

HƯỚNG DẪN CÀI ĐẶT (TRÊN UBUNTU 20.04)

1. Cài đặt Spark

Cài Java 8

- + Cài Java 8 (do Spark chạy trên máy ảo JVM của Java)
`sudo apt install default-jdk`

Download và cài Apache Spark

- + Download Apache Spark 2.4.5
`curl -O https://www.apache.org/dyn/closer.lua/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz`
- + Giải nén file vừa download
`tar xvf spark-2.4.5-bin-hadoop2.7.tgz`
- + Chuyển thư mục vừa giải nén vào thư mục /opt/
`sudo mv spark-2.4.5-bin-hadoop2.7/ /opt/spark`

Config biến môi trường Spark

- + Thêm vào file `.bashrc` như sau

```
export SPARK_HOME=/usr/local/spark
export PATH=$PATH:$SPARK_HOME/bin

export PYSARK_PYTHON=/usr/bin/python3.7
export PYSARK_DRIVER_PYTHON=/usr/bin/python3.7

export PYTHONPATH=$SPARK_HOME/python/:$PYTHONPATH
export PYTHONPATH=$SPARK_HOME/python/lib/py4j-0.10.7-src.zip:$PYTHONPATH
```

Chú ý config đường dẫn python đúng với bản python 3 có trên máy.

Kiểm tra

- + Để kiểm tra việc cài đặt thành công hay không, ta chạy lệnh
`pyspark`

```
mxnthng@latitude-e7450:~$ pyspark
Python 3.7.9 (default, Aug 18 2020, 02:07:21)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
21/01/28 21:58:18 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where 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

  ____      _
 / ___|  _ \| | | |
 \___ \ |_) | |_| |
  ___) ||  __| | | |
 |____||___|_|_|_|

 version 2.4.7

Using Python version 3.7.9 (default, Aug 18 2020 02:07:21)
SparkSession available as 'spark'.
>>> █
```

2. Cài đặt HDFS trên cụm 2 máy

Config các file

+ Config file `/etc/hosts` như sau

```
192.168.43.109 node1
192.168.43.139 node2
```

Chú ý sửa địa chỉ IP thành địa chỉ IP của 2 máy vật lý.

Config ssh-key

+ Config ssh-key để máy chủ có quyền truy cập đến datanode

```
ssh-keygen -b 4096
```

+ Ở node1 (master) ta copy public key trong

```
less /home/hadoop/.ssh/id_rsa.pub
```

+ Ở node2 (datanode) tạo file `master.pub` vào đường dẫn `/home/hadoop/.ssh`, dán public key vào file.

Cài Hadoop

+ Ở node1, ta download và unzip hadoop bằng chuỗi lệnh

```
cd
wget http://apache.cs.utah.edu/hadoop/common/current/hadoop-3.1.2.tar.gz
tar -xzf hadoop-3.1.2.tar.gz
mv hadoop-3.1.2 hadoop
```

Config các file

+ Thêm PATH vào file `.profile`

```
PATH=/home/bigdata/hadoop/bin:/home/bigdata/hadoop/sbin:$PATH
# ~/.profile: executed by the command interpreter for login shells.
# This file is not read by bash(1), if ~/.bash_profile or ~/.bash_login
```

+ Thêm PATH HADOOP_HOME trong file `.bashrc`

```
export HADOOP_HOME=/home/bigdata/hadoop
export PATH=${PATH}:${HADOOP_HOME}/bin:${HADOOP_HOME}/sbin
```

+ Chỉ định JAVA_HOME cho hadoop trong file `~/hadoop/etc/hadoop/hadoop-env.sh`

```
# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre
```

+ Cấu hình `~/hadoop/etc/hadoop/core-site.xml`

```
GNU nano 2.9.3 core-site.xml Modified

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://master:9000</value>
  </property>
</configuration>
```

+ Cấu hình ~/hadoop/etc/hadoop/hdfs-site.xml

```
GNU nano 2.9.3 hdfs-site.xml Modified
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/bigdata/data/nameNode</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/bigdata/data/dataNode</value>
  </property>

  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>0.0.0.0:9870</value>
  </property>
</configuration>
```

+ Cấu hình ~/hadoop/etc/hadoop/mapred-site.xml

```
GNU nano 2.9.3 mapred-site.xml Modified
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
  </property>
  <property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
  </property>
  <property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
  </property>
</configuration>
```

- + Cấu hình `~/hadoop/etc/hadoop/yarn-site.xml`

```
GNU nano 2.9.3                yarn-site.xml                Modified
<?xml version="1.0"?>
<configuration>

  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>

  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>13.76.132.16</value>
  </property>

  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

</configuration>
```

- + Cấu hình `~/hadoop/etc/hadoop/slaves`

```
GNU nano 2.9.3      slaves
node1
node2
```

+ Khởi động hadoop bằng câu lệnh:

```
hdfs namenode -format
start-dfs.sh
```

+ Sau khi chạy ta kiểm tra bằng cách mở cổng 9870 ở máy master. Ta thu được kết quả như hình:

[illegible]

