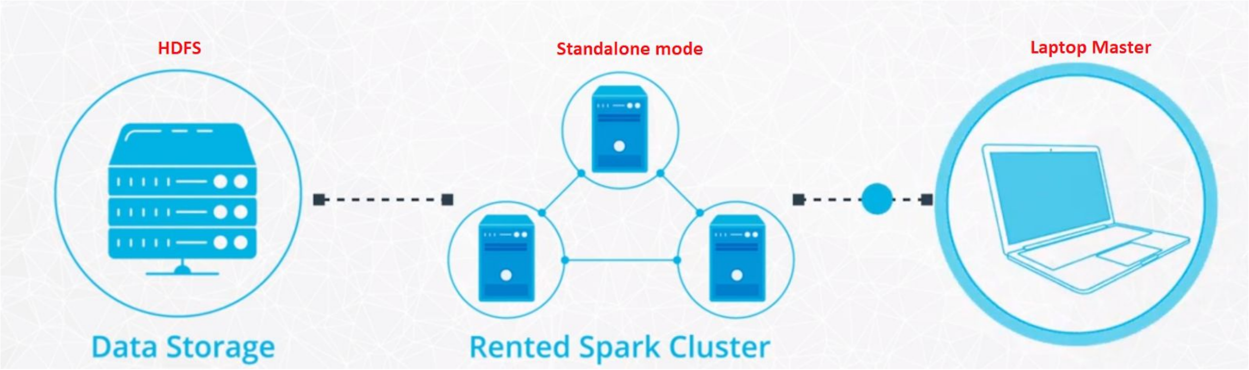
BÁO CÁO LAB 4
CÀI ĐẶT CỤM SPARK VÀ CHẠY CHƯƠNG TRÌNH WORDCOUNT
NHÓM: GIỮA CHÚNG TA

Tiếp nối bài cài đặt HDFS + YARN, tính WordCount bằng Hadoop MapReduce, trong bài này chúng ta sẽ cài đặt cụm spark và chạy trương trình WordCount.

Nguyên lý hoạt động: Cài đặt Spark Cluster ở Standalone mode. Master gửi job, load dữ liệu từ HDFS, phân tích dữ liệu bằng Spark. Dữ liệu khi phân tích xong được lưu lại vào HDFS, hiển thị kết quả.

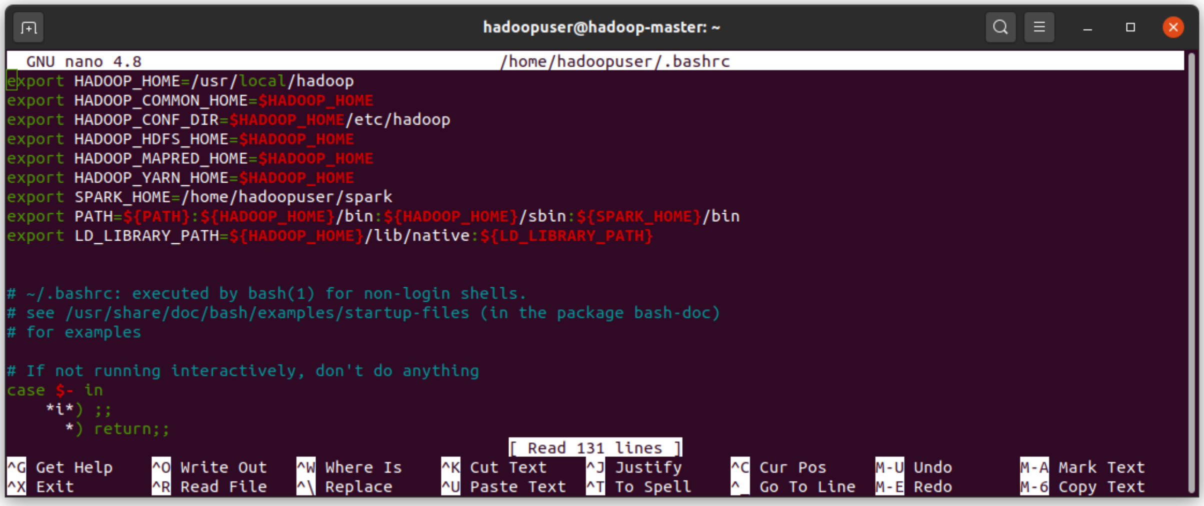


hadoopuser@hadoop-master: Master + Worker
hadoopuser@hadoop-slave1: Worker
hadoopuser@hadoop-slave2: Worker

