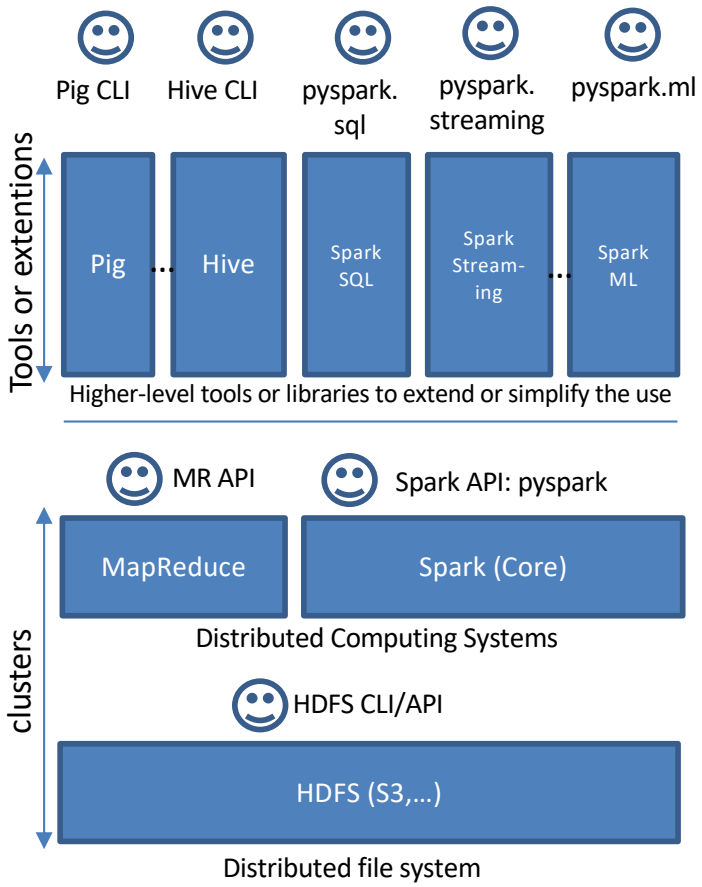
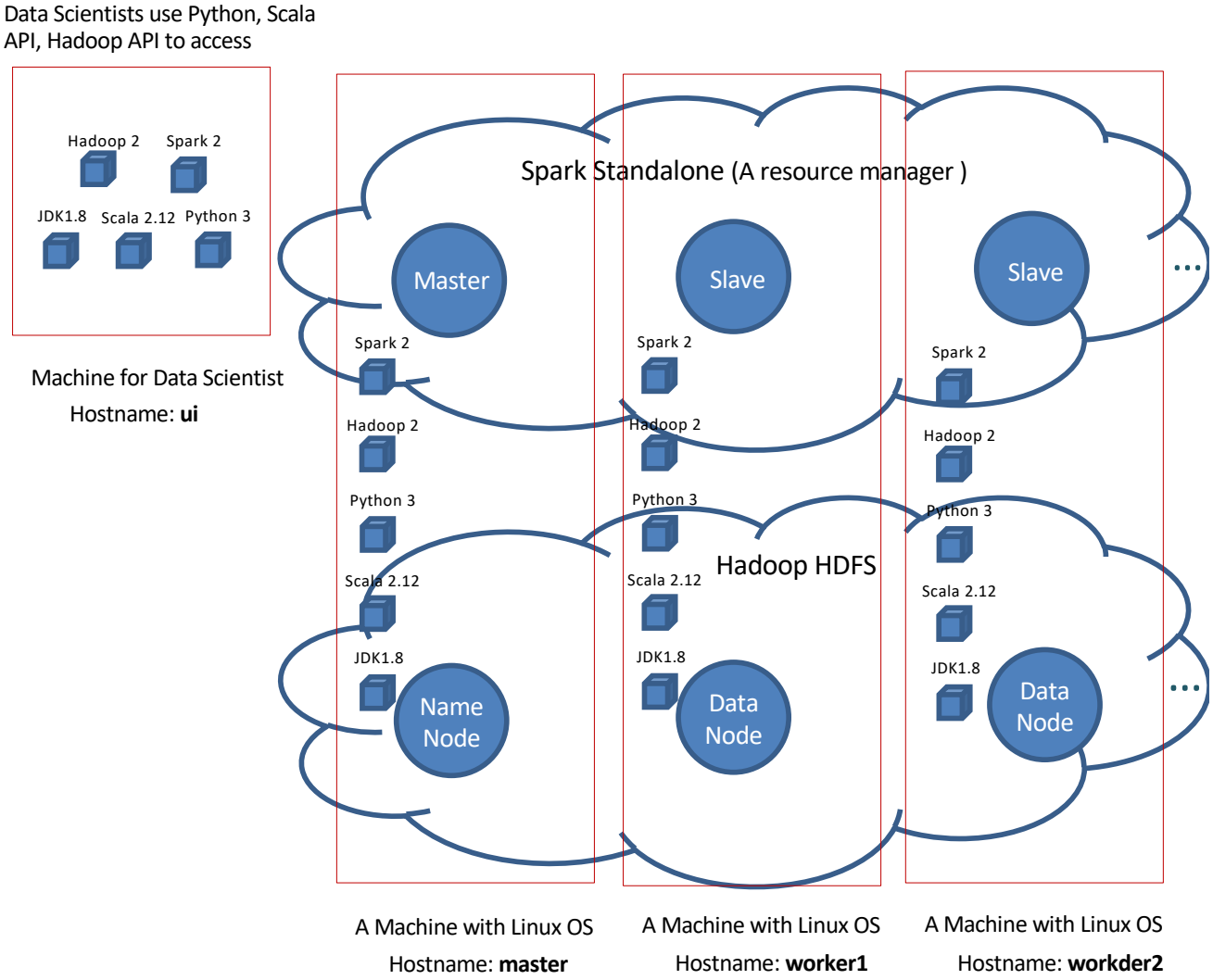
## Big Data Eco-System (Partial)

