**Assignment 3**

**Spark stand alone analytics.**

DINH HOANG VIET PHUONG – 301123263

1. Load two days of data

A screenshot of a computer

Description automatically generated

1. Check the UI at the local host port 4040 or the port that spark connects to when launched.

A screenshot of a computer

Description automatically generated

Time taken to load the data: 2.6232454776763916 seconds.

The data was successfully loaded in approximately 2.623 seconds and the resulting table displays transactions that originated from outside the United Kingdom.

A screenshot of a computer

Description automatically generated

The jobs seem to be lightweight since they are completed in milliseconds.

All tasks in the listed jobs have succeeded, indicating that there were no failures or issues during their execution.

To sum it up, the Spark UI provides an overview of the jobs executed in the Spark application. From the given screenshot, it appears that several lightweight operations or transformations were performed on the data in quick succession, all of which were completed successfully. The exact nature of these operations would be clearer with additional context from the code or application log.

1. Carry out some basic investigation: count the number of records, print the inferred schema. Record  
   the results in your analysis report.

A screenshot of a computer

Description automatically generated

1. Show all the transactions that are related to the purchase of stock id that starts with "227" with the  
   type of product “ALARM CLOCK” mentioned as part of the description or a unit price greater than 5.

A screenshot of a computer

Description automatically generated

1. Store the results into a new dataframe name it df2\_firstname

A screenshot of a computer

Description automatically generated

1. Show the sum of the quantities ordered and the minimum quantity order and the maximum quantity  
   order for the transactions you extracted in point 4 above. Investigate the UI take a screenshot of the  
   DAG plan and in your” Analysis report” add the number of stages the job required with the total time  
   required per stage in addition to the number of tasks required for each job. Finally, drill down on each  
   stage and produce the DAG graph for each stage and analyze the statistics, note the shuffle size and the number of partitions in your report.

A screenshot of a computer

Description automatically generated

Number of Stages and Time Required Per Stage:

From the "Completed Jobs" section of the Spark UI:

Job ID: 7

Stages: 1

Duration: 6 ms

Tasks: 81

A white text with black text

Description automatically generated with medium confidence

1. Show all the transactions mentioned in point 4 above that have originated form outside the United  
   Kingdom.

A screenshot of a computer

Description automatically generated