1. **Tải spark về**

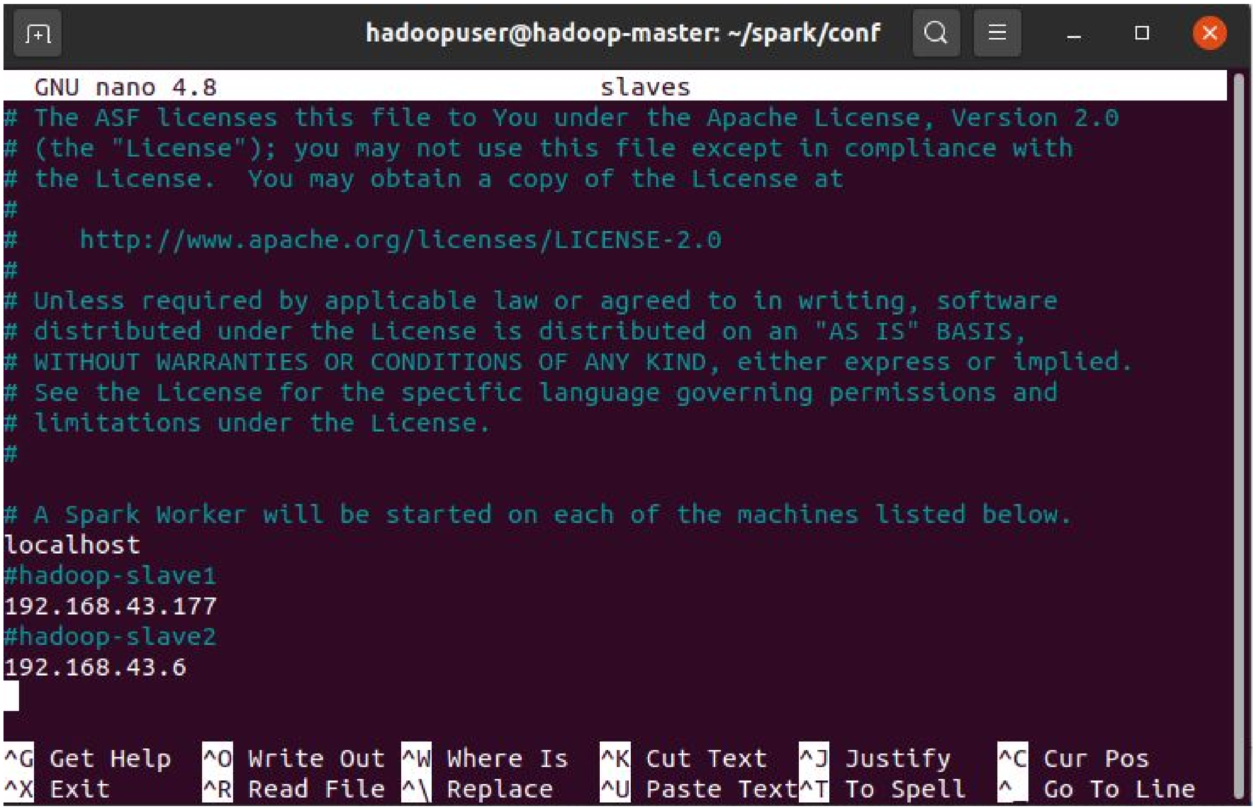
```
wget http://apache.claz.org/spark/spark-2.4.0/spark-2.4.0-bin-hadoop2.7.tgz

tar -xzf spark-2.4.0-bin-hadoop2.7.tgz

mv spark-2.4.0-bin-hadoop2.7 spark
```

2. **Tạo các biến môi trường**



```
  GNU nano 4.8                                  /home/hadoopuser/.bashrc
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export SPARK_HOME=/home/hadoopuser/spark
export PATH=${PATH}:${HADOOP_HOME}/bin:${HADOOP_HOME}/sbin:${SPARK_HOME}/bin
export LD_LIBRARY_PATH=${HADOOP_HOME}/lib/native:${LD_LIBRARY_PATH}


# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
case $- in
    *i*) ;;
      *) return;;
                                             [ Read 131 lines ]
^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text
```

