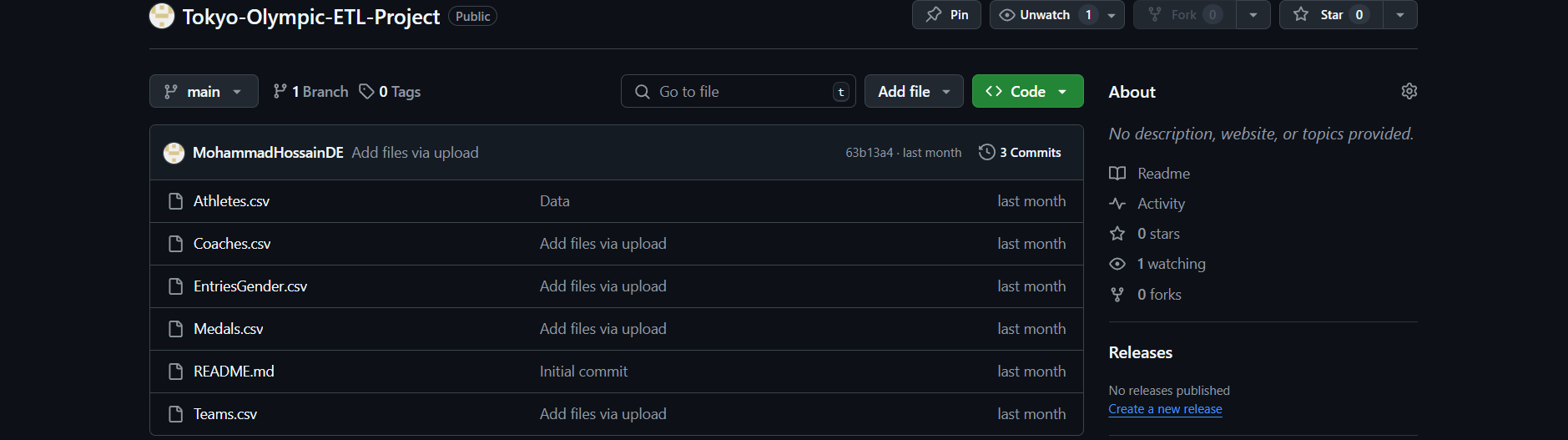
Mohammad Mubarak Hossain

**DE2023**

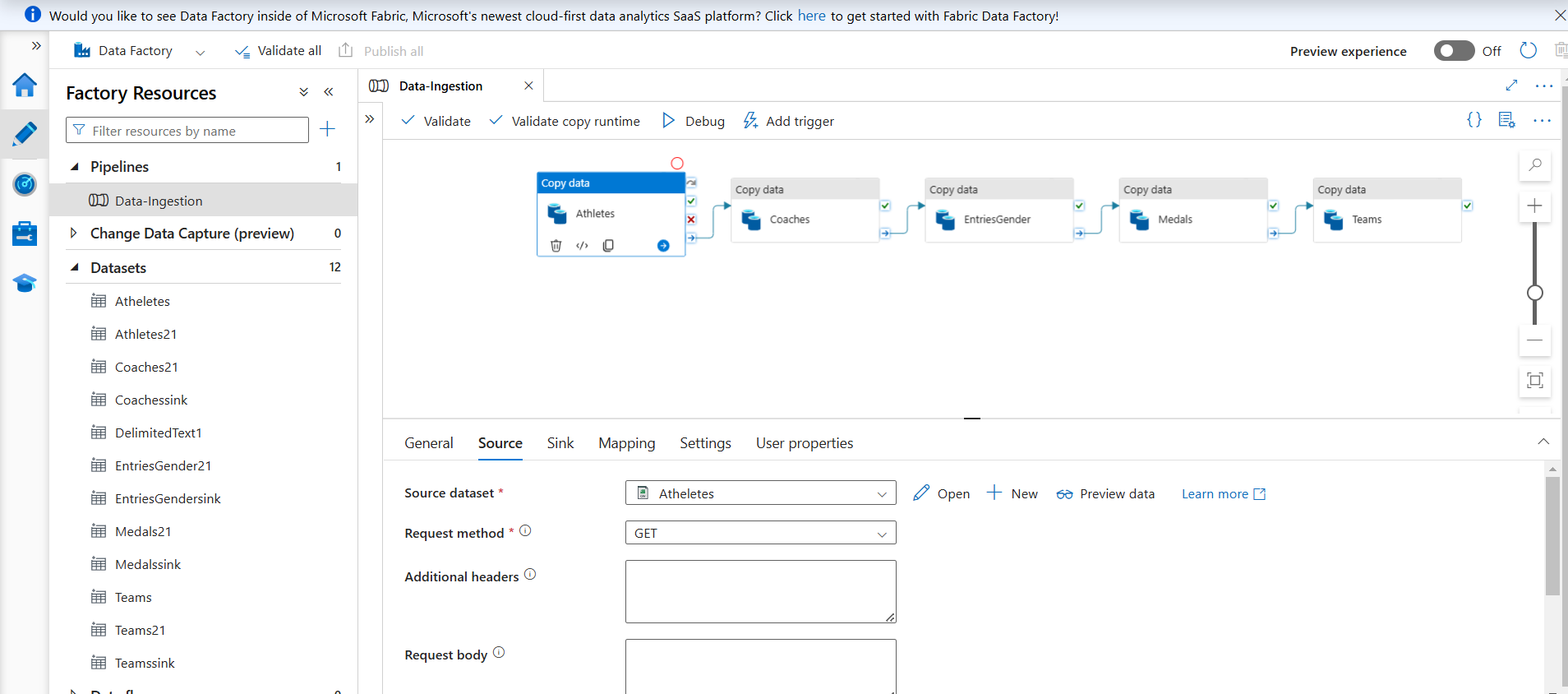
**Kurs:** [Datafångst, migrering och förädling (ETL/ELT)](https://studentportal.nackademin.se/course/view.php?id=930)

