BÁO CÁO LAB2 LƯU TRỮ XỬ LÝ DỮ LIỆU LỚN

Nhóm: Squad Game

Thành viên: Lại Ngọc Thăng Long 20183581

Nguyễn Đình Dũng 20183506

Nguyễn Thành Long 20183586

Nguyễn Khương Duy 20183513

BÀI LÀM:

I. Trên 3 máy master, slave1, slave2

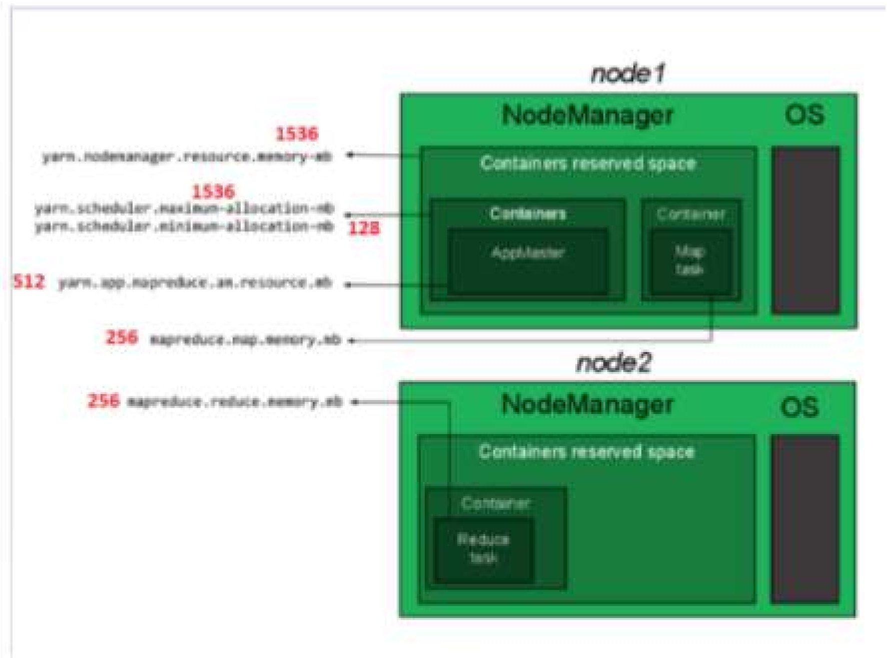
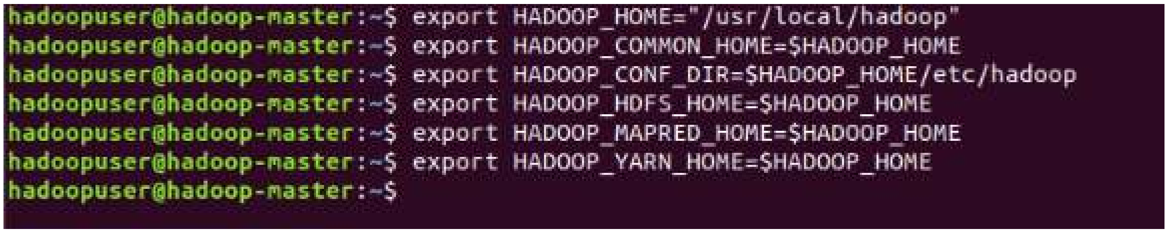

node1
NodeManager
OS
Containers reserved space
Containers
AppMaster
Container
Map task
node2
NodeManager
OS
Containers reserved space
Container
Reduce task
1536
yarn.nodemanager.resource.memory-mb
1536
yarn.scheduler.maximum-allocation-mb
yarn.scheduler.minimum-allocation-mb
128
512
yarn.app.mapreduce.am.resource.mb
256
mapreduce.map.memory.mb
256
mapreduce.reduce.memory.mb

Figure 8: Sample config for 2GB Nodes (will change in next lab)

1st Step: Configure yarn

```
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
```

