Data Engineering Batch 1

PROJECT 1

Hybrid Cloud Data Movement

Akilesh K

k.akilesh123@gmail.com

Project Statement

Implement a solution that involves moving data between on-premises data sources and Azure cloud using Azure Data Factory, and perform data processing tasks in Azure Databricks.

Project Overview

The project entails orchestrating a data pipeline to facilitate seamless data movement between on-premises data sources and the Azure cloud environment. Initially, a SQL server will be set up, along with the creation of a database to house the source data. Azure Data Factory will play a pivotal role in this process, acting as the conduit for transferring data from the SQL database to Blob storage. This transfer will be facilitated by utilizing a self-hosted integration runtime, ensuring secure and efficient data transmission. Subsequently, Azure Databricks will be leveraged to execute essential data transformation tasks using PySpark, enabling the manipulation and refinement of the ingested data. Overall, the solution aims to streamline data integration and processing workflows, enhancing data accessibility and usability within the Azure ecosystem.

About the Project

Database:

The Northwind database is a sample database that was originally created by Microsoft and used as the basis for their tutorials in a variety of database products for decades. The Northwind database contains the sales data for a fictitious company called “Northwind Traders,” which imports and exports specialty foods from around the world. The Northwind database is an excellent tutorial schema for a small-business ERP, with customers, orders, inventory, purchasing, suppliers, shipping, employees, and single-entry accounting.

The Northwind dataset includes sample data for the following.

Suppliers: Suppliers and vendors of Northwind

Customers: Customers who buy products from Northwind

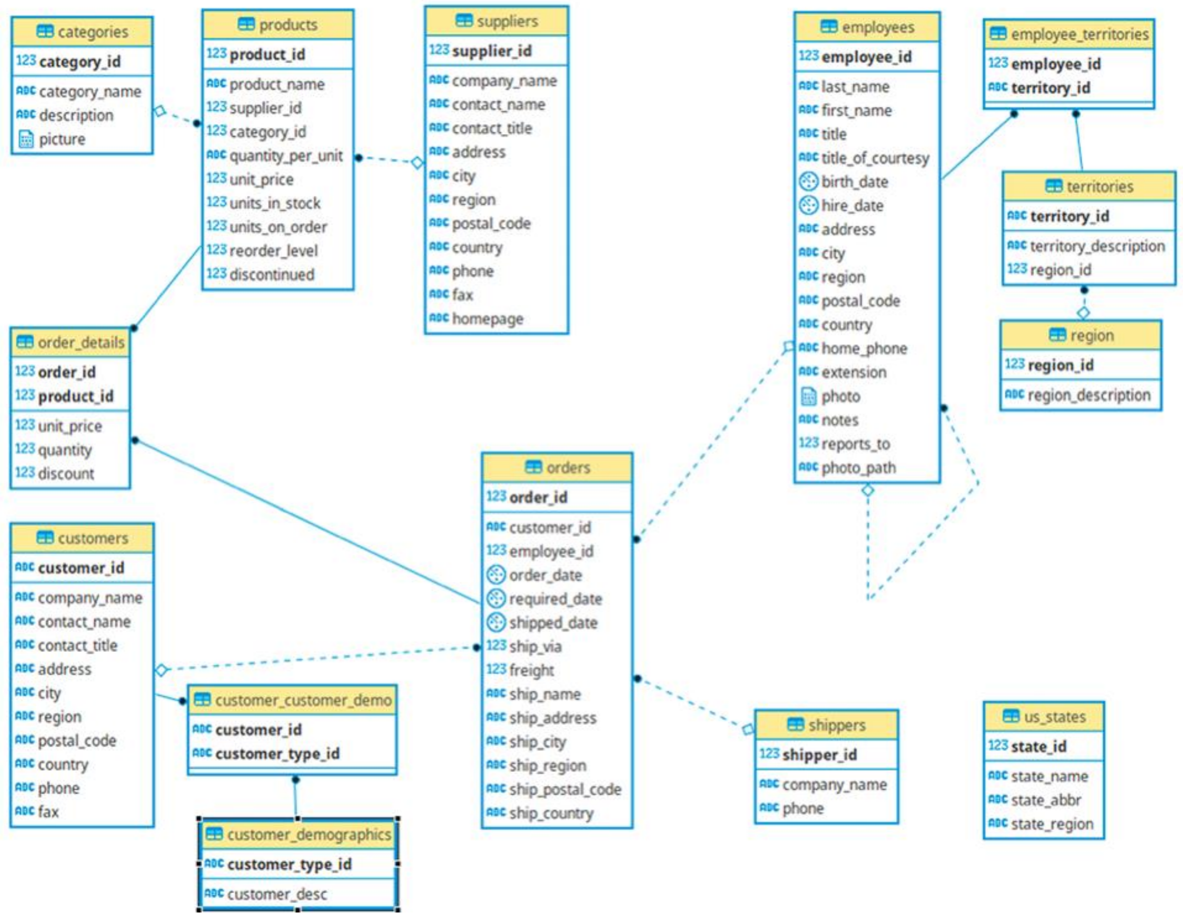
Employees: Employee details of Northwind traders

Products: Product information

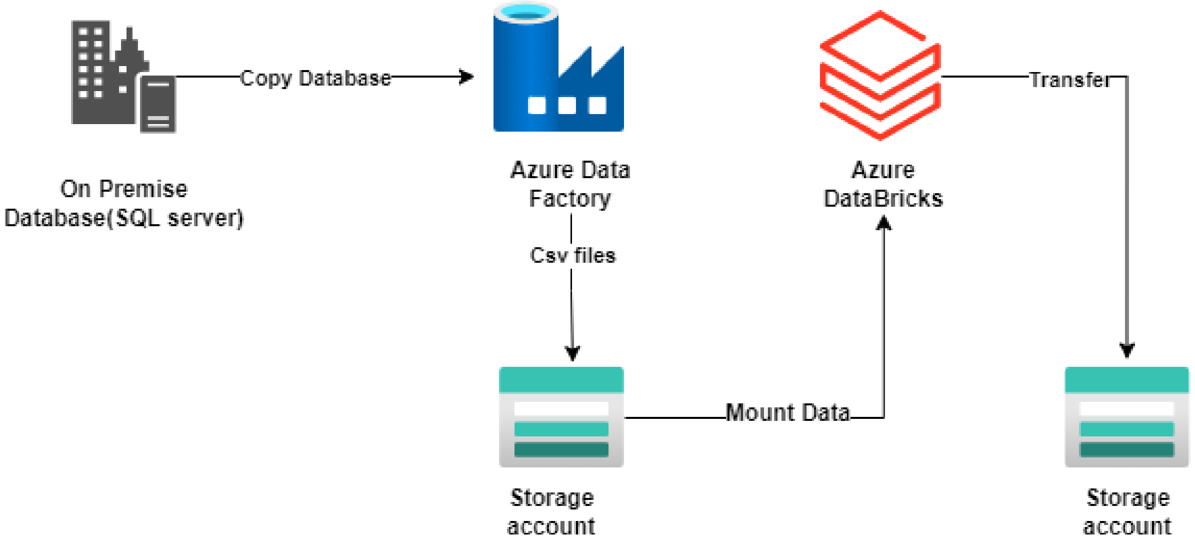
Shippers: The details of the shippers who ship the products from the traders to the end-customers

Orders and Order_Details: Sales Order transactions taking place between the customers & the company

Schema



Architecture diagram



The data flow in this architecture starts with SQL Server as the source, where data is extracted using ADF and loaded into Blob storage. Then, Azure Databricks performs the necessary data transformations using PySpark, leveraging the data stored in Blob storage. Overall, this architecture enables efficient data movement and processing between on-premises and cloud environments, facilitating analytics, reporting, and other data-driven tasks.

Azure Resources Used for this Project

- **Azure Data Factory (ADF):**

An ADF instance will be created to orchestrate and automate the data movement process.

Linked Services: Configured to establish connections to on-premises data sources and Azure Blob Storage.

Datasets: Defined to represent the data structures and schemas of the data sources.

Pipelines: Created to define the sequence of activities for copying data from on-premises to Azure Blob Storage.

- **Azure Blob Storage:**

Blob storage containers will be used as the destination for storing data transferred from on-premises sources.

The stored data will be accessed by Azure Databricks for data processing tasks.

- **Azure Databricks:**

A Databricks workspace will be provisioned to perform data processing tasks using PySpark.

Databricks Notebooks: Utilized to write and execute PySpark code for data transformation, cleaning, and analysis.

Project Requirements

- **Setting up Data Sources:**

Begin by identifying the on-premises data sources from which you want to extract data. These could be databases, files, or any other structured or unstructured data repositories.

Ensure that the necessary connectivity options are available for accessing these on-premises data sources securely from the Azure cloud environment.

- **Azure Data Factory Configuration:**

Create an Azure Data Factory (ADF) instance in your Azure subscription.

Configure linked services in ADF to establish connections to both the on-premises data sources and the Azure cloud environment. This involves providing authentication credentials and connection details.

Define datasets in ADF to represent the data structures and schemas of the data sources. This includes specifying the location, format, and schema of the data residing in both the on-premises and Azure cloud environments.

Create pipelines in ADF to orchestrate the data movement process. Pipelines define the sequence of activities required to copy data from the on-premises sources to the Azure cloud.

- **Data Movement with Azure Data Factory:**

Use ADF activities such as Copy Data to copy data from the on-premises sources to Azure Blob Storage. This activity handles the movement of data securely and efficiently, with options for parallelism, fault tolerance, and monitoring.

Configure the Copy Data activity with appropriate settings such as source and destination datasets, data integration runtime, scheduling options, and error handling mechanisms.

- **Data Processing with Azure Databricks:**

Provision an Azure Databricks workspace in your Azure subscription.

Define and implement data processing tasks using PySpark within the Databricks environment. PySpark provides a powerful framework for distributed data processing, enabling tasks such as data cleaning, transformation, aggregation, and analysis.

Use Databricks notebooks to write and execute PySpark code interactively, leveraging the scalability and performance of the Databricks runtime.

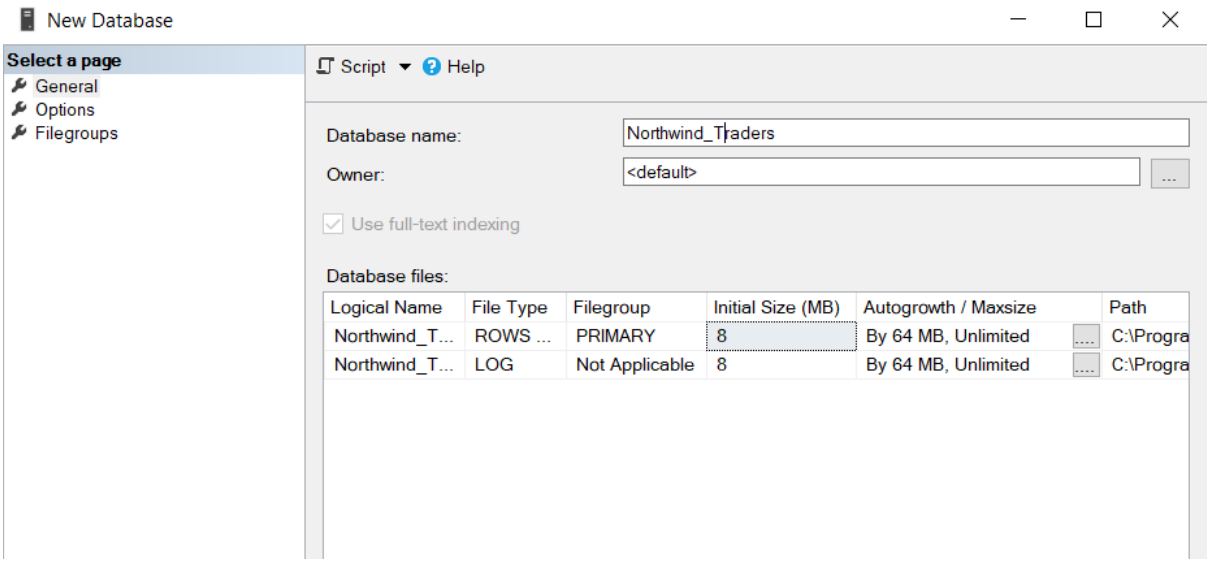
Tasks performed:

- Set up SQL Server with a database, create schema, and add data.
- Create a new user with SQL Server Authentication as System Admin.
- Create Azure storage account and blob container.
- Set up an Azure Data Factory account and configure a data pipeline for data transfer.
- Install self-hosted integration runtime on-premises for SQL Server connection.
- Mount blob storage to Azure Databricks and read CSV file into dataframe.

- Perform data transformations including aggregation, sorting, and joining.
- Profile data to assess quality and characteristics.
- Visualize data using pie chart representation.
- Transfer manipulated dataframe to blob storage

Implementation

- **For On premise database , set it up in SQL server**



Database name: Northwind_Traders

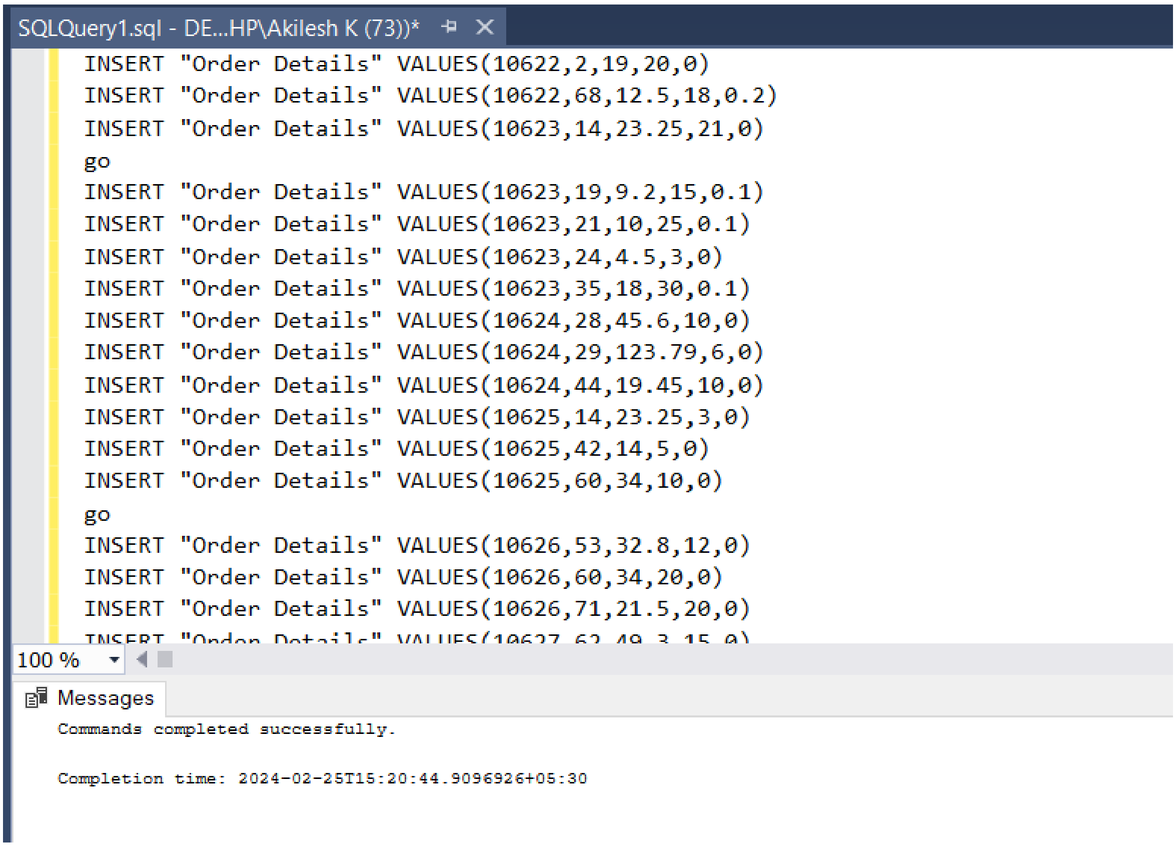
Owner: <default>

☑ Use full-text indexing

Database files:

Logical Name	File Type	Filegroup	Initial Size (MB)	Autogrowth / Maxsize		Path
Northwind_T...	ROWS ...	PRIMARY	8	By 64 MB, Unlimited	....	C:\Progra
Northwind_T...	LOG	Not Applicable	8	By 64 MB, Unlimited	....	C:\Progra