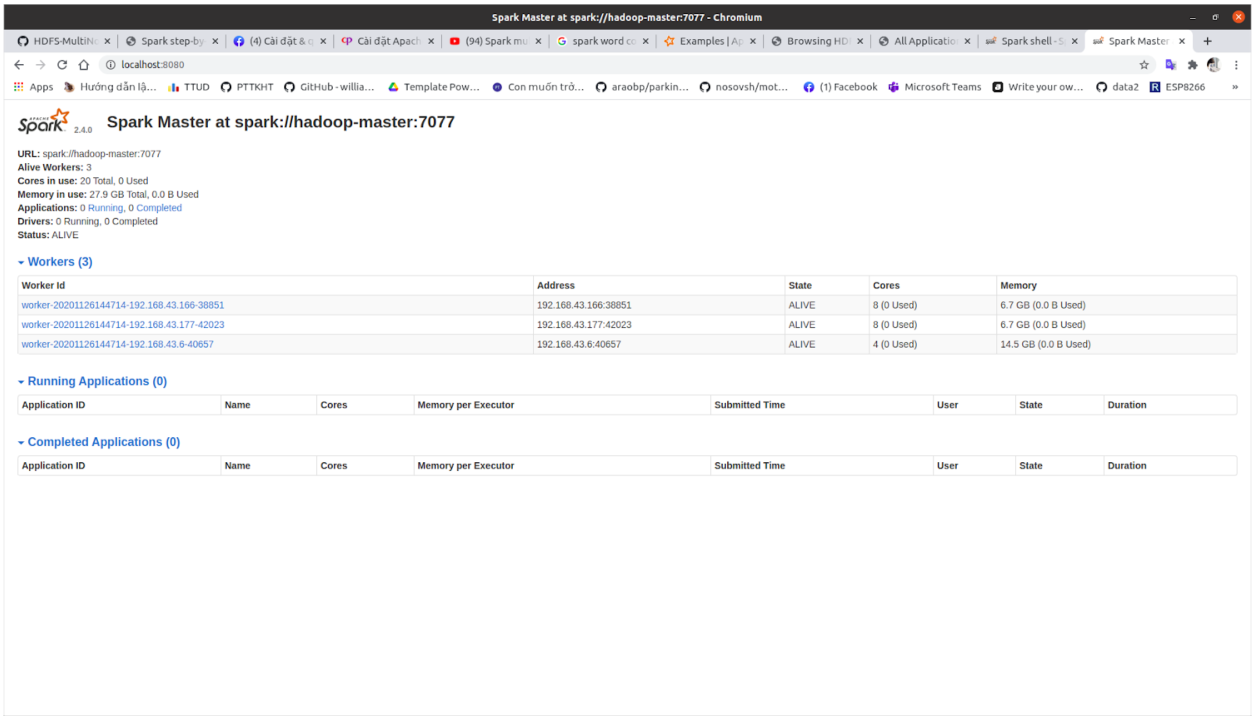
3. Tạo file slave ghi các địa chỉ slave



```
  GNU nano 4.8                          slaves
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

# A Spark Worker will be started on each of the machines listed below.
localhost
#hadoop-slave1
192.168.43.177
#hadoop-slave2
192.168.43.6


^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text^T To Spell    ^_ Go To Line
```

4. Start spark

```
$ cd $SPARK_HOME/sbin
$ ./start-all.sh
```



Spark Master at spark://hadoop-master:7077

URL: spark://hadoop-master:7077
Alive Workers: 3
Cores in use: 20 Total, 0 Used
Memory in use: 27.9 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20201126144714-192.168.43.166-38851	192.168.43.166:38851	ALIVE	8 (0 Used)	6.7 GB (0.0 B Used)
worker-20201126144714-192.168.43.177-42023	192.168.43.177:42023	ALIVE	8 (0 Used)	6.7 GB (0.0 B Used)
worker-20201126144714-192.168.43.6-40657	192.168.43.6:40657	ALIVE	4 (0 Used)	14.5 GB (0.0 B Used)

Running Applications (0)

