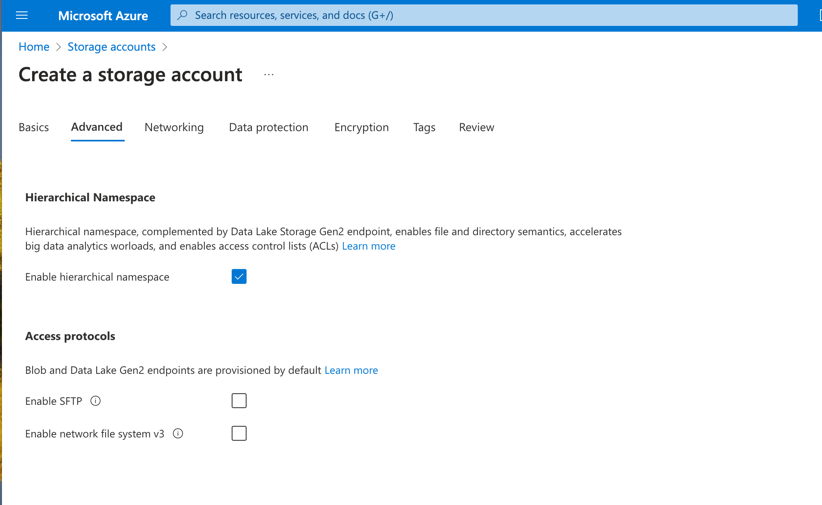
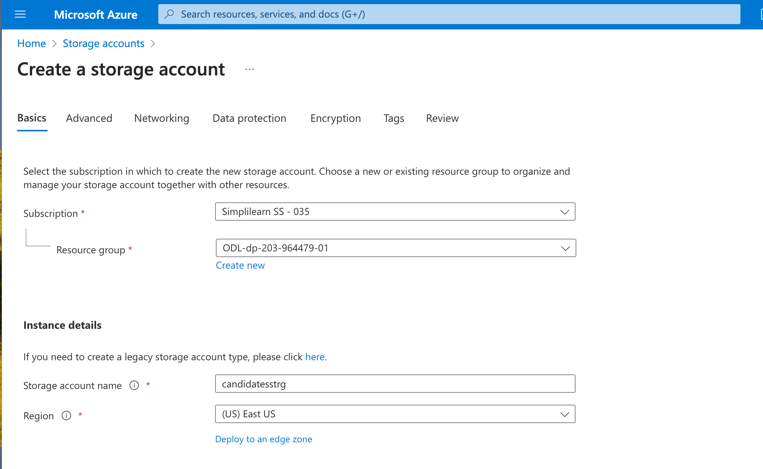
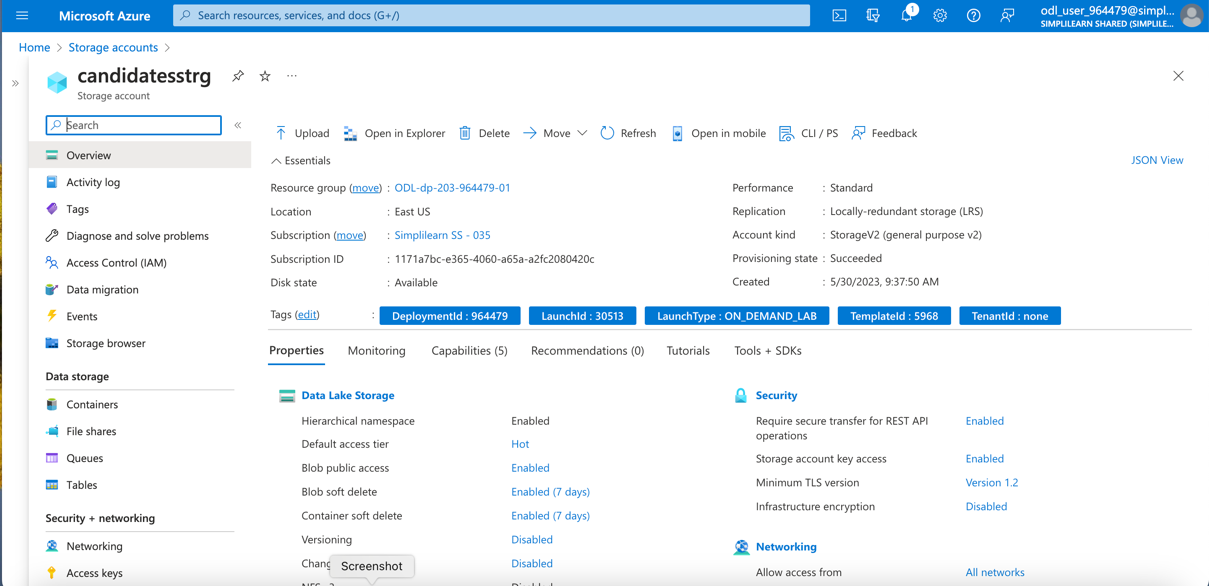
**Data Ingestion End-to-End Pipeline**

Course-end Project 1

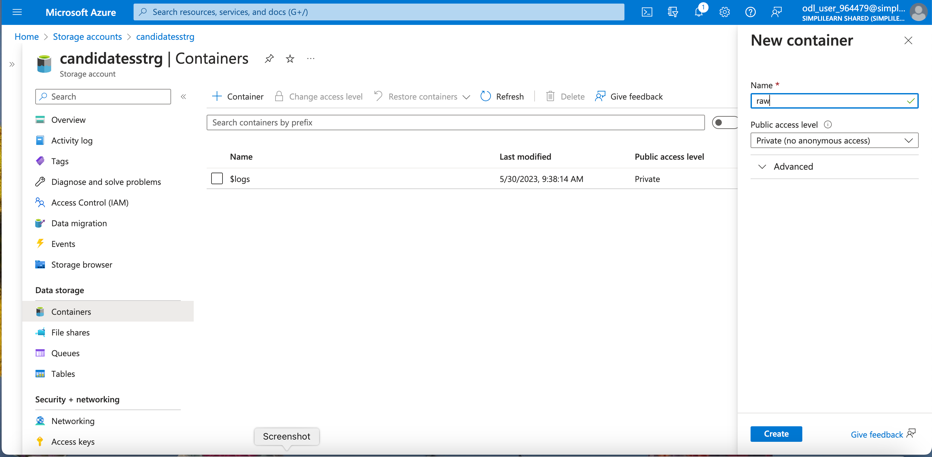
1. Create a landing storage account in Azure (go to storage Account)



