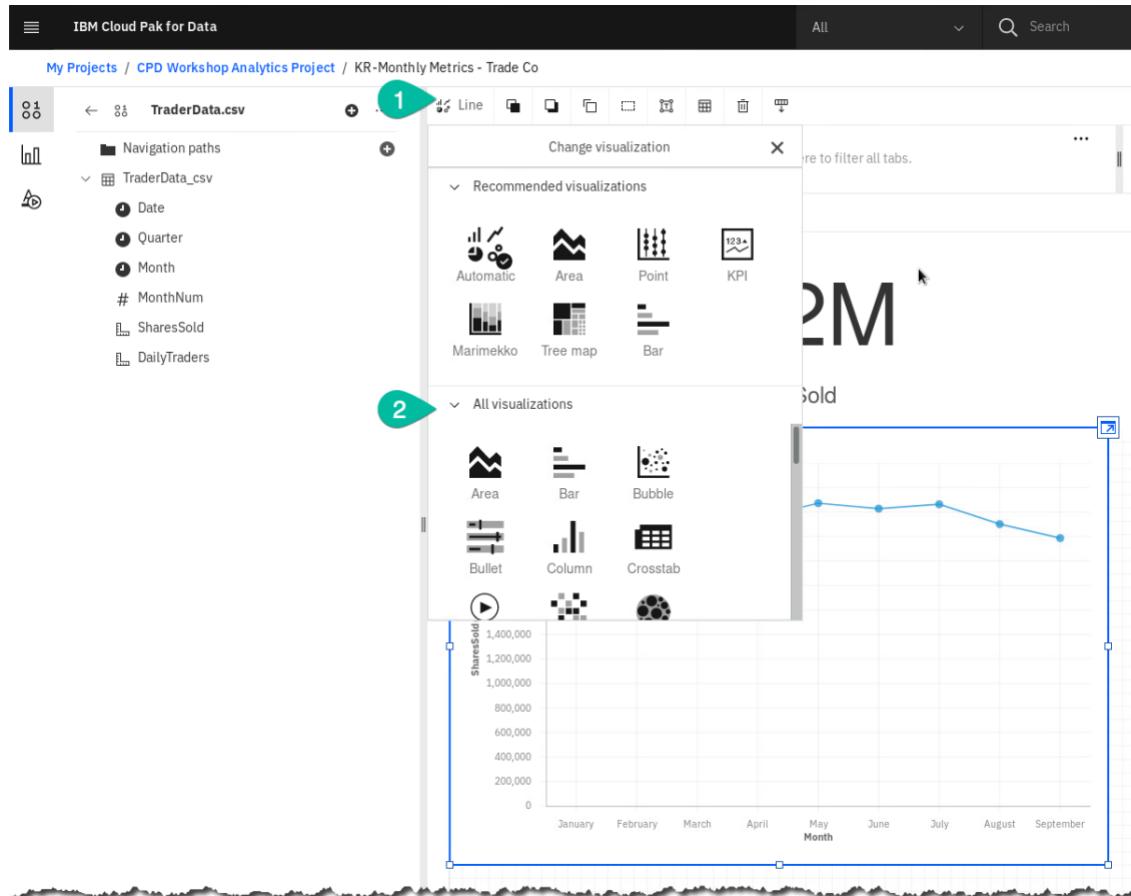


### 14.5.3 Building the line charts

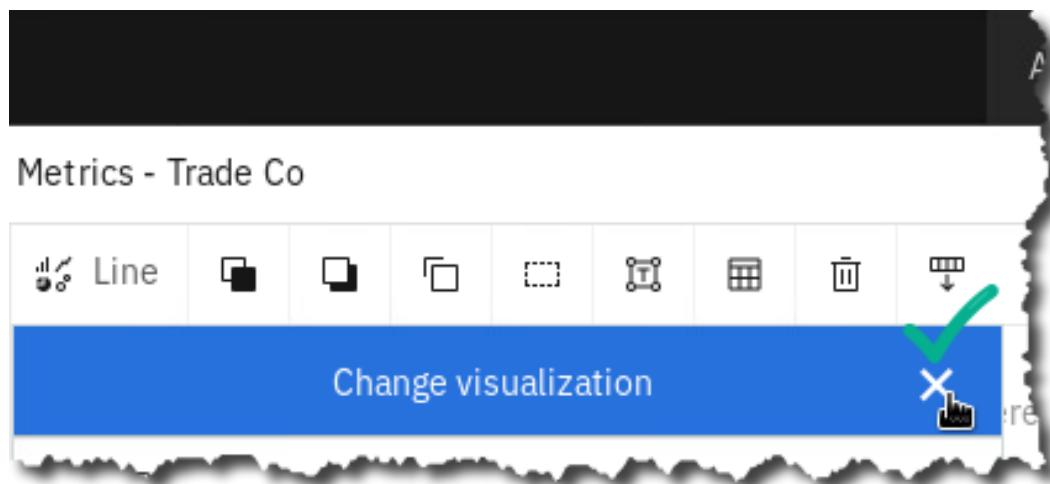
- 20. Hold the [Ctrl] key down (or Control-Option on Mac), click Month and SharesSold and drag the two onto the **lower left canvas area**. This time do not drop in the *Drop here to maximize* area.
- 21. When finished, it will automatically format like this. (Stretch the visualization wider if needed.)



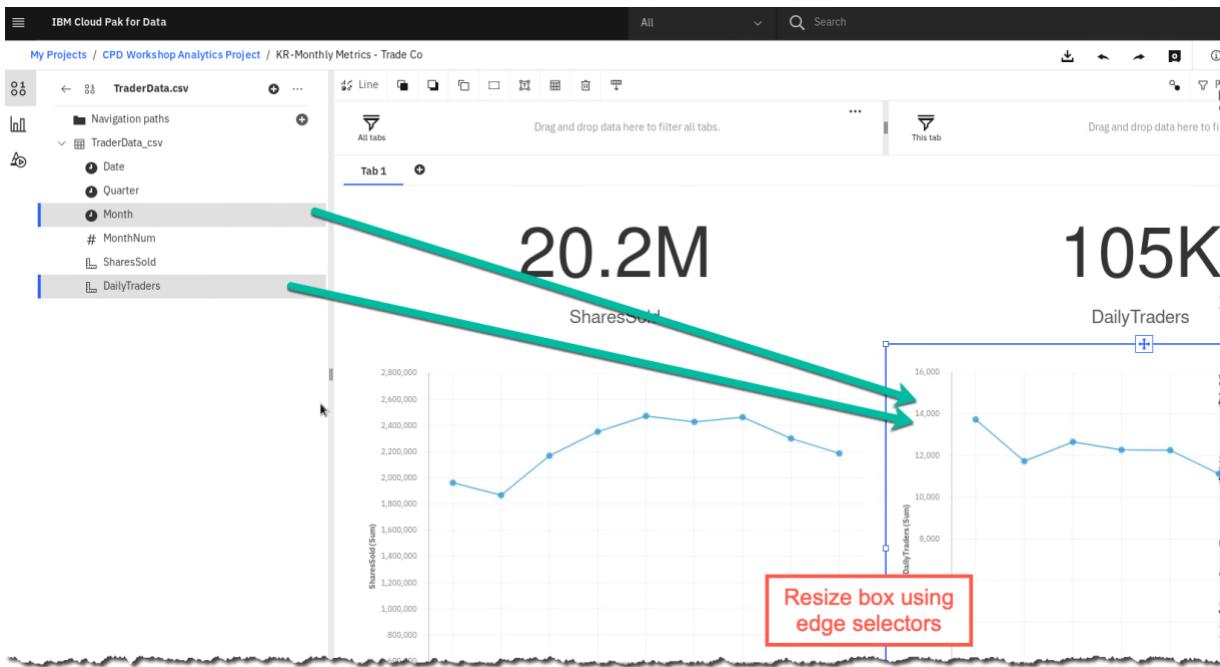
- \_\_22. Cognos Analytics uses augmented intelligence to automatically provide you with a chart and appearance that would most likely be represented by the data you select from your data source. You are not limited by this and can select other charts and graphs to represent your data. To see all the available visualizations available at your fingertips, simply select the **chart selector** (Line, in this case), and view **all the visualizations** available.



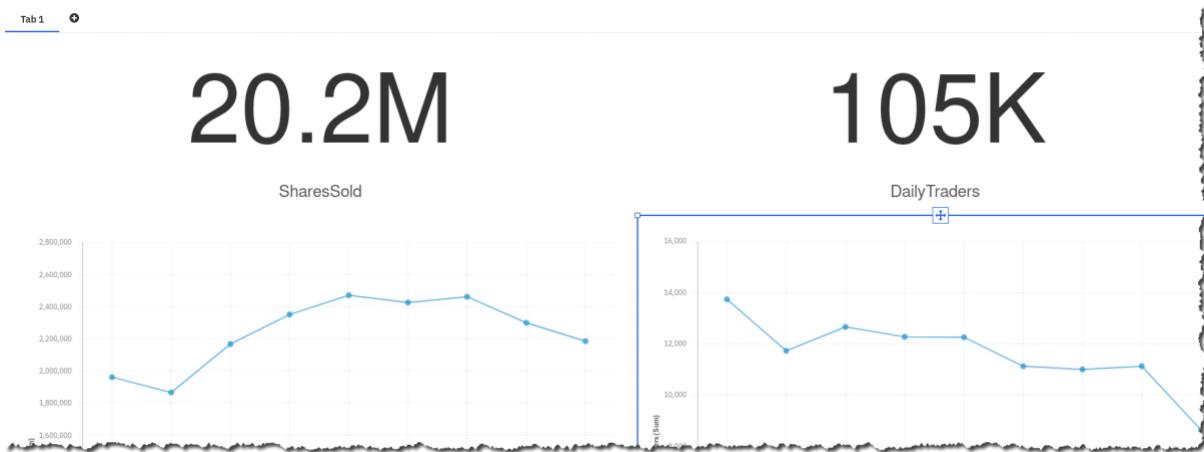
- \_\_23. Click the X to close the Change visualization screen.



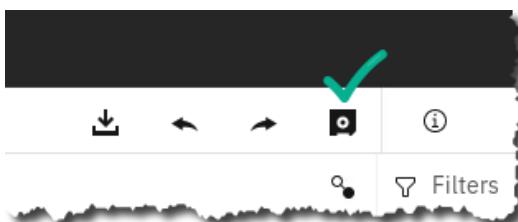
- \_24. Hold Control key down, click Month and DailyTraders and drag the two onto the lower canvas area. Again, do not drop in the *Drop here to maximize* area.



- \_25. Adjust the top, bottom, left and right of the chart boundaries so that all the boxes are aligned.  
\_26. The dashboard should display as below.

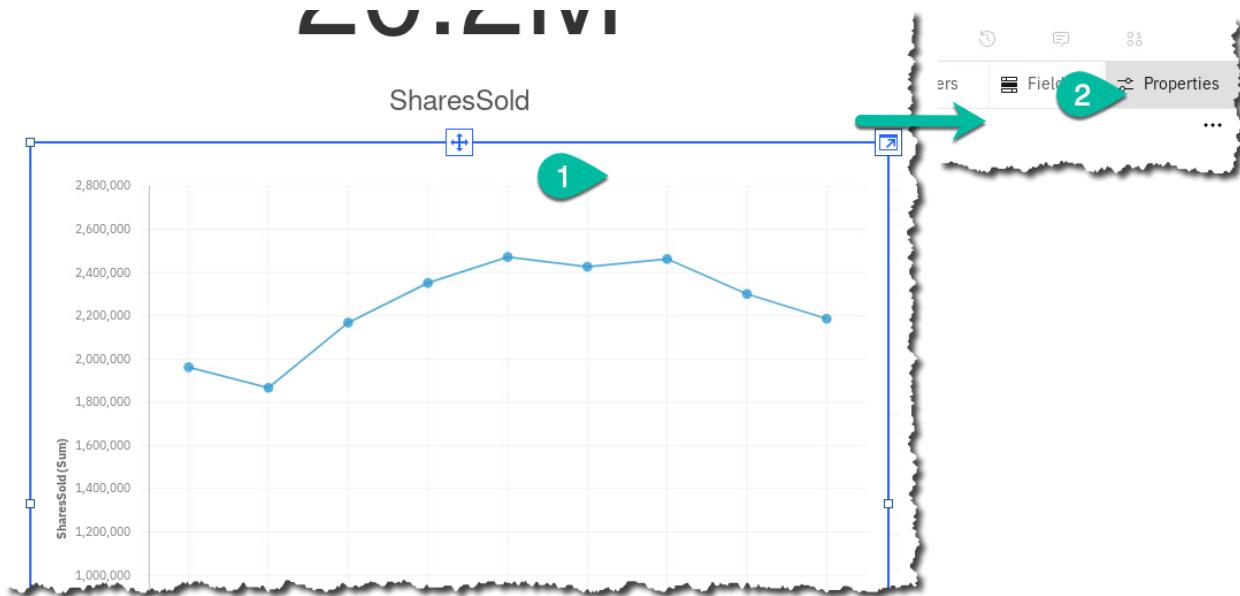


- \_27. Click the **Save** icon at the top of the screen to save your work.