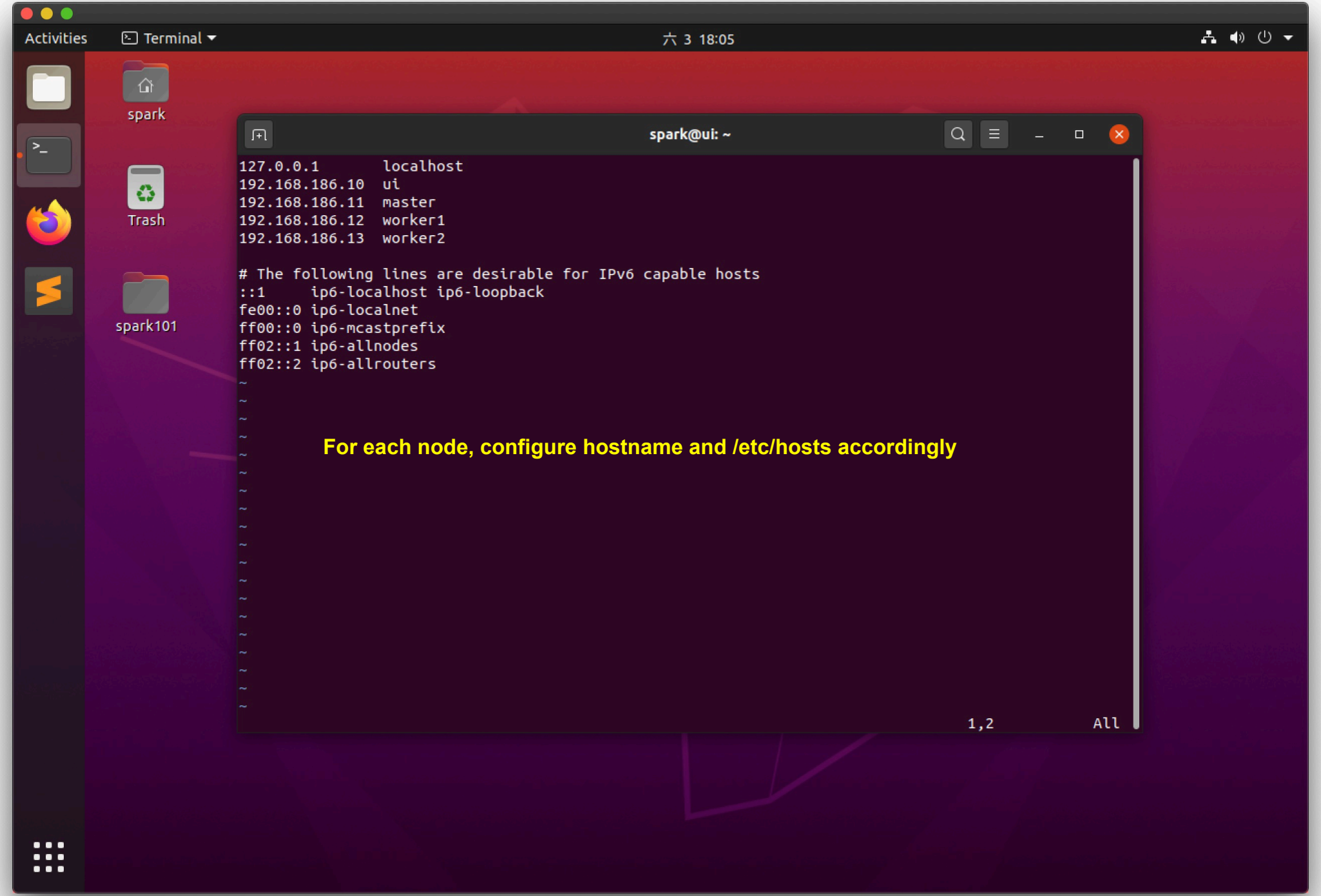


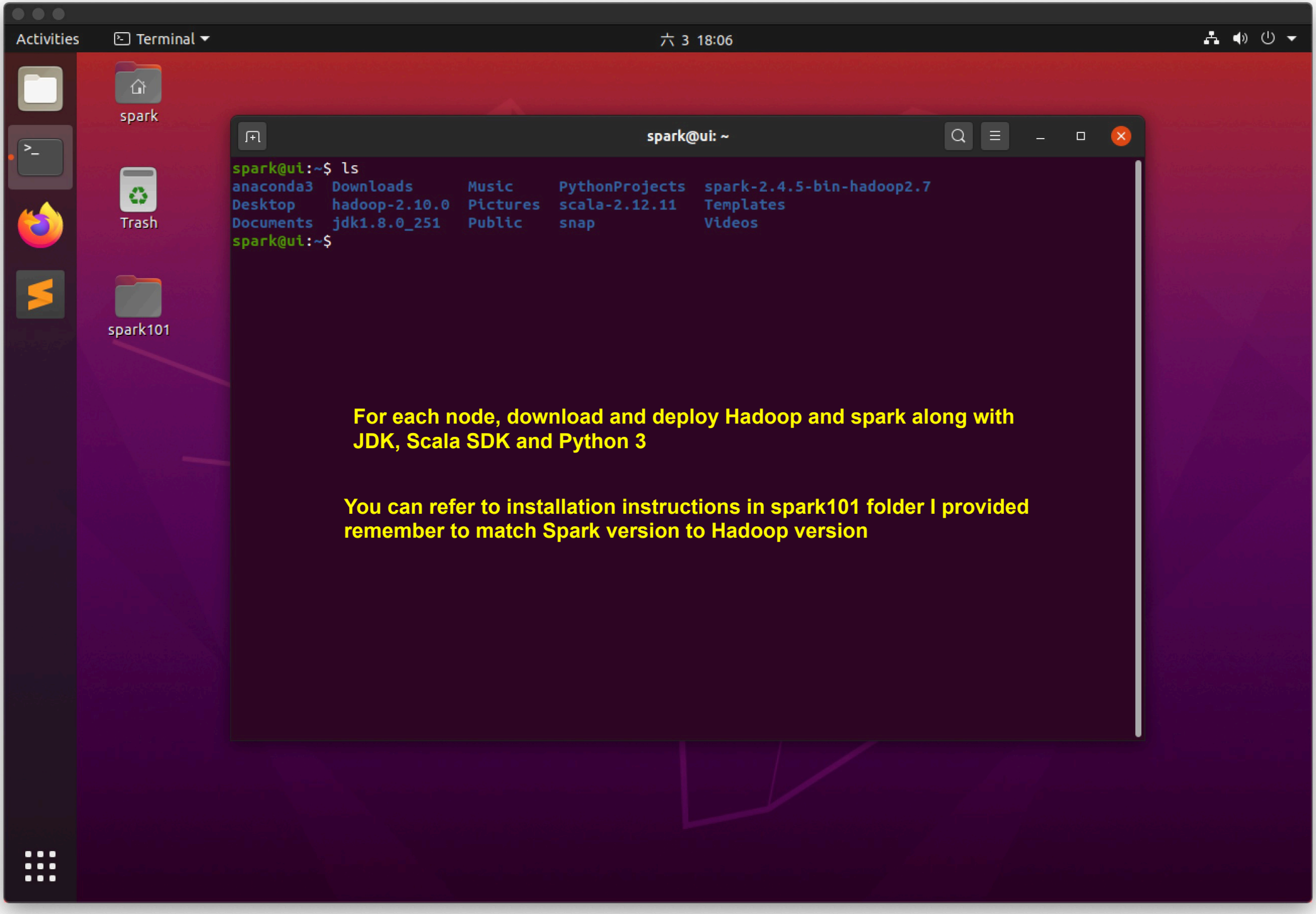

## Production Environment Setup

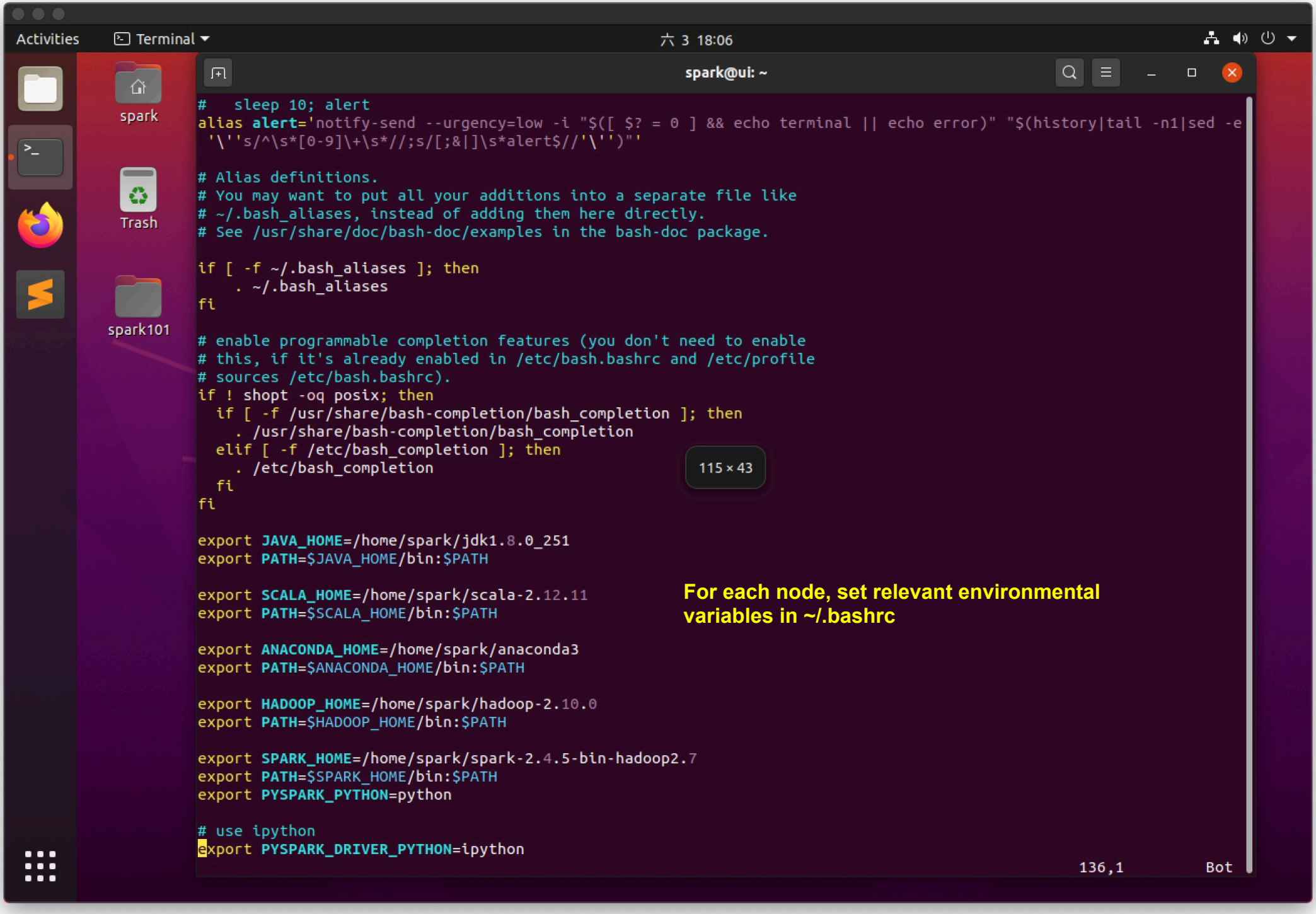


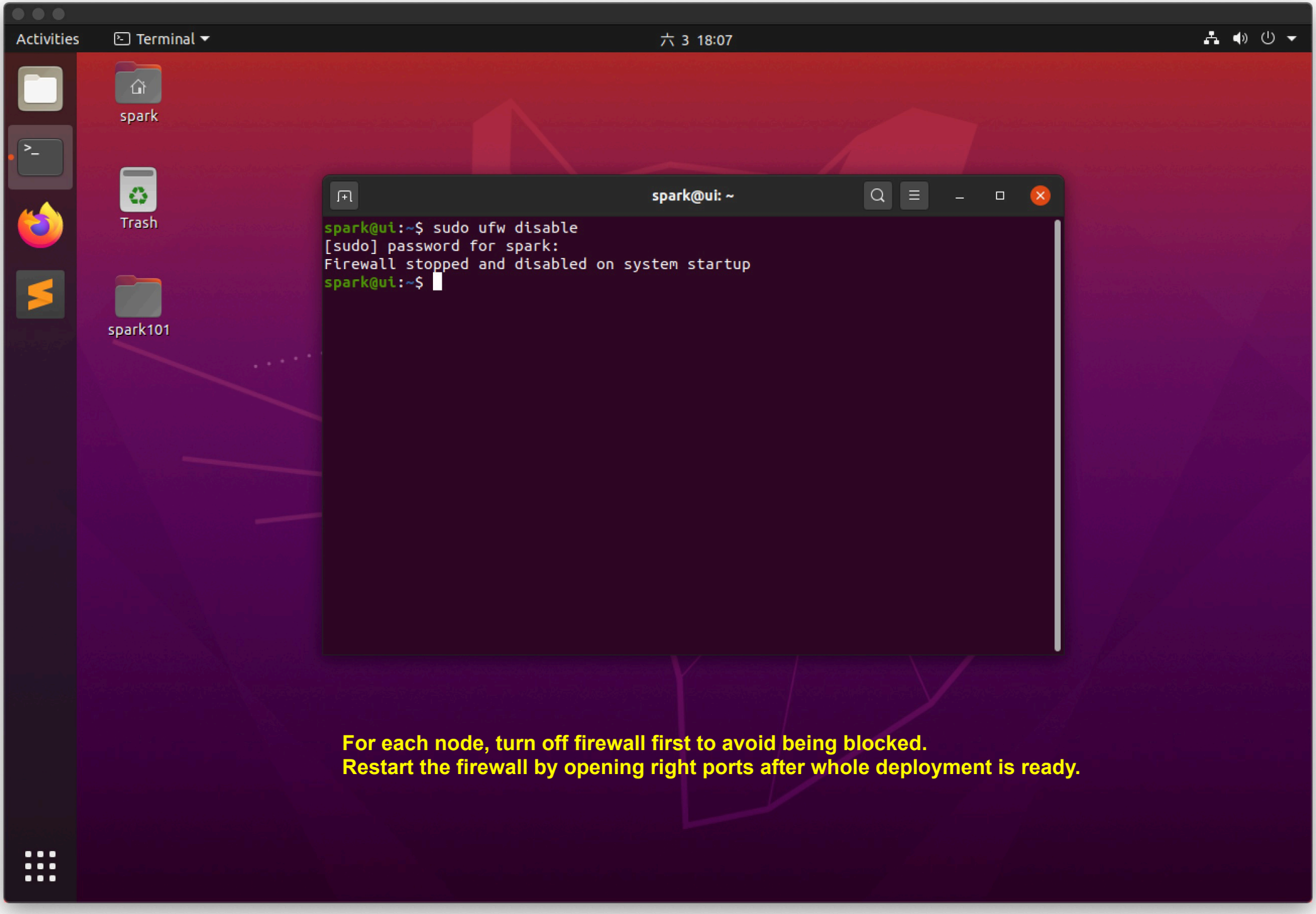

Note that:

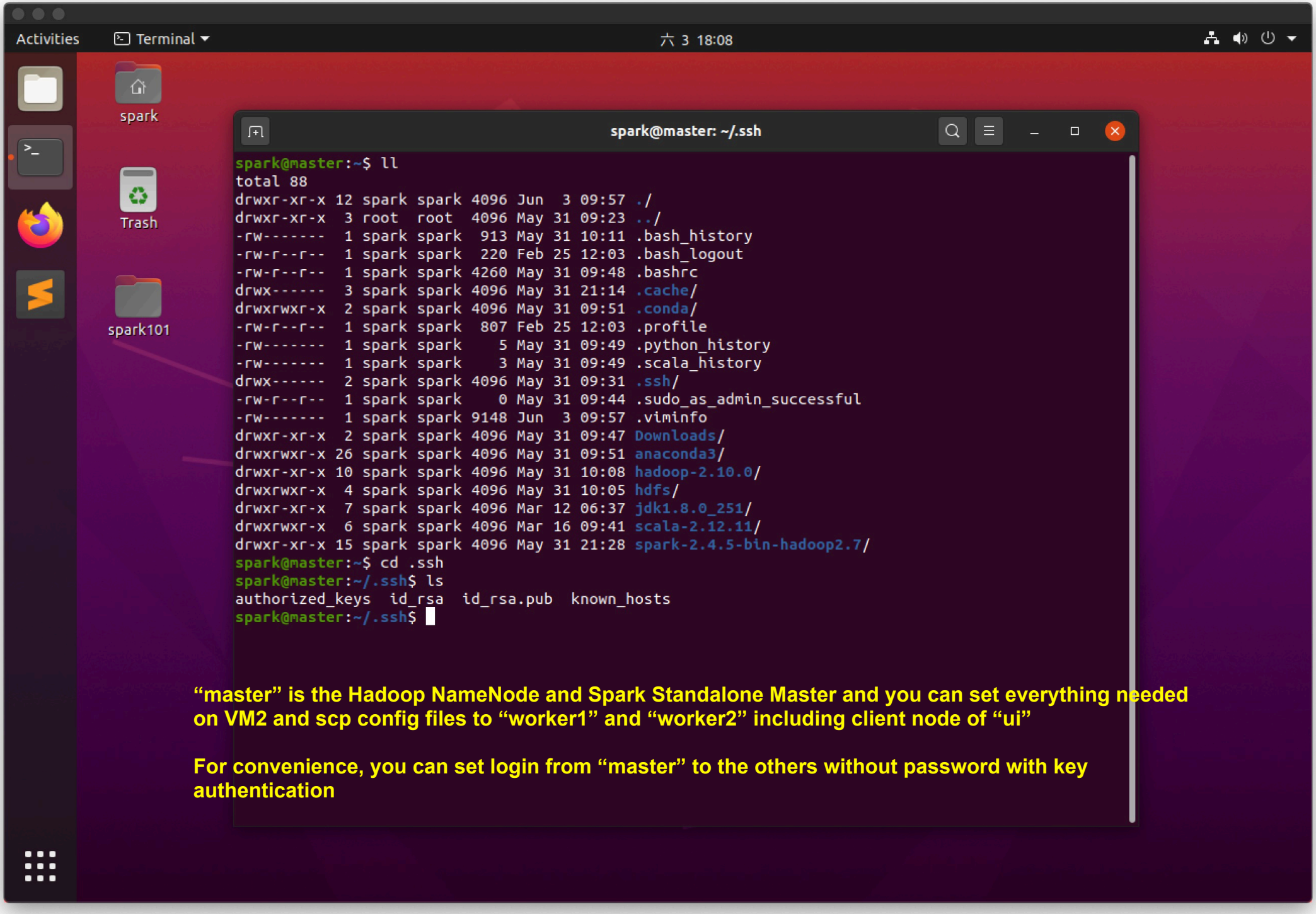
1. Make sure each machine in Spark/HDFS cluster has static IP address
2. Make sure the /etc/hosts has records for all nodes and synchronized on each node
3. The firewall on each node need to allow access to required ports
   (can turn off the firewall temporarily during the setup and put it on after the test is done)

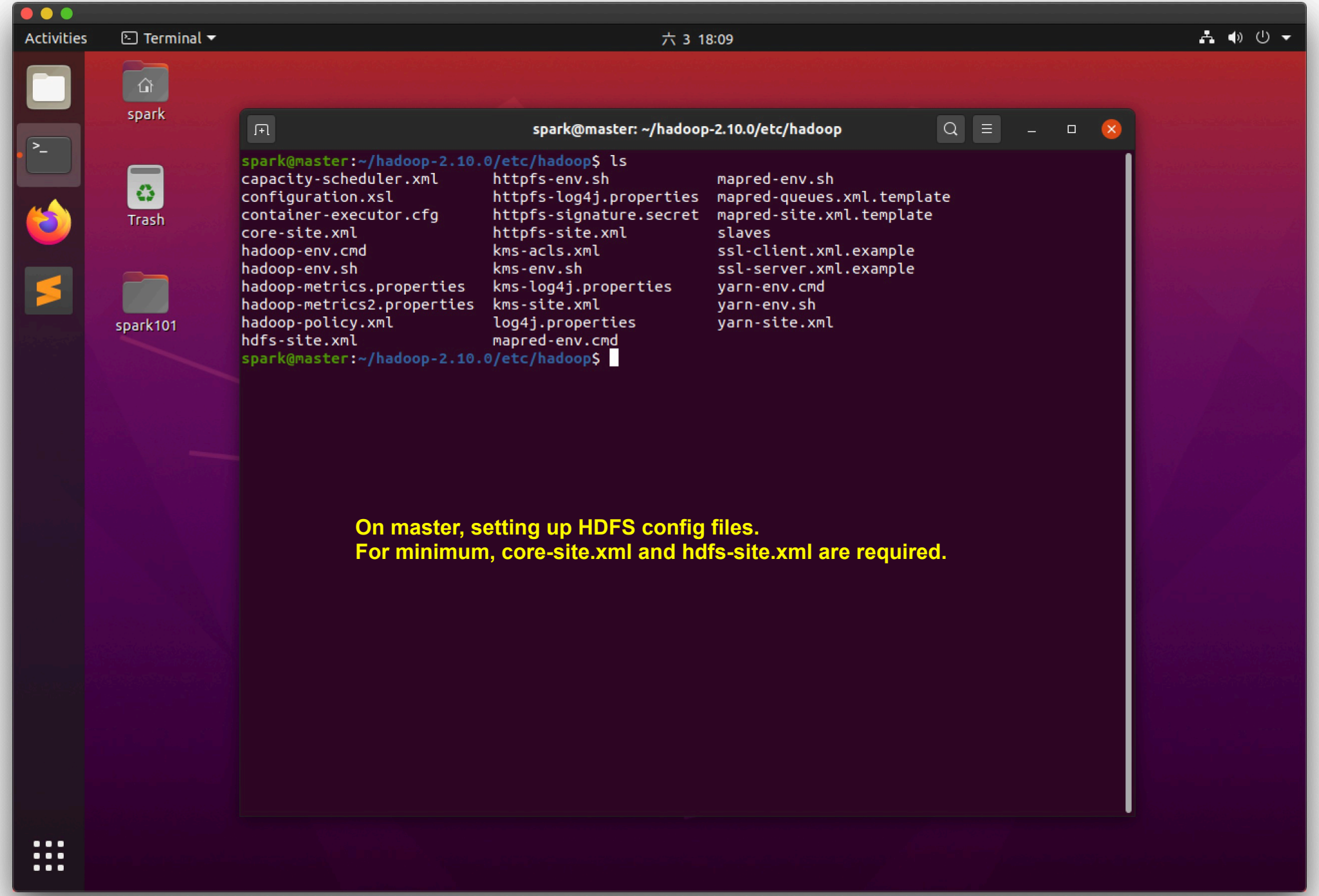

Activities
Terminal
六 3 18:05
spark
Trash
spark101
spark@ui: ~
127.0.0.1 localhost
192.168.186.10 ui
192.168.186.11 master
192.168.186.12 worker1
192.168.186.13 worker2
# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
1,2 All
For each node, configure hostname and /etc/hosts accordingly

