AhmadSultan_041034155

CST8921 – Cloud Industry Trends

Lab 4 Report

Title

Analyzing Data with Azure Databricks: A Hands-On Exploration.

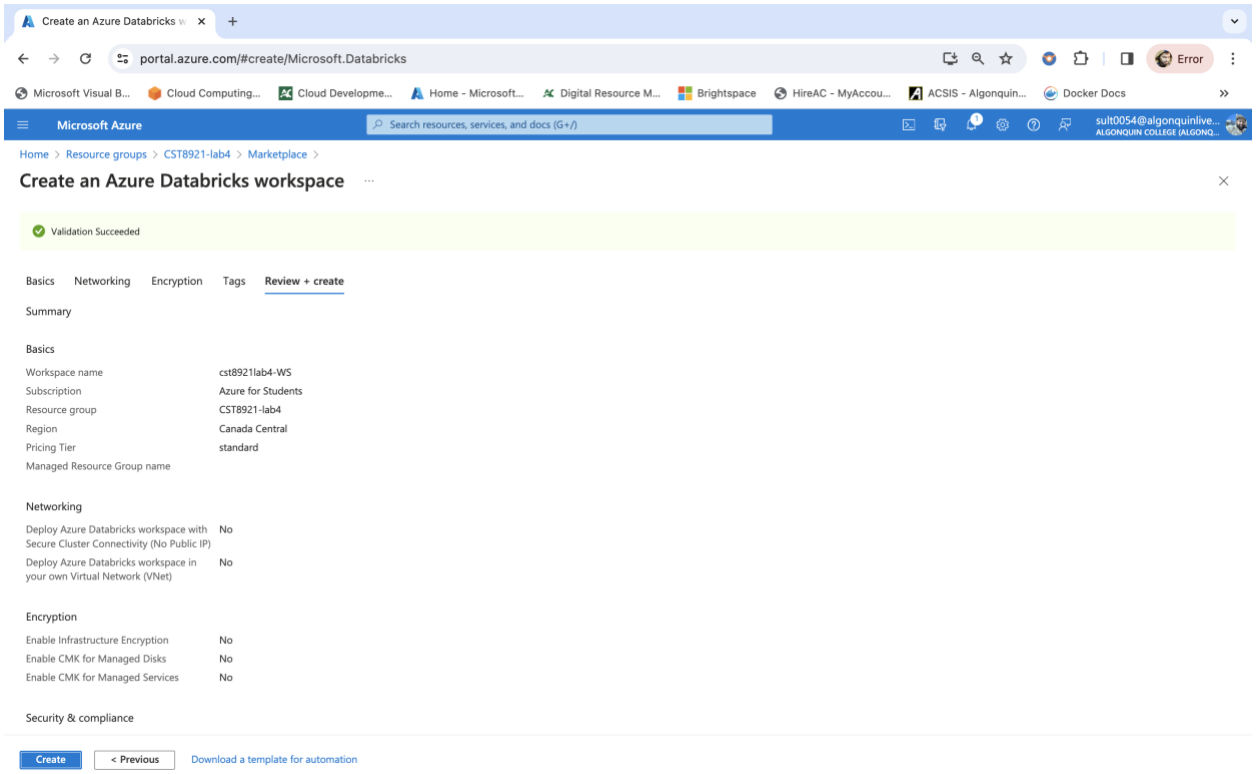
Introduction

Embark on a journey to master Azure Databricks, an Apache Spark-powered analytics platform. In this lab we will provision a workspace, analyze data using Spark, explore Delta Lake functionality, and execute Databricks notebooks seamlessly from Azure Data Factory.