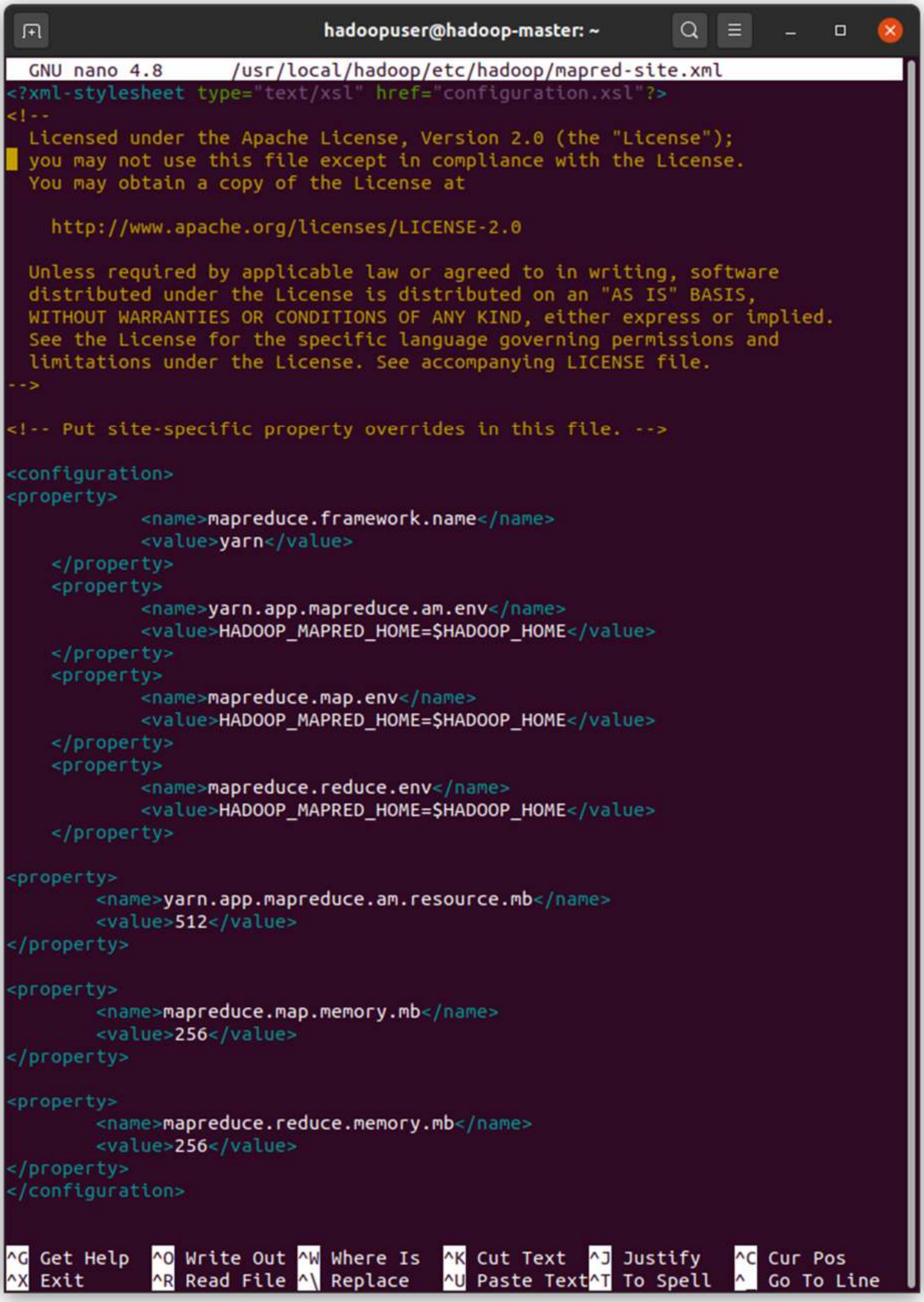


```
hadoopuser@hadoop-master:~$ export HADOOP_HOME="/usr/local/hadoop"
hadoopuser@hadoop-master:~$ export HADOOP_COMMON_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
hadoopuser@hadoop-master:~$ export HADOOP_HDFS_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_MAPRED_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_YARN_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$
```

2nd Step: Configure mapred-site.xml

```
sudo nano /usr/local/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>

<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>
</configuration>
```

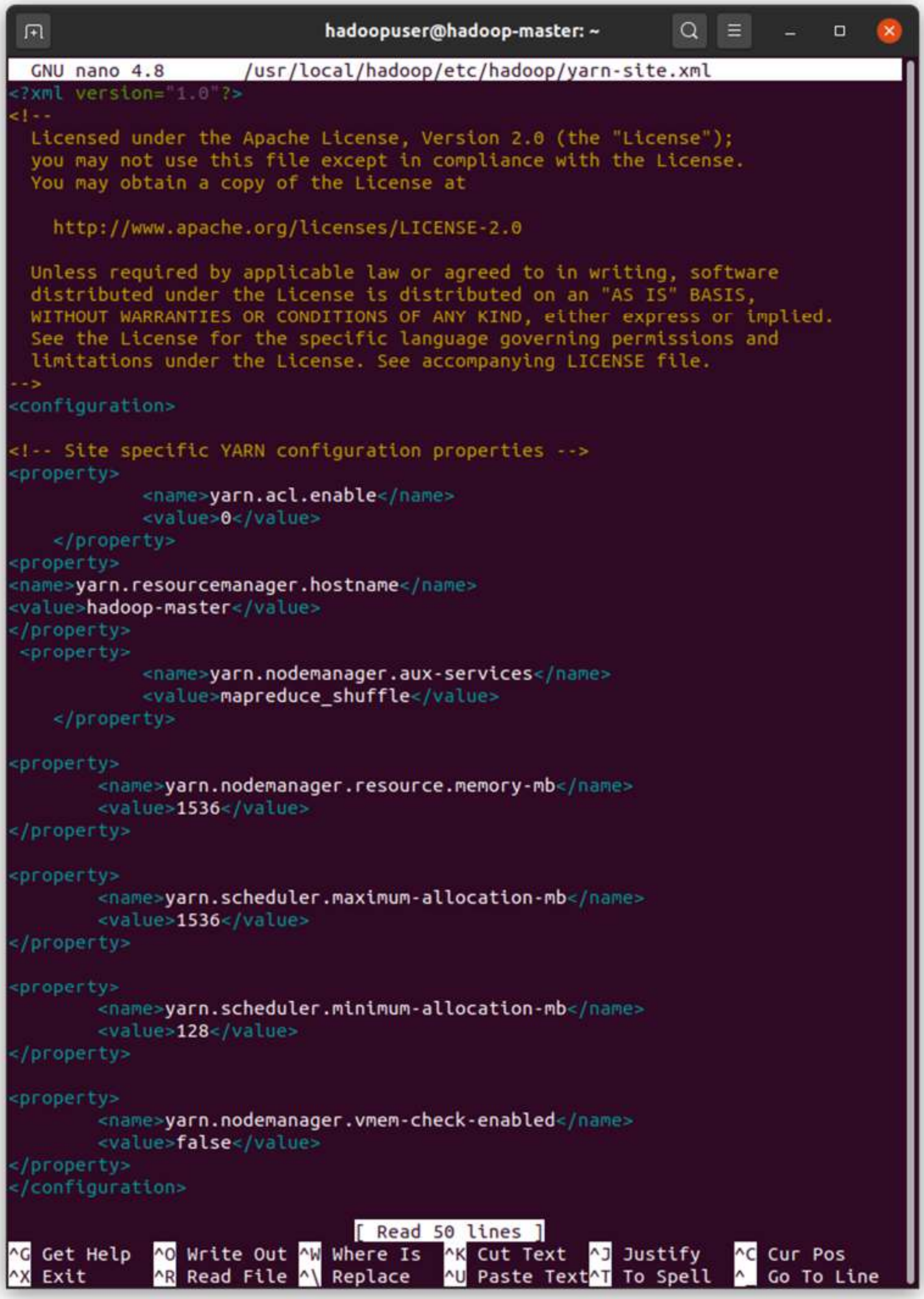


```
  GNU nano 4.8      /usr/local/hadoop/etc/hadoop/mapred-site.xml
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License 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>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>

<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>
</configuration>


^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
```