ETL Flow  
 Olympic Data Analysis

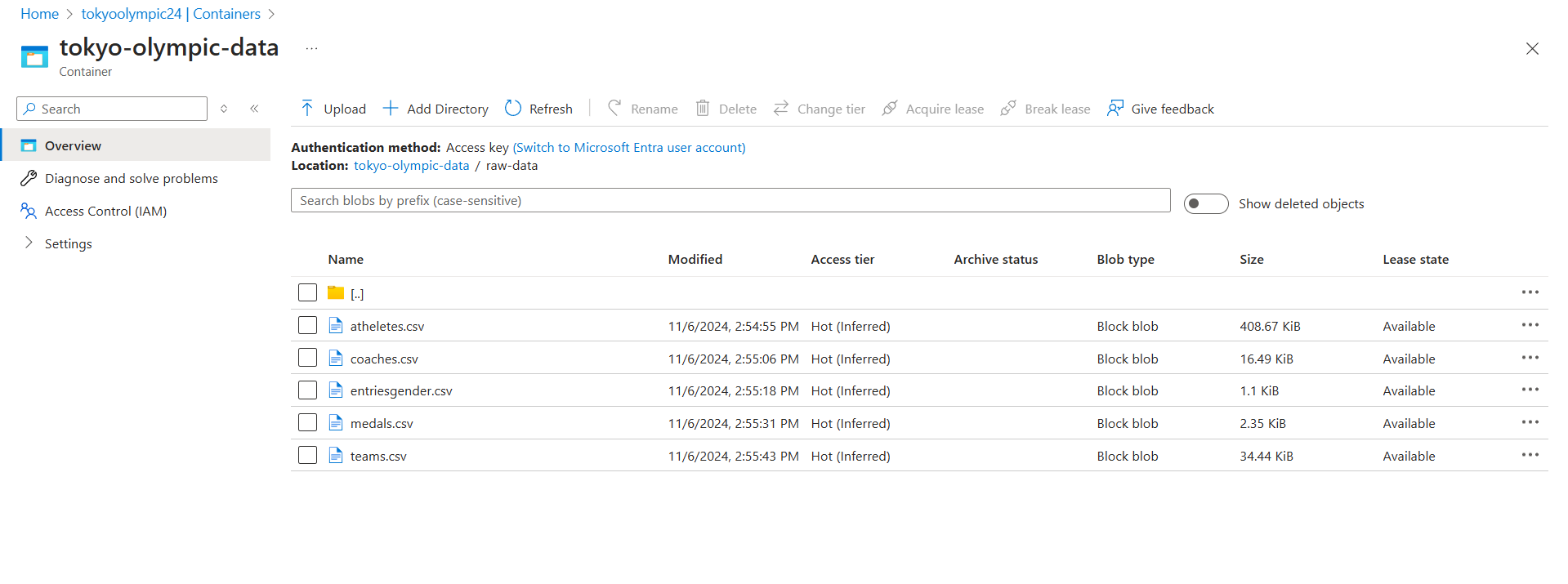
**Data Source:** Jag har tagit min data från Github och det är CSV filen.

