Computer Science & Information Systems

**Big Data Systems – Spark Lab Sheet 3**

**Word Count with Spark**

1. Objective:

Students should be able to

1. Get familiarity with the execution of Python programmes on the Spark cluster
2. Get hands-on experience with word count map reduce programme

This lab sheet provides a quick introduction of using Spark for Map Reduce programe with Python. This exercise will introduce the API through pySpark package, then next labs will show how to write applications in Python.

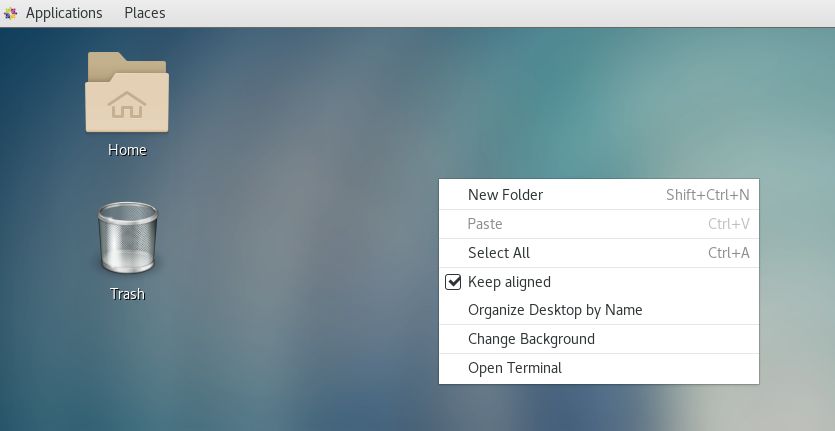
1. Steps to be performed:

Note - It’s assumed that student has made a slot reservation using the slot booking interface where Apache Spark framework was selected. The details of the Apache Spark systems to be used is received through an email. If not, please contact the administrators for the same.

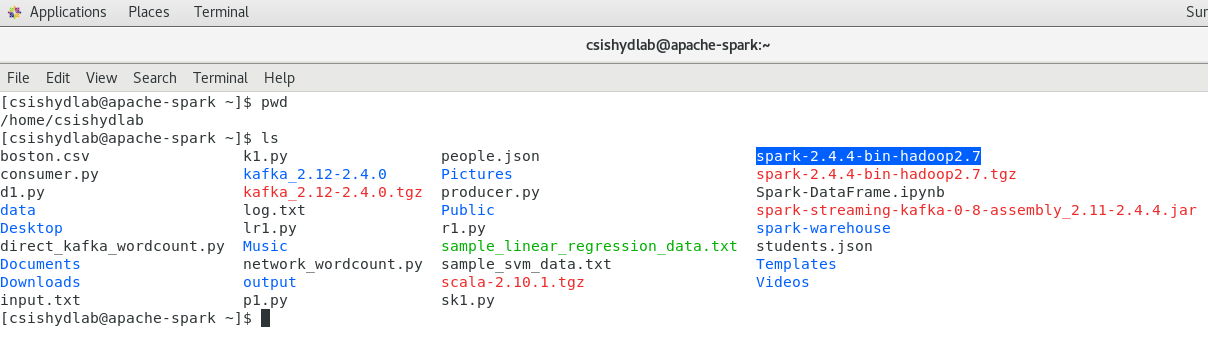
Also it’s assumed that students are aware of the process of logging into these virtual machines. If not, then get access to the user manual maintained for the usage of remote lab setup.

**Preparations -**

1. Open the terminal by right clicking on the desktop of the virtual machine.



1. Look at the current directory and also file listings in it. It must have a spark installation directory present in it. Commands like pwd, ls can be used for it.



1. Set the SPARK\_HOME and HOME variable to point to the spark installations.

[csishydlab@apache-spark bin]$ pwd

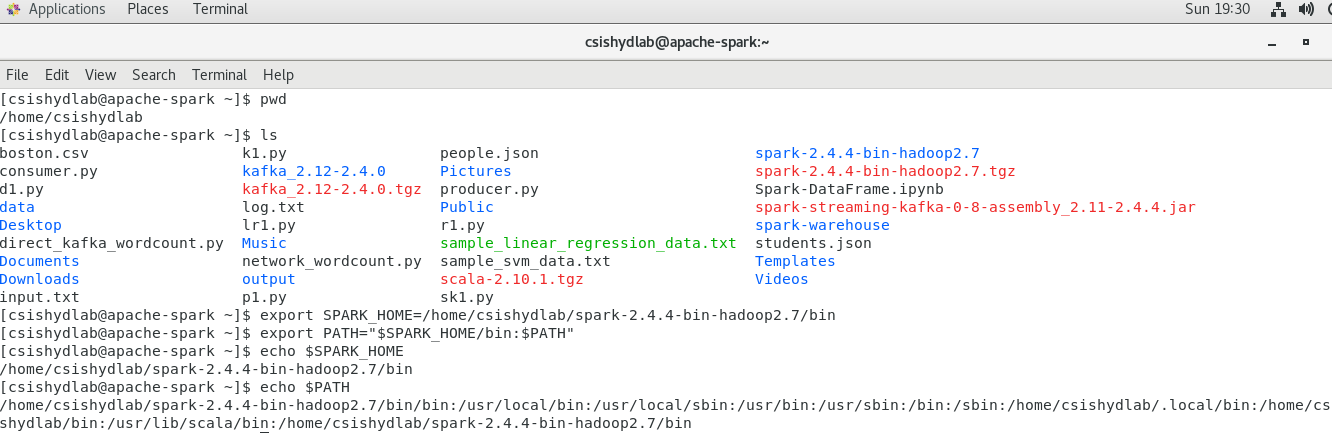
/home/csishydlab/spark-2.4.4-bin-hadoop2.7/bin

