Name: Khanh Nguyen

Date: 11/12/2022

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

Project 1

DESCRIPTION

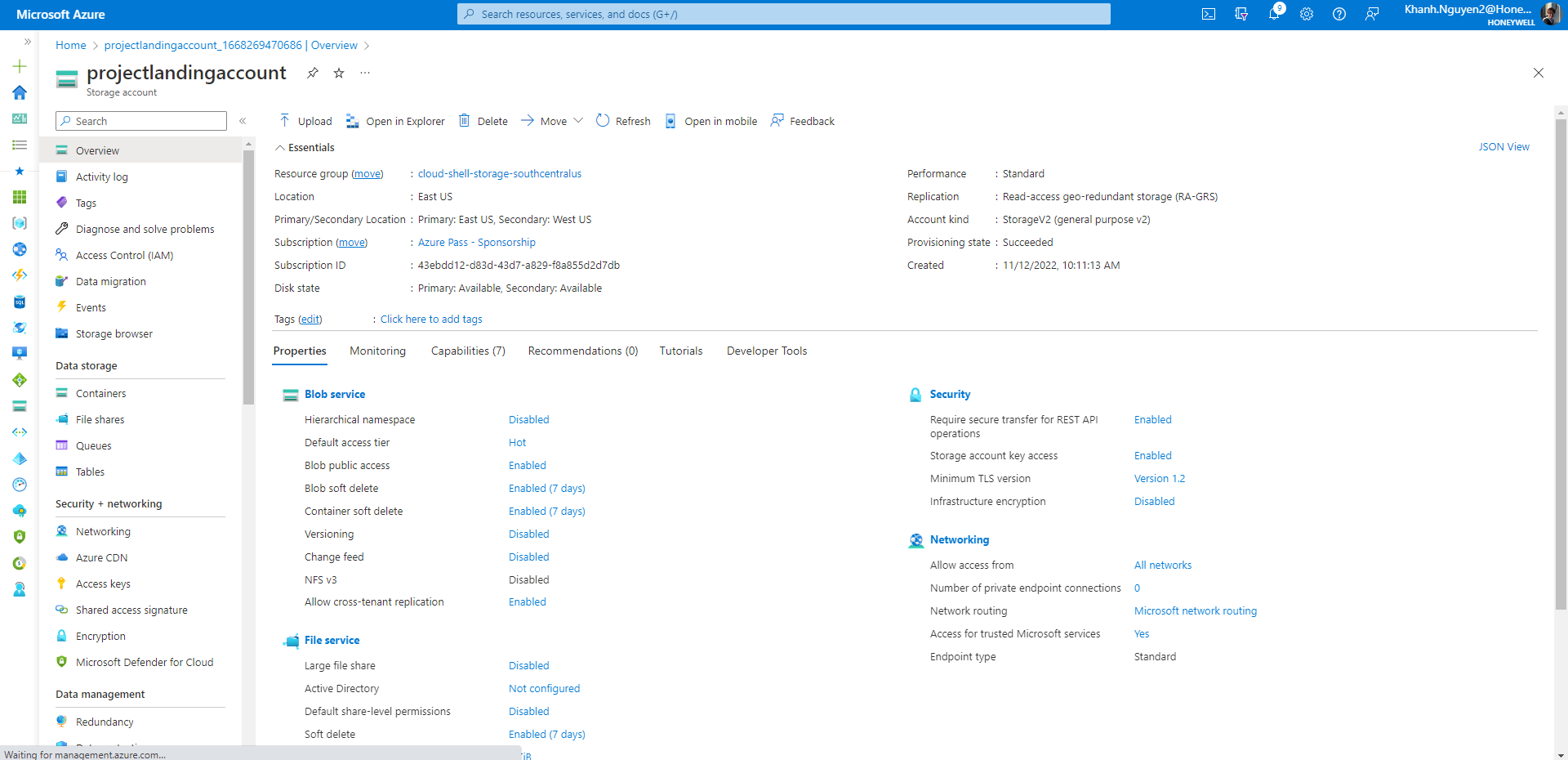
Your company is looking for a data engineer and is inviting candidates to apply for this position by providing a portal where applicants can add their credentials.

As thousands of candidates have applied for this position, the company has a huge amount of data that it needs to upload to its website. This data is moved to Azure Data Lake Storage parallelly. The company wants to save the contents of all CSV files to Delta Lake of Azure Databricks so that these files can be retrieved and accessed from Azure Databricks when required.