****

* **Data Factory:** Jag har skapat pipeline I data factory för att skicka min raw data till blob storage.

****

* **Azure Blob Storage:** Jag har skapat två mappen I blob storage container och en för raw data och en till för transform data. Data factory skickar min raw data till raw data mappen.

****

* **DataBricks:** Jag har skapat azure databricks för att data transformera och data analisera.

from pyspark.sql.functions import col

from pyspark.sql.types import IntegerType, DoubleType, BooleanType, DateType

configs = {

    "": ""

}

# Use wasbs for Blob Storage instead of abfss for ADLS Gen2

dbutils.fs.mount(

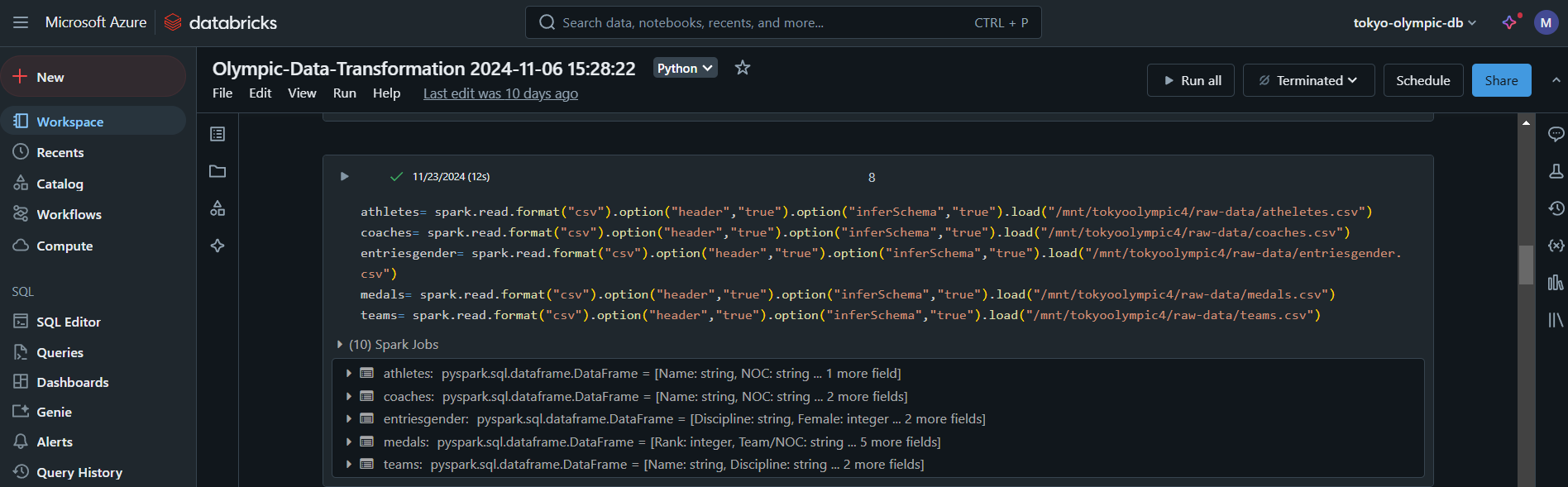
    source = "wasbs://tokyo-olympic-data@tokyoolympic24.blob.core.windows.net/",