- **Write query of the schema for the table and add data into it**

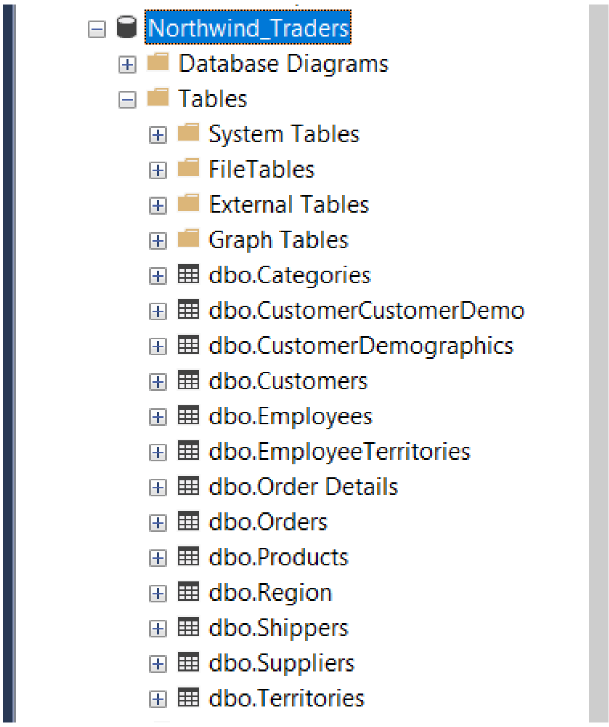


```
INSERT "Order Details" VALUES(10622,2,19,20,0)
INSERT "Order Details" VALUES(10622,68,12.5,18,0.2)
INSERT "Order Details" VALUES(10623,14,23.25,21,0)
go
INSERT "Order Details" VALUES(10623,19,9.2,15,0.1)
INSERT "Order Details" VALUES(10623,21,10,25,0.1)
INSERT "Order Details" VALUES(10623,24,4.5,3,0)
INSERT "Order Details" VALUES(10623,35,18,30,0.1)
INSERT "Order Details" VALUES(10624,28,45.6,10,0)
INSERT "Order Details" VALUES(10624,29,123.79,6,0)
INSERT "Order Details" VALUES(10624,44,19.45,10,0)
INSERT "Order Details" VALUES(10625,14,23.25,3,0)
INSERT "Order Details" VALUES(10625,42,14,5,0)
INSERT "Order Details" VALUES(10625,60,34,10,0)
go
INSERT "Order Details" VALUES(10626,53,32.8,12,0)
INSERT "Order Details" VALUES(10626,60,34,20,0)
INSERT "Order Details" VALUES(10626,71,21.5,20,0)
INSERT "Order Details" VALUES(10627,62,49.3,15,0)
```

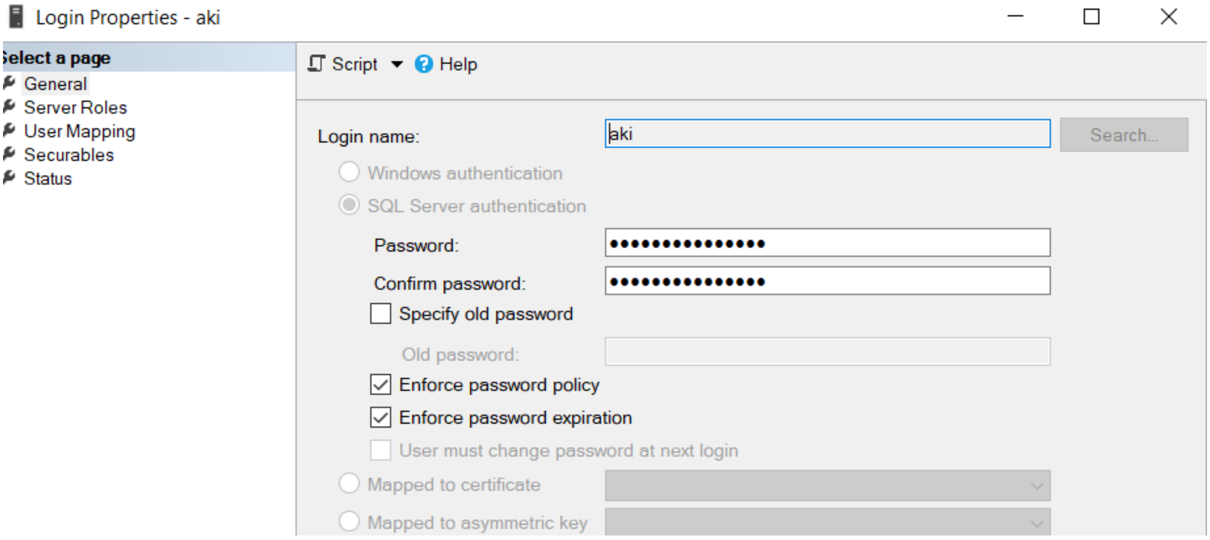
```
Commands completed successfully.

Completion time: 2024-02-25T15:20:44.9096926+05:30
```

- **Tables are created in the database of the SQl server**



- **Create a new user with SQL server Authentication with the role of System admin**



Login Properties - aki

Login name: aki

○ Windows authentication

● SQL Server authentication

Password: •••••••••••••••

Confirm password: •••••••••••••••

□ Specify old password

Old password:

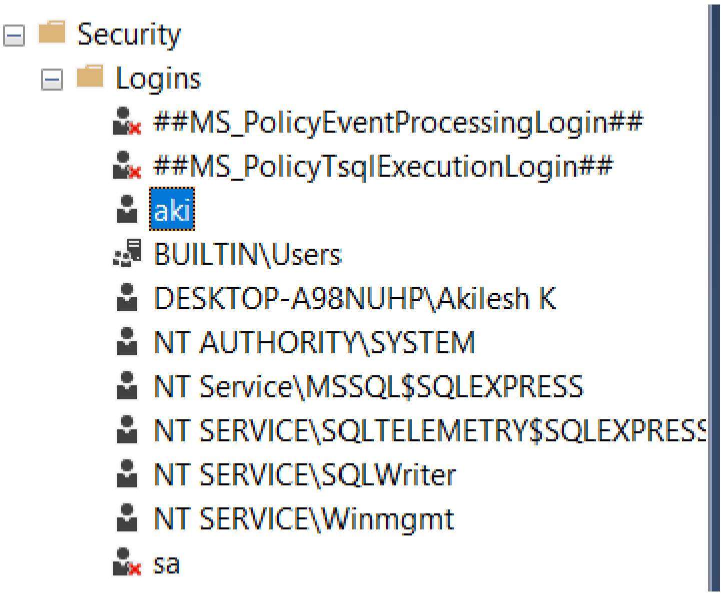
☑ Enforce password policy

☑ Enforce password expiration

□ User must change password at next login

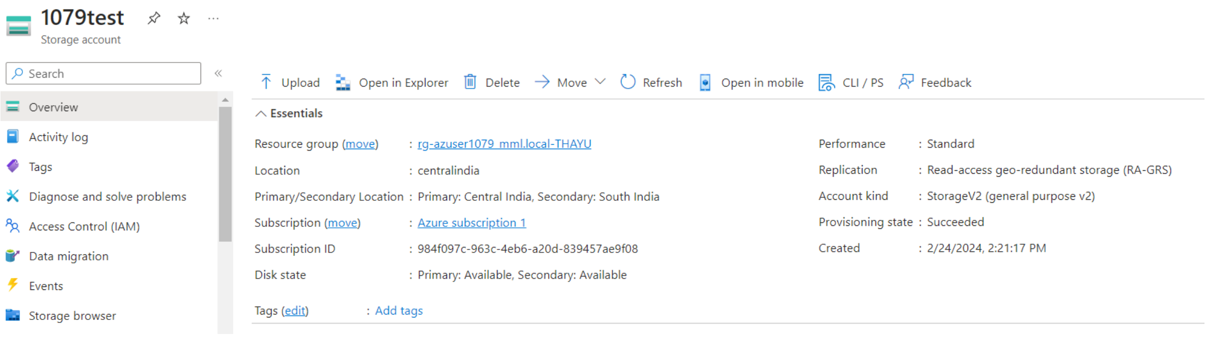
○ Mapped to certificate

○ Mapped to asymmetric key



```
Security
  Logins
    ##MS_PolicyEventProcessingLogin##
    ##MS_PolicyTsqlExecutionLogin##
    aki
    BUILTIN\Users
    DESKTOP-A98NUHP\Akilesh K
    NT AUTHORITY\SYSTEM
    NT Service\MSSQL$SQLEXPRESS
    NT SERVICE\SQLTELEMETRY$SQLEXPRESS
    NT SERVICE\SQLWriter
    NT SERVICE\Winmgmt
    sa
```

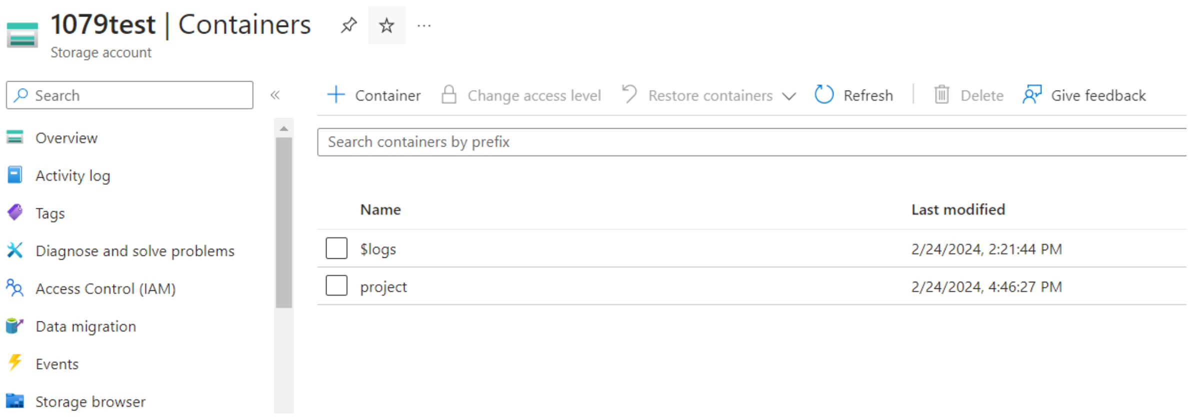
- **In azure ,create a storage account mentioning the location**



Resource group (move) : rg-azuser1079_mml.local-THAYU
Location : centralindia
Primary/Secondary Location : Primary: Central India, Secondary: South India
Subscription (move) : Azure subscription 1
Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08
Disk state : Primary: Available, Secondary: Available
Tags (edit) : Add tags

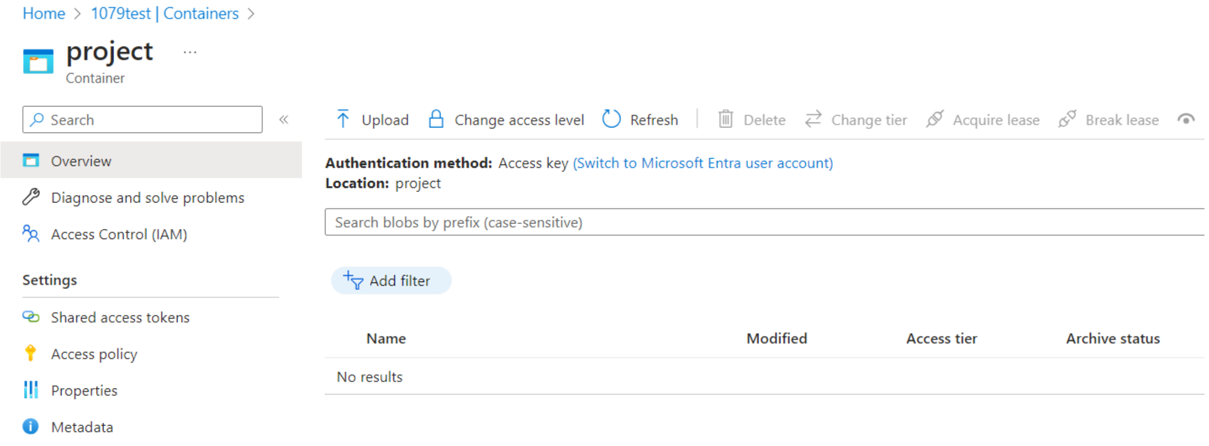
Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 2/24/2024, 2:21:17 PM

- **create a new blob container**



Name	Last modified
☐ $logs	2/24/2024, 2:21:44 PM
☐ project	2/24/2024, 4:46:27 PM

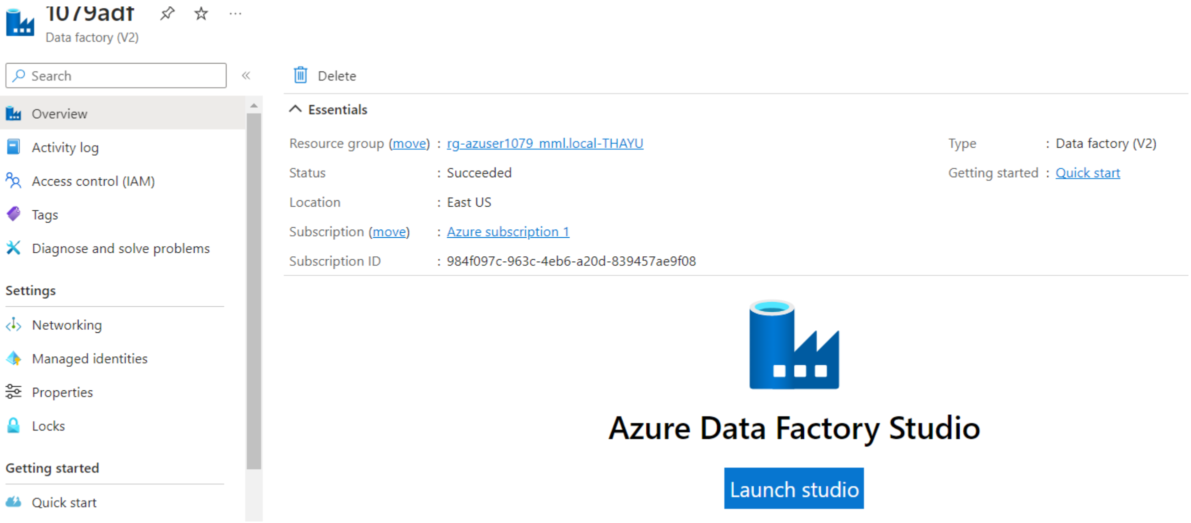
- **keep the container file empty**



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: project

Name	Modified	Access tier	Archive status
No results			

- **create a azure data factory account**



Resource group (move) : rg-azuser1079_mml.local-THAYU

Status : Succeeded

Location : East US

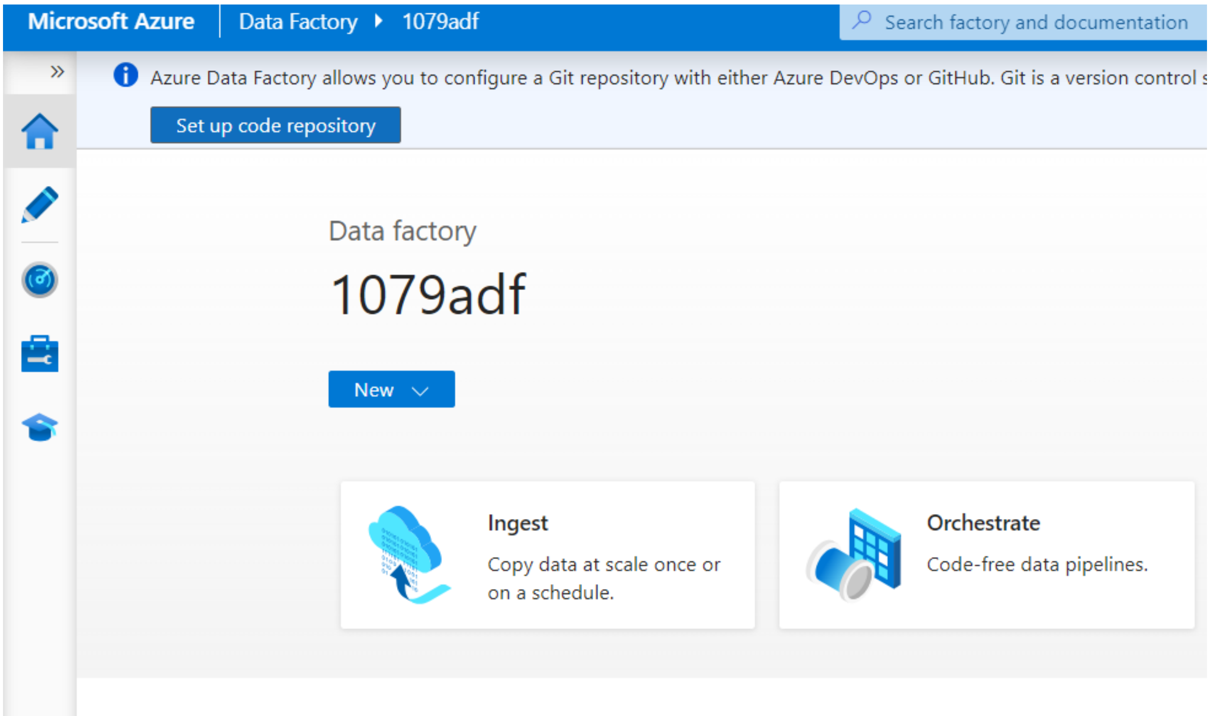
Subscription (move) : Azure subscription 1

Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08

Type : Data factory (V2)

Getting started : Quick start

- **select ingest data to create a new data pipeline**

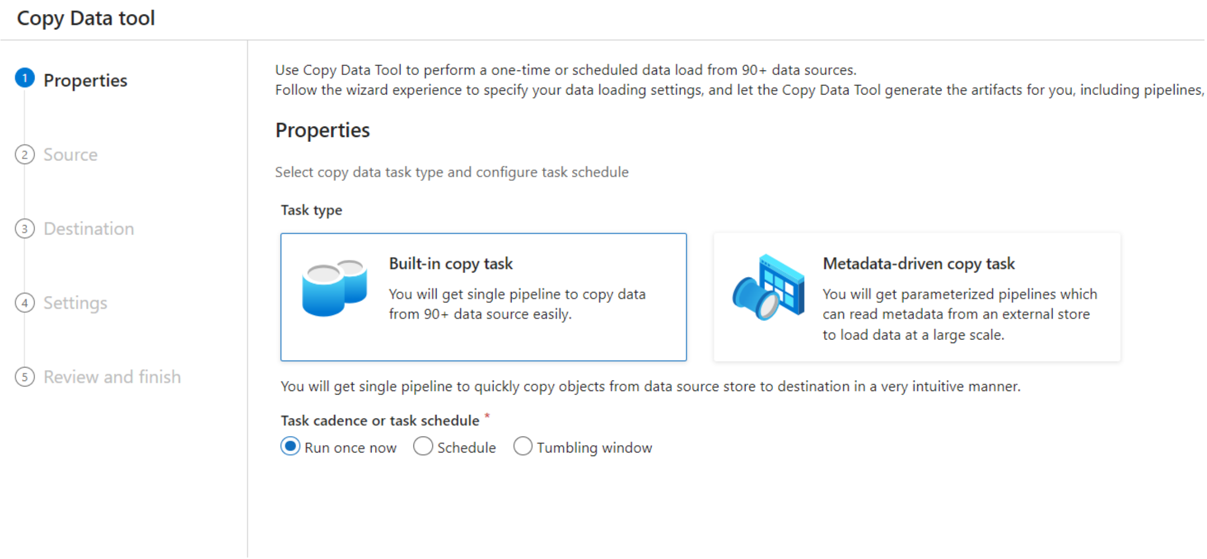


Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control s

Ingest: Copy data at scale once or on a schedule.

Orchestrate: Code-free data pipelines.

- **In Properties set type as built in copy task.**
- **task schedule as run once now**



Copy Data tool

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources.
Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines,

Properties

Select copy data task type and configure task schedule

Task type

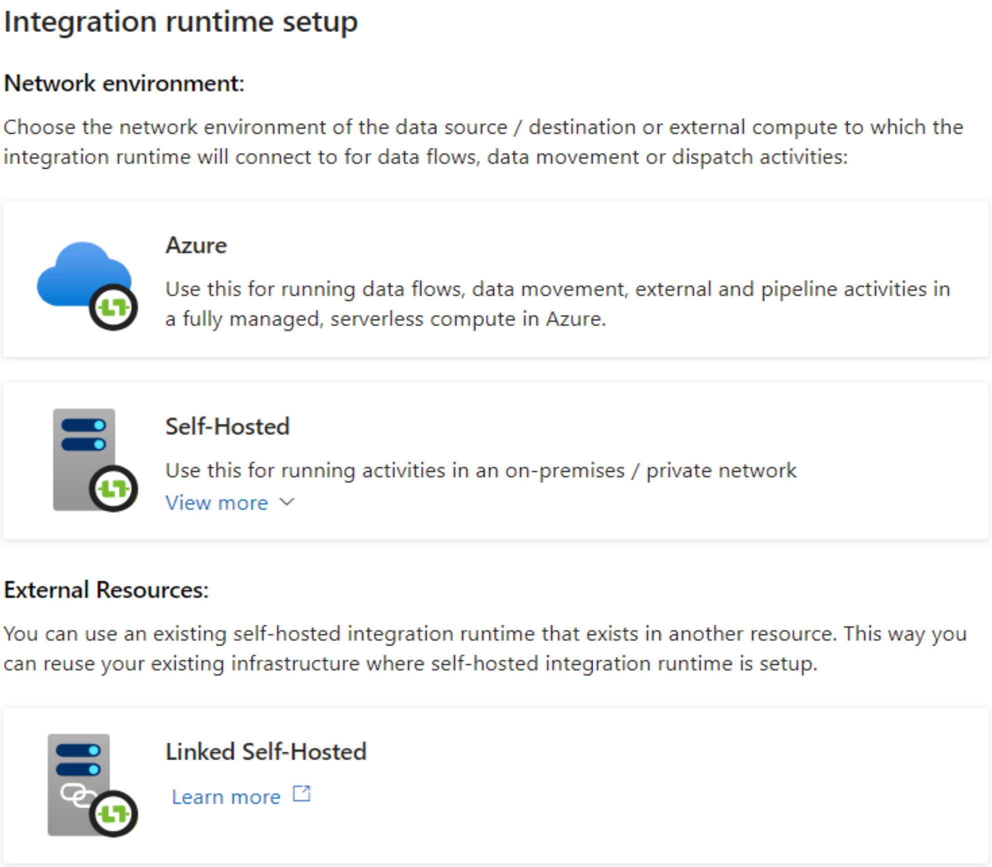
Built-in copy task
You will get single pipeline to copy data from 90+ data source easily.

Metadata-driven copy task
You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *
● Run once now ○ Schedule ○ Tumbling window

- **In integration runtime setup , select self Hosted for running on premises activities**



Integration runtime setup

Network environment:

Choose the network environment of the data source / destination or external compute to which the integration runtime will connect to for data flows, data movement or dispatch activities:

Azure
Use this for running data flows, data movement, external and pipeline activities in a fully managed, serverless compute in Azure.

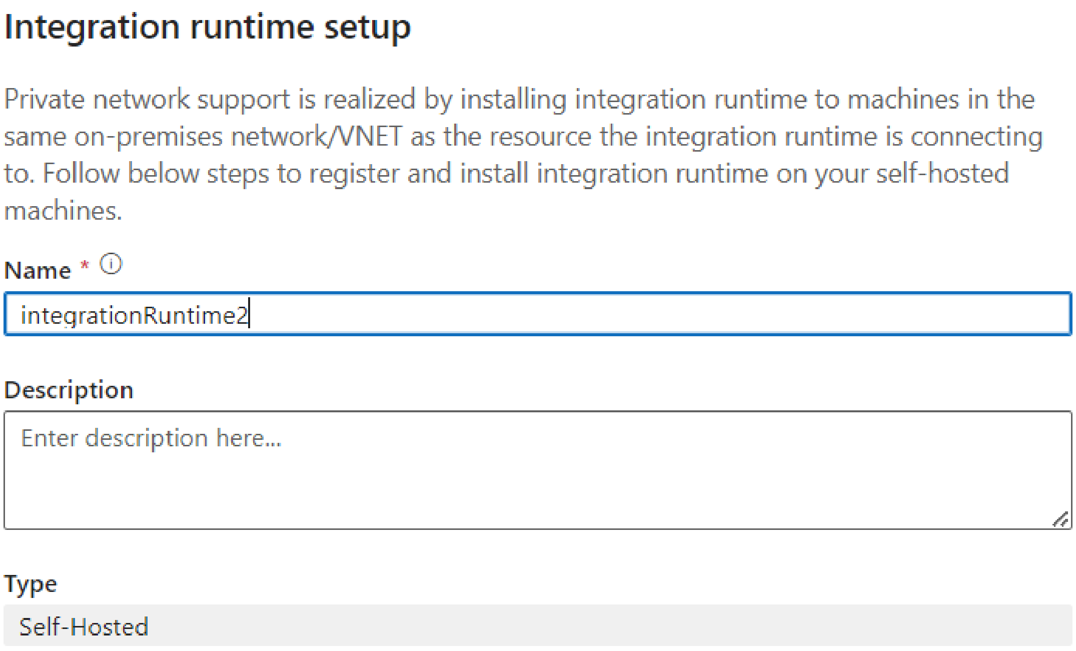
Self-Hosted
Use this for running activities in an on-premises / private network
View more

External Resources:

You can use an existing self-hosted integration runtime that exists in another resource. This way you can reuse your existing infrastructure where self-hosted integration runtime is setup.

Linked Self-Hosted
Learn more

- **Give a name for the integration runtime setup**



Integration runtime setup

Private network support is realized by installing integration runtime to machines in the same on-premises network/VNET as the resource the integration runtime is connecting to. Follow below steps to register and install integration runtime on your self-hosted machines.

Name *

integrationRuntime2

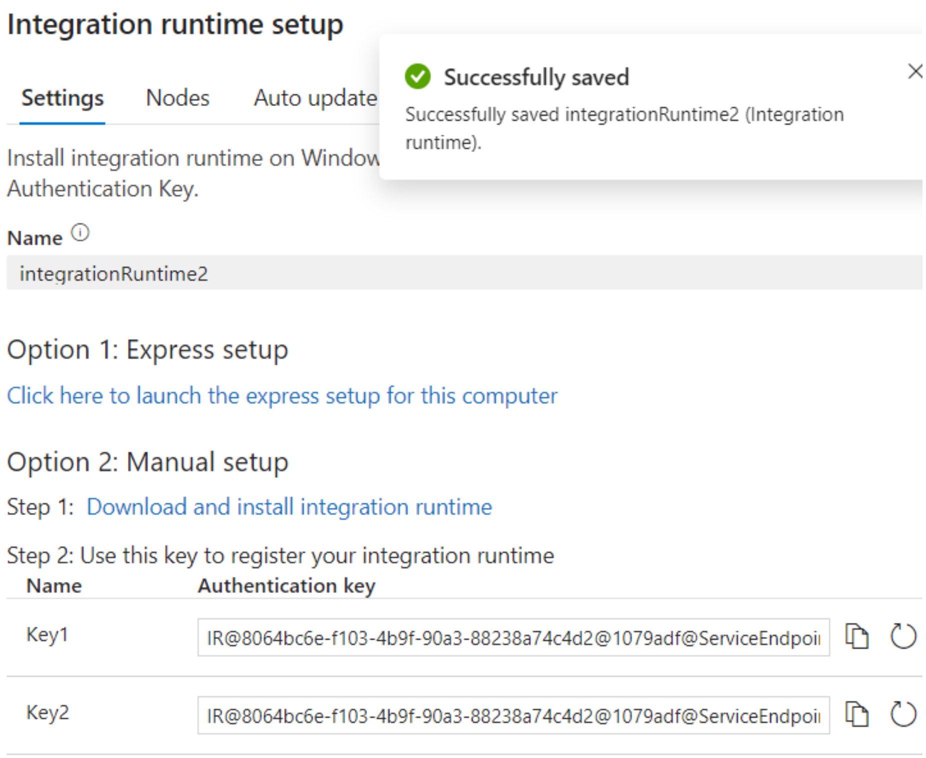
Description

Enter description here...

Type

Self-Hosted

- **In this setup, click to launch express setup for this computer.**



Integration runtime setup

Successfully saved

Successfully saved integrationRuntime2 (Integration runtime).

Install integration runtime on Window
Authentication Key.

Name

integrationRuntime2

Option 1: Express setup

Click here to launch the express setup for this computer

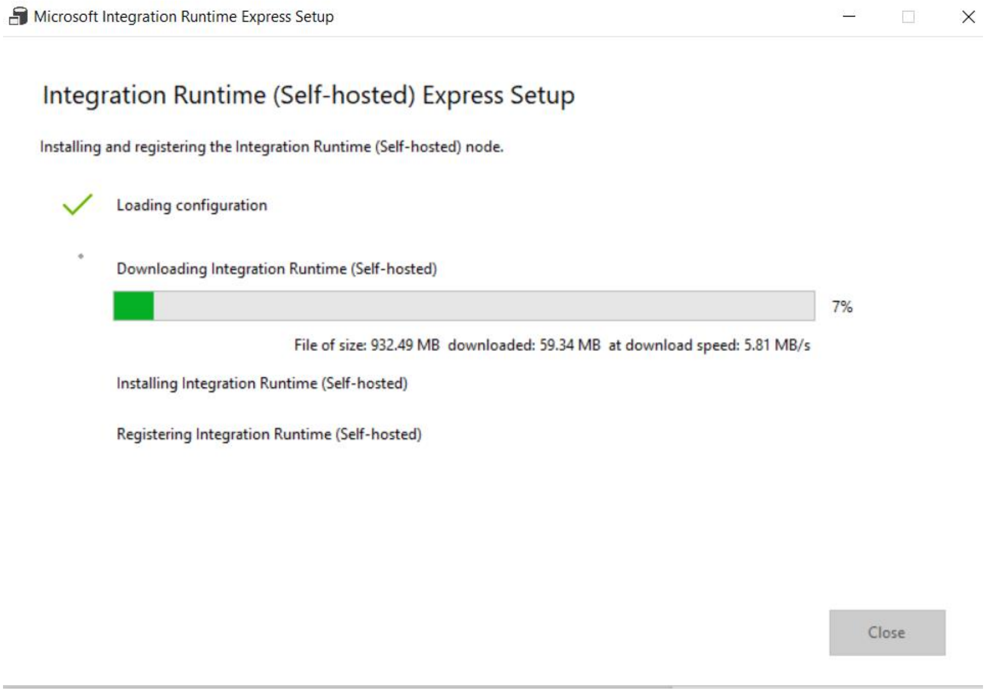
Option 2: Manual setup

Step 1: Download and install integration runtime

Step 2: Use this key to register your integration runtime

Name	Authentication key
Key1	IR@8064bc6e-f103-4b9f-90a3-88238a74c4d2@1079adf@ServiceEndpoi
Key2	IR@8064bc6e-f103-4b9f-90a3-88238a74c4d2@1079adf@ServiceEndpoi

- **Install this to connect with the premise database**



Microsoft Integration Runtime Express Setup

Integration Runtime (Self-hosted) Express Setup

Installing and registering the Integration Runtime (Self-hosted) node.

