

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



BÁO CÁO MÔN HỌC

**Lưu trữ và xử lý dữ liệu lớn**

**LAB04: Install Spark**

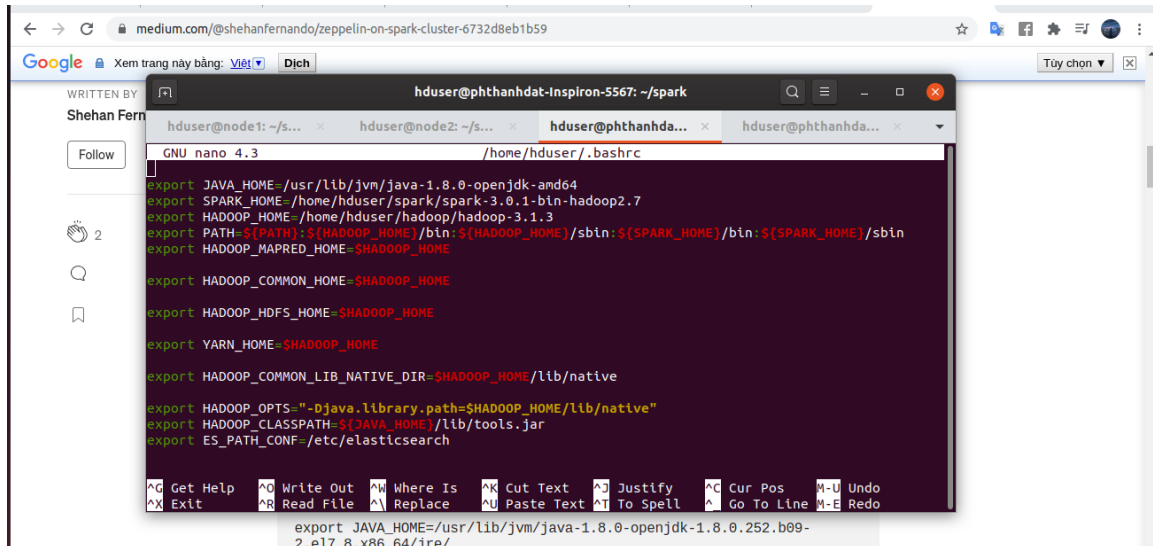
Nhóm: **RHUST**

Phan Thành Đạt	20173001
Hoàng Văn Chương	20172984
Đỗ Minh Vũ	20173471
Nguyễn Thị Bắc	20172963

# 1) Cài đặt cụm Spark

Tải Spark :

wget <https://downloads.apache.org/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz>