3rd Step: Config yarn-site.xml

```
sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml


<configuration>


<!-- Site specific YARN configuration properties -->
<property>
           <name>yarn.acl.enable</name>
           <value>0</value>
  </property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>hadoop-master</value>
</property>
 <property>
           <name>yarn.nodemanager.aux-services</name>
           <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.scheduler.minimum-allocation-mb</name>
        <value>128</value>
</property>

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

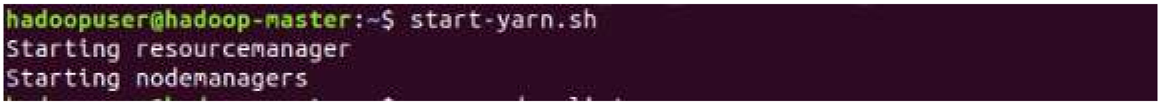


```
  GNU nano 4.8          /usr/local/hadoop/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
            <name>yarn.acl.enable</name>
            <value>0</value>
    </property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>hadoop-master</value>
</property>
 <property>
            <name>yarn.nodemanager.aux-services</name>
            <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.scheduler.minimum-allocation-mb</name>
        <value>128</value>
</property>

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

                               [ Read 50 lines ]
^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
```

II. Trên máy master

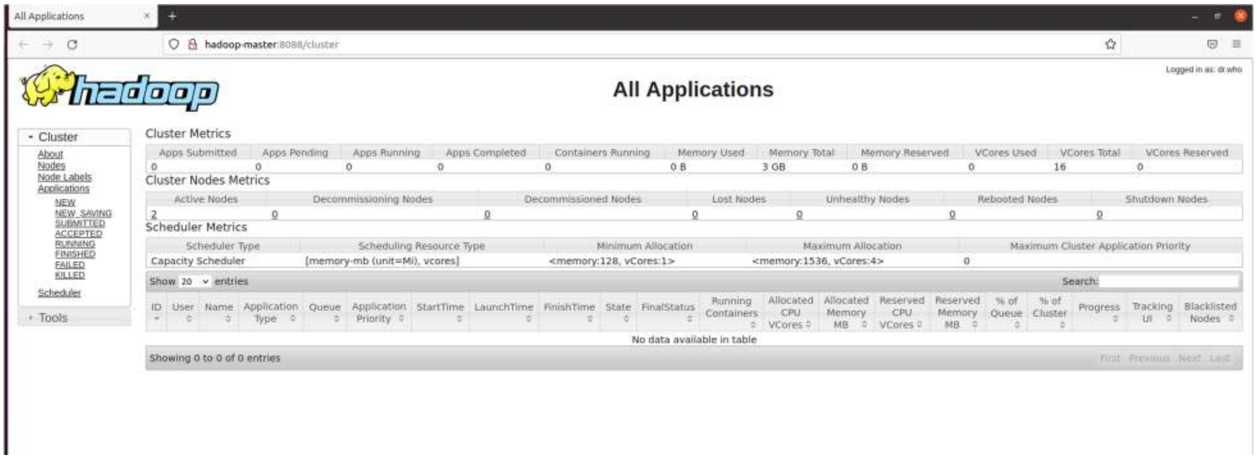
1st Step: Start yarn

```
start-yarn.sh
```



```
hadoopuser@hadoop-master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

Lên web, truy cập hadoop-master:8088/cluster



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	3 GB	0 B	0	16	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:128, vCores:1>	<memory:1536, vCores:4>	0

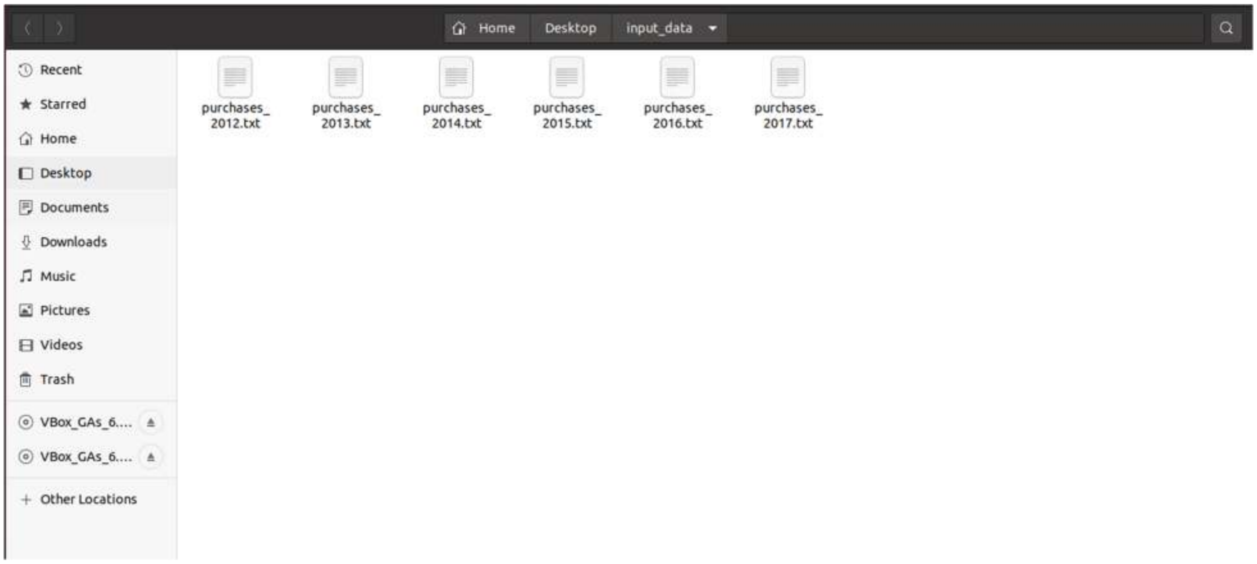
Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																				

