**Case Study Assessment**

**Name:** Akshay Purushanasa

**UID :** 261337

**Email :** Akshay.purushanasa@ust.com

**Case Study 1: Retail Sales Data Processing (using Azure blob storage and Azure Data Factory)**

1. Setting up Azure blob storage, created container and uploaded the CSV fileA computer screen shot of a computer

   Description automatically generated
2. **Configure Azure Data Factory**

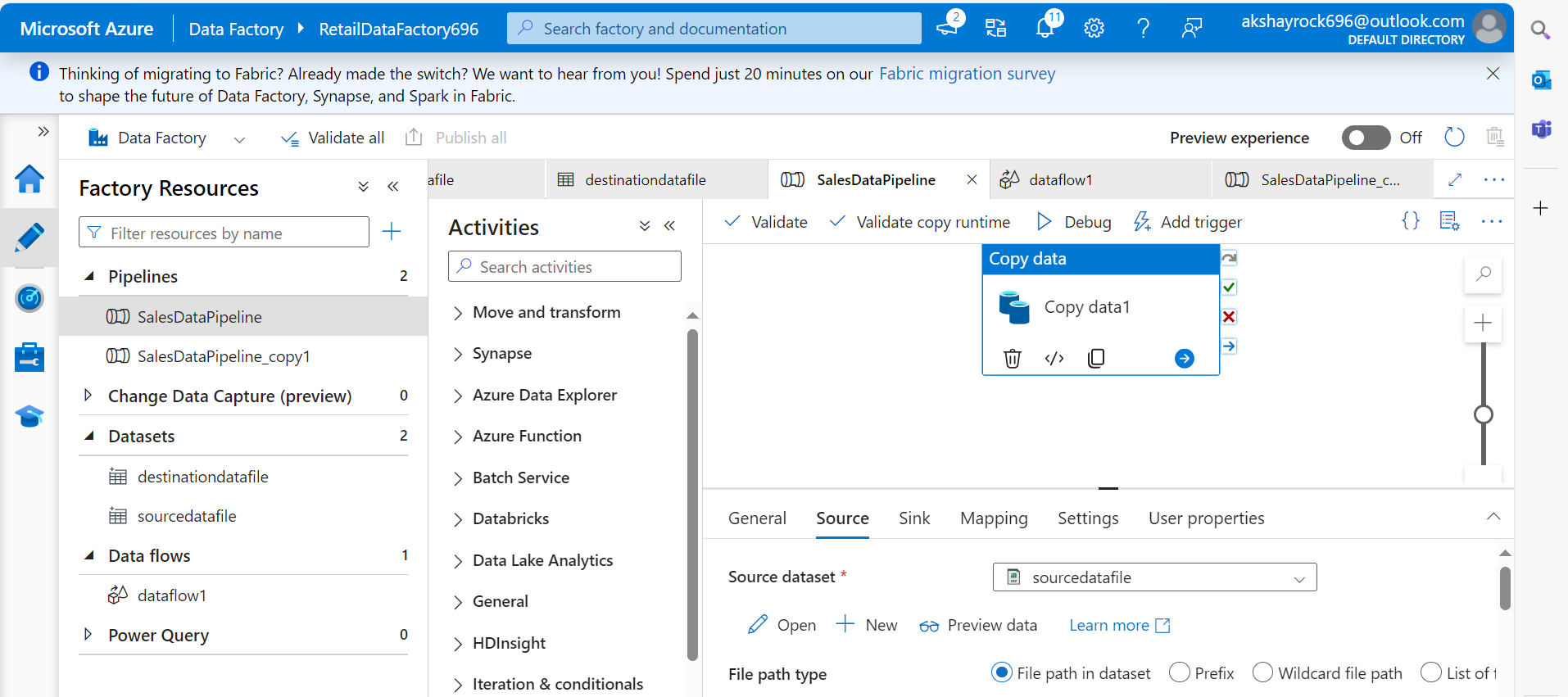
* Created Azure Factory A screenshot of a chat

  Description automatically generated
* Created Link Serivce A screenshot of a computer

  Description automatically generated
* Created Dataset Source and destination A screenshot of a computer

  Description automatically generated

1. Build and Configure Pipeline

* Created pipeline with Copy data activity select source and sink file 

**Case Study 2: Retail Sales Data Processing (using Azure DataBricks)**

1. Setting up Azure blob storage, created container and uploaded the CSV fileA computer screen shot of a computer

   Description automatically generated
2. **Set Up Azure Databricks**

* Create an Azure Databricks Workspace A screenshot of a computer

  Description automatically generated
* Creating cluster A screenshot of a computer

