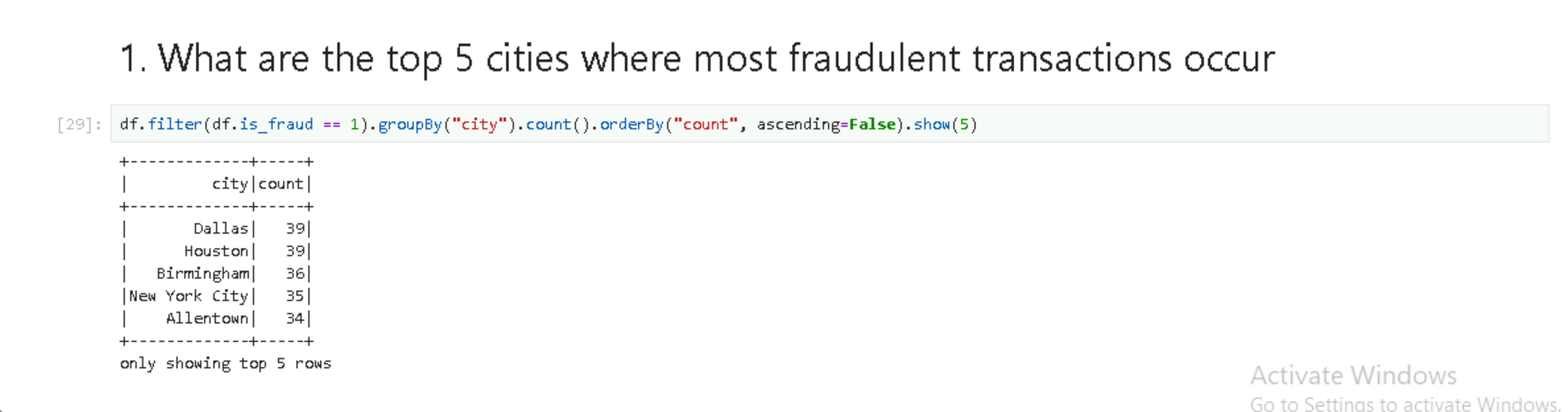
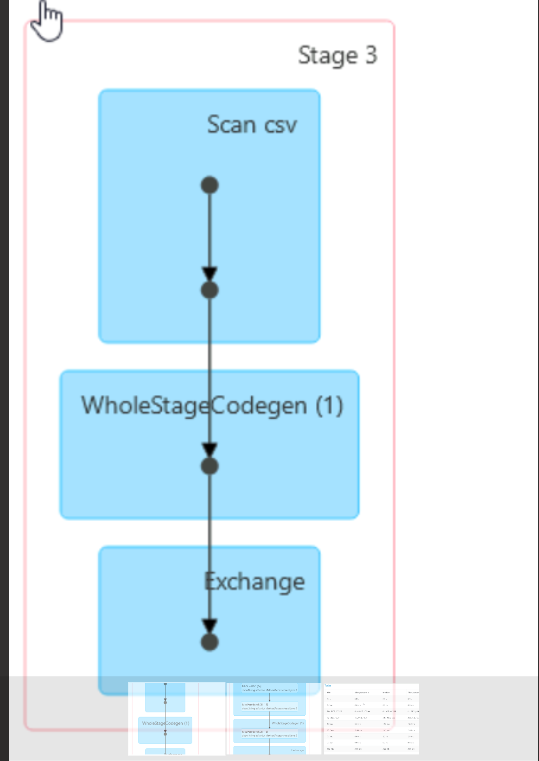
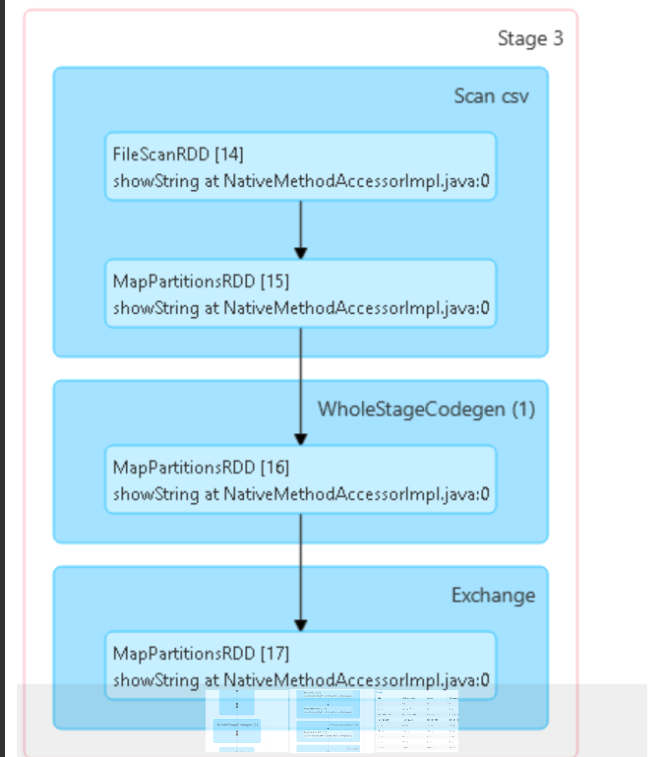
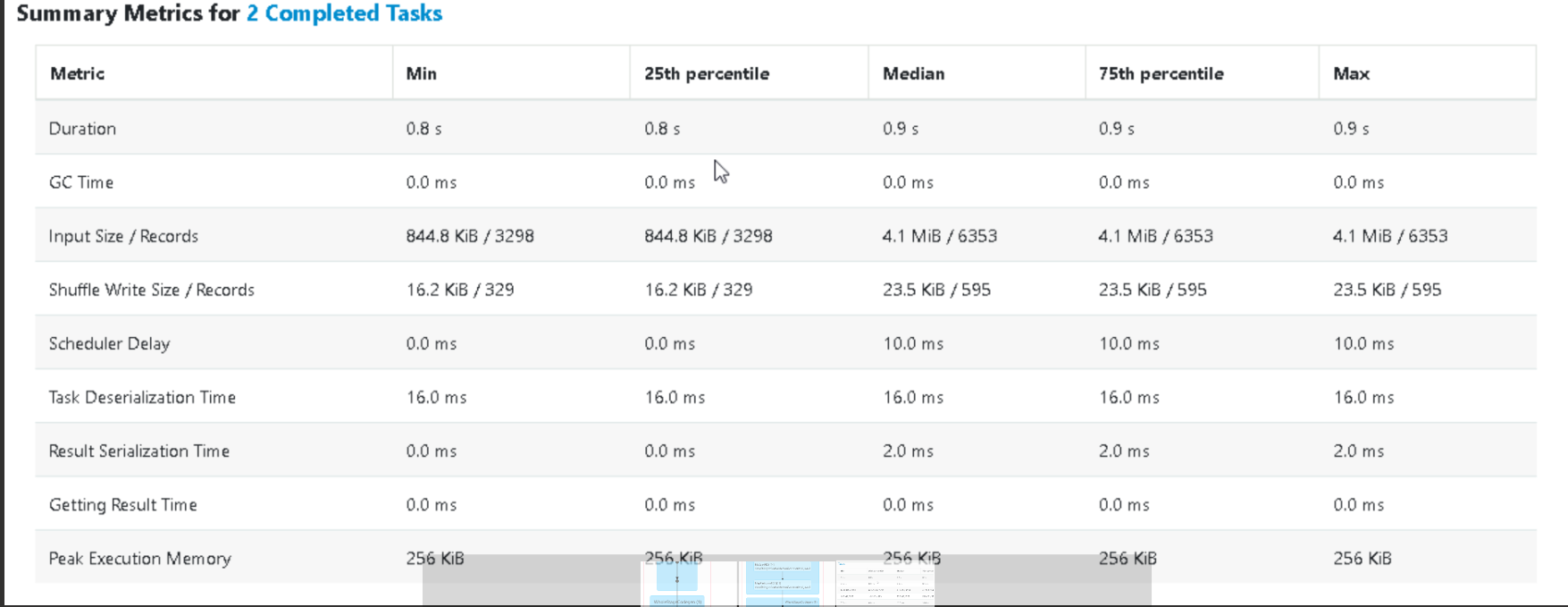
**GROUP E**

