**Mock Test 1 - Big Data Question no. 5**

**5.**Given a PySpark DataFrame named "logs" with columns "timestamp" (timestamp) and "event" (string), write a code to count the number of events that occurred in each hour and display the result sorted by the hour.

**Sol.**

from pyspark.sql import SparkSession

from pyspark.sql.functions import hour

from pyspark.sql.functions import count

spark = SparkSession.builder \

.appName("EventCountByHour") \

.getOrCreate()

# Assuming you already have the DataFrame named "logs"

# Extract hour from timestamp and count events by hour

hourly\_count\_df = logs.withColumn("hour", hour(logs.timestamp)) \

.groupBy("hour") \

.agg(count("event").alias("event\_count")) \

.orderBy("hour")

# Show the result

hourly\_count\_df.show()

# Stop the SparkSession

spark.stop()