[csishydlab@apache-spark bin]$ export SPARK\_HOME=/home/csishydlab/spark-2.4.4-bin-hadoop2.7/bin

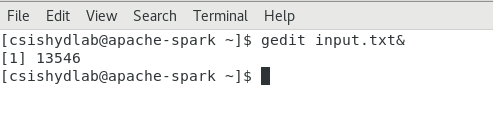
[csishydlab@apache-spark bin]$ export PATH="$SPARK\_HOME/bin:$PATH"

echo $SPARK\_HOME

echo $PATH



1. Prepare the input text file using any file editor. Copy and paste the content present in the attached input.txt file in this file.



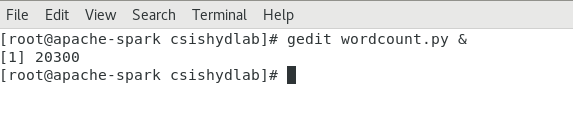
**Installing pySpark**

1. For the execution of python programmes on the Spark, a package named pyspark is required. Using the sudo previleges, install the packages with pip command.

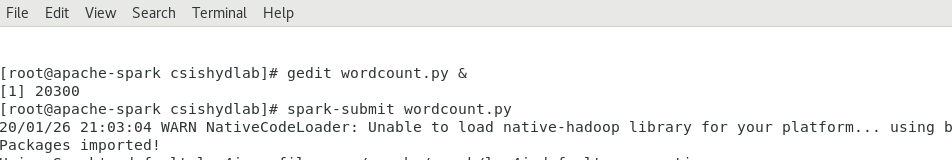
pip install pyspark

**Writing WordCount programme**

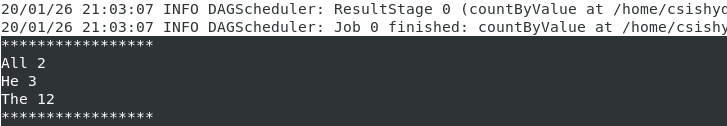
1. Open up the text editor and copy the code written in the attached wordcount.py file.



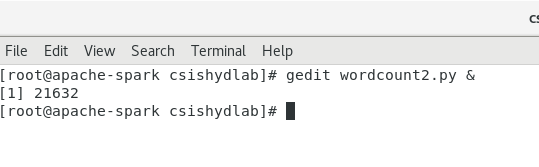
1. Execute the wordcount.py file using the spark-submit command.



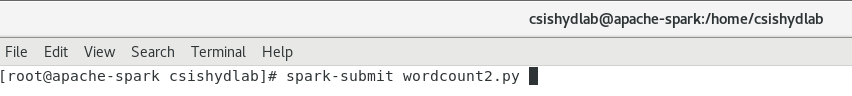
1. Look at the outcome printed while the program is getting executed on the Spark cluster. It shows how many times the first word of each lines has appeared.



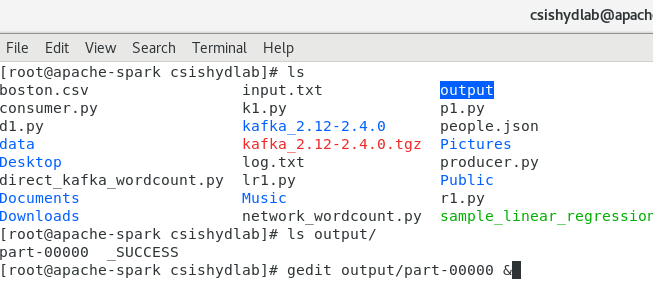
1. Open up the text editor and copy the code written in the attached wordcount2.py file.



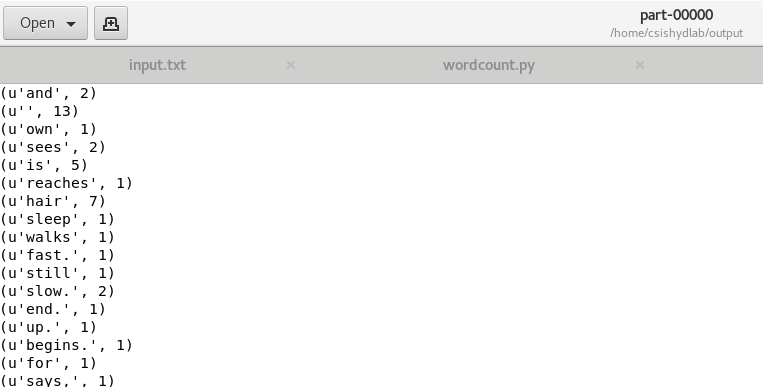
1. Execute the wordcount2.py file using the spark-submit command.



1. Look at the outcome printed while the program is getting executed on the Spark cluster. It shows how many times the word of each lines has appeared. The output will be stored in the “output” directory as follows



1. Look at the output in the file.



1. Outputs/Results:

Students should be able to

* Execute the python map reduce programme on Spark cluster
* See the word counts produced by the programme for the first word of every line of a file

1. Observations:

Students carefully needs to observe

* Details provided while spark application was running
* Number of maps executed
* Number of reducers used

1. References:
2. [Spark Documentation](https://spark.apache.org/docs/latest/quick-start.html)
3. [pySpark API Guide](https://spark.apache.org/docs/latest/api/python/index.html)