**Q1)** What are the top 5 cities where most fraudulent transactions occur?







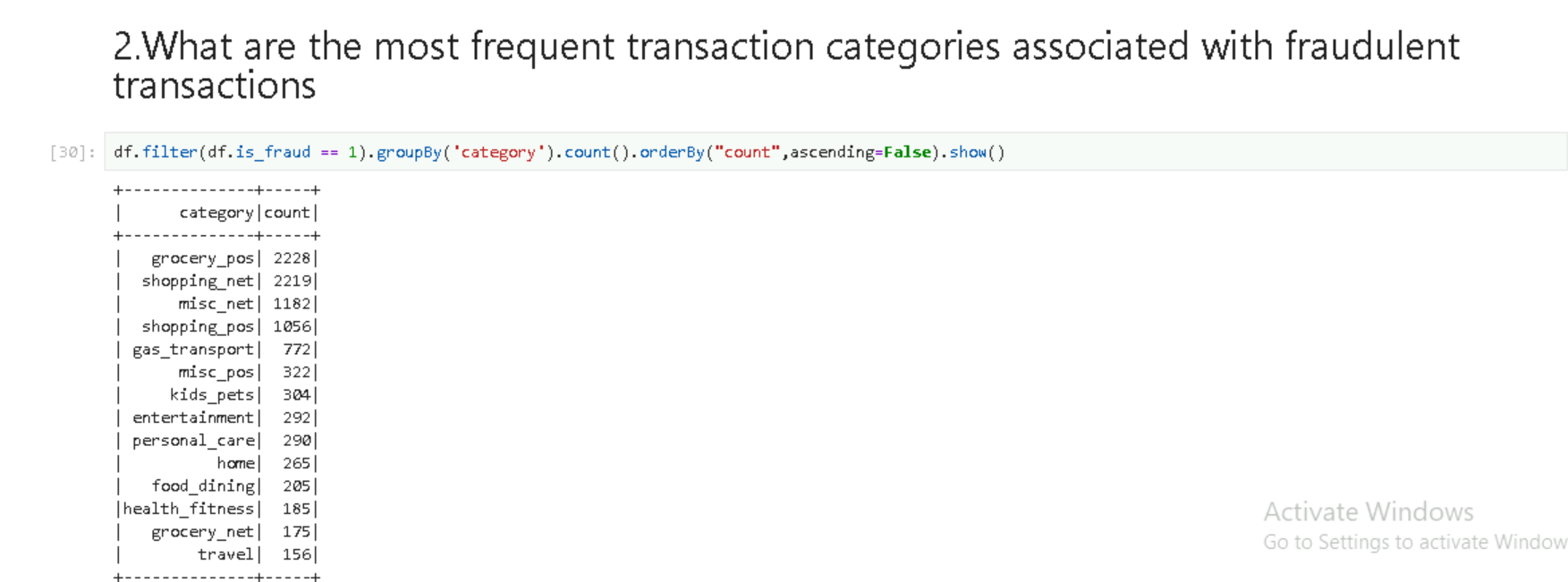


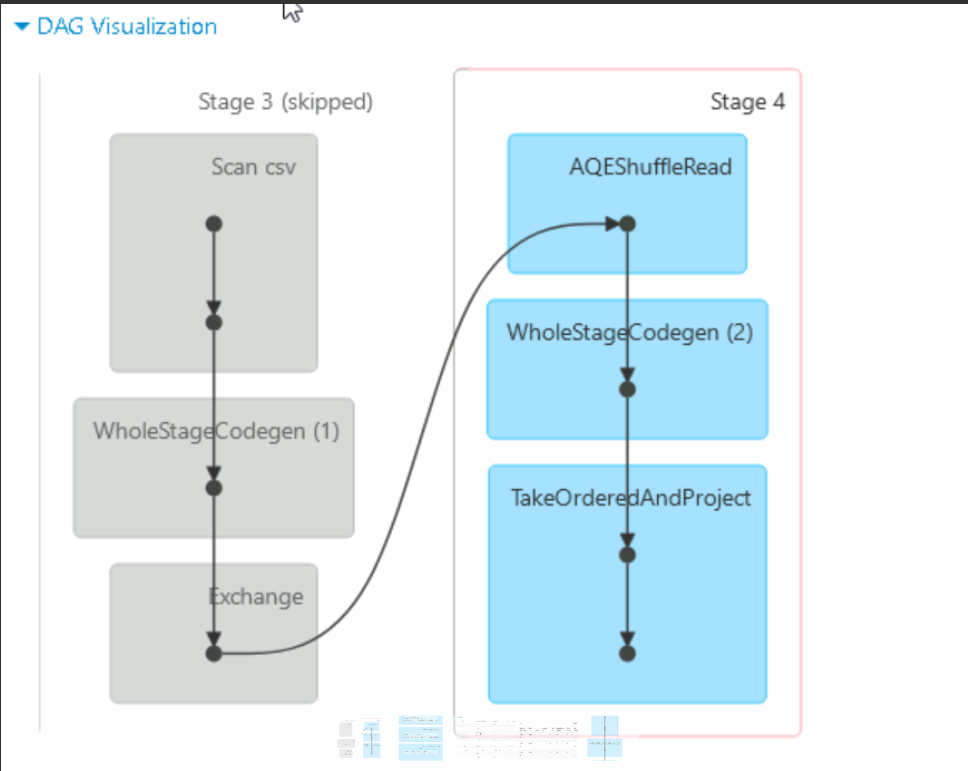
GC Time: Time spent on garbage collection during task execution.

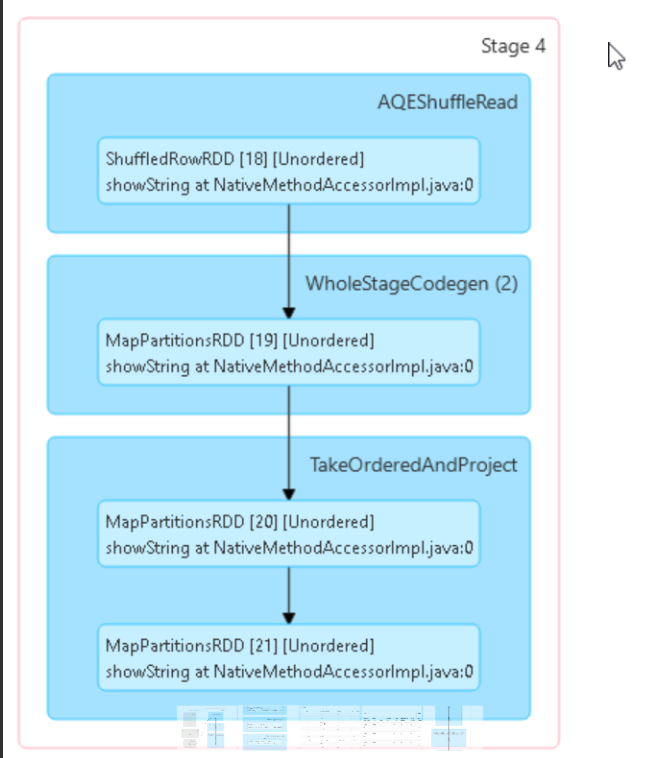
Result Serialization Time: Time taken to serialize the task result before sending it back to the driver.

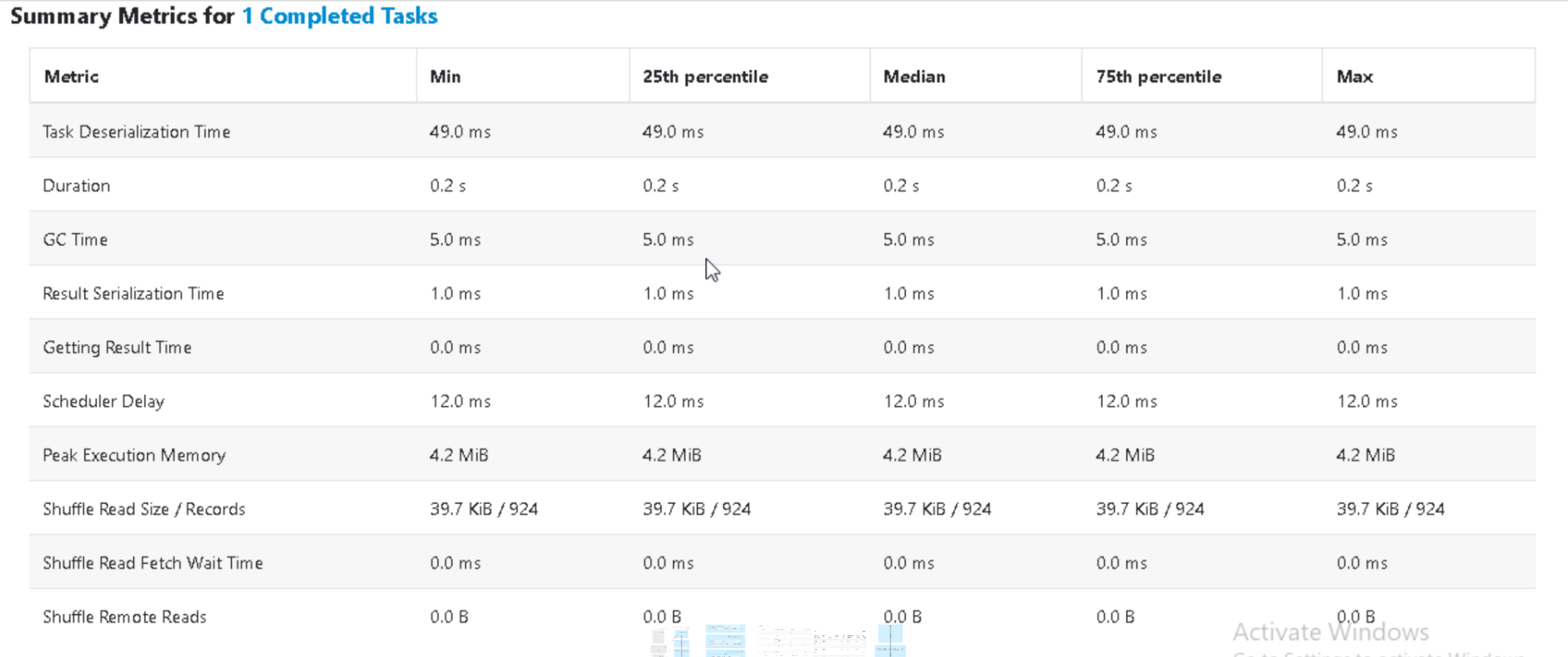
Task Deserialization Time: Time taken to deserialize a task before execution.

