from pyspark.sql import SparkSession

# create a SparkSession

spark = SparkSession.builder.appName("AirlineDelayAnalysis").getOrCreate()

# read the data from S3

df = spark.read.format("csv").option("header", "true").load("s3://pre-assighnment/dataset/DelayedFlights-updated.csv")

# register the DataFrame as a table

df.createOrReplaceTempView("airline\_delay")

# Show the schema and first few rows of the DataFrame

df.printSchema()

df.show(5)

#1)Year wise carrier delay from 2003-2010

import time

start\_time = time.time()

print("**1 - Year wise carrier delay from 2003-2010**")

# run the queries

year\_carrier\_delay = spark.sql("**SELECT Year, SUM(CarrierDelay) as TotalCarrierDelay FROM airline\_delay WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year ORDER BY Year**")

# show the results

year\_carrier\_delay.show()

end\_time = time.time()

print("Execution time for carrier\_delay query:", end\_time - start\_time, "seconds")

#2)Year wise NAS delay from 2003-2010

start\_time = time.time()

print("**2 - Year wise NAS delay from 2003-2010**")

# run the queries

year\_nas\_delay = spark.sql("**SELECT Year, SUM(NASDelay) as TotalNASDelay FROM airline\_delay WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year ORDER BY Year**")

# show the results

year\_nas\_delay.show()

end\_time = time.time()

print("Execution time for nas\_delay query:", end\_time - start\_time, "seconds")

#3)Year wise Weather delay from 2003-2010

start\_time = time.time()

print("**3 - Year wise Weather delay from 2003-2010**")

# run the queries

year\_weather\_delay = spark.sql("**SELECT Year, SUM(WeatherDelay) as TotalWeatherDelay FROM airline\_delay WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year ORDER BY Year**")

# show the results

year\_weather\_delay.show()

end\_time = time.time()

print("Execution time for weather\_delay query:", end\_time - start\_time, "seconds")

#4)Year wise late aircraft delay from 2003-2010

start\_time = time.time()

print("**4 - Year wise late aircraft delay from 2003-2010**")

# run the queries

year\_late\_aircraft\_delay = spark.sql("**SELECT Year, SUM(LateAircraftDelay) as TotalLateAircraftDelay FROM airline\_delay WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year ORDER BY Year**")

# show the results

year\_late\_aircraft\_delay.show()

end\_time = time.time()

print("Execution time for late\_aircraft\_delay query:", end\_time - start\_time, "seconds")

#5)Year wise security delay from 2003-2010

start\_time = time.time()

print("**5 - Year wise security delay from 2003-2010**")

# run the queries

year\_security\_delay = spark.sql("**SELECT Year, SUM(SecurityDelay) as TotalSecurityDelay FROM airline\_delay WHERE Year BETWEEN 2003 AND 2010 GROUP BY Year ORDER BY Year**")

# show the results

year\_security\_delay.show()

end\_time = time.time()

print("Execution time for security\_delay query:", end\_time - start\_time, "seconds")