**Spark Asgn Work**

**Student Name:** Shivani Sharma

**Course No :** ALY6110

**Title:** Data Management and Big Data

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Mr. Daya Rudhramoorthi

**Lab Work Completion Date:** 05-10-2020



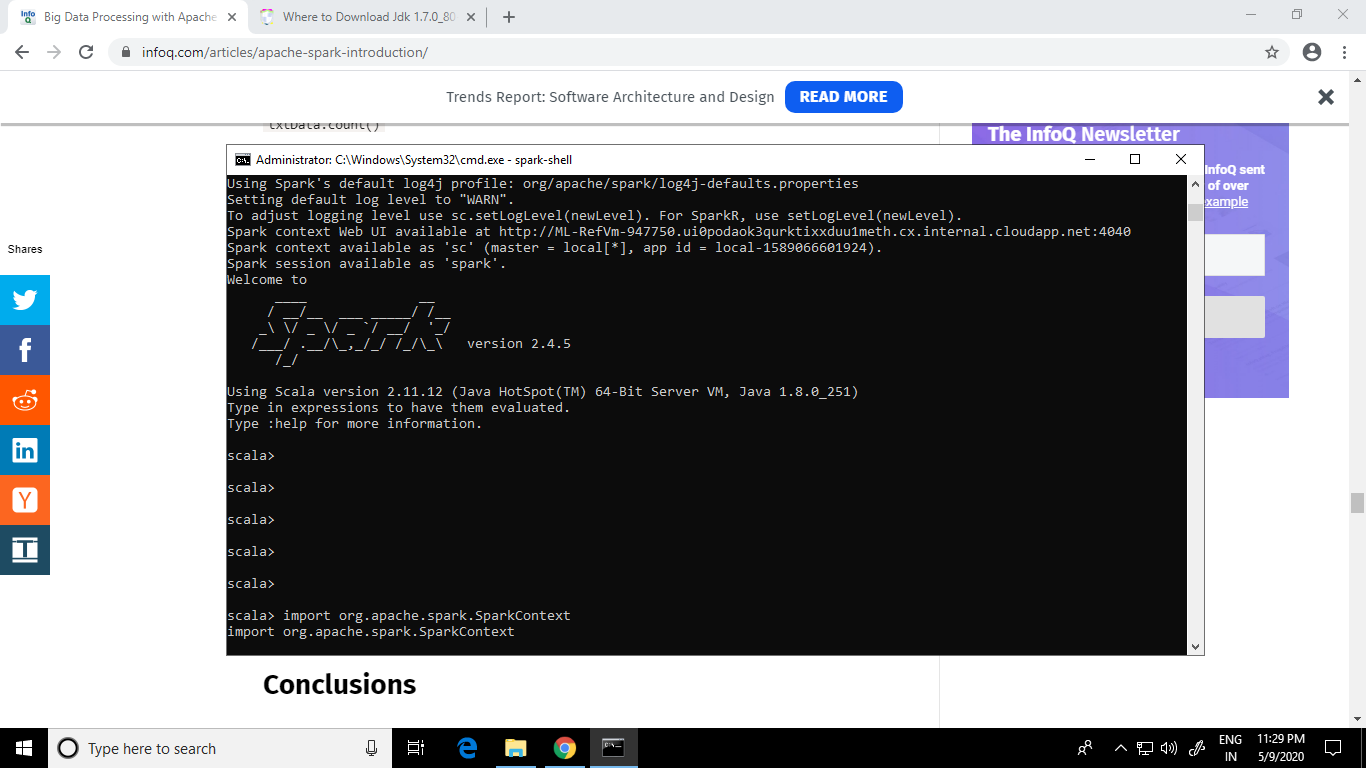
**Summary**

This lab assignment is based on one of the spark implementation of word count application, After installing jdk and spark on virtual machine scala command prompt opens up and then we need to execute a set of commands which I am going to explain in analysis section.