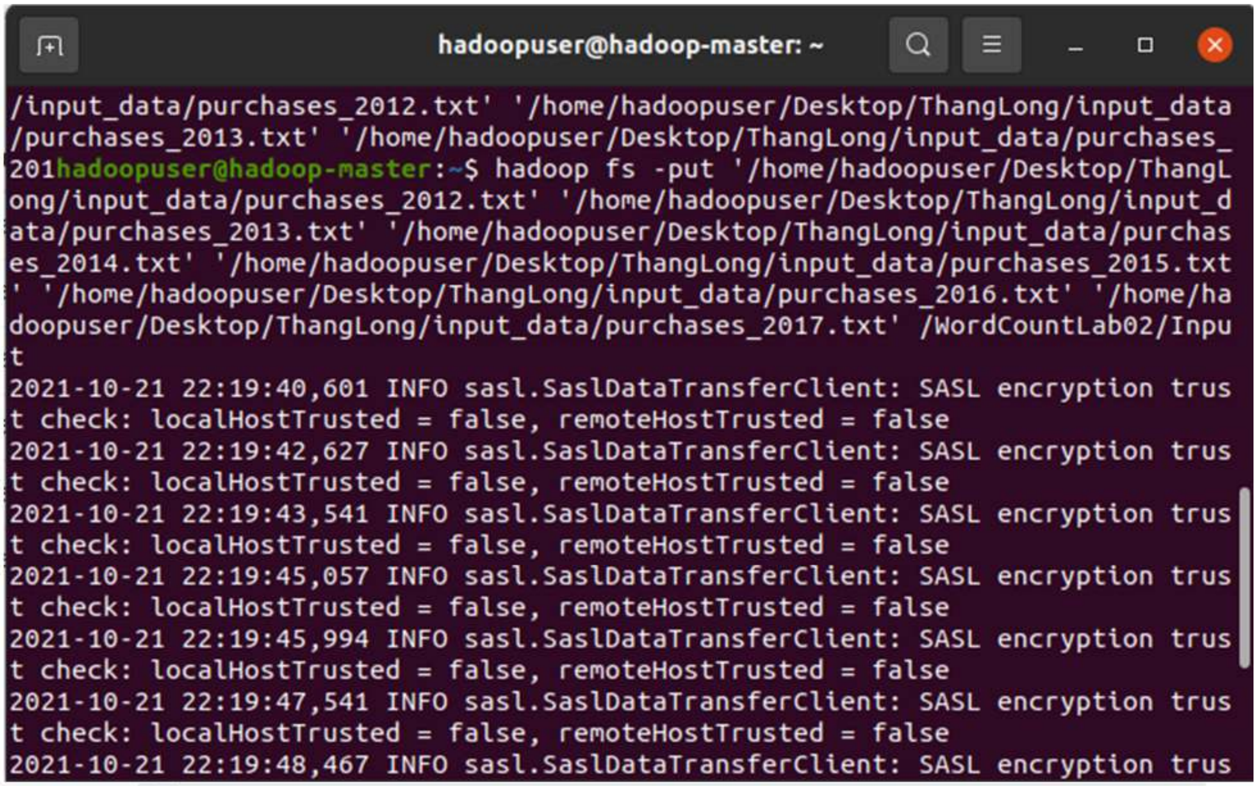
Showing 0 to 0 of 0 entries

2nd Step: Tạo MapReduce Job

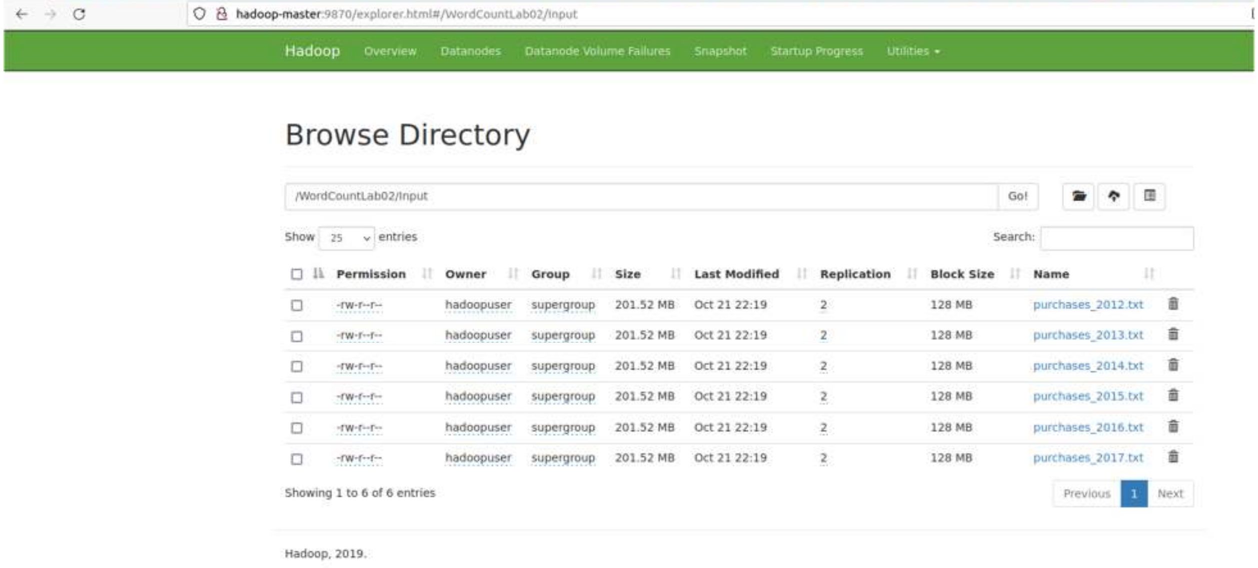
Chuẩn bị data



- Updata lên HDFS folder /WordCountLab02/Input



```
/input_data/purchases_2012.txt' '/home/hadoopuser/Desktop/ThangLong/input_data
/purchases_2013.txt' '/home/hadoopuser/Desktop/ThangLong/input_data/purchases_
201hadoopuser@hadoop-master:~$ hadoop fs -put '/home/hadoopuser/Desktop/ThangL
ong/input_data/purchases_2012.txt' '/home/hadoopuser/Desktop/ThangLong/input_d
ata/purchases_2013.txt' '/home/hadoopuser/Desktop/ThangLong/input_data/purchas
es_2014.txt' '/home/hadoopuser/Desktop/ThangLong/input_data/purchases_2015.txt
' '/home/hadoopuser/Desktop/ThangLong/input_data/purchases_2016.txt' '/home/ha
doopuser/Desktop/ThangLong/input_data/purchases_2017.txt' /WordCountLab02/Inpu
t