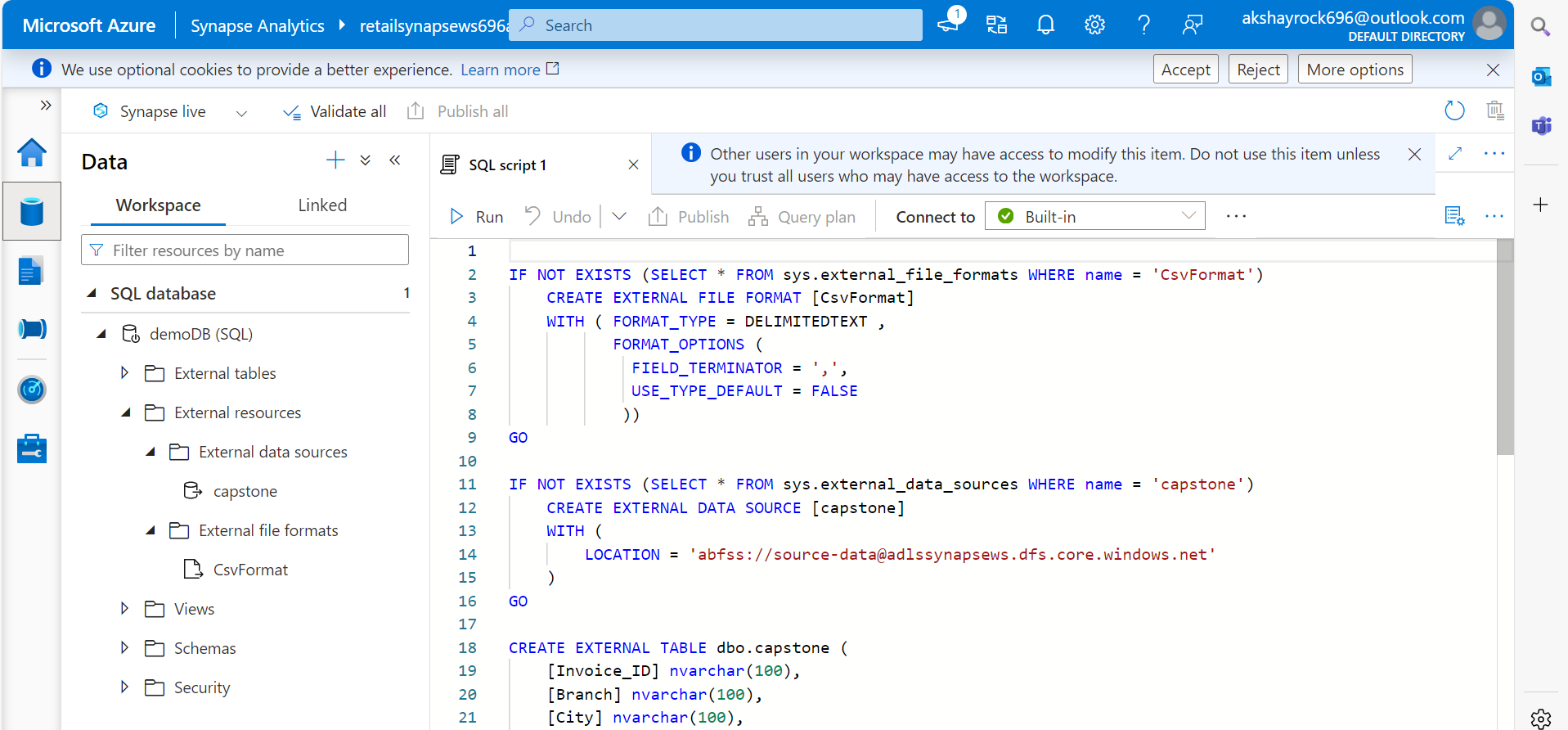
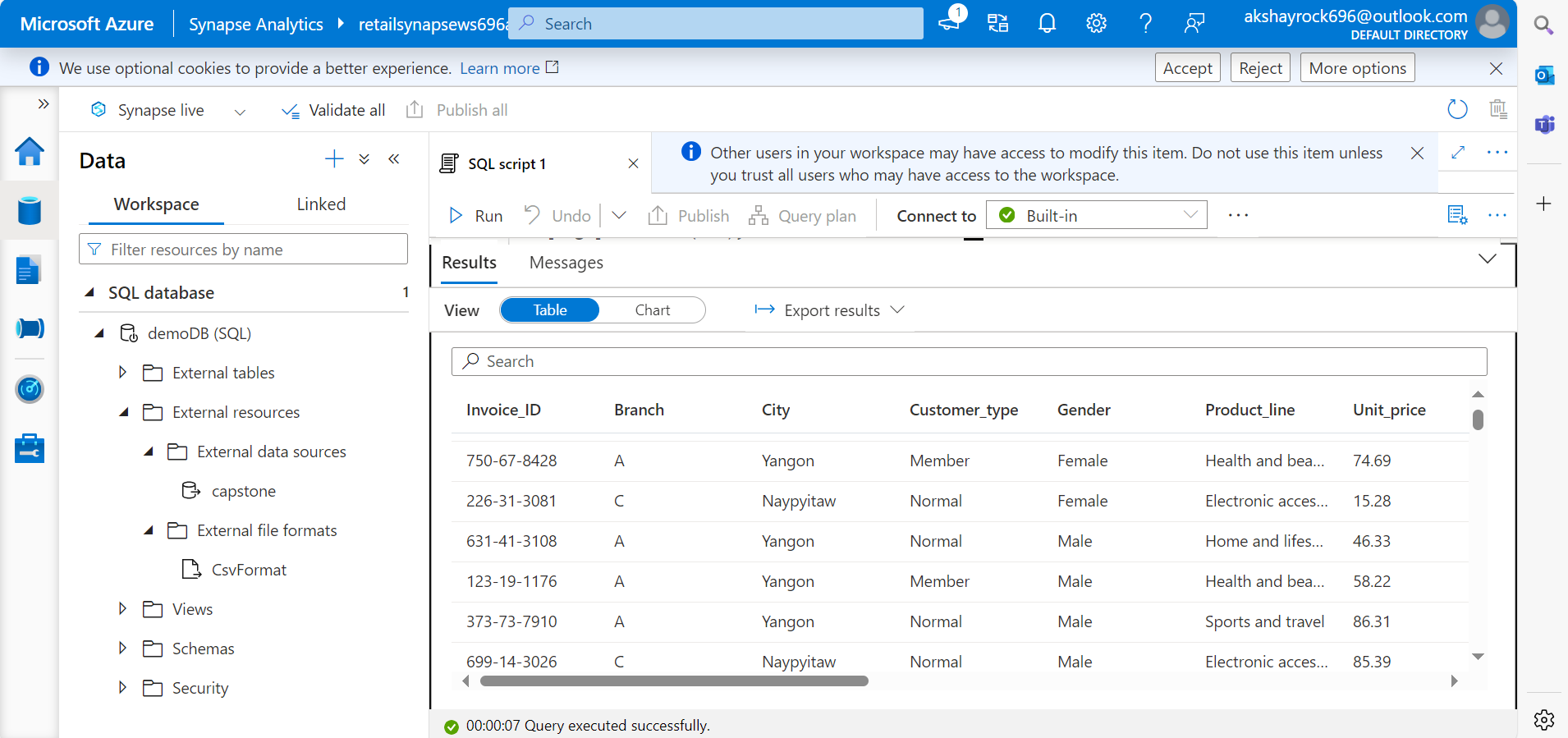
  Description automatically generated
* Created Databricks notebook A screenshot of a computer