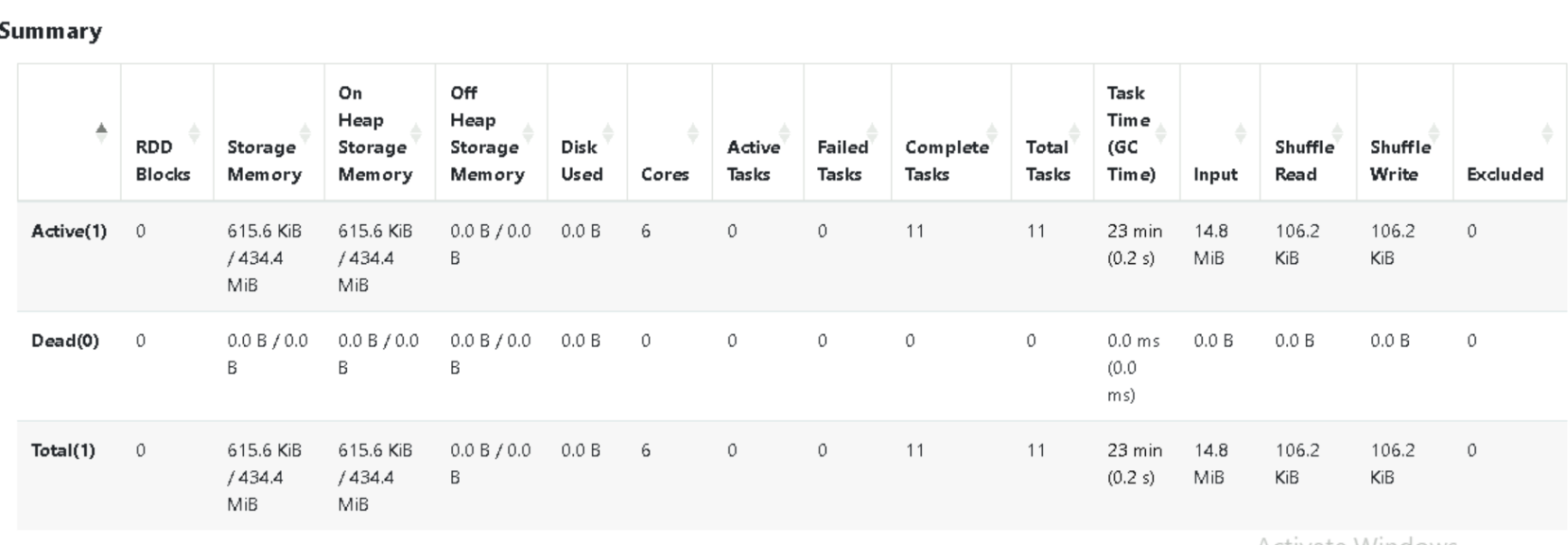
**Q2)** What are the most frequent transaction categories associated with fraudulent transactions?





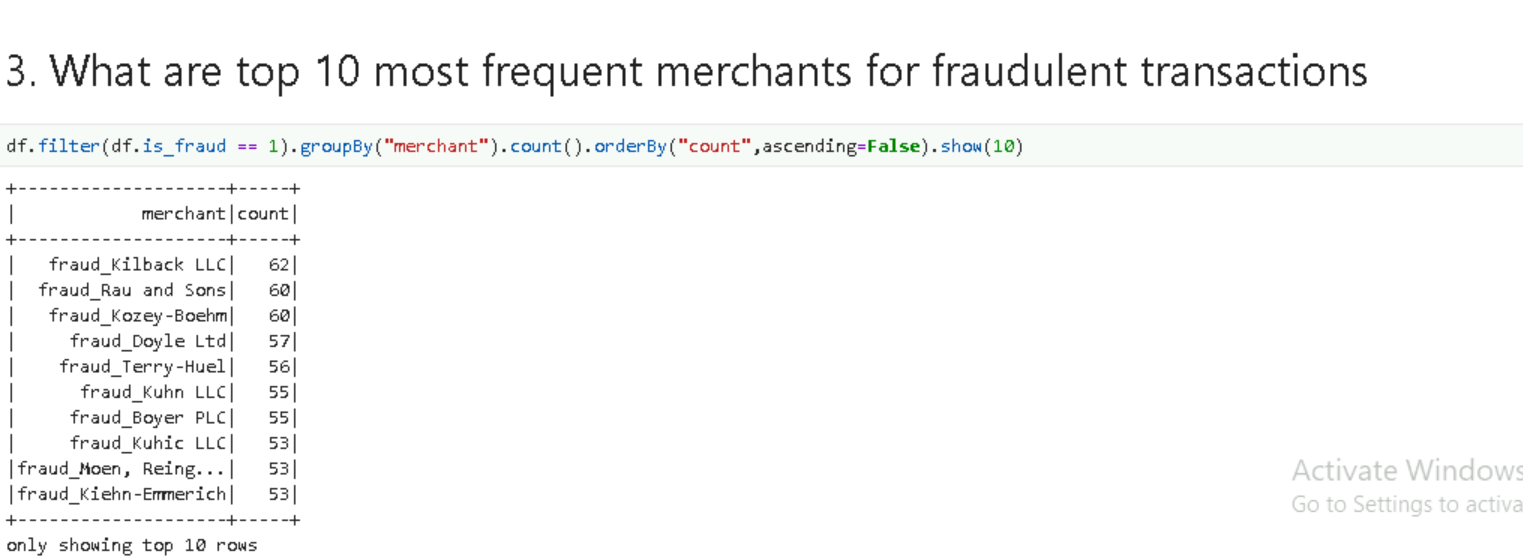


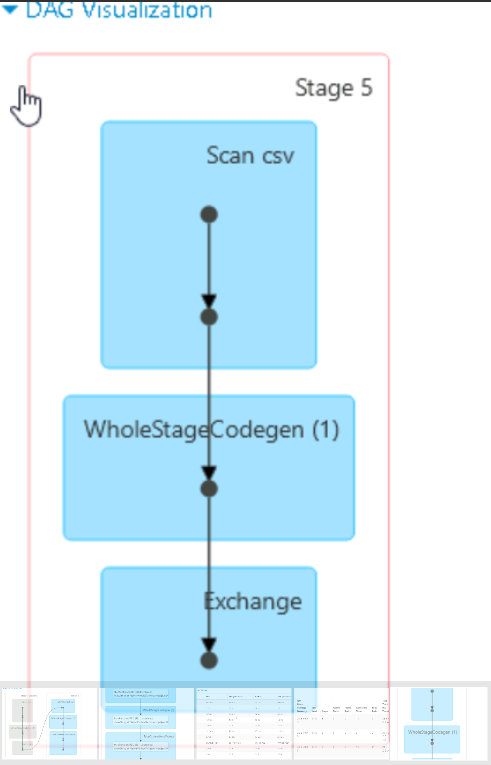




In Spark UI, a skipped stage refers to a stage that Spark does not execute because its output is already available.

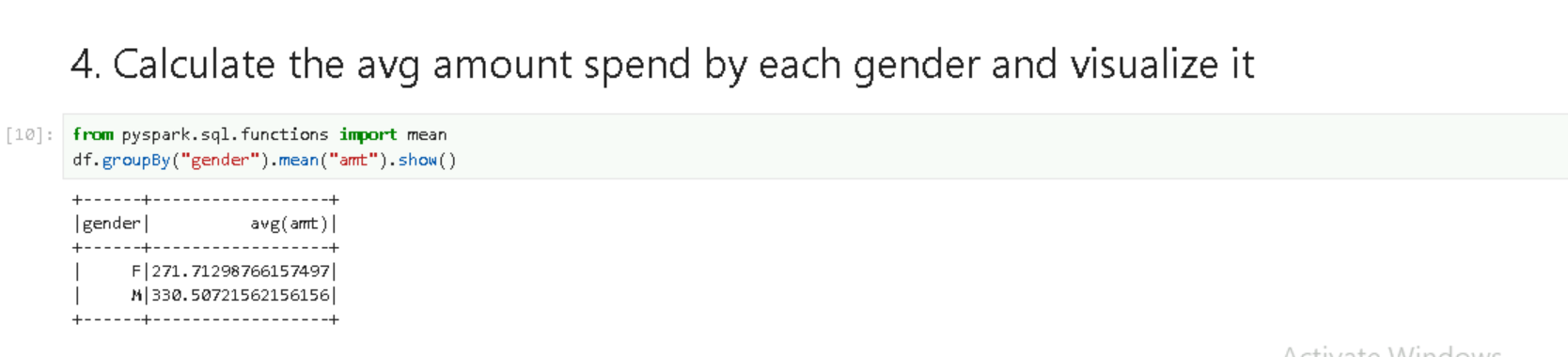
**Q3**) What are top 10 most frequent merchants for fraudulent transactions?