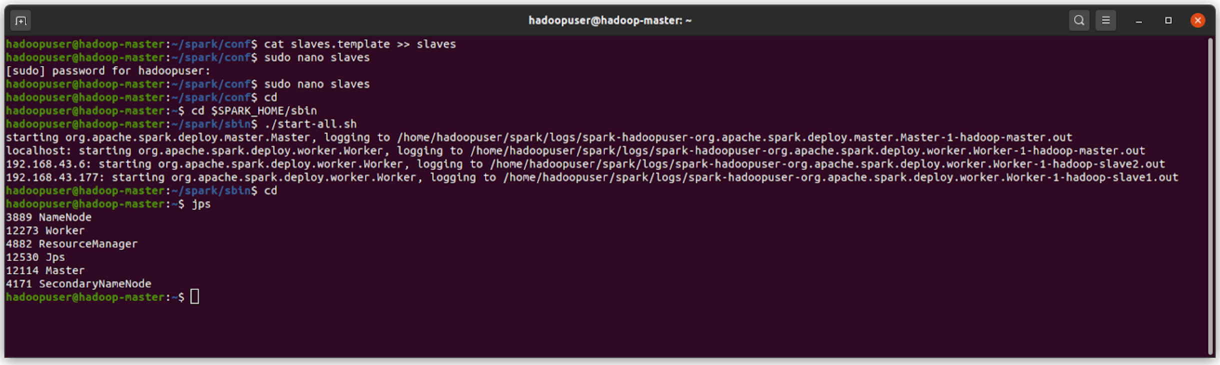
Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

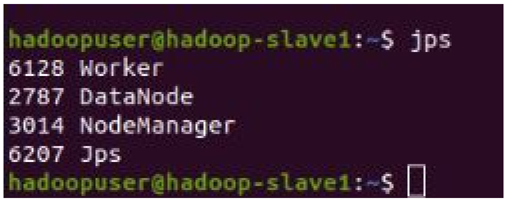
5. Check trên từng máy xem đã thành công chưa

- Với hadoopuser@hadoop-master: Master + Worker



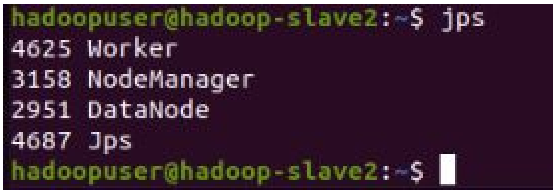
```
hadoopuser@hadoop-master:~/spark/conf$ cat slaves.template >> slaves
hadoopuser@hadoop-master:~/spark/conf$ sudo nano slaves
[sudo] password for hadoopuser:
hadoopuser@hadoop-master:~/spark/conf$ sudo nano slaves
hadoopuser@hadoop-master:~/spark/conf$ cd
hadoopuser@hadoop-master:~$ cd $SPARK_HOME/sbin
hadoopuser@hadoop-master:~/spark/sbin$ ./start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/hadoopuser/spark/logs/spark-hadoopuser-org.apache.spark.deploy.master.Master-1-hadoop-master.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/hadoopuser/spark/logs/spark-hadoopuser-org.apache.spark.deploy.worker.Worker-1-hadoop-master.out
192.168.43.6: starting org.apache.spark.deploy.worker.Worker, logging to /home/hadoopuser/spark/logs/spark-hadoopuser-org.apache.spark.deploy.worker.Worker-1-hadoop-slave2.out
192.168.43.177: starting org.apache.spark.deploy.worker.Worker, logging to /home/hadoopuser/spark/logs/spark-hadoopuser-org.apache.spark.deploy.worker.Worker-1-hadoop-slave1.out
hadoopuser@hadoop-master:~/spark/sbin$ cd
hadoopuser@hadoop-master:~$ jps
3889 NameNode
12273 Worker
4882 ResourceManager
12530 Jps
12114 Master
4171 SecondaryNameNode
hadoopuser@hadoop-master:~$ 
```

