# Databricks notebook source

from pyspark.sql.functions import col

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

# COMMAND ----------

service\_credential = dbutils.secrets.get(scope="my\_scope",key="secret1")

spark.conf.set("fs.azure.account.auth.type.olympicsdata187.dfs.core.windows.net", "OAuth")

spark.conf.set("fs.azure.account.oauth.provider.type.olympicsdata187.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")

spark.conf.set("fs.azure.account.oauth2.client.id.olympicsdata187.dfs.core.windows.net", “<client\_id>")

spark.conf.set("fs.azure.account.oauth2.client.secret.olympicsdata187.dfs.core.windows.net", service\_credential)

spark.conf.set("fs.azure.account.oauth2.client.endpoint.olympicsdata187.dfs.core.windows.net", “https://login.microsoftonline.com/<tenent id>/oauth2/v2.0/token")

# COMMAND ----------

spark.conf.set("fs.azure.account.key.olympicsdata187.dfs.core.windows.net", “<access key>")

# COMMAND ----------

athletes = spark.read.format("csv").option("header","true").load("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/raw-data/athletes.csv")

coaches = spark.read.format("csv").option("header","true").load("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/raw-data/coaches.csv")

entriesgender = spark.read.format("csv").option("header","true").load("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/raw-data/entriesgender.csv")

medals = spark.read.format("csv").option("header","true").load("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/raw-data/medals.csv")

teams = spark.read.format("csv").option("header","true").load("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/raw-data/teams.csv")

# COMMAND ----------

athletes.printSchema()

# COMMAND ----------

entriesgender.printSchema()

# COMMAND ----------

from pyspark.sql.functions import col, trim

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

entriesgender = entriesgender.select(

[col(c.strip()) for c in entriesgender.columns])

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

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

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

# COMMAND ----------

# COMMAND ----------

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

# COMMAND ----------

athletes.write.option("header","true").mode("overwrite").parquet("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/processed-data/athletes")

medals.write.option("header","true").mode("overwrite").parquet("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/processed-data/medals")

teams.write.option("header","true").mode("overwrite").parquet("abfss://tokyo-olympic-data@olympicsdata187.dfs.core.windows.net/processed-data/teams")