Steps

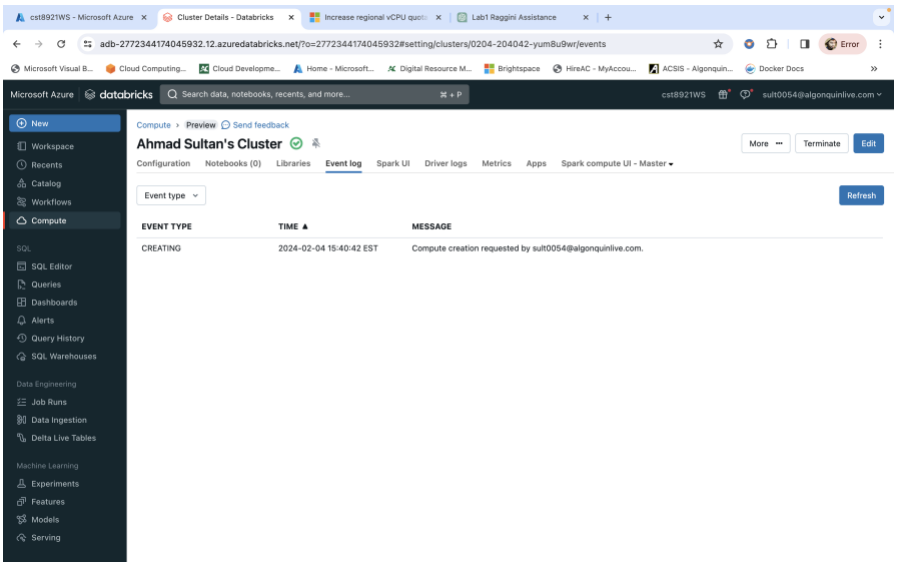
Part 1: Explore databricks notebooks

1. Provision azure databricks workspace environment.

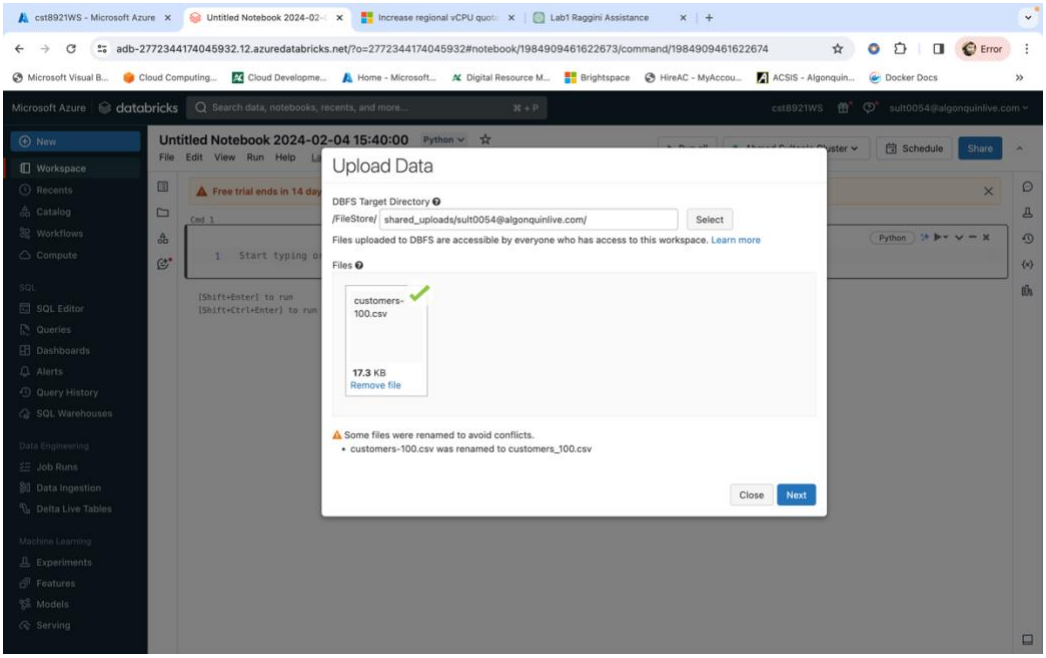


2. Create a single node cluster in the workspace

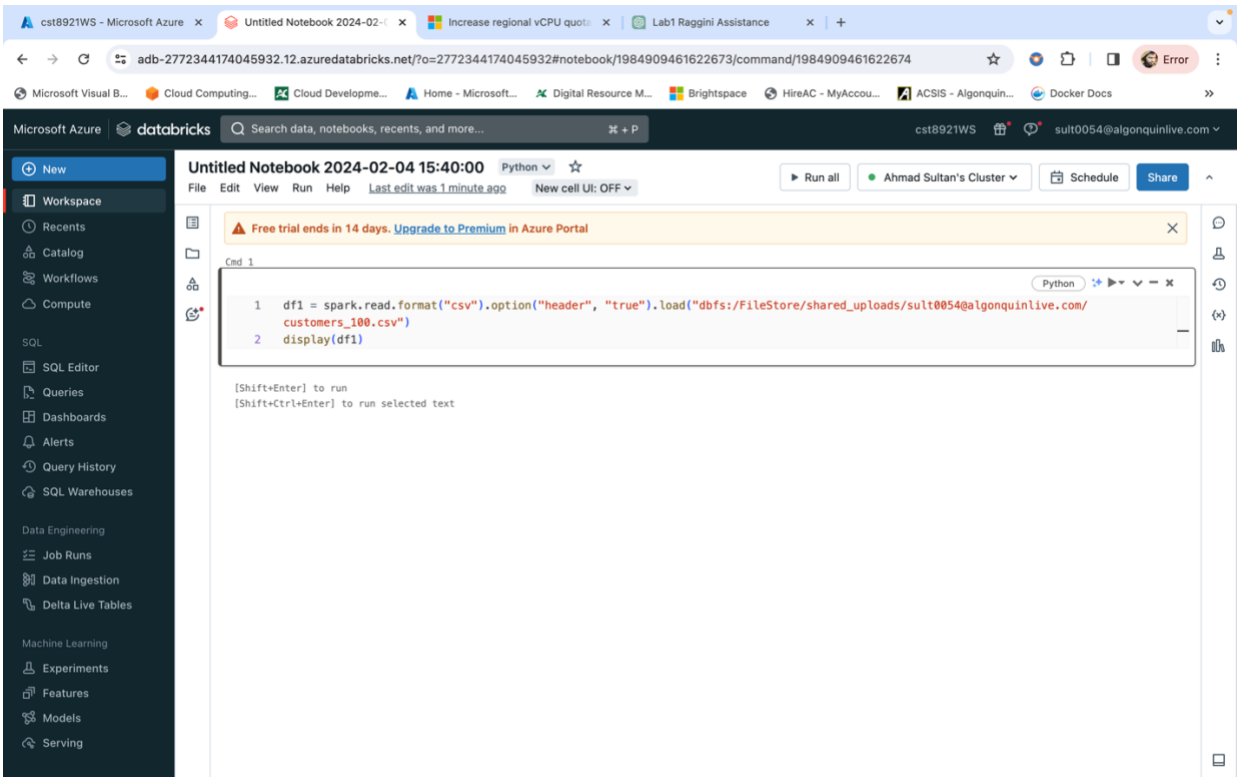
Note : Azure Databricks is a distributed processing platform that uses Apache Spark clusters to process data in parallel on multiple nodes. Each cluster consists of a driver node to coordinate the work, and worker nodes to perform processing tasks.



3. Use Spark to analyze data file – create a notebook to explore data. Download sample file from this link: https://www.datablist.com/learn/csv/download-sample-csv-files. Upload the file downloaded to DBFS directory in the workspace.



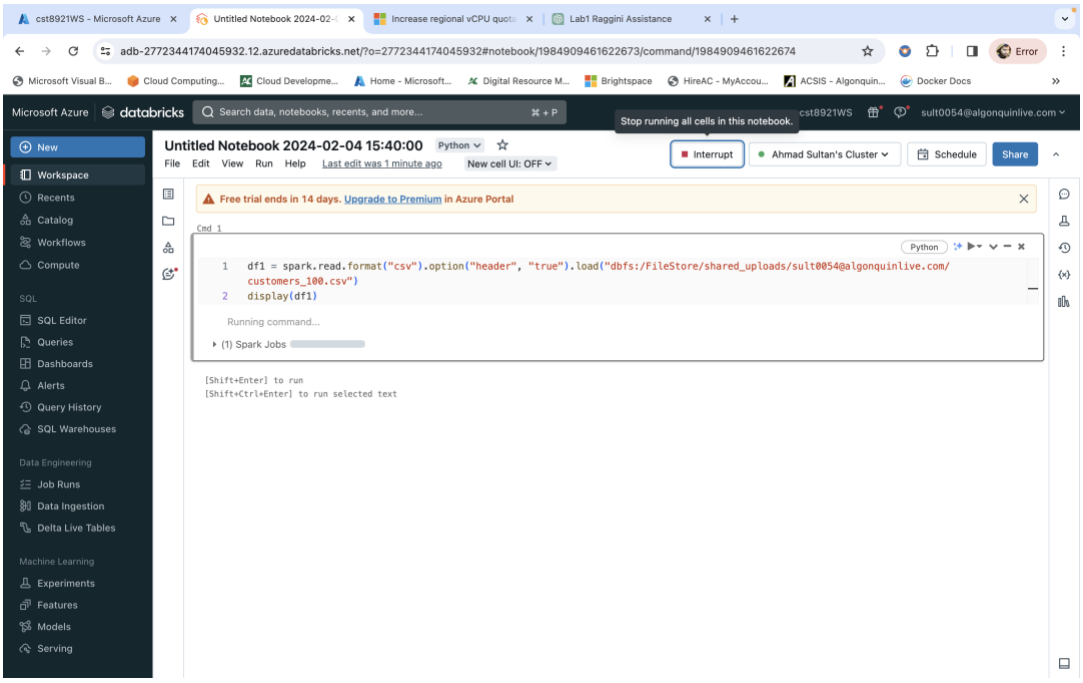
6. In the Access files from notebooks pane, select the sample PySpark code and copy it to the clipboard. You will use it to load the data from the file into a DataFrame. Then select Done. While exploring notebook load data in dataframe, change the file name from products to the file name you have uploaded in dbfs. Please change the below code as per your configuration in databricks notebook. **df1=spark.read.format("csv").option("header", "true").load("dbfs:/FileStore/shared_uploads/user@outlook.com/products.csv")**