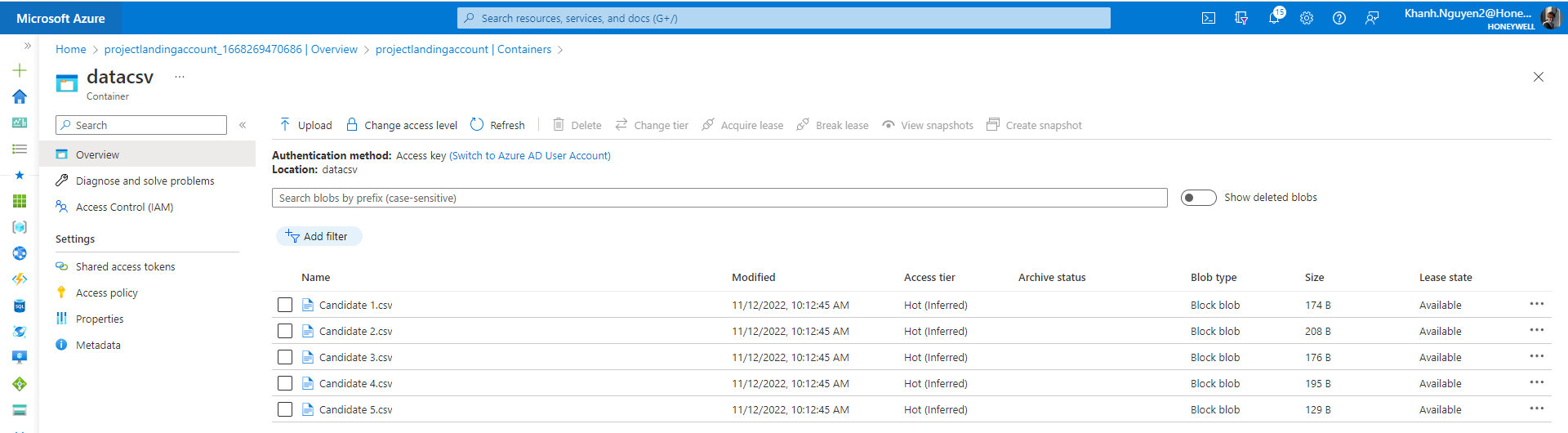
**Steps to Perform:**

1. Create a landing storage account in Azure

Enable Gen2 data lake

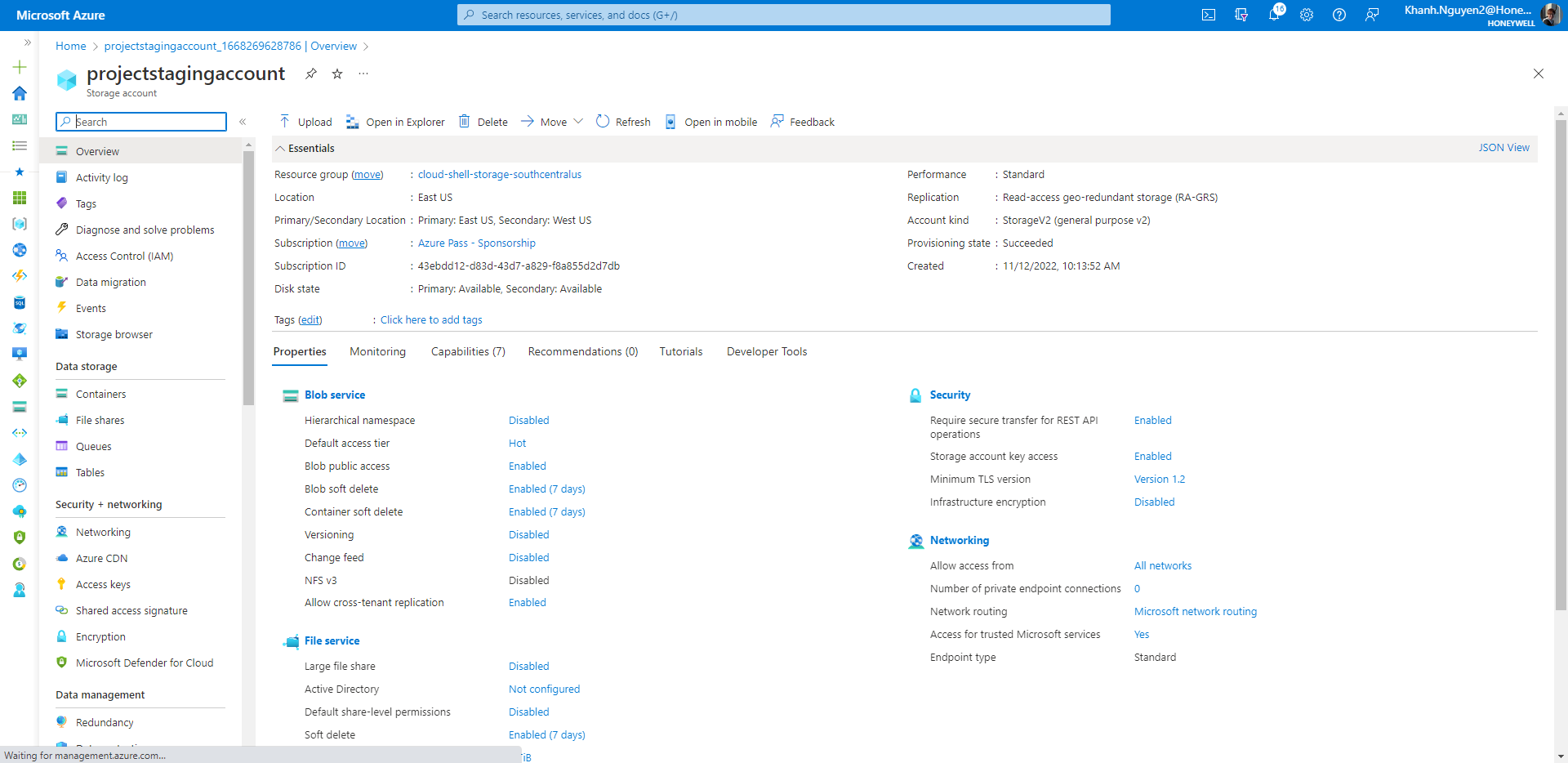