✓ Loading configuration

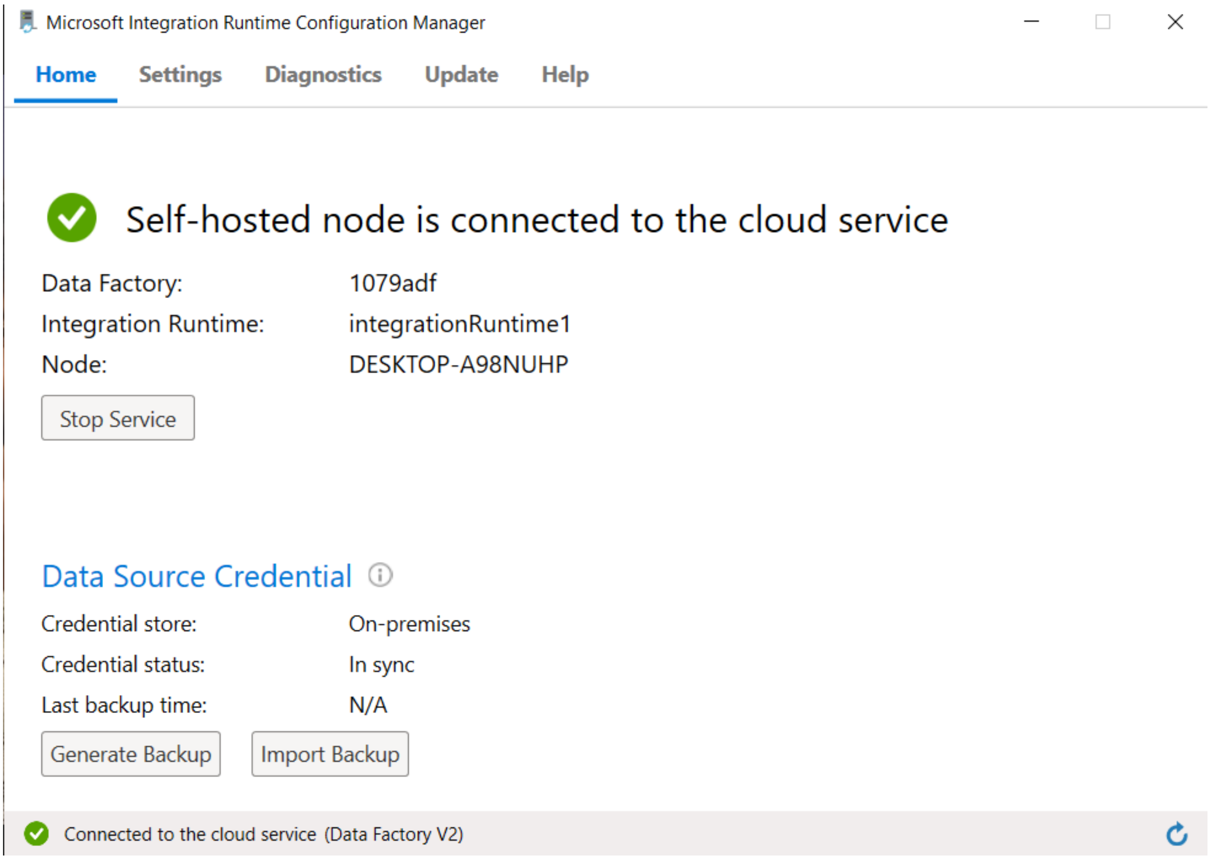
Downloading Integration Runtime (Self-hosted)

7%

File of size: 932.49 MB downloaded: 59.34 MB at download speed: 5.81 MB/s

Installing Integration Runtime (Self-hosted)

Registering Integration Runtime (Self-hosted)



Microsoft Integration Runtime Configuration Manager

✓ Self-hosted node is connected to the cloud service

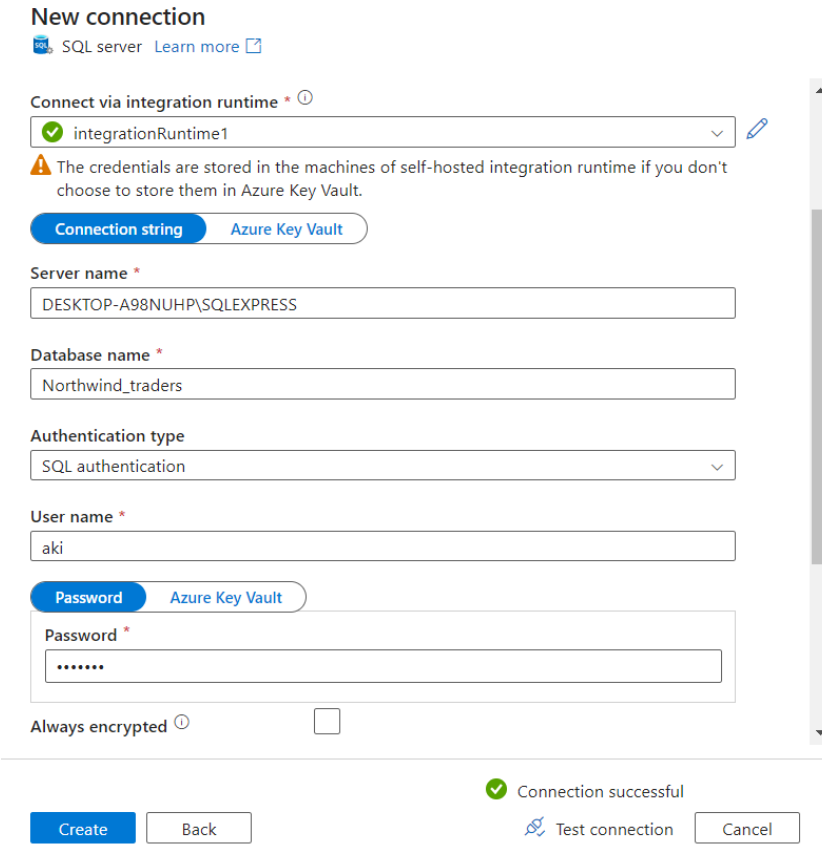
Data Factory:	1079adf
Integration Runtime:	integrationRuntime1
Node:	DESKTOP-A98NUHP

Data Source Credential

Credential store:	On-premises
Credential status:	In sync
Last backup time:	N/A

✓ Connected to the cloud service (Data Factory V2)

- **create the connection by mentioning server name and database name of the premise database**
- **select SQL authentication and enter user and password to access the database connection**



New connection

SQL server Learn more

Connect via integration runtime * integrationRuntime1

The credentials are stored in the machines of self-hosted integration runtime if you don't choose to store them in Azure Key Vault.

Server name * DESKTOP-A98NUHP\SQLEXPRESS

Database name * Northwind_traders

Authentication type SQL authentication

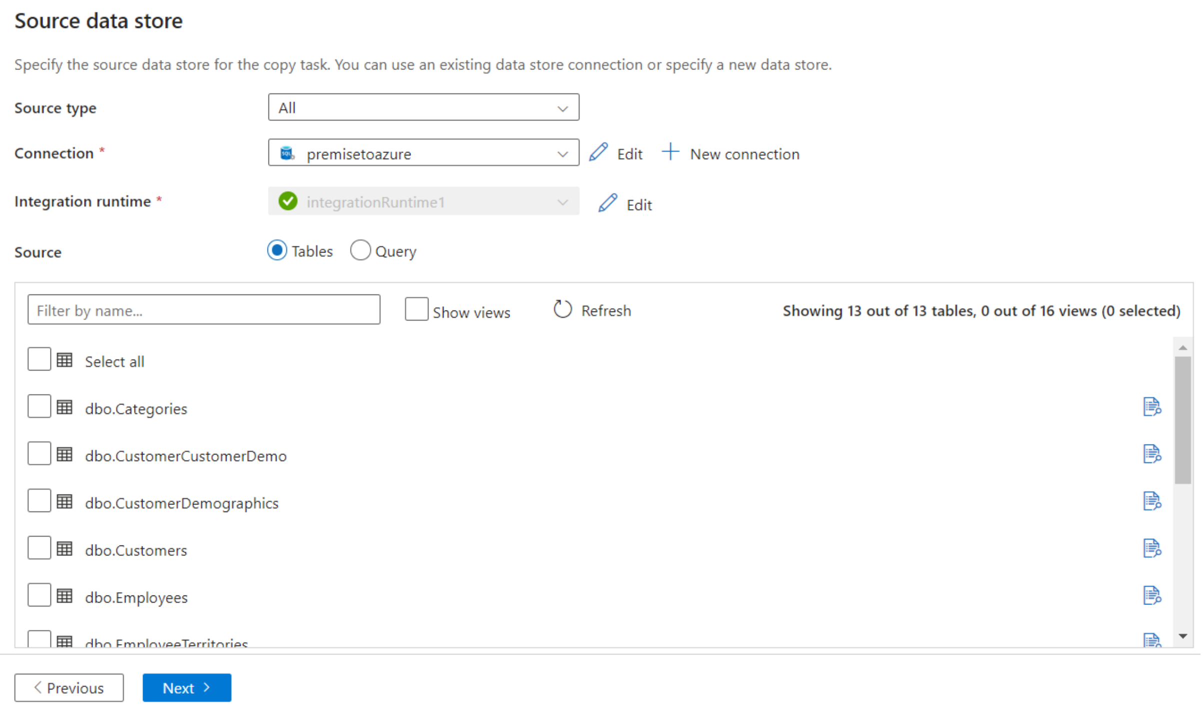
User name * aki

Password * •••••••

Always encrypted □

Connection successful

- **selected the tables in the database**



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: All

Connection *: premisetoazure

Integration runtime *: integrationRuntime1

Source: ● Tables ○ Query

□ Show views

Showing 13 out of 13 tables, 0 out of 16 views (0 selected)

□ Select all

□ dbo.Categories

□ dbo.CustomerCustomerDemo

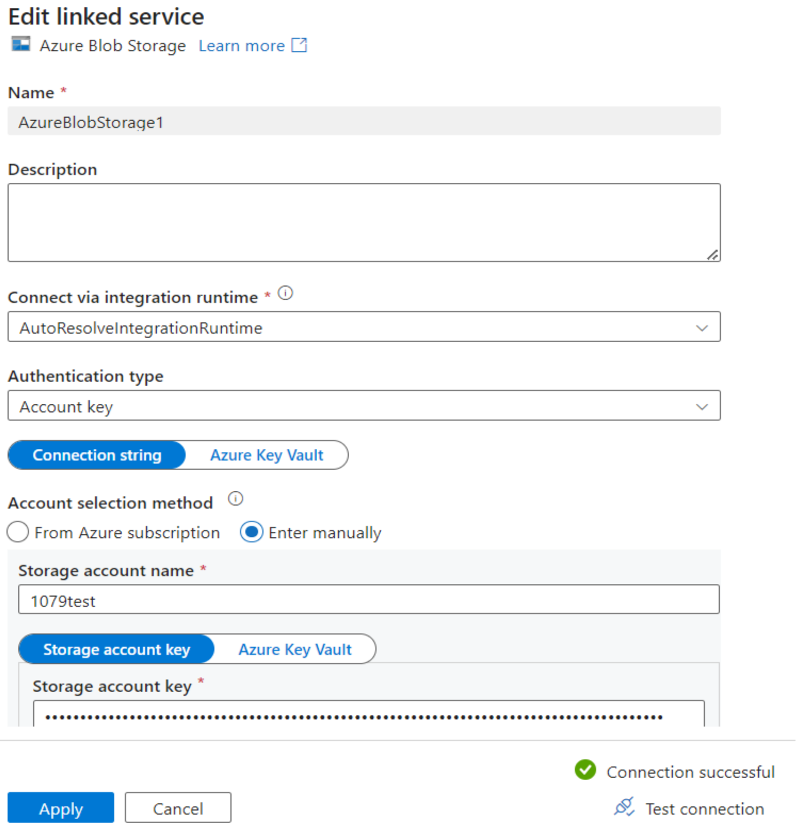
□ dbo.CustomerDemographics

□ dbo.Customers

□ dbo.Employees

□ dbo.EmployeeTerritories

- **Create a connection for destination storage blob**



Edit linked service

Azure Blob Storage

Name *: AzureBlobStorage1

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string

Account selection method: ○ From Azure subscription ● Enter manually

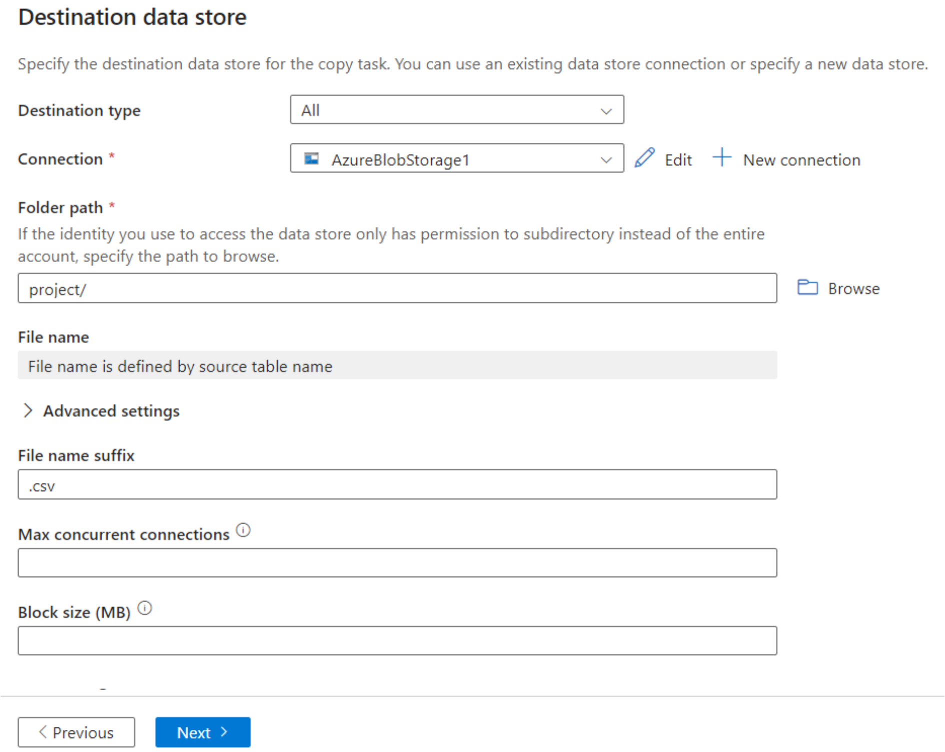
Storage account name *: 1079test

Storage account key

Storage account key *: ••

Connection successful

- **Select the container subfolder**



Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: All

Connection *: AzureBlobStorage1

Folder path *

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

project/

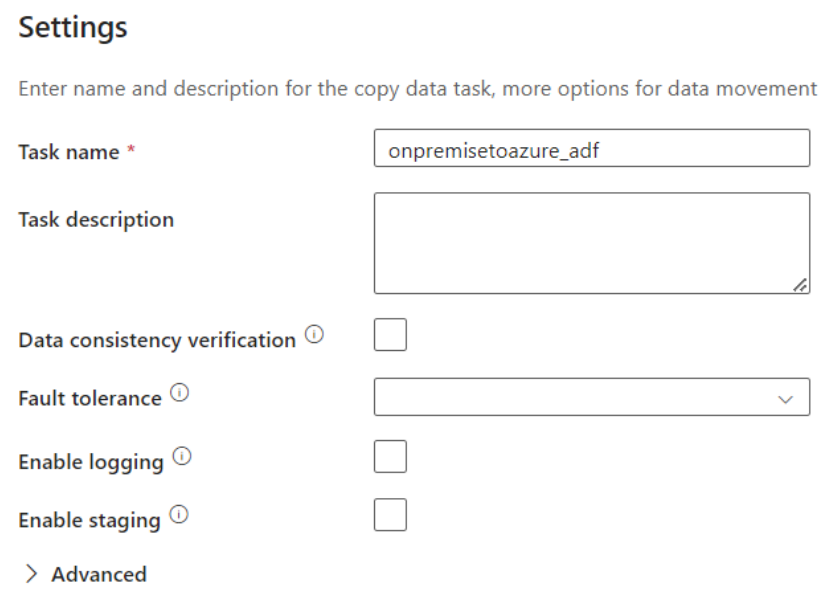
File name: File name is defined by source table name

File name suffix: .csv

Max concurrent connections:

Block size (MB):

- **Enter the data pipeline name**



Settings

Enter name and description for the copy data task, more options for data movement

Task name *: onpremisetoazure_adf

Task description:

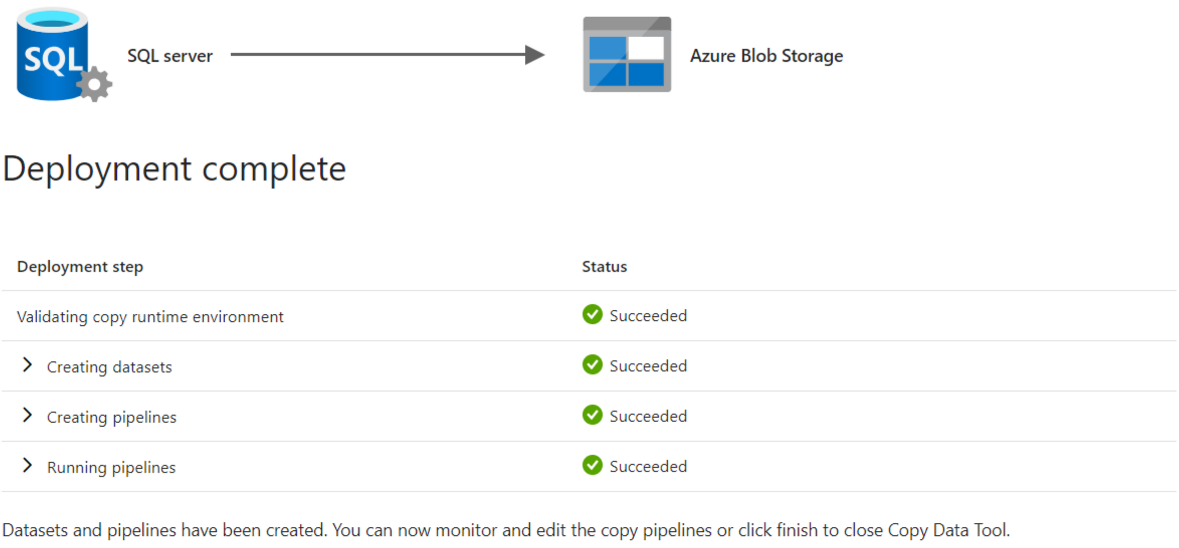
Data consistency verification: □

Fault tolerance:

Enable logging: □

Enable staging: □

- **Summary of the deployment of from SQL server to azure blob storage**

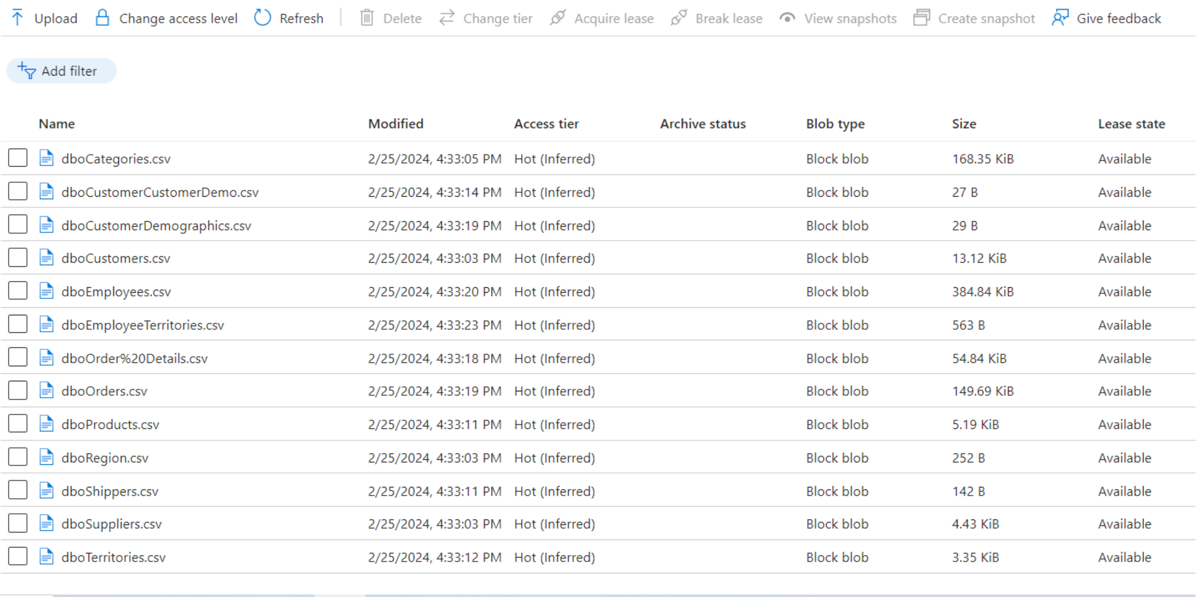


Deployment complete

Deployment step	Status
Validating copy runtime environment	✓ Succeeded
› Creating datasets	✓ Succeeded
› Creating pipelines	✓ Succeeded
› Running pipelines	✓ Succeeded

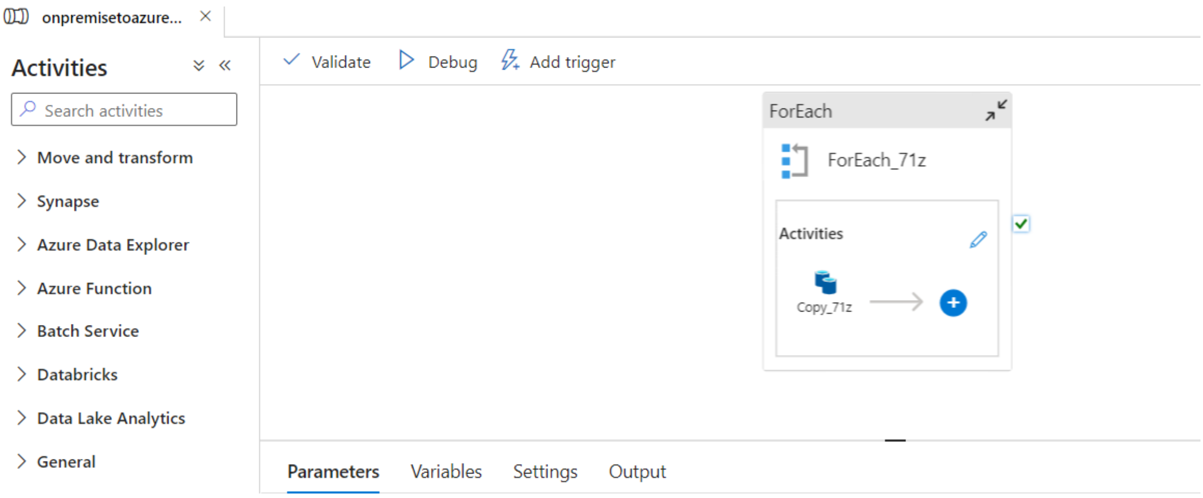
Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

- **The table is converted to CSV file and stored in the container**

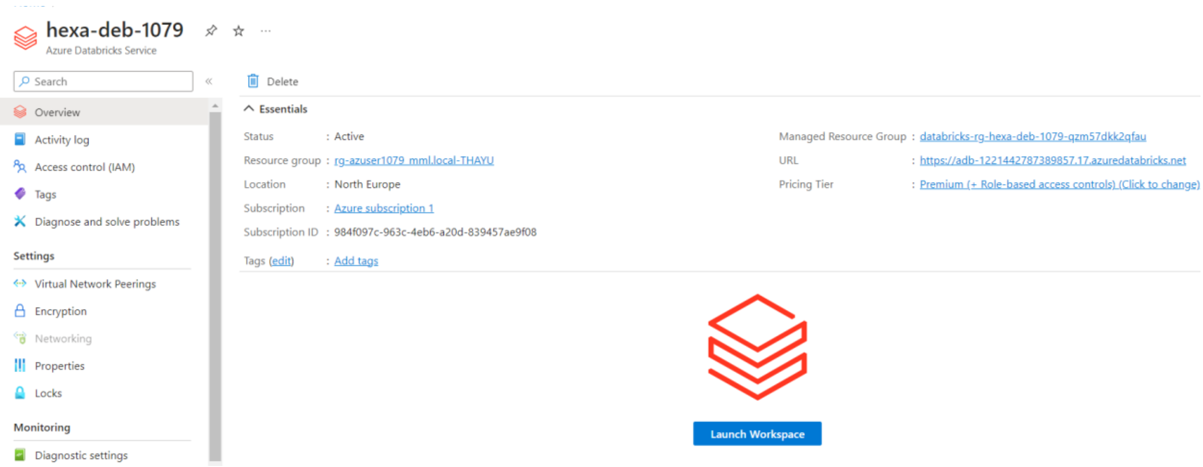


	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
□	dboCategories.csv	2/25/2024, 4:33:05 PM	Hot (Inferred)		Block blob	168.35 KiB	Available
□	dboCustomerCustomerDemo.csv	2/25/2024, 4:33:14 PM	Hot (Inferred)		Block blob	27 B	Available
□	dboCustomerDemographics.csv	2/25/2024, 4:33:19 PM	Hot (Inferred)		Block blob	29 B	Available
□	dboCustomers.csv	2/25/2024, 4:33:03 PM	Hot (Inferred)		Block blob	13.12 KiB	Available
□	dboEmployees.csv	2/25/2024, 4:33:20 PM	Hot (Inferred)		Block blob	384.84 KiB	Available
□	dboEmployeeTerritories.csv	2/25/2024, 4:33:23 PM	Hot (Inferred)		Block blob	563 B	Available
□	dboOrder%20Details.csv	2/25/2024, 4:33:18 PM	Hot (Inferred)		Block blob	54.84 KiB	Available
□	dboOrders.csv	2/25/2024, 4:33:19 PM	Hot (Inferred)		Block blob	149.69 KiB	Available
□	dboProducts.csv	2/25/2024, 4:33:11 PM	Hot (Inferred)		Block blob	5.19 KiB	Available
□	dboRegion.csv	2/25/2024, 4:33:03 PM	Hot (Inferred)		Block blob	252 B	Available
□	dboShippers.csv	2/25/2024, 4:33:11 PM	Hot (Inferred)		Block blob	142 B	Available
□	dboSuppliers.csv	2/25/2024, 4:33:03 PM	Hot (Inferred)		Block blob	4.43 KiB	Available
□	dboTerritories.csv	2/25/2024, 4:33:12 PM	Hot (Inferred)		Block blob	3.35 KiB	Available

- **Data pipeline of the activity**



- **Create a Azure databricks and launch it**



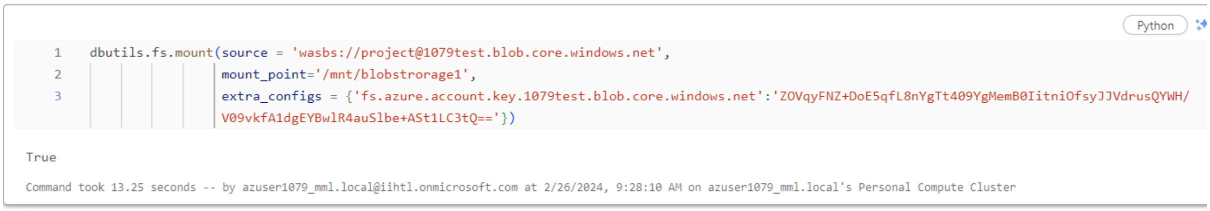
Essentials

Status : Active
Resource group : rg-azuser1079_mml.local-THAYU
Location : North Europe
Subscription : Azure subscription 1
Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08
Tags (edit) : Add tags

Managed Resource Group : databricks-rg-hexa-deb-1079-qzm57dkk2qfau
URL : https://adb-1221442787389857.17.azuredatabricks.net
Pricing Tier : Premium (+ Role-based access controls) (Click to change)

Mount blob storage to azure databricks

- **Add source line of the blob storage**
- **enter the mount points to store in databricks**
- **Configure extra_configs by mentioning the access key and pasting the key of the blob storage.**

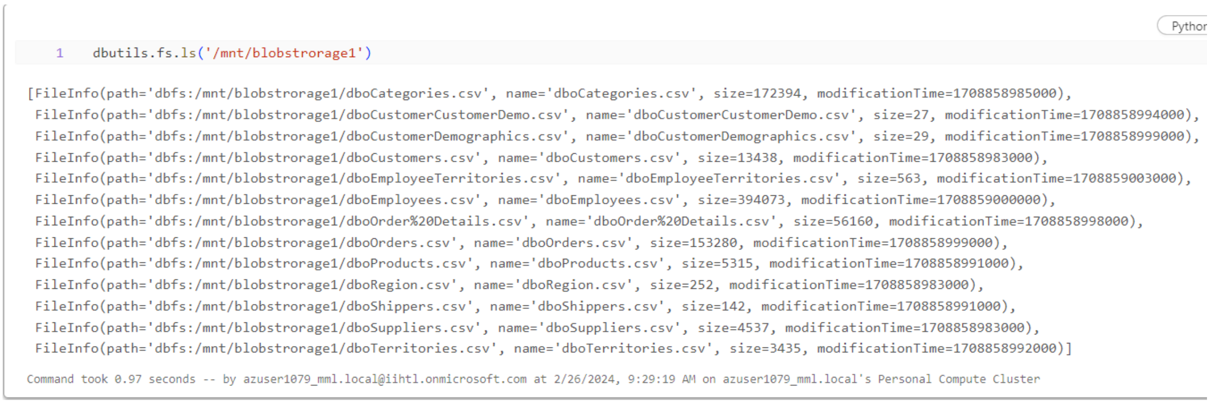


```python
dbutils.fs.mount(source = 'wasbs://project@1079test.blob.core.windows.net',
                 mount_point='/mnt/blobstrorage1',
                 extra_configs = {'fs.azure.account.key.1079test.blob.core.windows.net':'ZOVqyFNZ+DoE5qfL8nYgTt409YgMemB0IitniOfsyJJVdrusQYWH/
V09vkfA1dgEYBwlR4auSlbe+ASt1LC3tQ=='})
```

```
True

Command took 13.25 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 9:28:10 AM on azuser1079_mml.local's Personal Compute Cluster
```

- **List of the files mounted on Databricks**



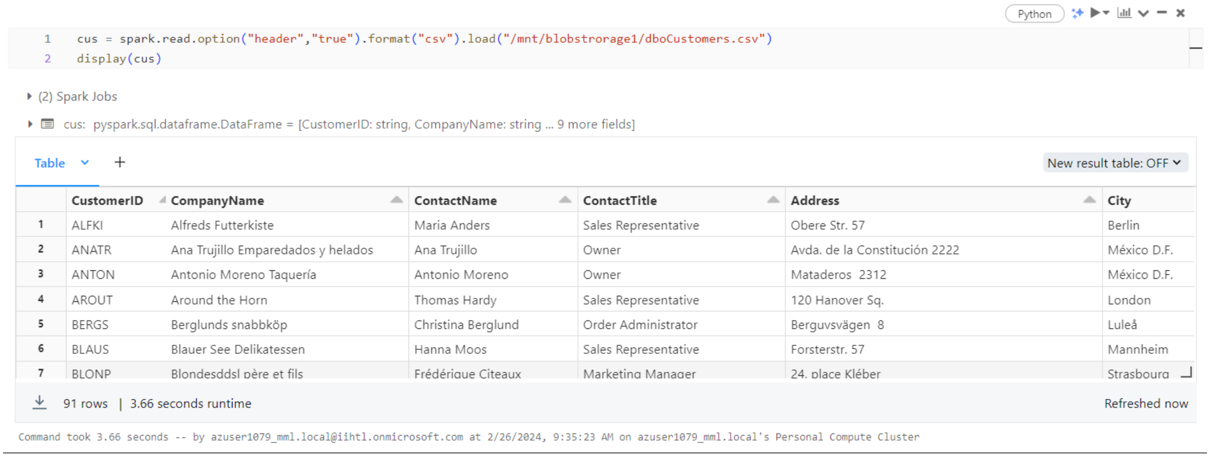
Python

```
dbutils.fs.ls('/mnt/blobstrorage1')
```

```
[FileInfo(path='dbfs:/mnt/blobstrorage1/dboCategories.csv', name='dboCategories.csv', size=172394, modificationTime=1708858985000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboCustomerCustomerDemo.csv', name='dboCustomerCustomerDemo.csv', size=27, modificationTime=1708858994000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboCustomerDemographics.csv', name='dboCustomerDemographics.csv', size=29, modificationTime=1708858999000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboCustomers.csv', name='dboCustomers.csv', size=13438, modificationTime=1708858983000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboEmployeeTerritories.csv', name='dboEmployeeTerritories.csv', size=563, modificationTime=1708859003000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboEmployees.csv', name='dboEmployees.csv', size=394073, modificationTime=1708859000000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboOrder%20Details.csv', name='dboOrder%20Details.csv', size=56160, modificationTime=1708858998000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboOrders.csv', name='dboOrders.csv', size=153280, modificationTime=1708858999000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboProducts.csv', name='dboProducts.csv', size=5315, modificationTime=1708858991000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboRegion.csv', name='dboRegion.csv', size=252, modificationTime=1708858983000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboShippers.csv', name='dboShippers.csv', size=142, modificationTime=1708858991000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboSuppliers.csv', name='dboSuppliers.csv', size=4537, modificationTime=1708858983000),
 FileInfo(path='dbfs:/mnt/blobstrorage1/dboTerritories.csv', name='dboTerritories.csv', size=3435, modificationTime=1708858992000)]

Command took 0.97 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 9:29:19 AM on azuser1079_mml.local's Personal Compute Cluster
```