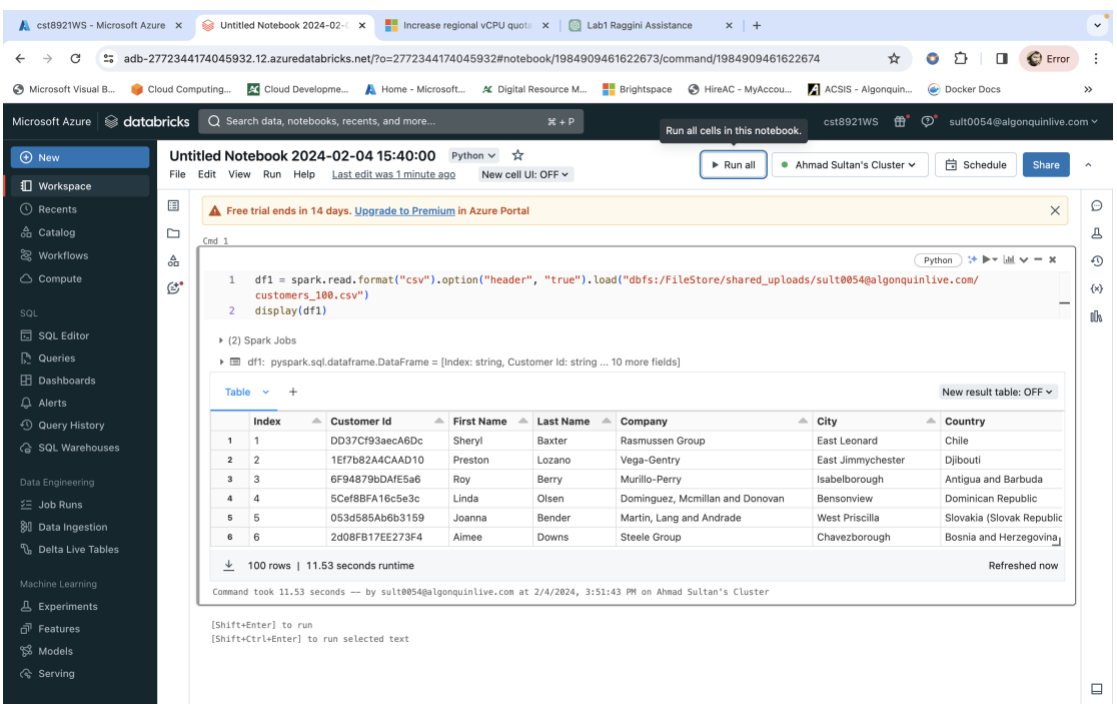


AhmadSultan_041034155

8.Explore the data using display(df1) command and visualize the results in notebook.

display(df1)





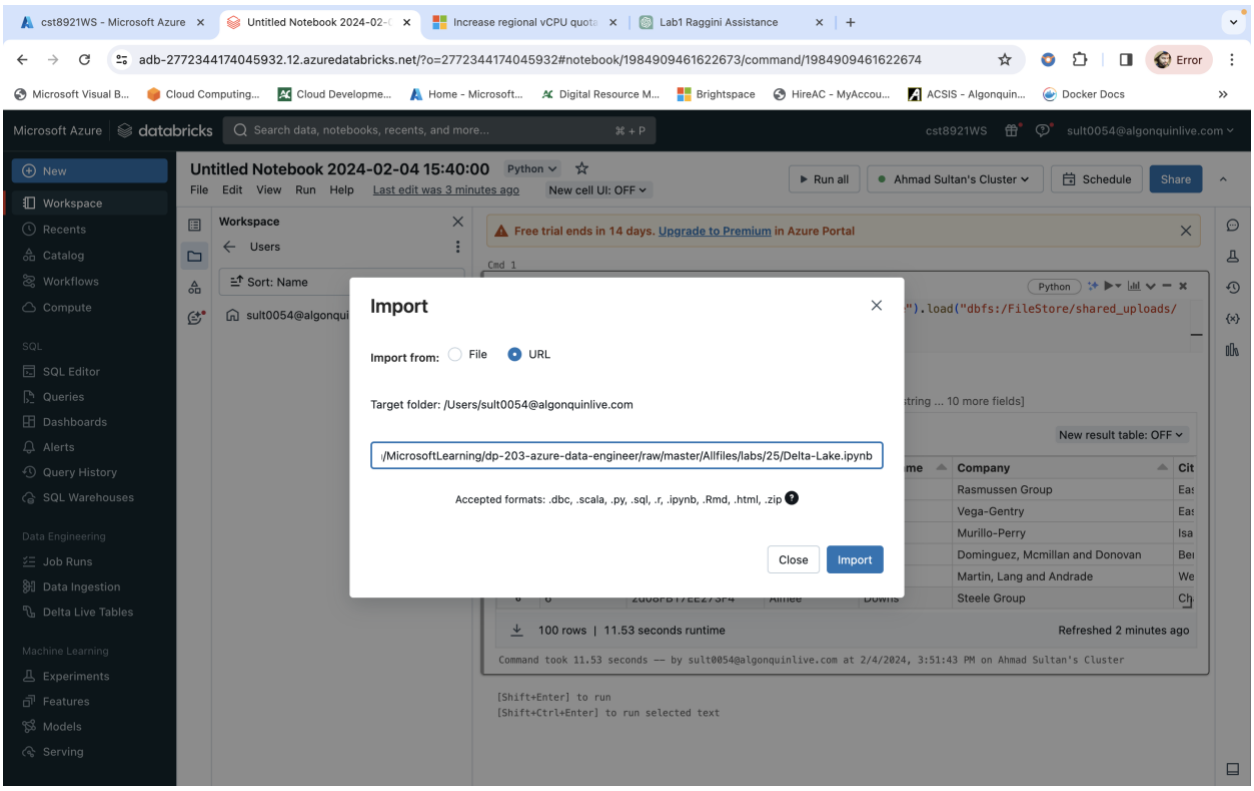
	Index	Customer Id	First Name	Last Name	Company	City	Country
1	1	DD37Cf93aecA6Dc	Sheryl	Baxter	Rasmussen Group	East Leonard	Chile
2	2	1Ef7b82A4CAAD10	Preston	Lozano	Vega-Gentry	East Jimmychester	Djibouti
3	3	6F94879bDAfE5a6	Roy	Berry	Murillo-Perry	Isabelborough	Antigua and Barbuda
4	4	5Cef8BFA16c5e3c	Linda	Olsen	Dominguez, Mcmillan and Donovan	Bensonview	Dominican Republic
5	5	053d585Ab6b3159	Joanna	Bender	Martin, Lang and Andrade	West Priscilla	Slovakia (Slovak Republic
6	6	2d08FB17EE273F4	Aimee	Downs	Steele Group	Chavezborough	Bosnia and Herzegovina