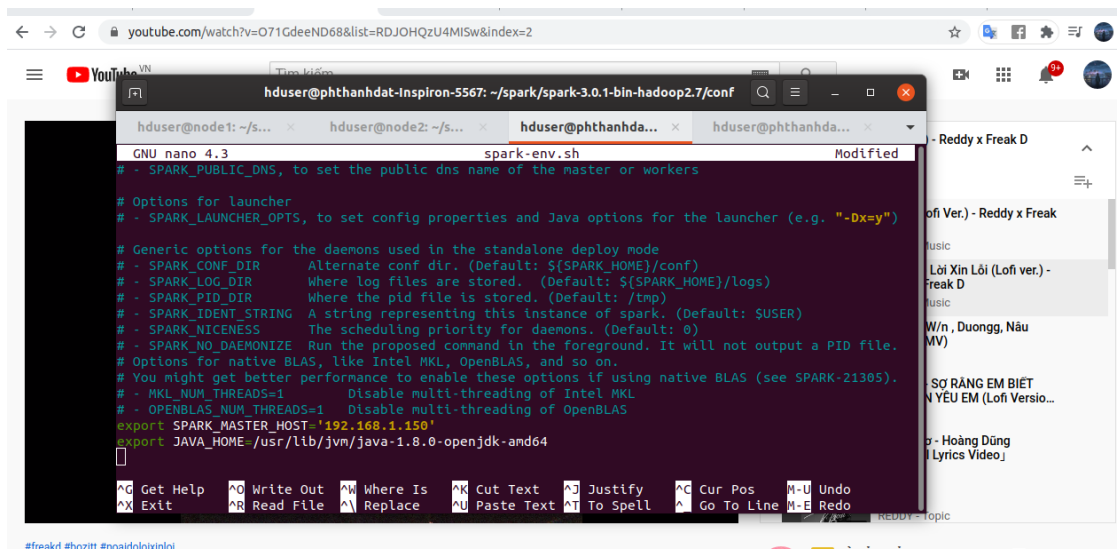


```
hduser@phthanhdat-Inspiron-5567: ~/spark
GNU nano 4.3 /home/hduser/.bashrc
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export SPARK_HOME=/home/hduser/spark/spark-3.0.1-bin-hadoop2.7
export HADOOP_HOME=/home/hduser/hadoop/hadoop-3.1.3
export PATH=${PATH}:${HADOOP_HOME}/bin:${SPARK_HOME}/bin:${SPARK_HOME}/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME

export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
export ES_PATH_CONF=/etc/elasticsearch

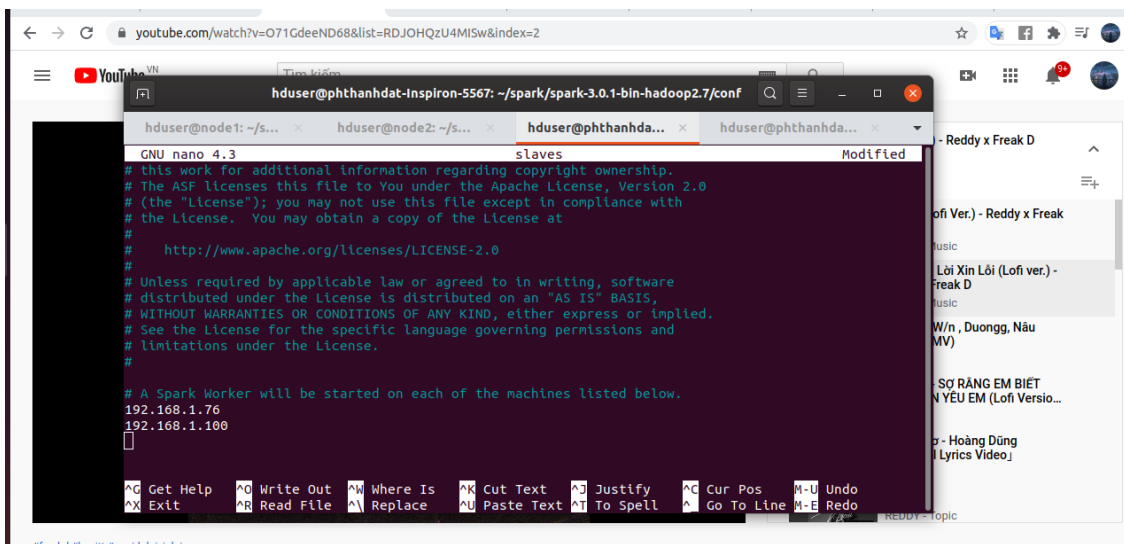
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.252.b09-2.el7_8.x86_64/jre/
```



```
hduser@phthanhdat-Inspiron-5567: ~/spark/spark-3.0.1-bin-hadoop2.7/conf
GNU nano 4.3 spark-env.sh Modified
# SPARK_PUBLIC_DNS, to set the public dns name of the master or workers

# Options for launcher
# - SPARK_LAUNCHER_OPTS, to set config properties and Java options for the launcher (e.g. "-Dx=y")

# Generic options for the daemons used in the standalone deploy mode
# - SPARK_CONF_DIR      Alternate conf dir. (Default: ${SPARK_HOME}/conf)
# - SPARK_LOG_DIR       Where log files are stored. (Default: ${SPARK_HOME}/logs)
# - SPARK_PID_DIR       Where the pid file is stored. (Default: /tmp)
# - SPARK_IDENT_STRING  A string representing this instance of spark. (Default: $USER)
# - SPARK_NICENESS       The scheduling priority for daemons. (Default: 0)
# - SPARK_NO_DAEMONIZE  Run the proposed command in the foreground. It will not output a PID file.
# Options for native BLAS, like Intel MKL, OpenBLAS, and so on.
# You might get better performance to enable these options if using native BLAS (see SPARK-21305).
# - MKL_NUM_THREADS=1   Disable multi-threading of Intel MKL
# - OPENBLAS_NUM_THREADS=1 Disable multi-threading of OpenBLAS
export SPARK_MASTER_HOST='192.168.1.150'
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
```



```
hduser@phthanhdat-Inspiron-5567: ~/spark/spark-3.0.1-bin-hadoop2.7/conf
GNU nano 4.3 slaves Modified
# this work for additional information regarding copyright ownership.
# 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.
192.168.1.76
192.168.1.100
```

