**Lab Exercise- Multi-Spilt**

from pyspark.sql import SparkSession

from awsglue.context import GlueContext

from pyspark.context import SparkContext

# Use the existing SparkContext and GlueContext

sc = SparkContext.getOrCreate()

glueContext = GlueContext(sc)

spark = glueContext.spark\_session

# Parameters

input\_path = "s3://hks-demo/input-customer-csv/customers.csv"

output\_path = "s3://hks-demo/output-multi-split/"

split\_size = 10 # Number of rows per file

# Read the large CSV file

df = spark.read.csv(input\_path, header=True, inferSchema=True)

# Calculate number of partitions

total\_rows = df.count()

num\_partitions = (total\_rows // split\_size) + 1

# Repartition the DataFrame

df = df.repartition(num\_partitions)

# Write the split files to S3

df.write \

.mode("overwrite") \

.option("header", "true") \

.csv(output\_path, compression='gzip')

print(f"Splitting complete. Files saved to {output\_path}")