Lab Assignment-02

ALY 6110 Data Management & Big Data

**CHIRAG SANJAYBHAI SHAH**

Northeastern University



**Note**:

Chirag Sanjaybhai Shah College of Professional Studies, Northeastern University, Boston, MA 02115.

This report was created as a part of the project to gain hands-on experience in PySpark.

**Instructor Name: Janos Mako**

Chirag Sanjaybhai Shah is now a student at Department of Analytics, Northeastern University

Contact:

[shah.chir@husky.neu.edu](mailto:shah.chir@husky.neu.edu)

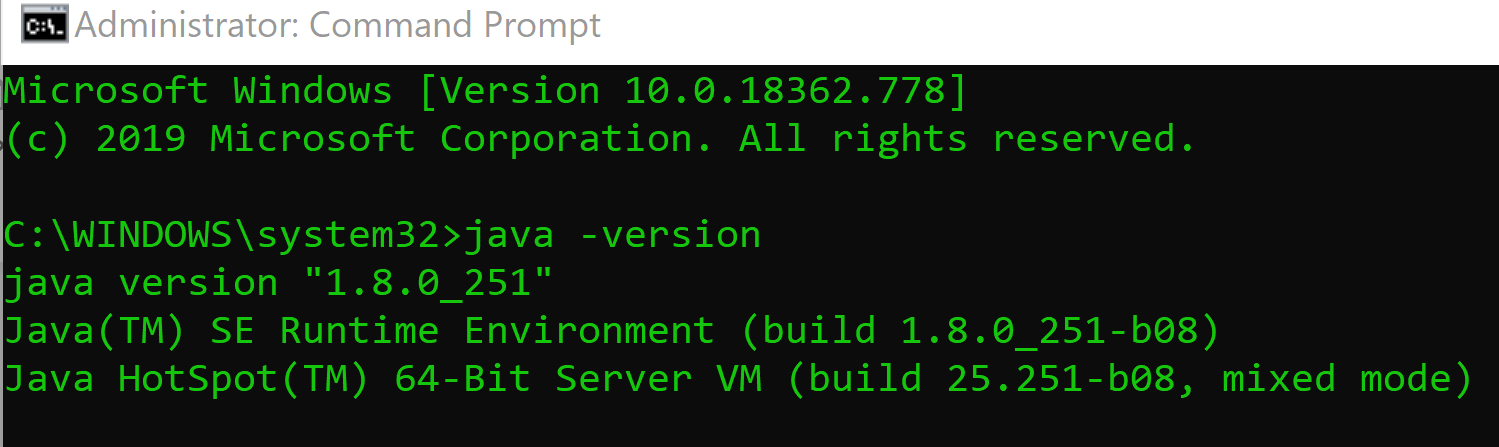
**INTRODUCTION**

We learnt pyspark in this week’s lab02 assignment. We first installed the required jdk version in C:/Java folder. The jdk 1.8 version was installed in the windows local machine. To install the spark, we downloaded the required .tar file from the spark website. We also set the required SPARK\_HOME, JAVA\_HOME environment variables in our windows machine. We then installed the jupyter notebook to run the pyspark version 2.4.5. In this assignment, we initialize the spark session and counted the words.