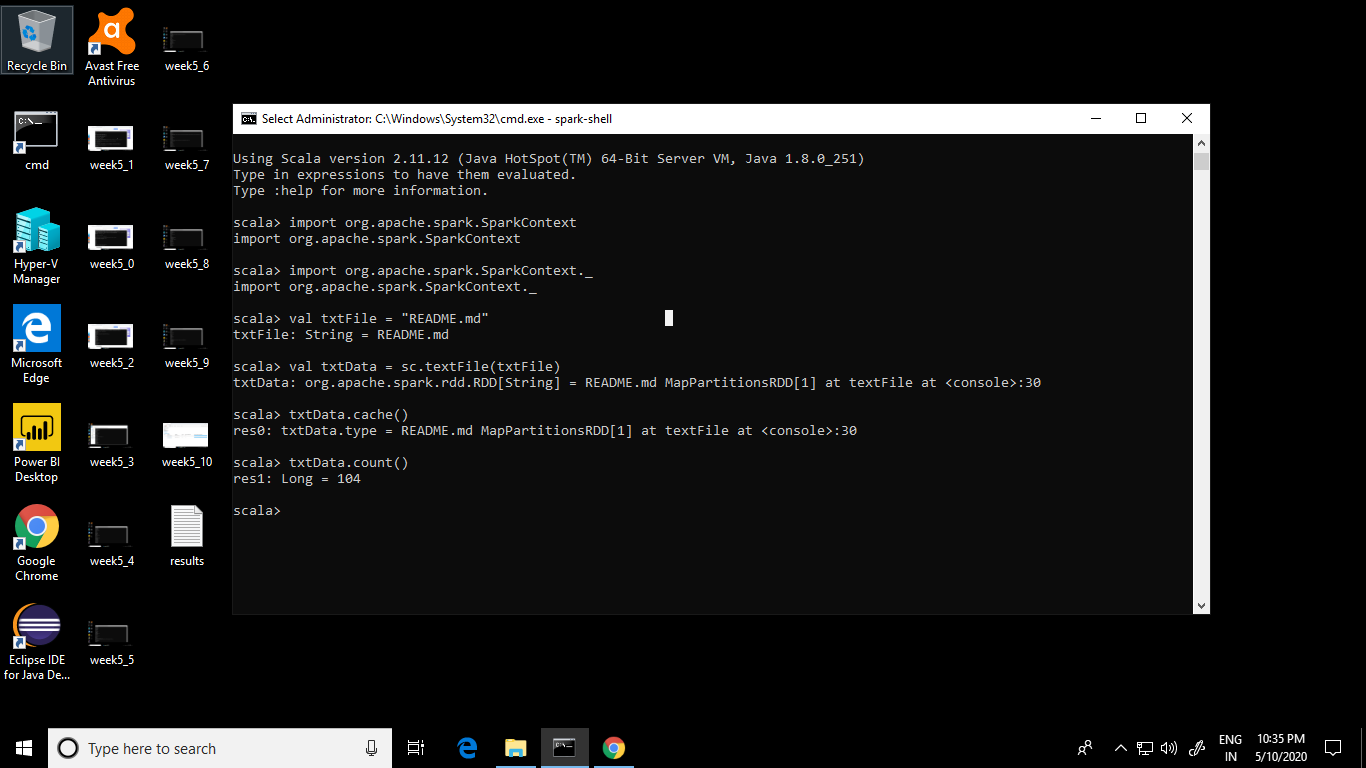
NOTE: This is to mention here that Hadoop is not required to be installed explicitly for this as it is inbuilt with the spark package. Just followed the instructions provided on the assignment reference material provided.