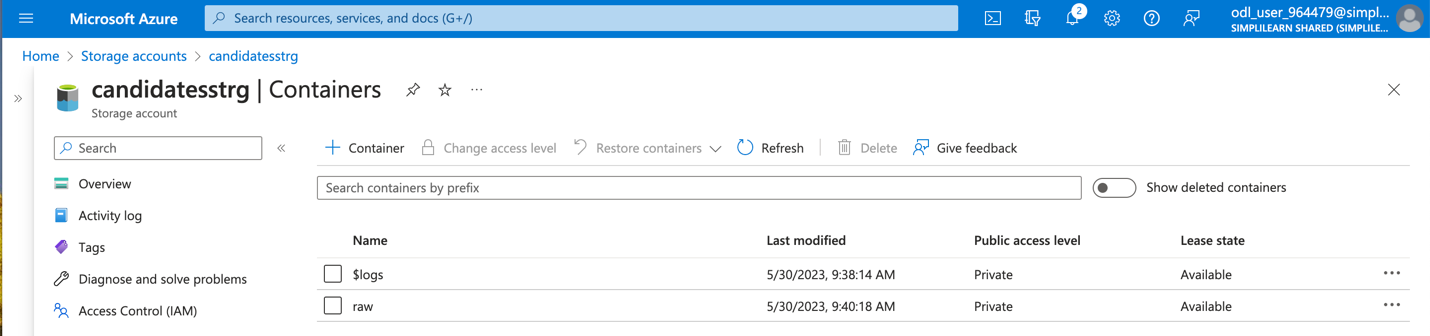




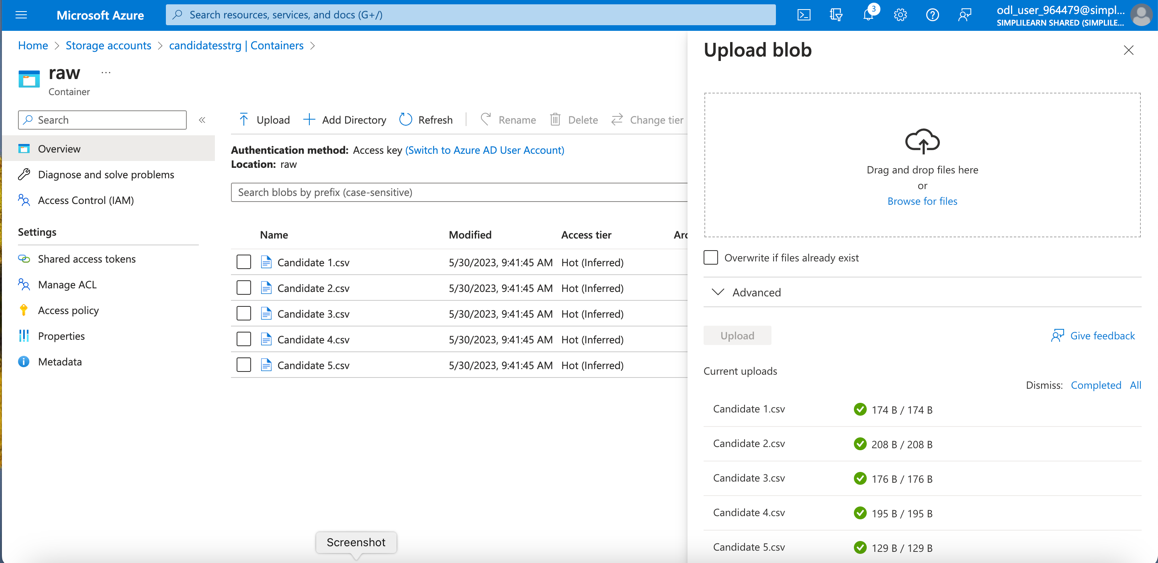
* Creating containers for the staging and landing files
  + Raw and candidates-parquet