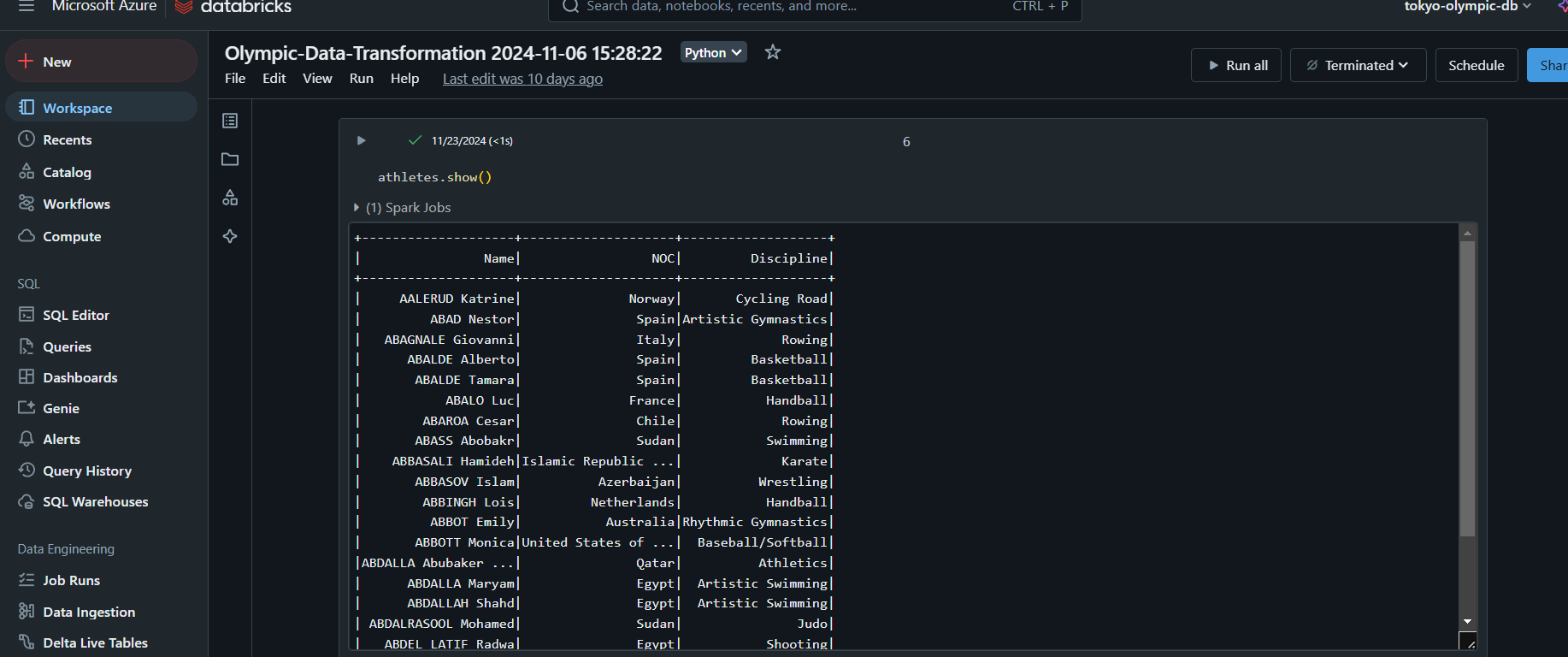
    mount\_point = "/mnt/tokyoolympic4",