Kiểm tra trên UI Web:

How to Install and Set Up | Facebook | 5. WordCount - Zeppelin | Spark Master at spark://192.168.1.150:7077

spark 3.0.1

## Spark Master at spark://192.168.1.150:7077

URL: spark://192.168.1.150:7077

Alive Workers: 2

Cores in use: 2 Total, 0 Used

Memory in use: 20.0 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20201127230731-192.168.1.100-40759	192.168.1.100:40759	ALIVE	1 (0 Used)	10.0 GiB (0.0 B Used)	
worker-20201127230746-192.168.1.76-34791	192.168.1.76:34791	ALIVE	1 (0 Used)	10.0 GiB (0.0 B Used)	

Running Applications (0)

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

Completed Applications (0)

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

Submit Job wordcount:

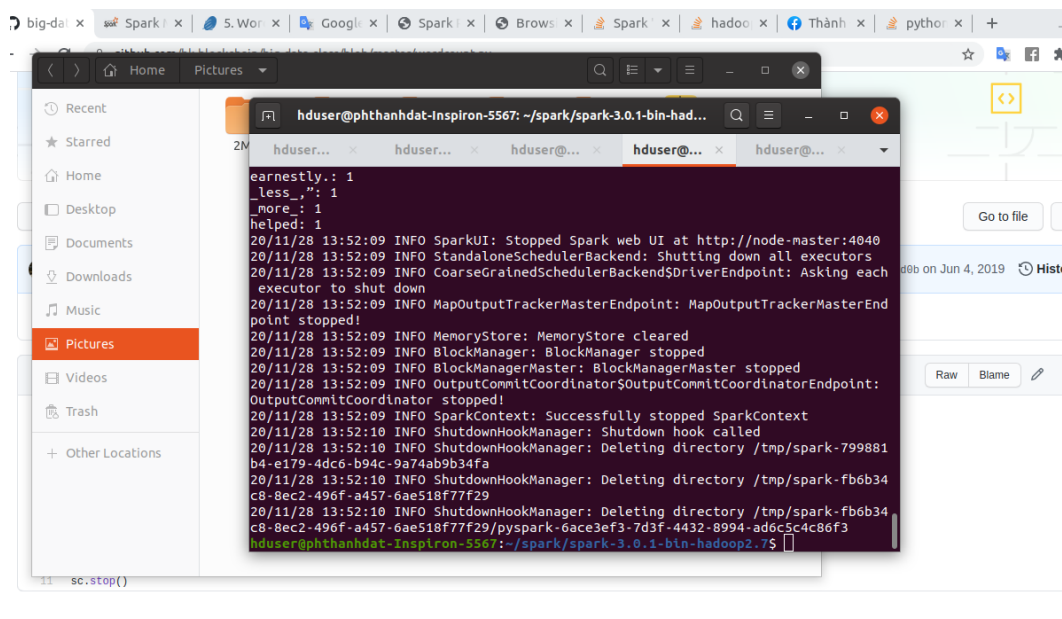
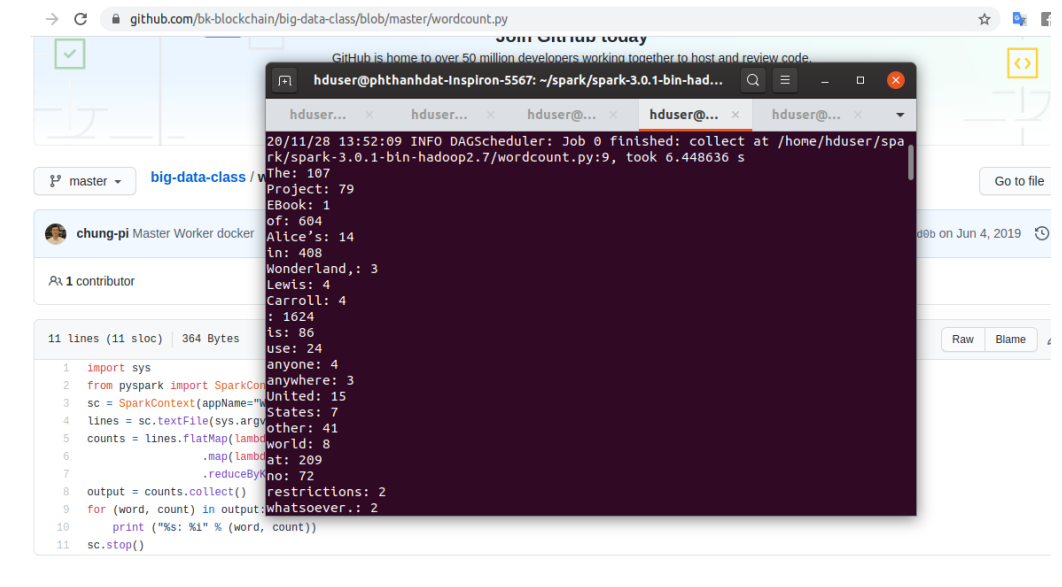
```
./bin/spark-submit wordcount.py
```