2- Store the CSV files in the storage account

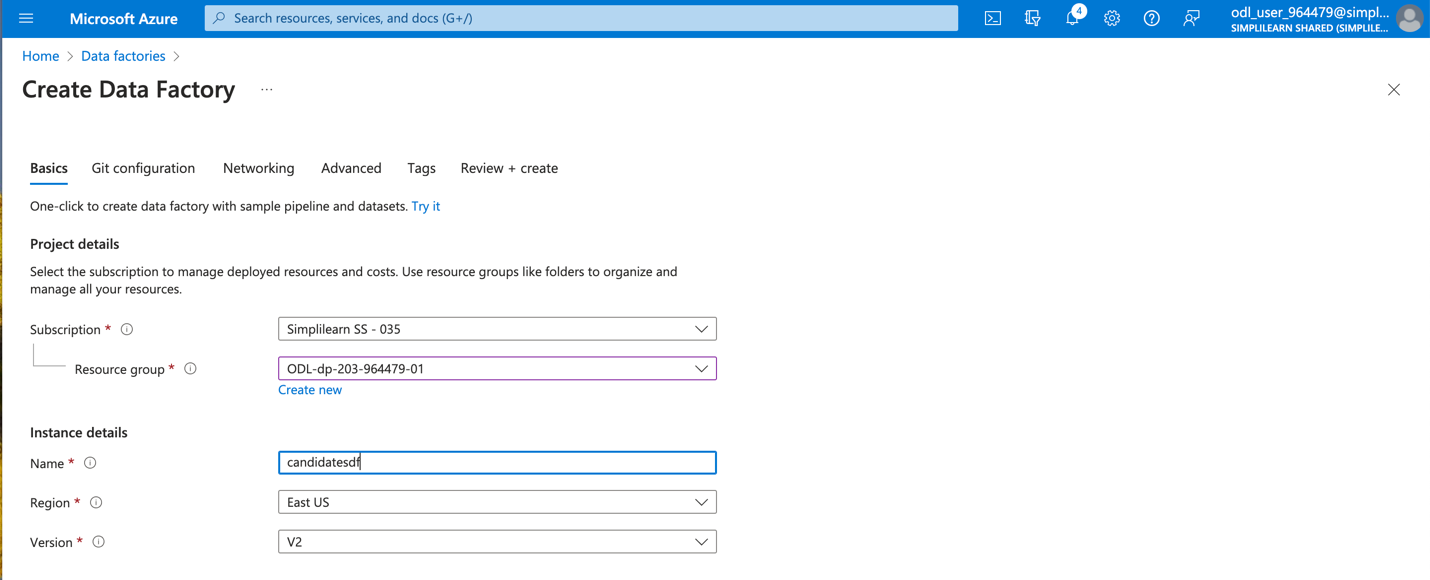


1. Create a staging storage account in Azure
   * Another folder created for Parquet file



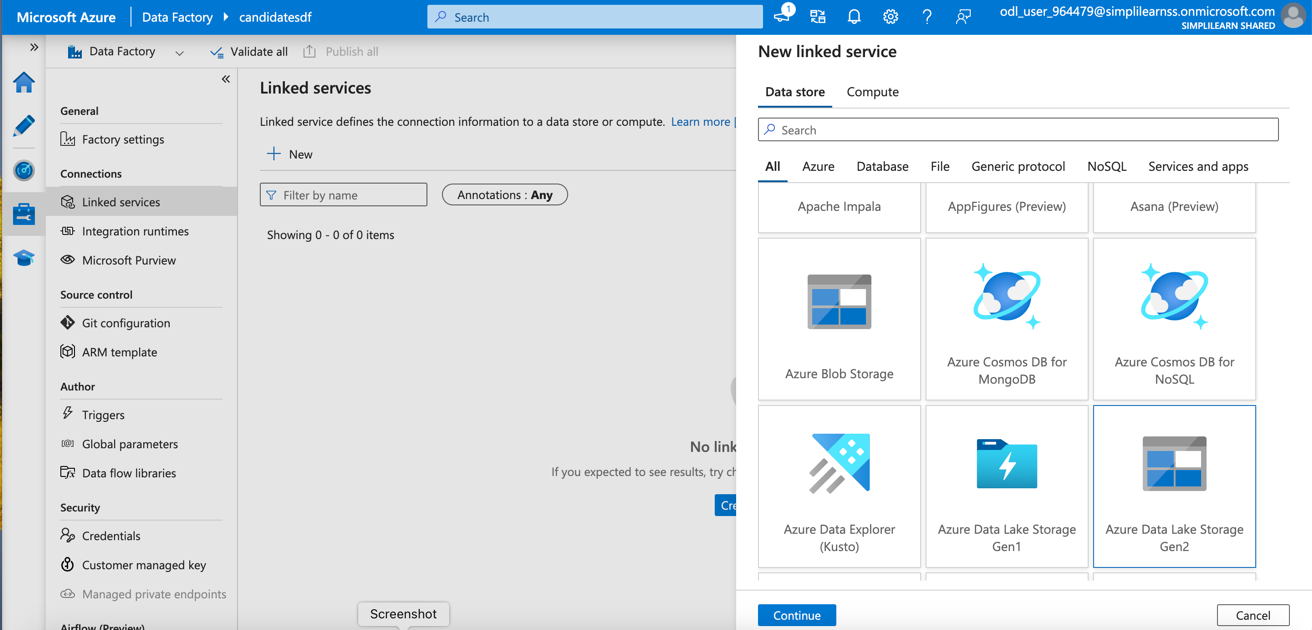


1. Create an Azure Data Factory resource and Azure Data Factory pipeline