  Description automatically generated
* Code which wrote in notebook
* from pyspark.sql import SparkSession
* spark = SparkSession.builder.appName("Data Analysis using Pyspark")\
* .config("spark.memory.offHeap.enabled","true").config("spark.memory.offHeap.size","20g").getOrCreate()
* file\_path = "wasbs://source-data@xyzretailstorage696.blob.core.windows.net/sales\_data.csv"
* spark.conf.set("fs.azure.account.key.xyzretailstorage696.blob.core.windows.net", "0P1gEkH1Tgf9KPuo7EOQ+nx/Kb2UPomZr3RZhGwxy0RhEoxmVyeAek5/j/UniFtGVHoj9Nm5wMEU+AStiBbhrQ==")
* #read file
* df = spark.read.format("csv").option("header", "true").load(file\_path)
* display(df)
* #deleted full row which is having null values
* df\_cleaned = df.dropna(how="any")
* display(df\_cleaned)
* #aggrigating function and
* df\_agg = df\_cleaned.groupBy("Branch").agg({"Total":"sum"}).withColumnRenamed("sum(Total)", "Total\_sales")
* display(df\_agg)
* spark.conf.set("fs.azure.account.key.xyzretailstorage696.blob.core.windows.net", "0P1gEkH1Tgf9KPuo7EOQ+nx/Kb2UPomZr3RZhGwxy0RhEoxmVyeAek5/j/UniFtGVHoj9Nm5wMEU+AStiBbhrQ==")
* #write file
* df\_agg.write.format("csv") \
* .option("header", "true") \
* .mode("overwrite") \
* .save("wasbs://transformed-data@xyzretailstorage696.blob.core.windows.net/sales\_data\_output.csv")
* df\_agg.show()

**Case Study 3 : Analyzing Retail Sales Data with Azure Synapse Analytics**

1. Setting up Azure data lake storage, created container and uploaded the CSV file A screenshot of a computer

   Description automatically generated
2. Creating Azure Synapse A screenshot of a computer

   Description automatically generated
3. Created Data Source 
4. Result

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*THANK YOU\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*