AhmadSultan_041034155

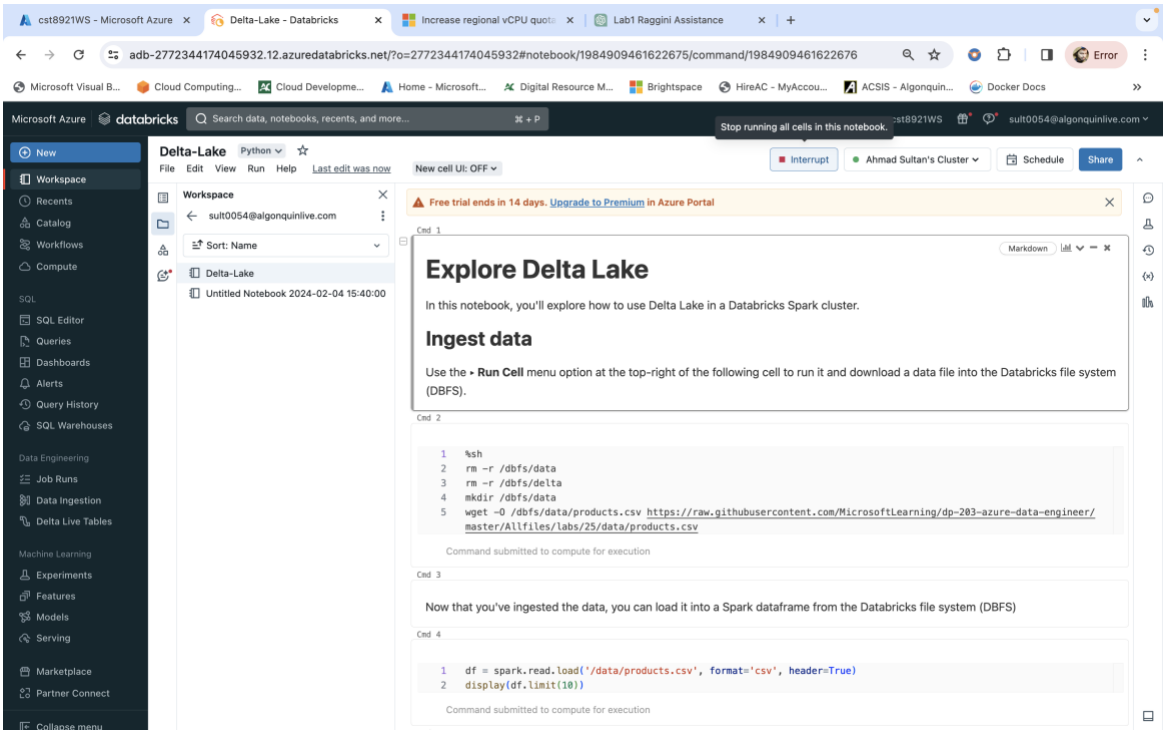
Part 2: Use Delta Lake in Azure Databricks

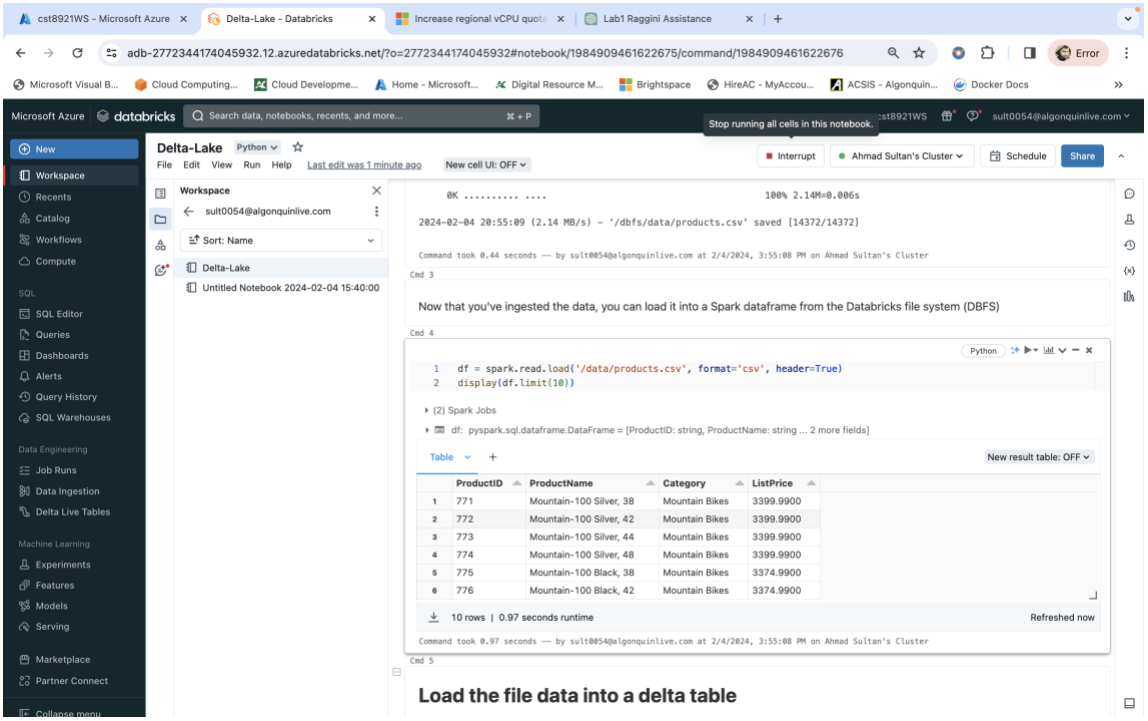
1. In the Azure Databricks workspace portal for your workspace, in the sidebar on the left, select Workspace. Then select the Home folder.