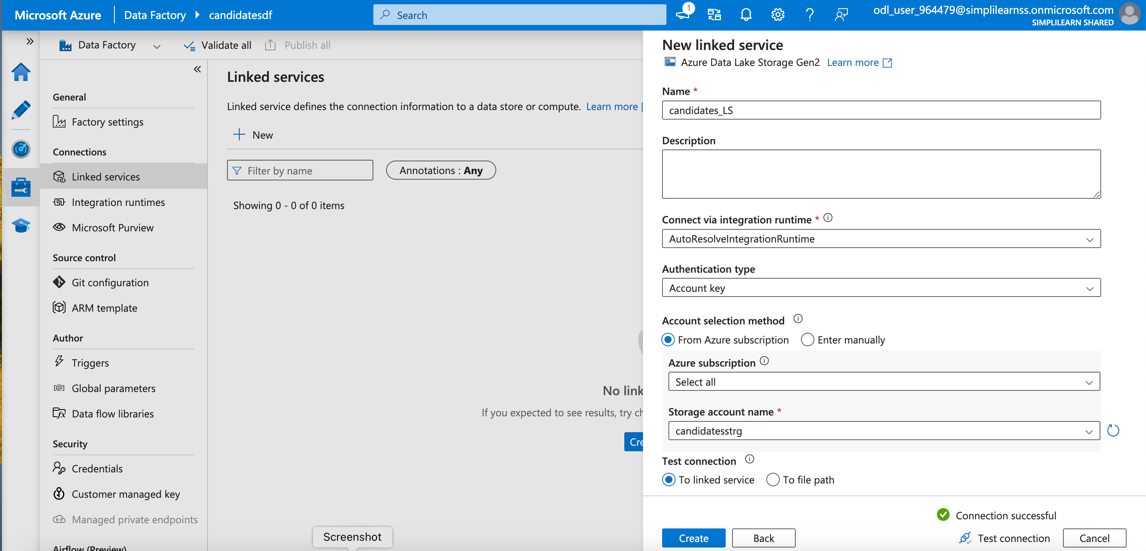




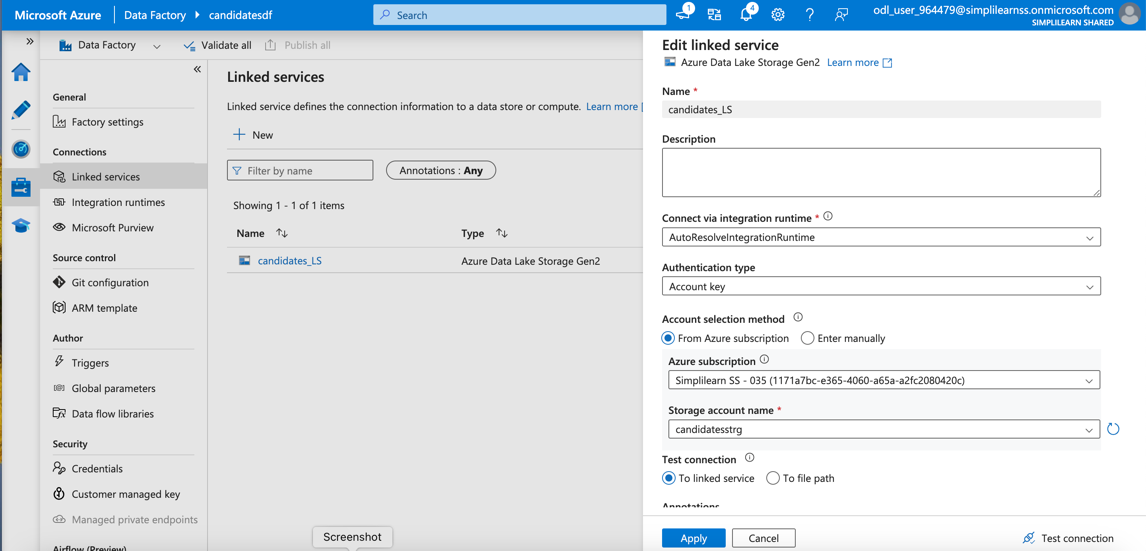




1. Create linked services for the storage accounts

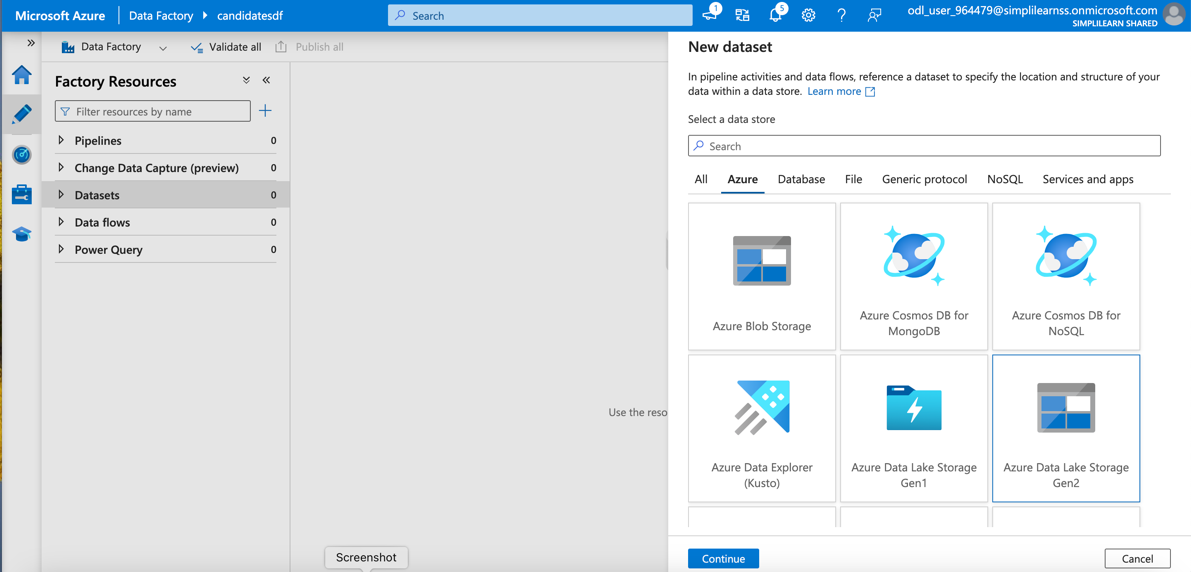




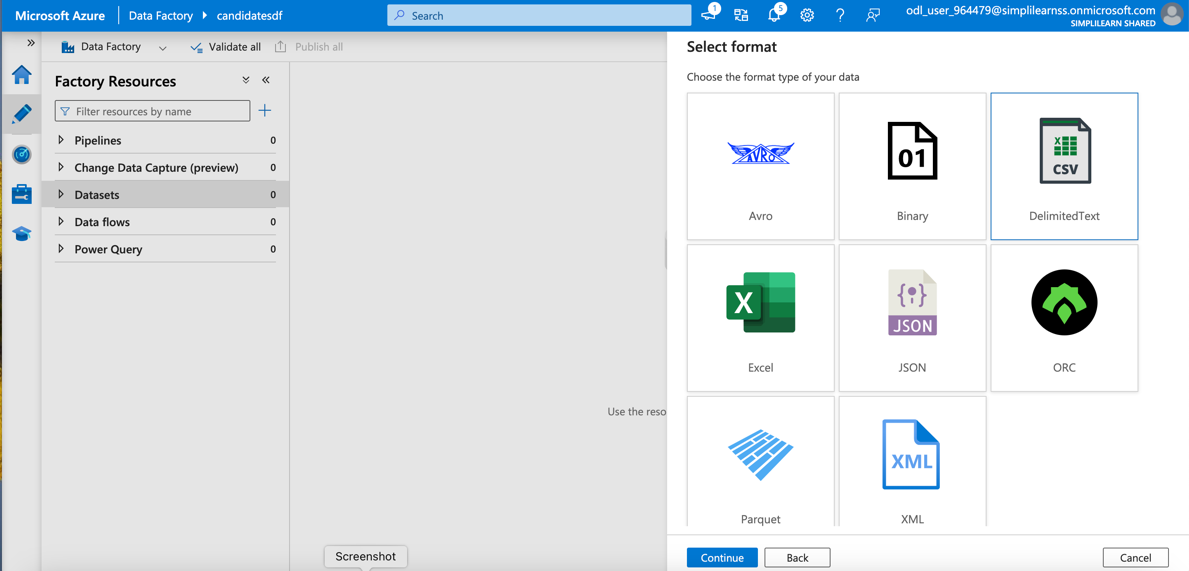