2. Store the CSV files in the storage account

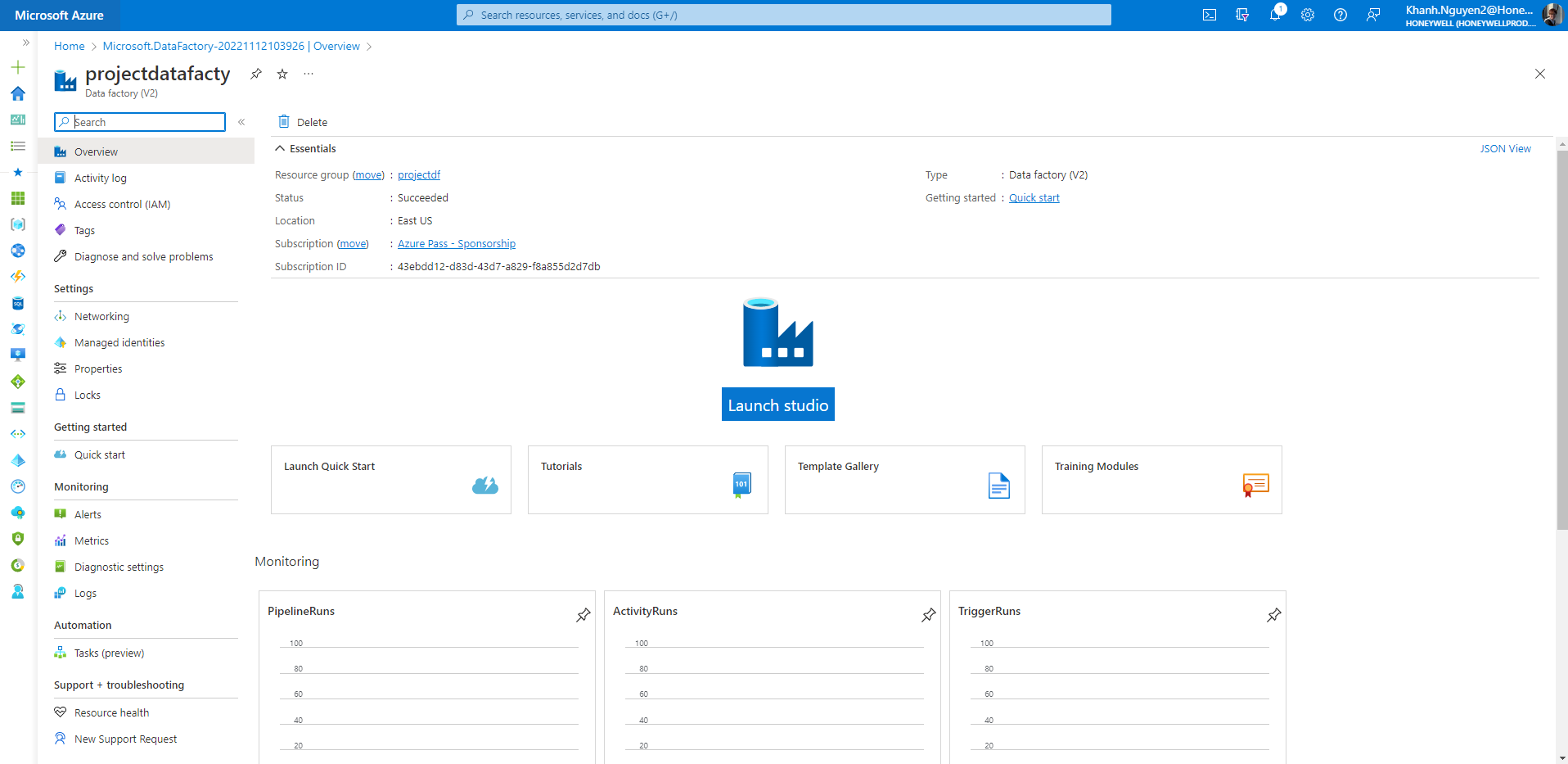


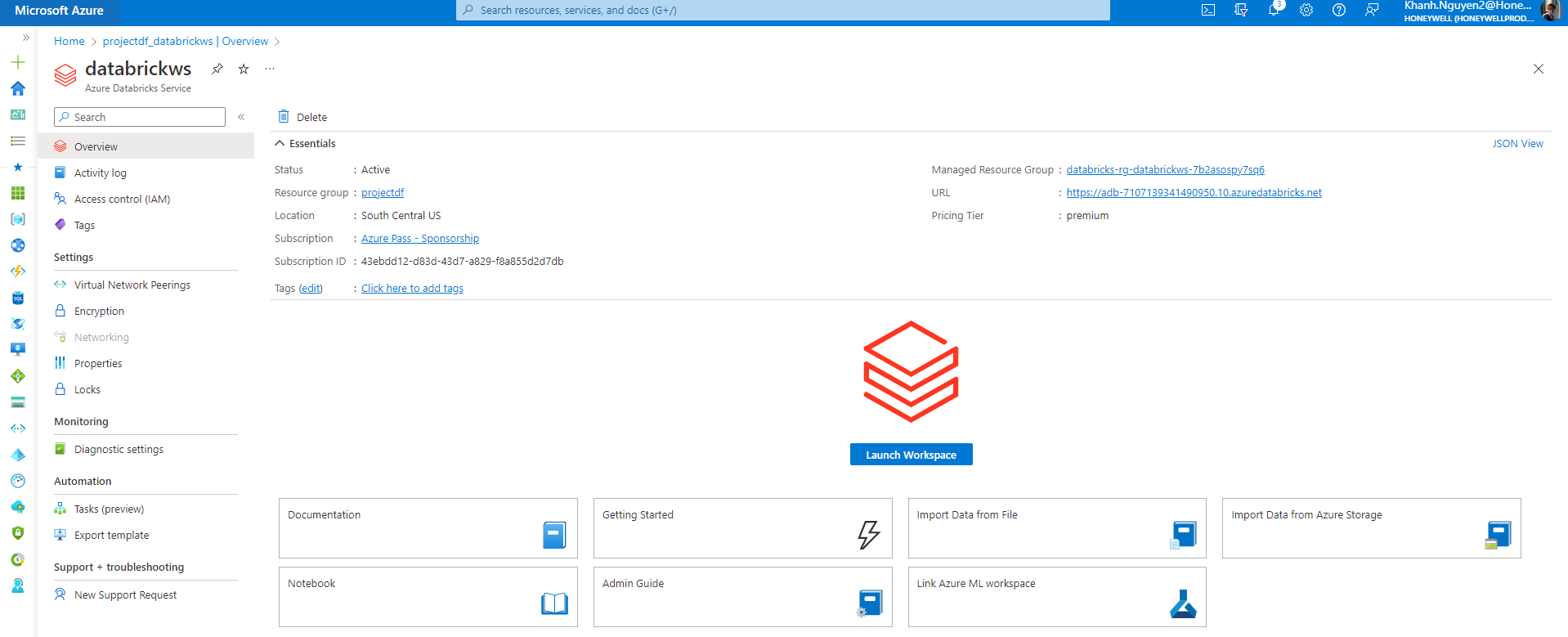
1. Create a staging storage account in Azure.