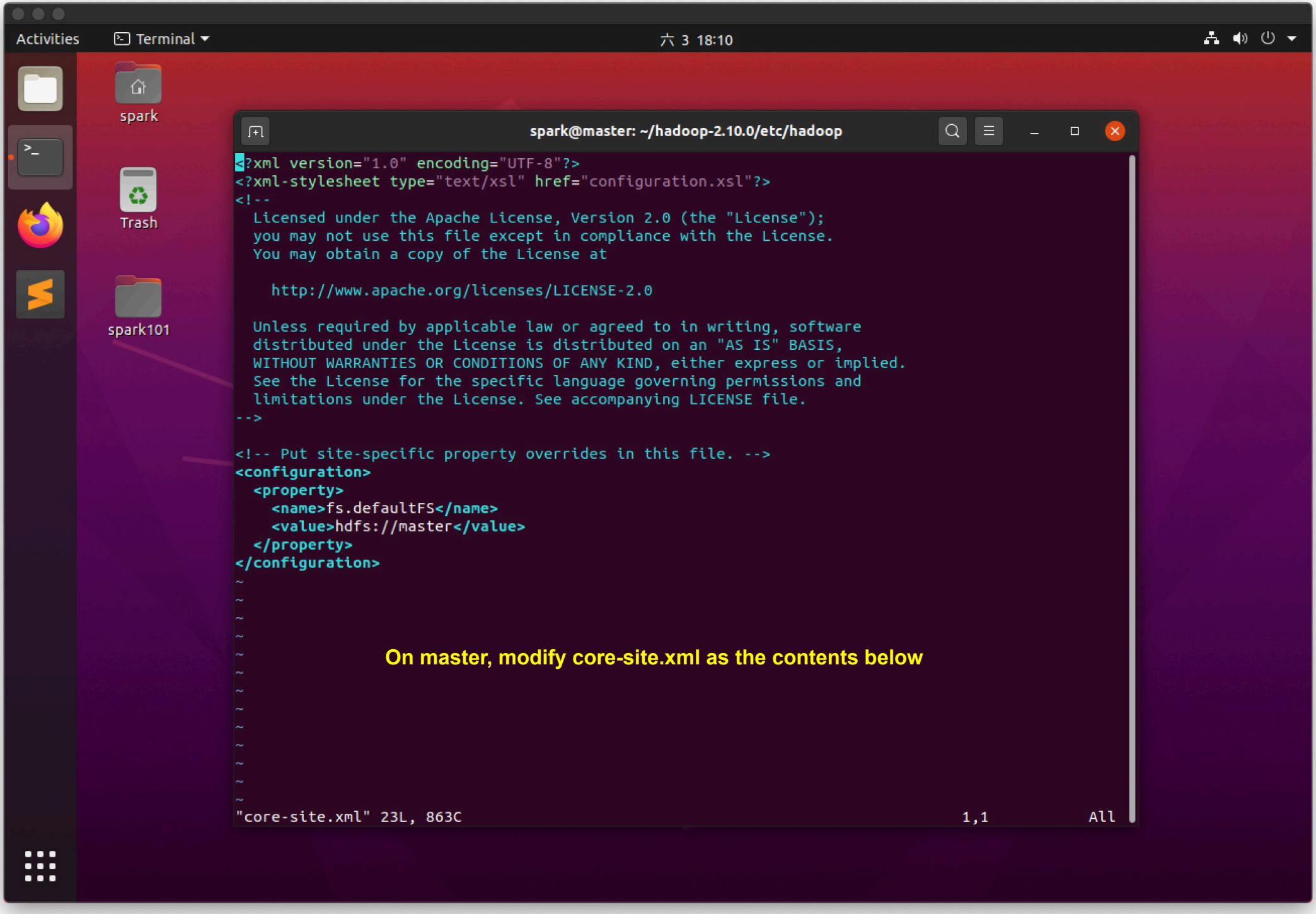

Activities
Terminal
六 3 18:06
spark
Trash
spark101
spark@ui: ~
spark@ui:~$ ls
anaconda3 Downloads Music PythonProjects spark-2.4.5-bin-hadoop2.7
Desktop hadoop-2.10.0 Pictures scala-2.12.11 Templates
Documents jdk1.8.0_251 Public snap Videos
spark@ui:~$
For each node, download and deploy Hadoop and spark along with JDK, Scala SDK and Python 3
You can refer to installation instructions in spark101 folder I provided remember to match Spark version to Hadoop version

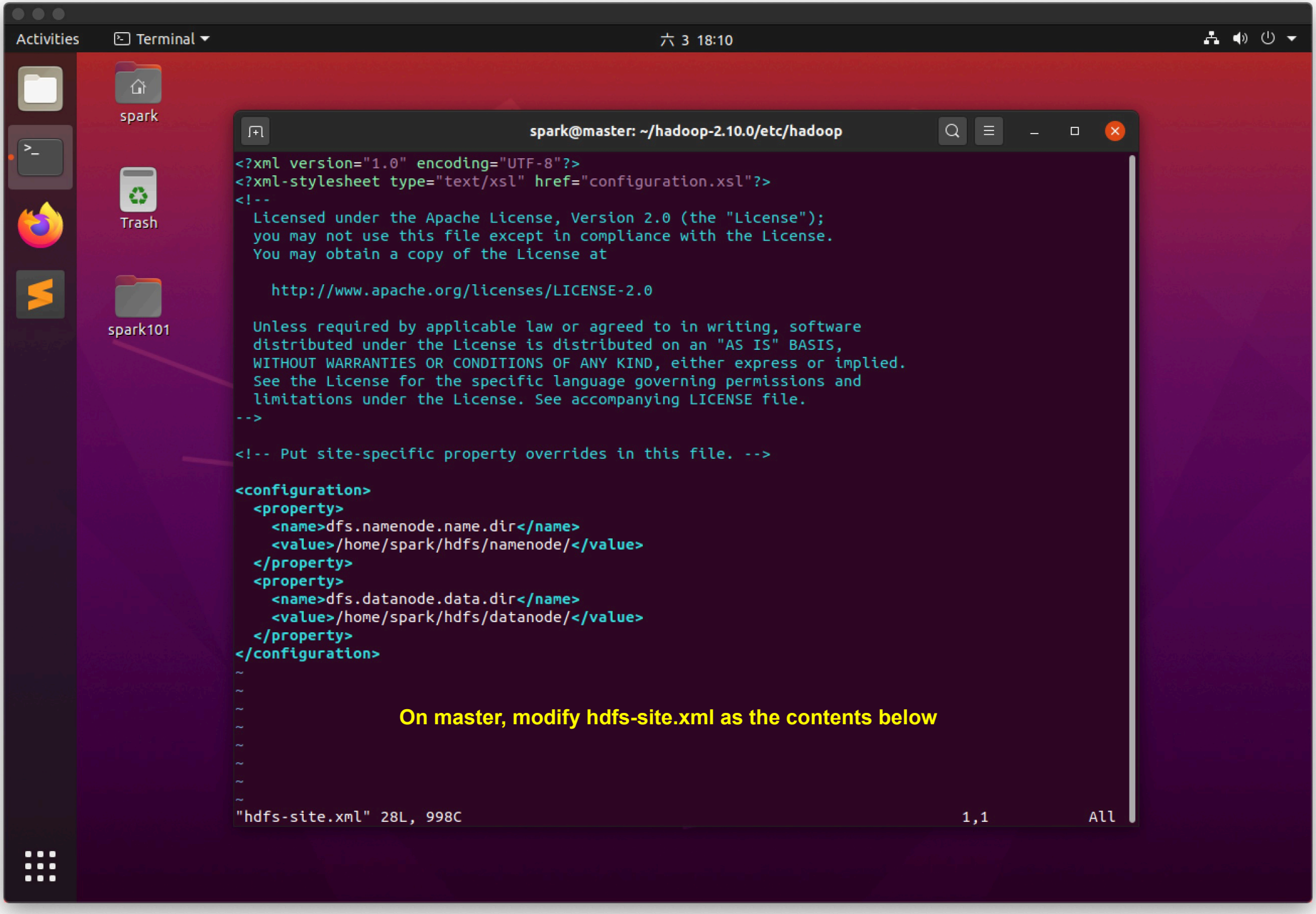

Activities
Terminal
六 3 18:06
spark
Trash
spark101
spark@ui: ~
# sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || echo error)" "$(history|tail -n1|sed -e
'\''s/^\s*[0-9]\+\s*//;s/[;&|]\s*alert$//'\'')"'
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.
if [ -f ~/.bash_aliases ]; then
. ~/.bash_aliases
fi
# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
if [ -f /usr/share/bash-completion/bash_completion ]; then
. /usr/share/bash-completion/bash_completion
elif [ -f /etc/bash_completion ]; then
. /etc/bash_completion
fi
fi
115 × 43
export JAVA_HOME=/home/spark/jdk1.8.0_251
export PATH=$JAVA_HOME/bin:$PATH
export SCALA_HOME=/home/spark/scala-2.12.11
export PATH=$SCALA_HOME/bin:$PATH
export ANACONDA_HOME=/home/spark/anaconda3
export PATH=$ANACONDA_HOME/bin:$PATH
export HADOOP_HOME=/home/spark/hadoop-2.10.0
export PATH=$HADOOP_HOME/bin:$PATH
export SPARK_HOME=/home/spark/spark-2.4.5-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH
export PYSPARK_PYTHON=python
# use ipython
export PYSPARK_DRIVER_PYTHON=ipython
136,1 Bot
For each node, set relevant environmental variables in ~/.bashrc

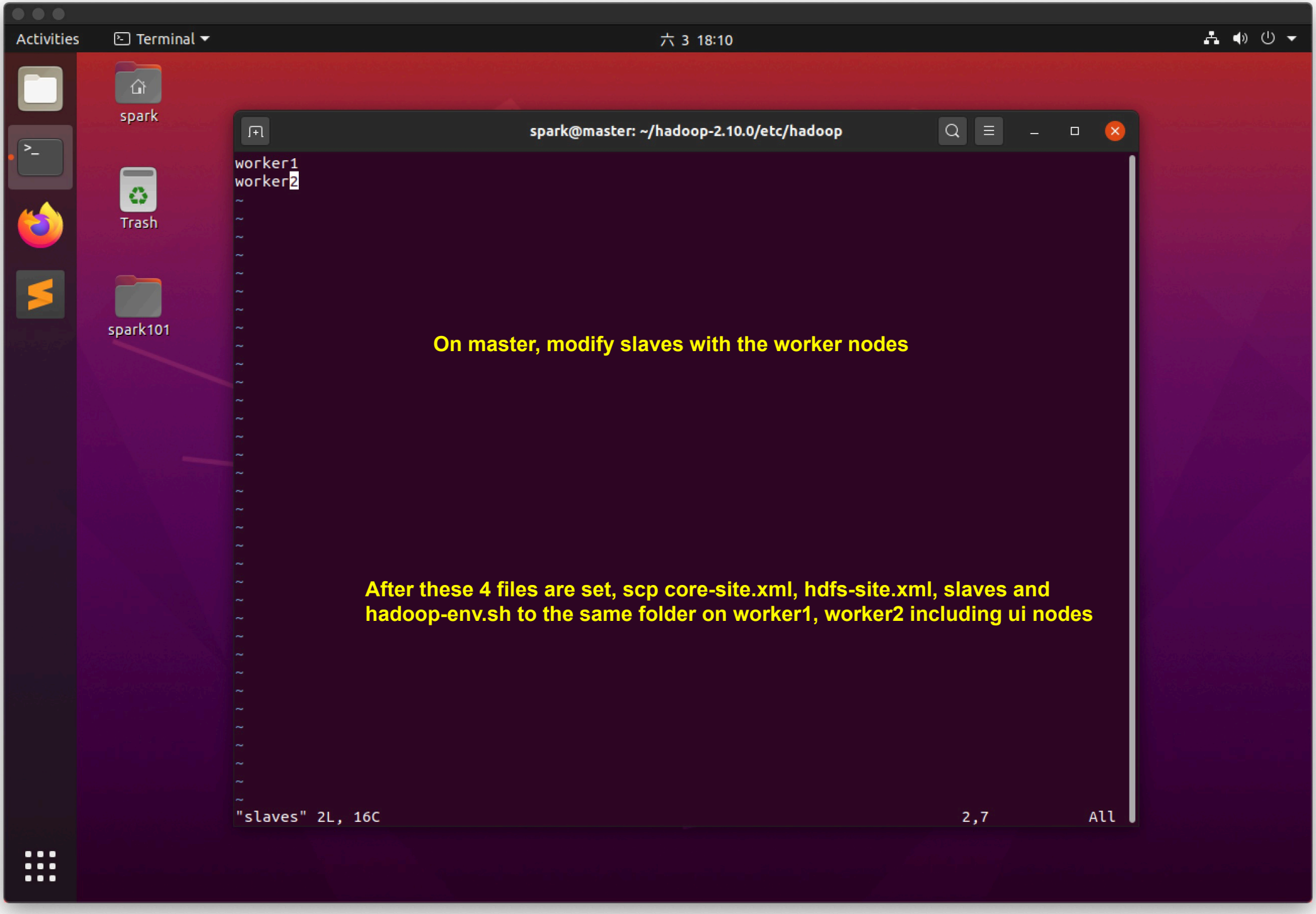

Activities
Terminal
六 3 18:07
spark
Trash
spark101
spark@ui: ~
spark@ui:~$ sudo ufw disable
[sudo] password for spark:
Firewall stopped and disabled on system startup
spark@ui:~$
For each node, turn off firewall first to avoid being blocked.
Restart the firewall by opening right ports after whole deployment is ready.