- **Read the Csv file from the mount and convert into a dataframe**



```
cus = spark.read.option("header","true").format("csv").load("/mnt/blobstrorage1/dboCustomers.csv")
display(cus)
```

(2) Spark Jobs

cus: pyspark.sql.dataframe.DataFrame = [CustomerID: string, CompanyName: string ... 9 more fields]

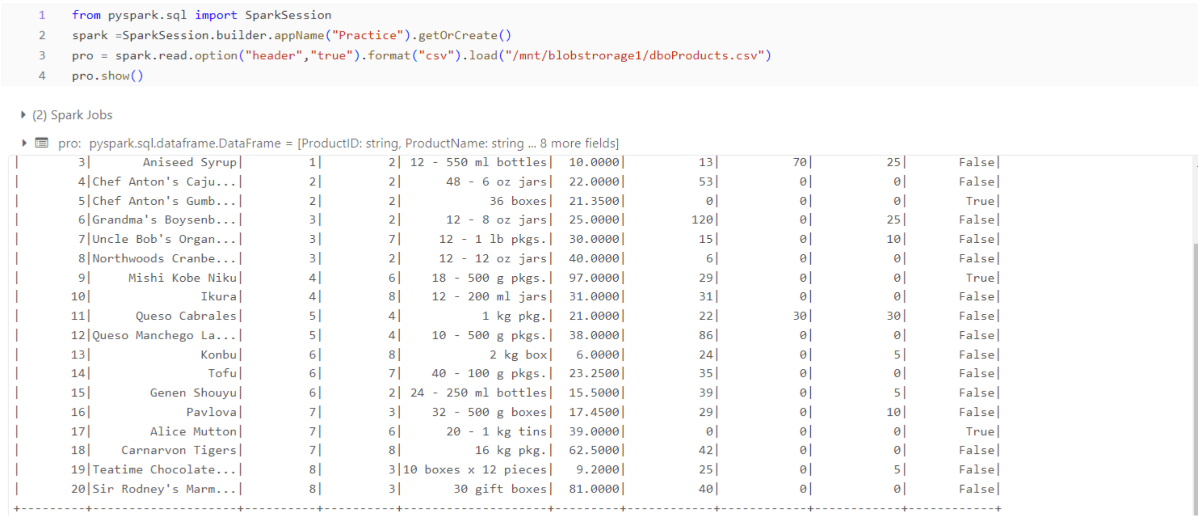
	CustomerID	CompanyName	ContactName	ContactTitle	Address	City
1	ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin
2	ANATR	Ana Trujillo Emparedados y helados	Ana Trujillo	Owner	Avda. de la Constitución 2222	México D.F.
3	ANTON	Antonio Moreno Taquería	Antonio Moreno	Owner	Mataderos 2312	México D.F.
4	AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London
5	BERGS	Berglunds snabbköp	Christina Berglund	Order Administrator	Berguvsvägen 8	Luleå
6	BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57	Mannheim
7	BLONP	Blondesddsl père et fils	Frédérique Citeaux	Marketing Manager	24, place Kléber	Strasbourg

91 rows | 3.66 seconds runtime

Refreshed now

```
Command took 3.66 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 9:35:23 AM on azuser1079_mml.local's Personal Compute Cluster
```

- **Create a spark session and display the data frame**



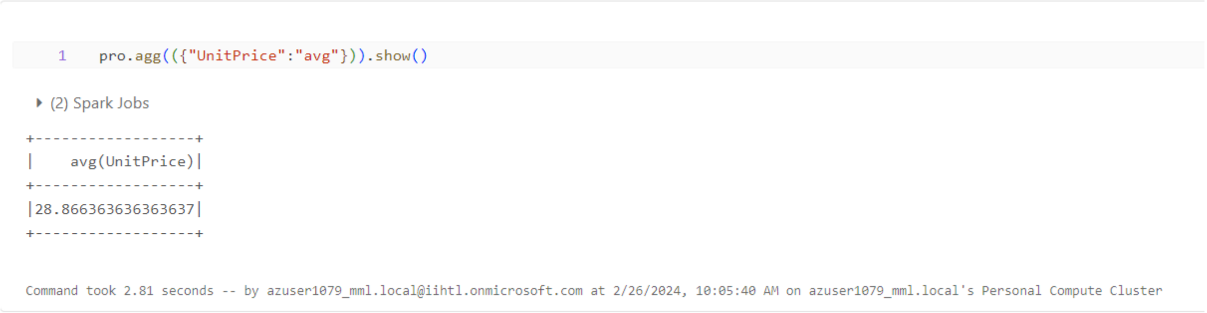
```
from pyspark.sql import SparkSession
spark =SparkSession.builder.appName("Practice").getOrCreate()
pro = spark.read.option("header","true").format("csv").load("/mnt/blobstrorage1/dboProducts.csv")
pro.show()
```

(2) Spark Jobs

pro: pyspark.sql.dataframe.DataFrame = [ProductID: string, ProductName: string ... 8 more fields]

```
|        3|       Aniseed Syrup|         1|         2| 12 - 550 ml bottles|  10.0000|          13|          70|          25|       False|
|        4|Chef Anton's Caju...|         2|         2|      48 - 6 oz jars|  22.0000|          53|           0|           0|       False|
|        5|Chef Anton's Gumb...|         2|         2|            36 boxes|  21.3500|           0|           0|           0|        True|
|        6|Grandma's Boysenb...|         3|         2|      12 - 8 oz jars|  25.0000|         120|           0|          25|       False|
|        7|Uncle Bob's Organ...|         3|         7|     12 - 1 lb pkgs.|  30.0000|          15|           0|          10|       False|
|        8|Northwoods Cranbe...|         3|         2|     12 - 12 oz jars|  40.0000|           6|           0|           0|       False|
|        9|     Mishi Kobe Niku|         4|         6|    18 - 500 g pkgs.|  97.0000|          29|           0|           0|        True|
|       10|               Ikura|         4|         8|    12 - 200 ml jars|  31.0000|          31|           0|           0|       False|
|       11|      Queso Cabrales|         5|         4|           1 kg pkg.|  21.0000|          22|          30|          30|       False|
|       12|Queso Manchego La...|         5|         4|    10 - 500 g pkgs.|  38.0000|          86|           0|           0|       False|
|       13|               Konbu|         6|         8|            2 kg box|   6.0000|          24|           0|           5|       False|
|       14|                Tofu|         6|         7|    40 - 100 g pkgs.|  23.2500|          35|           0|           0|       False|
|       15|        Genen Shouyu|         6|         2| 24 - 250 ml bottles|  15.5000|          39|           0|           5|       False|
|       16|             Pavlova|         7|         3|    32 - 500 g boxes|  17.4500|          29|           0|          10|       False|
|       17|        Alice Mutton|         7|         6|      20 - 1 kg tins|  39.0000|           0|           0|           0|        True|
|       18|    Carnarvon Tigers|         7|         8|          16 kg pkg.|  62.5000|          42|           0|           0|       False|
|       19|Teatime Chocolate...|         8|         3|10 boxes x 12 pieces|   9.2000|          25|           0|           5|       False|
|       20|Sir Rodney's Marm...|         8|         3|       30 gift boxes|  81.0000|          40|           0|           0|       False|
+---------+--------------------+----------+----------+--------------------+---------+------------+------------+------------+------------+
```

- **Aggregation of the average of the unit price**



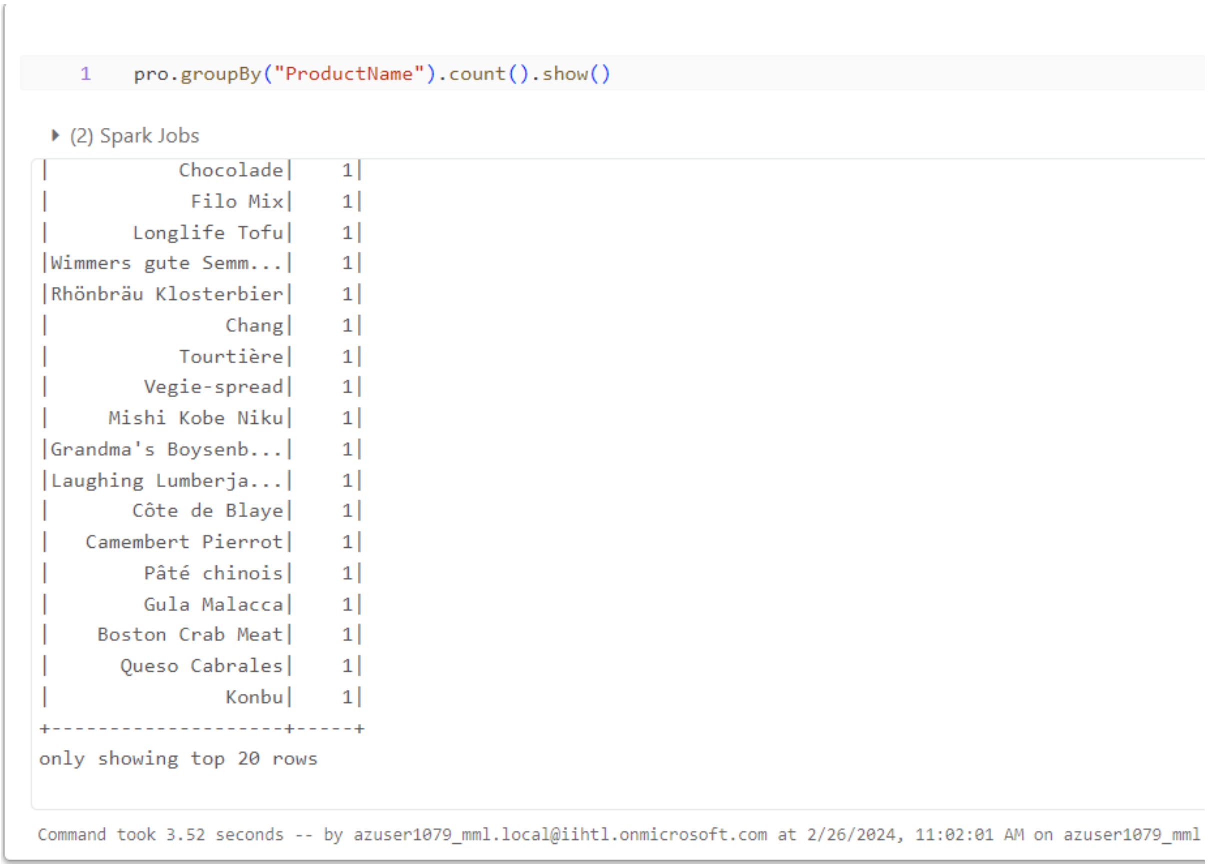
```
pro.agg(({"UnitPrice":"avg"})).show()
```

(2) Spark Jobs

```
+------------------+
|    avg(UnitPrice)|
+------------------+
|28.866363636363637|
+------------------+
```

Command took 2.81 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 10:05:40 AM on azuser1079_mml.local's Personal Compute Cluster

- **Display Grouped by product name**



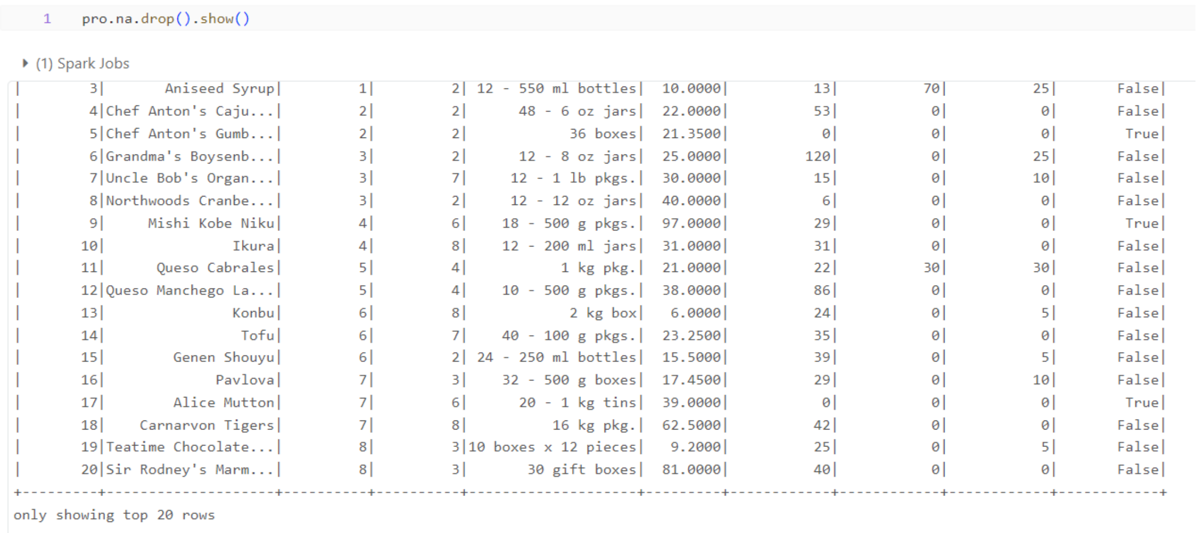
```
pro.groupBy("ProductName").count().show()
```

(2) Spark Jobs

```
|           Chocolade|    1|
|            Filo Mix|    1|
|        Longlife Tofu|    1|
|Wimmers gute Semm...|    1|
|Rhönbräu Klosterbier|    1|
|               Chang|    1|
|           Tourtière|    1|
|        Vegie-spread|    1|
|      Mishi Kobe Niku|    1|
|Grandma's Boysenb...|    1|
|Laughing Lumberja...|    1|
|        Côte de Blaye|    1|
|     Camembert Pierrot|    1|
|        Pâté chinois|    1|
|         Gula Malacca|    1|
|     Boston Crab Meat|    1|
|       Queso Cabrales|    1|
|               Konbu|    1|
+--------------------+-----+
only showing top 20 rows
```

Command took 3.52 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:02:01 AM on azuser1079_mml

- **Drop rows with Null values**

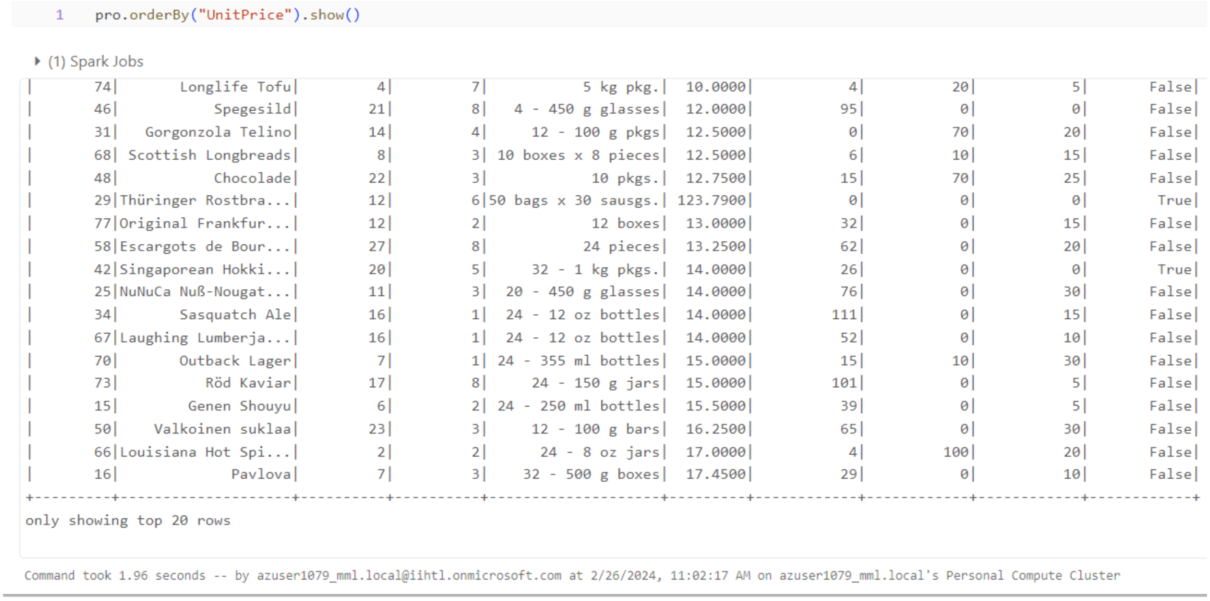


```
pro.na.drop().show()
```

(1) Spark Jobs

```
|        3|       Aniseed Syrup|         1|         2| 12 - 550 ml bottles| 10.0000|          13|          70|          25|       False|
|        4|Chef Anton's Caju...|         2|         2|      48 - 6 oz jars| 22.0000|          53|           0|           0|       False|
|        5|Chef Anton's Gumb...|         2|         2|            36 boxes| 21.3500|           0|           0|           0|        True|
|        6|Grandma's Boysenb...|         3|         2|      12 - 8 oz jars| 25.0000|         120|           0|          25|       False|
|        7|Uncle Bob's Organ...|         3|         7|     12 - 1 lb pkgs.| 30.0000|          15|           0|          10|       False|
|        8|Northwoods Cranbe...|         3|         2|     12 - 12 oz jars| 40.0000|           6|           0|           0|       False|
|        9|     Mishi Kobe Niku|         4|         6|    18 - 500 g pkgs.| 97.0000|          29|           0|           0|        True|
|       10|               Ikura|         4|         8|    12 - 200 ml jars| 31.0000|          31|           0|           0|       False|
|       11|      Queso Cabrales|         5|         4|           1 kg pkg.| 21.0000|          22|          30|          30|       False|
|       12|Queso Manchego La...|         5|         4|    10 - 500 g pkgs.| 38.0000|          86|           0|           0|       False|
|       13|               Konbu|         6|         8|            2 kg box|  6.0000|          24|           0|           5|       False|
|       14|                Tofu|         6|         7|    40 - 100 g pkgs.| 23.2500|          35|           0|           0|       False|
|       15|        Genen Shouyu|         6|         2| 24 - 250 ml bottles| 15.5000|          39|           0|           5|       False|
|       16|             Pavlova|         7|         3|    32 - 500 g boxes| 17.4500|          29|           0|          10|       False|
|       17|        Alice Mutton|         7|         6|      20 - 1 kg tins| 39.0000|           0|           0|           0|        True|
|       18|    Carnarvon Tigers|         7|         8|          16 kg pkg.| 62.5000|          42|           0|           0|       False|
|       19|Teatime Chocolate...|         8|         3|10 boxes x 12 pieces|  9.2000|          25|           0|           5|       False|
|       20|Sir Rodney's Marm...|         8|         3|       30 gift boxes| 81.0000|          40|           0|           0|       False|
+---------+--------------------+----------+----------+--------------------+--------+------------+------------+------------+------------+
only showing top 20 rows
```