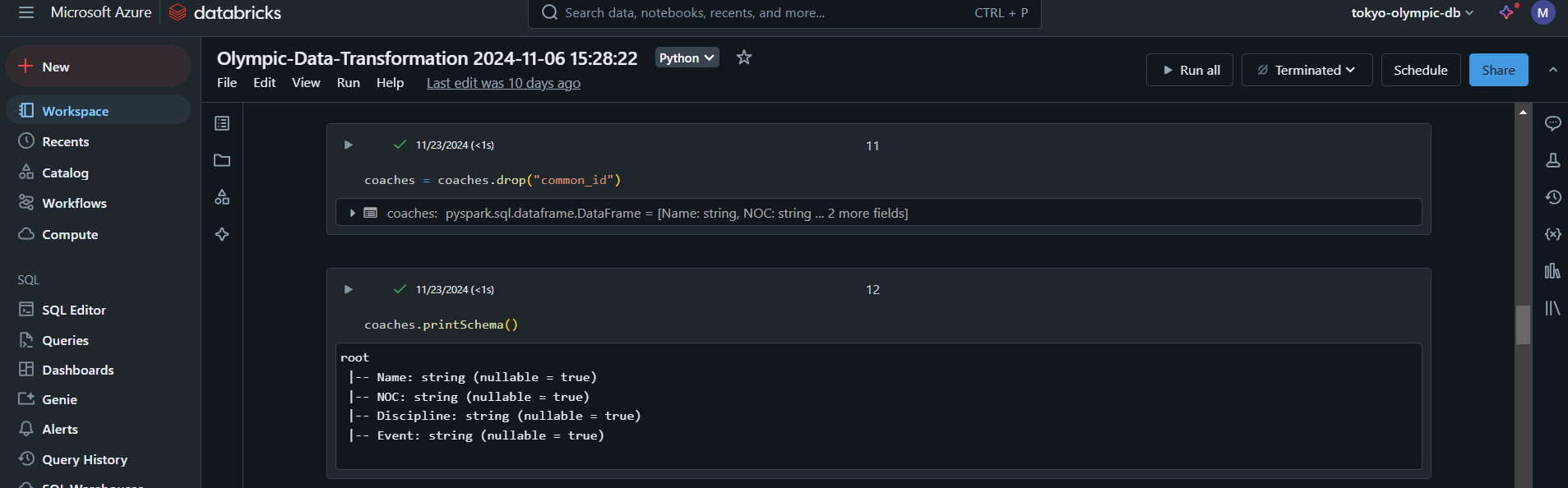
    extra\_configs = configs

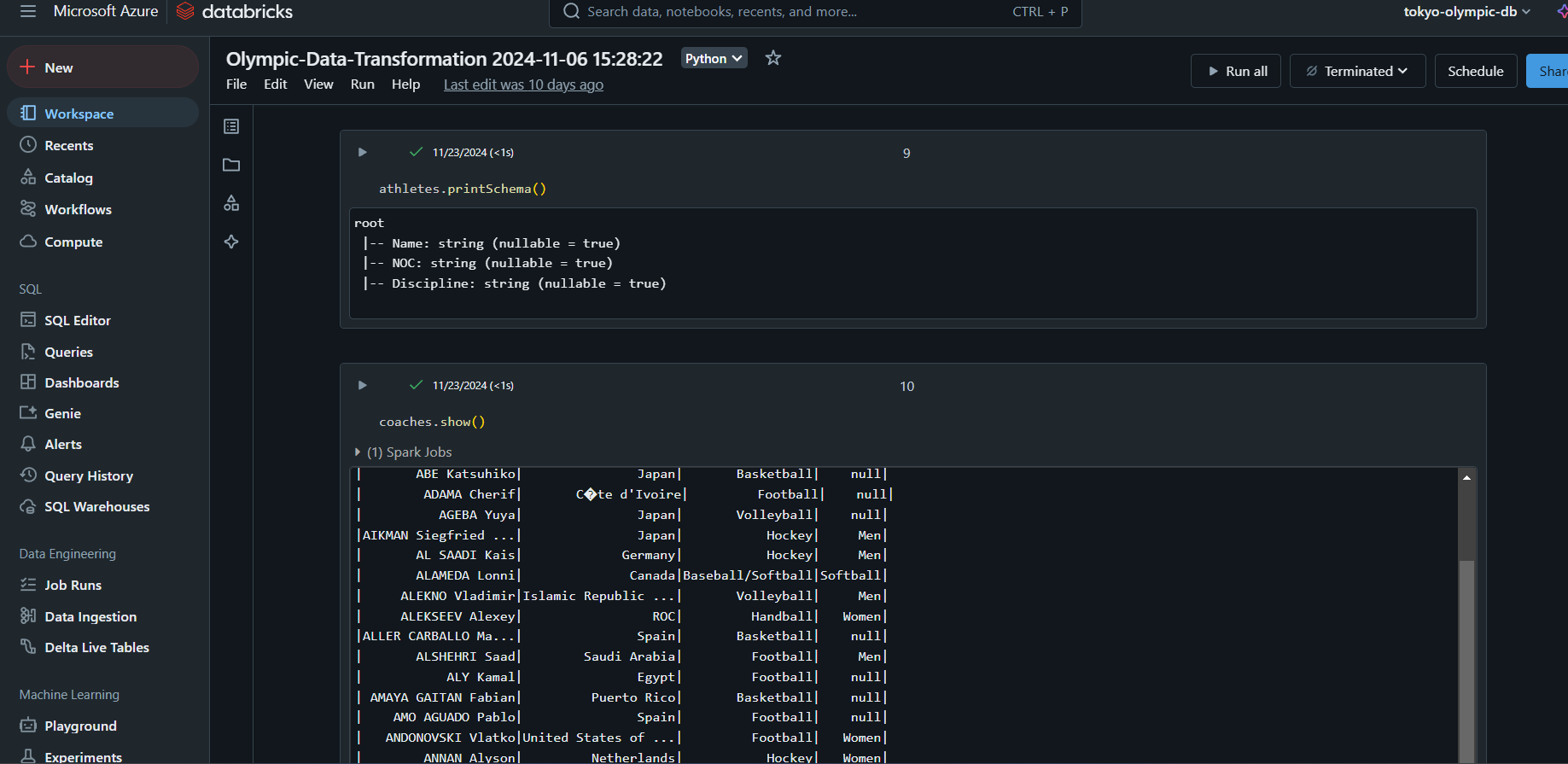
)

dbutils.fs.ls("/mnt/tokyoolympic4/")

****

****

****

****

coaches = coaches.drop("common\_id")

coaches.printSchema()

entriesgender.show()

entriesgender.printSchema()

entriesgender = entriesgender.withColumn("Female",col("Female").cast(IntegerType()))\