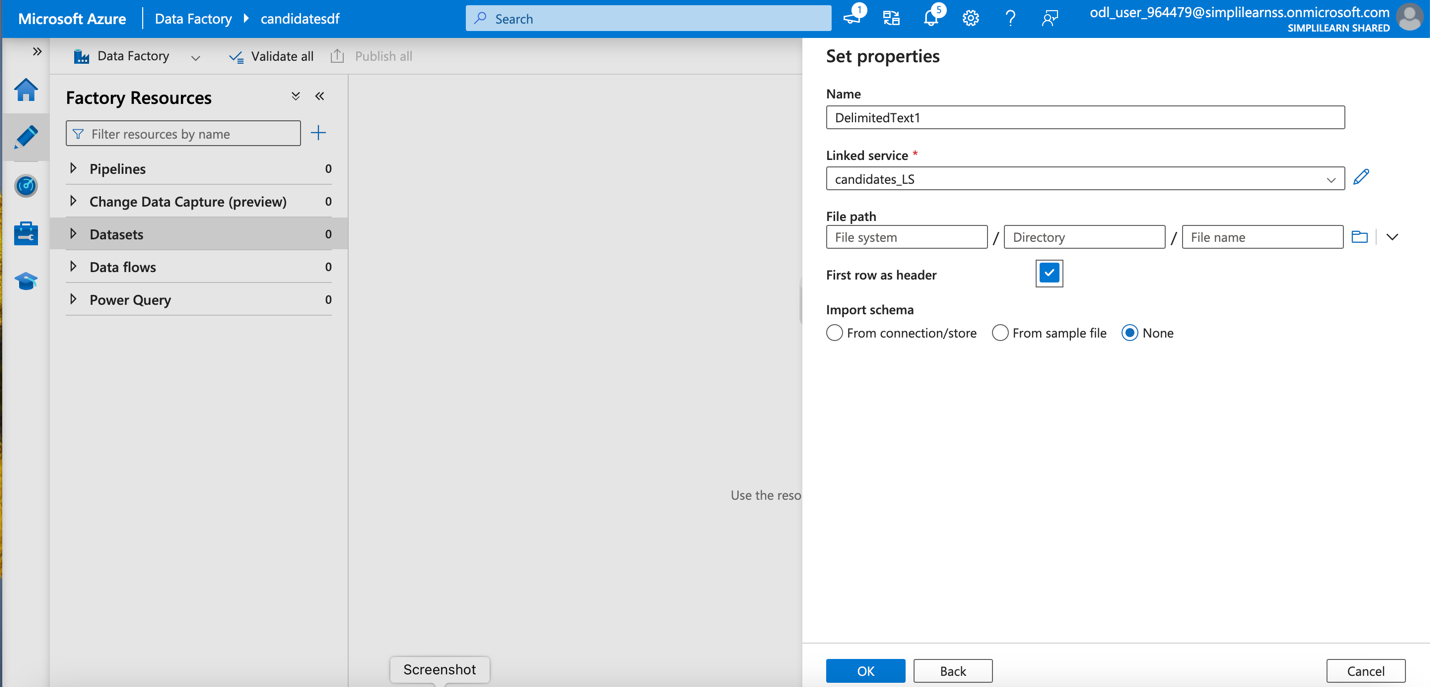
* Create first dataset for CSV files





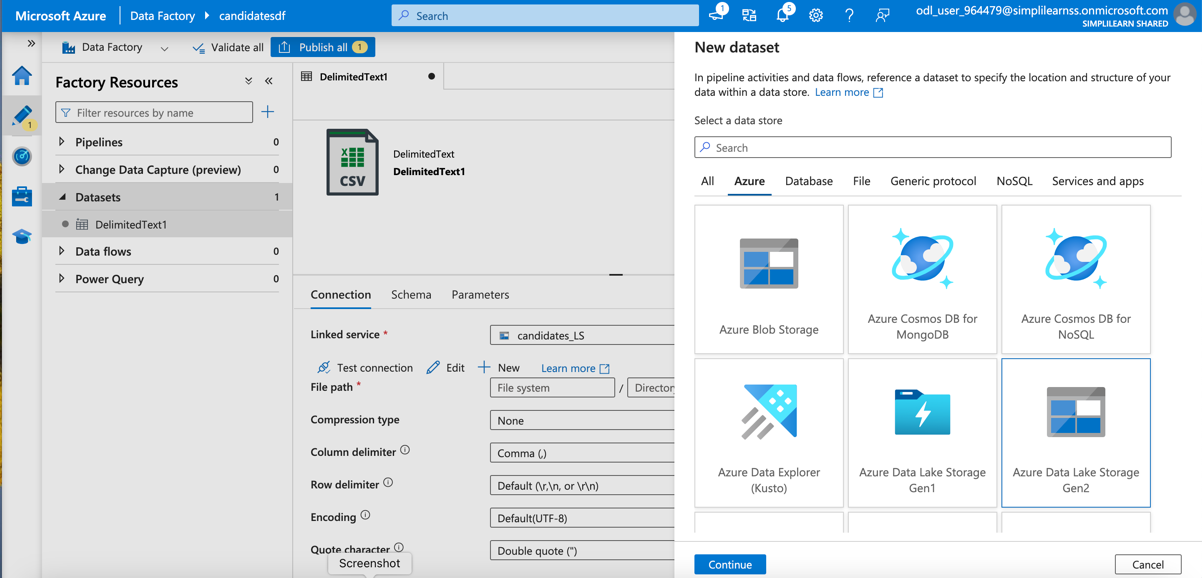




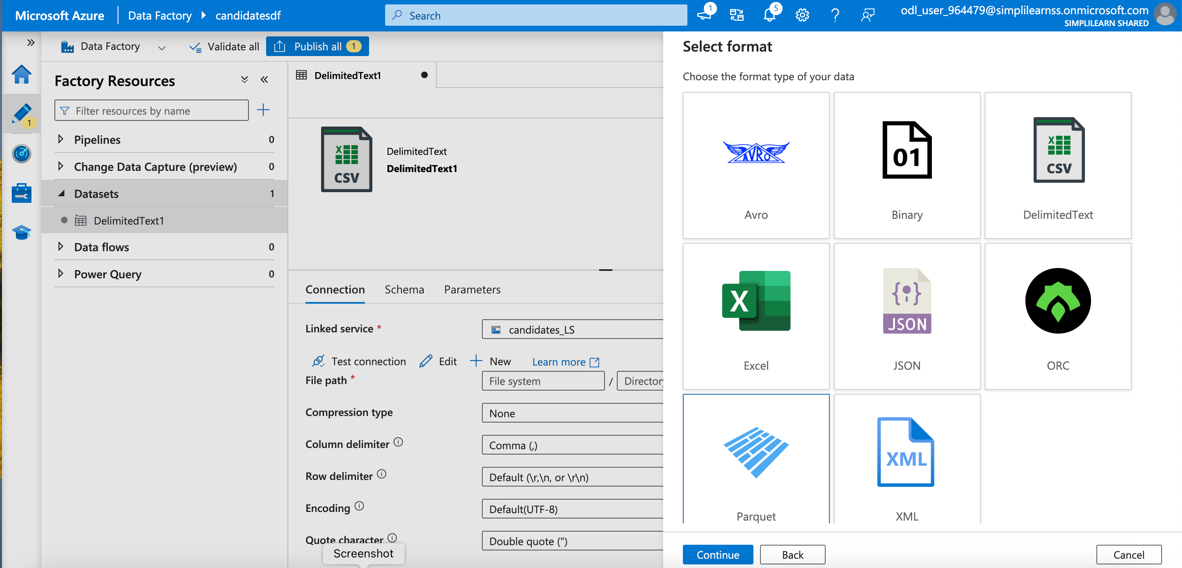




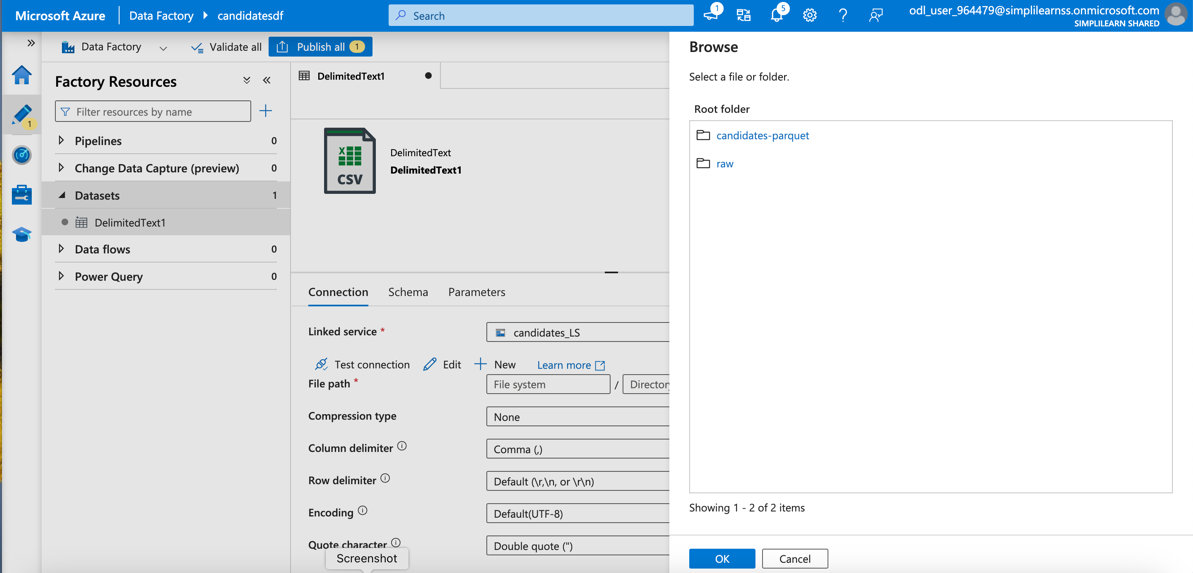
* Create second dataset for parquet file