2021-10-21 22:19:40,601 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:42,627 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:43,541 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:45,057 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:45,994 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:47,541 INFO sasl.SaslDataTransferClient: SASL encryption trus
t check: localHostTrusted = false, remoteHostTrusted = false
2021-10-21 22:19:48,467 INFO sasl.SaslDataTransferClient: SASL encryption trus
```



Browse Directory

/WordCountLab02/Input Go!

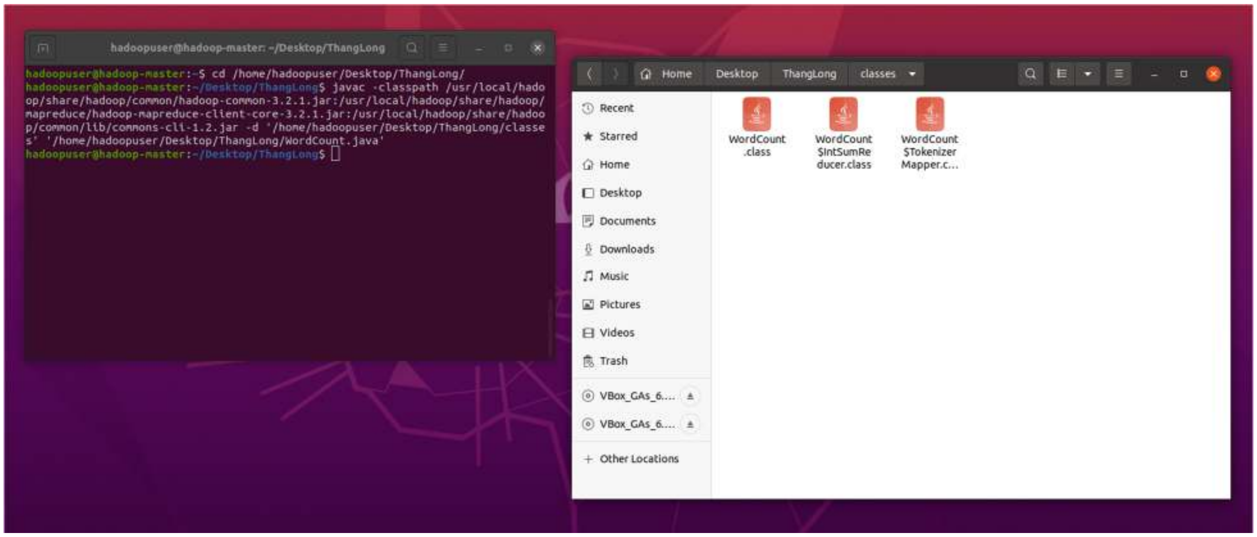
Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2012.txt
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2013.txt
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2014.txt
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2015.txt
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2016.txt
□	-rw-r--r--	hadoopuser	supergroup	201.52 MB	Oct 21 22:19	2	128 MB	purchases_2017.txt

Showing 1 to 6 of 6 entries

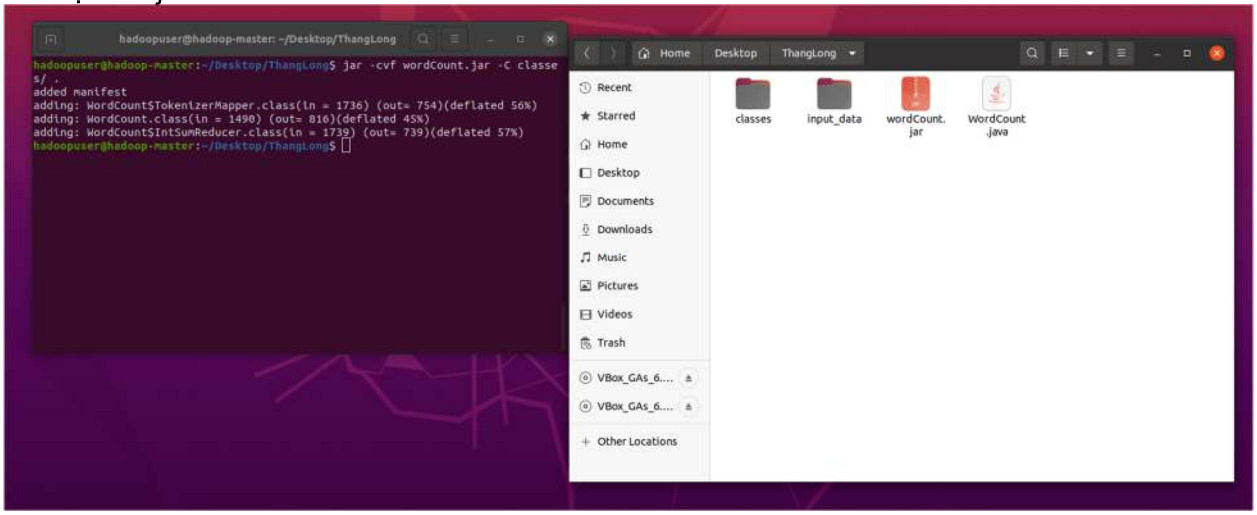
Hadoop, 2019.