    .withColumn("Male",col("Male").cast(IntegerType()))\

    .withColumn("Total",col("Total").cast(IntegerType()))

medals.show()

medals = medals.withColumnRenamed("Team", "NOC").withColumnRenamed("Rank by Total", "Rank\_by\_Total")

medals.show()

medals.printSchema()

teams.show()

teams.printSchema()

#Find the top countries with the highest number of gold medals

top\_gold\_medal\_countries = medals.orderBy("Gold", ascending=False).select("NOC", "Gold")

top\_gold\_medal\_countries.show()

#calculate the average number of entires by gender for each discipline

average\_entries\_by\_gender = entriesgender.withColumn(

    'Avg\_female', entriesgender['Female'] / entriesgender['Total']

).withColumn(

    'Avg\_Male', entriesgender['Male'] / entriesgender['Total']

)

average\_entries\_by\_gender.show()

average\_entries\_by\_gender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/entriesgender")

entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/entriesgender")

athletes.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/athletes")

coaches.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/coaches")

entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/entriesgender")

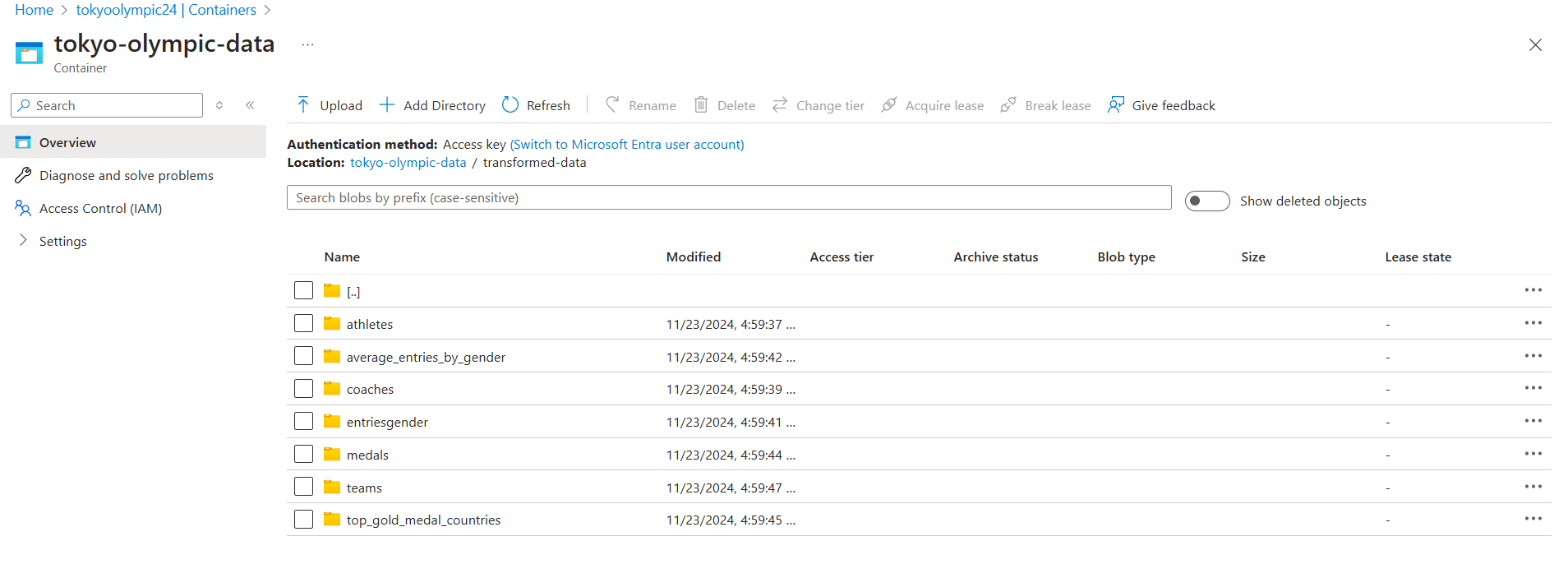
average\_entries\_by\_gender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/average\_entries\_by\_gender")

medals.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/medals")