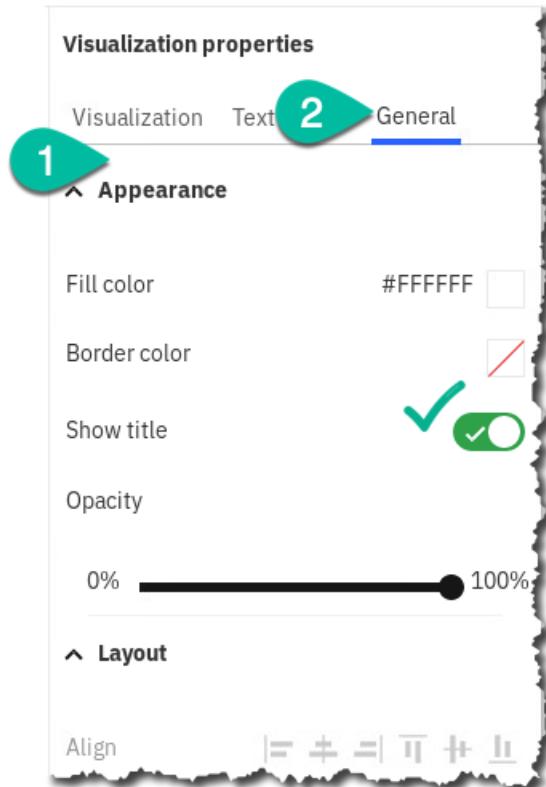


#### 14.5.4 Making the finishing touches

- 28. Select (by clicking on) the bottom left chart visualization, then select the **Properties** button at the top of the screen to format that visualization.

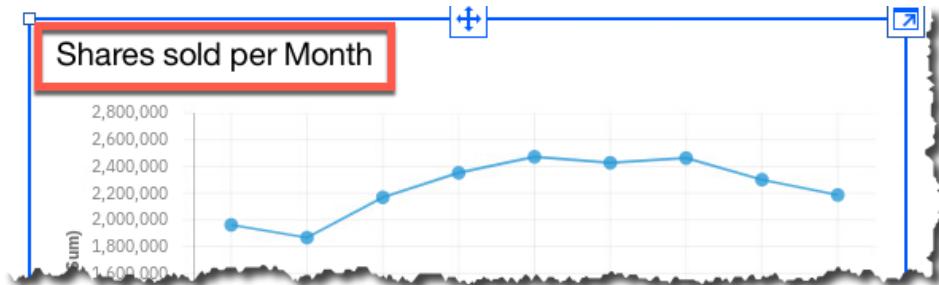


- 29. Expand **Appearance**, then from the **General** tab, check **Show title**.



\_\_30. In the bottom left visualization itself you can now type in a title for it.

Enter **Shares sold per month**.

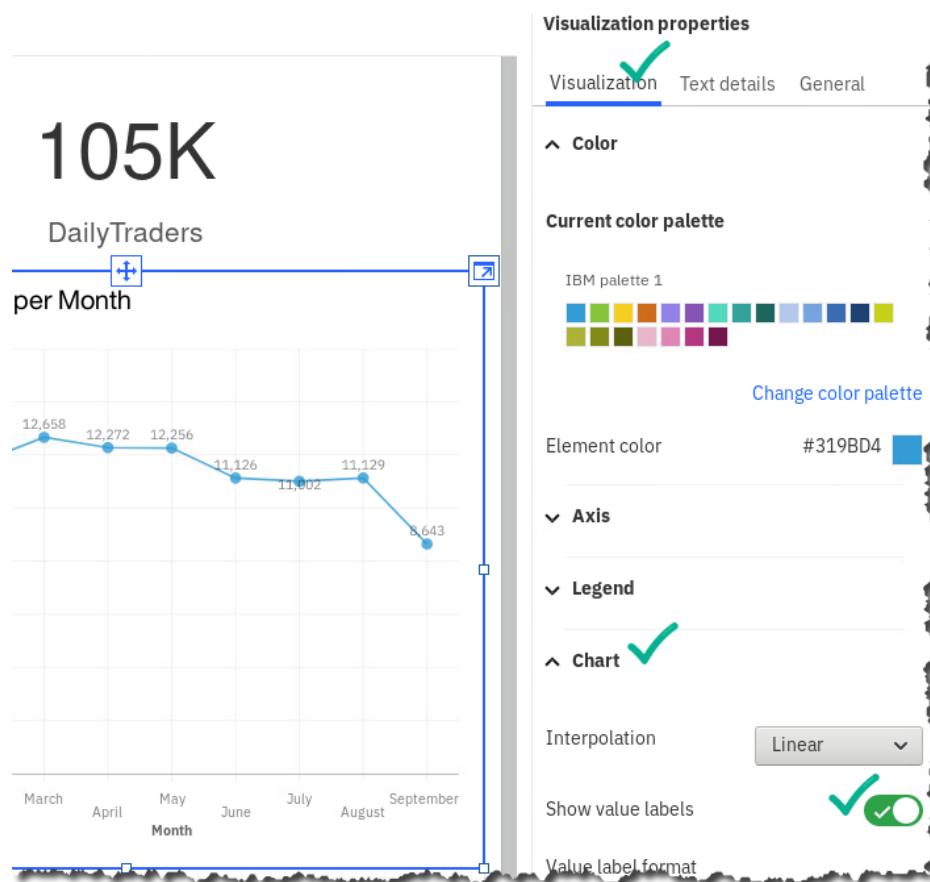


\_\_31. Click the bottom right visualization, select Properties (if not selected), Appearance, General tab, Show title, and enter **Number of Traders per month**.

The screenshot shows the "Visualization properties" panel open. It has tabs for "General", "Text", and "Appearance". A green callout "1" points to the "Appearance" tab, and another green callout "2" points to the "General" tab. Under "Appearance", the "Show title" checkbox is checked, indicated by a green checkmark. The "General" tab is selected, showing the title "Number of Traders per Month" highlighted with a red box. To the right is the chart titled "DailyTraders" showing the number of traders per month:

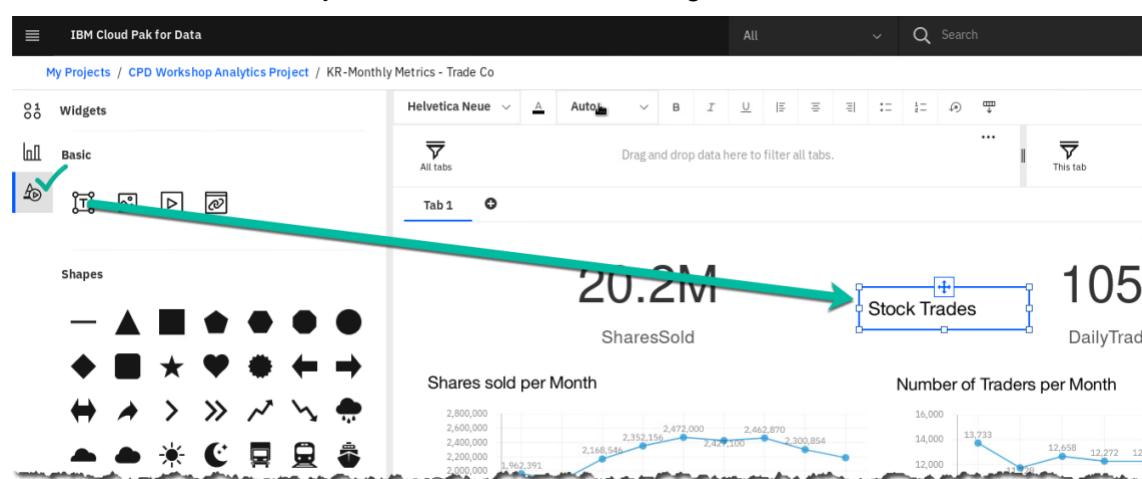
Month	DailyTraders (Sum) (Approx.)
January	13,500
February	11,500
March	12,500
April	12,000
May	12,000
June	11,000
July	11,000
August	11,000
September	8,500

- \_\_32. From the **Visualization** tab, expand **Chart**, then check **Show Value Labels**.

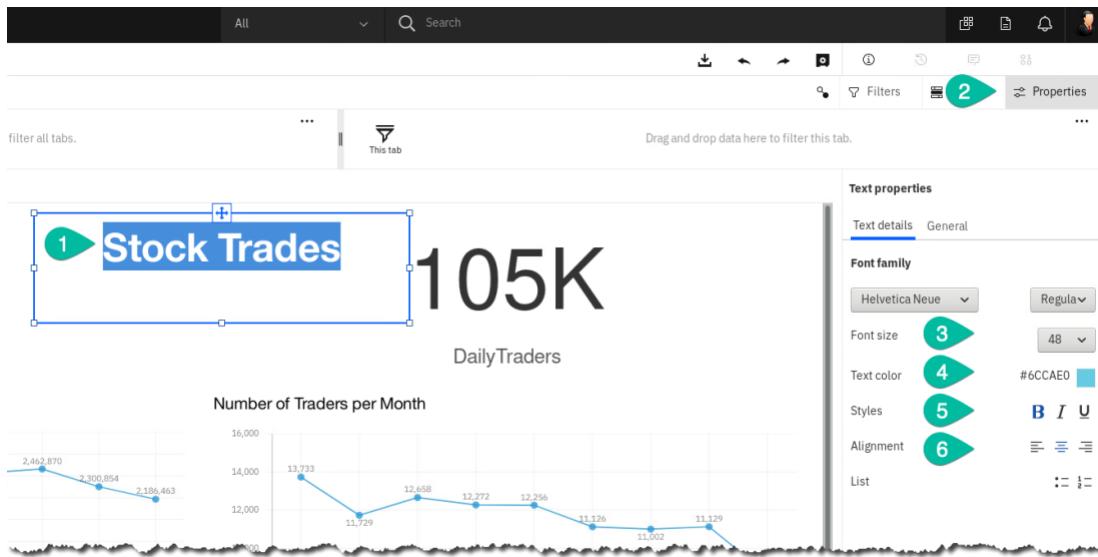


- \_\_33. Select the **Shares sold per Month** chart then check **Show Value Labels**.

- \_\_34. On the left side of the canvas, click the **Widgets** menu and drag a **Text box** between the top two visualizations.  
Title it **Stock Trades**. Adjust the box so the title is large and readable.



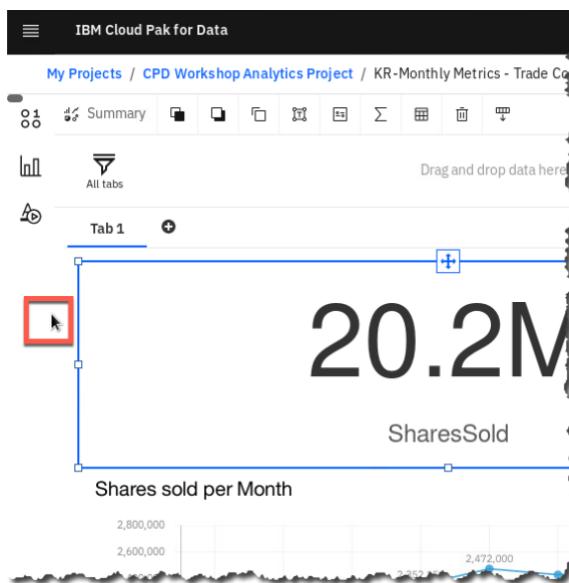
- \_\_35. Format the properties of the title **Stock Trades**. Make the text **font size 48, bold**, choose **light teal** **text color** and **center** aligned.



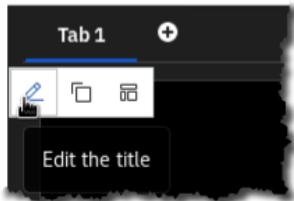
- \_\_36. Close the Widgets panel.



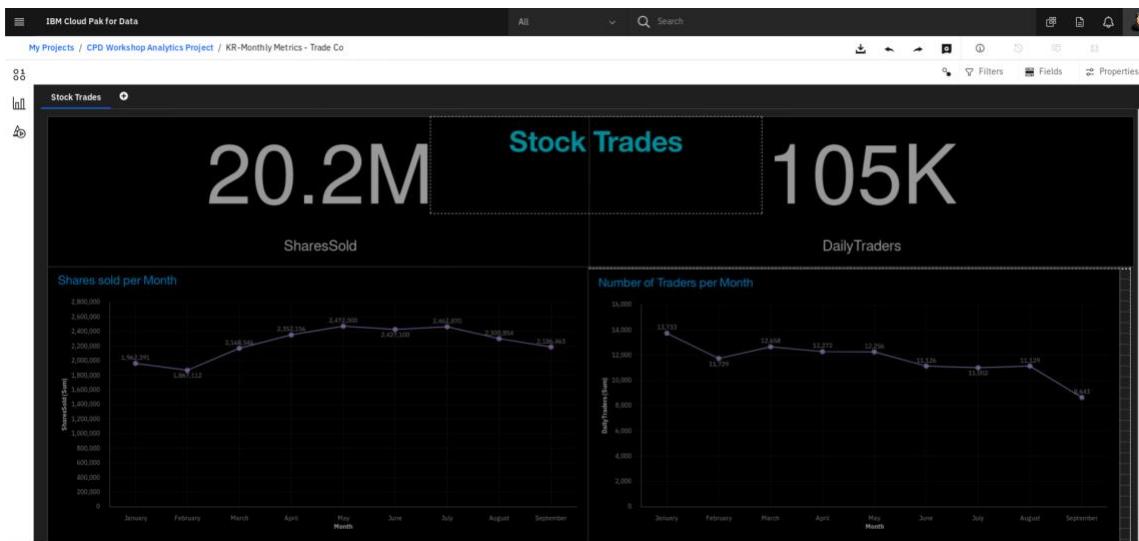
- \_\_37. Next, let's change the overall look of the dashboard to match our desired look and feel. Select somewhere outside the boxes to select the overall dashboard itself. For example, click to the left of the top left visualization as shown:



- \_\_38. This selects the entire dashboard. Select [Properties](#) again, expand [Color and theme](#) and choose [Dark](#).
- \_\_39. Close Properties by selecting [Properties](#).
- \_\_40. Select [tab](#) at top left and choose the [Pencil icon](#) to Edit the title. Give it a title of [Stock Trades](#).



- \_\_41. The final results should look something like this:



- \_\_42. Click [Save](#) again to mark your progress.

You will notice from examining the charts that [Shares sold](#) is relatively flat and daily trades are falling off. We need to use CPD to discover the WHY behind this trend.

## 14.6 Lab conclusion

Dashboard creation and analysis is just one part of the Analyze phase, but a business analyst can do this relatively quickly and simply so that they can get information from your organization's data without relying on data scientists for everything.