2. At the top of the page, in the ⋮ menu next to your user name, select Import. Then in the Import dialog box, select URL and import the notebook from https://github.com/MicrosoftLearning/dp-203-azure-data-engineer/raw/master/Allfiles/labs/25/Delta-Lake.ipynb

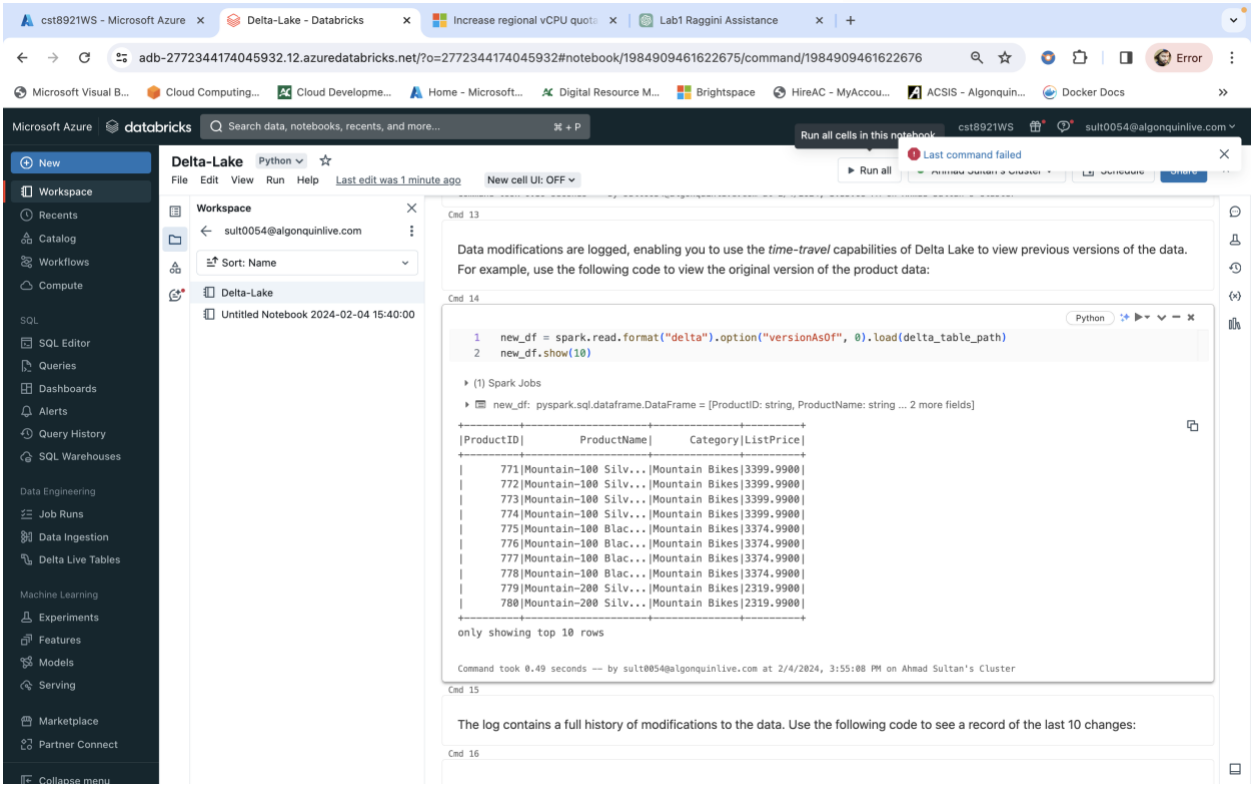


2. Connect the notebook to your cluster, and follow the instructions it contains; running the cells it contains to explore delta lake functionality.



