

ASSIGNMENT-3

BIG DATA PROGRAMING

1:

Installation of spark in kvm. As a First step check for updates using **\$ sudo apt update**. Before installing spark, the spark is written in scale so scale need to be installed first. This can be done using **\$sudo apt install scala** in the terminal. Once scala is installed check for its version using **\$scala -version** and to conform whether it's installed successfully.

```
vmalapati1@kvm_vmalapati1:~$ scala -version
Scala code runner version 2.11.12 -- Copyright 2002-2017, LAMP/EPFL
vmalapati1@kvm_vmalapati1:~$
```

Fig: scale version

Once scale is installed update the .bashrc such that the OS will recognize the scale using below commands adding to .bashrc file in bash script.

```
export SCALA_HOME=/home/vmalapati1/SCALA_HOME
export PATH=$SCALA_HOME/bin:$PATH
```

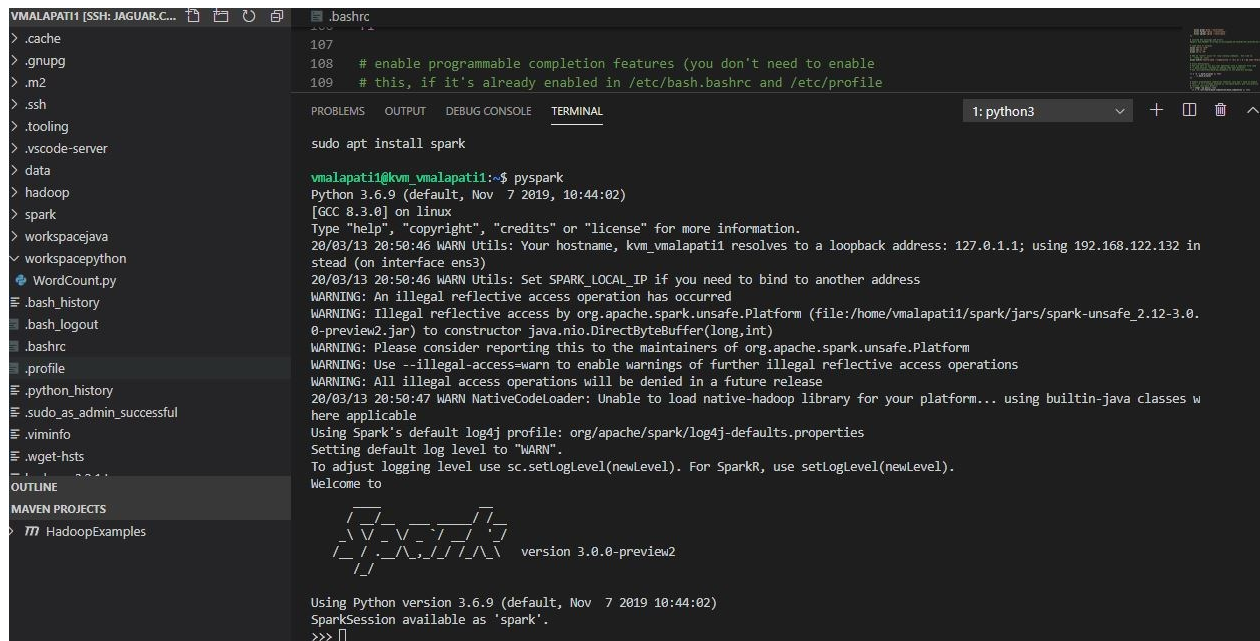
Installation of spark can be done in two ways, Chosee the specific version of spark as per requirement from the Apache spark org and download the .tgz spark file directly into local PC or KVM. Once downloaded into local PC unzip it in local pc and copy the entire spark folder into KVM using VS code in the home of kvm. Other way is directly download into kvm and unzip it from the terminal of kvm into home of kvm.