- Với hadoopuser@hadoop-slave1: Worker



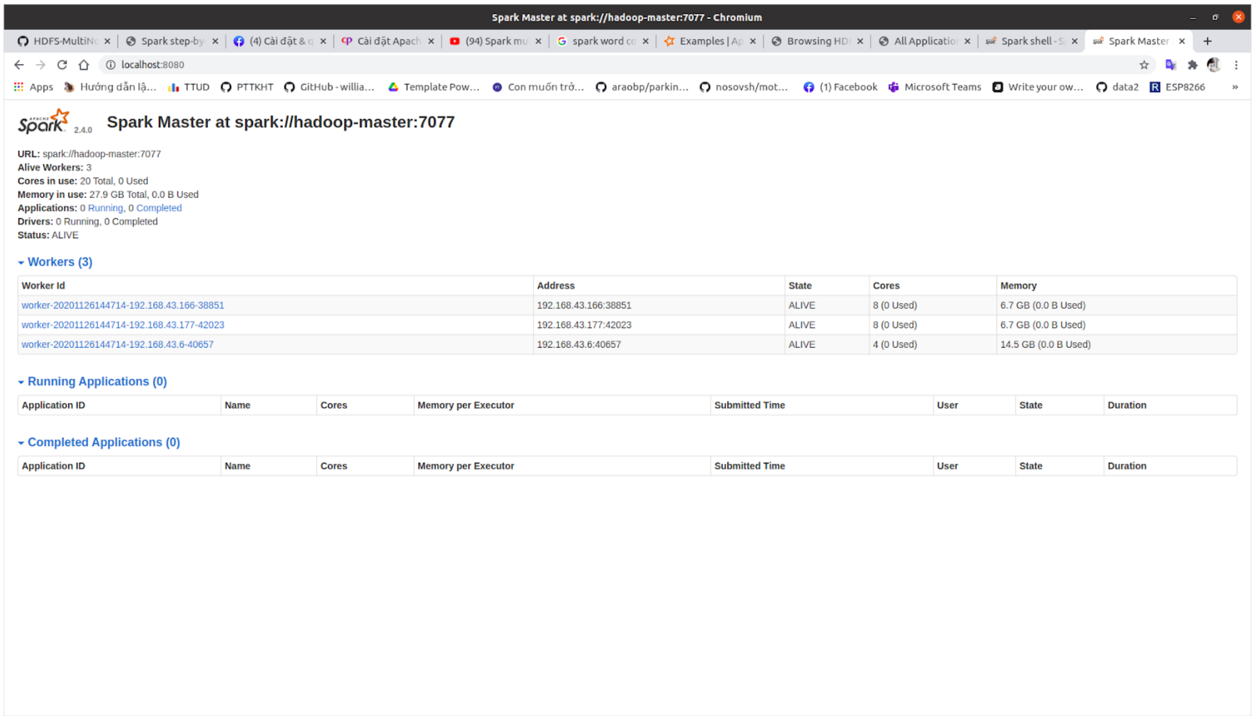
```
hadoopuser@hadoop-slave1:~$ jps
6128 Worker
2787 DataNode
3014 NodeManager
6207 Jps
hadoopuser@hadoop-slave1:~$ 
```

- Với hadoopuser@hadoop-slave2: Worker



```
hadoopuser@hadoop-slave2:~$ jps
4625 Worker
3158 NodeManager
2951 DataNode
4687 Jps
hadoopuser@hadoop-slave2:~$ 
```

6. Check trên từng máy thành công, kiểm tra tiếp trên webUI



Spark Master at spark://hadoop-master:7077

URL: spark://hadoop-master:7077
Alive Workers: 3
Cores in use: 20 Total, 0 Used
Memory in use: 27.9 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20201126144714-192.168.43.166-38851	192.168.43.166:38851	ALIVE	8 (0 Used)	6.7 GB (0.0 B Used)
worker-20201126144714-192.168.43.177-42023	192.168.43.177:42023	ALIVE	8 (0 Used)	6.7 GB (0.0 B Used)
worker-20201126144714-192.168.43.6-40657	192.168.43.6:40657	ALIVE	4 (0 Used)	14.5 GB (0.0 B Used)

Running Applications (0)

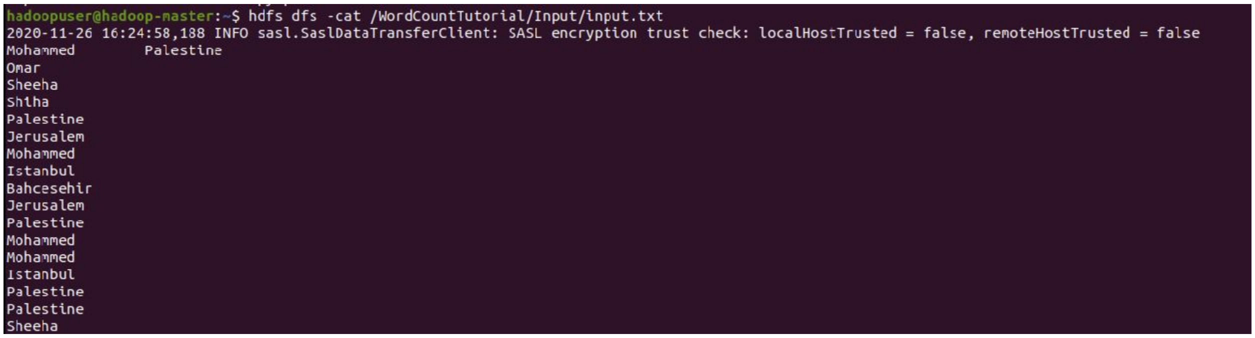
Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

