

CÀI ĐẶT HADOOP CLUSTER

1. Tạo nút master và các nút khác trong /etc/hosts: `nano /etc/hosts`

```
GNU nano 6.2 /etc/hosts
127.0.0.1 localhost
127.0.1.1 ubuntu-bigdata
10.10.47.48 node-master
10.10.47.47 node1
10.10.47.49 node2
```

(thêm 3 dòng cuối, với IP chính là địa chỉ IP riêng của từng máy!)

2. Phân phối cặp khóa xác thực cho người dùng Hadoop (là public key mà node-master đã tạo ra ở phần cài single hadoop)

➔ Đăng nhập vào tài khoản Hadoop trên nút master và thực hiện copy cặp khóa đến các node thành viên.

`ssh-copy-id hadoop@node1` (hoặc dùng địa chỉ IP thay chỗ node1)

`ssh-copy-id hadoop@node2` (hoặc dùng địa chỉ IP thay chỗ node2)

```
hadoop@doanb2013527-BD:~$ ssh-copy-id hadoop@node1
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hadoop/.ssh/id_rsa.pub"
The authenticity of host 'node1 (10.10.40.254)' can't be established.
ED25519 key fingerprint is SHA256:4MfeLmZF0LGcxbUh4d6PXPUFh2bF7Q5BcT+b2BtP89U.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install all the new keys
hadoop@node1's password:

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'hadoop@node1'"
and check to make sure that only the key(s) you wanted were added.
```

3. Chỉnh sửa các file **.xml** đã tạo ở phần single hadoop. (Vẫn còn đang ở người dùng Hadoop).

- 3.1. Cài đặt NameNode Location đến nút master tại cổng 9000:

`nano ~/hadoop/etc/hadoop/core-site.xml`

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://node-master:9000</value>
  </property>
</configuration>
```

```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/core-site.xml
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
See the License for the specific language governing permissions
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://node-master:9000</value>
  </property>
</configuration>

```

3.2. Cài đặt đường dẫn PATH cho HDFS:

nano ~/hadoop/etc/hadoop/hdfs-site.xml

```
<configuration>
```

```
  <property>
```

```
    <name>dfs.replication</name>
```

```
    <value>2</value>
```

(bao nhiêu cluster thì thay đổi số cho phù hợp)

```
  </property>
```

```
  <property>
```

```
    <name>dfs.namenode.name.dir</name>
```

```
    <value>/home/hadoop/data/nameNode</value>
```

```
  </property>
```

```
  <property>
```

```
    <name>dfs.datanode.data.dir</name>
```

```
    <value>home/hadoop/data/dataNode</value>
```

```
  </property>
```

```
</configuration>
```

```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/hadoop/data/nameNode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/hadoop/data/dataNode</value>
  </property>
</configuration>

```

3.3. Cài đặt YARN như framework mặc định cho MapReduce:

nano ~/hadoop/etc/hadoop/mapred-site.xml

```

<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>

```

```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
<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>

```

3.4. Cấu hình YARN:

nano ~/hadoop/etc/hadoop/yarn-site.xml

```

<configuration>

```

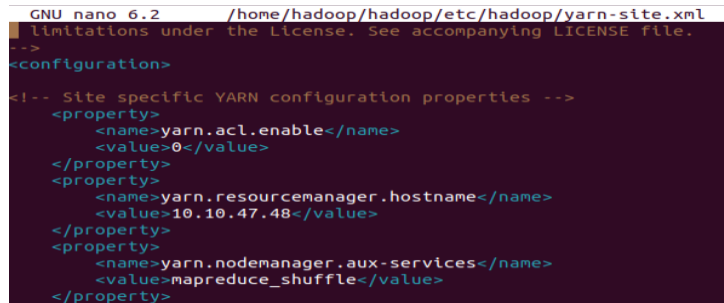
```

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

<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>10.10.47.48</value>
  (địa chỉ IP của máy node-master)
</property>

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

```



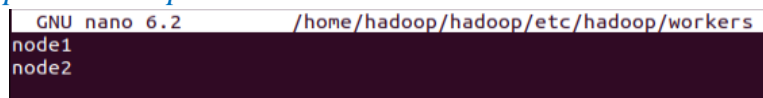
```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>10.10.47.48</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>

```

4. Cấu hình các nút **worker** (Node1, Node2), thêm 2 nút vào file workers:

nano ~/hadoop/etc/hadoop/workers



```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/workers
node1
node2

```

5. Cấu hình phân bổ bộ nhớ trong 2 file **mapred-site.xml** và **yarn.xml**

- 5.1. Bổ sung các dòng sau vào file **yarn.xml**:

nano ~/hadoop/etc/hadoop/yarn-site.xml

```

<property>
  <name>yarn.nodemanager.resource.memory-mb</name>
  <value>1536</value>
</property>

<property>
  <name>yarn.scheduler.maximum-allocation-mb</name>
  <value>1536</value>
</property>

<property>

```

```

        <name>yarn.scheduler.minimum-allocation-mb</name>
        <value>128</value>
    </property>