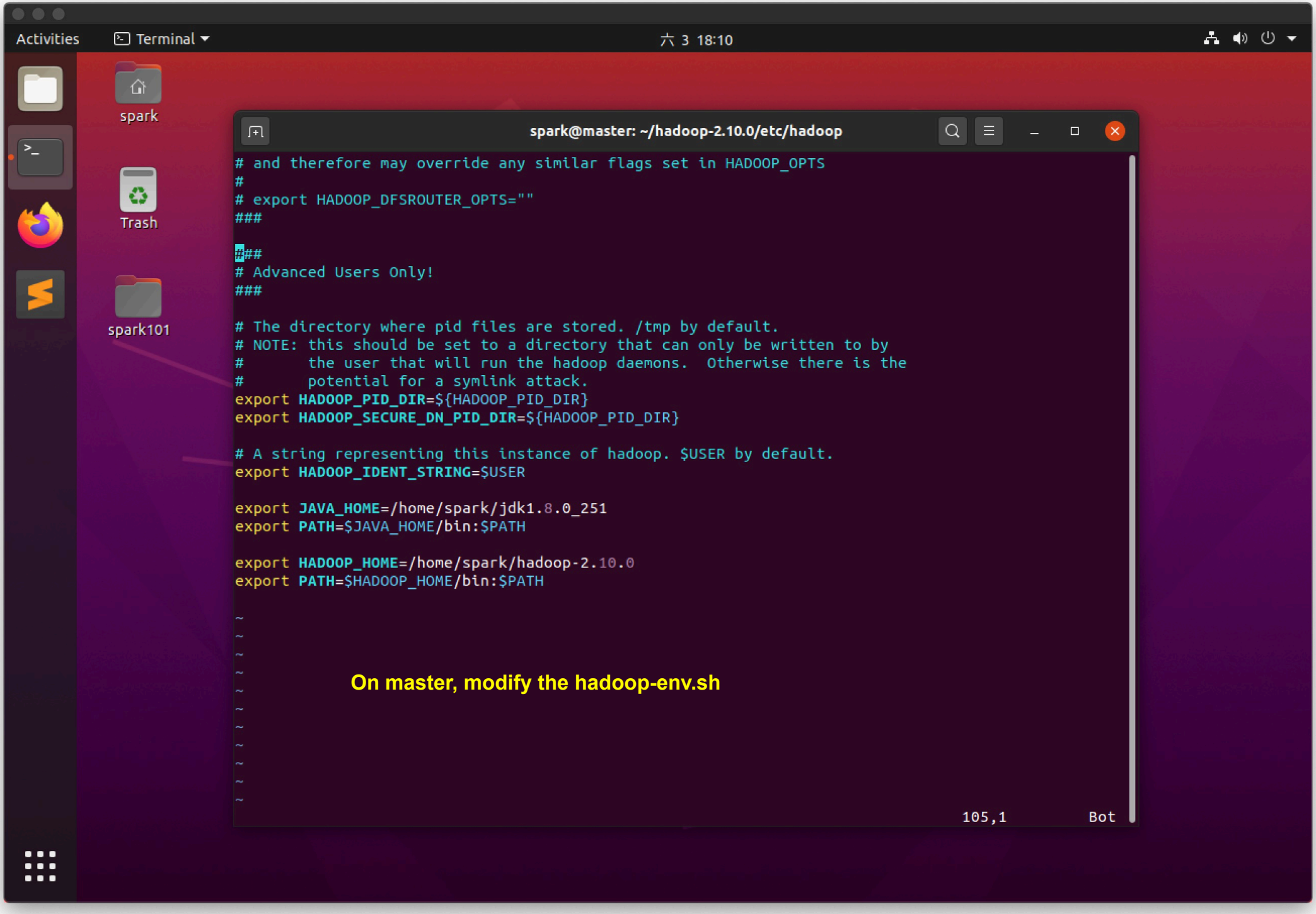

Activities
Terminal
六 3 18:08
spark
Trash
spark101
spark@master: ~/.ssh
spark@master:~$ ll
total 88
drwxr-xr-x 12 spark spark 4096 Jun  3 09:57 ./
drwxr-xr-x  3 root  root  4096 May 31 09:23 ../
-rw-------  1 spark spark  913 May 31 10:11 .bash_history
-rw-r--r--  1 spark spark  220 Feb 25 12:03 .bash_logout
-rw-r--r--  1 spark spark 4260 May 31 09:48 .bashrc
drwx------  3 spark spark 4096 May 31 21:14 .cache/
drwxrwxr-x  2 spark spark 4096 May 31 09:51 .conda/
-rw-r--r--  1 spark spark  807 Feb 25 12:03 .profile
-rw-------  1 spark spark    5 May 31 09:49 .python_history
-rw-------  1 spark spark    3 May 31 09:49 .scala_history
drwx------  2 spark spark 4096 May 31 09:31 .ssh/
-rw-r--r--  1 spark spark    0 May 31 09:44 .sudo_as_admin_successful
-rw-------  1 spark spark 9148 Jun  3 09:57 .viminfo
drwxr-xr-x  2 spark spark 4096 May 31 09:47 Downloads/
drwxrwxr-x 26 spark spark 4096 May 31 09:51 anaconda3/
drwxr-xr-x 10 spark spark 4096 May 31 10:08 hadoop-2.10.0/
drwxrwxr-x  4 spark spark 4096 May 31 10:05 hdfs/
drwxr-xr-x  7 spark spark 4096 Mar 12 06:37 jdk1.8.0_251/
drwxrwxr-x  6 spark spark 4096 Mar 16 09:41 scala-2.12.11/
drwxr-xr-x 15 spark spark 4096 May 31 21:28 spark-2.4.5-bin-hadoop2.7/
spark@master:~$ cd .ssh
spark@master:~/.ssh$ ls
authorized_keys  id_rsa  id_rsa.pub  known_hosts
spark@master:~/.ssh$
“master” is the Hadoop NameNode and Spark Standalone Master and you can set everything needed on VM2 and scp config files to “worker1” and “worker2” including client node of “ui”
For convenience, you can set login from “master” to the others without password with key authentication

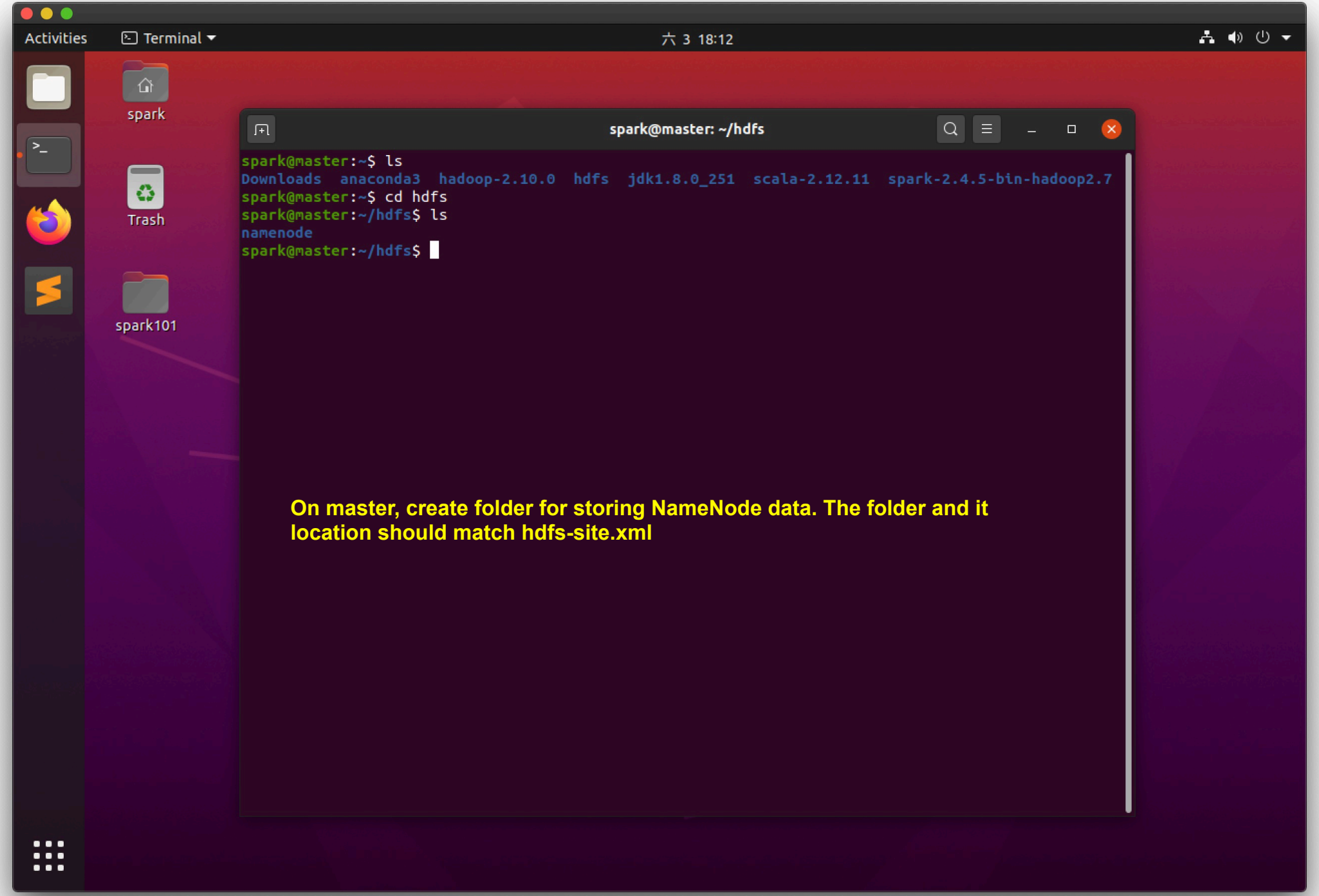

Activities
Terminal
六 3 18:09
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/etc/hadoop
spark@master:~/hadoop-2.10.0/etc/hadoop$ ls
capacity-scheduler.xml httpfs-env.sh mapred-env.sh
configuration.xsl httpfs-log4j.properties mapred-queues.xml.template
container-executor.cfg httpfs-signature.secret mapred-site.xml.template
core-site.xml httpfs-site.xml slaves
hadoop-env.cmd kms-acls.xml ssl-client.xml.example
hadoop-env.sh kms-env.sh ssl-server.xml.example
hadoop-metrics.properties kms-log4j.properties yarn-env.cmd
hadoop-metrics2.properties kms-site.xml yarn-env.sh
hadoop-policy.xml log4j.properties yarn-site.xml
hdfs-site.xml mapred-env.cmd
spark@master:~/hadoop-2.10.0/etc/hadoop$
On master, setting up HDFS config files.
For minimum, core-site.xml and hdfs-site.xml are required.

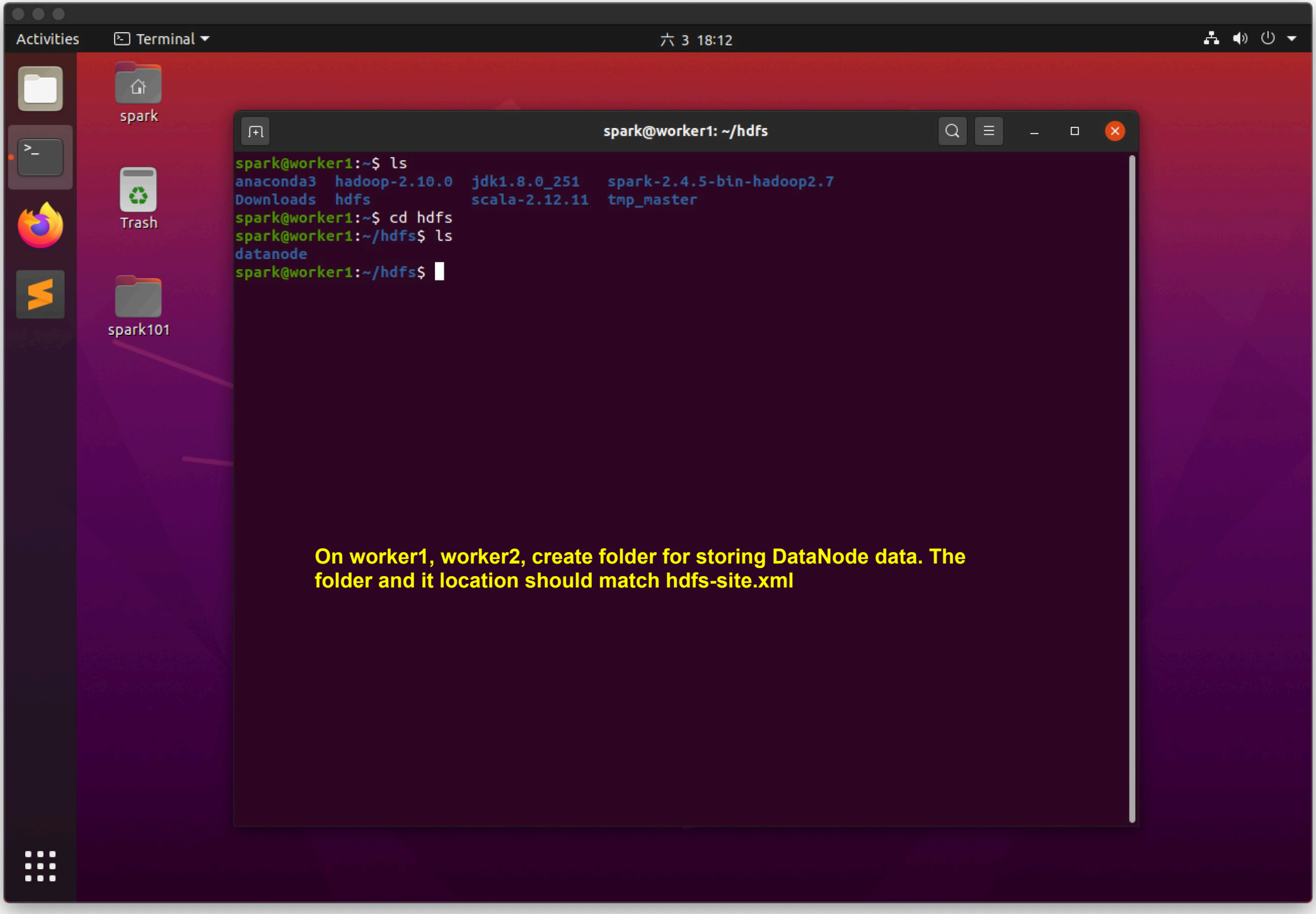

Activities
Terminal
六 3 18:10
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/etc/hadoop
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance 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. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://master</value>
</property>
</configuration>
"core-site.xml" 23L, 863C
1,1
All
On master, modify core-site.xml as the contents below