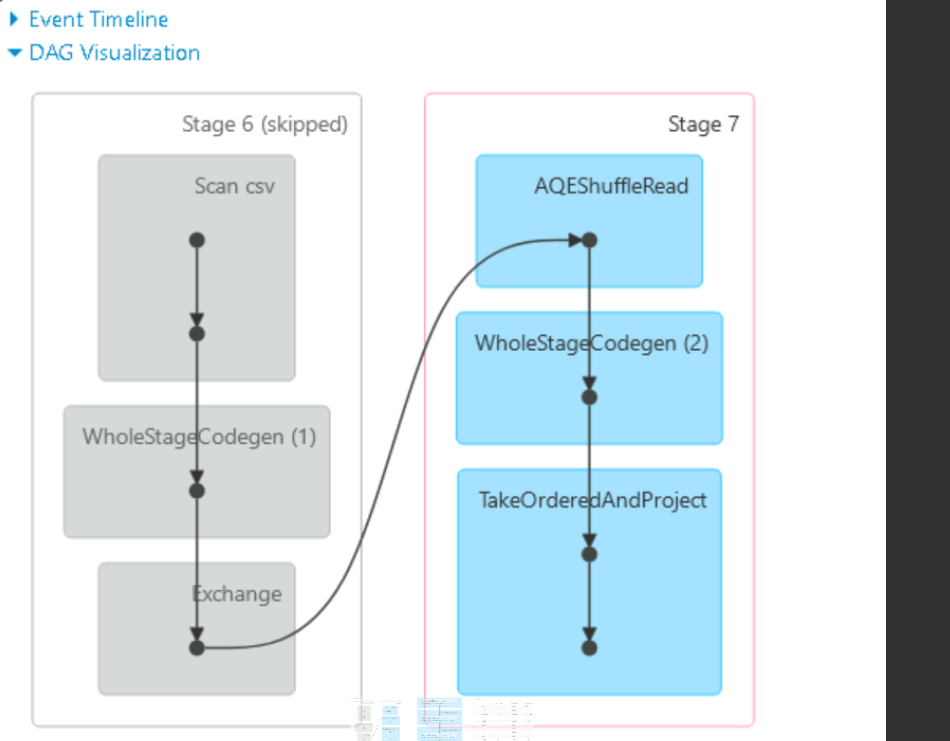


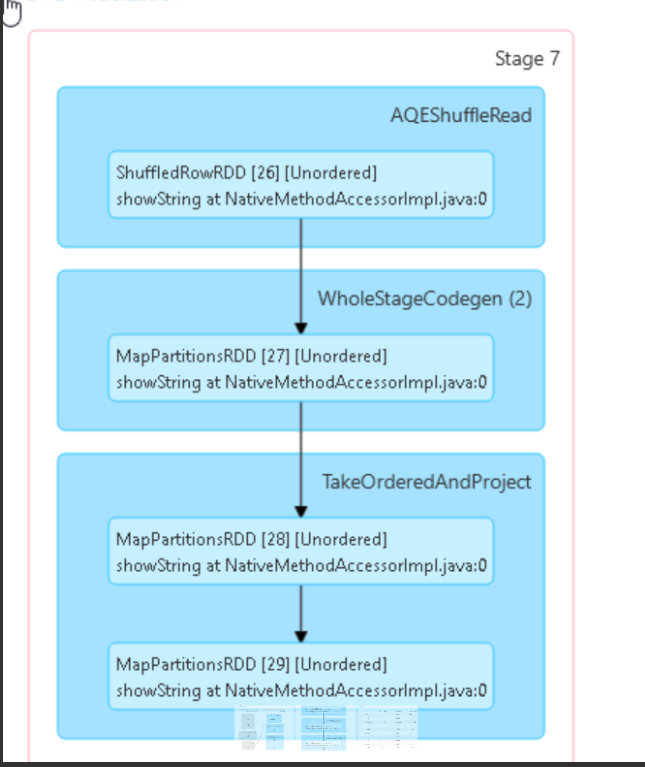


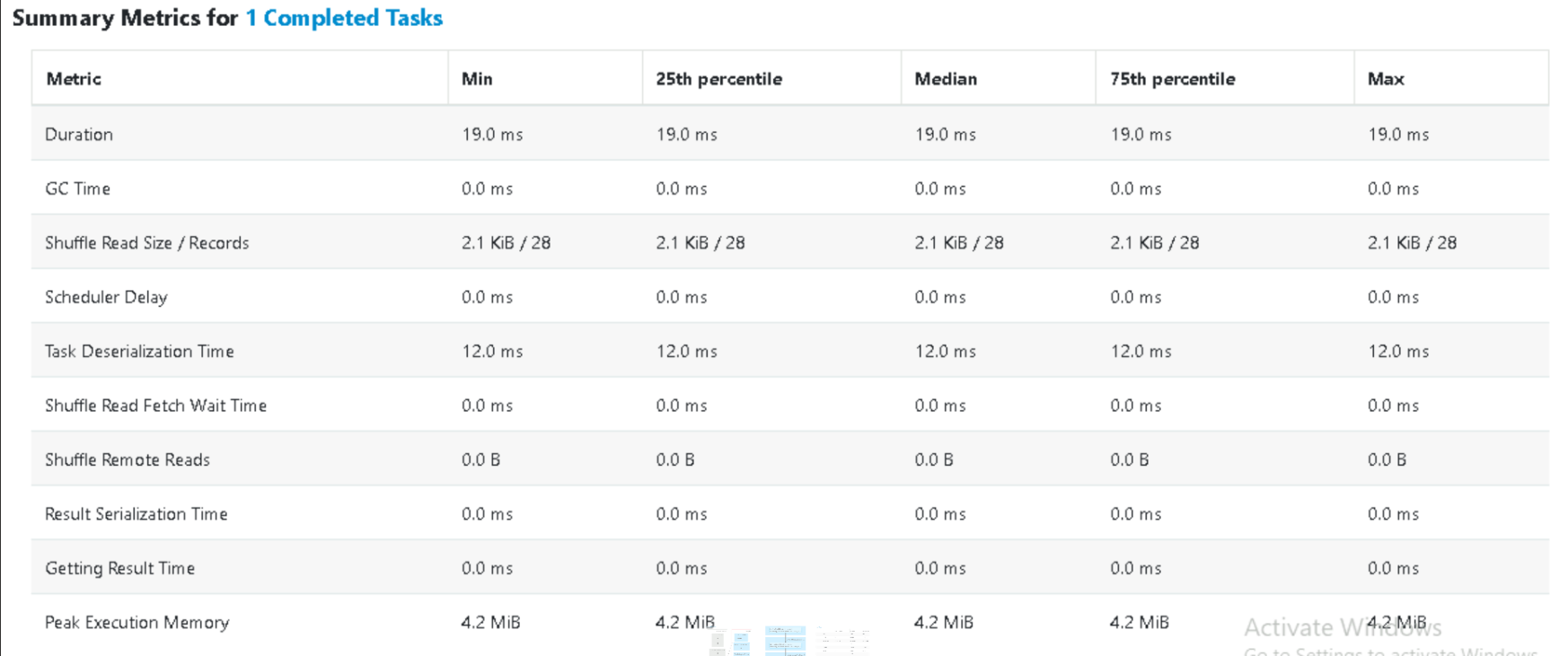


**Q4)** Calculate the avg amount by each gender and visualize it

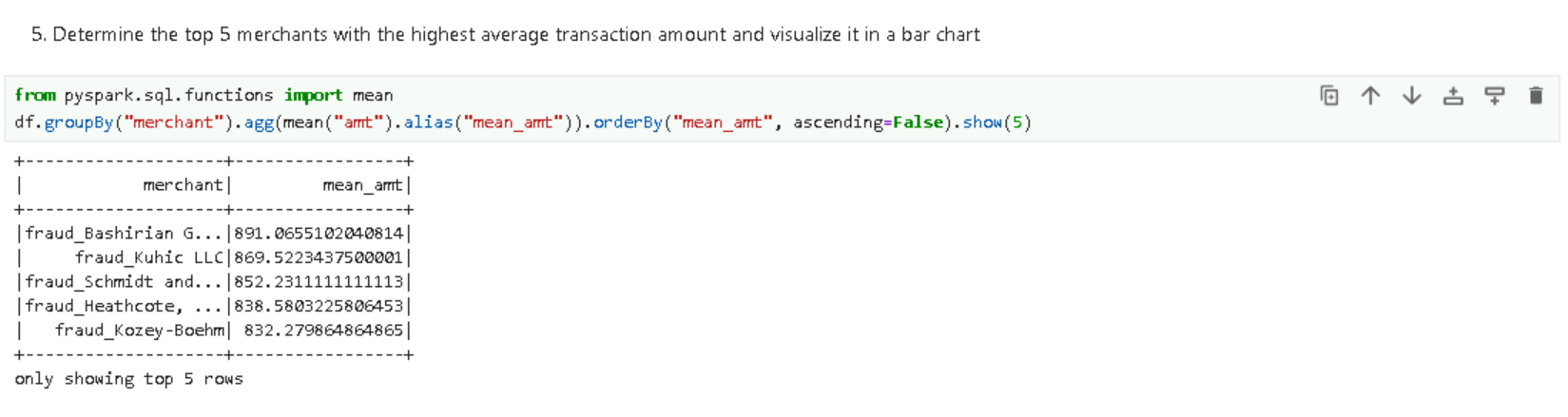


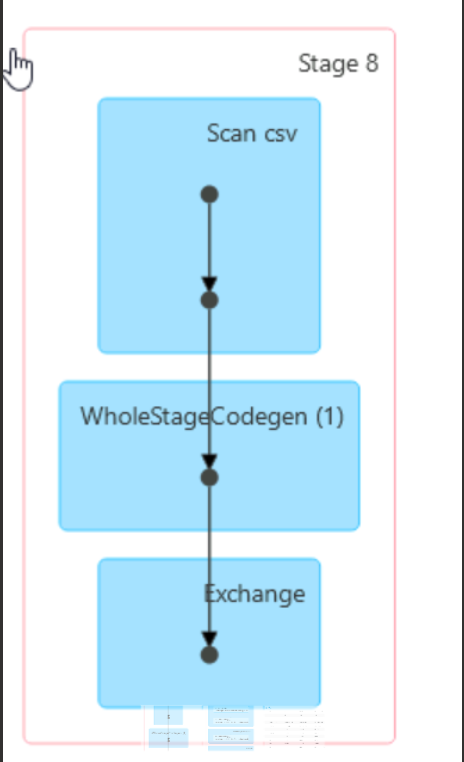


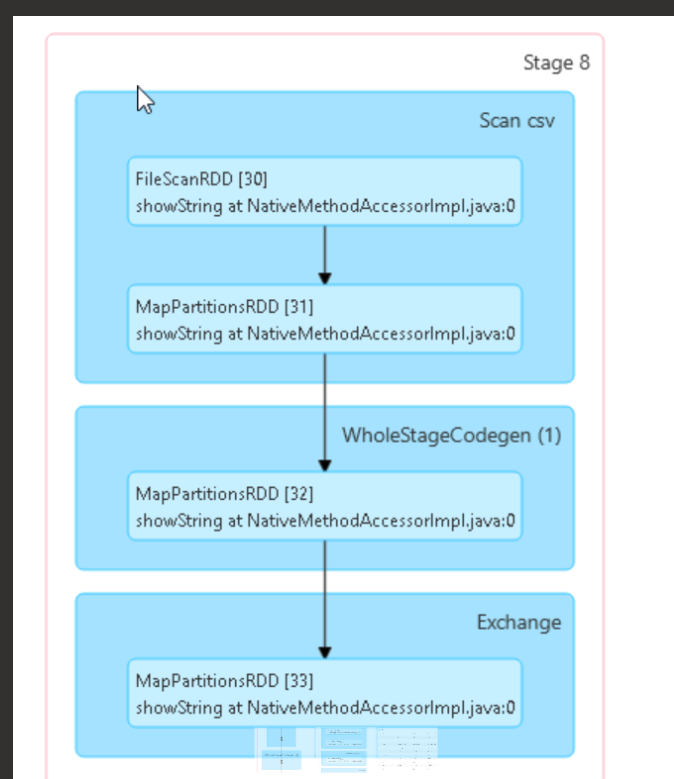


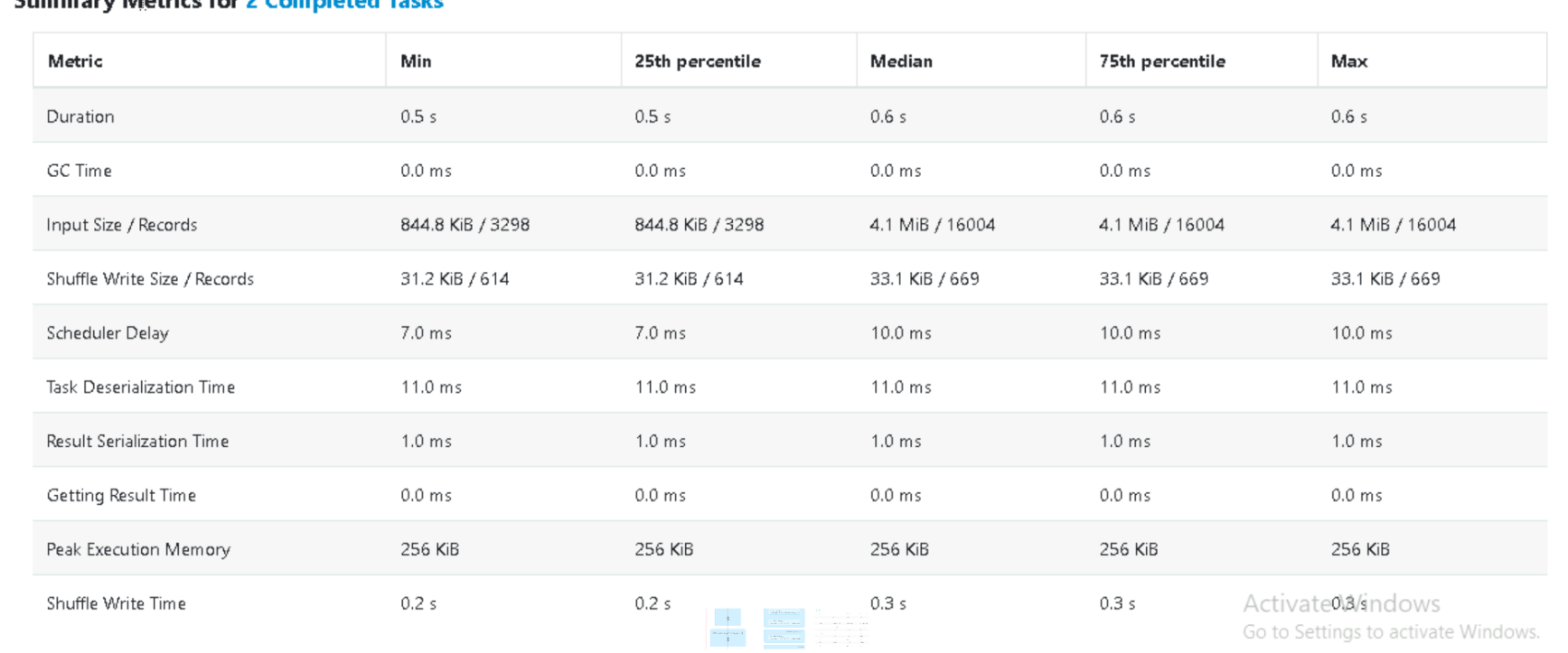