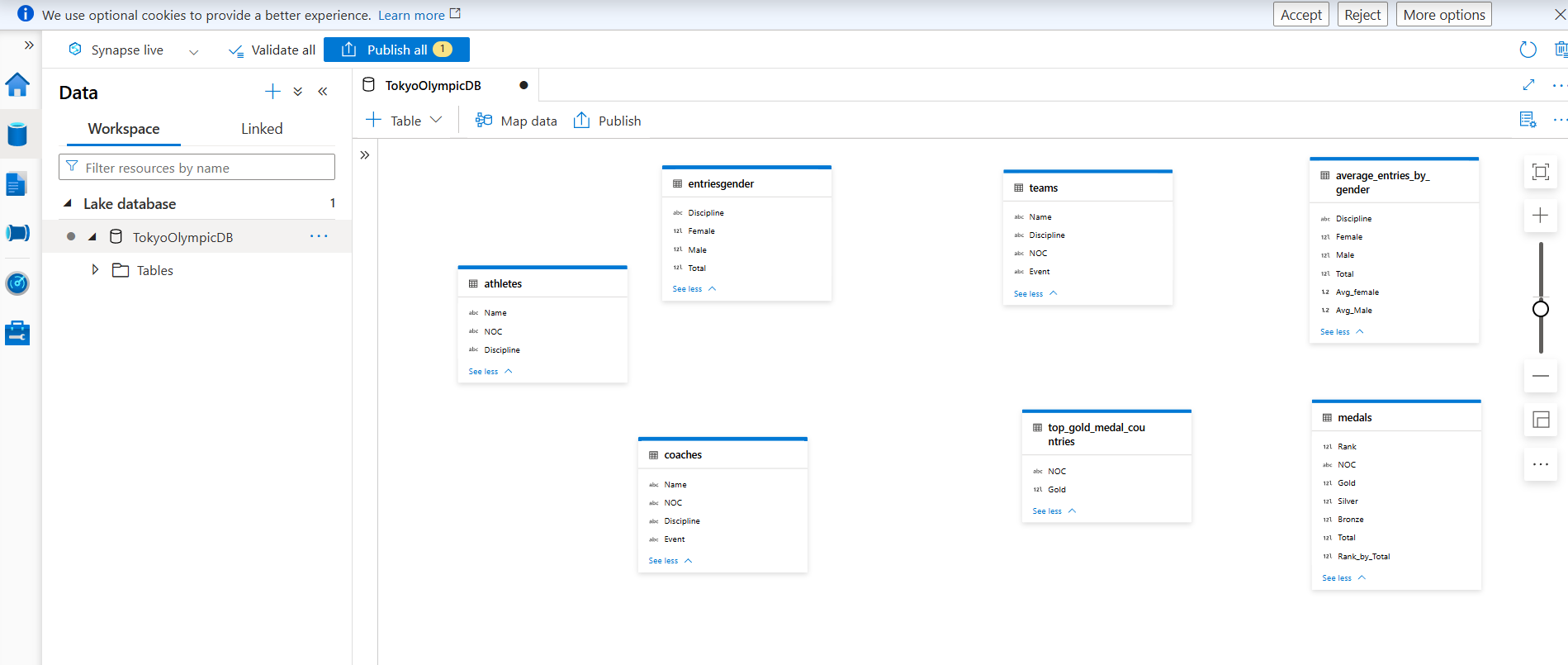
top\_gold\_medal\_countries.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/top\_gold\_medal\_countries")

teams.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympic4/transformed-data/teams")

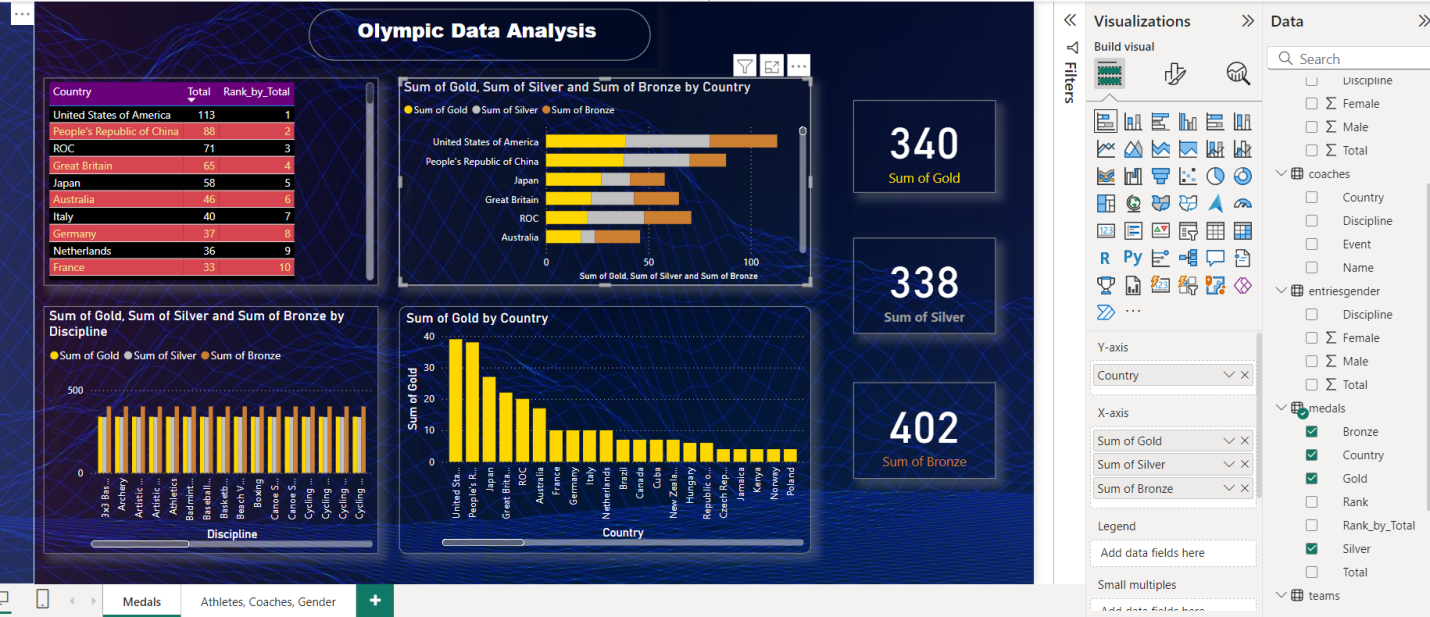
**Azure Blob Storage:** After analys skickar jag min data till transform mappen i azure blob storage