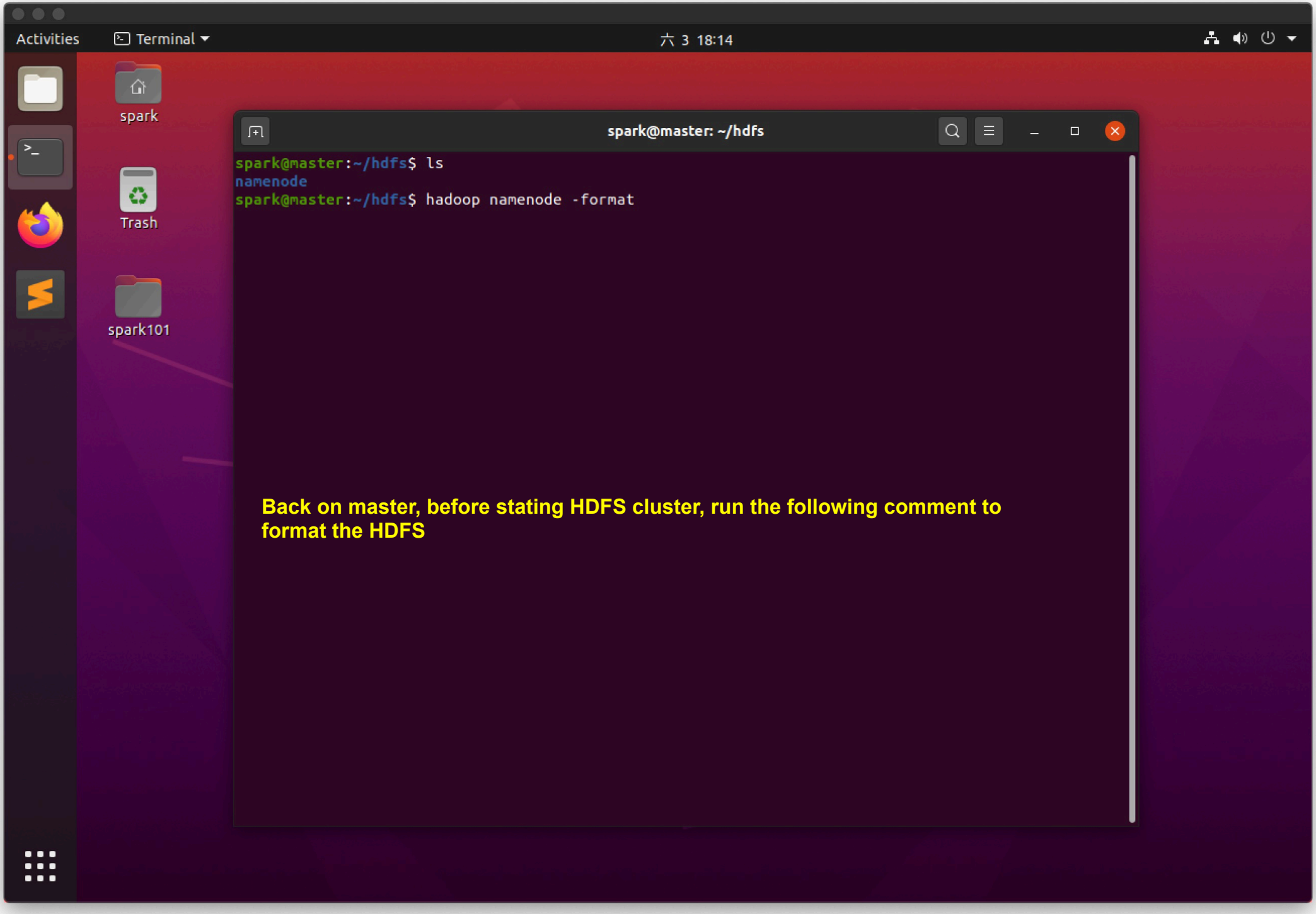

Activities
Terminal
六 3 18:10
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/etc/hadoop
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance 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. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/spark/hdfs/namenode/</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/spark/hdfs/datanode/</value>
</property>
</configuration>
On master, modify hdfs-site.xml as the contents below
"hdfs-site.xml" 28L, 998C
1,1
All

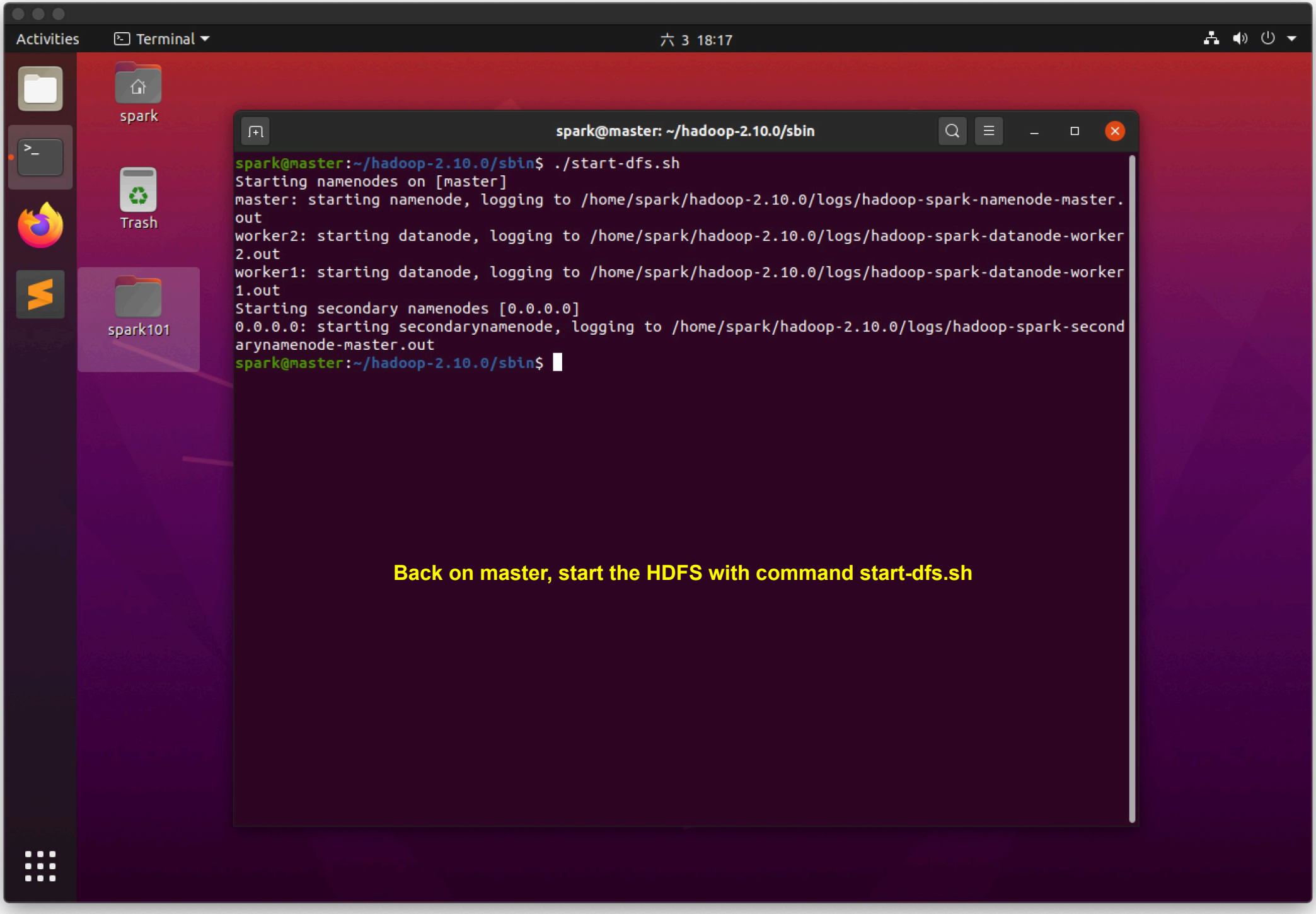

Activities
Terminal
六 3 18:10
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/etc/hadoop
worker1
worker2
On master, modify slaves with the worker nodes
After these 4 files are set, scp core-site.xml, hdfs-site.xml, slaves and hadoop-env.sh to the same folder on worker1, worker2 including ui nodes
"slaves" 2L, 16C
2,7
All

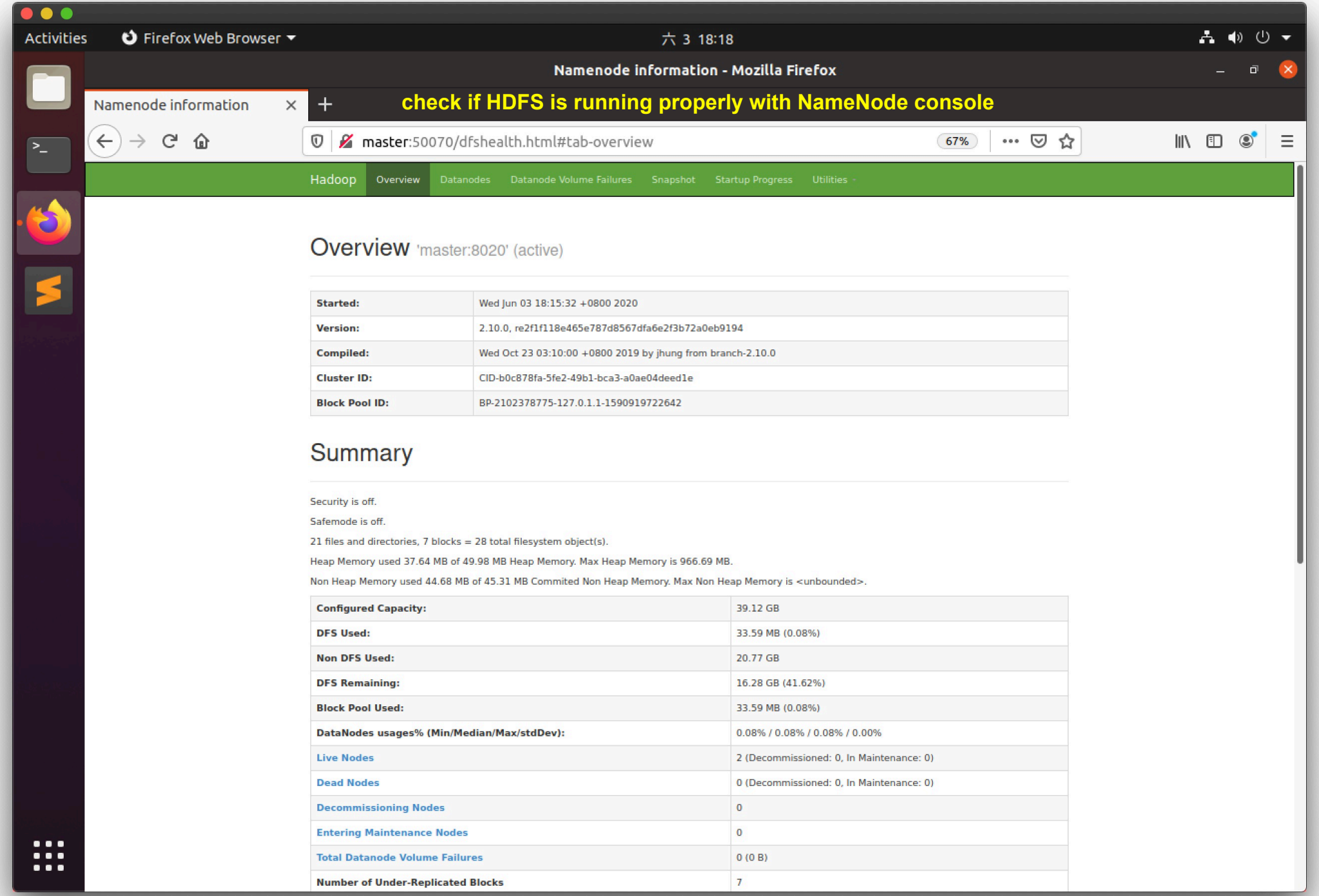

Activities
Terminal
六 3 18:10
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/etc/hadoop
# and therefore may override any similar flags set in HADOOP_OPTS
#
# export HADOOP_DFSROUTER_OPTS=""
###
##
# Advanced Users Only!
###
# The directory where pid files are stored. /tmp by default.
# NOTE: this should be set to a directory that can only be written to by
#       the user that will run the hadoop daemons.  Otherwise there is the
#       potential for a symlink attack.
export HADOOP_PID_DIR=${HADOOP_PID_DIR}
export HADOOP_SECURE_DN_PID_DIR=${HADOOP_PID_DIR}
# A string representing this instance of hadoop. $USER by default.
export HADOOP_IDENT_STRING=$USER
export JAVA_HOME=/home/spark/jdk1.8.0_251
export PATH=$JAVA_HOME/bin:$PATH
export HADOOP_HOME=/home/spark/hadoop-2.10.0
export PATH=$HADOOP_HOME/bin:$PATH
105,1 Bot
On master, modify the hadoop-env.sh