Enable Gen2 data lake

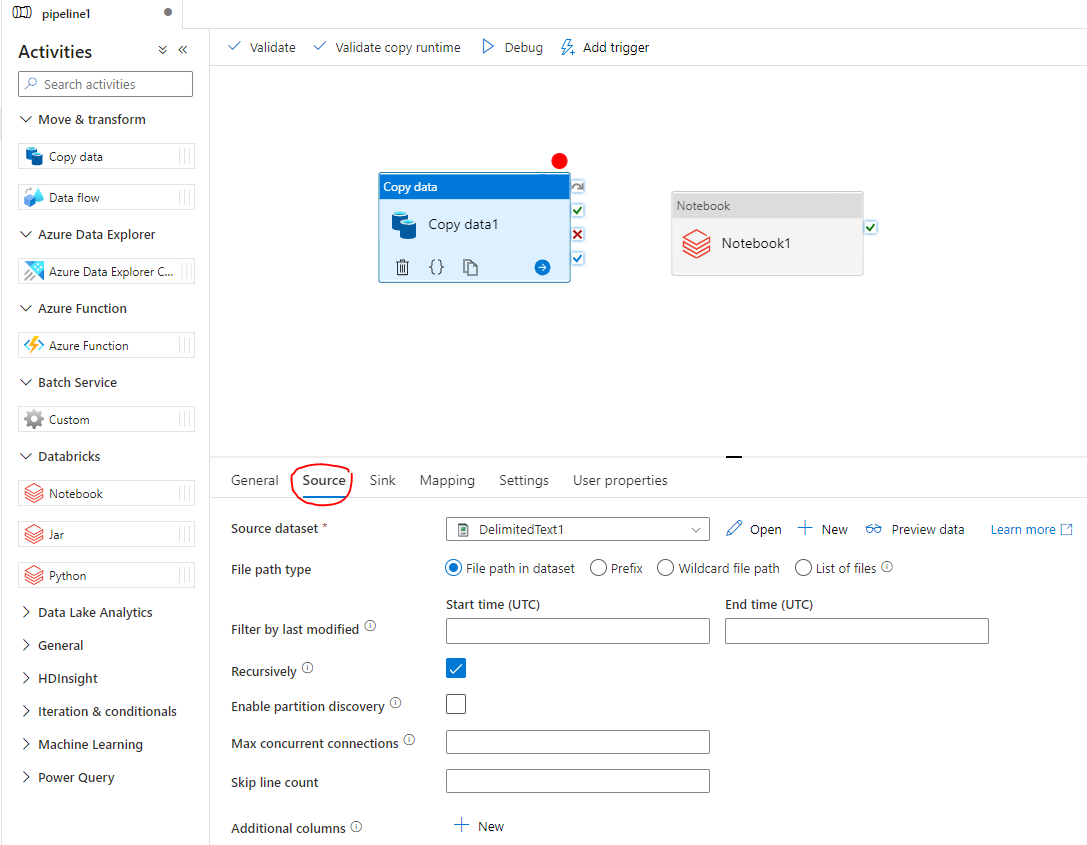


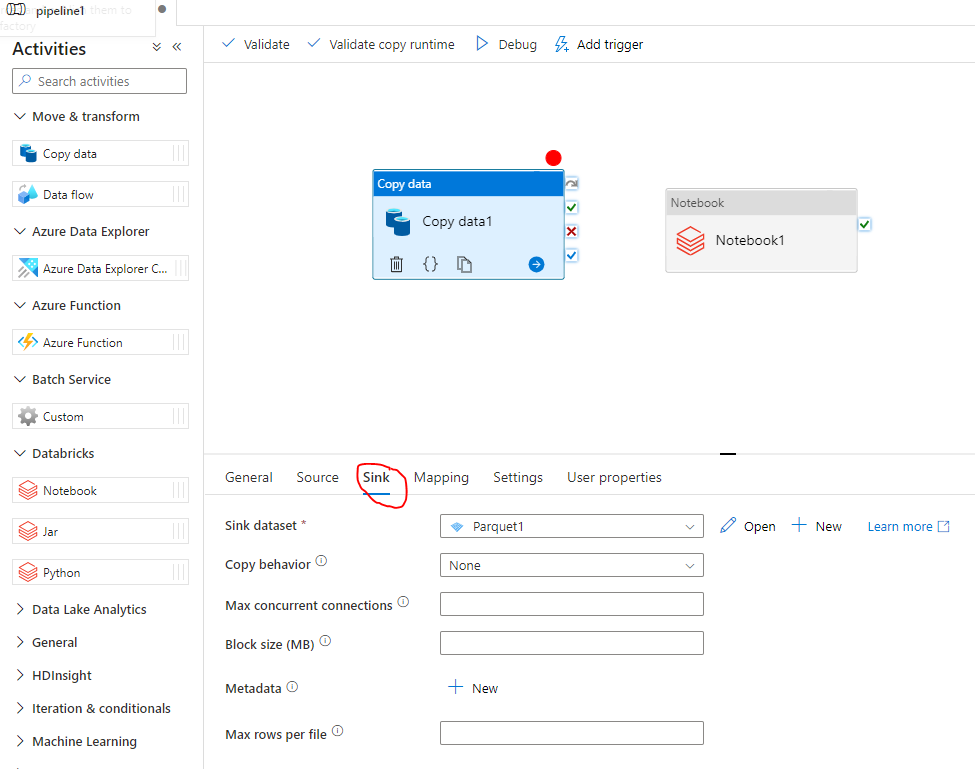
1. Create an Azure Data Factory resource and Azure Data Factory pipeline