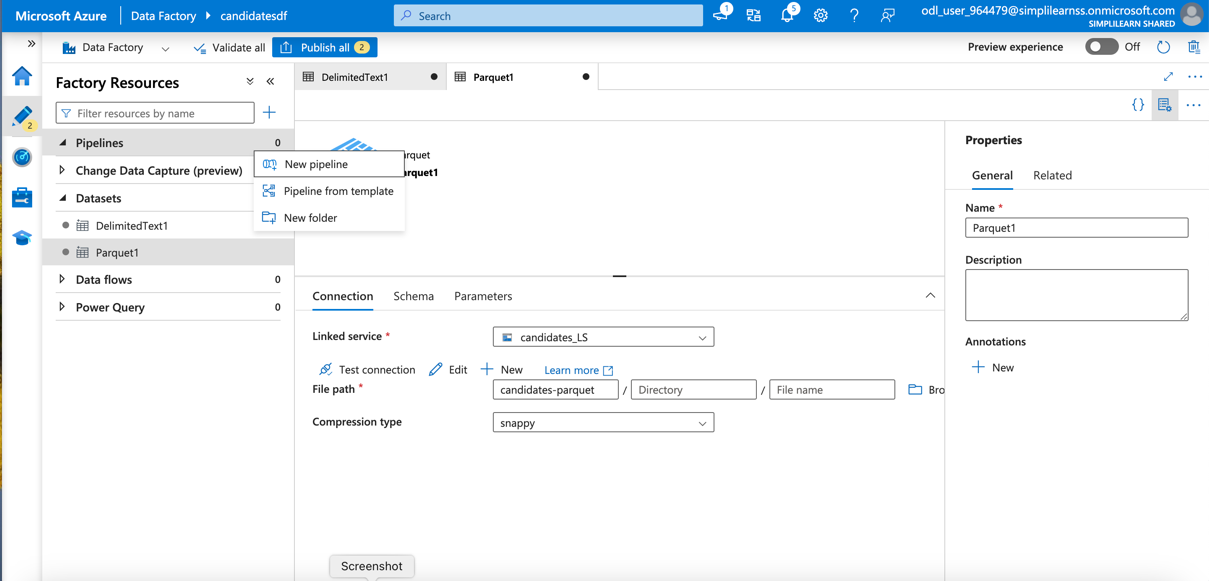




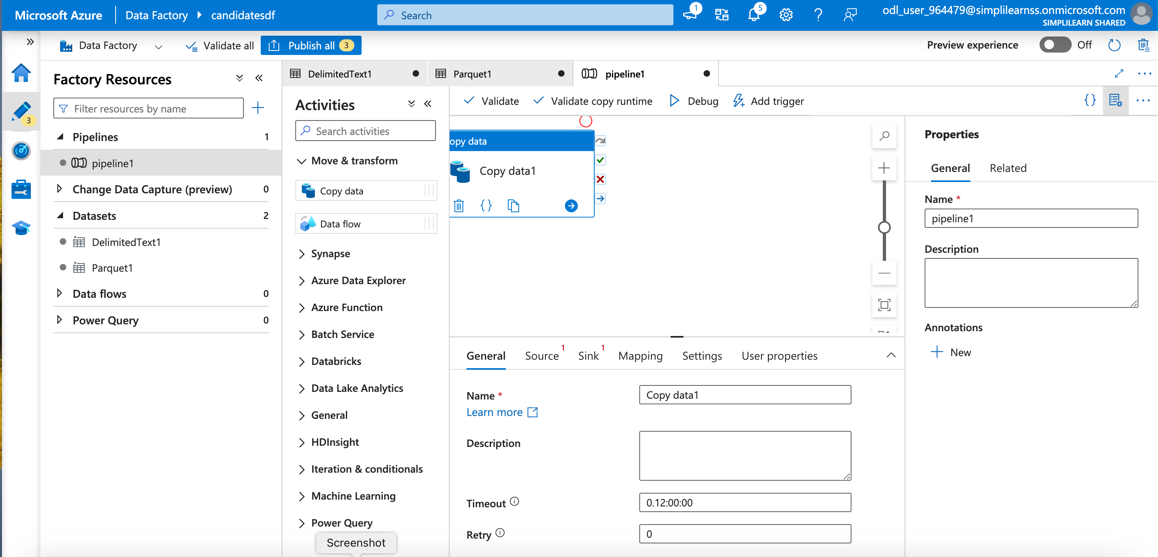


1. Use Azure Databricks as a part of the ADF pipeline
2. Create a linked service in ADF for Databricks

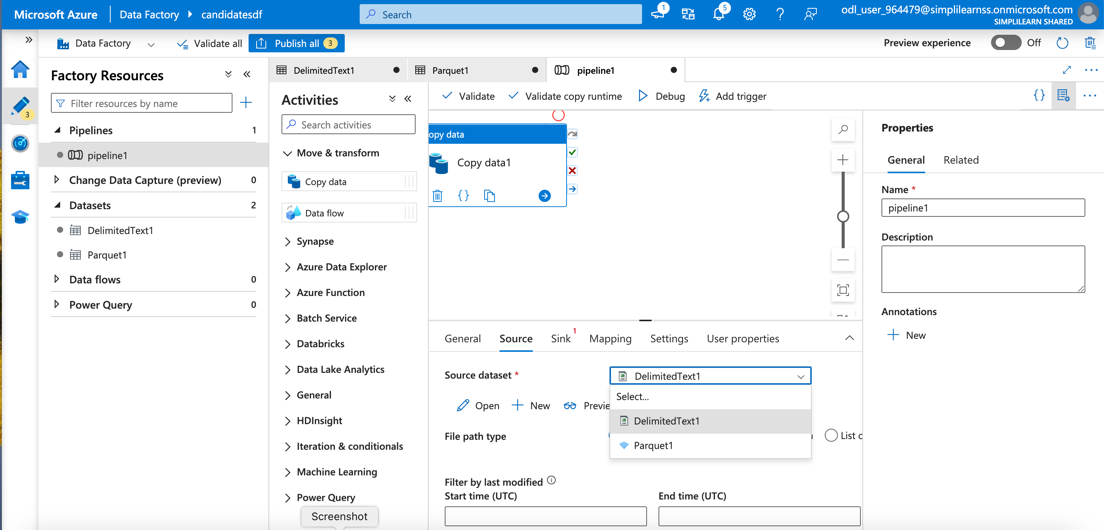
* Creating Pipeline



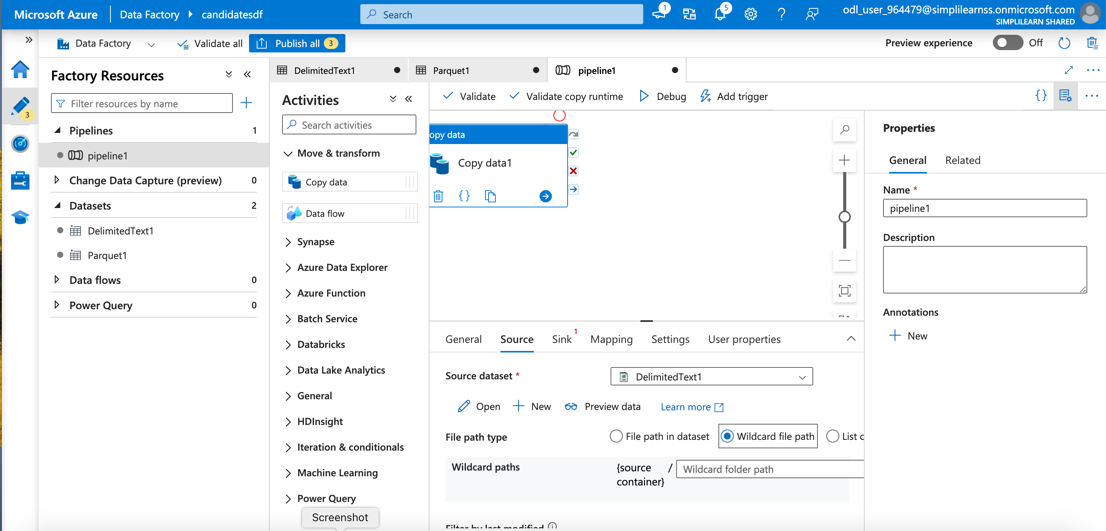




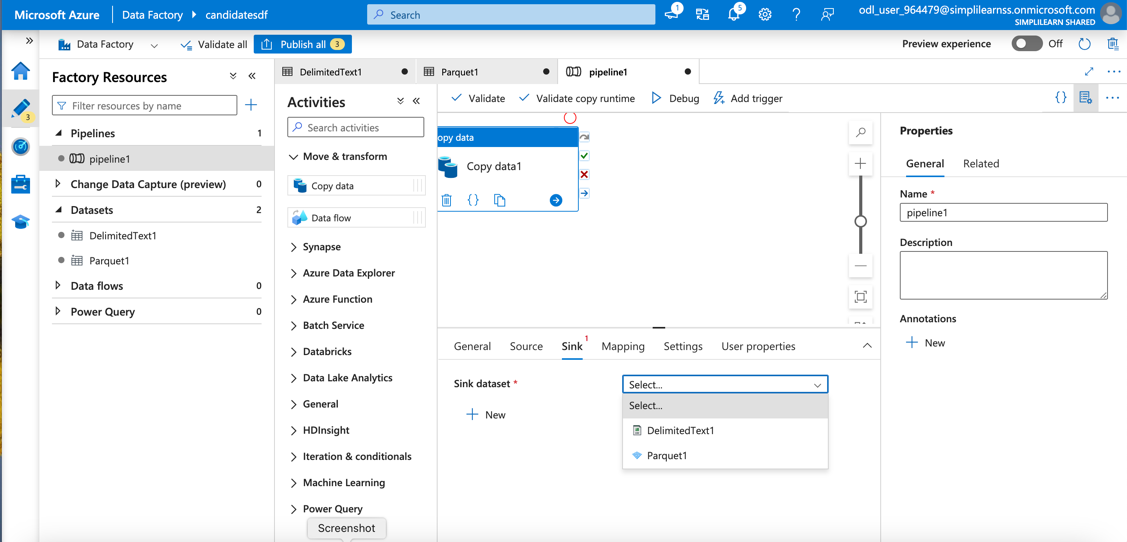




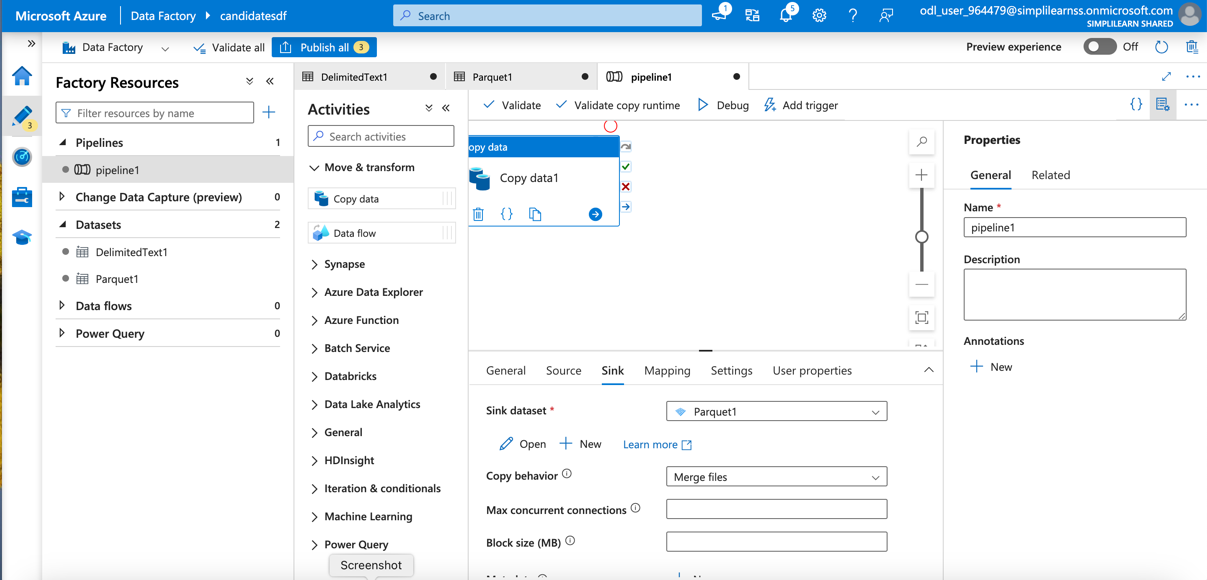






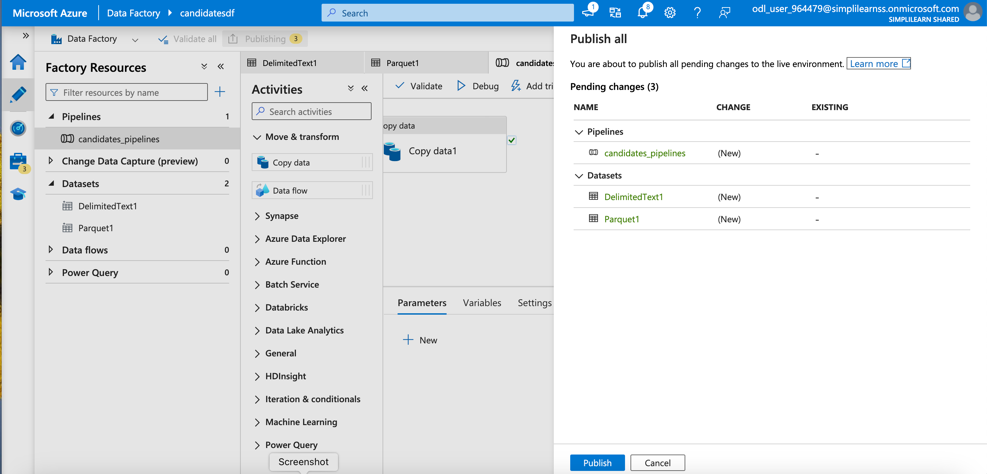