- **Display data in Ascending order**



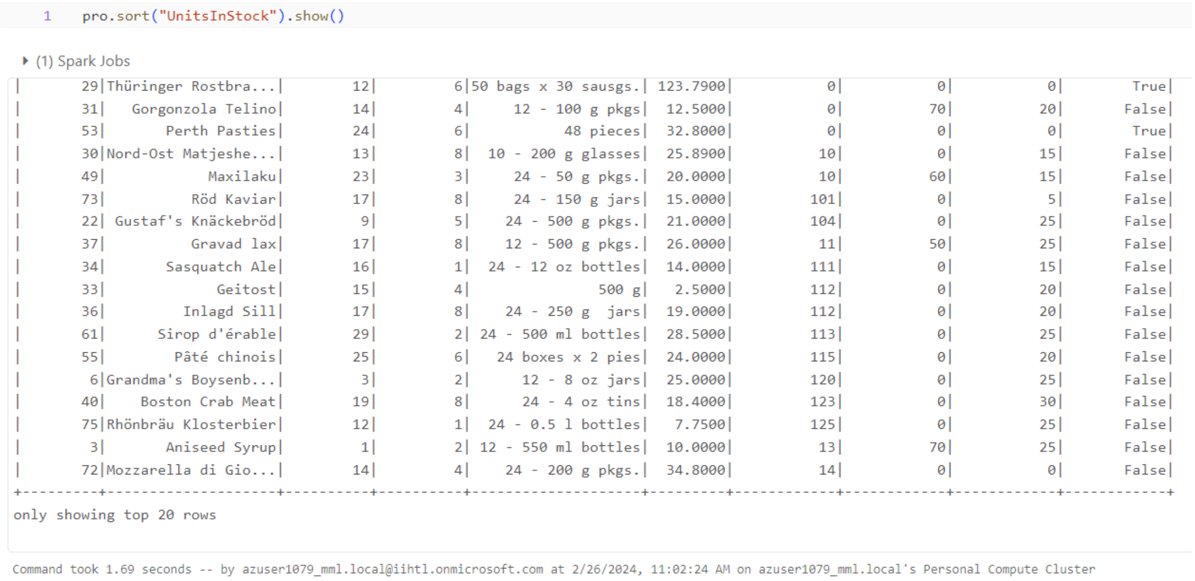
```
pro.orderBy("UnitPrice").show()
```

(1) Spark Jobs

```
|       74|       Longlife Tofu|         4|         7|           5 kg pkg.| 10.0000|           4|          20|           5|       False|
|       46|           Spegesild|        21|         8|   4 - 450 g glasses| 12.0000|          95|           0|           0|       False|
|       31|   Gorgonzola Telino|        14|         4|     12 - 100 g pkgs| 12.5000|           0|          70|          20|       False|
|       68| Scottish Longbreads|         8|         3| 10 boxes x 8 pieces| 12.5000|           6|          10|          15|       False|
|       48|           Chocolade|        22|         3|            10 pkgs.| 12.7500|          15|          70|          25|       False|
|       29|Thüringer Rostbra...|        12|         6|50 bags x 30 sausgs.|123.7900|           0|           0|           0|        True|
|       77|Original Frankfur...|        12|         2|            12 boxes| 13.0000|          32|           0|          15|       False|
|       58|Escargots de Bour...|        27|         8|           24 pieces| 13.2500|          62|           0|          20|       False|
|       42|Singaporean Hokki...|        20|         5|     32 - 1 kg pkgs.| 14.0000|          26|           0|           0|        True|
|       25|NuNuCa Nuß-Nougat...|        11|         3|  20 - 450 g glasses| 14.0000|          76|           0|          30|       False|
|       34|       Sasquatch Ale|        16|         1|  24 - 12 oz bottles| 14.0000|         111|           0|          15|       False|
|       67|Laughing Lumberja...|        16|         1|  24 - 12 oz bottles| 14.0000|          52|           0|          10|       False|
|       70|       Outback Lager|         7|         1| 24 - 355 ml bottles| 15.0000|          15|          10|          30|       False|
|       73|          Röd Kaviar|        17|         8|     24 - 150 g jars| 15.0000|         101|           0|           5|       False|
|       15|        Genen Shouyu|         6|         2| 24 - 250 ml bottles| 15.5000|          39|           0|           5|       False|
|       50|    Valkoinen suklaa|        23|         3|     12 - 100 g bars| 16.2500|          65|           0|          30|       False|
|       66|Louisiana Hot Spi...|         2|         2|      24 - 8 oz jars| 17.0000|           4|         100|          20|       False|
|       16|             Pavlova|         7|         3|    32 - 500 g boxes| 17.4500|          29|           0|          10|       False|
+---------+--------------------+----------+----------+--------------------+--------+------------+------------+------------+------------+
only showing top 20 rows
```

Command took 1.96 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:02:17 AM on azuser1079_mml.local's Personal Compute Cluster

- **Sort the data based on unit in stock**



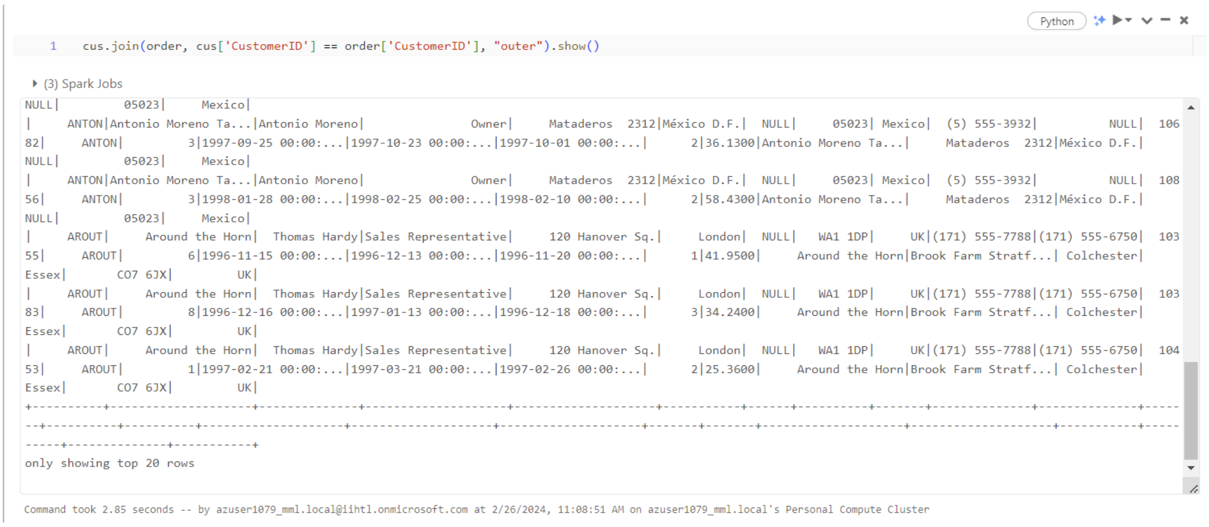
```
pro.sort("UnitsInStock").show()
```

▸ (1) Spark Jobs

```
|       29|Thüringer Rostbra...|         12|         6|50 bags x 30 sausgs.| 123.7900|           0|           0|           0|        True|
|       31|   Gorgonzola Telino|         14|         4|      12 - 100 g pkgs|  12.5000|           0|          70|          20|       False|
|       53|        Perth Pasties|         24|         6|           48 pieces|  32.8000|           0|           0|           0|        True|
|       30|Nord-Ost Matjeshe...|         13|         8|  10 - 200 g glasses|  25.8900|          10|           0|          15|       False|
|       49|            Maxilaku|         23|         3|     24 - 50 g pkgs.|  20.0000|          10|          60|          15|       False|
|       73|           Röd Kaviar|         17|         8|     24 - 150 g jars|  15.0000|         101|           0|           5|       False|
|       22|  Gustaf's Knäckebröd|          9|         5|    24 - 500 g pkgs.|  21.0000|         104|           0|          25|       False|
|       37|          Gravad lax|         17|         8|    12 - 500 g pkgs.|  26.0000|          11|          50|          25|       False|
|       34|       Sasquatch Ale|         16|         1|  24 - 12 oz bottles|  14.0000|         111|           0|          15|       False|
|       33|             Geitost|         15|         4|               500 g|   2.5000|         112|           0|          20|       False|
|       36|         Inlagd Sill|         17|         8|     24 - 250 g  jars|  19.0000|         112|           0|          20|       False|
|       61|      Sirop d'érable|         29|         2| 24 - 500 ml bottles|  28.5000|         113|           0|          25|       False|
|       55|        Pâté chinois|         25|         6|   24 boxes x 2 pies|  24.0000|         115|           0|          20|       False|
|        6|Grandma's Boysenb...|          3|         2|      12 - 8 oz jars|  25.0000|         120|           0|          25|       False|
|       40|     Boston Crab Meat|         19|         8|      24 - 4 oz tins|  18.4000|         123|           0|          30|       False|
|       75|Rhönbräu Klosterbier|         12|         1|   24 - 0.5 l bottles|   7.7500|         125|           0|          25|       False|
|        3|       Aniseed Syrup|          1|         2| 12 - 550 ml bottles|  10.0000|          13|          70|          25|       False|
|       72|Mozzarella di Gio...|         14|         4|    24 - 200 g pkgs.|  34.8000|          14|           0|           0|       False|
+---------+--------------------+-----------+----------+--------------------+---------+------------+------------+------------+------------+
only showing top 20 rows
```

Command took 1.69 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:02:24 AM on azuser1079_mml.local's Personal Compute Cluster

- **Inner join of customer and order table based on customer ID**



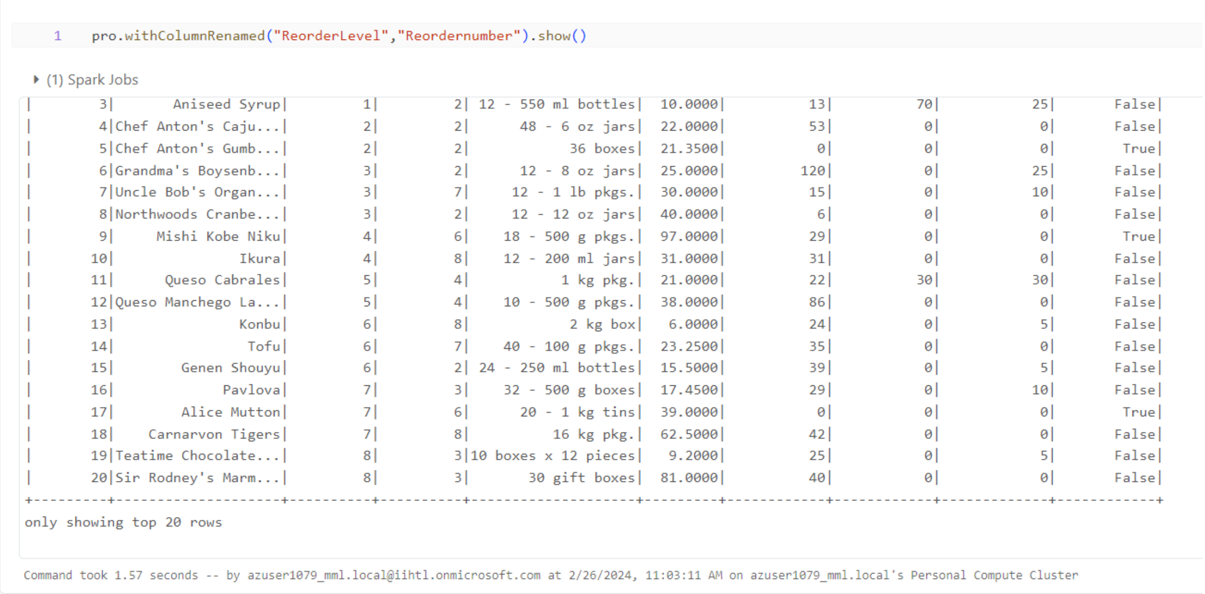
```
cus.join(order, cus['CustomerID'] == order['CustomerID'], "outer").show()
```

▸ (3) Spark Jobs

```
NULL|       05023|     Mexico|
|     ANTON|Antonio Moreno Ta...|Antonio Moreno|                Owner|     Mataderos  2312|México D.F.|  NULL|     05023| Mexico|  (5) 555-3932|           NULL|  106
82|     ANTON|         3|1997-09-25 00:00:...|1997-10-23 00:00:...|1997-10-01 00:00:...|      2|36.1300|Antonio Moreno Ta...|     Mataderos  2312|México D.F.|
NULL|       05023|     Mexico|
|     ANTON|Antonio Moreno Ta...|Antonio Moreno|                Owner|     Mataderos  2312|México D.F.|  NULL|     05023| Mexico|  (5) 555-3932|           NULL|  108
56|     ANTON|         3|1998-01-28 00:00:...|1998-02-25 00:00:...|1998-02-10 00:00:...|      2|58.4300|Antonio Moreno Ta...|     Mataderos  2312|México D.F.|
NULL|       05023|     Mexico|
|     AROUT|     Around the Horn|  Thomas Hardy|Sales Representative|     120 Hanover Sq.|     London|  NULL|   WA1 1DP|     UK|(171) 555-7788|(171) 555-6750|  103
55|     AROUT|         6|1996-11-15 00:00:...|1996-12-13 00:00:...|1996-11-20 00:00:...|      1|41.9500|     Around the Horn|Brook Farm Stratf...| Colchester|
Essex|       CO7 6JX|         UK|
|     AROUT|     Around the Horn|  Thomas Hardy|Sales Representative|     120 Hanover Sq.|     London|  NULL|   WA1 1DP|     UK|(171) 555-7788|(171) 555-6750|  103
83|     AROUT|         8|1996-12-16 00:00:...|1997-01-13 00:00:...|1996-12-18 00:00:...|      3|34.2400|     Around the Horn|Brook Farm Stratf...| Colchester|
Essex|       CO7 6JX|         UK|
|     AROUT|     Around the Horn|  Thomas Hardy|Sales Representative|     120 Hanover Sq.|     London|  NULL|   WA1 1DP|     UK|(171) 555-7788|(171) 555-6750|  104
53|     AROUT|         1|1997-02-21 00:00:...|1997-03-21 00:00:...|1997-02-26 00:00:...|      2|25.3600|     Around the Horn|Brook Farm Stratf...| Colchester|
Essex|       CO7 6JX|         UK|
+----------+--------------------+-------------+--------------------+--------------------+-----------+------+----------+-------+--------------+--------------+-----
--+----------+----------+--------------------+--------------------+--------------------+-------+-------+--------------------+--------------------+-----------+-----
-----+-------------+-----------+
only showing top 20 rows
```