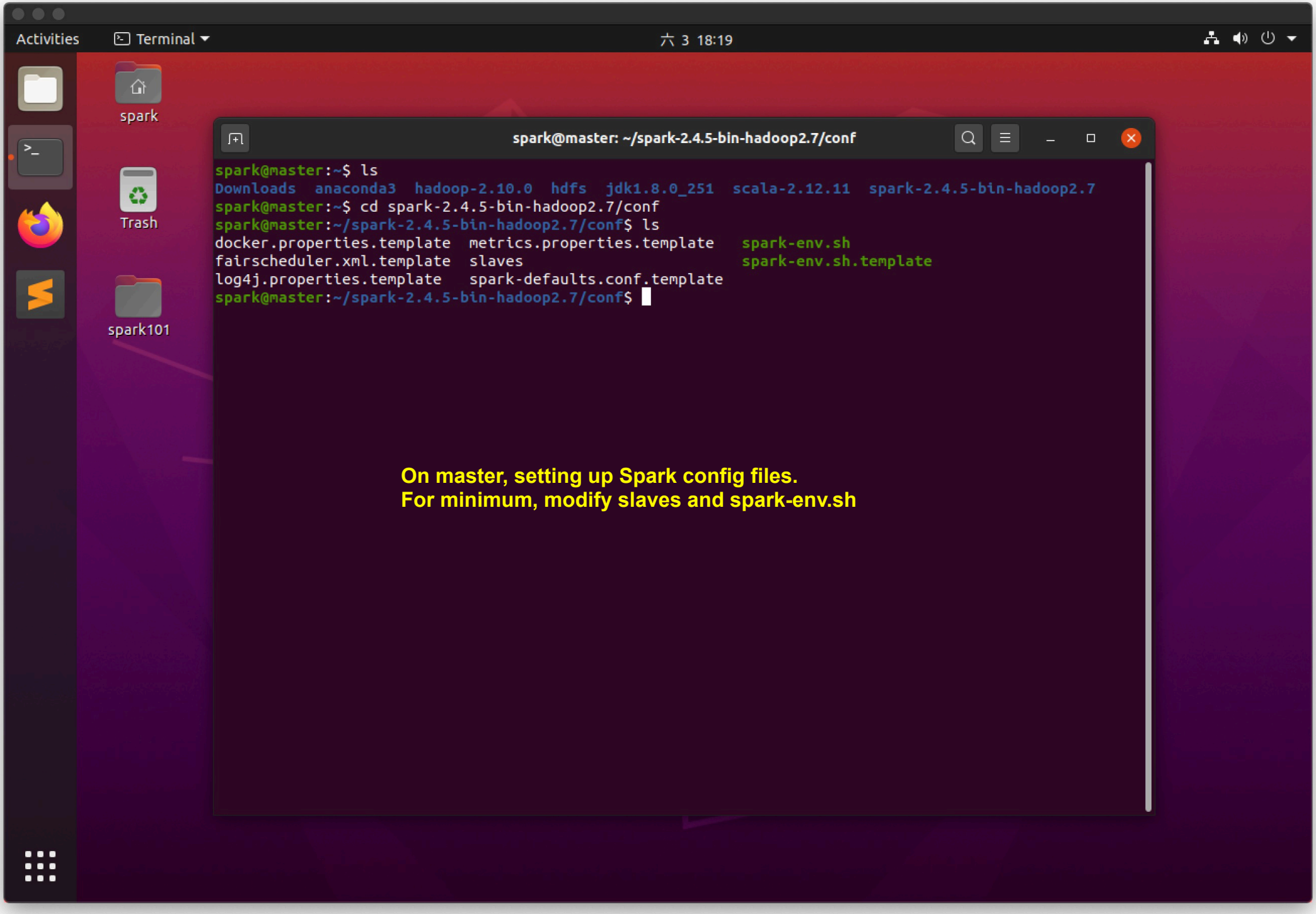

Activities
Terminal
六 3 18:12
spark
Trash
spark101
spark@master: ~/hdfs
spark@master:~$ ls
Downloads anaconda3 hadoop-2.10.0 hdfs jdk1.8.0_251 scala-2.12.11 spark-2.4.5-bin-hadoop2.7
spark@master:~$ cd hdfs
spark@master:~/hdfs$ ls
namenode
spark@master:~/hdfs$
On master, create folder for storing NameNode data. The folder and it location should match hdfs-site.xml

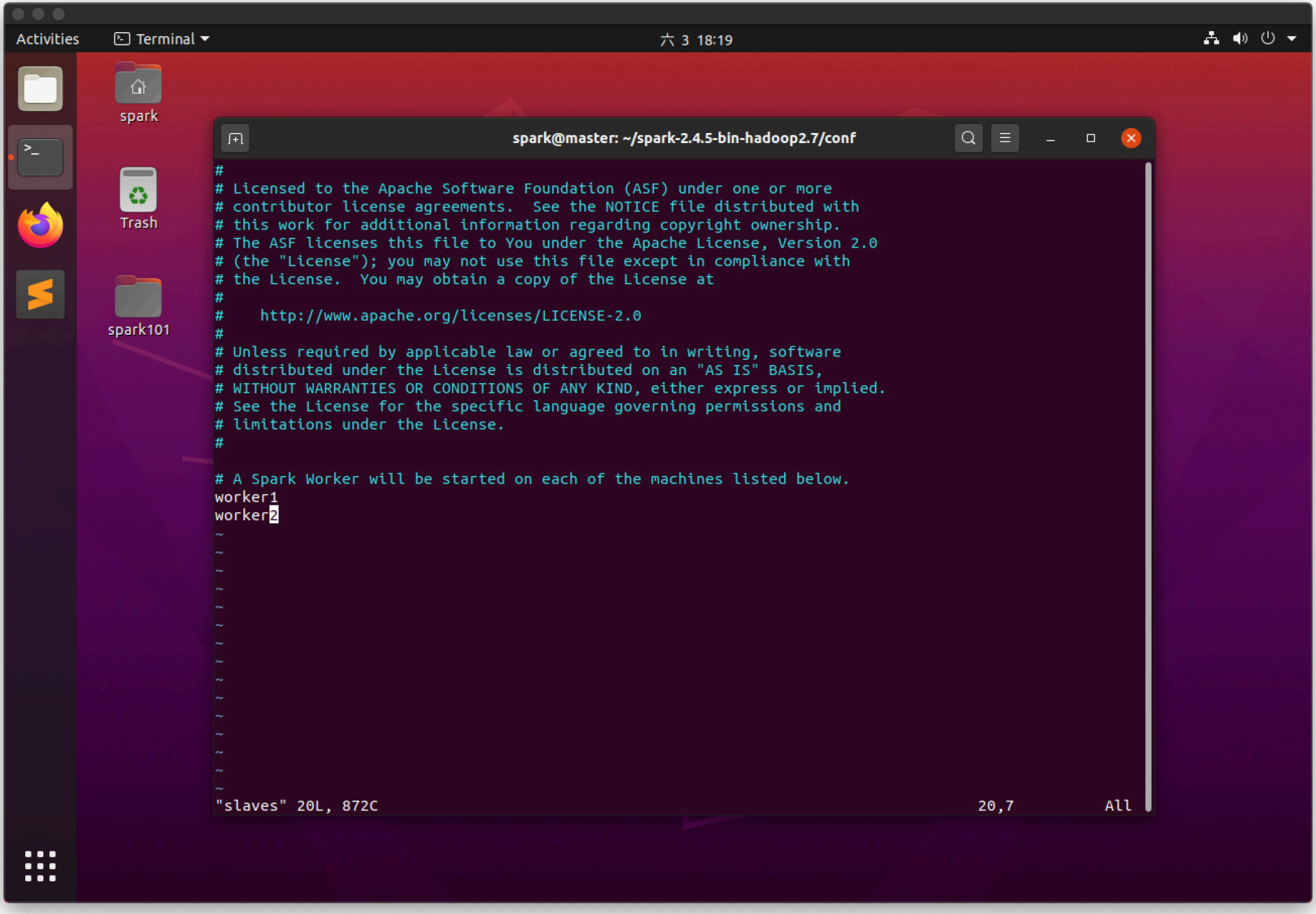

Activities
Terminal
六 3 18:12
spark
Trash
spark101
spark@worker1: ~/hdfs
spark@worker1:~$ ls
anaconda3 hadoop-2.10.0 jdk1.8.0_251 spark-2.4.5-bin-hadoop2.7
Downloads hdfs scala-2.12.11 tmp_master
spark@worker1:~$ cd hdfs
spark@worker1:~/hdfs$ ls
datanode
spark@worker1:~/hdfs$
On worker1, worker2, create folder for storing DataNode data. The folder and it location should match hdfs-site.xml

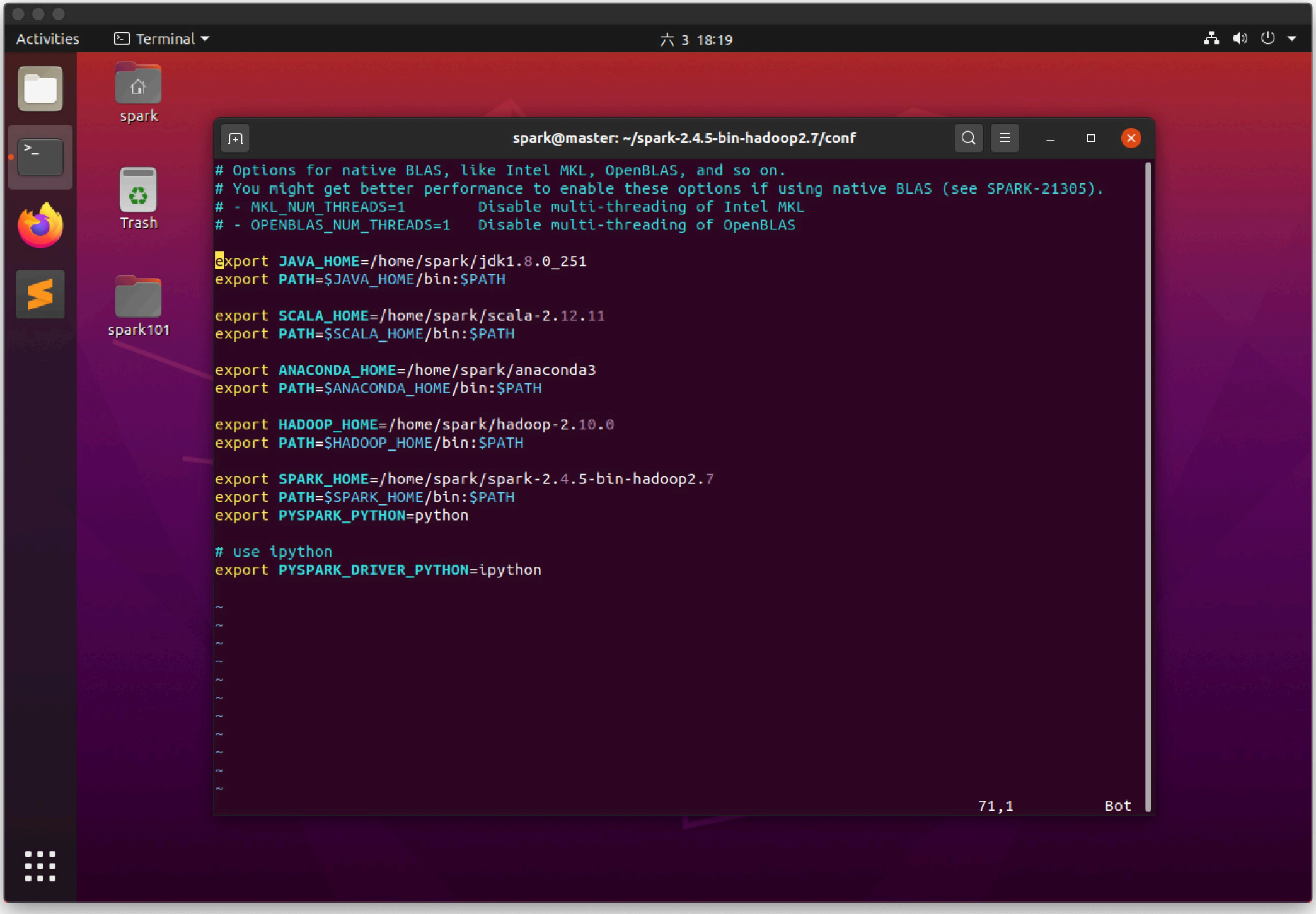

Activities
Terminal
六 3 18:14
spark
Trash
spark101
spark@master: ~/hdfs
spark@master:~/hdfs$ ls
namenode
spark@master:~/hdfs$ hadoop namenode -format
Back on master, before stating HDFS cluster, run the following comment to format the HDFS

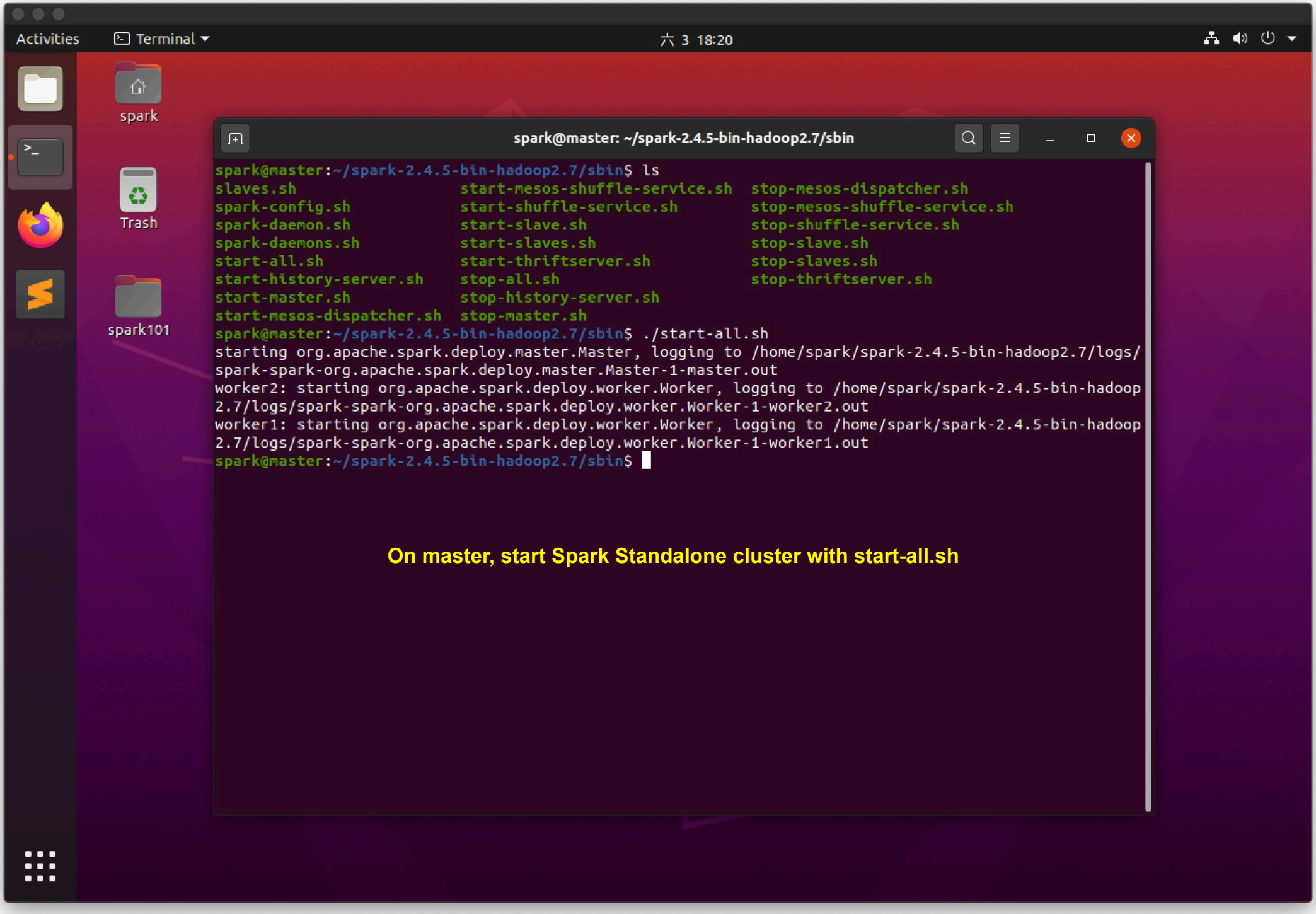

Activities
Terminal
六 3 18:17
spark
Trash
spark101
spark@master: ~/hadoop-2.10.0/sbin
spark@master:~/hadoop-2.10.0/sbin$ ./start-dfs.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/spark/hadoop-2.10.0/logs/hadoop-spark-namenode-master.out
worker2: starting datanode, logging to /home/spark/hadoop-2.10.0/logs/hadoop-spark-datanode-worker2.out
worker1: starting datanode, logging to /home/spark/hadoop-2.10.0/logs/hadoop-spark-datanode-worker1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/spark/hadoop-2.10.0/logs/hadoop-spark-secondarynamenode-master.out
spark@master:~/hadoop-2.10.0/sbin$
Back on master, start the HDFS with command start-dfs.sh