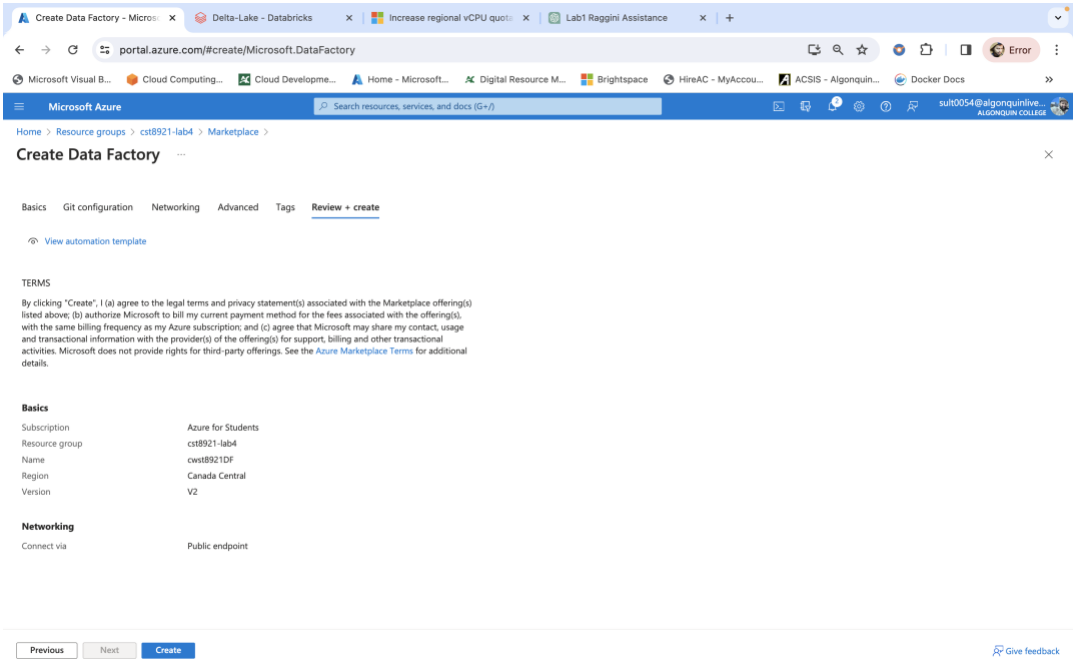


AhmadSultan_041034155

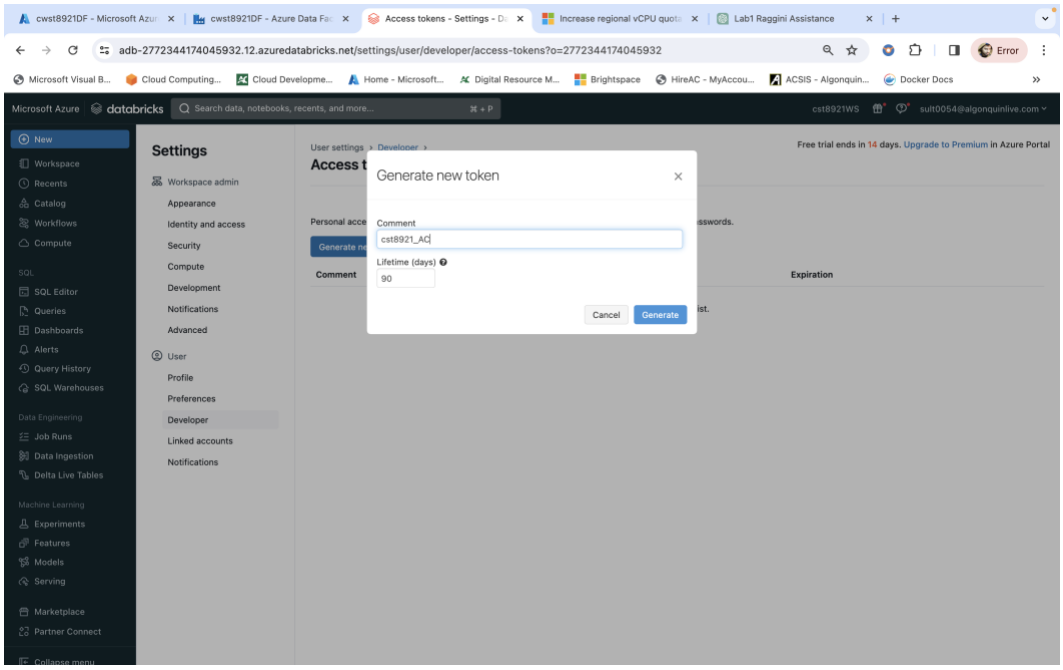


Part 3: Execute Databricks notebook from Azure data factory

1. Use the notebook created in the part 1 to execute from azure data factory

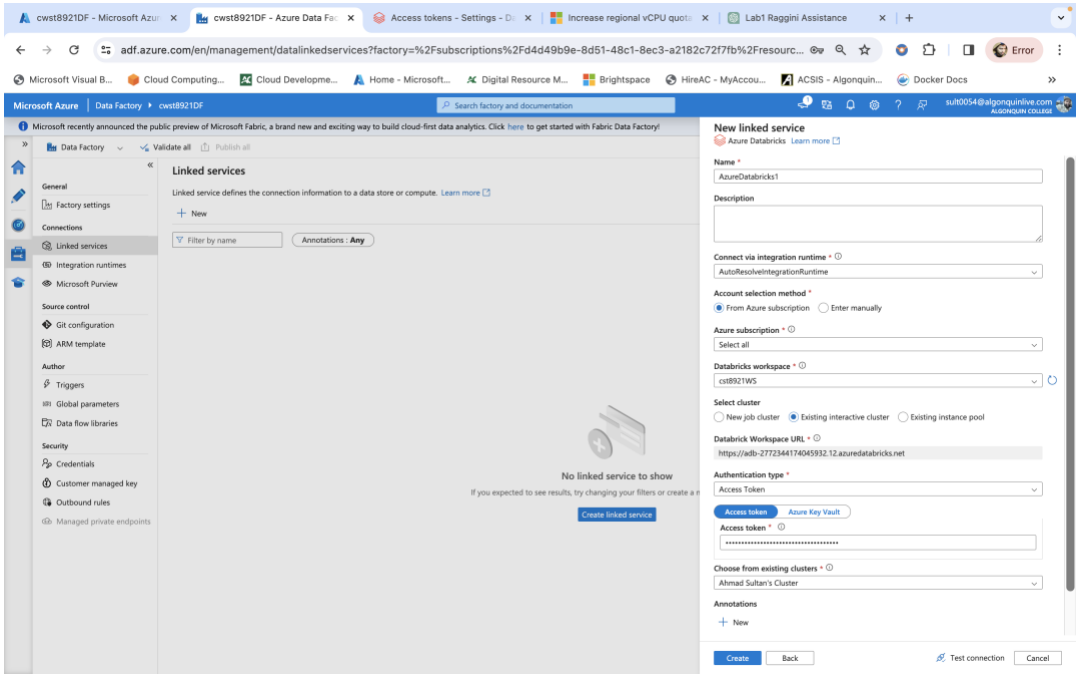


2. generate access token from databricks notebook