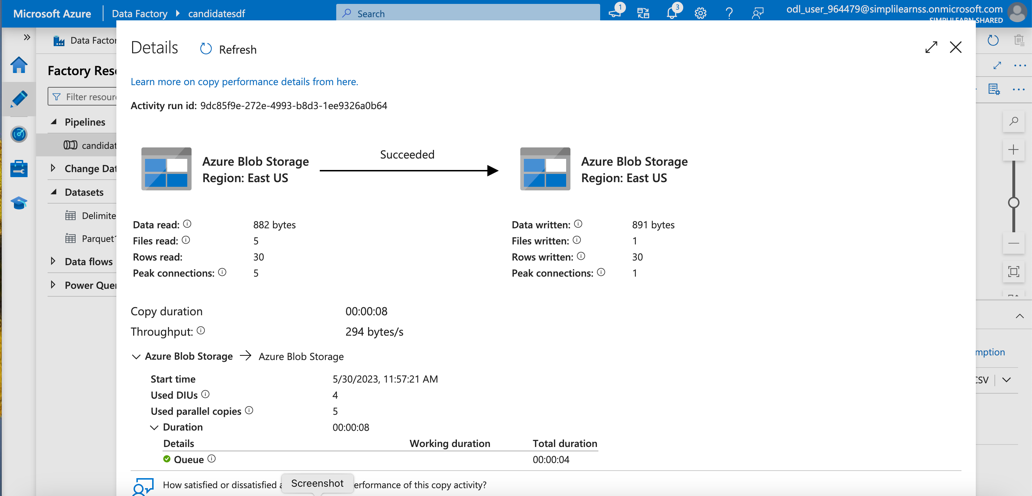


* Validate and Publish the Pipeline



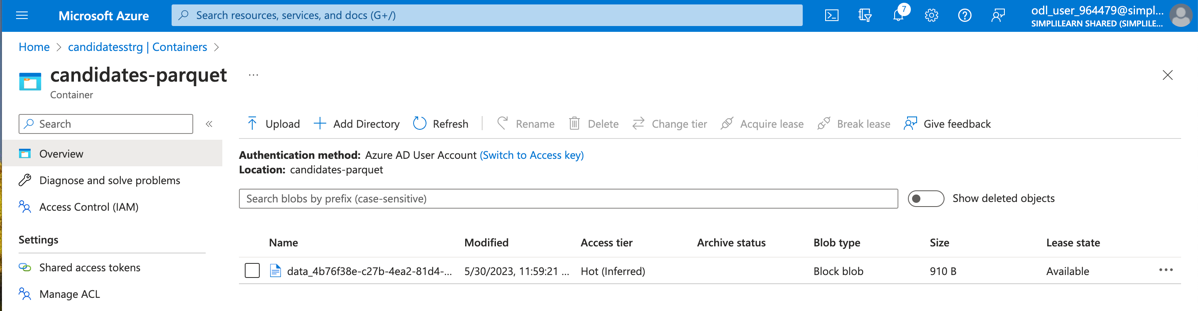


1. Convert the CSV files to Parquet files in staging storage



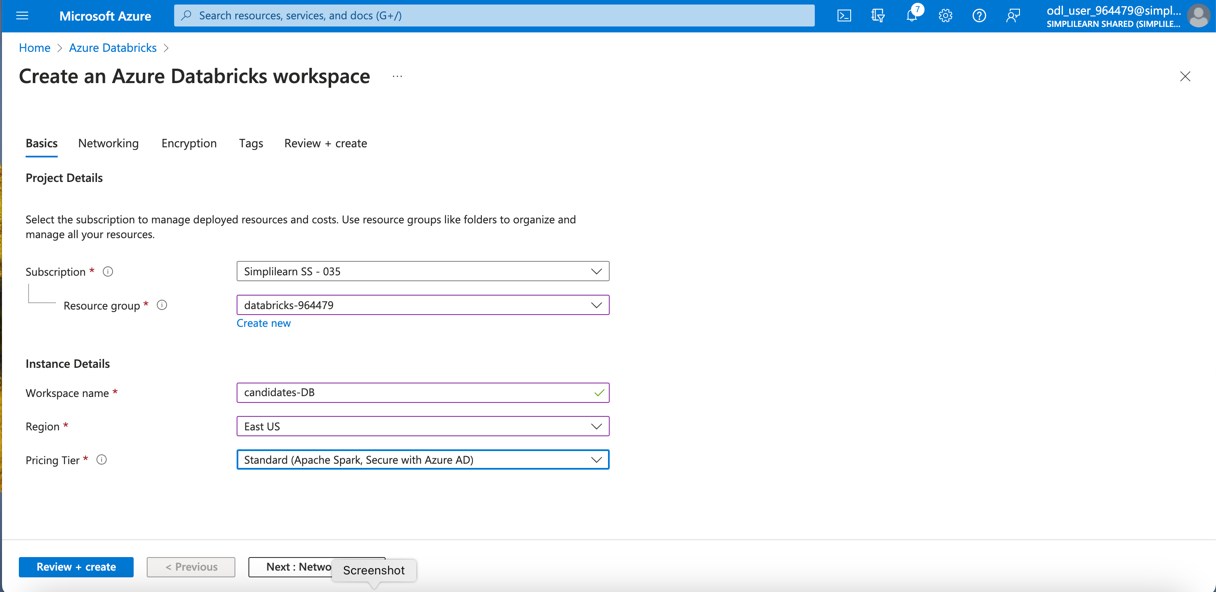


* Parquet file successfully created and saved in storage



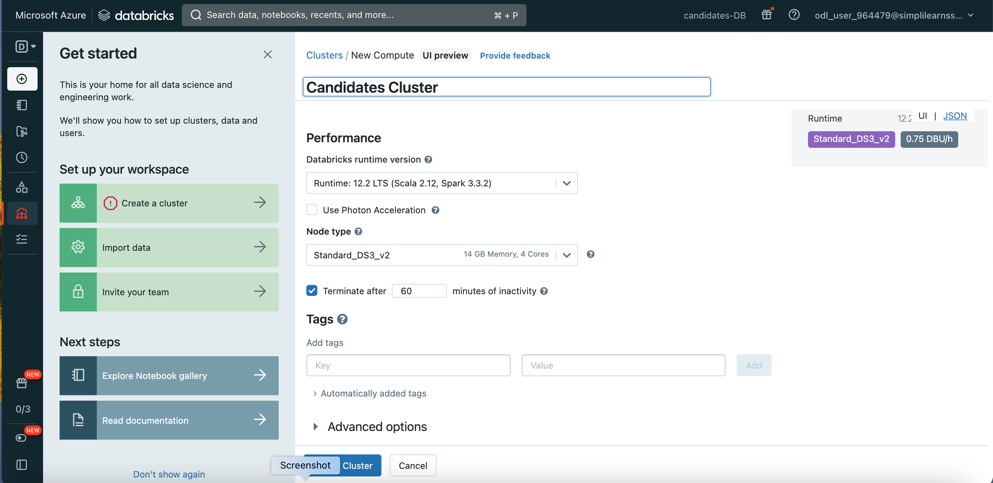


1. Access Parquet files from the staging account in Azure Databricks

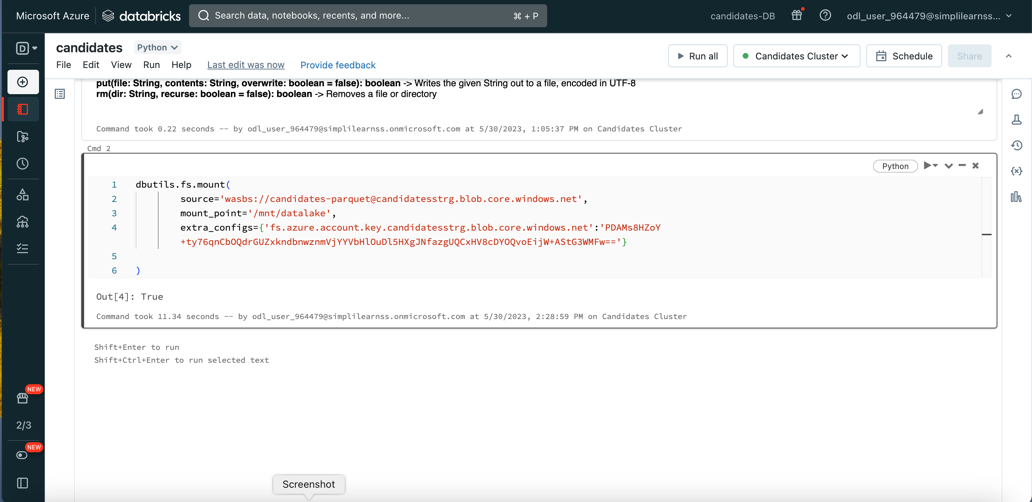