**Q5)** Determine the top 5 merchants with the highest average transaction amount and visualize it in a bar chart.

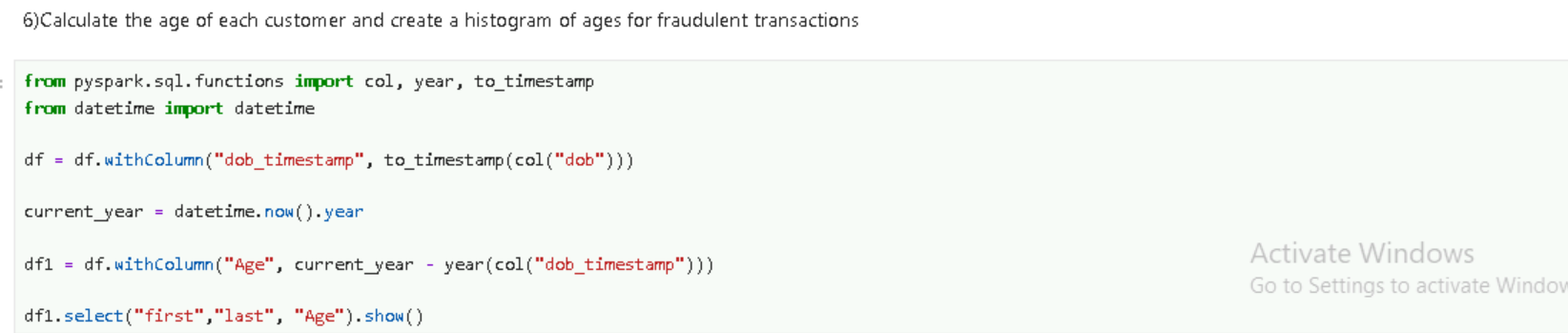
****

****

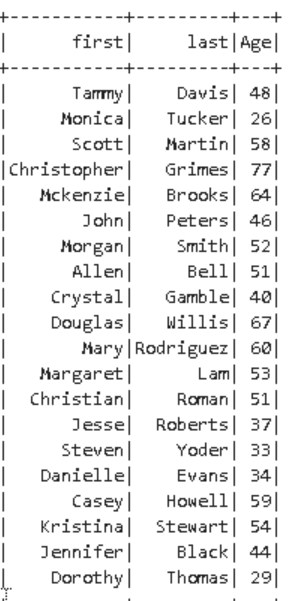
****

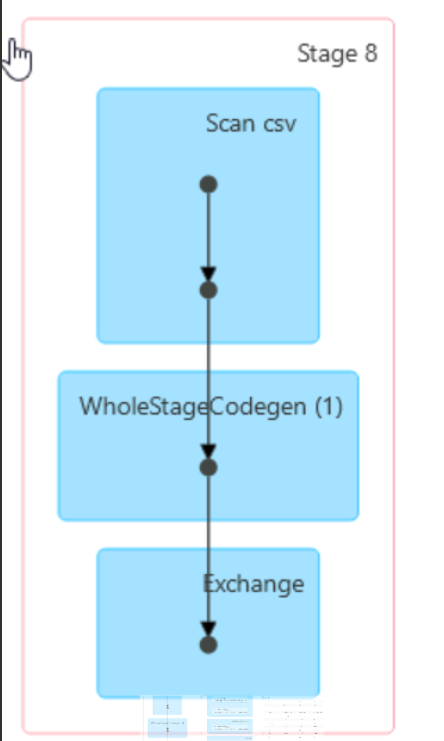
****

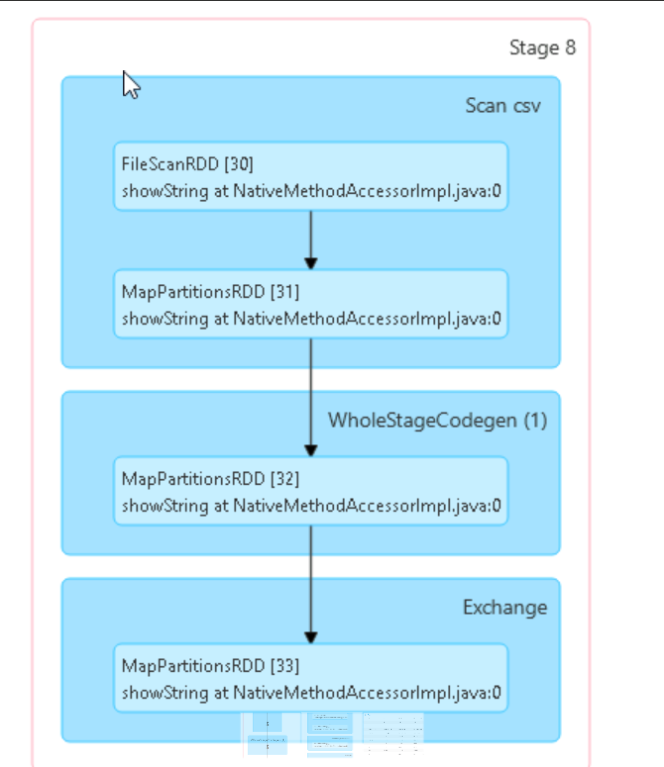
**Q6)** Calculate the age of each customer and create a histogram of ages for fraudulent transactions.