- Thực thi file wordcount.java
- Xuất hiện 3 file class trong folder /classes



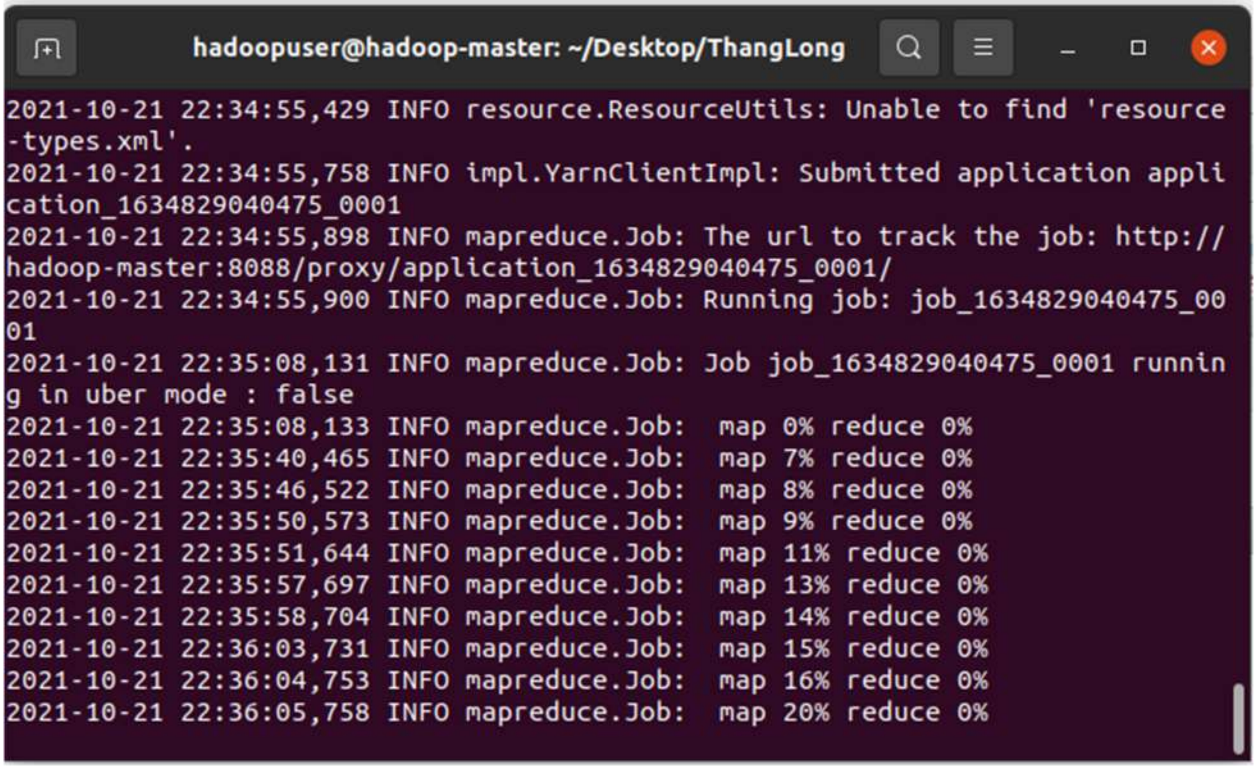
```
hadoopuser@hadoop-master:~$ cd /home/hadoopuser/Desktop/ThangLong/
hadoopuser@hadoop-master:~/Desktop/ThangLong$ javac -classpath /usr/local/hado
op/share/hadoop/common/hadoop-common-3.2.1.jar:/usr/local/hadoop/share/hadoop/
mapreduce/hadoop-mapreduce-client-core-3.2.1.jar:/usr/local/hadoop/share/hadoo
p/common/lib/commons-cli-1.2.jar -d '/home/hadoopuser/Desktop/ThangLong/classe
s' '/home/hadoopuser/Desktop/ThangLong/WordCount.java'
hadoopuser@hadoop-master:~/Desktop/ThangLong$ 
```

- Tạo file jar



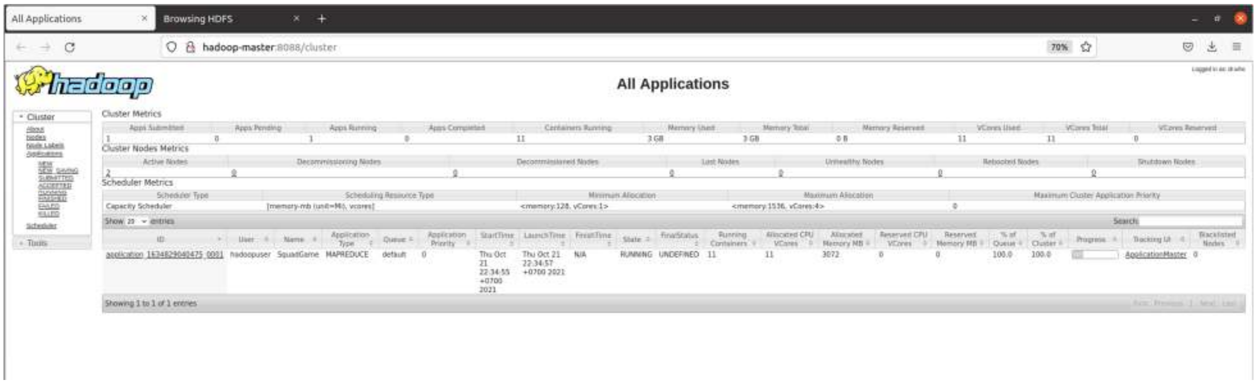
```
hadoopuser@hadoop-master:~/Desktop/ThangLong$ jar -cvf wordCount.jar -C classe
s/ .
added manifest
adding: WordCount$TokenizerMapper.class(in = 1736) (out= 754)(deflated 56%)
adding: WordCount.class(in = 1490) (out= 816)(deflated 45%)
adding: WordCount$IntSumReducer.class(in = 1739) (out= 739)(deflated 57%)
hadoopuser@hadoop-master:~/Desktop/ThangLong$ 
```