However, the Data Scientist can take this information and create machine learning models from the knowledge gained by this kind of analysis.

### \*\* End of Lab 14 - Cognos Dashboard Embedded – Deeper Dive

Lab by Burt Vialpando and Kent Rubin

## Lab 15 SPSS USING NPS – DEEPER DIVE

### 15.1 Lab overview

In this lab you will use SPSS Modeler (premium cartridge) to build a churn model using training data stored in Netezza Performance Server (NPS). You will also use a Jupyter notebook to build a churn model using training data stored in NPS.

These two methods for building models are distinctly different: SPSS uses a clickable paradigm, where Jupyter notebooks use coding. After building the models you will deploy them in a Cloud Pak for Data deployment space.

You will then create two types of deployments: batch and online. Finally, you will execute the deployments to complete an end-to-end example of training, deploying and using predictive models via the Cloud Pak for Data platform with NPS data.

### 15.2 Persona represented in this lab

The [Data Scientist](#) persona is the likely role to perform all the tasks shown in this lab.

Persona (Role)	Capabilities
 Data Scientist	Data Scientists bring expertise in statistics and the process of building ML/AI models to make predictions and answer key business questions.

### 15.3 Logging into the CPD web client (if you have not already done so)

- \_\_1. If you are starting this lab stand-alone (without going through previous labs) do the following:
- \_\_2. Click the desktop icon: [Cloud Pak for Data Web Client](#).



- \_\_3. The CPD web client GUI displays as shown. Use `cpduser` and `cpduser` for the *Username* and *Password* and click [Sign In](#).



## 15.4 Working with NPS Data in SPSS

- \_\_4. Click [Navigation Menu](#)  $\Rightarrow$  [Projects](#)  $\Rightarrow$  [NPS Analytics Project](#).

The screenshot shows the 'IBM Cloud Pak for Data' interface. On the left, a sidebar has 'Projects' selected, indicated by a green checkmark. A large green arrow points from this selection to the 'NPS Analytics Project' entry in the main list. The list also includes 'DataAnalysisProject' and 'AMW-DV-Project'. The interface includes a 'Filter by' dropdown set to 'All projects' and a search bar.

- \_\_5. Click [Assets](#).

The screenshot shows the 'My projects / NPS Analytics Project' screen. The 'Assets' tab is selected, indicated by a green checkmark. Below it, a search bar asks 'What assets are you looking for?' and a section titled 'Data assets' is expanded, indicated by a green arrow.

- \_\_6. Scroll down to the Modeler flows Section and click [Churn Prediction Flow](#).

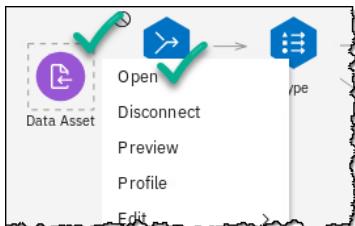
The screenshot shows the 'Modeler flows' section. It lists a single flow: 'Churn Prediction Flow', which is highlighted with a green checkmark. The table columns are 'Name' and 'Type', with 'SPSS Modeler' listed under 'Type'.

 Data Scientist	<p>Note: This SPSS Modeler flow performs data preparation, model building, and model evaluation, all without requiring coding, by using a graphical user interface.</p>
---	---

### 15.4.1 Importing data

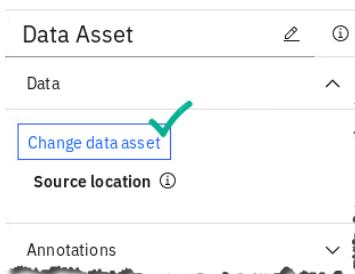
- \_\_7. You'll need to connect to three tables stored in NPS. To connect to the first table, do the following:

Mouse over the **Data Asset Import Node**  $\Rightarrow$  **Right click** (or click the **ellipsis**)  $\Rightarrow$  **Open**.

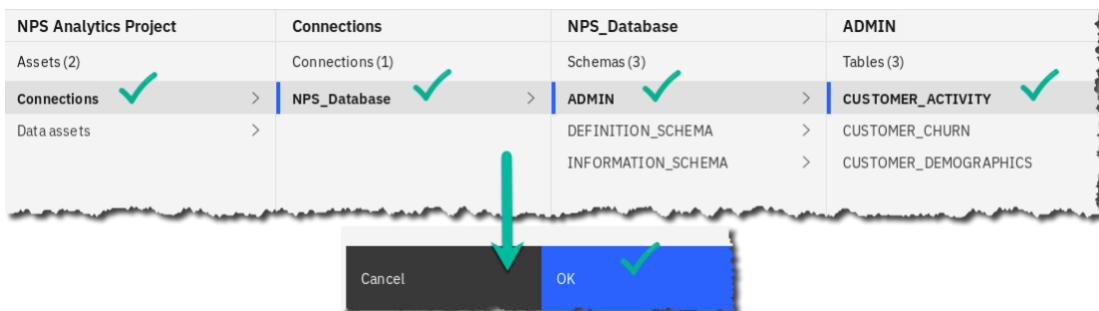


- \_\_8. On the right side of the screen a pop-up will appear.

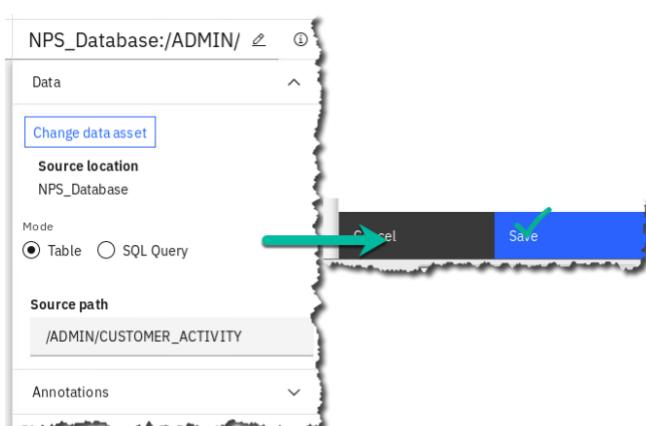
Select **Change data asset**.



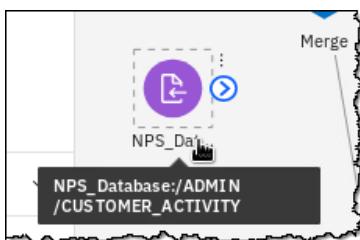
- \_\_9. Click **Connections**  $\Rightarrow$  **NPS\_Database**  $\Rightarrow$  **ADMIN**  $\Rightarrow$  **CUSTOMER\_ACTIVITY**  $\Rightarrow$  **OK**.



- \_\_10. Click **Save** in the lower right to finish setting up the Data Asset Import node to the first table in NPS.

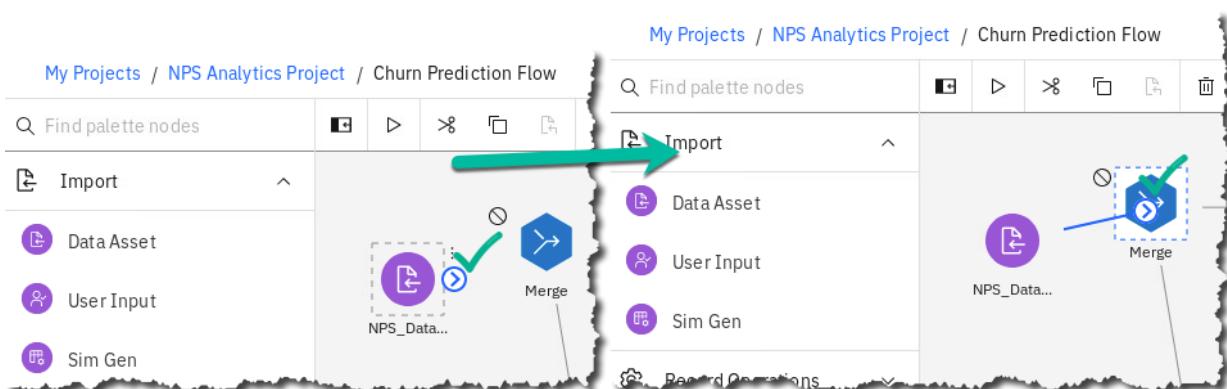


- \_11. Hover over the node on the canvas and notice it has changed its name to **NPS\_Database:/ADMIN/CUSTOMER\_ACTIVITY**.

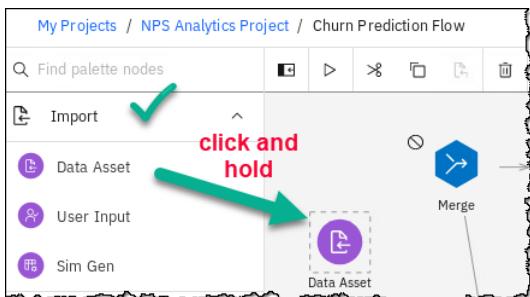


- \_12. Mouse over the **Data Asset Import** node you are working with until an arrow in a circle appears.

Click and hold the arrow while dragging to the **Merge** node located just to the right. Release the mouse button to connect the two nodes.

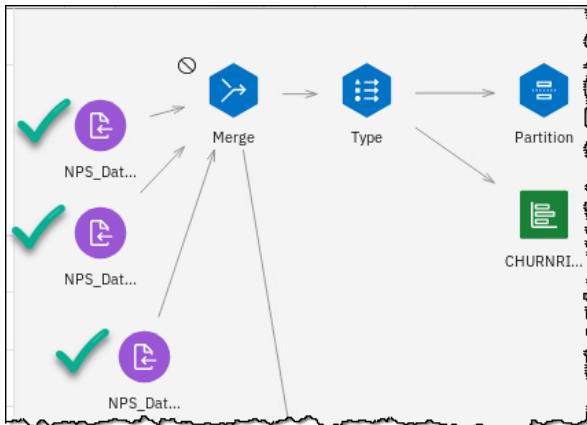


- \_13. Now create and connect to the remaining two tables by clicking the dropdown menu [Import](#), then click and hold [Data Asset](#) and drag and drop it onto the canvas to the right. Follow the [Importing Data](#) steps above to create connection for [CUSTOMER\\_CHURN](#) and [CUSTOMER\\_DEMOGRAPHICS](#).



When done, you will have three Data Asset Import nodes connected to these tables in NPS [CUSTOMER\\_ACTIVITY](#), [CUSTOMER\\_CHURN](#), [CUSTOMER\\_DEMOGRAPHICS](#).

Tip: Hover over each node to make sure you have three different tables represented by the three nodes you just created and subsequently merged.



#### 15.4.2 Testing the merged data

- \_14. To test your work, do a preview of the merged data.

On the [Merge](#) node click the [ellipsis](#)  $\Rightarrow$  [Preview](#).

At this point you have merged all the training data needed to train and test the churn model.



Click the [X](#) in the upper right of the [Data preview](#) to close the window.

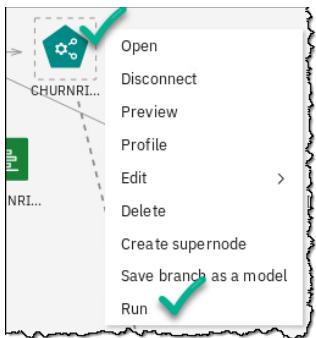
## 15.5 Building, evaluating, and saving the SPSS model

In this section you will retrain the SPSS model using the training data from NPS. Then, you will evaluate the goodness of the model. Finally, you will deploy the model and save a batch of data for testing the deployed model.

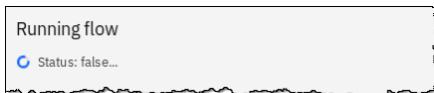
### 15.5.1 Rebuild the model

- 15. Rebuild the model by first finding the **green CHURNRISK** model building node.

Click **green CHURNRISK (node)**  $\Rightarrow$  **ellipsis**  $\Rightarrow$  **Run**.

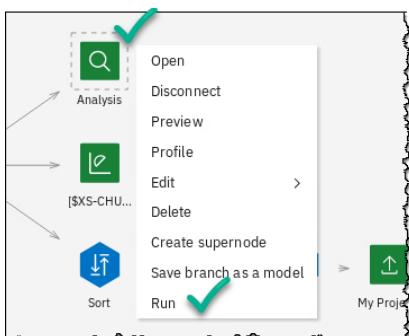


- 16. You will see the flow running...

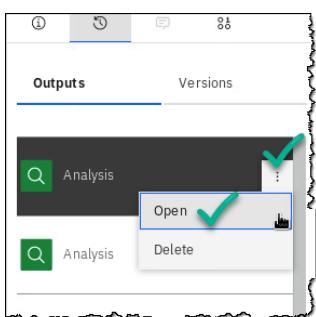


- 17. Evaluate the goodness of the model by hovering over the ellipsis on the **Analysis** node.

Click **Analysis (node)**  $\Rightarrow$  **Ellipsis**  $\Rightarrow$  **Run**.

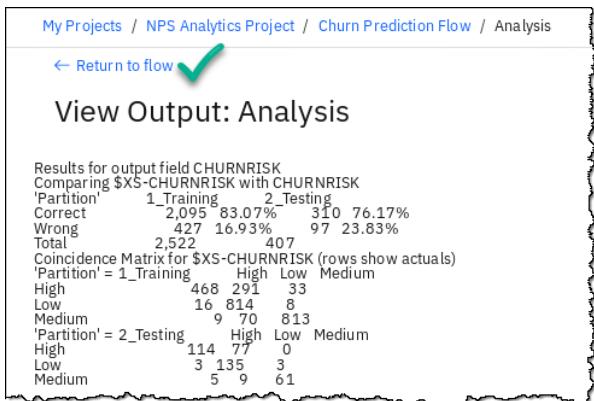


- 18. In the pop up box **Outputs** section, click **Analysis**  $\Rightarrow$  **Open**.