**Analysis**

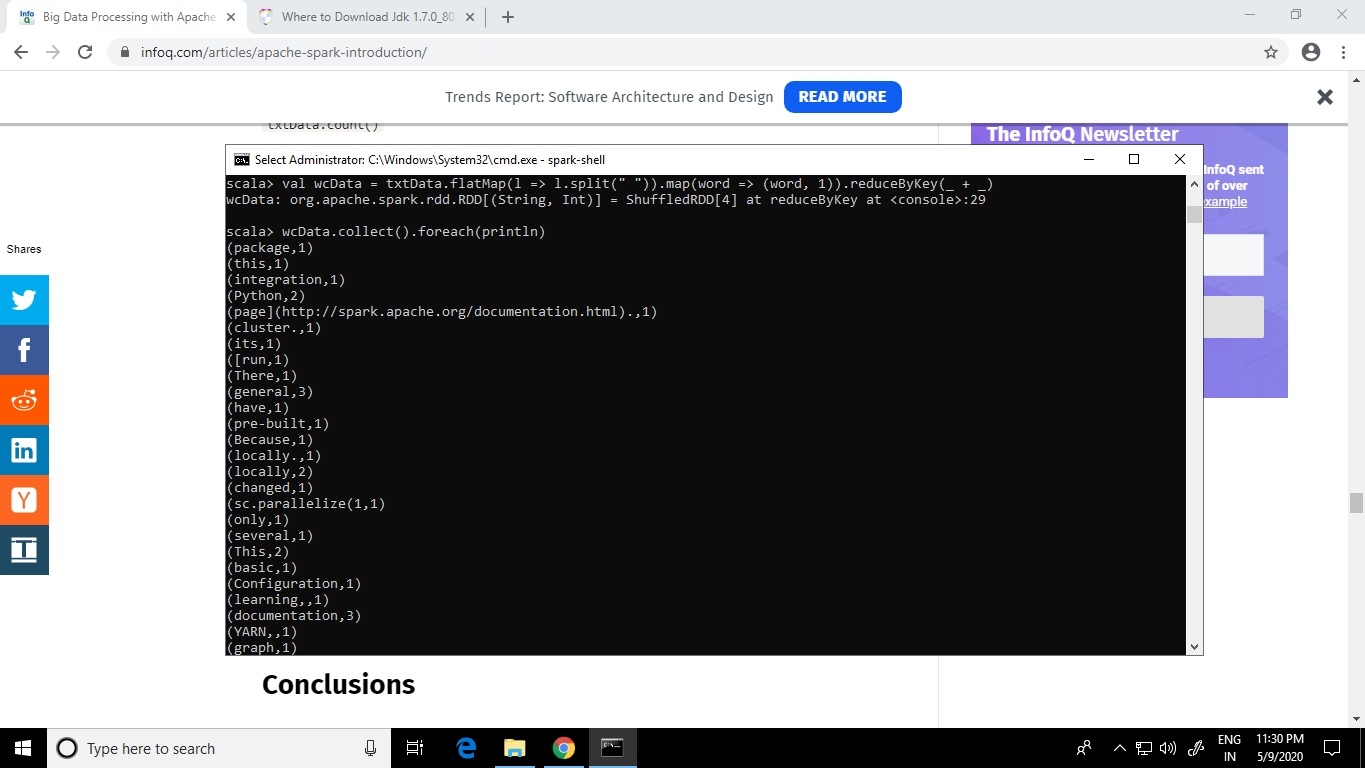
First step would be to execute the import statements for spark context and the logic for the same is as follows:

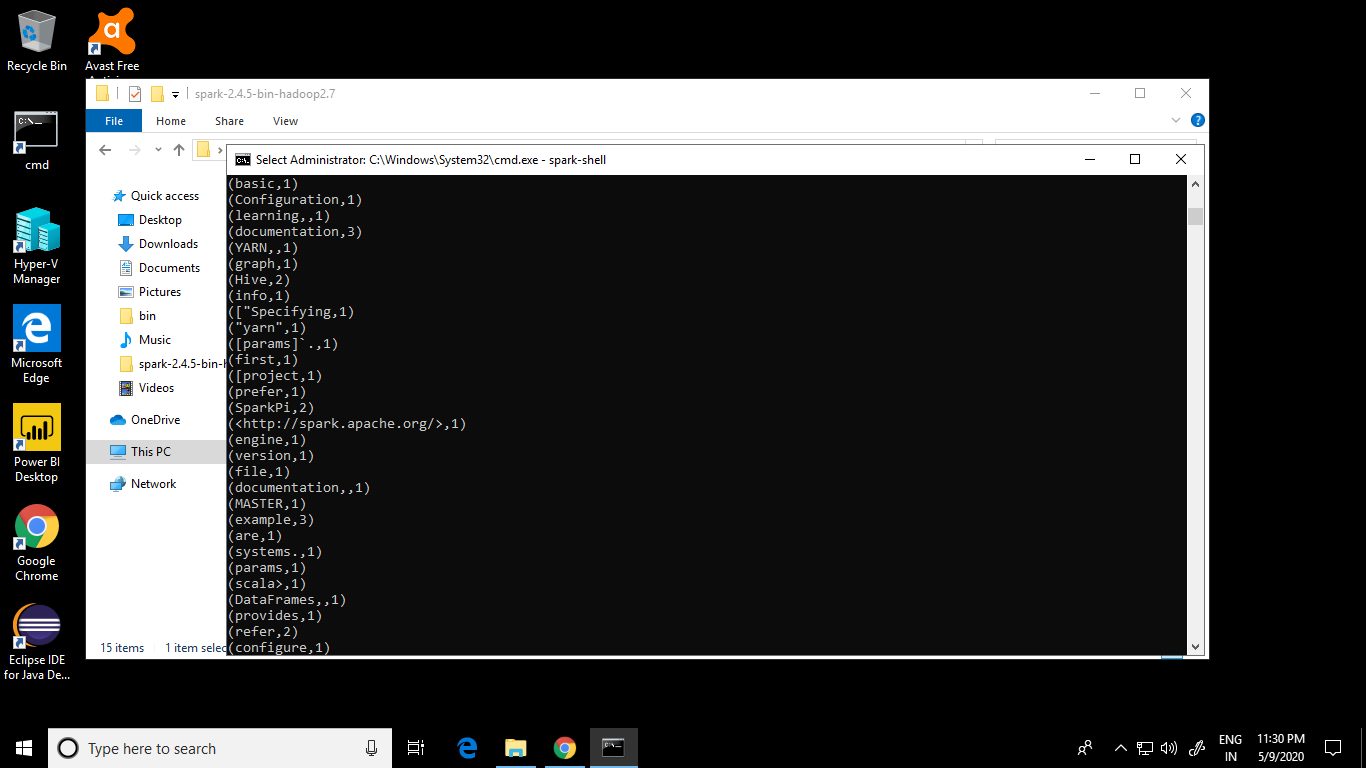


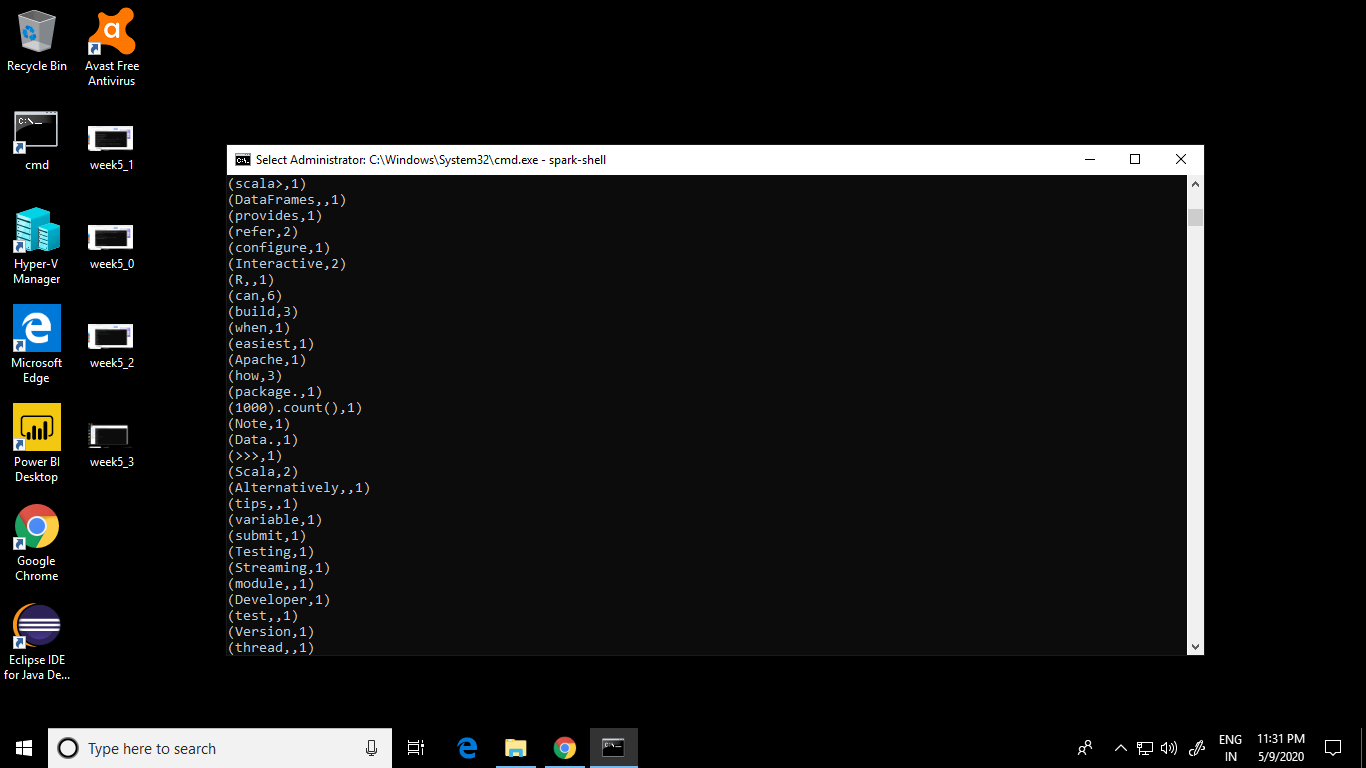
Then we need to import the file and logic is as follows:

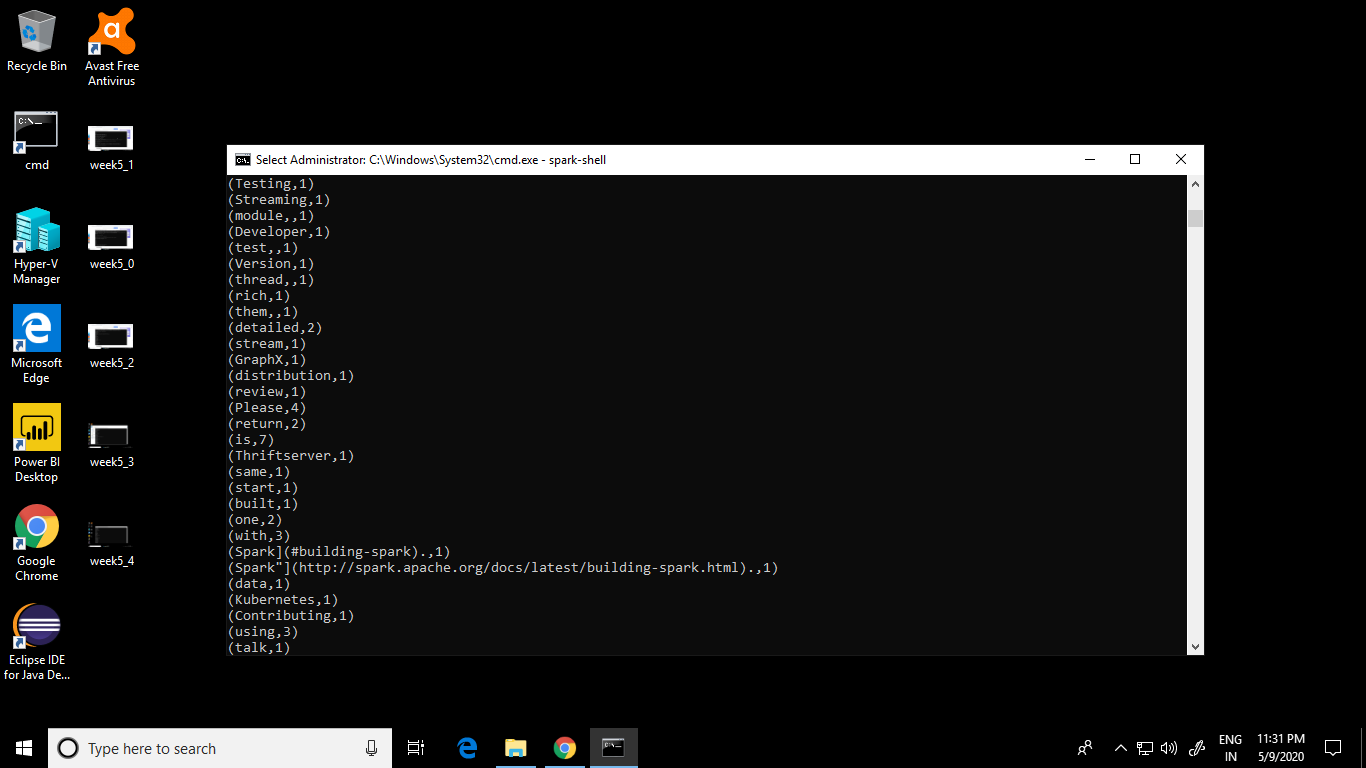


Following are the results for the same:

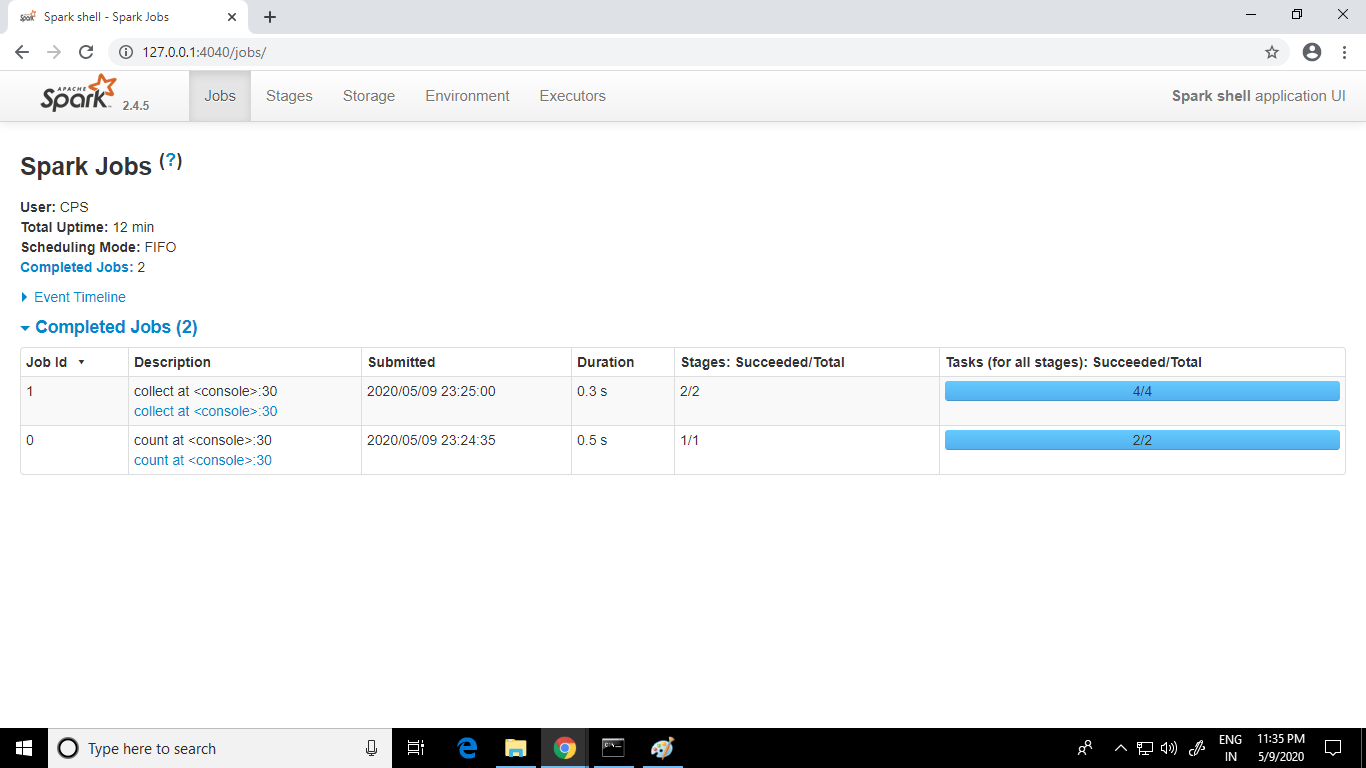








Following is the spark job details which is being executed at the backend for the word count program.



In this assignment we have been asked following questions:

1. Count of word “Hadoop” : “Hadoop” word is appearing at 2 places with count values as 2 and 3 with a total count of 5.
2. Most commonly used word is “to” which is appearing 17 times in the READMe file.
3. There are more than 1 word which are repeating with a count value of 1 like package, its, # etc which is the least count among other words.
4. From the web console logs of spark job we can see that the word count job completed in 0.5 seconds.

# References

1. Penchikala, S. (2015, January 30). Big Data Processing with Apache Spark – Part 1: Introduction. Retrieved from https://www.infoq.com/articles/apache-spark-introduction/