Once spark is installed update the .bashrc such that OS will know the spark and also python3 for spark, by adding below lines of code.

```
export SPARK_HOME=/home/vmalapati1/spark
export PATH=$SPARK_HOME/bin:$PATH
export PYSARK_PYTHON=python3
```

Once the code is added just run **\$source ~/.bashrc**

After these steps open spark terminal by typing pyspark command in terminal, if there exists any error related to permission use this command in terminal to give **permission chmod +x /home/rob/spark/bin/***



```
.bashrc
107
108 # enable programmable completion features (you don't need to enable
109 # this, if it's already enabled in /etc/bash.bashrc and /etc/profile

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL 1: python3 + [ ] [x] [ ] [ ]

sudo apt install spark

vmalapati1@kvm_vmalapati1:~$ pyspark
Python 3.6.9 (default, Nov 7 2019, 10:44:02)
[GCC 8.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
20/03/13 20:50:46 WARN Utils: Your hostname, kvm_vmalapati1 resolves to a loopback address: 127.0.1.1; using 192.168.122.132 in
stead (on interface ens3)
20/03/13 20:50:46 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/vmalapati1/spark/jars/spark-unsafe_2.12-3.0.
0-preview2.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/03/13 20:50:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      _ _ _ _ _ _ _ _ _ _
     / V _ _ _ _ _ _ _ \
    / _ / _ _ _ _ _ _ \ \
   / _ / _ _ _ _ _ _ \ \ \
  / _ / _ _ _ _ _ _ \ \ \ \
 / _ / _ _ _ _ _ _ \ \ \ \ \
/_ _ / _ _ _ _ _ _ \ \ \ \ \ \

version 3.0.0-preview2

Using Python version 3.6.9 (default, Nov 7 2019 10:44:02)
SparkSession available as 'spark'.
>>> |
```

Fig: showing spark shell opened in terminal of kvm

2.

Create a folder workspacepython to store the .py files

Get the wordcount.py file and running it on test and peterpan text files and printing the top K most frequency words on the terminal

The code takes Two input system arguments one the data file path and the other is the K value how many top k words to be printed out. The command for running python file in spark is as shown below. A python file can be run on spark by using spark-submit file location.

**spark-submit /home/vmalapati1/workspacepython/WordCount.py
/home/vmalapati1/data/test.txt 5**

**spark-submit /home/vmalapati1/workspacepython/WordCount.py
/home/vmalapati1/data/peterpan.txt 30**

The Three main blocks of code in wordcount.py are

- Configure the spark APP and set this configuration to context,
- Read the text file and apply transformation and actions on it, I mean map and reduce and splitting line by line with space
- After this sort the top K frequency words.

The output pictures after running wordcount on above to files are shown below

```
vmalapati@SSH: JAGUAR.CS.GSU.EDU workspacepython > WordCount.py ...
> .cache
> .gnupg / private-keys-v1.d
> .m2
> .ssh
> .tooling
> .vscode-server
> data
> output
> output_test
> Pagerank_output
> pr_test
01InitialPRValues.txt
02AdjacencyList.txt
peterpan.txt
test.txt
hadoop
spark
workspacejava
workspacepython
WordCount.py
OUTLINE
MAVEN PROJECTS
HadoopExamples

20/03/04 20:32:45 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
20/03/04 20:32:45 INFO ShuffleBlockFetcherIterator: Getting 1 (82.3 KiB) non-empty blocks including 1 (82.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
20/03/04 20:32:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 17 ms
20/03/04 20:32:45 INFO PythonRunner: Times: total = 75, boot = -1038, init = 1063, finish = 50
20/03/04 20:32:45 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 2270 bytes result sent to driver
20/03/04 20:32:45 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 248 ms on 192.168.122.132 (executor driver) (1 /1)
20/03/04 20:32:45 INFO DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:154) finished in 0.309 s
20/03/04 20:32:45 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
20/03/04 20:32:45 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
20/03/04 20:32:45 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
20/03/04 20:32:45 INFO DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:154, took 2.645537 s
[('the', 2331), ('', 2259), ('and', 1396), ('to', 1214), ('a', 962), ('of', 929), ('was', 890), ('he', 866), ('in', 683), ('that', 564), ('had', 498), ('it', 473), ('they', 465), ('she', 465), ('his', 455), ('you', 483), ('but', 378), ('for', 377), ('not', 375), ('with', 361), ('her', 361), ('is', 350), ('on', 329), ('at', 322), ('as', 315), ('I', 253), ('be', 249), ('have', 247), ('were', 243), ('Peter', 238)]
20/03/04 20:32:45 INFO SparkContext: Invoking stop() from shutdown hook
20/03/04 20:32:45 INFO SparkUI: Stopped Spark web UI at http://192.168.122.132:4040
20/03/04 20:32:45 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
20/03/04 20:32:45 INFO MemoryStore: MemoryStore cleared
20/03/04 20:32:45 INFO BlockManager: BlockManager stopped
20/03/04 20:32:45 INFO BlockManagerMaster: BlockManagerMaster stopped
20/03/04 20:32:45 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/03/04 20:32:45 INFO SparkContext: Successfully stopped SparkContext
20/03/04 20:32:45 INFO ShutdownHookManager: Shutdown hook called
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-243218a7-4b5a-47e1-aebc-393140d52161
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-5dc1b5a4-6f58-45e6-8fdb-fdbac2bfff11a/pyspark-073b2bff
-2f96-4fd2-aa57-00537ffb639
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-5dc1b5a4-6f58-45e6-8fdb-fdbac2bfff11a
vmalapati@kvm_vmalapati1:~$
vmalapati@kvm_vmalapati1:~$
vmalapati@kvm_vmalapati1:~$
```