Command took 2.85 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:08:51 AM on azuser1079_mml.local's Personal Compute Cluster

- **Display the table with a changed column name**

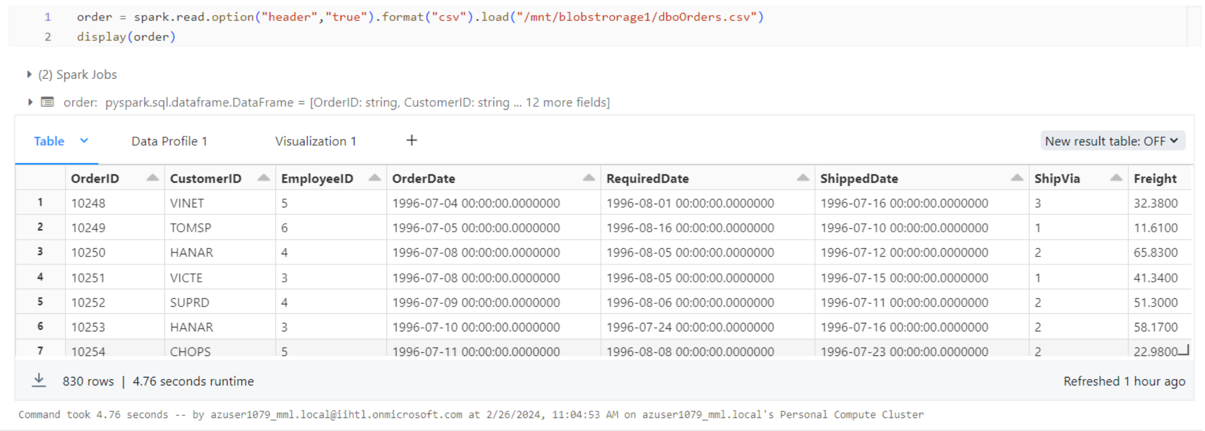


```
pro.withColumnRenamed("ReorderLevel","Reordernumber").show()
```

```
|       3|       Aniseed Syrup|         1|         2| 12 - 550 ml bottles|  10.0000|          13|          70|           25|       False|
|       4|Chef Anton's Caju...|         2|         2|     48 - 6 oz jars|  22.0000|          53|           0|            0|       False|
|       5|Chef Anton's Gumb...|         2|         2|           36 boxes|  21.3500|           0|           0|            0|        True|
|       6|Grandma's Boysenb...|         3|         2|     12 - 8 oz jars|  25.0000|         120|           0|           25|       False|
|       7|Uncle Bob's Organ...|         3|         7|     12 - 1 lb pkgs.|  30.0000|          15|           0|           10|       False|
|       8|Northwoods Cranbe...|         3|         2|    12 - 12 oz jars|  40.0000|           6|           0|            0|       False|
|       9|     Mishi Kobe Niku|         4|         6|    18 - 500 g pkgs.|  97.0000|          29|           0|            0|        True|
|      10|               Ikura|         4|         8|    12 - 200 ml jars|  31.0000|          31|           0|            0|       False|
|      11|      Queso Cabrales|         5|         4|          1 kg pkg.|  21.0000|          22|          30|           30|       False|
|      12|Queso Manchego La...|         5|         4|    10 - 500 g pkgs.|  38.0000|          86|           0|            0|       False|
|      13|               Konbu|         6|         8|           2 kg box|   6.0000|          24|           0|            5|       False|
|      14|                Tofu|         6|         7|    40 - 100 g pkgs.|  23.2500|          35|           0|            0|       False|
|      15|        Genen Shouyu|         6|         2| 24 - 250 ml bottles|  15.5000|          39|           0|            5|       False|
|      16|             Pavlova|         7|         3|   32 - 500 g boxes|  17.4500|          29|           0|           10|       False|
|      17|        Alice Mutton|         7|         6|     20 - 1 kg tins|  39.0000|           0|           0|            0|        True|
|      18|    Carnarvon Tigers|         7|         8|         16 kg pkg.|  62.5000|          42|           0|            0|       False|
|      19|Teatime Chocolate...|         8|         3|10 boxes x 12 pieces|   9.2000|          25|           0|            5|       False|
|      20|Sir Rodney's Marm...|         8|         3|      30 gift boxes|  81.0000|          40|           0|            0|       False|
+--------+--------------------+----------+----------+--------------------+---------+------------+------------+-------------+------------+
only showing top 20 rows
```

```
Command took 1.57 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:03:11 AM on azuser1079_mml.local's Personal Compute Cluster
```

- **Display dataframe of order table**



```
order = spark.read.option("header","true").format("csv").load("/mnt/blobstrorage1/dboOrders.csv")
display(order)
```

order: pyspark.sql.dataframe.DataFrame = [OrderID: string, CustomerID: string ... 12 more fields]

	OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight
1	10248	VINET	5	1996-07-04 00:00:00.0000000	1996-08-01 00:00:00.0000000	1996-07-16 00:00:00.0000000	3	32.3800
2	10249	TOMSP	6	1996-07-05 00:00:00.0000000	1996-08-16 00:00:00.0000000	1996-07-10 00:00:00.0000000	1	11.6100
3	10250	HANAR	4	1996-07-08 00:00:00.0000000	1996-08-05 00:00:00.0000000	1996-07-12 00:00:00.0000000	2	65.8300
4	10251	VICTE	3	1996-07-08 00:00:00.0000000	1996-08-05 00:00:00.0000000	1996-07-15 00:00:00.0000000	1	41.3400
5	10252	SUPRD	4	1996-07-09 00:00:00.0000000	1996-08-06 00:00:00.0000000	1996-07-11 00:00:00.0000000	2	51.3000
6	10253	HANAR	3	1996-07-10 00:00:00.0000000	1996-07-24 00:00:00.0000000	1996-07-16 00:00:00.0000000	2	58.1700
7	10254	CHOPS	5	1996-07-11 00:00:00.0000000	1996-08-08 00:00:00.0000000	1996-07-23 00:00:00.0000000	2	22.9800

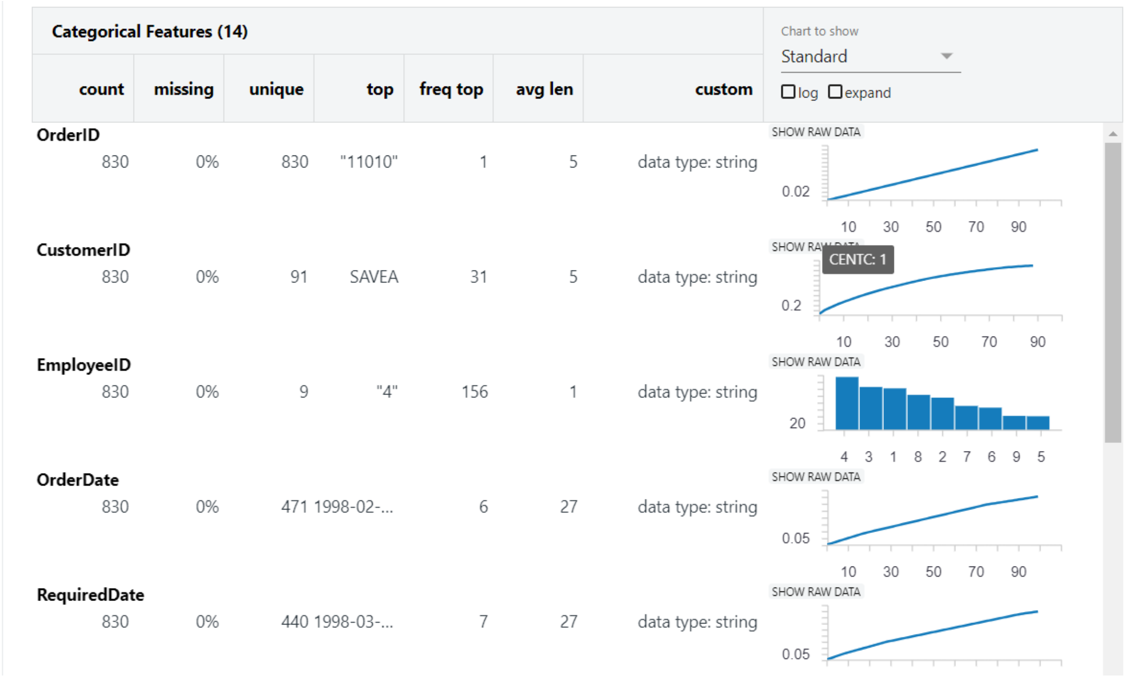
830 rows | 4.76 seconds runtime

Refreshed 1 hour ago

```
Command took 4.76 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:04:53 AM on azuser1079_mml.local's Personal Compute Cluster
```

Display the profiling of the data in the table

- **Data profiling is the process of collecting statistics and summaries of data to assess its quality and other characteristics**



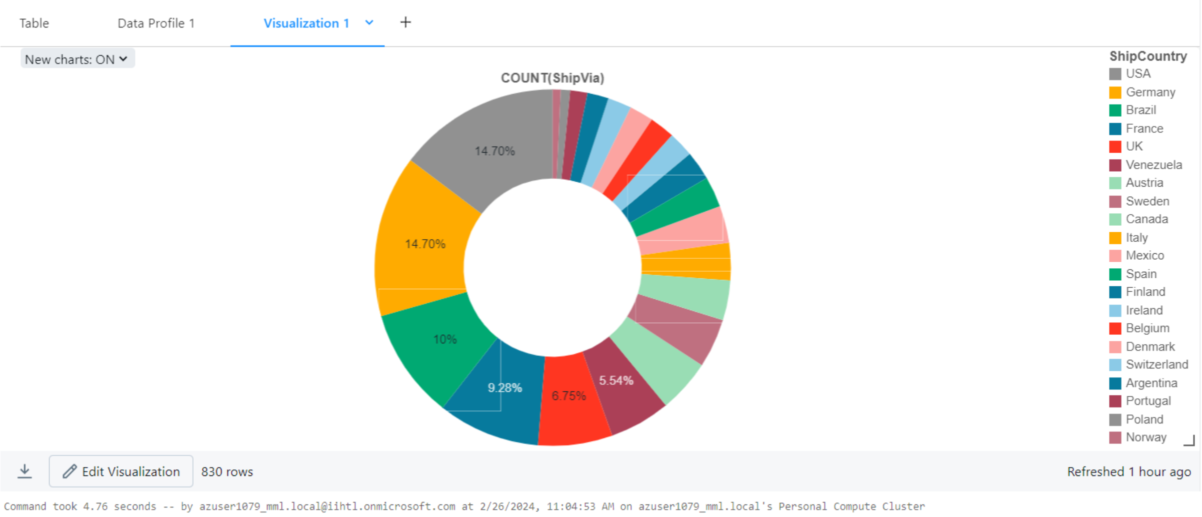
Categorical Features (14)

	count	missing	unique	top	freq top	avg len	custom
OrderID	830	0%	830	"11010"	1	5	data type: string
CustomerID	830	0%	91	SAVEA	31	5	data type: string
EmployeeID	830	0%	9	"4"	156	1	data type: string
OrderDate	830	0%	471	1998-02-...	6	27	data type: string
RequiredDate	830	0%	440	1998-03-...	7	27	data type: string

Chart to show: Standard

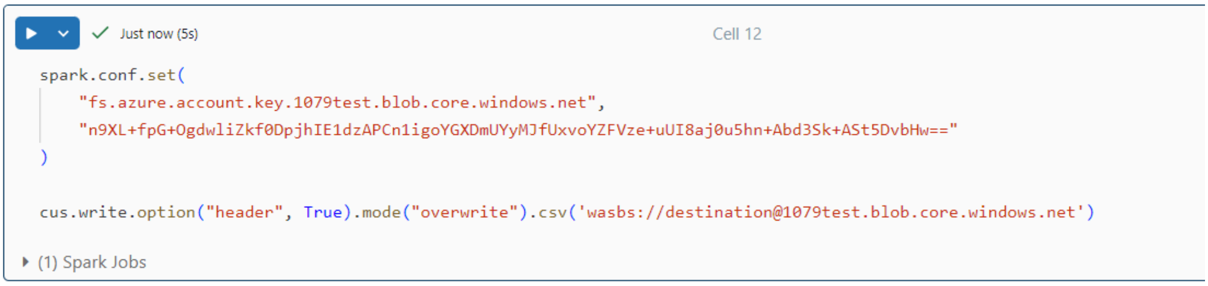
□ log □ expand

- **Visualization of the country data represented in the form of the pie chart.**



```
Command took 4.76 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/26/2024, 11:04:53 AM on azuser1079_mml.local's Personal Compute Cluster
```

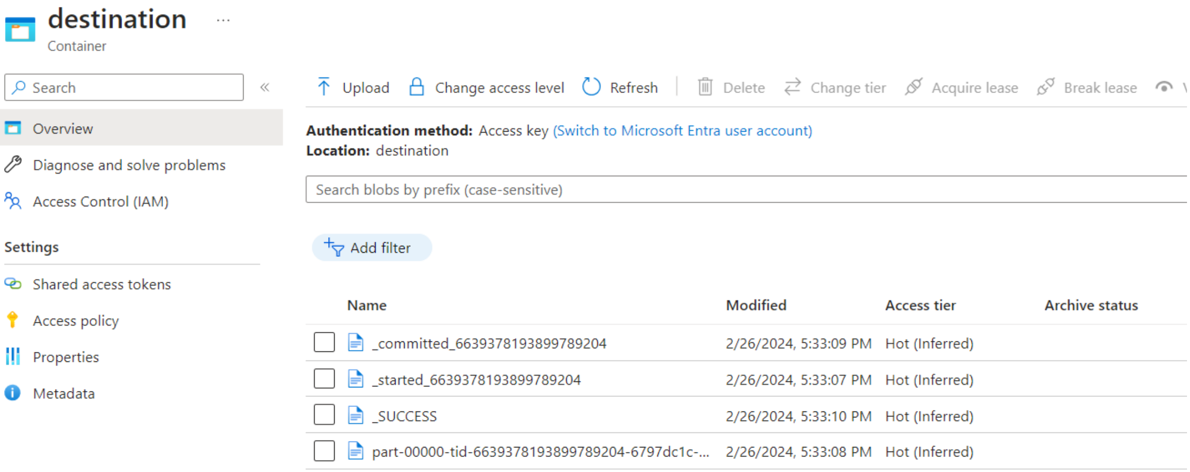
- **Transfer of the dataframe which we manipulated to a blob storage thorough access key.**



```
spark.conf.set(
    "fs.azure.account.key.1079test.blob.core.windows.net",
    "n9XL+fpG+OgdwliZkf0DpjhIE1dzAPCn1igoYGXDmUYyMJfUxvoYZFVze+uUI8aj0u5hn+Abd3Sk+ASt5DvbHw=="
)

cus.write.option("header", True).mode("overwrite").csv('wasbs://destination@1079test.blob.core.windows.net')
```

- **The Dataframe gets committed to a new storage**



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: destination

	Name	Modified	Access tier	Archive status
□	_committed_6639378193899789204	2/26/2024, 5:33:09 PM	Hot (Inferred)	
□	_started_6639378193899789204	2/26/2024, 5:33:07 PM	Hot (Inferred)	
□	_SUCCESS	2/26/2024, 5:33:10 PM	Hot (Inferred)	
□	part-00000-tid-6639378193899789204-6797dc1c-...	2/26/2024, 5:33:08 PM	Hot (Inferred)	

Conclusion

In conclusion, the project successfully implemented a robust data pipeline leveraging various Azure services. By setting up a SQL server with a database, data was securely stored on-premises. Azure Data Factory was then employed to seamlessly transfer data from the SQL database to Blob storage in the Azure cloud, ensuring scalability and reliability. Additionally, Azure Databricks played a crucial role in performing data transformations and enabling visualization using PySpark, empowering data analysts and engineers to derive insights and make informed decisions. Overall, this project demonstrates the power of Azure services in building end-to-end data solutions, from ingestion to transformation and visualization, paving the way for efficient data-driven workflows and analytics.