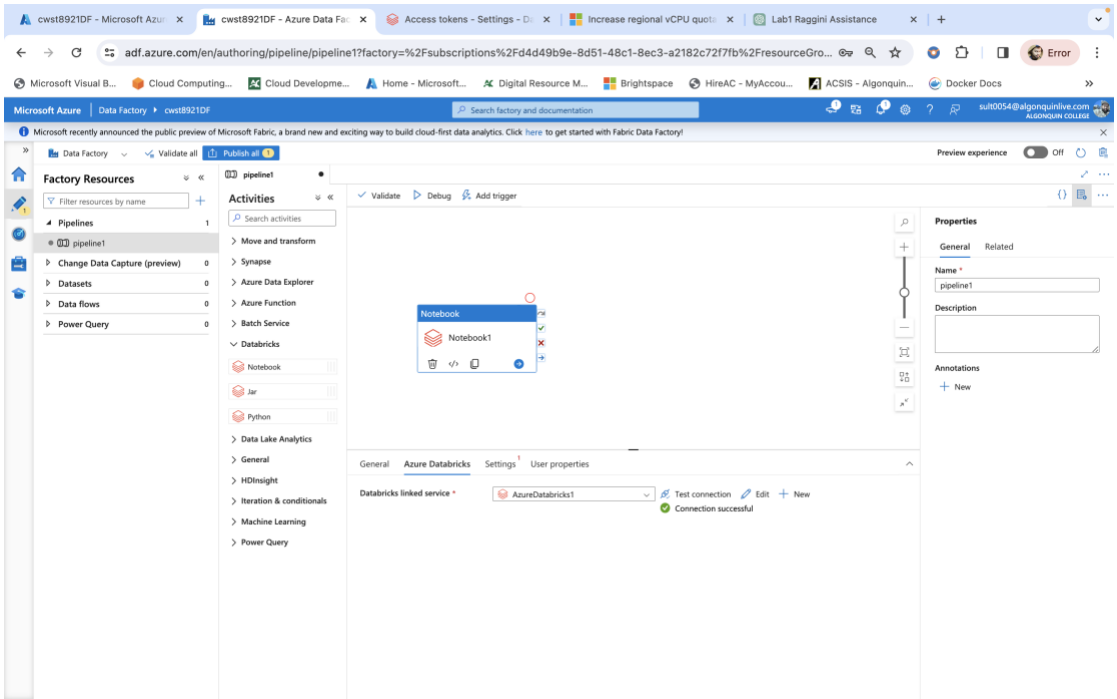


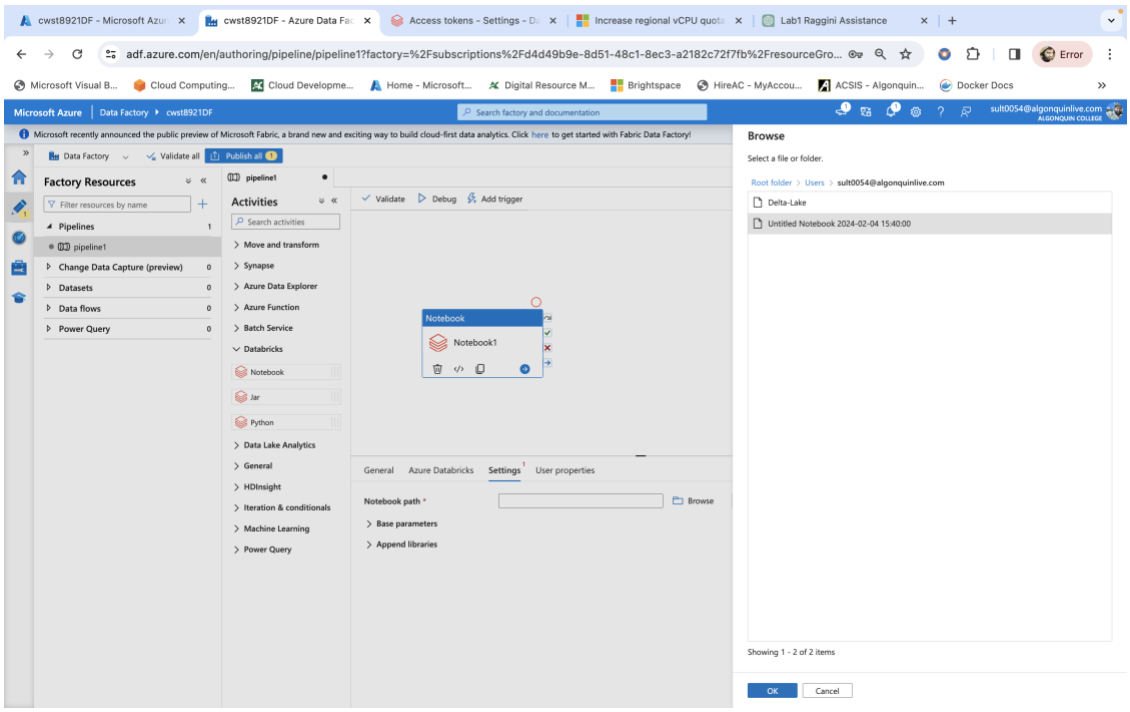
dapib8464e08757e053a4b332057ce1aa74b

3. create azure data factory instance and create a linked service to enable access to databricks workspace.



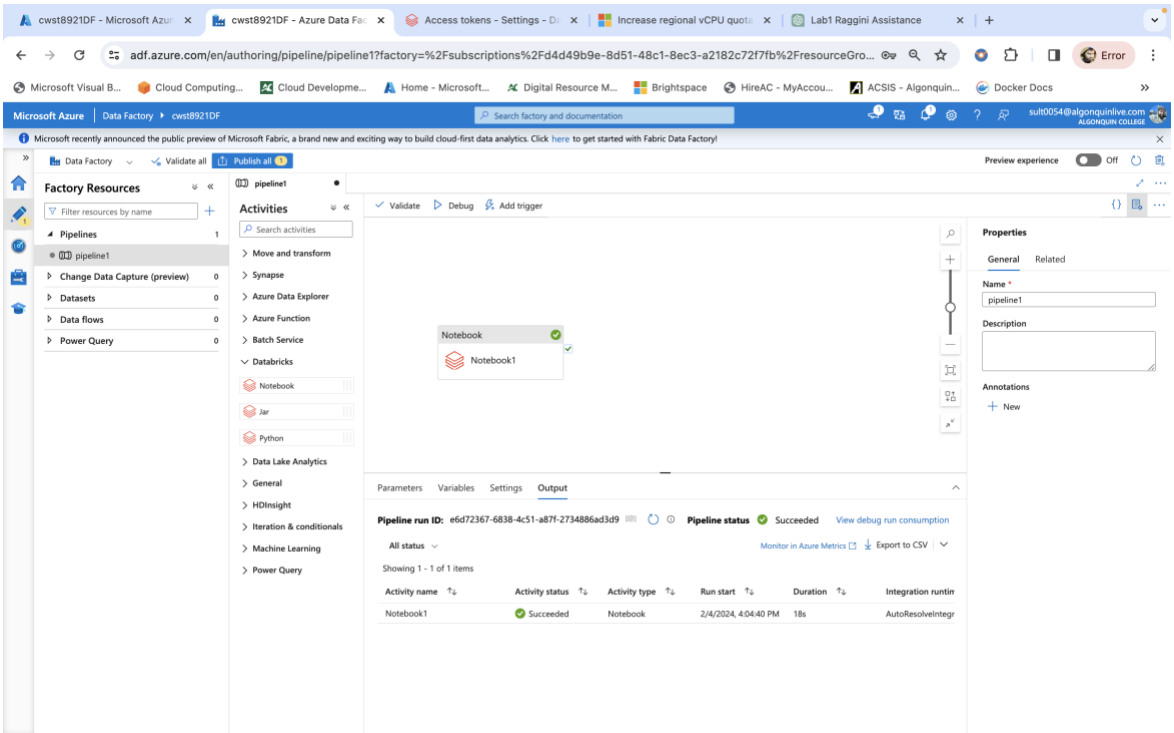
4. Create a pipeline to run the notebook from data factory.