The screenshot shows a GitHub repository page for 'blue-data-class' by 'chung-pi'. A terminal window is open, displaying the execution of a Spark application. The terminal output shows the Spark context starting, security manager initialization, and the successful submission of a WordCountExample application. The application is running on a Spark cluster with 39191 cores and 39191 GB of memory. The logs indicate that the application is successfully started and is running on port 8080.

```

hduser@phthanhdot-Inspiron-5567: ~/spark/spark-3.0.1-bin-had...
hduser... x hduser... x hduser@... x hduser@... x hduser@... x
hduser@phthanhdot-Inspiron-5567:~/spark/spark-3.0.1-bin-hadoop2.7$ ./bin/spark-s
submit wordcount.py
20/11/28 13:51:57 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/11/28 13:51:58 INFO SparkContext: Running Spark version 3.0.1
20/11/28 13:51:58 INFO ResourceUtils: =====
20/11/28 13:51:58 INFO ResourceUtils: Resources for spark.driver:
=====
20/11/28 13:51:58 INFO ResourceUtils: =====
20/11/28 13:51:58 INFO SparkContext: Submitted application: WordCountExample
20/11/28 13:51:58 INFO SecurityManager: Changing view acls to: hduser
20/11/28 13:51:58 INFO SecurityManager: Changing modify acls to: hduser
20/11/28 13:51:58 INFO SecurityManager: Changing view acls groups to:
20/11/28 13:51:58 INFO SecurityManager: Changing modify acls groups to:
20/11/28 13:51:58 INFO SecurityManager: SecurityManager: authentication disabled
; ut acls disabled; users with view permissions: Set(hduser); groups with view
permissions: Set(); users with modify permissions: Set(hduser); groups with mod
ify permissions: Set()
20/11/28 13:51:58 INFO Utils: Successfully started service 'sparkDriver' on port
39191.
20/11/28 13:51:58 INFO SparkEnv: Registering MapOutputTracker
19 print ("%s: %1" % (word, count))
20 for (word, count) in output:
21     sc.stop()

```



Kiểm tra trên web:

**Spark Master at spark://192.168.1.150:7077**

URL: spark://192.168.1.150:7077  
 Alive Workers: 2  
 Cores in use: 2 Total, 0 Used  
 Memory in use: 20.0 GiB Total, 0.0 B Used  
 Resources in use:  
 Applications: 0 Running, 1 Completed  
 Drivers: 0 Running, 0 Completed  
 Status: ALIVE

▼ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20201127230731-192.168.1.100-40759	192.168.1.100:40759	ALIVE	1 (0 Used)	10.0 GiB (0.0 B Used)	
worker-20201127230746-192.168.1.76-34791	192.168.1.76:34791	ALIVE	1 (0 Used)	10.0 GiB (0.0 B Used)	

▼ Running Applications (0)

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

▼ Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20201130095609-0000	WordCountExample	2	1024.0 MiB		2020/11/30 09:56:09	hduser	FINISHED	10 s

## 2) Cài đặt Zeppelin

Tải Zeppelin

wget <https://downloads.apache.org/zeppelin/zeppelin-0.9.0-preview2/zeppelin-0.9.0-preview2-bin-all.tgz>