**ANALYSIS**

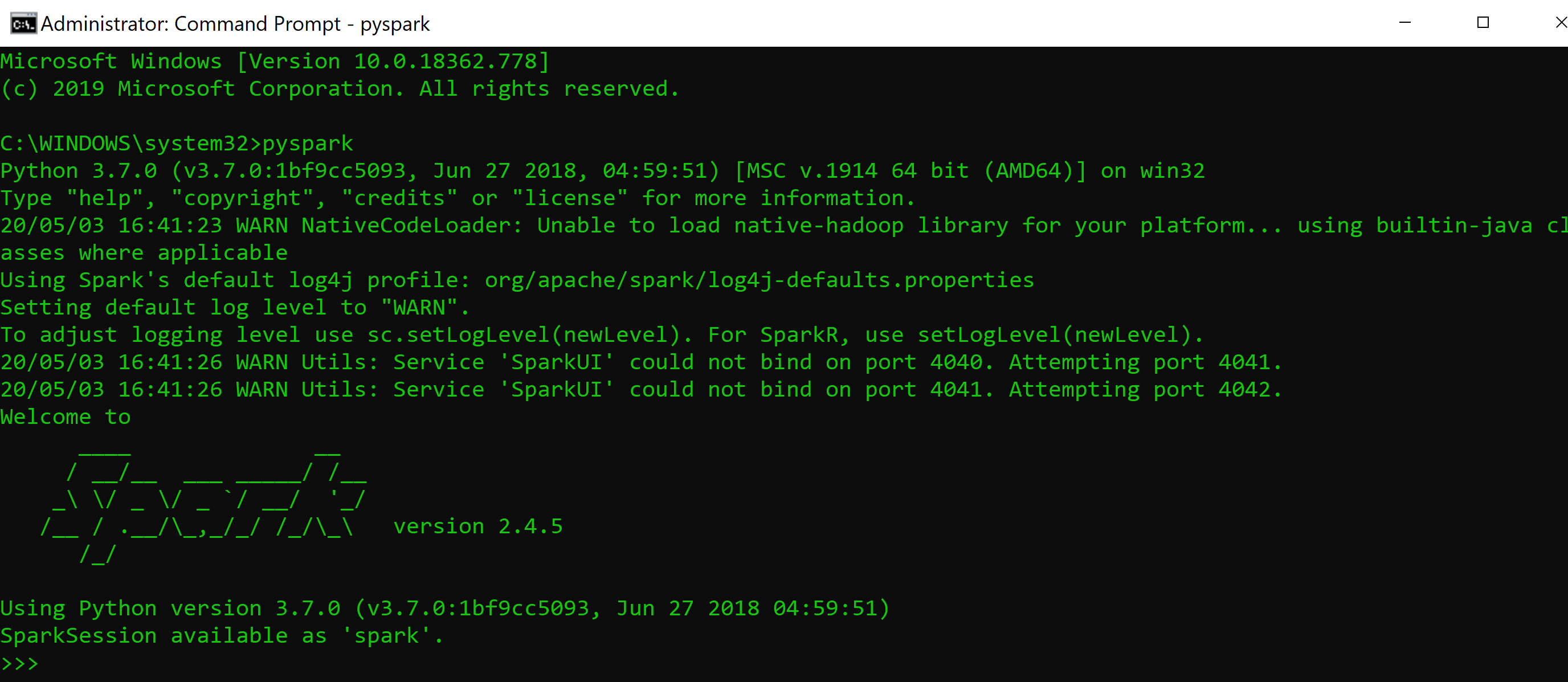
1. Installing Java and JDK

We install the JDK SE kit from the oracle website. The version is installed as below:



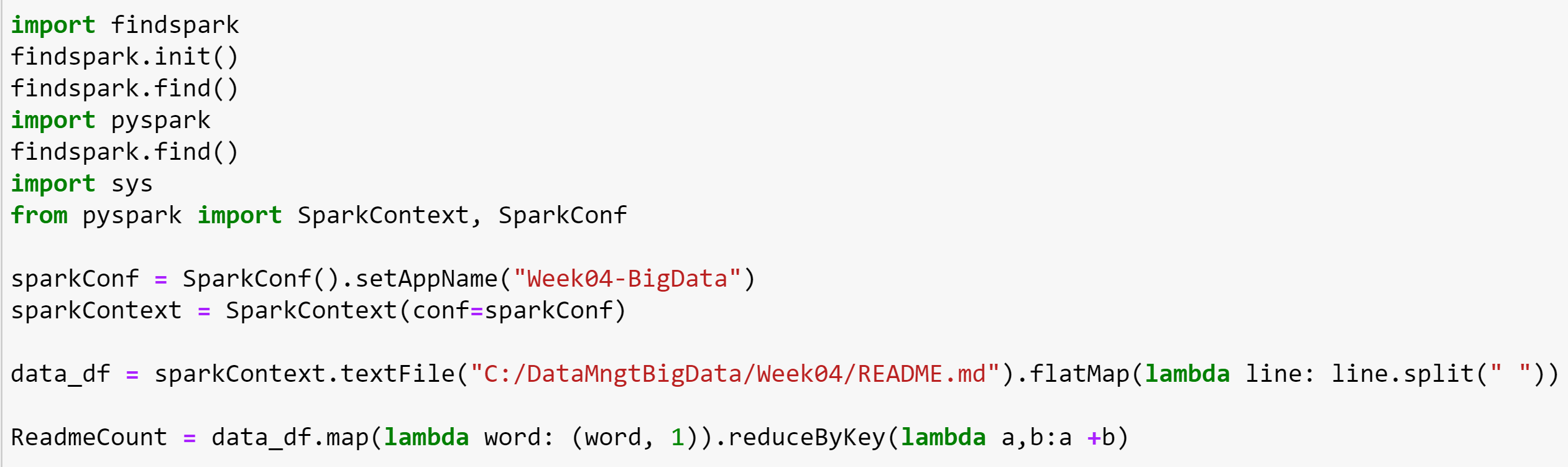
1. Installing and Loading Apache Spark

The Spark version 2.4.5 has been downloaded from the Apache Spark website and the README.MD file has been cached for the wordcount application. To install Spark in the Python environment. We then used pip install pyspark to install the pyspark in jupyter notebook.



1. Counting words in README.MD file

To accomplish this process, we initialise the sparkContext and sparkConf variables. We used the flatMap() to read the words:



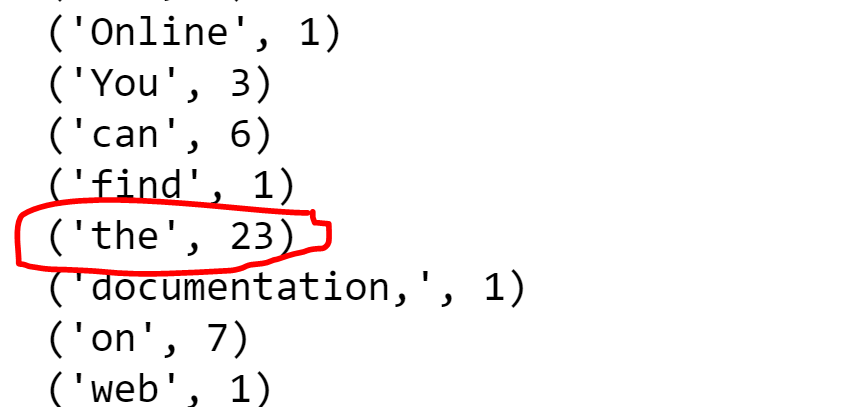
1. Count of word Hadoop in the file:

The count of Hadoop is 3.



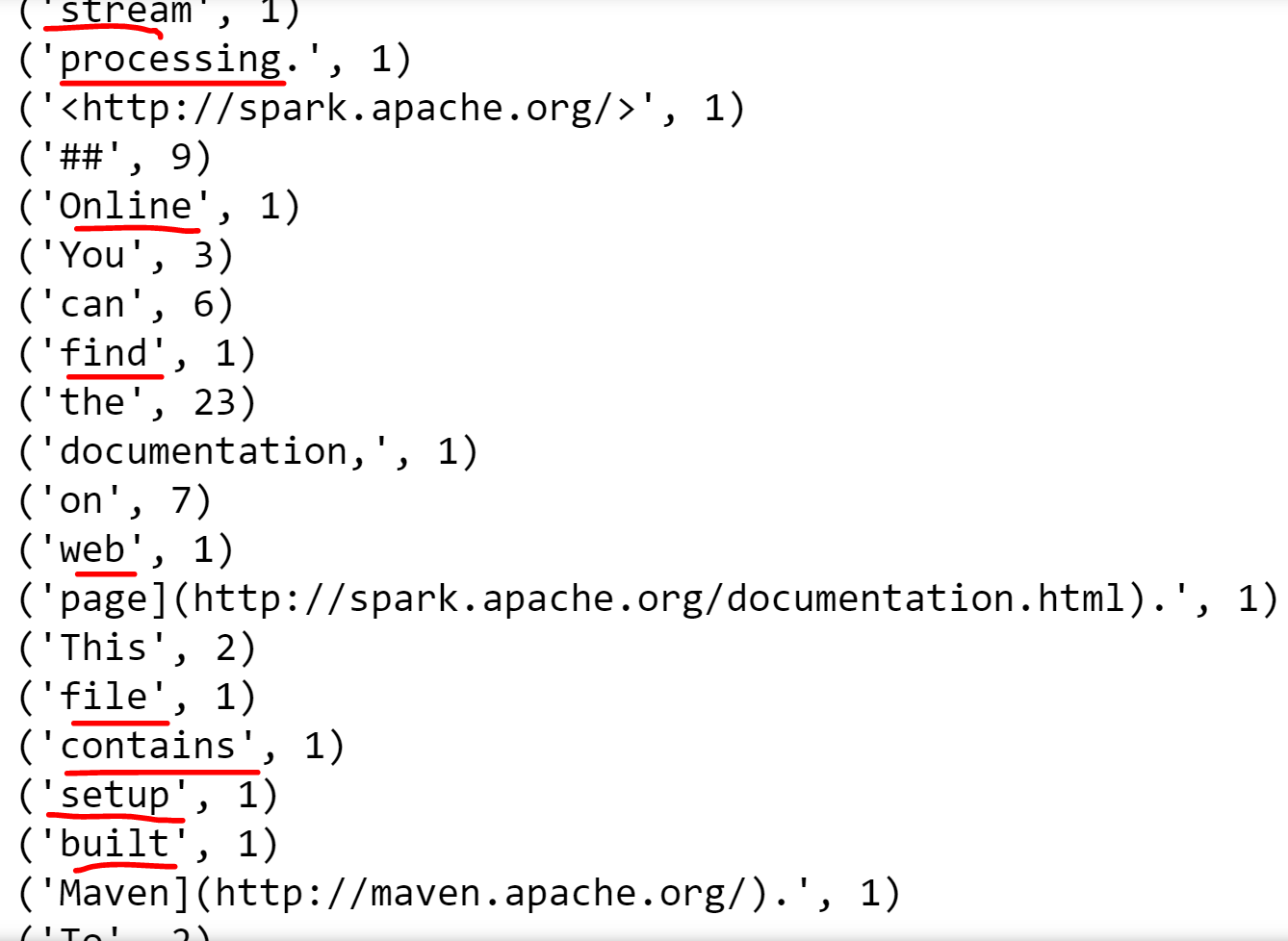
1. Most Common Word in the file:

‘the’ is the common word with 23:



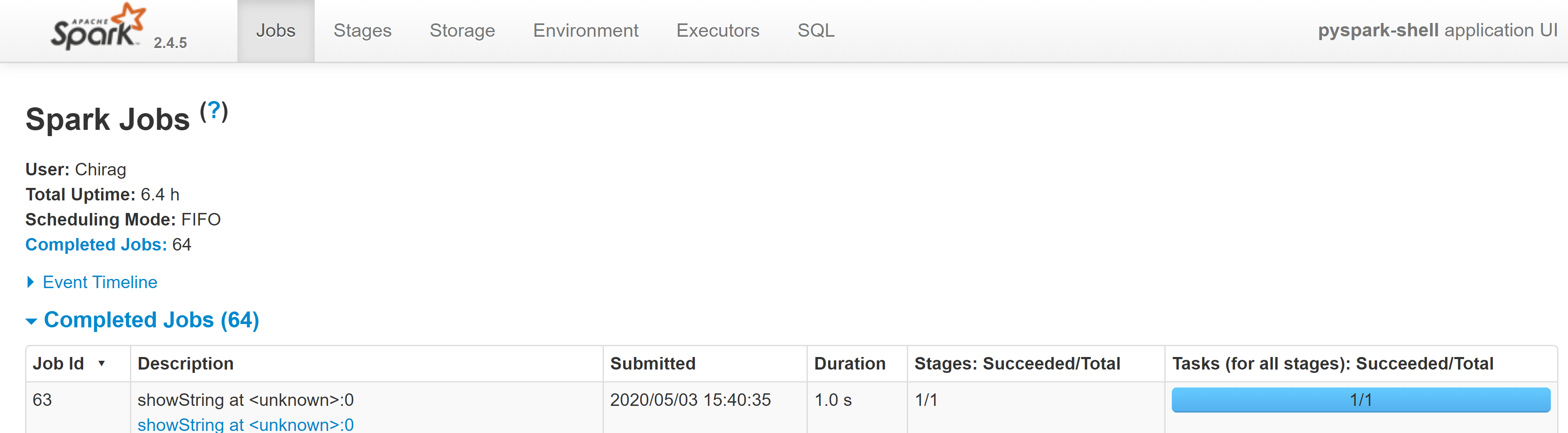
1. Multiple words occur once in the file:

There are many words which occur once in file as:



1. Total time (in seconds) to count the words in the file:

Total time to load and print the file is 1s as below:



**CONCULSION**

We installed the pyspark. With the help of Readme file, we counted the words and did word count analysis.

**REFERENCES**

1. <https://www.infoq.com/articles/apache-spark-introduction/>