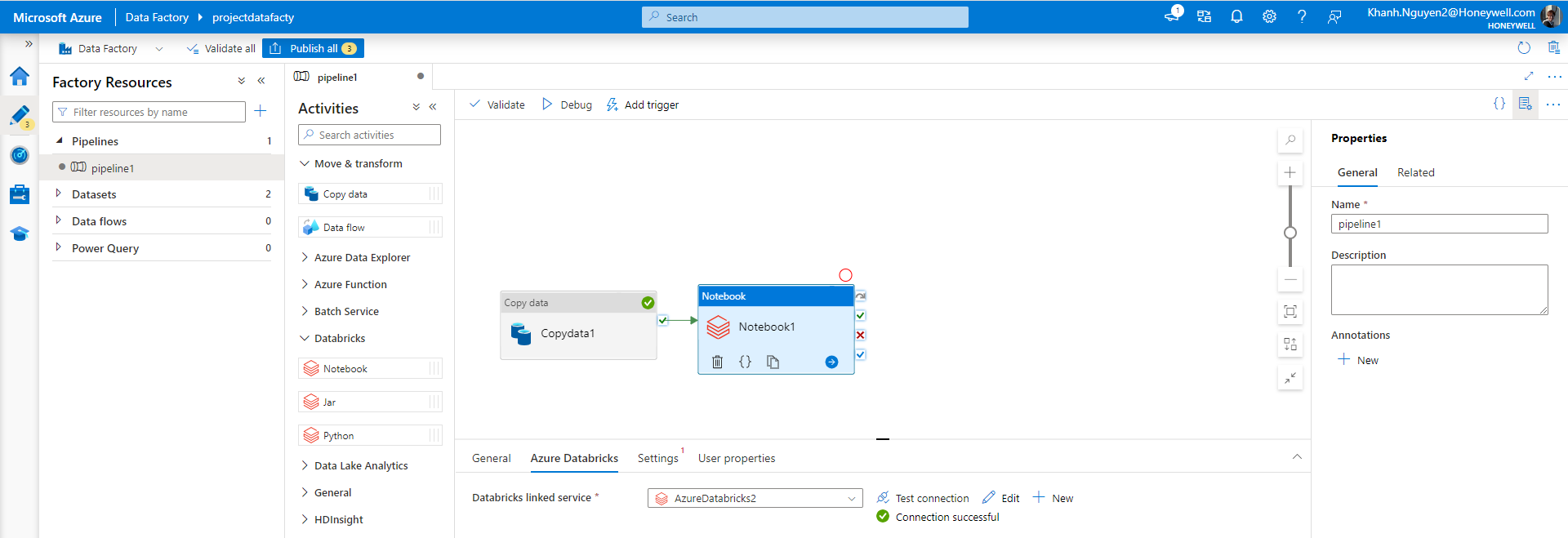


5. Create linked services for the storage accounts



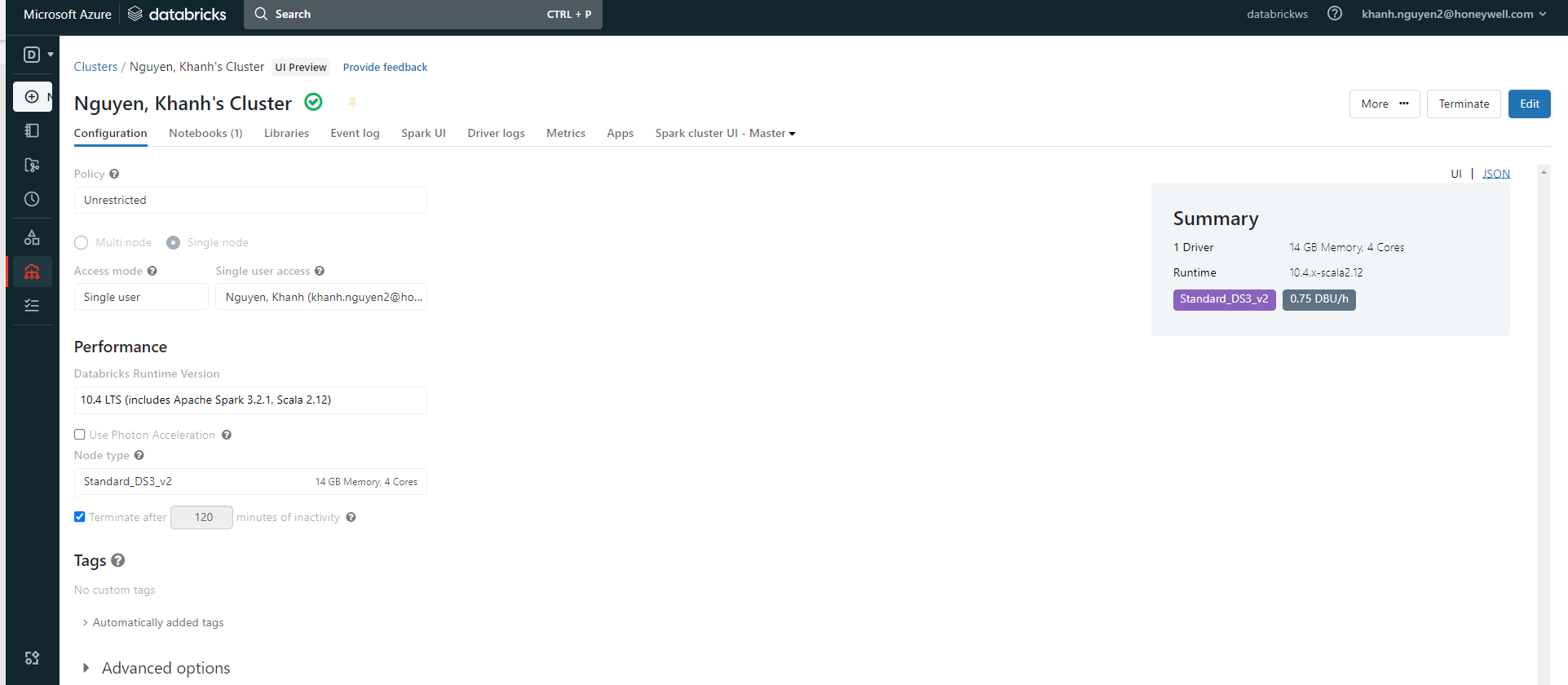