****

**OUTPUT**

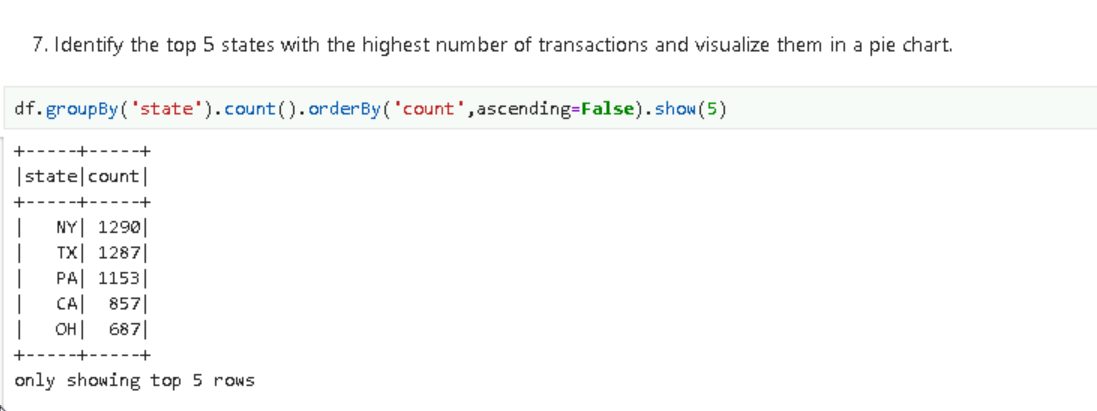
****

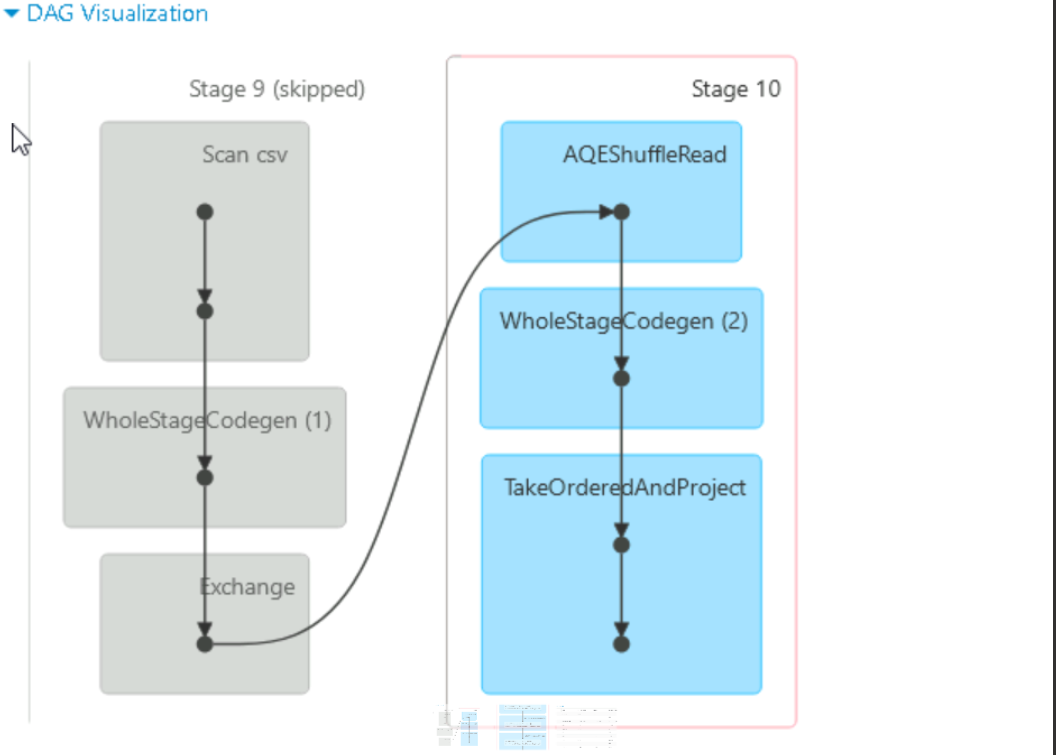
****

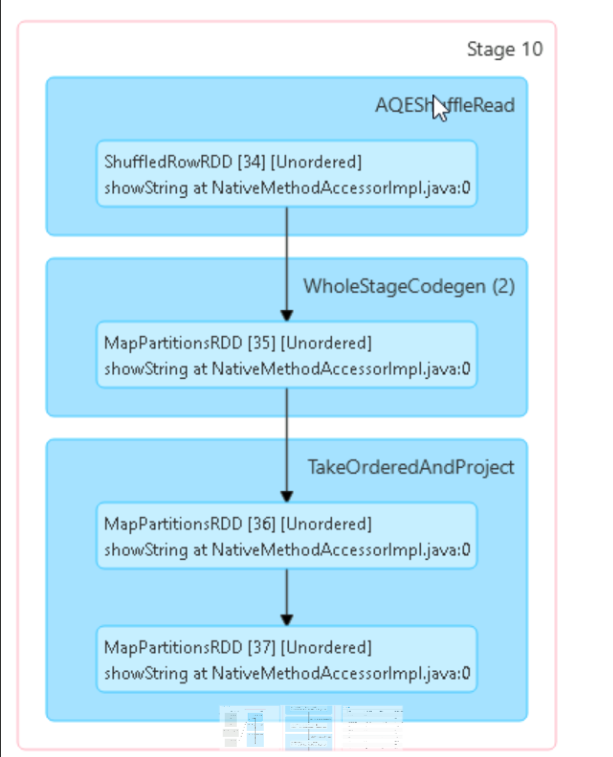
****

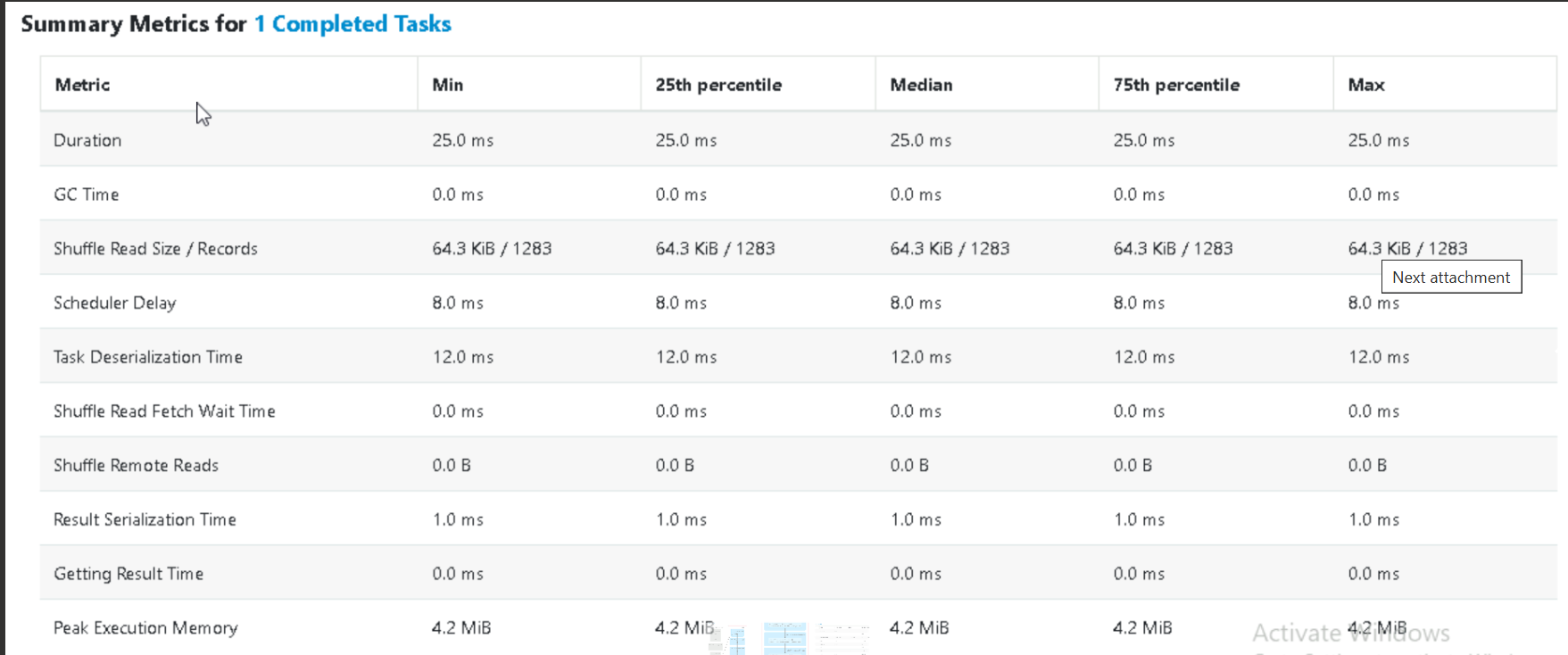
****

**Q7)** Identify the top 5 states with the highest number of transactions and visualize them in a pie chart

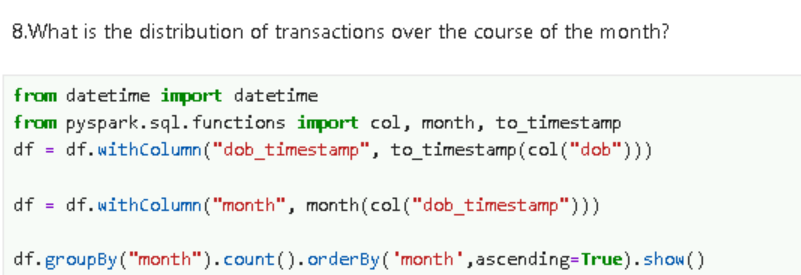