    <property>
        <name>yarn.nodemanager.vmem-check-enabled</name>
        <value>>false</value>
    </property>

```

```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>128</value>
  </property>
  <property>
    <name>yarn.nodemanager.vmem-check-enabled</name>
    <value>>false</value>
  </property>

```

5.2. Bổ sung các dòng sau vào file **mapred.xml**:

nano ~/hadoop/etc/hadoop/mapred-site.xml

```

    <property>
        <name>yarn.app.mapreduce.am.resource.mb</name>
        <value>512</value>
    </property>

    <property>
        <name>mapreduce.map.memory.mb</name>
        <value>256</value>
    </property>

    <property>
        <name>mapreduce.reduce.memory.mb</name>
        <value>256</value>
    </property>

```

```

GNU nano 6.2 /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
    <property>
        <name>mapreduce.reduce.env</name>
        <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
    </property>
    <property>
        <name>yarn.app.mapreduce.am.resource.mb</name>
        <value>512</value>
    </property>
    <property>
        <name>mapreduce.map.memory.mb</name>
        <value>256</value>
    </property>
    <property>
        <name>mapreduce.reduce.memory.mb</name>
        <value>256</value>
    </property>

```

6. Copy các file đã sửa ở mục 3 cho các nút workers: (dùng vòng lặp for cho tiện cho các trường hợp có nhiều nút cho sau này)

for node in node1 node2; do

scp ~/hadoop/etc/hadoop/ \$node:~/hadoop/etc/hadoop/;*

done

(lệnh này viết liên tục không enter xuống dòng nhé, do lệnh dài nên khi trình bày mới xuống dòng thôi).

7. Format thư mục NameNode: *hdfs namenode -format*

```
2023-09-14 14:40:26,055 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-09-14 14:40:26,065 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retr
y cache entry expiry time is 600000 millis
2023-09-14 14:40:26,069 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-09-14 14:40:26,069 INFO util.GSet: VM type = 64-bit
2023-09-14 14:40:26,069 INFO util.GSet: 0.0299999999329447746% max memory 479.5 MB = 147.3 KB
2023-09-14 14:40:26,069 INFO util.GSet: capacity = 2^14 = 16384 entries
2023-09-14 14:40:26,125 INFO namenode.FSImage: Allocated new BlockPoolId: BP-183793974-127.0.1.1-169
4677226102
2023-09-14 14:40:26,380 INFO common.Storage: Storage directory /home/hadoop/data/nameNode has been s
uccessfully formatted.
2023-09-14 14:40:26,476 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/data/nam
eNode/current/fsimage.ckpt_000000000000000000 using no compression
2023-09-14 14:40:26,715 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/data/nameNode/c
urrent/fsimage.ckpt_000000000000000000 of size 401 bytes saved in 0 seconds .
2023-09-14 14:40:26,817 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid
0
2023-09-14 14:40:26,881 INFO namenode.FSNamesystem: Stopping services started for active state
2023-09-14 14:40:26,881 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-09-14 14:40:26,893 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutd
own.
2023-09-14 14:40:26,895 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at doanb2013527-BD/127.0.1.1
*****/
hadoop@doanb2013527-BD: $
```

(màn hình như v là thành công)

8. Start Hadoop cluster: *start-all.sh*

(7, 8 sẽ chạy thực thi như trong phần cài single hadoop)

➔ Khi bắt đầu chạy hadoop ở mục 8, nó sẽ bắt đầu NameNode và SecondNameNode ở nút master và DataNode ở nút 1,2.

```
hadoop@doanb2013527-BD:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [node-master]
node-master: Warning: Permanently added 'node-master' (ED25519) to the list of known hosts.
Starting datanodes
node2: datanode is running as process 3820. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pi
d file is empty before retry.
node1: datanode is running as process 61026. Stop it first and ensure /tmp/hadoop-hadoop-datanode.p
id file is empty before retry.
Starting secondary namenodes [doanb2013527-BD]
Starting resourcemanager
Starting nodemanagers
node2: nodemanager is running as process 4335. Stop it first and ensure /tmp/hadoop-hadoop-nodeman
ager.pid file is empty before retry.
node1: nodemanager is running as process 61807. Stop it first and ensure /tmp/hadoop-hadoop-nodeman
ager.pid file is empty before retry.
```

9. Kiểm tra các tiến trình đang chạy trên mỗi nút (lệnh này thực hiện trên nút master, nút 1 và 2 để thấy sự khác biệt): *jps*

- ➔ Nút master sẽ thấy 3 dòng với các số ID tiến trình khác nhau của Jps, NameNode và SecondNameNode.
- ➔ Nút 1, 2 sẽ thấy 2 dòng với các số ID tiến trình khác nhau của Jps và DataNode.

10. Để ngưng chạy HDFS trên các nút: [*stop-dfs.sh*](#)

```
hadoop@doanb2013527-BD:~$ start-dfs.sh
Starting namenodes on [node-master]
Starting datanodes
Starting secondary namenodes [doanb2013527-BD]
hadoop@doanb2013527-BD:~$
```