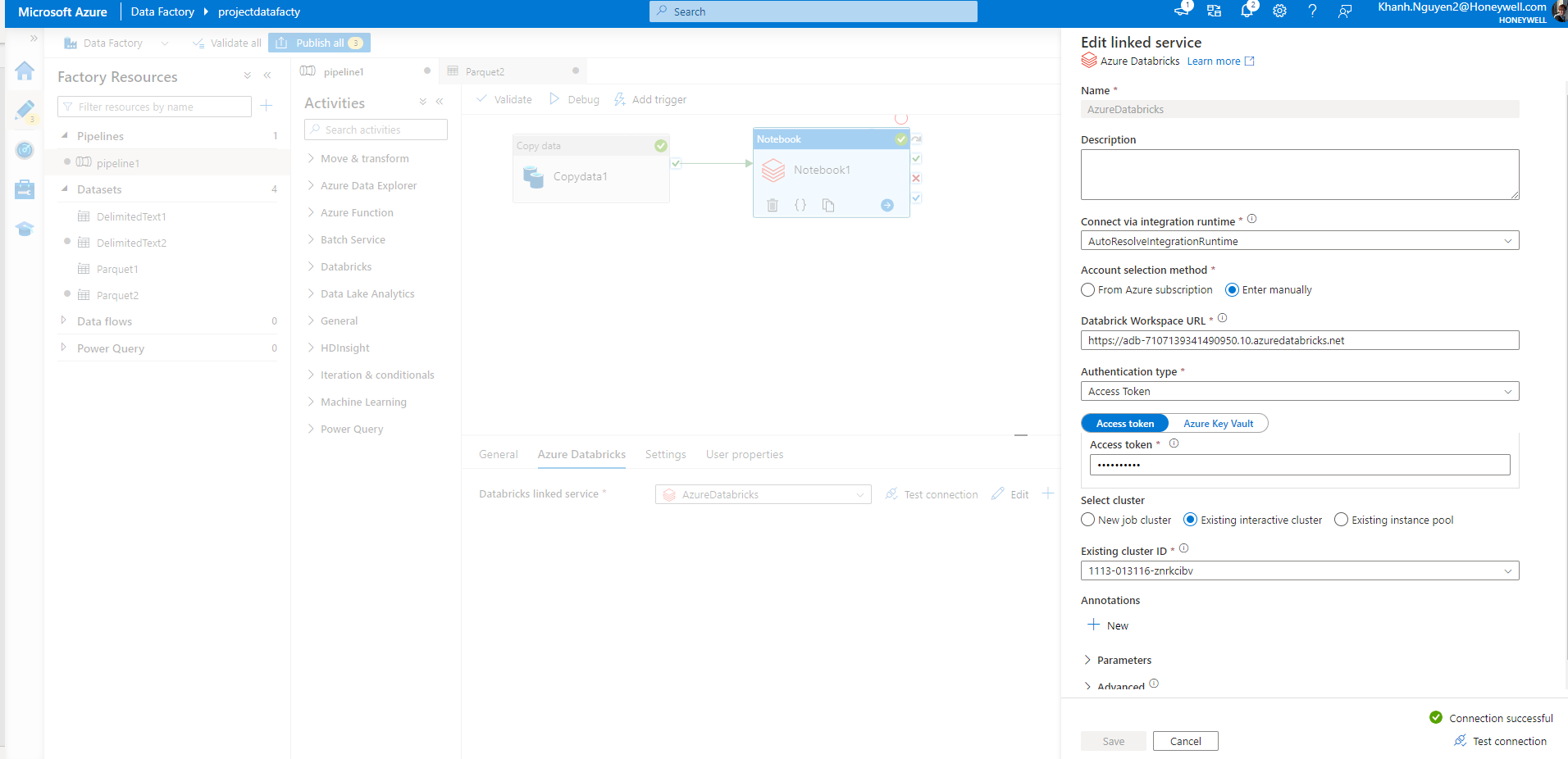
1. Use Azure Databricks as a part of the ADF pipeline