****

****

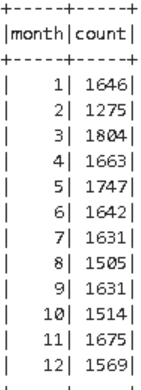
****

****

**Q8)** What is the distribution of transactions over the course of the month?

****

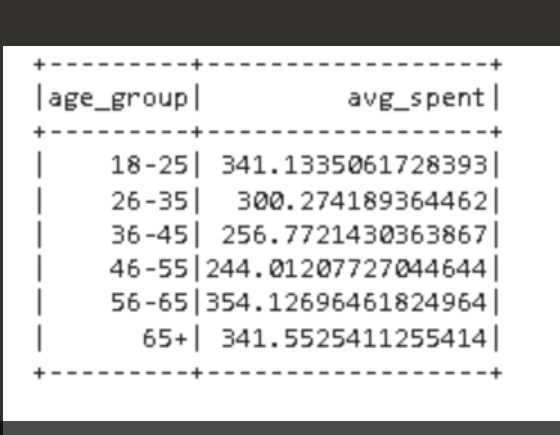
OUTPUT



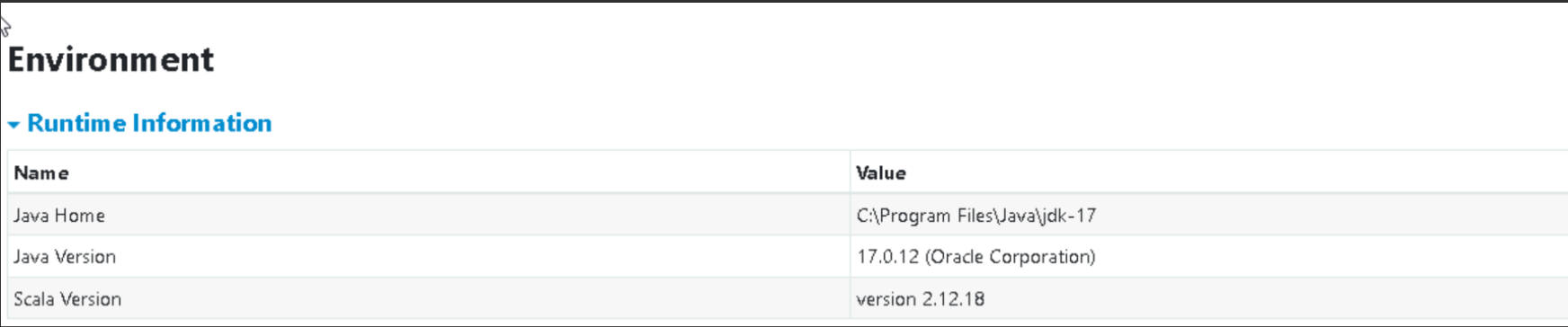
**Q10**) Analyse spending pattern based on age groups and avg spending amount for each group