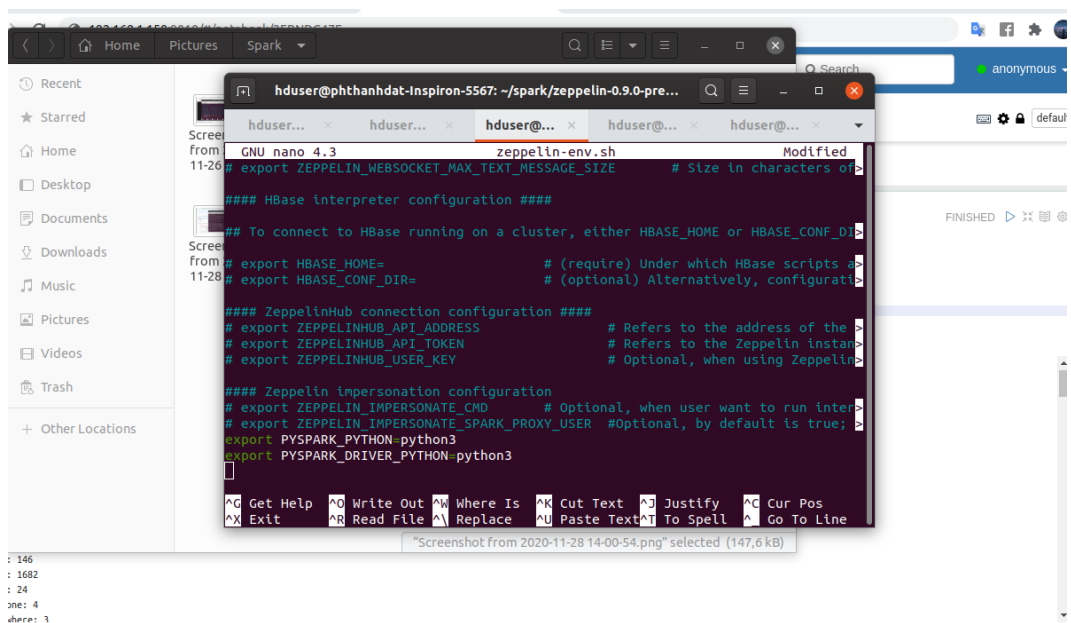
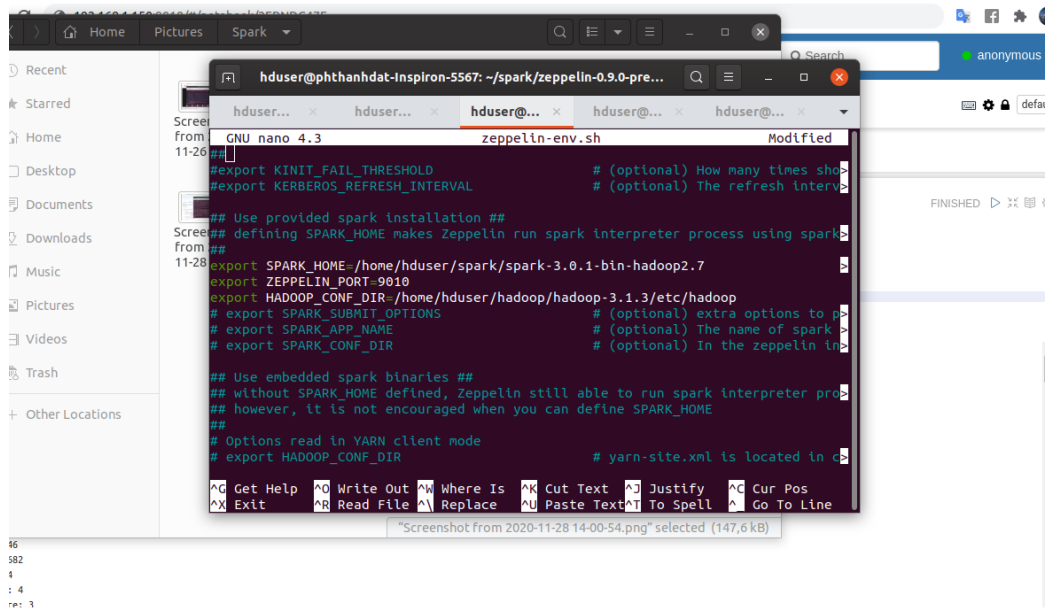
```

GNU nano 4.3 zeppelin-env.sh Modified
# 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.

export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export SPARK_MASTER=spark://node-master:7077 # Spark master url. eg. spark://>
# export ZEPPELIN_ADDR # Bind address (default 127.0.0.1)
# export ZEPPELIN_PORT # port number to listen (default 8443)
# export ZEPPELIN_LOCAL_IP # Zeppelin's thrift server ip address
# export ZEPPELIN_JAVA_OPTS # Additional jvm options. for example: -Xmx1g
# export ZEPPELIN_MEM # Zeppelin jvm mem options Default: -Xmx1g
# export ZEPPELIN_INTP_MEM # zeppelin interpreter process memory
# export ZEPPELIN_INTP_JAVA_OPTS # zeppelin interpreter process jvm options
# export ZEPPELIN_SSL_PORT # ssl port (used when ssl environment is enabled)

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