\_19. Review the information in [View Output: Analysis](#).

When finished, click [Return to Flow](#).

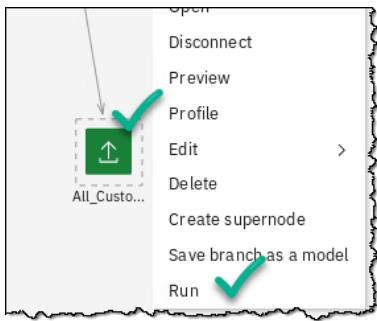


#### 15.5.2 Prepare test data

The portion of an SPSS flow that is to be deployed to Cloud Pak for Data will receive its input data from a CSV file. You will prepare for this by creating a small sample file in CSV format.

You will save this file for use by the portion of the flow to be deployed as a model.

\_20. Find the [All\\_Customer\\_Data](#) node  $\Rightarrow$  Click the ellipsis  $\Rightarrow$  Run.



\_21. The node to the immediate right of the Data Asset Export node that you just ran reads the CSV file you created and provides it as input to a copy of the Churn model you built earlier.

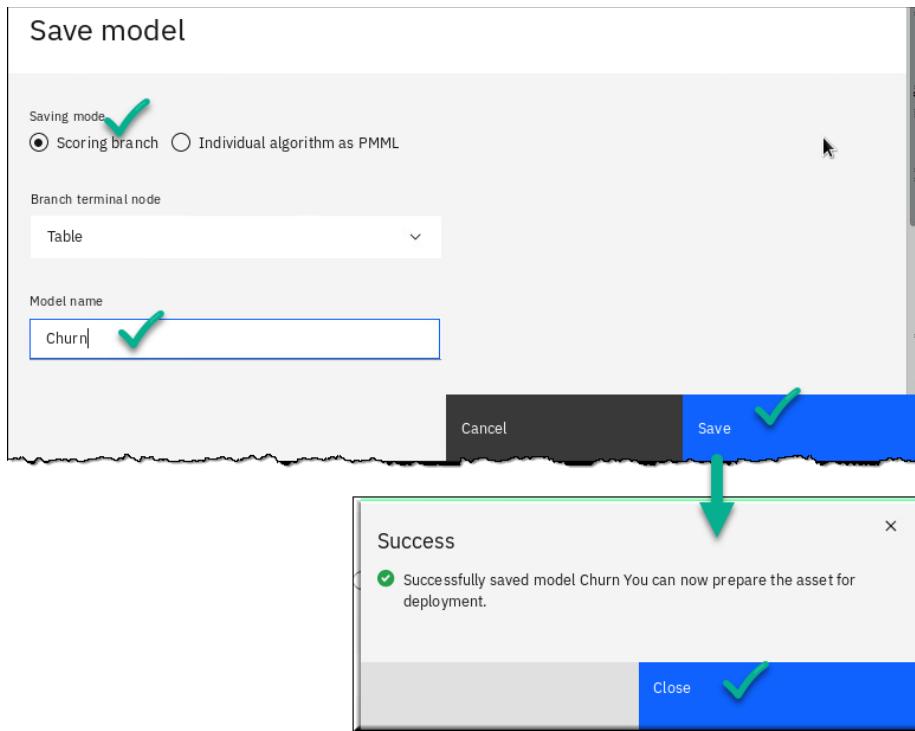
Hover over the [Table](#) node  $\Rightarrow$  Click on the ellipsis  $\Rightarrow$  Save branch as a model



\_\_22. Leave the Saving mode as Scoring branch.

Name the model **Churn**.

**Save** ⇒ **Close**.

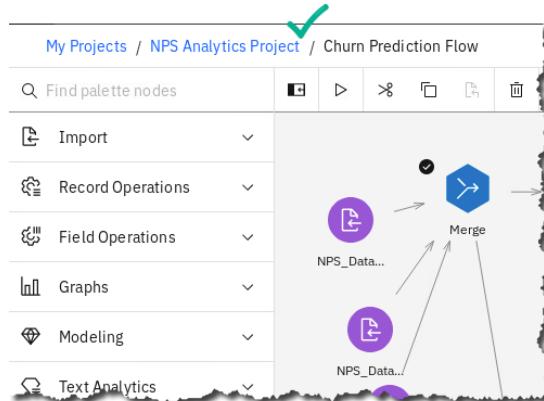


## 15.6 Creating and testing an online model deployment

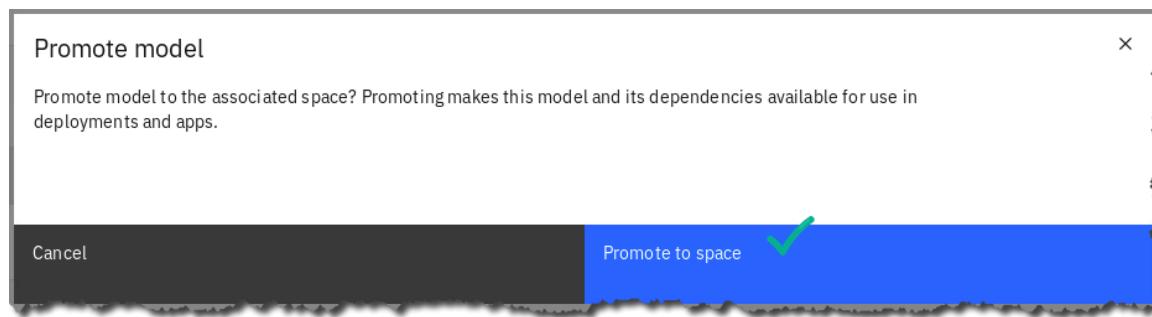
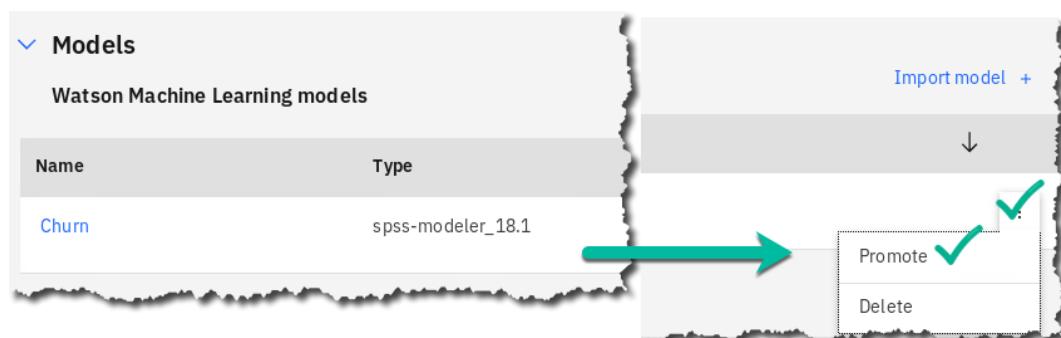
Now you will create and test an online deployment using the model you just saved.

### 15.6.1 Create the model deployment

- \_23. Click on [NPS Analytics Project](#).



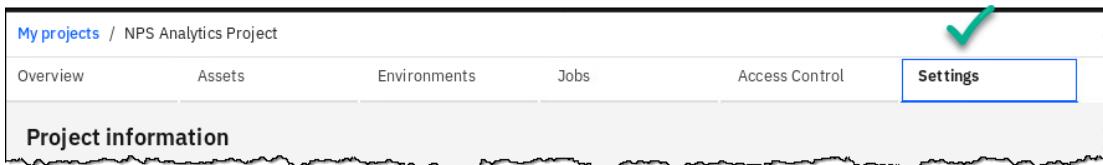
- \_24. Scroll down to the Models Section and click on the [ellipsis](#) for the Churn model  $\Rightarrow$  [Promote](#)  $\Rightarrow$  [Promote to space](#).



You will see a message near the top of your screen indicating success.

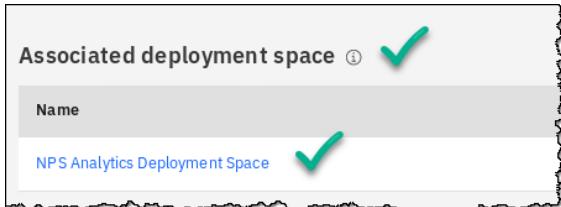


\_25. Scroll to the top of the screen and click **Settings**.



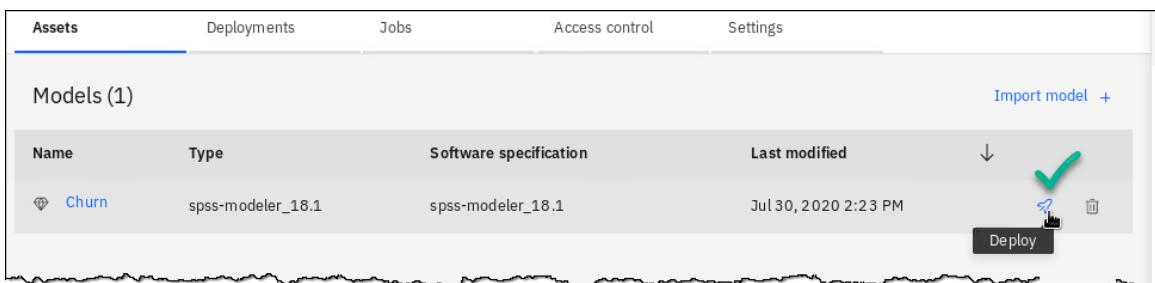
\_26. Scroll down to find section **Associated deployment space**.

Click **NPS Analytics Deployment Space**.



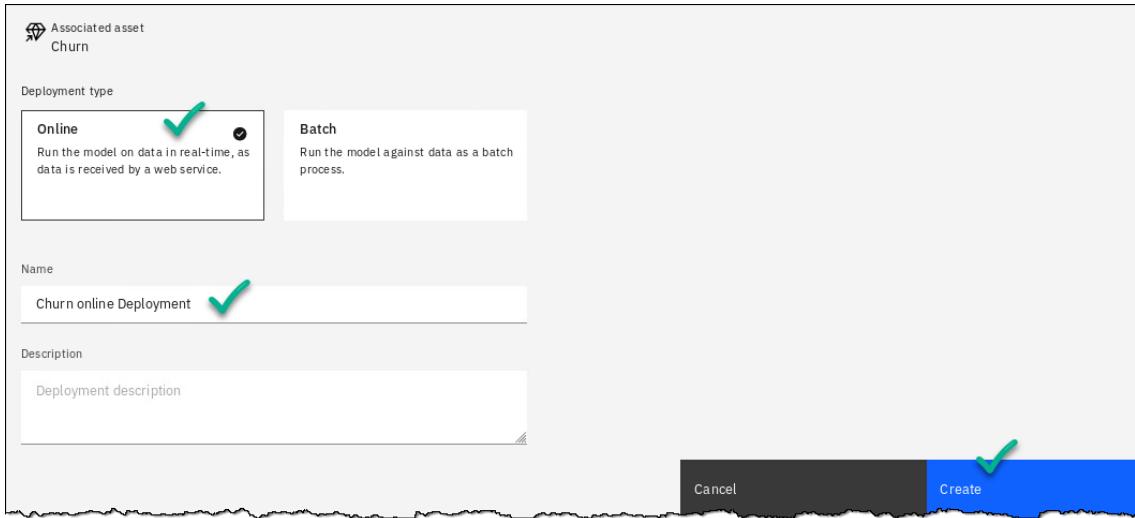
\_27. Locate the model **Churn** and **hover** at the end of the line.

Click **Deploy**.



\_28. Create an online deployment.

Select **Online** ⇒ Name it **Churn online Deployment** ⇒ **Create**.



\_29. Wait for the Status to change to Deployed.

### 15.6.2 Test the online model deployment

- \_\_30. Click [Churn online Deployment](#).

DEPLOYMENT TYPES		1 Online Deployment(s)	
Online	(1)	Name	Status
Batch	(0)	Churn online De...	Deployed

- \_\_31. Select tab [Test](#).

- \_\_32. Enter input data by first scrolling down until you can see the [ESTINCOME](#) and [AGE](#) fields, enter 60000 for [ESTINCOME](#) and 25 for [AGE](#).

Click [Predict](#) ⇒

Enter input data

Integer
ESTINCOME 60000
HOMEOWNER
String
AGE 25
<a href="#">Predict</a>

- \_\_33. Scroll down in the [Results](#) window and you'll see the churn prediction along with the confidence in the prediction.

```

56           null,
57           "High",
58           0.5496875798260724
59
]
```

## 15.7 Working with NPS Data in Jupyter Notebooks

In this section you will use a Jupyter notebook to train and deploy a predictive model. This is a coding approach and this notebook is written in Python.

The notebook is complete except for the connections to the three tables containing training data stored in NPS. You need to add the code to accomplish this. However, you'll find that the code will be automatically generated for you.

### 15.7.1 Launch the notebook

- \_34. Click [Navigation Menu](#)  $\Rightarrow$  [Projects](#)  $\Rightarrow$  [NPS Analytics Project](#).

The screenshot shows the 'Projects' screen in the IBM Cloud Pak for Data interface. On the left, there's a navigation sidebar with links for Home, Projects (which has a green checkmark), Connections, My instances, and Collect. The main area is titled 'Projects' and shows a list of projects. The first project listed is 'DataAnalysisProject', followed by 'AMW-DV-Project' and 'NPS Analytics Project' at the bottom. A green checkmark is placed next to 'DataAnalysisProject'. A large green arrow points from the 'Projects' link in the sidebar to the 'DataAnalysisProject' entry in the list.

- \_35. Click [Assets](#).