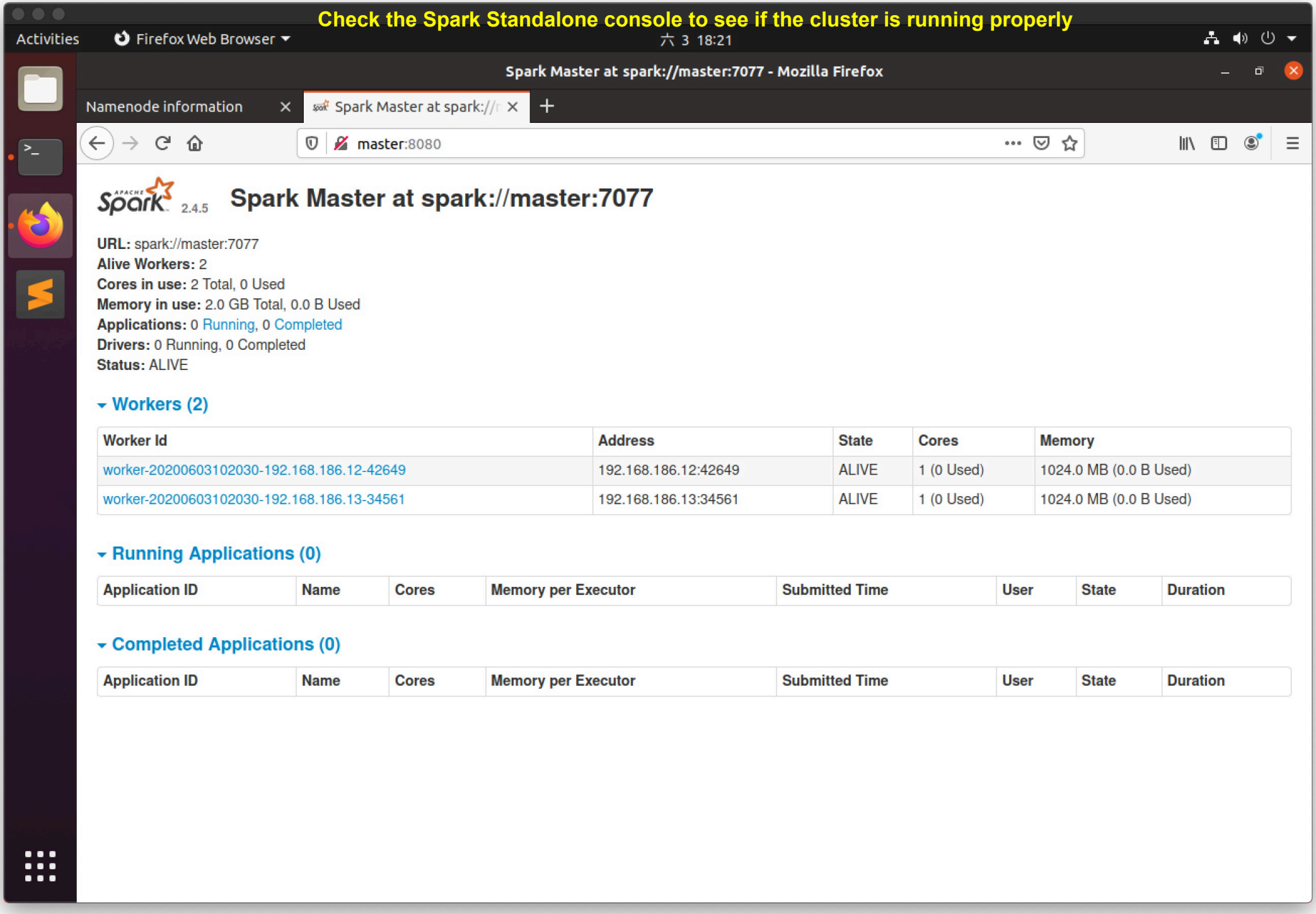

Activities
Firefox Web Browser
六 3 18:18
Namenode information - Mozilla Firefox
Namenode information
check if HDFS is running properly with NameNode console
master:50070/dfshealth.html#tab-overview
67%
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'master:8020' (active)
Started: Wed Jun 03 18:15:32 +0800 2020
Version: 2.10.0, re2f1f118e465e787d8567dfa6e2f3b72a0eb9194
Compiled: Wed Oct 23 03:10:00 +0800 2019 by jhung from branch-2.10.0
Cluster ID: CID-b0c878fa-5fe2-49b1-bca3-a0ae04deed1e
Block Pool ID: BP-2102378775-127.0.1.1-1590919722642
Summary
Security is off.
Safemode is off.
21 files and directories, 7 blocks = 28 total filesystem object(s).
Heap Memory used 37.64 MB of 49.98 MB Heap Memory. Max Heap Memory is 966.69 MB.
Non Heap Memory used 44.68 MB of 45.31 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 39.12 GB
DFS Used: 33.59 MB (0.08%)
Non DFS Used: 20.77 GB
DFS Remaining: 16.28 GB (41.62%)
Block Pool Used: 33.59 MB (0.08%)
DataNodes usages% (Min/Median/Max/stdDev): 0.08% / 0.08% / 0.08% / 0.00%
Live Nodes 2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes 0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes 0
Entering Maintenance Nodes 0
Total Datanode Volume Failures 0 (0 B)
Number of Under-Replicated Blocks 7

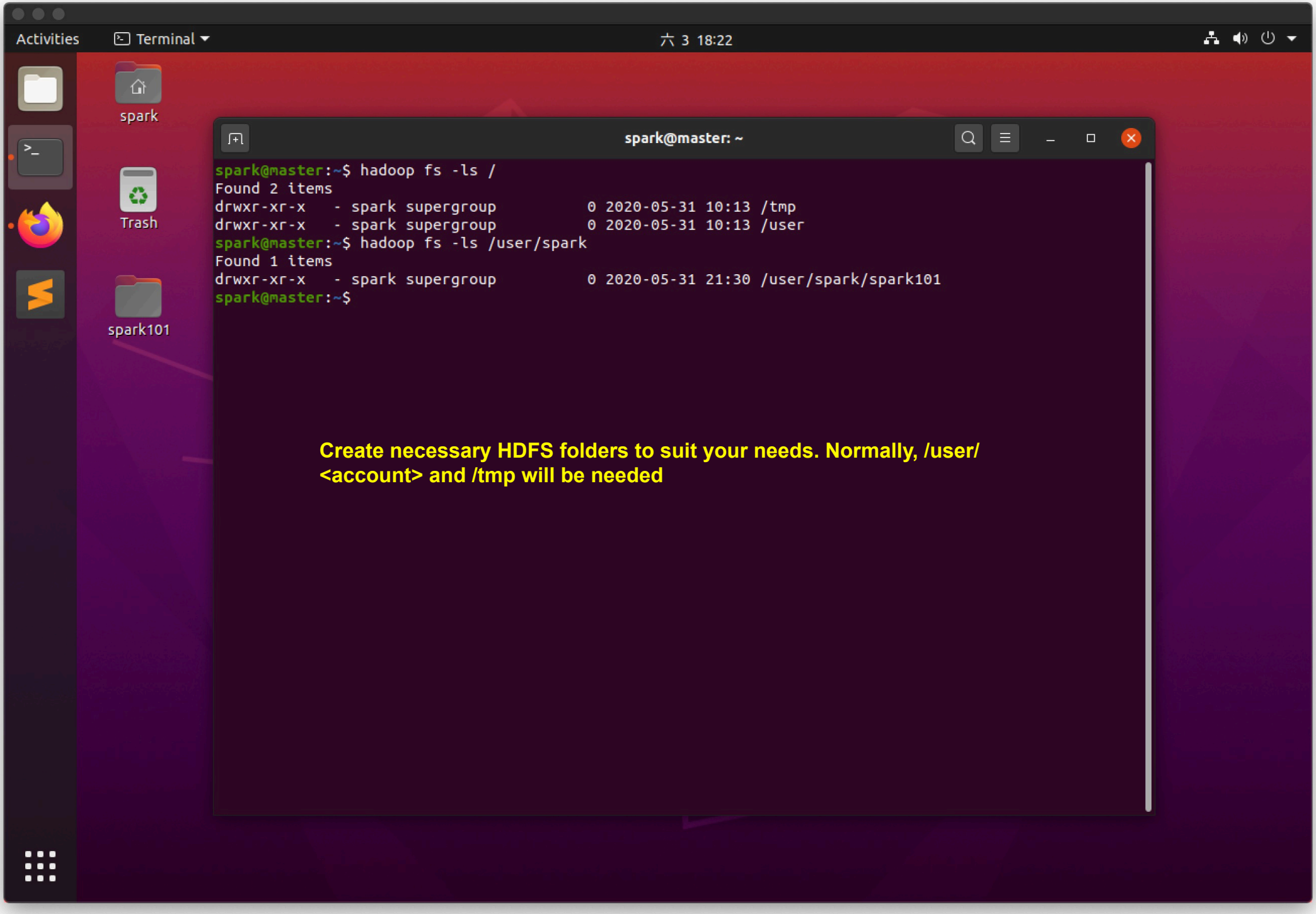

Activities
Terminal
六 3 18:19
spark
Trash
spark101
spark@master: ~/spark-2.4.5-bin-hadoop2.7/conf
spark@master:~$ ls
Downloads anaconda3 hadoop-2.10.0 hdfs jdk1.8.0_251 scala-2.12.11 spark-2.4.5-bin-hadoop2.7
spark@master:~$ cd spark-2.4.5-bin-hadoop2.7/conf
spark@master:~/spark-2.4.5-bin-hadoop2.7/conf$ ls
docker.properties.template metrics.properties.template spark-env.sh
fairscheduler.xml.template slaves spark-env.sh.template
log4j.properties.template spark-defaults.conf.template
spark@master:~/spark-2.4.5-bin-hadoop2.7/conf$
On master, setting up Spark config files.
For minimum, modify slaves and spark-env.sh

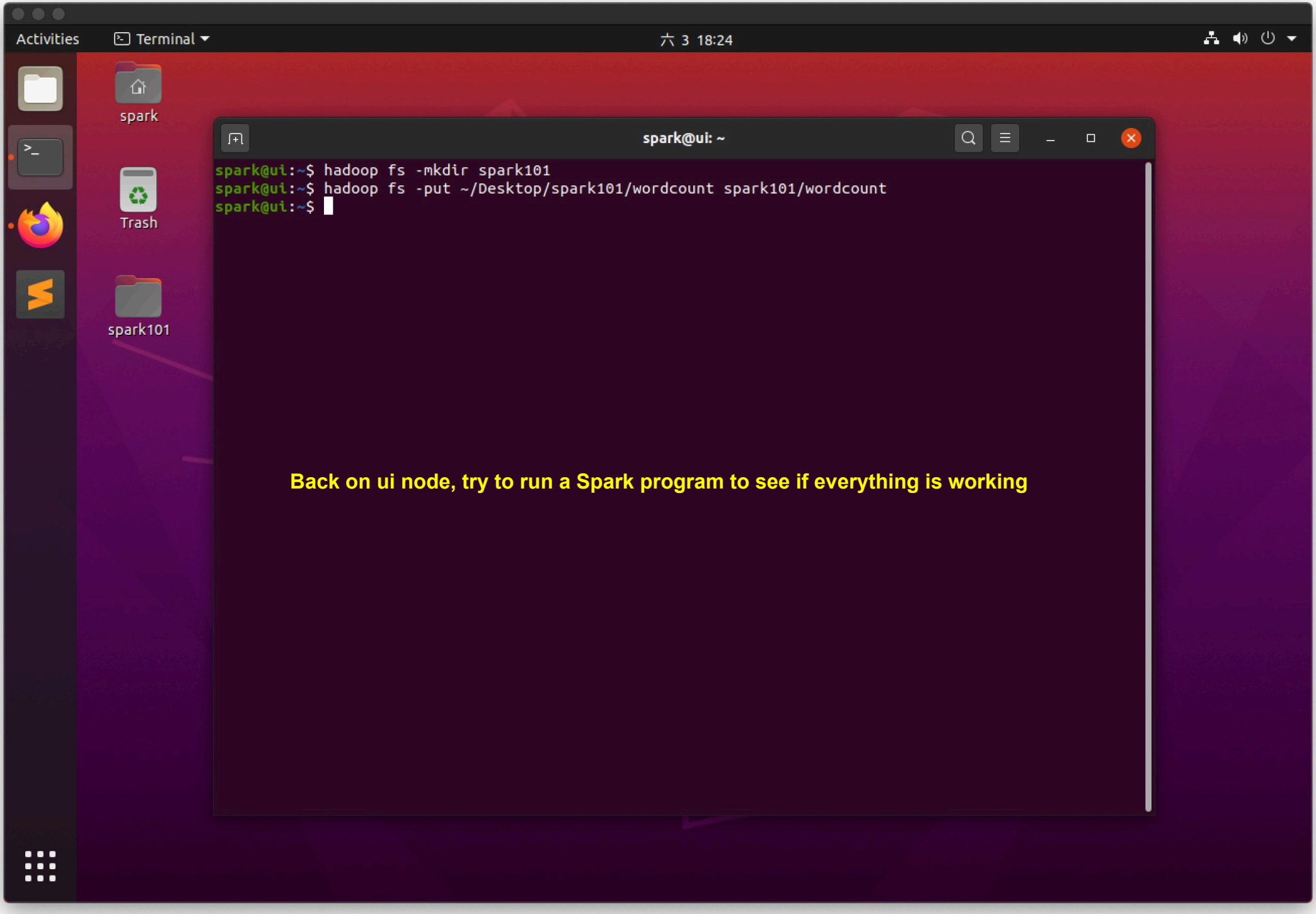

Activities
Terminal
六 3 18:19
spark
Trash
spark101
spark@master: ~/spark-2.4.5-bin-hadoop2.7/conf
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements.  See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# 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.
worker1
worker2
"slaves" 20L, 872C
20,7
All