```



```
big-data x Spark M x 5. Word x Google x Spark Re x Browsin x Spark 'Fi x hadoop x (1) Facet x +
github.com/bk-blockchain/big-data-class/blob/master/docker-compose.yml
76
77 spark_worker_02:
78   container_name: spark_w
79   image: spark_master:late
80   hostname: worker_02
81   build:
82     context: .
83     dockerfile: Dockerfile
84   labels:
85     com.bkwallet.description: spark_worker_02
86     com.bkwallet.maintainer: hdu
87   environment:
88     - ZEPPELIN_PORT=8081
89     - SPARK_MASTER=spark://192.168.1.150:7070
90     - MASTER=spark://master:7070
91     - SPARK_WORKER_CORES=4
92     - SPARK_WORKER_MEMORY=4g
93     - SPARK_DRIVER_MEMORY=4g
94     - SPARK_EXECUTOR_MEMORY=4g
95   healthcheck:
96     disable: true
97   dns:
98     - 8.8.8.8
99     - 8.8.4.4
100   ports:
101     - "8083:8081"
102   logging:
103     driver: json-file
104     options:
105       max-size: "50m"
106       max-file: "2"
107     restart: "always"
108
```

```
hduser@phthanhdat-Inspiron-5567: ~/spark/zeppelin-0.9.0-pre...
hduser... x hduser... x hduser@... x hduser@... x
hduser@phthanhdat-Inspiron-5567: ~/spark/zeppelin-0.9.0-preview2-bin-all$ jps
27065 Jps
12266 SparkSubmit
hduser@phthanhdat-Inspiron-5567: ~/spark/zeppelin-0.9.0-preview2-bin-all$ ls
alice.txt      k8s      notebook  zeppelin-web-0.9.0-preview2.war
bin            lib      NOTICE   zeppelin-web-angular-0.9.0-preview2.war
conf           LICENSE  plugins
holmes.txt     licenses README.md
interpreter    logs     run
hduser@phthanhdat-Inspiron-5567: ~/spark/zeppelin-0.9.0-preview2-bin-all$ ./bin/z
zeppelin-daemon.sh start
Zeppelin start [ OK ]
hduser@phthanhdat-Inspiron-5567: ~/spark/zeppelin-0.9.0-preview2-bin-all$
```

big-data-class/wordcount x Facebook x 5. WordCount - Zeppelin x +

192.168.1.150:9010/#/notebook/2FRNDC47F

**Zeppelin** Notebook Job Search anonymous

## 5. WordCount

Finished

```
%sh
id
pwd

uid=1002(hduser) gid=1002(hadoop) groups=1002(hadoop),27(sudo)
/home/hduser/spark/zeppelin-0.9.0-preview2-bin-all

Took 0 sec. Last updated by anonymous at November 27 2020, 10:45:13 PM.
```

Finished

```
%python
import sys
from pyspark import SparkContext
sc = SparkContext(appName="WordCountExample")
lines = sc.textFile("hdfs://192.168.1.150:9000/books/alice.txt")
counts = lines.flatMap(lambda x: x.split(' ')) \
               .map(lambda x: (x, 1)) \
               .reduceByKey(lambda x,y:x+y)
output = counts.collect()
for (word, count) in output:
    print("%s: %i" % (word, count))
sc.stop()
```

The: 107  
Project: 79  
Gutenberg: 21  
EBook: 1  
of: 604  
Alice's: 14  
Adventures: 5  
in: 488  
Wonderland,: 3  
by: 81  
Lewis: 4  
Carroll: 4  
+ 14234