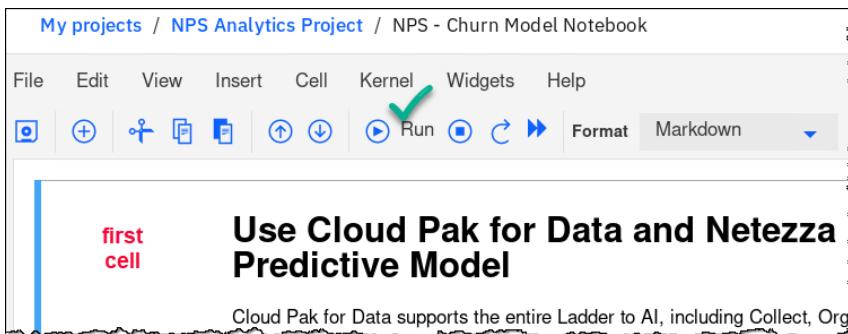
The screenshot shows the 'Assets' screen for the 'NPS Analytics Project'. At the top, there are tabs for 'Overview' and 'Assets' (which has a green checkmark). Below the tabs is a search bar with the placeholder 'What assets are you looking for?'. Underneath the search bar, there's a section titled 'Data assets' with a green checkmark next to it.

- \_36. Scroll down to the [Notebooks Section](#).

Click the [Pencil icon](#) next to the Notebook named [NPS - Churn Model Notebook](#).

The screenshot shows the 'Notebooks' section. It has a header with a green checkmark next to 'Notebooks'. Below the header is a table with columns: Name, Shared, Scheduled, Status, Language, Last editor, and Last modified. There is one row in the table, representing the 'NPS - Churn Model Notebook'. The 'Name' column contains the notebook name. The 'Last modified' column shows the date 'Jul 29, 2020'. To the right of the last column, there are three icons: a green checkmark, a pencil, and a blue question mark.

- \_\_37. You will start by selecting the first cell by [clicking](#) on it  $\Rightarrow$  Read the contents of the cell then click [Run](#).



You can tell when the notebook is running a cell by observing a solid black circle in the upper right-hand corner of the screen. It turns to an open circle when the notebook has completed its run. After any individual cell has run, any output from the code is displayed immediately after that cell, and the next cell is automatically selected for you.



If you receive a 403: Forbidden error restart the notebook by finding the information icon at the top right of the screen and clicking as shown below.

403 : Forbidden

Starting runtime for NPS - Churn Model Notebook  
The selected runtime has 1 vCPU and 2 GB RAM.

A screenshot of the Jupyter Notebook environment settings. It shows the "Information" tab with "General" and "Environment" tabs. The "Environment" tab is active, showing "Environment definition: Default Python 3.6". Below that is the "Runtime status" section, which shows a dropdown menu set to "Running". A green arrow points to the "Running" status. Another green arrow points to the "Restart" button in the dropdown menu.

Data Scientist

### 15.7.2 Load data into the notebook

- 38. Continue reading and running each cell until you reach the three cells containing the text.

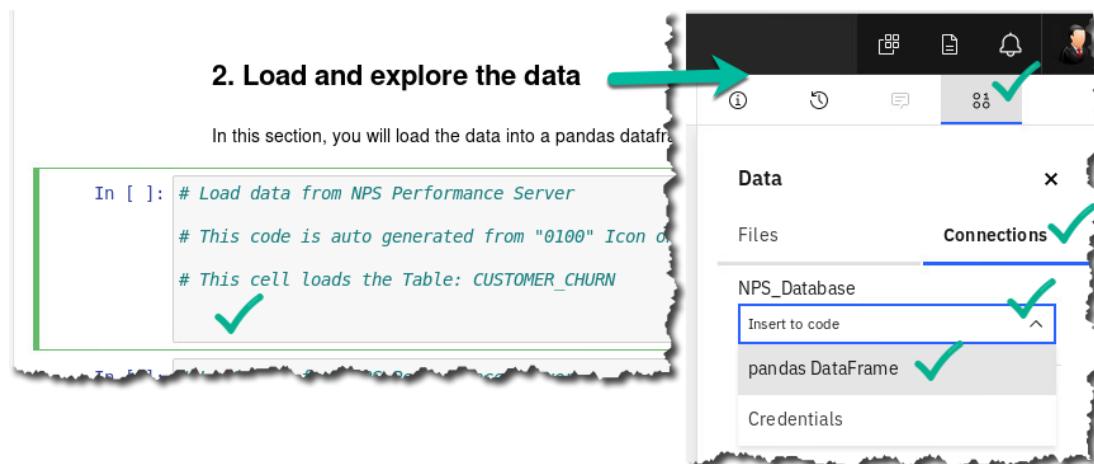
```
# Load data from NPS Performance Server
```

**2. Load and explore the data**

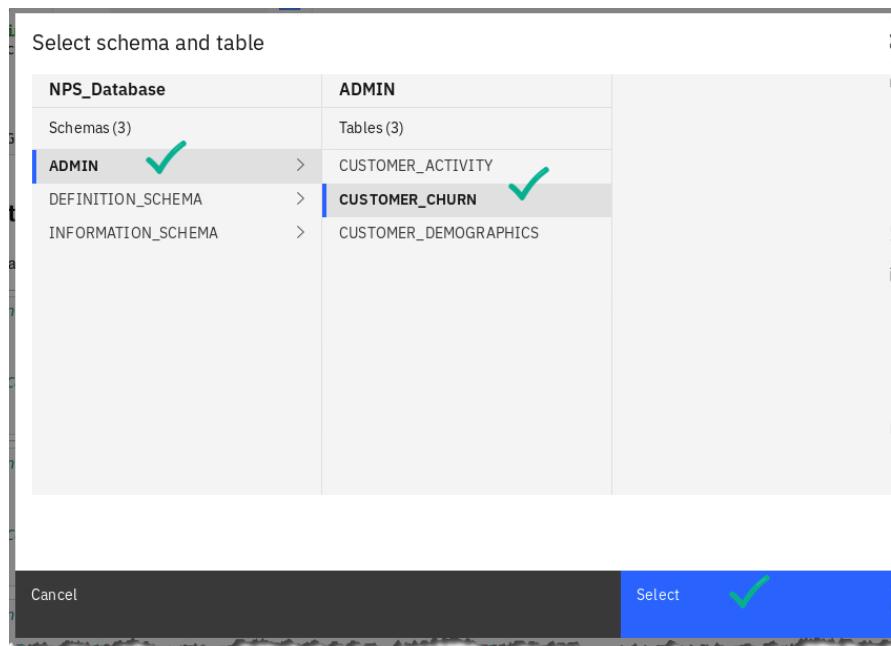
In this section, you will load the data into a pandas dataframe and perform an exploratory data analysis.

```
# Load data from NPS Performance Server
```

- 39. Position and **Click** your cursor on the last blank line inside the first of the three data cells under the notebook Section named “Load and explore the data.” Then, Click **Data** (the 0100 Icon)  $\Rightarrow$  **Connections**  $\Rightarrow$  **Insert to code**  $\Rightarrow$  **Insert pandas DataFrame**.

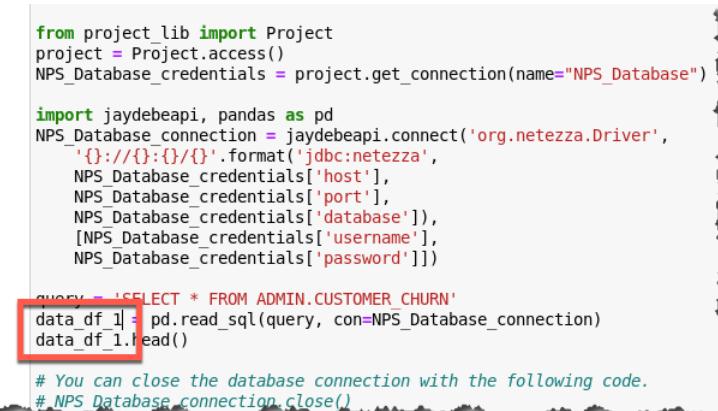


- 40. In the pop-up to Select schema and table, Click **ADMIN**  $\Rightarrow$  **CUSTOMER\_CHURN**  $\Rightarrow$  **Select**. This will insert the code needed to load the data table from NPS into the notebook.



- \_\_41. Near the end of the inserted code there are two lines of code that start with `data_df_#`, where the # is a number. If the number is not equal to **1**, then change it to **1** in both lines of code.

`data_df_4`  $\Rightarrow$  `data_df_1`



```

from project_lib import Project
project = Project.access()
NPS_Database_credentials = project.get_connection(name="NPS_Database")

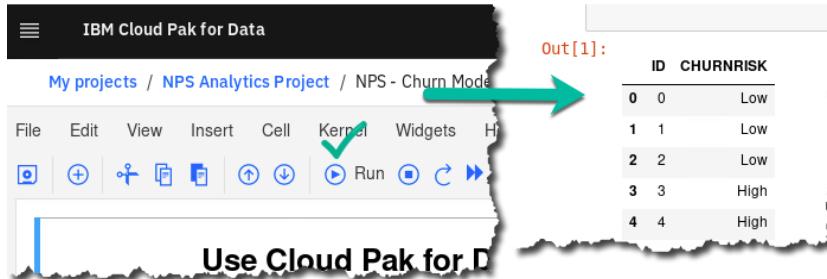
import jaydebeapi, pandas as pd
NPS_Database_connection = jaydebeapi.connect('org.netezza.Driver',
    '{}://{}:{}{}'.format('jdbc:netezza',
    NPS_Database_credentials['host'],
    NPS_Database_credentials['port'],
    NPS_Database_credentials['database'],
    [NPS_Database_credentials['username'],
    NPS_Database_credentials['password']]))

QUERY = 'SELECT * FROM ADMIN.CUSTOMER_CHURN'
data_df_1 = pd.read_sql(query, con=NPS_Database_connection)
data_df_1.head()

# You can close the database connection with the following code.
# NPS_Database_connection.close()

```

- \_\_42. Click **Run** to load the table into the notebook and display the first few rows.



IBM Cloud Pak for Data

My projects / NPS Analytics Project / NPS - Churn Model

File Edit View Insert Cell Kernel Widgets Help

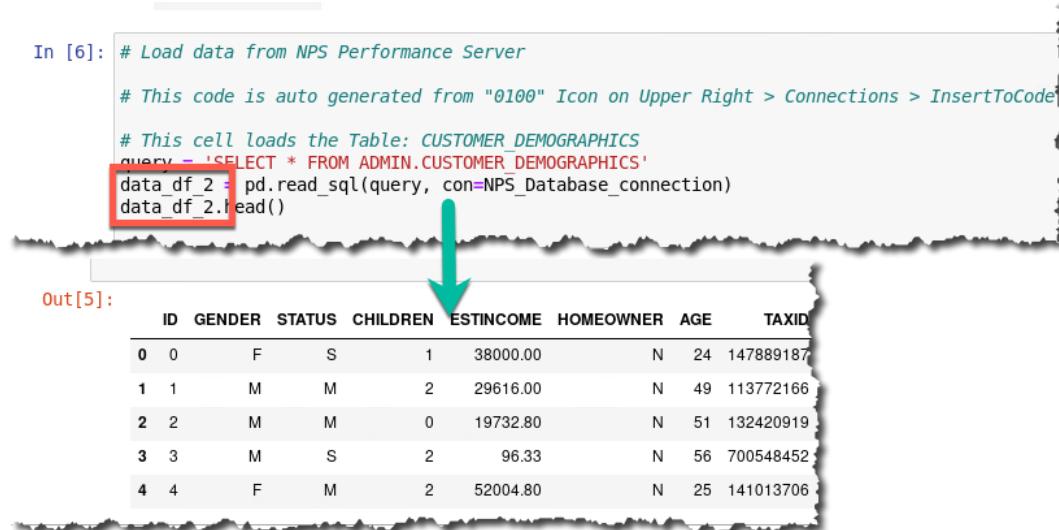
Run

Out[1]:

ID	CHURNRISK
0	Low
1	Low
2	Low
3	High
4	High

- \_\_43. Now you should be on the cell that loads the `CUSTOMER_DEMOGRAPHICS` table from NPS. To do so, repeat steps from the top of this sub-section [Loading data into the Notebook](#).

Be sure to select the `CUSTOMER_DEMOGRAPHICS` table and use `data_df_2` as the DataFrame variable name. Note that the code to load this table is shorter than the code to load the first table. This is because the code for the first table included loading the database credentials, which is only required once.



In [6]: # Load data from NPS Performance Server

# This code is auto generated from "0100" Icon on Upper Right > Connections > InsertToCode

# This cell loads the Table: CUSTOMER DEMOGRAPHICS

query = 'SELECT \* FROM ADMIN.CUSTOMER\_DEMOGRAPHICS'

data\_df\_2 = pd.read\_sql(query, con=NPS\_Database\_connection)

data\_df\_2.head()

Out[5]:

ID	GENDER	STATUS	CHILDREN	ESTINCOME	HOMEOWNER	AGE	TAXID
0	F	S	1	38000.00	N	24	147889187
1	M	M	2	29616.00	N	49	113772166
2	M	M	0	19732.80	N	51	132420919
3	M	S	2	96.33	N	56	700548452
4	F	M	2	52004.80	N	25	141013706



Data Scientist

Note: The code to load this table is shorter than the code to load the first table. This is because the code for the first table included loading the database credentials, which is only required once.

- 44. Now you should be on the cell that loads the third and final table, CUSTOMER\_ACTIVITY, from NPS. To do so repeat the steps again from sub-section [Loading data into the notebook](#).