- Sử dụng file jar thực thi chạy job mapreduce



```
2021-10-21 22:34:55,429 INFO resource.ResourceUtils: Unable to find 'resource
-types.xml'.
2021-10-21 22:34:55,758 INFO impl.YarnClientImpl: Submitted application appli
cation_1634829040475_0001
2021-10-21 22:34:55,898 INFO mapreduce.Job: The url to track the job: http://
hadoop-master:8088/proxy/application_1634829040475_0001/
2021-10-21 22:34:55,900 INFO mapreduce.Job: Running job: job_1634829040475_00
01
2021-10-21 22:35:08,131 INFO mapreduce.Job: Job job_1634829040475_0001 runnin
g in uber mode : false
2021-10-21 22:35:08,133 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-21 22:35:40,465 INFO mapreduce.Job:  map 7% reduce 0%
2021-10-21 22:35:46,522 INFO mapreduce.Job:  map 8% reduce 0%
2021-10-21 22:35:50,573 INFO mapreduce.Job:  map 9% reduce 0%
2021-10-21 22:35:51,644 INFO mapreduce.Job:  map 11% reduce 0%
2021-10-21 22:35:57,697 INFO mapreduce.Job:  map 13% reduce 0%
2021-10-21 22:35:58,704 INFO mapreduce.Job:  map 14% reduce 0%
2021-10-21 22:36:03,731 INFO mapreduce.Job:  map 15% reduce 0%
2021-10-21 22:36:04,753 INFO mapreduce.Job:  map 16% reduce 0%
2021-10-21 22:36:05,758 INFO mapreduce.Job:  map 20% reduce 0%
```

- Lên web, check job đang chạy



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
1	0	1	0	11	3 GB	3 GB	0 B	11	11	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics

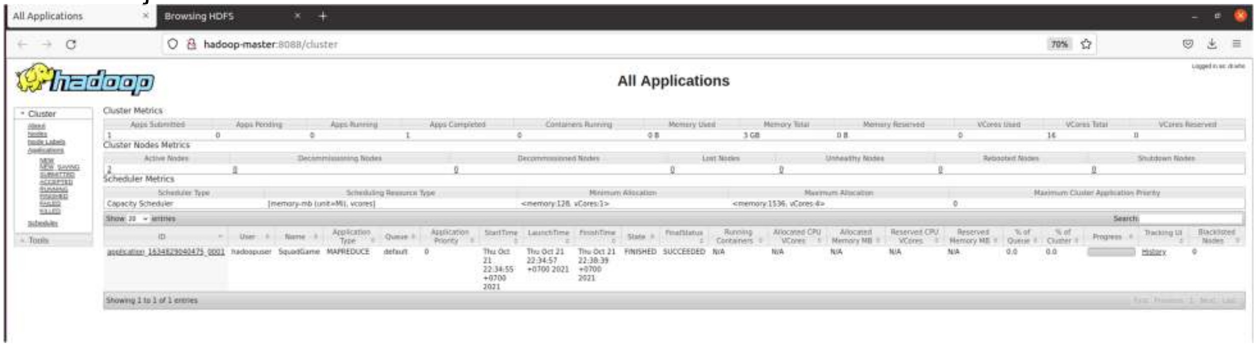
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:128, vCores:1>	<memory:1536, vCores:4>	0

Show 20 entries   Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1634829040475_0001	hadoopuser	SquadGame	MAPREDUCE	default	0	Thu Oct 21 22:34:55 +0700 2021	Thu Oct 21 22:34:57 +0700 2021	N/A	RUNNING	UNDEFINED	11	11	3072	0	0	100.0	100.0		ApplicationMaster	0

Showing 1 to 1 of 1 entries

- Check job finished



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
1	0	0	1	0	0 B	3 GB	0 B	0	16	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics

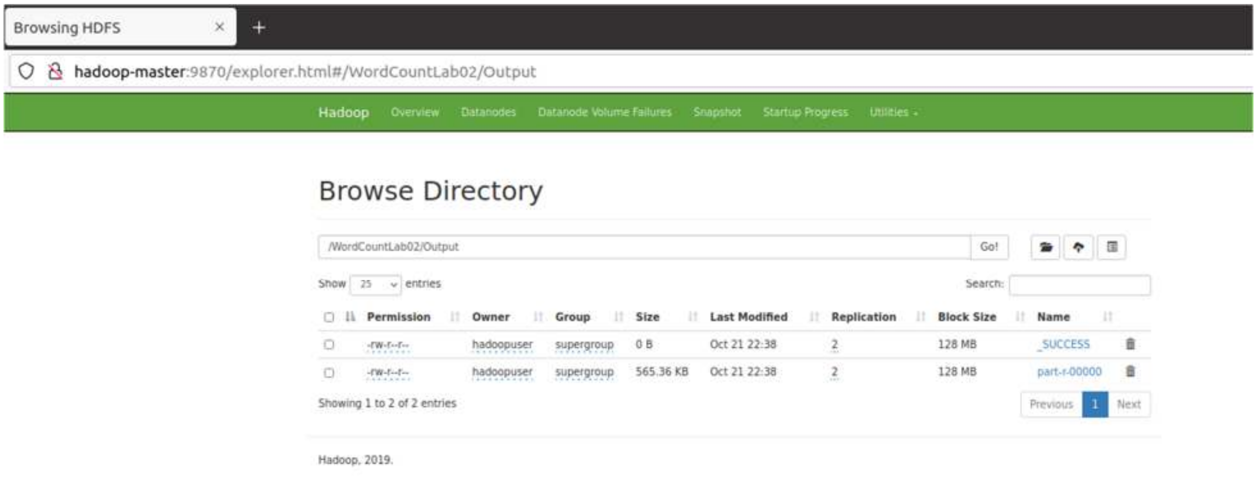
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:128, vCores:1>	<memory:1536, vCores:4>	0

Show 20 entries   Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1634829040475_0001	hadoopuser	SquadGame	MAPREDUCE	default	0	Thu Oct 21 22:34:55 +0700 2021	Thu Oct 21 22:34:57 +0700 2021	Thu Oct 21 22:38:39 +0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0

Showing 1 to 1 of 1 entries

- Xem kết quả



Browse Directory

/WordCountLab02/Output Go!