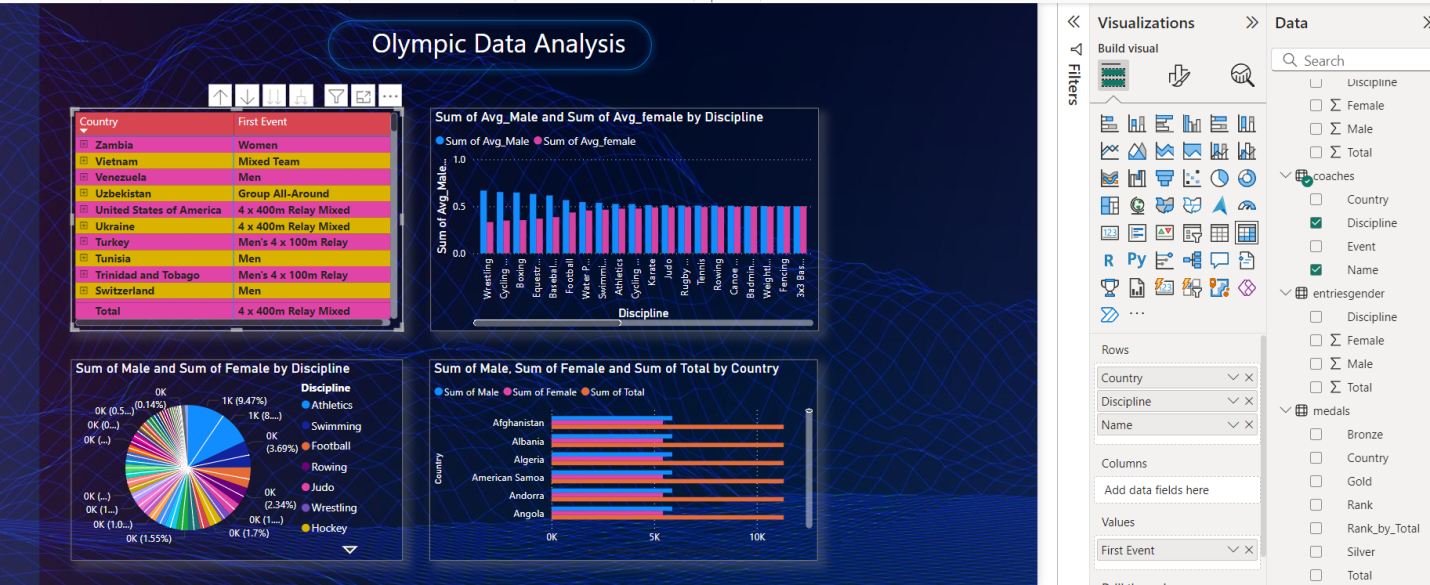
****

**Azure synapse Analytics:** Jag har skapat azure synapse analytics för att skapa azure sql database.

****

**Power BI:** därefter skickar min data till Power Bi från azure synapse för att bygga dashboard.

****

****

**Det var min ETL process för Tokyo olympic data analysis.**