The only changes are to make sure to select the `CUSTOMER_ACTIVITY` table and use `data_df_3` as the DataFrame variable name.

```
In [7]: # Load data from NPS Performance Server
# This code is auto generated from "0100" Icon on Upper Right > Connections > InsertToCode
# This cell loads the Table: CUSTOMER_ACTIVITY
query = 'SELECT * FROM ADMIN.CUSTOMER_ACTIVITY'
data_df_3 = pd.read_sql(query, con=NPS_Database_connection)
data_df_3.head()
```

ID	TOTALDALLARVALUETRADED	TOTALUNITSTRADED	LARGESTSINGLETRANSACTION	S
0	593	1184043.01	488	9595.18
1	702	1301994.91	520	9420.20
2	1115	1207386.21	479	9395.26
3	1780	1194102.87	521	9158.42
4	1286	1272265.01	501	9518.61

- 45. Continue running the notebook one cell at a time until you reach the end. Near the end of the notebook a new deployment space will be created and the model that you've trained will be deployed as an online deployment. You'll work with that deployed model in the next Section.

```
'id': '211e6bbf-4c2a-4b45-9e79-055d2a7ecd7e',
'modified_at': '2020-07-29T16:56:27.636Z',
'name': 'Churn-Model-Deployment',
'parent': {'href': ''},
'space_id': 'ccfeeee5b-79c1-47ef-bcf1-eac8ba35048f',
'tags': ['Churn-Model-Deployment-Tag']}
```

```
In [50]: payload = [{"values": [1, 38000, 24, 1200000, 509, 9400, 940, 51, 3, 10, 0, -81000]}]
```

```
In [51]: payload_metadata = {client.deployments.ScoringMetaNames.INPUT_DATA: payload}
# score
predictions = client.deployments.score(scoring_deployment_id, payload_metadata)
predictions
```

```
Out[51]: {'predictions': [{'fields': ['prediction', 'probability'],
  'values': [[0, [0.6454002261161804, 0.3545997738838196]]]}]}
```

## 5. Conclusion

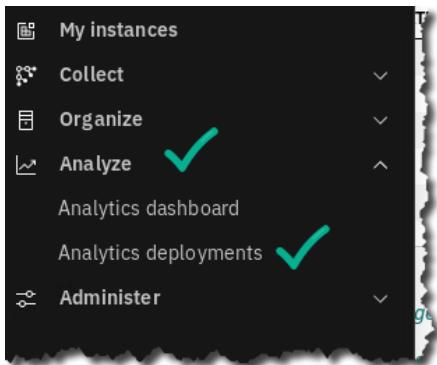


You have successfully completed this notebook!

## 15.8 Working with Batch Deployments

In this section we'll work with the model you saved from the Jupyter notebook.

- 46. Click [Navigation Menu](#)  $\Rightarrow$  [Analyze](#)  $\Rightarrow$  [Analytics deployments](#).



- 47. Click [Deployment-Space-Created-From-Notebook](#)  $\Rightarrow$  [Churn-Model](#)  $\Rightarrow$  [Create deployment](#)  $\Rightarrow$  [Batch](#)  $\Rightarrow$  name it [Churn Batch Deployment](#)  $\Rightarrow$  select [One Standard CPU](#)  $\Rightarrow$  Click [Create](#)

Analytics deployment spaces

The screenshot shows the 'Deployment-Space-Created-From-Notebook' interface. It includes a search bar, a 'Name' field containing 'Deployment-Space-Created-From-Notebook', and tabs for 'Assets', 'Deployments', 'Jobs', and 'Access control'. Under the 'Models (1)' section, there is a table with one row for 'Churn-Model'. A green arrow points to the 'Create deployment' button in the top right corner of this section. The next screen shows the 'Create deployment' dialog. It has sections for 'Deployment type' (with 'Batch' selected), 'Hardware definition' (with '1 standard CPU, 4 GB RAM' selected), and a 'Name' field containing 'Churn Batch Deployment'. A final green arrow points to the 'Create' button at the bottom of the dialog.

—48. Before creating the batch job, we need to get some input data.

Click **Navigation Menu** ⇒ **Click Projects** ⇒ **Click NPS Analytics Project** ⇒ Under Data assets click **Payload to test deployment.json** ⇒ Carefully copy the entire contents of the json file (making sure to highlight everything).

The first screenshot shows the 'IBM Cloud Pak for Data' navigation menu with 'Projects' selected. A green arrow points to the 'Name' filter field in the search bar above the project list. The second screenshot shows the 'Data assets' section with 'Payload to test deployment.json' selected. A green checkmark is next to it. The third screenshot shows the 'Copy' button highlighted in red in the 'Project / Payload to test deployment.json' interface. A green arrow points from the selected JSON file to the 'Copy' button.

—49. Click **Navigation Menu** ⇒ **Analyze** ⇒ **Analytics deployments** ⇒ **Deployment-Space-Created-From-Notebook** ⇒ **Deployments** ⇒ **Churn Batch Deployment**.

The first screenshot shows the 'My instances' section of the navigation menu with 'Organize' and 'Analyze' selected. A green arrow points to the 'Analytics deployments' link. The second screenshot shows the 'Deployment-Space-Created-From-Notebook' interface with 'Deployment-Space-Created-From-Notebook' selected. A green checkmark is next to it. The third screenshot shows the 'Deployment-Space-Created-From-Notebook' interface with the 'Deployments' tab selected. It lists two deployments: 'Churn Batch Deployment' (Batch, Deployed) and 'Churn-Model-Deployment' (Online, Deployed).

- \_\_50. Click **Create Job** ⇒ name the job **Churn Batch Job** ⇒ **Specify input data** ⇒ Click **Inline input data** ⇒ Paste the JSON data you copied in the previous step ⇒ **Confirm** ⇒ **Create and Run**.



\_\_51. Let the job complete (takes about 20 seconds).

Click the [start time for the job run](#).

Click [Show More](#) ⇒ scroll to the bottom of the log to view the prediction and probability.

Start Time	Status
Aug 05, 2020 7:33:54 PM	Completed

```

{
  "run_id": "ccfeeee5b-79c1-47ef-bcf1-eac8ba35048f",
  "start_time": "2020-07-29T17:19:34.190670Z",
  "end_time": "2020-07-29T17:19:34.170938Z",
  "status": "completed",
  "model_version": "v1",
  "model_name": "Churn Model",
  "model_type": "Classification",
  "model_subtype": "Logistic Regression",
  "model_architecture": "Linear",
  "model_parameters": {
    "C": 1.0,
    "penalty": "l2",
    "solver": "lbfgs",
    "max_iter": 100
  },
  "model_coefficients": [
    {
      "feature": "Satisfaction Score",
      "coefficient": 0.6454002261161804
    },
    {
      "feature": "Tenure (in months)",
      "coefficient": 0.3545997738838196
    }
  ],
  "model_intercept": -81000.0,
  "model_accuracy": 0.85,
  "model_f1_score": 0.82,
  "model_precision": 0.88,
  "model_recall": 0.78,
  "model_roc_auc": 0.9
}
  
```

[Show more](#)

## 15.9 Lab Conclusion

In this lab, you used tables stored in NPS as a source of training data for building churn models.

Using SPSS Modeler, part of the SPSS Modeler Premium Cartridge in Cloud Pak for Data, you built a churn model using the clickable paradigm for modeling. You deployed the model for online use and tested the deployment.

Using a Jupyter notebook written in Python you built another churn model using the coding paradigm for modeling. You deployed this model programmatically from the notebook and manually from the deployment space. You then tested the batch deployment using a JSON payload.

### \*\* End of Lab 15 – SPSS using NPS – Deeper Dive

Lab by Tom Konchan, Edited by Burt Vialpando, John Lucas and Kent Rubin

## Lab 16 NETEZZA PERFORMANCE SERVER (NPS) GETTING STARTED

### 16.1 Lab overview

This lab will introduce you to the Netezza Performance Server (NPS) Data Warehouse service. We will also refer to NPS as Netezza. You will learn how to access `nzsql` from the Netezza command line tool and perform these basic operations:

- Create a database
- Create a table
- Load a table
- Unload a table

These tasks provide an initial experience of interacting with the database engine. Other labs (outside of this workshop) are available to provide users experience with the advanced features of Netezza.

### 16.2 Understanding Netezza Users

The NPS environment consists of two types of users - operating system users and database users. Operating system users can access the command line on the Netezza Linux host containers. Database users can access NPS to perform database-related tasks.

The Netezza Linux Container has the following default users:

Users	Capabilities
 root	Linux user with login permission to NPS Has ownership to various NPS files in the installation directory Usually not exposed to client users
 nz	Linux user with login permission to NPS Has ownership to various NPS files in the installation directory Access to the Netezza command tools
 admin	Database super-user (not a Linux user) Full access to all database administration privileges and object

## 16.3 Connecting to the Netezza Performance Server Command Line

To access the Netezza host Linux container command line, you will use the Terminal application available on the Unified Desktop.

The Netezza host Linux container is running the service `sshd` which is configured to allow remote secure login. In our environment `ssh` is listening on port `22`.

 nz	<p>Note: In Cloud Pak for Data System with Netezza Performance Server, <code>sshd</code> is listening on port <code>51022</code>. In Netezza on Cloud <code>ssh</code> is disabled as the cloud solution provides other mechanisms to access the command line tools.</p>
--	--

### 16.3.1 SSH to the Netezza Host Container using the Terminal

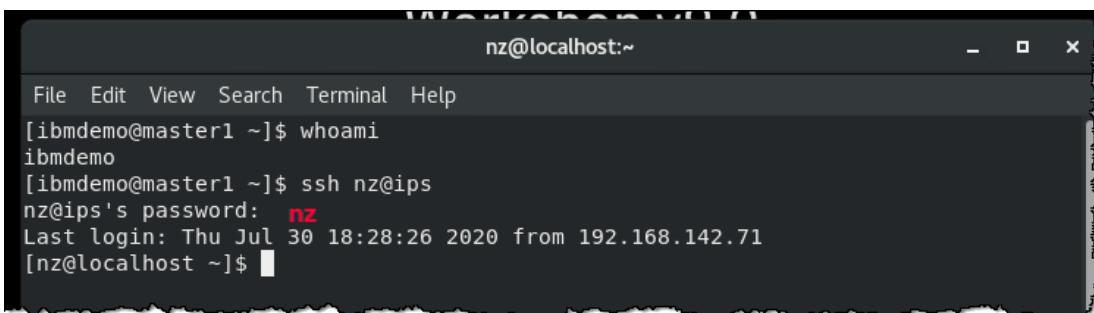
- \_\_1 Double-click the `Terminal` desktop icon.



 nz	<p>Note: You can be user <code>ibmdemo</code> to ssh into node <code>ips</code>.</p>
--	--

- \_\_2 Next, `ssh` into the Netezza Linux host container:

```
$ whoami
$ ssh nz@ips (password = nz)
```



```
nz@localhost:~
File Edit View Search Terminal Help
[ibmdemo@master1 ~]$ whoami
ibmdemo
[ibmdemo@master1 ~]$ ssh nz@ips
nz@ips's password: nz
Last login: Thu Jul 30 18:28:26 2020 from 192.168.142.71
[nz@localhost ~]$
```

- \_\_3 You are now logged into the Netezza host Linux container as user `nz`.

 nz	<p>Note: In this lab the root user is not needed on the Netezza host Linux container.</p> <p>For reference, the <code>root</code> password is <code>netezza</code>.</p>
--	---

## 16.4 Checking the state of NPS

- \_\_4 Check the state of the Netezza engine by running `nzstate`.

```
$ nzstate
```

```
Last login: Thu Jul 30 18:31:42 2020 from 192.168.142.71
[nz@localhost ~]$ nzstate
System state is 'Online'.
[nz@localhost ~]$ █
```

If the Netezza system is offline (Stopped) then start Netezza with the `nzstart` command.



```
nz@localhost:~$ File Edit View Search Terminal Help
[nz@localhost ~]$ nzstate
System state is 'Stopped'.
[nz@localhost ~]$ nzstart
nzstart: Warning: Using auto-generated topology: /tmp/initTopology.autodb.sim.cfg

(startupsrv) Info: NZ-00022: --- program 'startupsrv' (25246) starting on host 'localhost.localdomain' ... ---
[nz@localhost ~]$ nzstate
System state is 'Online'.
[nz@localhost ~]$ █
```

## 16.5 Setting up the lab database

- \_\_5 Execute the script to set up the database used in this lab:

```
$ cd ./labs/base/cli/setupLab/
$ ls -l
$ ./setupLab.sh
```

```
nz@localhost:~/labs/base/cli/setupLab
File Edit View Search Terminal Help
[nz@localhost ~]$ cd ./labs/base/cli/setupLab
[nz@localhost setupLab]$ ls -l
total 8
-rwxr-xr-x. 1 nz nz 2607 May 11 07:47 labdb_tables.sql
-rwxr-xr-x. 1 nz nz 1045 May 11 07:47 setupLab.sh
[nz@localhost setupLab]$ ./setupLab.sh
ERROR: DROP DATABASE: object LABDB does not exist.
CREATE DATABASE
ERROR: DROP USER: object LABADMIN does not exist.
CREATE USER
ALTER USER
ALTER DATABASE
CREATE TABLE
Load session of table 'NATION' completed successfully
Load session of table 'REGION' completed successfully
Load session of table 'CUSTOMER' completed successfully
Load session of table 'SUPPLIER' completed successfully
Load session of table 'PART' completed successfully
Load session of table 'PARTSUPP' completed successfully
Load session of table 'ORDERS' completed successfully
Load session of table 'LINEITEM' completed successfully
[nz@localhost setupLab]$ █
```

## 16.6 Connecting to the Netezza system database using nzsql

Since you have not created any user or databases yet, you will connect to the default database as the default user, with the following credentials:

Database: **system**  
 Username: **admin**  
 Password: **password**

When issuing the **nzsql** command, the user supplies the user account, password and the database to connect to, using the example syntax below. Do not try to execute this command as it is just demonstrating the syntax:

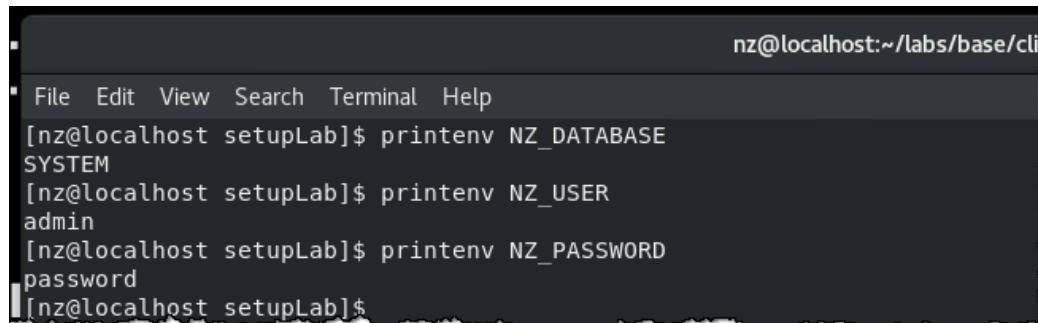
```
nzsql -d [db_name] -u [user] -pw [password]
```

Alternatively, these values (db\_name, user, password) can be stored in the command shell and passed to the nzsql command when nzsql is issued without any arguments.