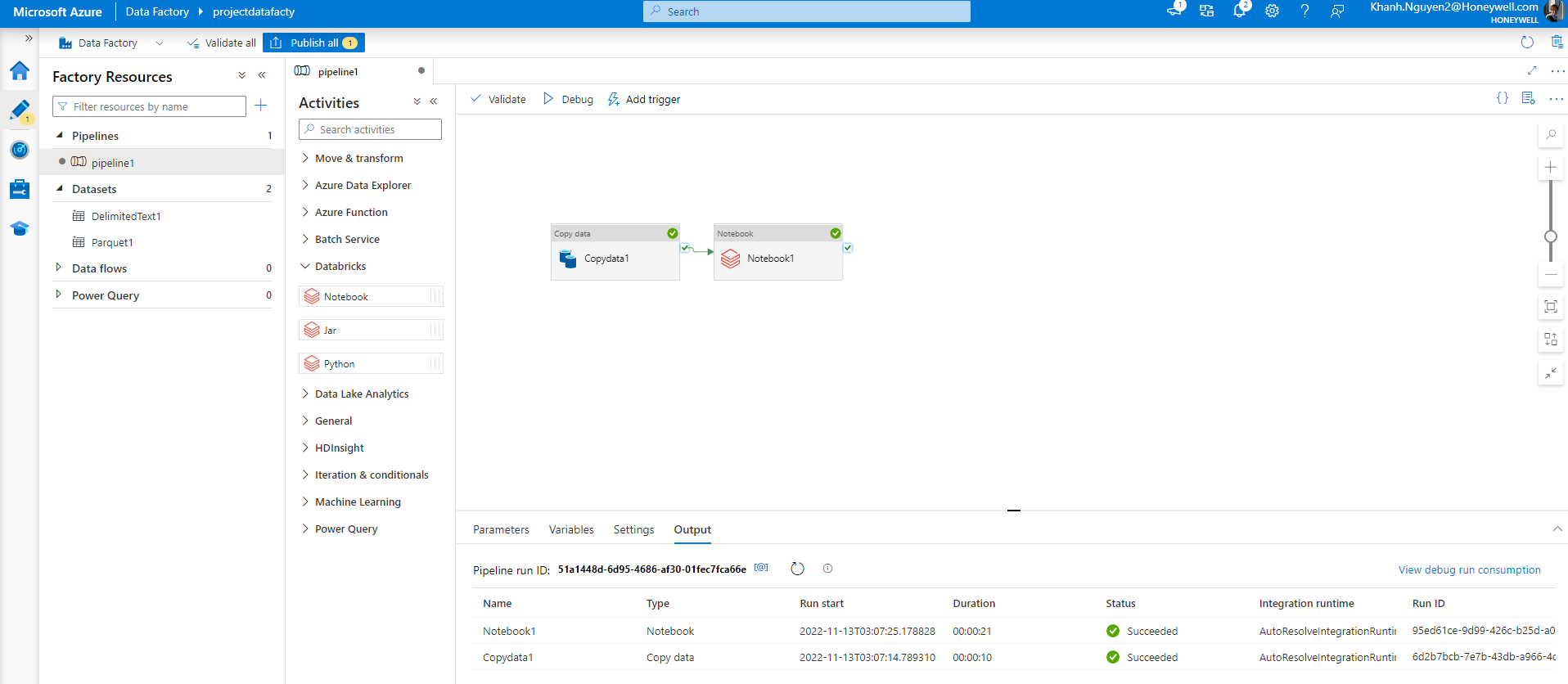
7. Create a linked service in ADF for Databricks.