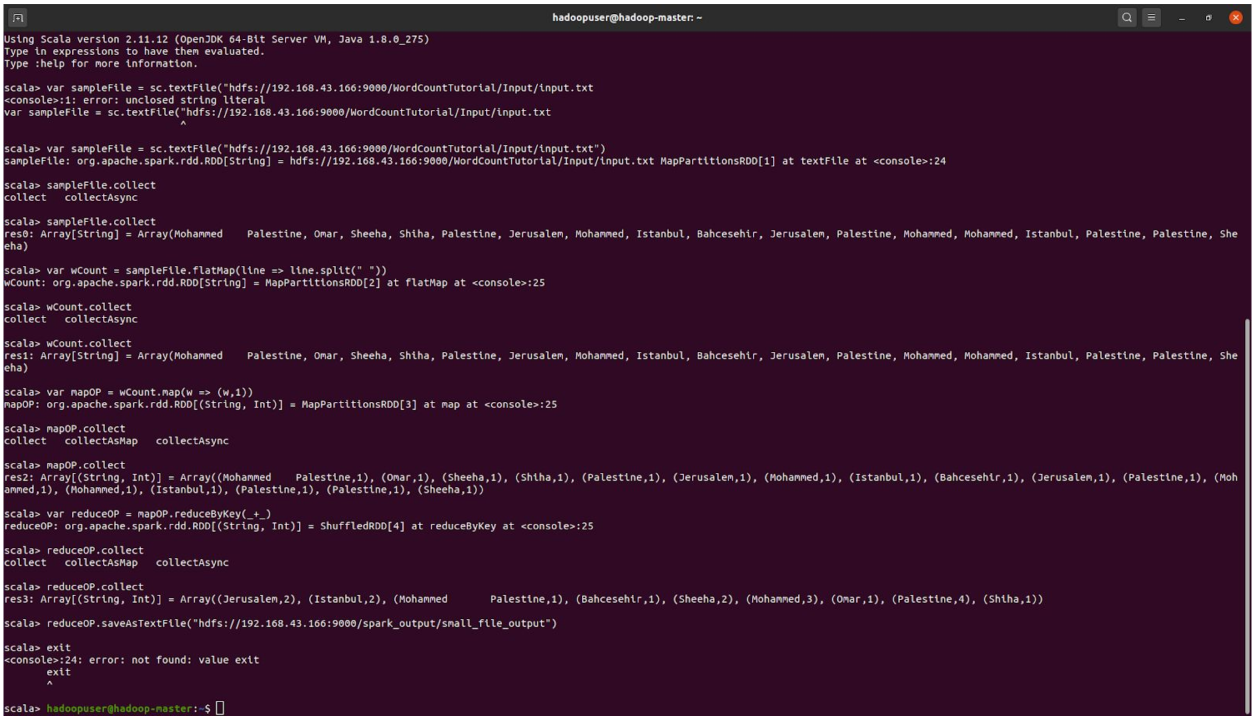
7. Chạy chương trình WordCount

- Trước tiên kiểm tra file từ cụm HDFS



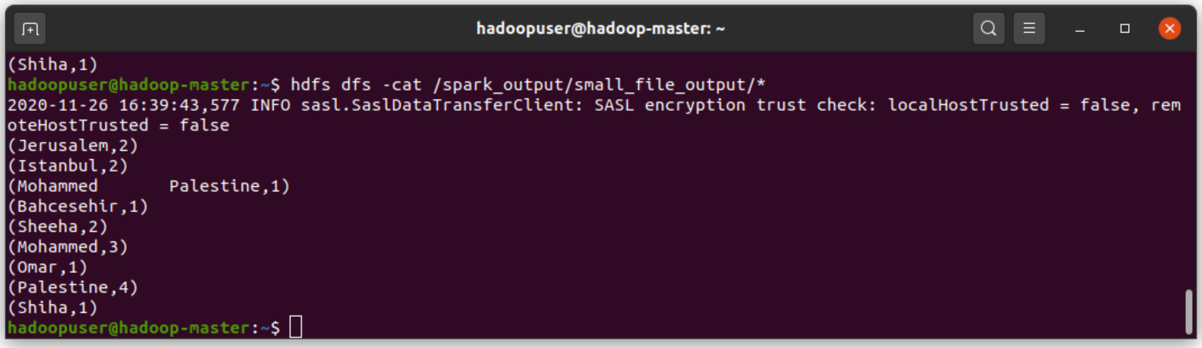
```
hadoopuser@hadoop-master:~$ hdfs dfs -cat /WordCountTutorial/Input/input.txt
2020-11-26 16:24:58,188 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Mohammed        Palestine
Omar
Sheeha
Shiha
Palestine
Jerusalem
Mohammed
Istanbul
Bahcesehir
Jerusalem
Palestine
Mohammed
Mohammed
Istanbul
Palestine
Palestine
Sheeha
```

- Thực thi chương trình WordCount và up lên cụm HDFS



```
Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_275)
Type in expressions to have them evaluated.
Type :help for more information.

scala> var sampleFile = sc.textFile("hdfs://192.168.43.166:9000/WordCountTutorial/Input/input.txt
<console>:1: error: unclosed string literal
var sampleFile = sc.textFile("hdfs://192.168.43.166:9000/WordCountTutorial/Input/input.txt
                             ^

scala> var sampleFile = sc.textFile("hdfs://192.168.43.166:9000/WordCountTutorial/Input/input.txt")
sampleFile: org.apache.spark.rdd.RDD[String] = hdfs://192.168.43.166:9000/WordCountTutorial/Input/input.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> sampleFile.collect
collect   collectAsync

scala> sampleFile.collect
res0: Array[String] = Array(Mohammed    Palestine, Omar, Sheeha, Shiha, Palestine, Jerusalem, Mohammed, Istanbul, Bahcesehir, Jerusalem, Palestine, Mohammed, Mohammed, Istanbul, Palestine, Palestine, She
eha)

scala> var wCount = sampleFile.flatMap(line => line.split(" "))
wCount: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:25

scala> wCount.collect
collect   collectAsync