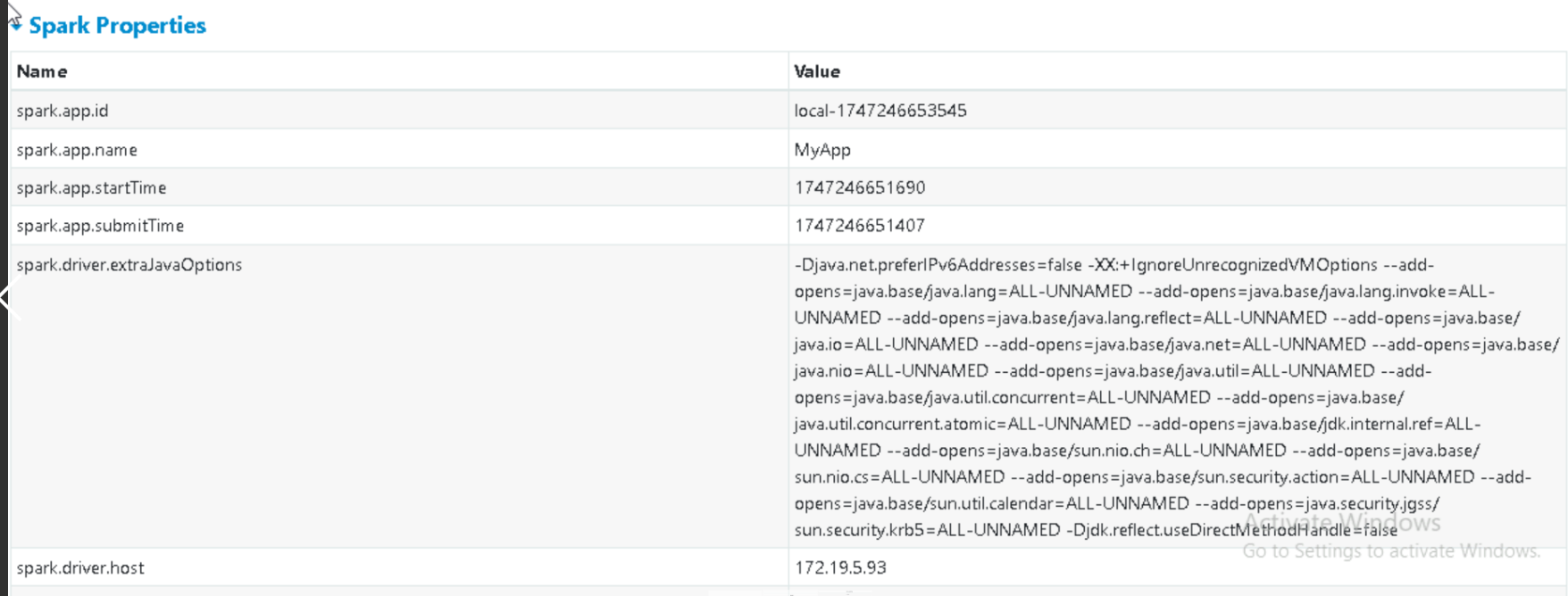


OUTPUT



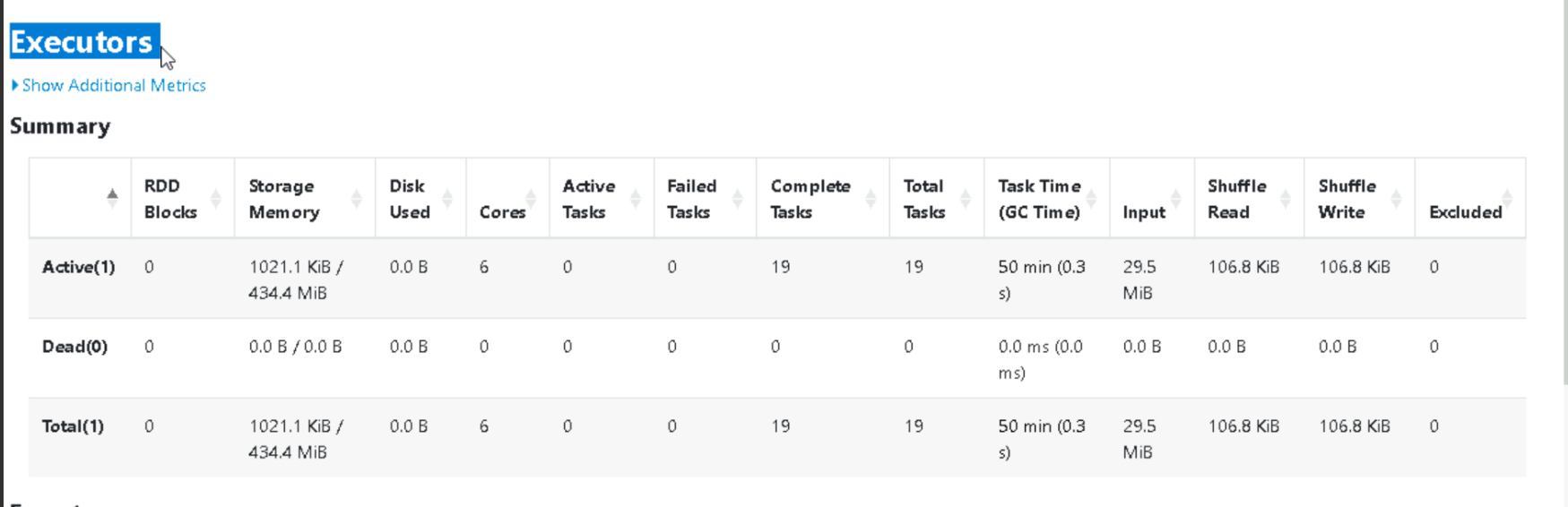
EXPLORATIONS

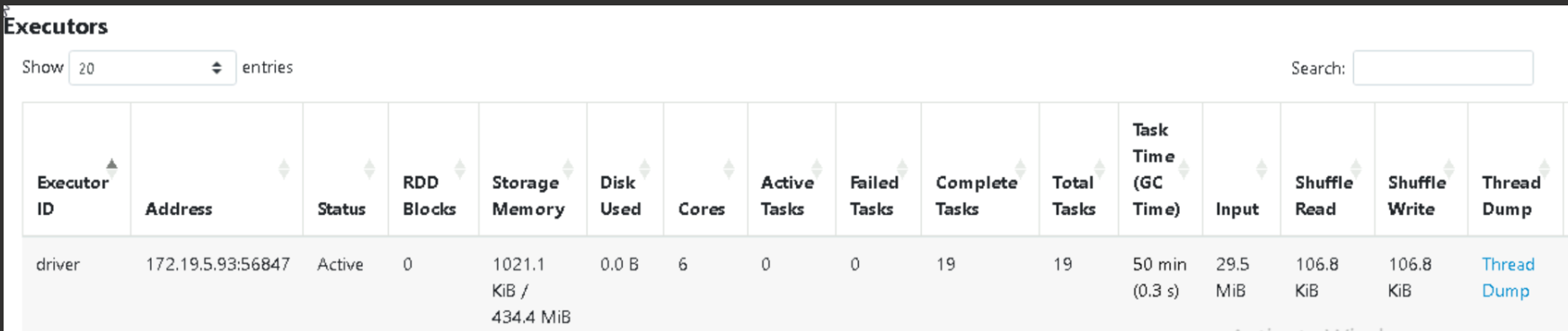




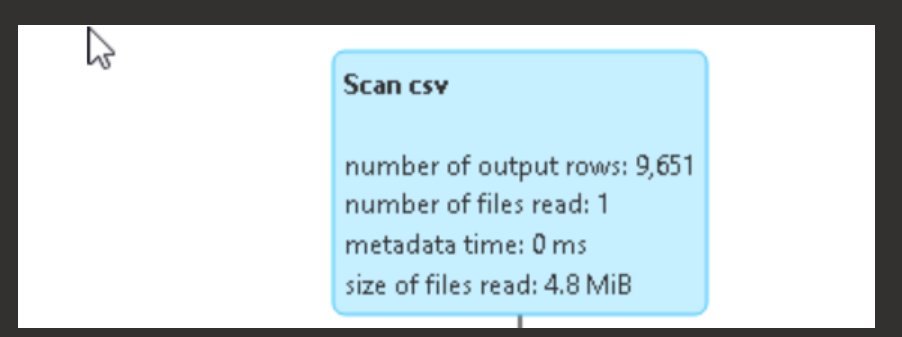


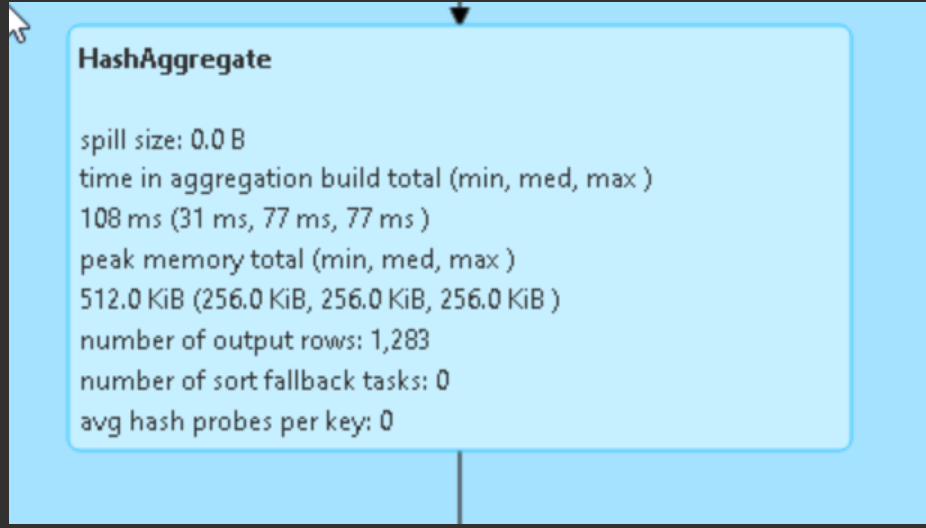
Executors





HASH AGGREGATE AND CSV





Scan csv scans the csv files processing of reading and analysing csv data using the sparks built in capabilities

Hash aggregation is physical execution strategy used for performing aggregation operations efficiently. Commonly seen in query plans when spark processes grouping and aggregation function like GROUP BY, COUNT, SUM, etc.