Create a single node cluster in Databricks and input information to ADF

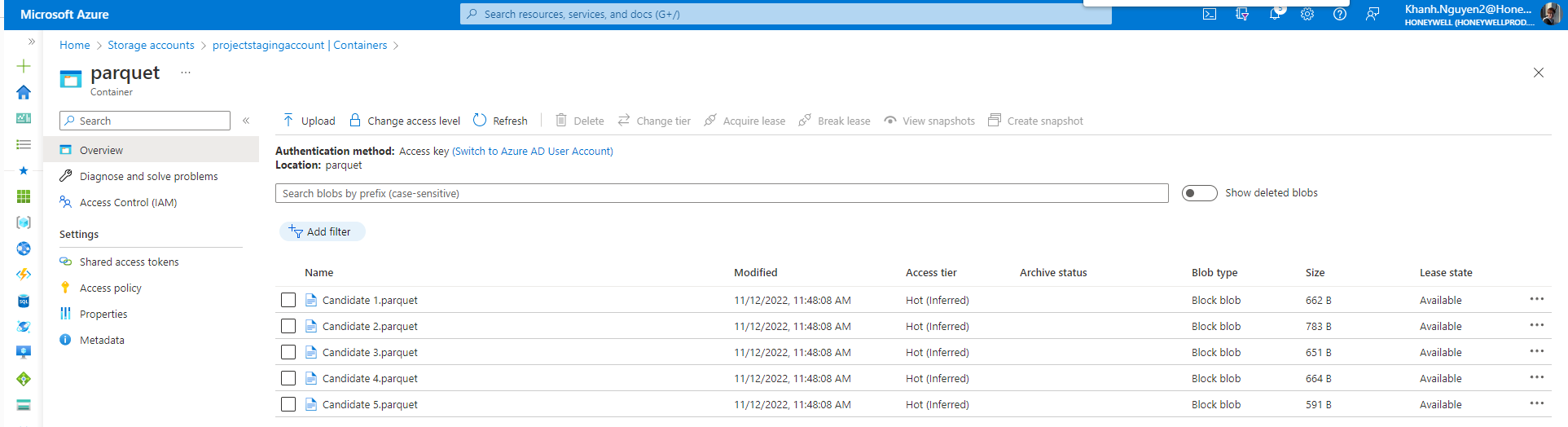




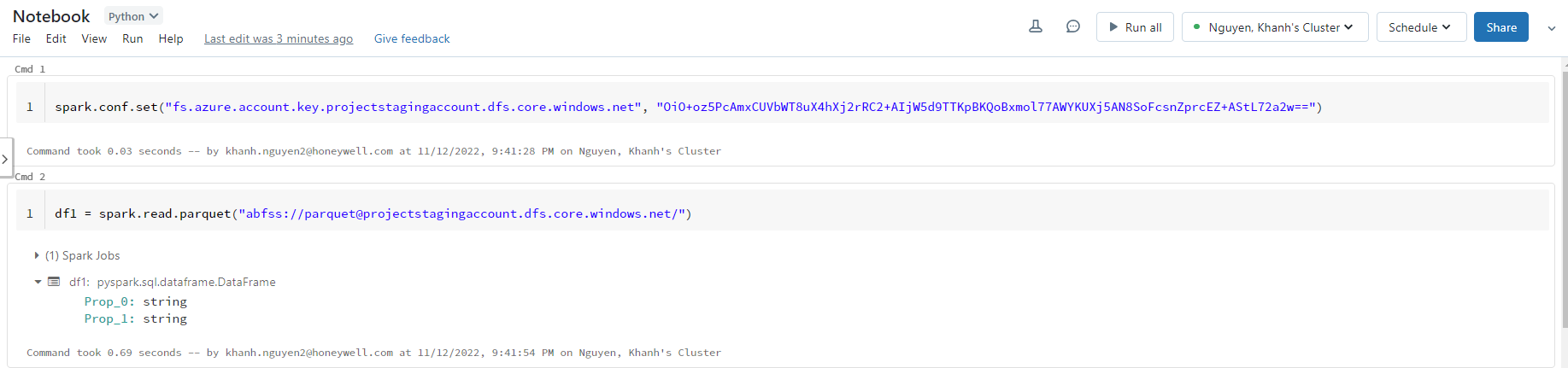
Hit ‘Validate’ to ensure no error and Hit ‘Debug’



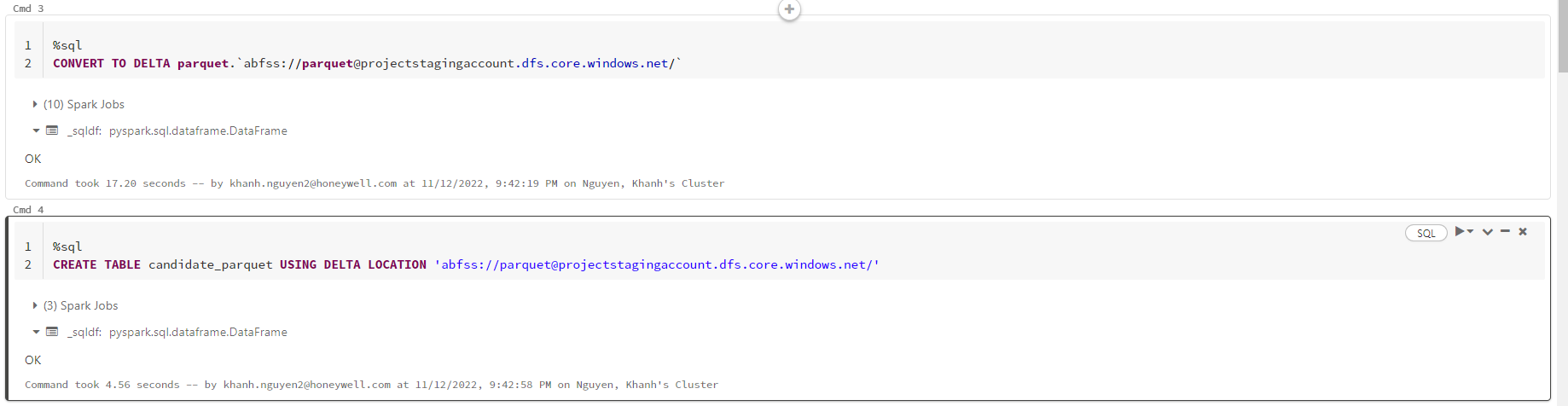
8. Convert the CSV files to Parquet files in staging storage



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



10. Convert the Parquet files to Azure Databricks Delta tables



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