Show 25 entries Search:

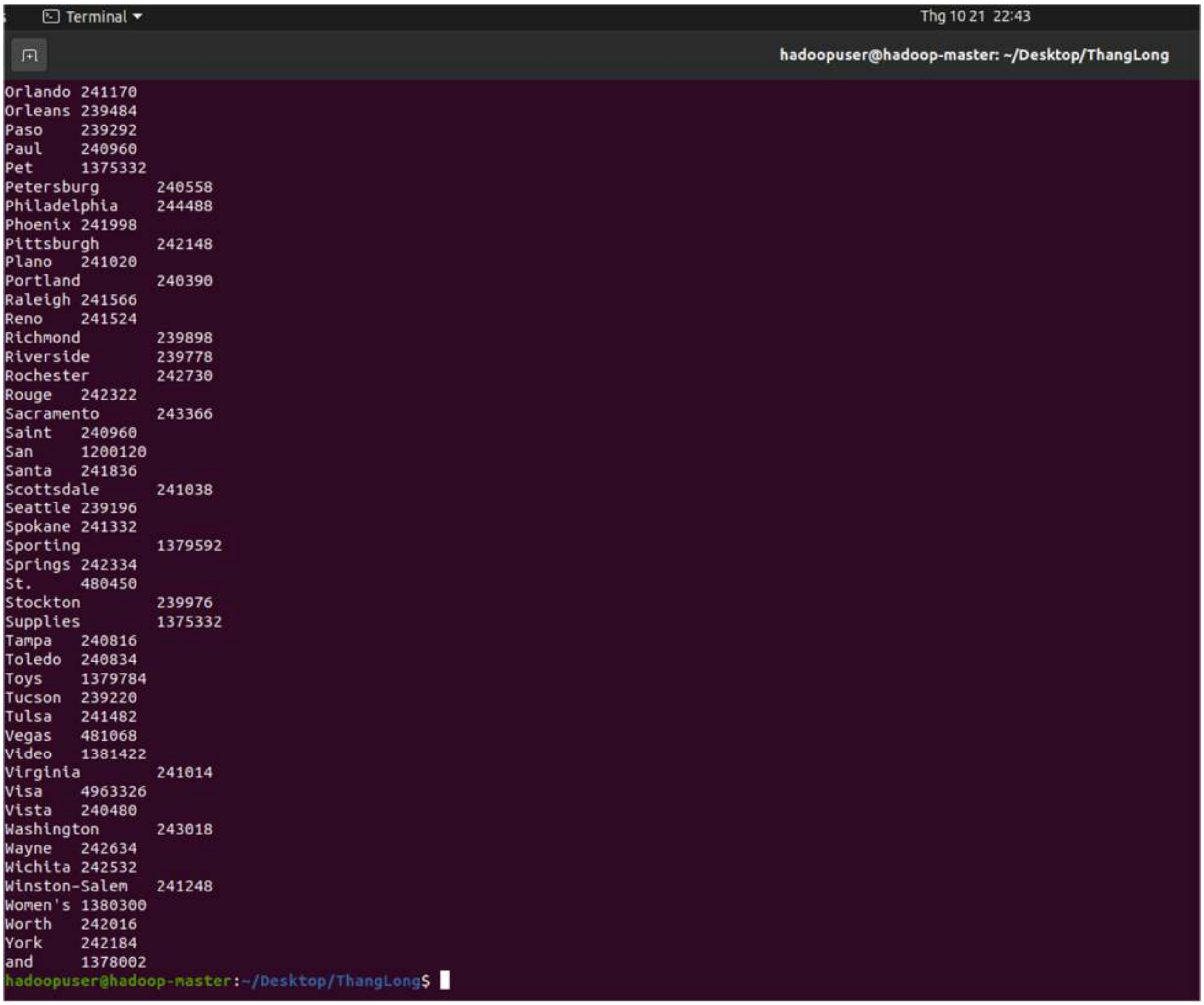
☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	hadoopuser	supergroup	0 B	Oct 21 22:38	2	128 MB	_SUCCESS
☐	-rw-r--r--	hadoopuser	supergroup	565.36 KB	Oct 21 22:38	2	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2019.

Chạy lệnh:
Hdfs dfs -cat /WordCountLab02/Output/*



```
Orlando 241170
Orleans 239484
Paso    239292
Paul    240960
Pet     1375332
Petersburg      240558
Philadelphia    244488
Phoenix 241998
Pittsburgh      242148
Plano   241020
Portland        240390
Raleigh 241566
Reno    241524
Richmond        239898
Riverside       239778
Rochester       242730
Rouge   242322
Sacramento      243366
Saint   240960
San     1200120
Santa   241836
Scottsdale      241038
Seattle 239196
Spokane 241332
Sporting        1379592
Springs 242334
St.     480450
Stockton        239976
Supplies        1375332
Tampa   240816
Toledo  240834
Toys    1379784
Tucson  239220
Tulsa   241482
Vegas   481068
Video   1381422
Virginia        241014
Visa    4963326
Vista   240480
Washington      243018
Wayne   242634
Wichita 242532
Winston-Salem   241248
Women's 1380300
Worth   242016
York    242184
and     1378002
hadoopuser@hadoop-master:~/Desktop/ThangLong$ 
```