+ Sau khi cài đặt thành công ta đẩy 47 file dữ liệu lên HDFS. Sau khi đẩy lên và dùng lệnh *hdfs dfs -ls /* để kiểm tra thì thu được kết quả như hình

```
Found 47 items
-rw-r--r--  2 bigdata supergroup 46646946 2020-12-27 20:43 /data/data1.txt
-rw-r--r--  2 bigdata supergroup 20367001 2020-12-27 20:43 /data/data10.txt
-rw-r--r--  2 bigdata supergroup 19241527 2020-12-27 20:43 /data/data11.txt
-rw-r--r--  2 bigdata supergroup 16759435 2020-12-27 20:43 /data/data12.txt
-rw-r--r--  2 bigdata supergroup 16966458 2020-12-27 20:43 /data/data13.txt
-rw-r--r--  2 bigdata supergroup 26576265 2020-12-27 20:43 /data/data14.txt
-rw-r--r--  2 bigdata supergroup 20829610 2020-12-27 20:43 /data/data15.txt
-rw-r--r--  2 bigdata supergroup 18923814 2020-12-27 20:43 /data/data16.txt
-rw-r--r--  2 bigdata supergroup 21561638 2020-12-27 20:43 /data/data17.txt
-rw-r--r--  2 bigdata supergroup 15569851 2020-12-27 20:43 /data/data18.txt
-rw-r--r--  2 bigdata supergroup 16582964 2020-12-27 20:43 /data/data19.txt
-rw-r--r--  2 bigdata supergroup 62544652 2020-12-27 20:43 /data/data2.txt
-rw-r--r--  2 bigdata supergroup 22071598 2020-12-27 20:43 /data/data20.txt
-rw-r--r--  2 bigdata supergroup 128933289 2020-12-27 20:43 /data/data21.txt
-rw-r--r--  2 bigdata supergroup 21759212 2020-12-27 20:43 /data/data22.txt
-rw-r--r--  2 bigdata supergroup 20103132 2020-12-27 20:43 /data/data23.txt
-rw-r--r--  2 bigdata supergroup 19405523 2020-12-27 20:43 /data/data24.txt
-rw-r--r--  2 bigdata supergroup 18213834 2020-12-27 20:43 /data/data25.txt
-rw-r--r--  2 bigdata supergroup 19557616 2020-12-27 20:43 /data/data26.txt
-rw-r--r--  2 bigdata supergroup 18785140 2020-12-27 20:43 /data/data27.txt
-rw-r--r--  2 bigdata supergroup 18475520 2020-12-27 20:43 /data/data28.txt
-rw-r--r--  2 bigdata supergroup 19238524 2020-12-27 20:43 /data/data29.txt
-rw-r--r--  2 bigdata supergroup 45906230 2020-12-27 20:43 /data/data3.txt
-rw-r--r--  2 bigdata supergroup 17789656 2020-12-27 20:43 /data/data30.txt
-rw-r--r--  2 bigdata supergroup 17819422 2020-12-27 20:43 /data/data31.txt
-rw-r--r--  2 bigdata supergroup 18946046 2020-12-27 20:43 /data/data32.txt
-rw-r--r--  2 bigdata supergroup 17124986 2020-12-27 20:43 /data/data33.txt
-rw-r--r--  2 bigdata supergroup 19759803 2020-12-27 20:43 /data/data34.txt
-rw-r--r--  2 bigdata supergroup 15636295 2020-12-27 20:43 /data/data35.txt
-rw-r--r--  2 bigdata supergroup 37731903 2020-12-27 20:43 /data/data36.txt
-rw-r--r--  2 bigdata supergroup 40230293 2020-12-27 20:43 /data/data37.txt
-rw-r--r--  2 bigdata supergroup 56628709 2020-12-27 20:43 /data/data38.txt
-rw-r--r--  2 bigdata supergroup 41177767 2020-12-27 20:43 /data/data39.txt
-rw-r--r--  2 bigdata supergroup 46860302 2020-12-27 20:43 /data/data4.txt
-rw-r--r--  2 bigdata supergroup 56072810 2020-12-27 20:43 /data/data40.txt
-rw-r--r--  2 bigdata supergroup 43340354 2020-12-27 20:43 /data/data41.txt
-rw-r--r--  2 bigdata supergroup 38966846 2020-12-27 20:43 /data/data42.txt
-rw-r--r--  2 bigdata supergroup 64864995 2020-12-27 20:43 /data/data43.txt
-rw-r--r--  2 bigdata supergroup 40490627 2020-12-27 20:43 /data/data44.txt
-rw-r--r--  2 bigdata supergroup 89922574 2020-12-27 20:43 /data/data45.txt
-rw-r--r--  2 bigdata supergroup 91515755 2020-12-27 20:43 /data/data46.txt
-rw-r--r--  2 bigdata supergroup 20484603 2020-12-27 20:43 /data/data47.txt
-rw-r--r--  2 bigdata supergroup 42778192 2020-12-27 20:43 /data/data5.txt
-rw-r--r--  2 bigdata supergroup 35522193 2020-12-27 20:43 /data/data6.txt
-rw-r--r--  2 bigdata supergroup 30423763 2020-12-27 20:43 /data/data7.txt
-rw-r--r--  2 bigdata supergroup 29789324 2020-12-27 20:43 /data/data8.txt
-rw-r--r--  2 bigdata supergroup 25079705 2020-12-27 20:43 /data/data9.txt
```

3. Chạy code

Chạy bằng Python thuần

- + Chạy file *wordcount_python.py*.
- + Ta thu được kết quả như hình:

```
mxnthng@latitude-e7450:~$ /usr/bin/python3 /home/mxnthng/Desktop/20201/Project3/wordcount_python.py
Total time: 0:02:16.557642
```

Chạy bằng Spark trên máy local

- + Chạy file *wordcount_spark.py* thông qua spark-submit bằng câu lệnh
spark-submit /home/mxnthng/Desktop/20201/Project3/wordcount_spark.py

Chạy bằng Spark Cluster trên HDFS

- + Phải push data lên HDFS trước
- + Chạy tương tự như trên, nhưng trong file *wordcount_spark.py* thay *local_path* thành *hdfs_path*.