* Go to Azure DataBricks Workspace
  + Create new cluster

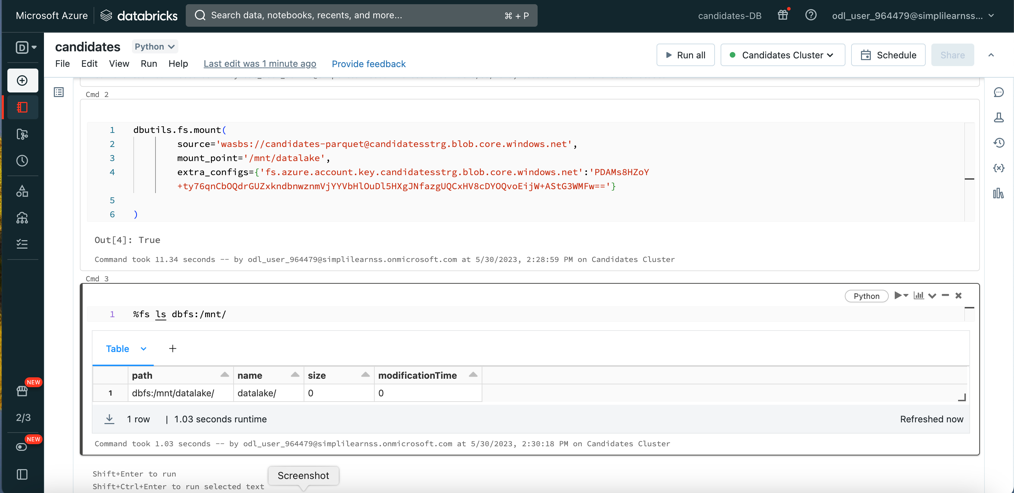


* Create new notebook inside workspace
* Mount the Parquet files from Azure storage



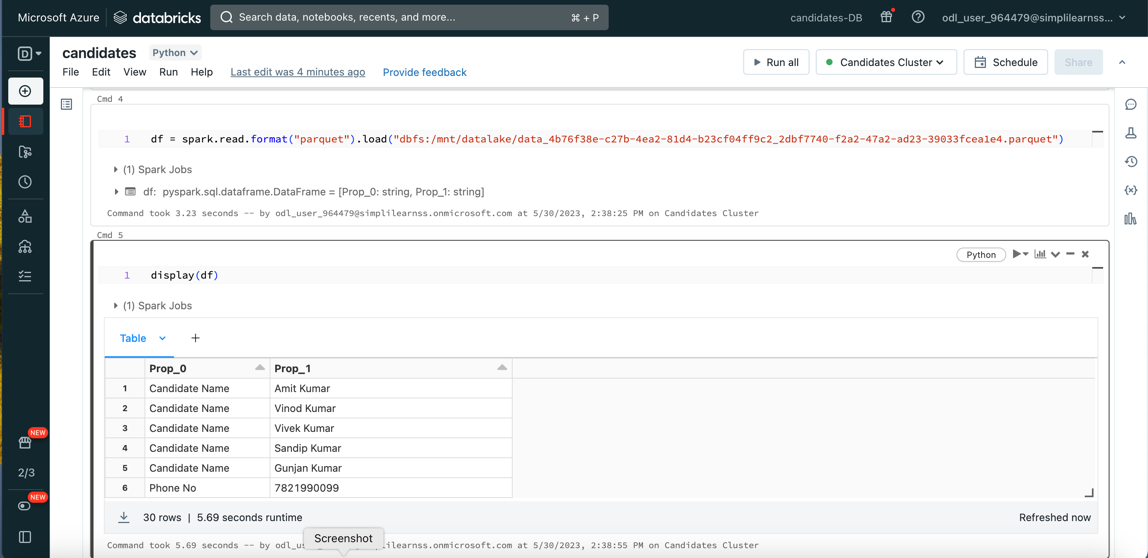


`





1. Convert the Parquet files to Azure Databricks Delta tables
   1. Create a data frame and load the parquet file in it
   2. Display the loaded file



1. Store and visualize the data from Azure Databricks Delta tables