### 16.6.1 Verify the current database, user and password values

- \_\_6 Let's verify the current database, user and password values stored in the command shell by issuing **printenv NZ\_DATABASE**, **printenv NZ\_USER**, and **printenv NZ\_PASSWORD** commands:

```
$ printenv NZ_DATABASE
$ printenv NZ_USER
$ printenv NZ_PASSWORD
```



The screenshot shows a terminal window with a dark background and light-colored text. The title bar reads "nz@localhost:~/labs/base/cli". The window contains the following text:

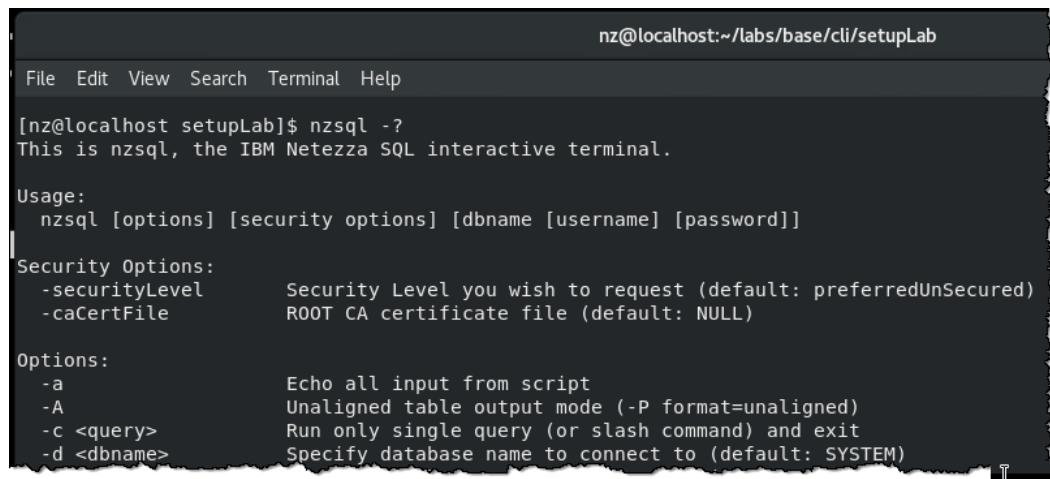
```
nz@localhost:~/labs/base/cli
File Edit View Search Terminal Help
[nz@localhost setupLab]$ printenv NZ_DATABASE
SYSTEM
[nz@localhost setupLab]$ printenv NZ_USER
admin
[nz@localhost setupLab]$ printenv NZ_PASSWORD
password
[nz@localhost setupLab]$
```

Since the current values correspond to our desired values, no modification is required. These environment variables are stored in the nz user's **.bashrc** file.

### 16.6.2 Review nzssql options and using the console

\_\_7 Let's take a look at what options are available to start nzssql.

```
$ nzssql -?
```



The terminal window title is "nz@localhost:~/labs/base/cli/setupLab". The output shows:

```
[nz@localhost setupLab]$ nzssql -?
This is nzssql, the IBM Netezza SQL interactive terminal.

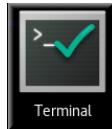
Usage:
  nzssql [options] [security options] [dbname [username] [password]]

Security Options:
  -securityLevel      Security Level you wish to request (default: preferredUnSecured)
  -caCertFile         ROOT CA certificate file (default: NULL)

Options:
  -a                  Echo all input from script
  -A                  Unaligned table output mode (-P format=unaligned)
  -c <query>          Run only single query (or slash command) and exit
  -d <dbname>          Specify database name to connect to (default: SYSTEM)
```

\_\_8 The `-?` option will list the usage and all options for the `nzssql` command. In this next exercise, you will start `nzssql` without arguments.

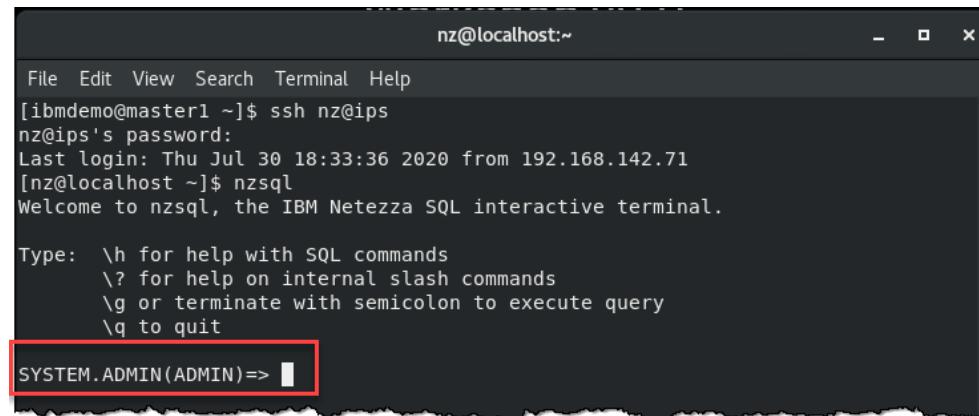
Open up another terminal window by double-clicking the `Terminal` desktop icon.



\_\_9 Next, `ssh` into the Netezza Linux host container:

```
$ ssh nz@ips (password = nz)
```

```
$ nzssql
```



The terminal window title is "nz@localhost:~". The output shows:

```
[ibmdemo@master1 ~]$ ssh nz@ips
nz@ips's password:
Last login: Thu Jul 30 18:33:36 2020 from 192.168.142.71
[nz@localhost ~]$ nzssql
Welcome to nzssql, the IBM Netezza SQL interactive terminal.

Type:  \h for help with SQL commands
      \? for help on internal slash commands
      \g or terminate with semicolon to execute query
      \q to quit

SYSTEM.ADMIN(ADMIN)=> 
```

This will bring up the `nzssql` prompt below that shows a connection to the system database as user admin:

The `nzssql` console prompt format: `DATABASE . SCHEMA (USER)`

## 16.7 Commonly used commands and SQL statements

There are commonly used commands that start with \ which we will demonstrate in this section. First, you will run the two help commands to familiarize yourself with these handy commands. The \h command lists the available SQL commands, while the \? command is used to list the internal slash commands.

- 10 Execute these commands in the [nzsql console](#):

```
SYSTEM.ADMIN(ADMIN)=> \h
```

```
nz@localhost:~$ ./nzsql -P 5432 -U SYSTEM -W ADMIN
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)=> \h
Available help:
  ALTER AGGREGATE          CREATE ROLE           GROOM TABLE
  ALTER CATEGORY           CREATE SCHEDULER RULE   INSERT
  ALTER COHORT             CREATE SCHEMA         MERGE
  ALTER DATABASE            CREATE SECURITY LEVEL  RESET
  ALTER FUNCTION            CREATE SEQUENCE       REVERT
  ALTER GROUP               CREATE SYNONYM        REVOKE
  ALTER HISTORY CONFIGURATION CREATE TABLE        ROLLBACK
  ALTER PROCEDURE           CREATE TABLE AS      SELECT
```

```
SYSTEM.ADMIN(ADMIN)=> \?
```

```
nz@localhost:~$ ./nzsql -P 5432 -U SYSTEM -W ADMIN
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)=> \?
\a                  toggle between unaligned and aligned mode
\act                show current active sessions
\c[onnect] [dbname [user] [password]]      connect to new database (currently 'SYSTEM')
\c <title>          HTML table title
\copy ...           perform SQL COPY with data stream to the client machine
\dt <table>          describe table (or view, index, sequence, synonym)
\do <table>          describe table or view in sorted order
\dt[v|i|s|e|x]      list tables/views/indices/sequences/temp tables/external tables

\f <sep>          change field separator
\g [file]           send query to backend (and results in [file] or |pipe)
\h [cmd]            help on syntax of sql commands, * for all commands
\H                  toggle HTML mode (currently off)
\i <file>           read and execute queries from <file>
\l[+]              list all databases, + for additional fields
```

From the output of the \? command, you found the [\l](#) internal command you can use to find out all the databases.

\_\_11 Let's look at the list of databases that have been created:

SYSTEM.ADMIN(ADMIN)=> \l

```
nz@localhost:~ 
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)-> \l
  List of databases
  DATABASE   | OWNER
  -----
CHURN      | ADMIN
LABDB      | LABADMIN
MYWORKSHOPDB | ADMIN
SYSTEM      | ADMIN
WORKSHOP    | ADMIN
(5 rows)

SYSTEM.ADMIN(ADMIN)->
```

Secondly, you will use \dSv to find the system views within the system database.

 nz	Note: There are system tables, but it is not recommended to directly access those tables as they can change from release to release and are restricted from the normal user.
--	--

\_\_12 Let's find the system views:

SYSTEM.ADMIN(ADMIN)=> \dSv

```
nz@localhost:~ 
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)-> \dSv
  List of relations
  Schema   |           Name          |   Type   | Owner
  -----
DEFINITION_SCHEMA | V_ACL_DATA        | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_ATTRIBUTE        | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_ATTRIBUTE2       | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_AUTHENTICATION  | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_AUTHENTICATION_SETTINGS | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_BACKUP_GROUP     | SYSTEM VIEW | ADMIN
DEFINITION_SCHEMA | V_BACKUP_GROUP_HISTORY | SYSTEM VIEW | ADMIN
--More--
```

 nz	Note: Press the space bar to scroll down the result set when you see --More-- on the screen.
--	--

The list of the system views was truncated in the output above due to the length. Here are the primary views to investigate for DBAs/users new to Netezza:

```

_V_USER
_V_SESSION_V_DATABASE
_V_SCHEMA
_V_TABLE
_V_VIEW
_V_ATTRIBUTE
_V_SEQUENCE
_V_SYNONYM
_V_FUNCTION
_V_AGGREGATE
_V_PROCEDURE

```

- 13 From the previous command, you can see that there is a user table called `_V_USER`. To find out what is stored in that table, you will use the describe command `\d`.

`SYSTEM.ADMIN(ADMIN)=> \d _V_USER`

Attribute	Type	Modifier	Default Value
OBJID	OID	NOT NULL	
USERNAME	NAME	NOT NULL	
OWNER	NAME		
VALIDUNTIL	TIMESTAMP		
CREATEDATE	ABSTIME	NOT NULL	
ROWLIMIT	INTEGER		
ACCT_LOCKED	BOOLEAN		
INV_CONN_CNT	SMALLINT		
PWD_INVALID	BOOLEAN		
PWD_LAST_CHGED	DATE		
--More--			

This will return all the columns of the `_V_USER` system table.

- 14 Next, find the existing users stored in the table. In case too many rows are returned at once, you can first calculate the number of rows it contains by executing the following query:

`SYSTEM.ADMIN(ADMIN)=> SELECT COUNT(*) FROM (SELECT * FROM _V_USER) AS "Wrapper";`

COUNT
2 (1 row)

- \_\_15 The above query is essentially the same as `SELECT COUNT (*) FROM _V_USER`. We have demonstrated the subselect syntax in case there is a complex query that needs to have the result set evaluated. The result should show the two entries in the user table.

You can next enter the following query to list the usernames:

```
SYSTEM.ADMIN(ADMIN)=> SELECT OBJID, USERNAME, OWNER, CREATEDATE, USEAUTH,
PWD_EXPIRY FROM _V_USER;
```

```
nz@localhost:~
```

OBJID	USERNAME	OWNER	CREATEDATE	USEAUTH	PWD_EXPIRY
4900	ADMIN	ADMIN	2020-02-12 19:04:55	DEFAULT	0
201246	LABADMIN	ADMIN	2020-07-30 19:38:17	DEFAULT	0

```
(2 rows)
```

```
SYSTEM.ADMIN(ADMIN)=> 
```

- \_\_16 To exit `nzsql`, use the command `\q` to return to NPS.

```
SYSTEM.ADMIN(ADMIN)=> \q
```

```
File Edit View Search Terminal Help
```

```
SYSTEM.ADMIN(ADMIN)=> \q
```

```
[nz@localhost ~]$ 
```

- \_\_17 Re-enter `nzsql`:

```
$ nzsql
```

```
nz@localhost:~
```

OBJID	USERNAME	OWNER	CREATEDATE	USEAUTH	PWD_EXPIRY
4900	ADMIN	ADMIN	2020-02-12 19:04:55	DEFAULT	0
201246	LABADMIN	ADMIN	2020-07-30 19:38:17	DEFAULT	0

```
(2 rows)
```

```
SYSTEM.ADMIN(ADMIN)=> 
```

You will be using the `nzsql` command line tool throughout the rest of this lab.

For a complete list of the nzsql internal slash options, see: <https://ibm.biz/nzsql-slash-options>.

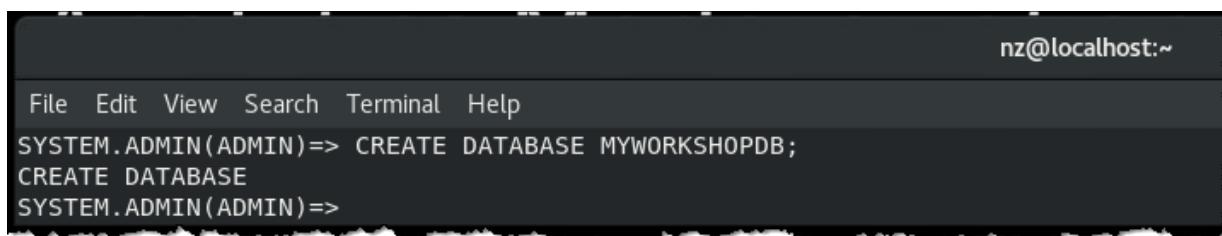
## 16.8 Creating a database

A database in Netezza is a collection of entities such as schemas, tables, views, synonyms, and other objects.

