Data Science Infrastructures – Exercise 06

**Assignment I – Creating a Spark Cluster**

Answer: After exporting the modulepaht and loading the spark 3.4.0, I started the interactive shell with one cpu. The final version of my terminal with the beginning of the Scala expressions within the cell is shown below.

A screenshot of a computer screen

Description automatically generated

Figure 1 Creating a Spark Cluster

**Assignment II – Interactive Use of Spark**

Answer: This is how I uploaded my melville.txt file into the HPC system.

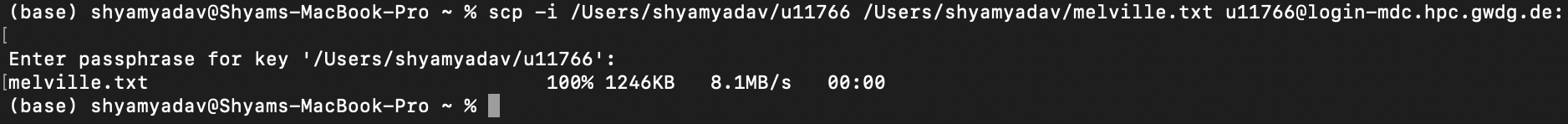
****

Figure 2 Uploading melville.txt into hpc

Then, I performed MapReduce on Spark Cluster as shown below.

A computer screen shot of a black screen

Description automatically generated

Figure 3 MapReduce on Spark

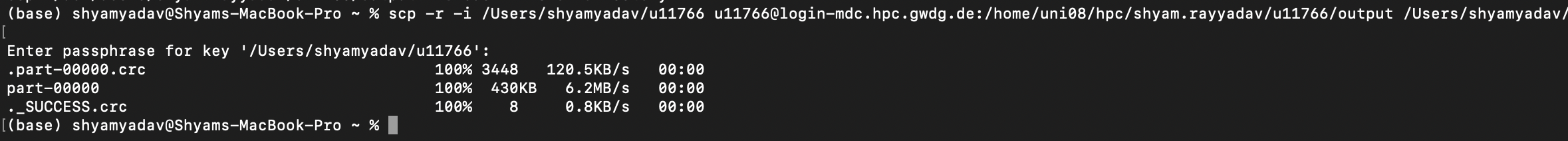
Now, the last portion of the output file “part-00000” is shown below.

A screen shot of a computer program

Description automatically generated

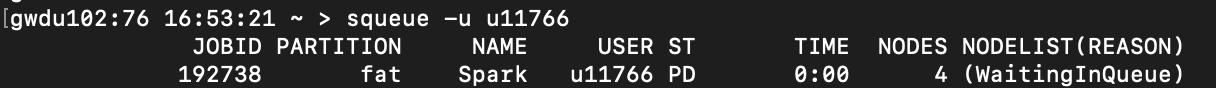
Figure 4 Last portion of the output file "part-00000"

Also, I tried to download my output file using the following command. The results are also shown below.



**Assignment III – Run a Spark Job**

Answer: At first the cluster with 4 workers was created. The batch job on HPC cluster, which contains Spark Cluster, in queue waiting is shown below.



After checking the staus of my squeue, we can see that my job lies in the last corner of the queue.

A screenshot of a computer program

Description automatically generated