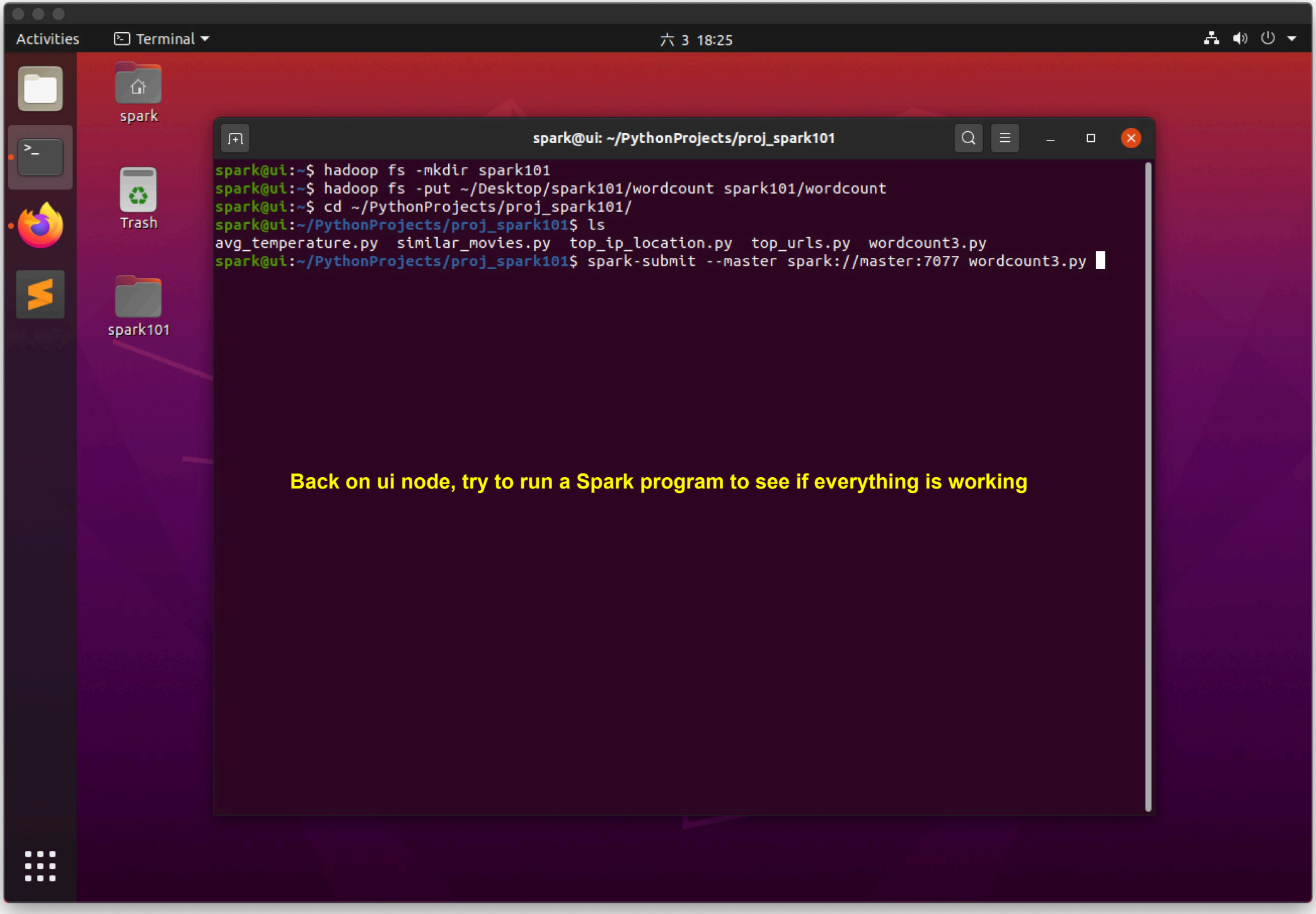

Activities
Terminal
六 3 18:19
spark
Trash
spark101
spark@master: ~/spark-2.4.5-bin-hadoop2.7/conf
# Options for native BLAS, like Intel MKL, OpenBLAS, and so on.
# You might get better performance to enable these options if using native BLAS (see SPARK-21305).
# - MKL_NUM_THREADS=1 Disable multi-threading of Intel MKL
# - OPENBLAS_NUM_THREADS=1 Disable multi-threading of OpenBLAS
export JAVA_HOME=/home/spark/jdk1.8.0_251
export PATH=$JAVA_HOME/bin:$PATH
export SCALA_HOME=/home/spark/scala-2.12.11
export PATH=$SCALA_HOME/bin:$PATH
export ANACONDA_HOME=/home/spark/anaconda3
export PATH=$ANACONDA_HOME/bin:$PATH
export HADOOP_HOME=/home/spark/hadoop-2.10.0
export PATH=$HADOOP_HOME/bin:$PATH
export SPARK_HOME=/home/spark/spark-2.4.5-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH
export PYSPARK_PYTHON=python
# use ipython
export PYSPARK_DRIVER_PYTHON=ipython
71,1 Bot

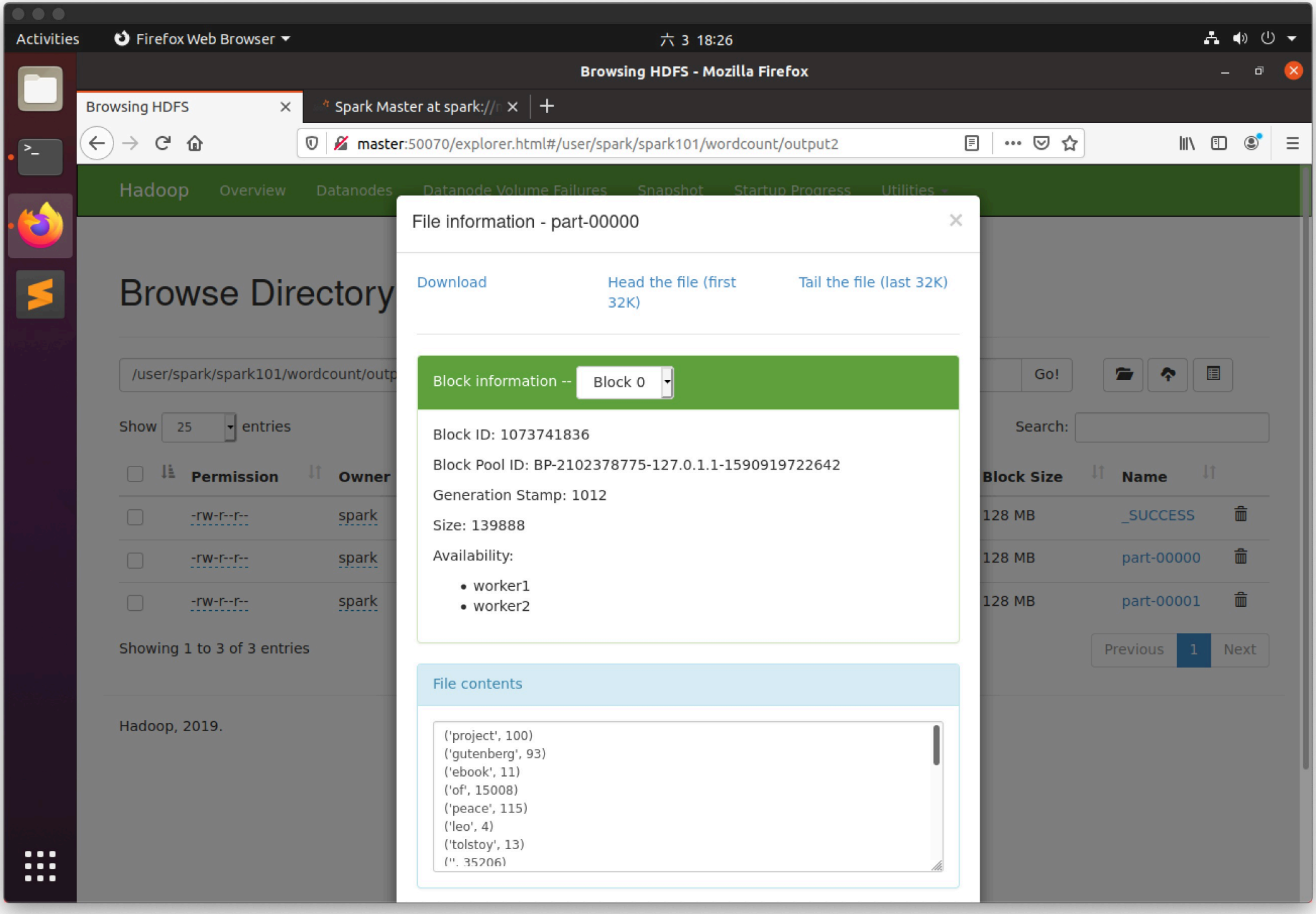

Activities
Terminal
六 3 18:20
spark
Trash
spark101
spark@master: ~/spark-2.4.5-bin-hadoop2.7/sbin
spark@master:~/spark-2.4.5-bin-hadoop2.7/sbin$ ls
slaves.sh start-mesos-shuffle-service.sh stop-mesos-dispatcher.sh
spark-config.sh start-shuffle-service.sh stop-mesos-shuffle-service.sh
spark-daemon.sh start-slave.sh stop-shuffle-service.sh
spark-daemons.sh start-slaves.sh stop-slave.sh
start-all.sh start-thriftserver.sh stop-slaves.sh
start-history-server.sh stop-all.sh stop-thriftserver.sh
start-master.sh stop-history-server.sh
start-mesos-dispatcher.sh stop-master.sh
spark@master:~/spark-2.4.5-bin-hadoop2.7/sbin$ ./start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/spark/spark-2.4.5-bin-hadoop2.7/logs/spark-spark-org.apache.spark.deploy.master.Master-1-master.out
worker2: starting org.apache.spark.deploy.worker.Worker, logging to /home/spark/spark-2.4.5-bin-hadoop2.7/logs/spark-spark-org.apache.spark.deploy.worker.Worker-1-worker2.out
worker1: starting org.apache.spark.deploy.worker.Worker, logging to /home/spark/spark-2.4.5-bin-hadoop2.7/logs/spark-spark-org.apache.spark.deploy.worker.Worker-1-worker1.out
spark@master:~/spark-2.4.5-bin-hadoop2.7/sbin$
On master, start Spark Standalone cluster with start-all.sh


Check the Spark Standalone console to see if the cluster is running properly
Activities
Firefox Web Browser
六 3 18:21
Spark Master at spark://master:7077 - Mozilla Firefox
Namenode information
Spark Master at spark://
master:8080
Spark Master at spark://master:7077
2.4.5
URL: spark://master:7077
Alive Workers: 2
Cores in use: 2 Total, 0 Used
Memory in use: 2.0 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE
Workers (2)
Worker Id
Address
State
Cores
Memory
worker-20200603102030-192.168.186.12-42649
192.168.186.12:42649
ALIVE
1 (0 Used)
1024.0 MB (0.0 B Used)
worker-20200603102030-192.168.186.13-34561
192.168.186.13:34561
ALIVE
1 (0 Used)
1024.0 MB (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


Activities
Terminal
六 3 18:22
spark
Trash
spark101
spark@master: ~
spark@master:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x - spark supergroup 0 2020-05-31 10:13 /tmp
drwxr-xr-x - spark supergroup 0 2020-05-31 10:13 /user
spark@master:~$ hadoop fs -ls /user/spark
Found 1 items
drwxr-xr-x - spark supergroup 0 2020-05-31 21:30 /user/spark/spark101
spark@master:~$
Create necessary HDFS folders to suit your needs. Normally, /user/<account> and /tmp will be needed


Activities
Terminal
六 3 18:24
spark
Trash
spark101
spark@ui: ~
spark@ui:~$ hadoop fs -mkdir spark101
spark@ui:~$ hadoop fs -put ~/Desktop/spark101/wordcount spark101/wordcount
spark@ui:~$
Back on ui node, try to run a Spark program to see if everything is working


Activities
Terminal
六 3 18:25
spark
Trash
spark101
spark@ui: ~/PythonProjects/proj_spark101
spark@ui:~$ hadoop fs -mkdir spark101
spark@ui:~$ hadoop fs -put ~/Desktop/spark101/wordcount spark101/wordcount
spark@ui:~$ cd ~/PythonProjects/proj_spark101/
spark@ui:~/PythonProjects/proj_spark101$ ls
avg_temperature.py  similar_movies.py  top_ip_location.py  top_urls.py  wordcount3.py
spark@ui:~/PythonProjects/proj_spark101$ spark-submit --master spark://master:7077 wordcount3.py
Back on ui node, try to run a Spark program to see if everything is working


Activities
Firefox Web Browser
六 3 18:26
Browsing HDFS - Mozilla Firefox
Browsing HDFS
Spark Master at spark://
master:50070/explorer.html#/user/spark/spark101/wordcount/output2
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/user/spark/spark101/wordcount/outp
Go!
Show 25 entries
Search:
Permission
Owner
Block Size
Name
-rw-r--r--
spark
128 MB
_SUCCESS
-rw-r--r--
spark
128 MB
part-00000
-rw-r--r--
spark
128 MB
part-00001
Showing 1 to 3 of 3 entries
Previous
1
Next
Hadoop, 2019.
File information - part-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information -- Block 0
Block ID: 1073741836
Block Pool ID: BP-2102378775-127.0.1.1-1590919722642
Generation Stamp: 1012
Size: 139888
Availability:
worker1
worker2
File contents
('project', 100)
('gutenberg', 93)
('ebook', 11)
('of', 15008)
('peace', 115)
('leo', 4)
('tolstoy', 13)