Fig: showing result on peterpan

```
20/03/04 20:42:50 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
20/03/04 20:42:50 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, 192.168.122.132, executor driver, partition 0, NO DE LOCAL, 7143 bytes)
20/03/04 20:42:50 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
20/03/04 20:42:50 INFO ShuffleBlockFetcherIterator: Getting 1 (129.0 B) non-empty blocks including 1 (129.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
20/03/04 20:42:50 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 17 ms
20/03/04 20:42:50 INFO PythonRunner: Times: total = 50, boot = -755, init = 805, finish = 0
20/03/04 20:42:50 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1915 bytes result sent to driver
20/03/04 20:42:50 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 198 ms on 192.168.122.132 (executor driver) (1 /1)
20/03/04 20:42:50 INFO DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:154) finished in 0.235 s
20/03/04 20:42:50 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
20/03/04 20:42:50 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
20/03/04 20:42:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
20/03/04 20:42:50 INFO DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:154, took 2.113393 s
[('hadoop', 4), ('spark', 3), ('pig', 2), ('hive', 1), ('hbase', 1)]
20/03/04 20:42:50 INFO SparkContext: Invoking stop() from shutdown hook
20/03/04 20:42:50 INFO SparkUI: Stopped Spark web UI at http://192.168.122.132:4040
20/03/04 20:42:50 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
20/03/04 20:42:50 INFO MemoryStore: MemoryStore cleared
20/03/04 20:42:50 INFO BlockManager: BlockManager stopped
20/03/04 20:42:50 INFO BlockManagerMaster: BlockManagerMaster stopped
20/03/04 20:42:50 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/03/04 20:42:50 INFO SparkContext: Successfully stopped SparkContext
20/03/04 20:42:50 INFO ShutdownHookManager: Shutdown hook called
20/03/04 20:42:50 INFO ShutdownHookManager: Deleting directory /tmp/spark-f1e8af26-f27f-47f0-970d-5df2072bbbcf/pyspark-fdf5a68a
-7ff3-4ad9-8f8b-ed68d19af8e4
20/03/04 20:42:50 INFO ShutdownHookManager: Deleting directory /tmp/spark-b323c2c1-12f0-4e83-97a1-5ba620799ef8
20/03/04 20:42:50 INFO ShutdownHookManager: Deleting directory /tmp/spark-f1e8af26-f27f-47f0-970d-5df2072bbbcf
vmalapati@kvm_vmalapati1:~$
```