scala> wCount.collect
res1: Array[String] = Array(Mohammed    Palestine, Omar, Sheeha, Shiha, Palestine, Jerusalem, Mohammed, Istanbul, Bahcesehir, Jerusalem, Palestine, Mohammed, Mohammed, Istanbul, Palestine, Palestine, She
eha)

scala> var mapOP = wCount.map(w => (w,1))
mapOP: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:25

scala> mapOP.collect
collect   collectAsMap   collectAsync

scala> mapOP.collect
res2: Array[(String, Int)] = Array((Mohammed    Palestine,1), (Omar,1), (Sheeha,1), (Shiha,1), (Palestine,1), (Jerusalem,1), (Mohammed,1), (Istanbul,1), (Bahcesehir,1), (Jerusalem,1), (Palestine,1), (Moh
ammed,1), (Mohammed,1), (Istanbul,1), (Palestine,1), (Palestine,1), (Sheeha,1))

scala> var reduceOP = mapOP.reduceByKey(_+_)
reduceOP: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:25

scala> reduceOP.collect
collect   collectAsMap   collectAsync

scala> reduceOP.collect
res3: Array[(String, Int)] = Array((Jerusalem,2), (Istanbul,2), (Mohammed       Palestine,1), (Bahcesehir,1), (Sheeha,2), (Mohammed,3), (Omar,1), (Palestine,4), (Shiha,1))

scala> reduceOP.saveAsTextFile("hdfs://192.168.43.166:9000/spark_output/small_file_output")

scala> exit
<console>:24: error: not found: value exit
       exit
       ^

scala> hadoopuser@hadoop-master:~$ 
```

- Kết quả thu được trên cụm HDFS



```
(Shiha,1)
hadoopuser@hadoop-master:~$ hdfs dfs -cat /spark_output/small_file_output/*
2020-11-26 16:39:43,577 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, rem
oteHostTrusted = false
(Jerusalem,2)
(Istanbul,2)
(Mohammed       Palestine,1)
(Bahcesehir,1)
(Sheeha,2)
(Mohammed,3)
(Omar,1)
(Palestine,4)
(Shiha,1)
hadoopuser@hadoop-master:~$ 
```