The initial system consists of the SYSTEM database containing system tables and views. Initially, only the admin user can create databases, but the admin user can grant other users permission to create databases as described in security and access control. You cannot delete the system database. The admin user can also make another user the owner of a database, giving that user admin-like control over that database and its contents. Security is out of scope for this lab as you will perform all operations as the ADMIN database super-user.

- 18 The database creator becomes the default owner of the database. The owner can remove the database and all of its objects, even if other users own objects within the database.

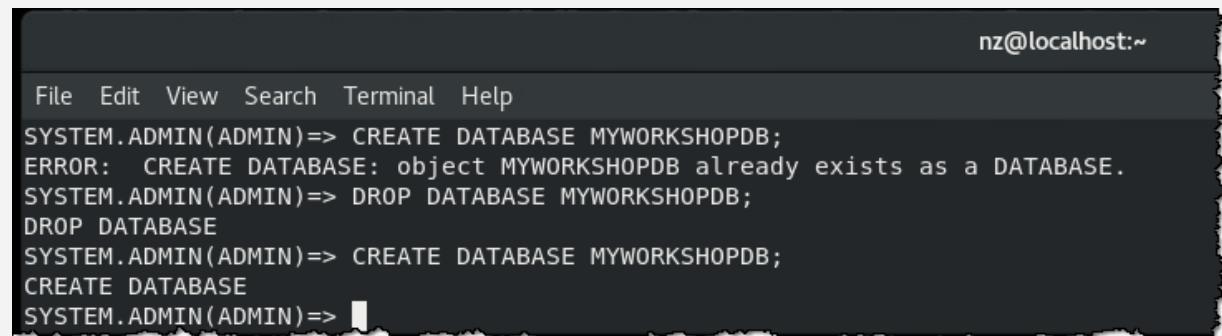
```
SYSTEM.ADMIN(ADMIN)=> CREATE DATABASE MYWORKSHOPDB;
```



The screenshot shows a terminal window with a dark background and light-colored text. At the top right, it says "nz@localhost:~". Below that is a menu bar with "File Edit View Search Terminal Help". The main area of the terminal shows the command "CREATE DATABASE MYWORKSHOPDB;" being entered by the user "SYSTEM.ADMIN(ADMIN)". After the command is run, the terminal prompt "SYSTEM.ADMIN(ADMIN)=>" appears again.

If database **MYWORKSHOPDB** already exists, you will get an error on the CREATE statement.

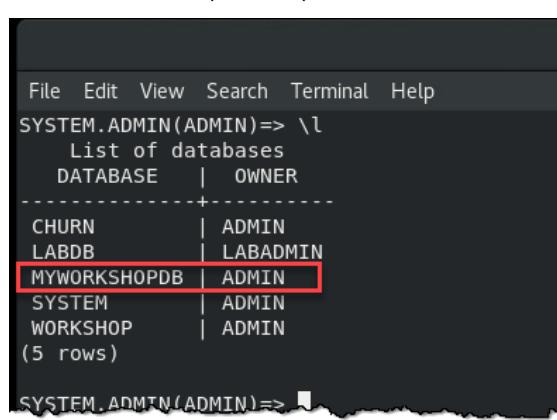
Simply drop it and then create it like this:



The screenshot shows a terminal window with a dark background and light-colored text. At the top right, it says "nz@localhost:~". Below that is a menu bar with "File Edit View Search Terminal Help". The main area of the terminal shows the user attempting to create a database named "MYWORKSHOPDB". An error message "ERROR: CREATE DATABASE: object MYWORKSHOPDB already exists as a DATABASE." is displayed. The user then enters the command "DROP DATABASE MYWORKSHOPDB;" to remove the existing database. Finally, the user runs the "CREATE DATABASE MYWORKSHOPDB;" command again, which succeeds without an error.

\_\_19 List the databases in the Netezza system:

```
SYSTEM.ADMIN(ADMIN)=> \l
```



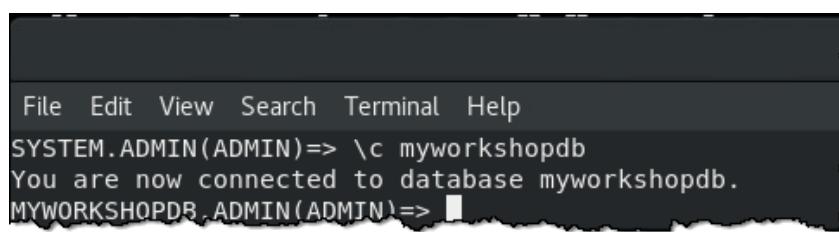
```
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)=> \l
  List of databases
  DATABASE | OWNER
  -----+-----
  CHURN   | ADMIN
  LABDB   | LABADMIN
  MYWORKSHOPDB | ADMIN
  SYSTEM   | ADMIN
  WORKSHOP | ADMIN
(5 rows)

SYSTEM.ADMIN(ADMIN)=>
```

Notice the end of the `\` commands do not require a terminator `;` this is because the `\` commands terminate automatically on a single line. SQL statements can span more than one line in `nzssql`.

\_\_20 Connect to the new database:

```
SYSTEM.ADMIN(ADMIN)=> \c myworkshopdb
```



```
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)=> \c myworkshopdb
You are now connected to database myworkshopdb.
MYWORKSHOPDB.ADMIN(ADMIN)=>
```

## 16.9 Creating a table

Now that you have a database, you can create a table to store your data. Tables are specific to a Netezza database and have one or more columns, each with attributes or data types.

**Netezza supported data types:**

Data type	Example	Description
<code>byteint</code> <code>smallint</code> <code>integer</code> <code>bigint</code>	120 0 256 1290985	See <a href="#">Integer data types</a> .
<code>numeric</code> <code>decimal</code>	-99.56 123.679	See <a href="#">Fixed-point data types</a> .
<code>real</code> <code>double precision</code>	-81293.35	See <a href="#">Floating-point data types</a> .
<code>char (n)</code>	<code>salary</code>	See <a href="#">Character strings</a> .
<code>varchar (n)</code>	<code>this is a variable string</code>	See <a href="#">Character strings</a> .
<code>boolean</code>	<code>true</code>	This is an ASCII string that contains one of the following values: True yes 1 t y False no 0 f n
<code>JSON</code>		See <a href="#">Document data types</a>
<code>date</code>	2002-02-04	The date is an exact 4-byte data type. The system recognizes a range of dates that are composed of year, month, and day. See <a href="#">DateStyle option</a> .
<code>time</code>	01:59:45 23:00:01	See <a href="#">time data type</a> .
<code>time with time zone</code>	01:15:33 -05	See <a href="#">Time with time zone (timetz) data type</a> .
<code>timestamp</code>	2002-02-04 01:15:33	See <a href="#">timestamp data type</a> .
<code>interval</code>	2 years 3 months 4 days	See <a href="#">Interval data types</a> .

- \_\_21 To create our lab table, make sure you are connected to the database you created called MYWORKSHOPDB. A DDL file was prepared for you in </home/nz/labs/base/cli>.

View the file while in nzsql by running an operating system command with `\!` option:

```
MYWORKSHOPDB.ADMIN(ADMIN)=> \!
/home/nz/labs/base/cli/customer_transactions.ddl
```

```
nz@localhost:~
```

```
File Edit View Search Terminal Help
SYSTEM.ADMIN(ADMIN)=> \c myworkshopdb
You are now connected to database myworkshopdb.
MYWORKSHOPDB.ADMIN(ADMIN)=> \! cat /home/nz/labs/base/cli/customer_transactions.ddl
\echo
\echo ***** Creating schema: "APPWORKSHOP_18_19"
CREATE SCHEMA APPWORKSHOP_18_19
;
\echo
\echo ***** Setting schema: "APPWORKSHOP_18_19"
```

- \_\_22 Now create the table with the `\i` option:

```
MYWORKSHOPDB.ADMIN(ADMIN)=> \i
/home/nz/labs/base/cli/customer_transactions.ddl
```

```
nz@localhost:~
```

```
File Edit View Search Terminal Help
MYWORKSHOPDB.ADMIN(ADMIN)=> \i /home/nz/labs/base/cli/customer_transactions.ddl
***** Creating schema: "APPWORKSHOP_18_19"
CREATE SCHEMA
***** Setting schema: "APPWORKSHOP_18_19"
SET SCHEMA
***** Creating table: "CUSTOMER_TRANSACTIONS"
CREATE TABLE
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=>
```

- \_\_23 Check that the table exists: (Note: there is a space before the `\d`):

```
MYWORKSHOPDB.ADMIN(ADMIN)=> \d
```

```
nz@localhost:~
```

```
File Edit View Search Terminal Help
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> \d
      List of relations
 Schema |           Name           | Type | Owner
-----+-----+-----+
 APPWORKSHOP_18_19 | CUSTOMER_TRANSACTIONS | TABLE | ADMIN
(1 row)

MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=>
```

Notice that the `nzsql` prompt schema changed, this was due to the `SET SCHEMA` command that was specified in `customer_transactions.ddl`.

\_\_24 Select the table data:

```
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> SELECT * FROM CUSTOMER_TRANSACTIONS;
```

```
nz@localhost:~  
File Edit View Search Terminal Help  
(1 row)  
  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> SELECT * FROM CUSTOMER_TRANSACTIONS;  
CUSTID | TRANSACTION_DATE | SYMBOL | STOCK_PRICE | UNITS_TRADED | REALIZED_LOSS | REALIZED_GAIN  
-----+-----+-----+-----+-----+-----+-----+  
(0 rows)  
  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> ■
```

There is no data in the table yet because you just created it. In the next section, you will load some data.

## 16.10 Loading data into a table

\_\_25 Load data into our `CUSTOMER_TRANSACTIONS` table using an external table (a csv file.).

You will be using a three line command for doing this, terminated on the last line with the ; terminator:

```
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> INSERT INTO CUSTOMER_TRANSACTIONS  
SELECT * FROM EXTERNAL
```

```
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=>  
'/home/nz/labs/data/customer_activity_transactions-1H2020.csv'
```

```
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> USING (DELIMITER ',' SkipRows 1  
QuotedValue double);
```

```
nz@localhost:~  
File Edit View Search Terminal Help  
  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> INSERT INTO CUSTOMER_TRANSACTIONS SELECT * FROM EXTERNAL  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)-> '/home/nz/labs/data/customer_activity_transactions-1H2020.csv'  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)-> USING (DELIMITER ',' SkipRows 1 QuotedValue double);  
INSERT 0 500000  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=>
```

\_\_26 Select 5 rows from the table `CUSTOMER_TRANSACTIONS`:

```
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> SELECT * FROM CUSTOMER_TRANSACTIONS  
LIMIT 5;
```

```
nz@localhost:~  
File Edit View Search Terminal Help  
  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> SELECT * FROM CUSTOMER_TRANSACTIONS LIMIT 5;  
CUSTID | TRANSACTION_DATE | SYMBOL | STOCK_PRICE | UNITS_TRADED | REALIZED_LOSS | REALIZED_GAIN  
-----+-----+-----+-----+-----+-----+-----+  
437 | 2020-01-01 | GS | 24.93 | 53 | 0.00 | 1321.29  
1730 | 2020-01-05 | JPM | 94.62 | 14 | -1324.68 | 0.00  
1961 | 2020-01-09 | V | 40.93 | 27 | 0.00 | 1105.11  
402 | 2020-01-13 | UNH | 78.19 | 2 | -156.38 | 0.00  
2012 | 2020-01-17 | WBA | 73.17 | 31 | 0.00 | 2268.27  
(5 rows)  
  
MYWORKSHOPDB.APPWORKSHOP_18_19(ADMIN)=> ■
```

## 16.11 Lab conclusion

In this lab you learned how to use the [nzsql](#) tool to create a database, a table and load data into that table. These are the basic functions of the Netezza database.

### **\*\* End of Lab 16 – Netezza Performance Server (NPS) Getting Started**

Lab by Dan Hancock, Edited by Burt Vialpando and Kent Rubin

---

## Back Page: Notices

This information was developed for products and services offered in the USA.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing  
IBM Corporation  
North Castle Drive  
Armonk, NY 10504-1785  
USA

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation  
Licensing  
2-31 Roppongi 3-chome, Minato-ku  
Tokyo 106-0032, Japan

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.** Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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 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.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental. All references to fictitious companies or individuals are used for illustration purposes only.

#### COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

---

## Back Page: Trademarks and Copyrights

The following terms are trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM	AIX	CICS	ClearCase	ClearQuest	Cloudscape
Cube Views	Db2	developerWorks	DRDA	IMS	IMS/ESA
Informix	Lotus	Lotus Workflow	MQSeries	OmniFind	
Rational	Redbooks	Red Brick	RequisitePro	System i	
System z	Tivoli	WebSphere	Workplace	System p	

Adobe, Acrobat, Portable Document Format (PDF), and PostScript are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, other countries, or both.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both. See Java Guidelines

Microsoft, Windows, Windows NT, and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

ITIL is a registered trademark and a registered community trademark of the Office of Government Commerce and is registered in the U.S. Patent and Trademark Office.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Other company, product and service names may be trademarks or service marks of others.

## **NOTES**

## **NOTES**



---

© Copyright IBM Corporation 2020.

The information contained in these materials is provided for informational purposes only and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way.

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at:

<https://www.ibm.com/legal/us/en/copytrade.shtml>

Other company, product and service names may be trademarks or service marks of others.



Please Recycle