Fig: showing result on text


```
20/03/04 20:32:45 INFO DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:194, took 2.045557 s
[('the', 2331), ('', 2259), ('and', 1396), ('to', 1214), ('a', 962), ('of', 929), ('was', 898), ('he', 866), ('in', 683), ('tha
t', 564), ('had', 498), ('it', 473), ('they', 465), ('she', 465), ('his', 455), ('you', 403), ('but', 378), ('for', 377), ('not
', 375), ('with', 361), ('her', 361), ('is', 350), ('on', 329), ('at', 322), ('as', 315), ('I', 253), ('be', 249), ('have', 247
), ('were', 243), ('Peter', 238)]
20/03/04 20:32:45 INFO SparkContext: Invoking stop() from shutdown hook
20/03/04 20:32:45 INFO SparkUI: Stopped Spark web UI at http://192.168.122.132:4040
20/03/04 20:32:45 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
20/03/04 20:32:45 INFO MemoryStore: MemoryStore cleared
20/03/04 20:32:45 INFO BlockManager: BlockManager stopped
20/03/04 20:32:45 INFO BlockManagerMaster: BlockManagerMaster stopped
20/03/04 20:32:45 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/03/04 20:32:45 INFO SparkContext: Successfully stopped SparkContext
20/03/04 20:32:45 INFO ShutdownHookManager: Shutdown hook called
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-243218a7-4b5a-47e1-aebc-393140d52161
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-5dc1b5a4-6f58-45e6-8fdb-fdbac2bff11a/pyspark-073b2bff
-2f96-4fd2-aa57-00537ffb4639
20/03/04 20:32:45 INFO ShutdownHookManager: Deleting directory /tmp/spark-5dc1b5a4-6f58-45e6-8fdb-fdbac2bff11a
vmalapati1@kvm_vmalapati1:~$
vmalapati1@kvm_vmalapati1:~$
vmalapati1@kvm_vmalapati1:~$
```

Fig: sowing result on peterpan

The output results consists of printed top K most frequency words in a list of tuples as shown in above figures.

As Further more I have implemented one more system argument that takes path for writing the output of wordcount i.e top K most frequency words in a text file in the data/sparkoutput folder these output files contain output as follows

New command for running wordcount with output file path:

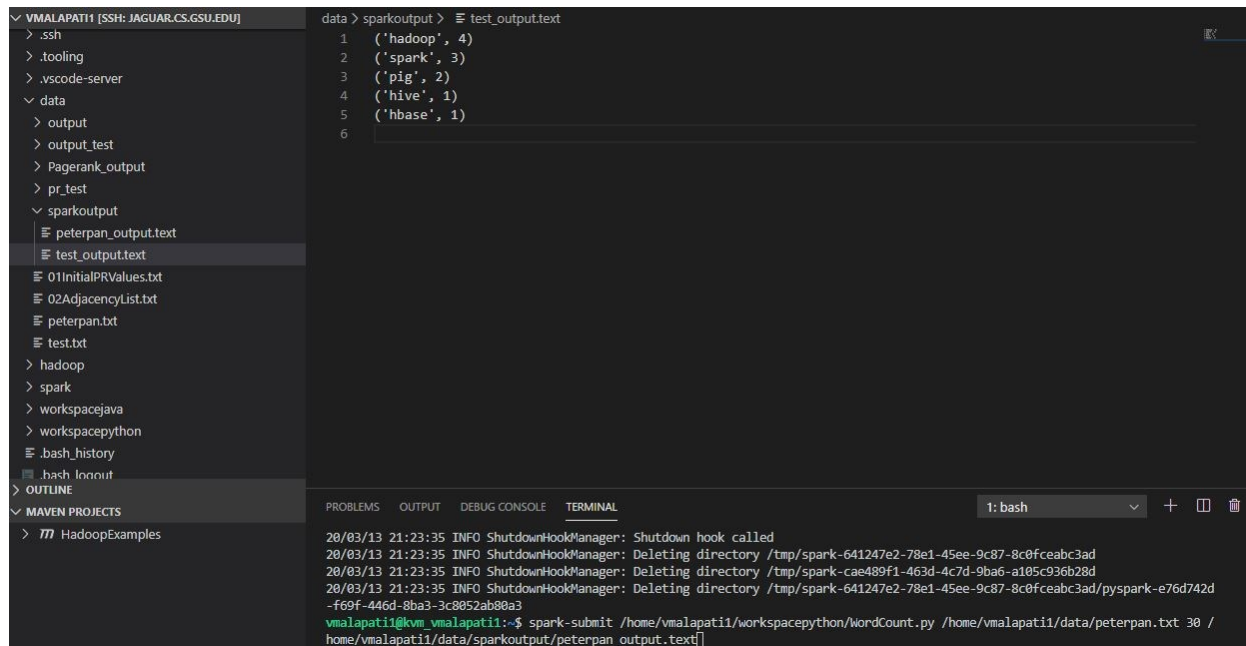
```
spark-submit /home/vmalapati1/workspacepython/WordCount.py
/home/vmalapati1/data/peterpan.txt 30
/home/vmalapati1/data/sparkoutput/peterpan_output.text
```

```
▼ VMALAPATI1 [SSH: JAGUARLC...  ▢ ▢ ▢
> .ssh
> .tooling
> .vscode-server
▼ data
> output
> output_test
> Pagerank_output
> pr_test
▼ sparkoutput
  peterpan_output.text
  test_output.text
  01InitialPRValues.txt
  02AdjacencyList.txt
  peterpan.txt
  test.txt
> hadoop
> spark
> workspacejava
> workspacepython
  .bash_history
  .bash_logout
▼ OUTLINE
▼ MAVEN PROJECTS
  HadoopExamples

data > sparkoutput > peterpan_output.text
1 ('the', 2331)
2 ('', 2259)
3 ('and', 1396)
4 ('to', 1214)
5 ('a', 962)
6 ('of', 929)
7 ('was', 898)
8 ('he', 866)
9 ('in', 683)
10 ('that', 564)
11 ('had', 498)
12 ('it', 473)
13 ('they', 465)
14 ('she', 465)
15 ('his', 455)
16 ('you', 403)
17 ('but', 378)
18 ('for', 377)
19 ('not', 375)
20 ('with', 361)
21 ('her', 361)
22 ('is', 350)
23 ('on', 329)
24 ('at', 322)

20/03/13 21:23:35 INFO ShutdownHookManager: Shutdown hook called
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-641247e2-78e1-45ee-9c87-8c0fceb3ad
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-cae489f1-463d-4c7d-9ba6-a105c936b28d
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-641247e2-78e1-45ee-9c87-8c0fceb3ad/pyspark-e76d742d
-f69f-446d-8ba3-3c8052ab80a3
vmalapati1@kvm_vmalapati1:~$ spark-submit /home/vmalapati1/workspacepython/WordCount.py /home/vmalapati1/data/peterpan.txt 30 /
/home/vmalapati1/data/sparkoutput/peterpan_output.text]
```

Fig: showing output file for peterpan data in KVM



```
data > sparkoutput > test_output.txt
1 ('hadoop', 4)
2 ('spark', 3)
3 ('pig', 2)
4 ('hive', 1)
5 ('hbase', 1)
6

20/03/13 21:23:35 INFO ShutdownHookManager: Shutdown hook called
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-641247e2-78e1-45ee-9c87-8c0fceb3ad
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-cae489f1-463d-4c7d-9ba6-a105c936b28d
20/03/13 21:23:35 INFO ShutdownHookManager: Deleting directory /tmp/spark-641247e2-78e1-45ee-9c87-8c0fceb3ad/pyspark-e76d742d-f69f-446d-8ba3-3c8052ab80a3
vmalapati1@kvm_vmalapati1:~$ spark-submit /home/vmalapati1/workspacepython/wordCount.py /home/vmalapati1/data/peterpan.txt 30 /
home/vmalapati1/data/sparkoutput/peterpan_output.txt[]
```

Fig: Showing output file for test