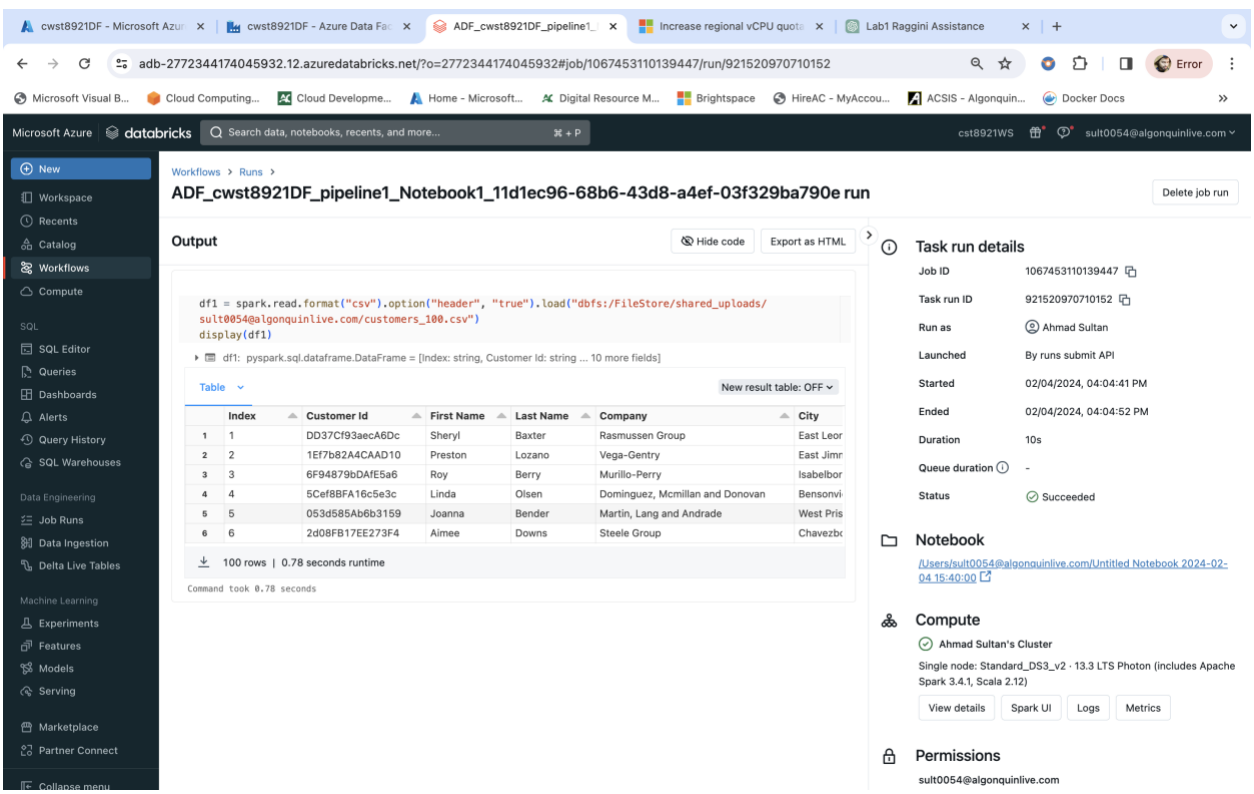




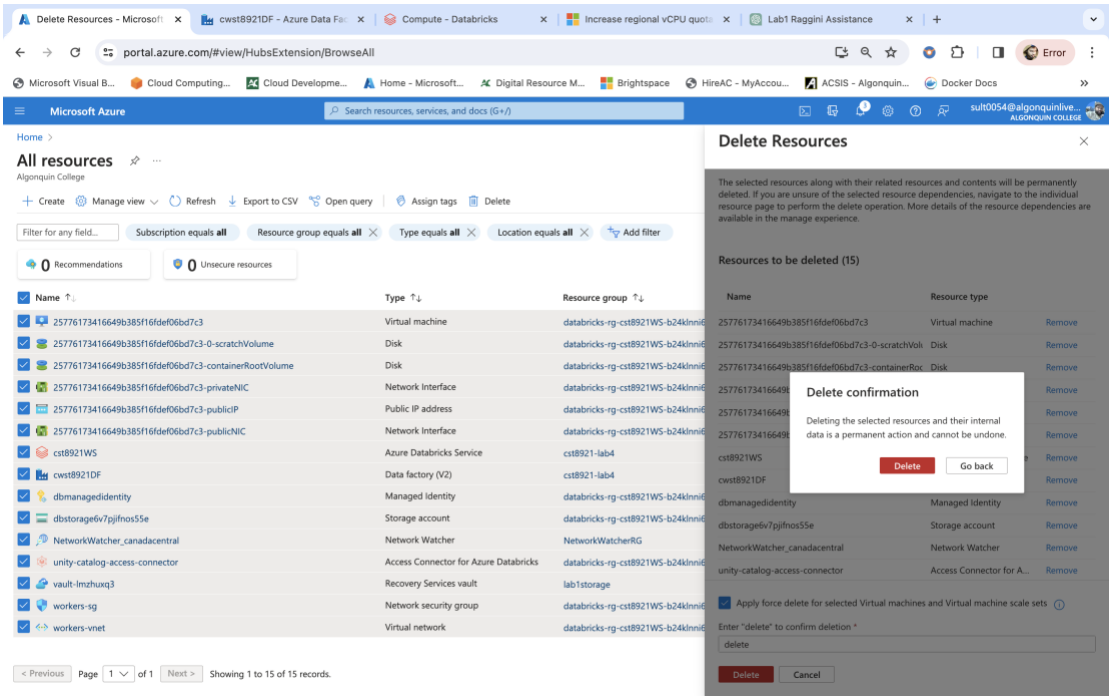
AhmadSultan_041034155

5.Execute the pipeline and monitor the run to see the status of pipeline if it is successful or not.





6. Delete all the resources created in the lab.



Results

1. Gain hands-on experience in provisioning Azure Databricks, creating clusters, and analyzing data using Spark and PySpark.
2. Explore Delta Lake capabilities, understanding how it enhances data reliability and performance.
3. Learn to integrate Azure Data Factory with Databricks, executing notebooks through pipelines and monitoring runs for efficient data processing.