Matheus Vinicius Ferreira Figueiredo Teixeira

301236904

# Exercise 1

## Check the UI at the local host port 4040 or the port that spark connects to when launched and record the following in your analysis report:

### The time it took to load the data.

The data loading process was broken down into 2 jobs. Job 0 took 2.0 seconds to be executed. Job 1 took 0.8 seconds to be executed. Therefore, the total time to load the data was 2.8 seconds.

### The number of tasks and try to explain in your own words what happened in the analysis report.

Again, the loading data process was broken into 2 jobs. Job 0 has only one stage, with only one task; this job is scanning the files. Job 2, also has only one stage, but this stage has 2 tasks, meaning that the data was split in 2 partitions; this job is where the data is being deserialized and that is when the schema is being inferred.

### Take a screenshot of the DAG execution and add it to your analysis report.

Figure 1  
*DAG of job 0*

A screenshot of a computer

Description automatically generated

Figure 2  
*DAG of stage 0*

A screenshot of a computer

Description automatically generated

Figure 3  
*DAG for Job 1*

A screenshot of a computer

Description automatically generated

Figure 4  
*DAG for stage 1*

A screenshot of a computer

Description automatically generated

## Carry out some basic investigation

### count the number of records

**Number of records:** 5649

### print the inferred schema. Record the results in your analysis report.

root

|-- InvoiceNo: string (nullable = true)

|-- StockCode: string (nullable = true)

|-- Description: string (nullable = true)

|-- Quantity: integer (nullable = true)

|-- InvoiceDate: string (nullable = true)

|-- UnitPrice: double (nullable = true)

|-- CustomerID: double (nullable = true)

|-- Country: string (nullable = true)

## Investigate the UI take a screenshot of the DAG plan

Figure 5  
*DAG for Job 6*

A screenshot of a computer

Description automatically generated

### 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.

The job required 2 stages

|  |  |  |
| --- | --- | --- |
| Stage | Time | Number of tasks |
| 8 | 200 ms | 2 |
| 9 | 13 ms | 1 |

### Finally, drill down on each stage and produce the DAG graph for each stage

Figure 6  
*DAG for stage 8*

A screenshot of a computer

Description automatically generated

Figure 7  
DAG for stage 9

A screenshot of a computer

Description automatically generated

### and analyze the statistics, note the shuffle size and the number of partitions in your report.

We know that Stage 8 executed 2 tasks, meaning the data was read from 2 partitions, and From the Stage 8 DAG figure, we can see it read the data and processed it leading to a shuffle with size of 141 B across 2 partitions. From the Stage 9 DAG figure, we can see again that the shuffle size was 141 B across 2 partitions; Stage 9, is the exchange operation; it executed 1 task meaning that the data was aggregated in 1 partition.