

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



BÀI TẬP TRÊN LỚP
MÔN HỌC: LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN
LAB 2: MAPREDUCE

NHÓM: GIỮA CHÚNG TA

Hà Nội, 11-2020

Set Up a MapReduce in Hadoop 3.2.1 Multi-Node Cluster on Ubuntu

INPUT DATA

2012-01-01 12:01 San Jose Music 12.99 Amex

I. Trên 3 máy master, slave1, slave2

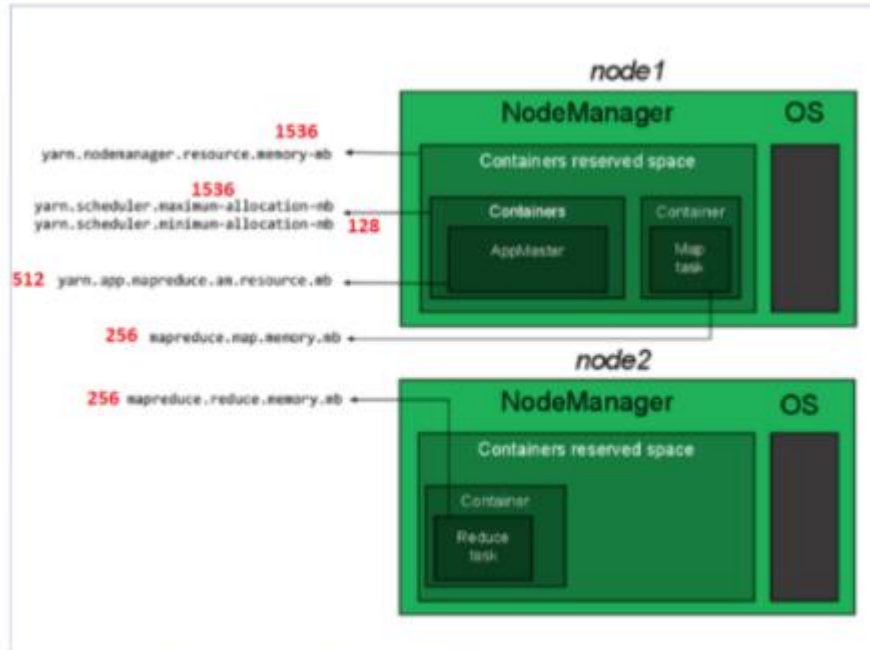


Figure 8: Sample config for 2GB Nodes (will change in next lab)

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>
```

```
hadoopuser@hadoop-master: ~
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>
```

```
hadoopuser@hadoop-master: ~  
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
```

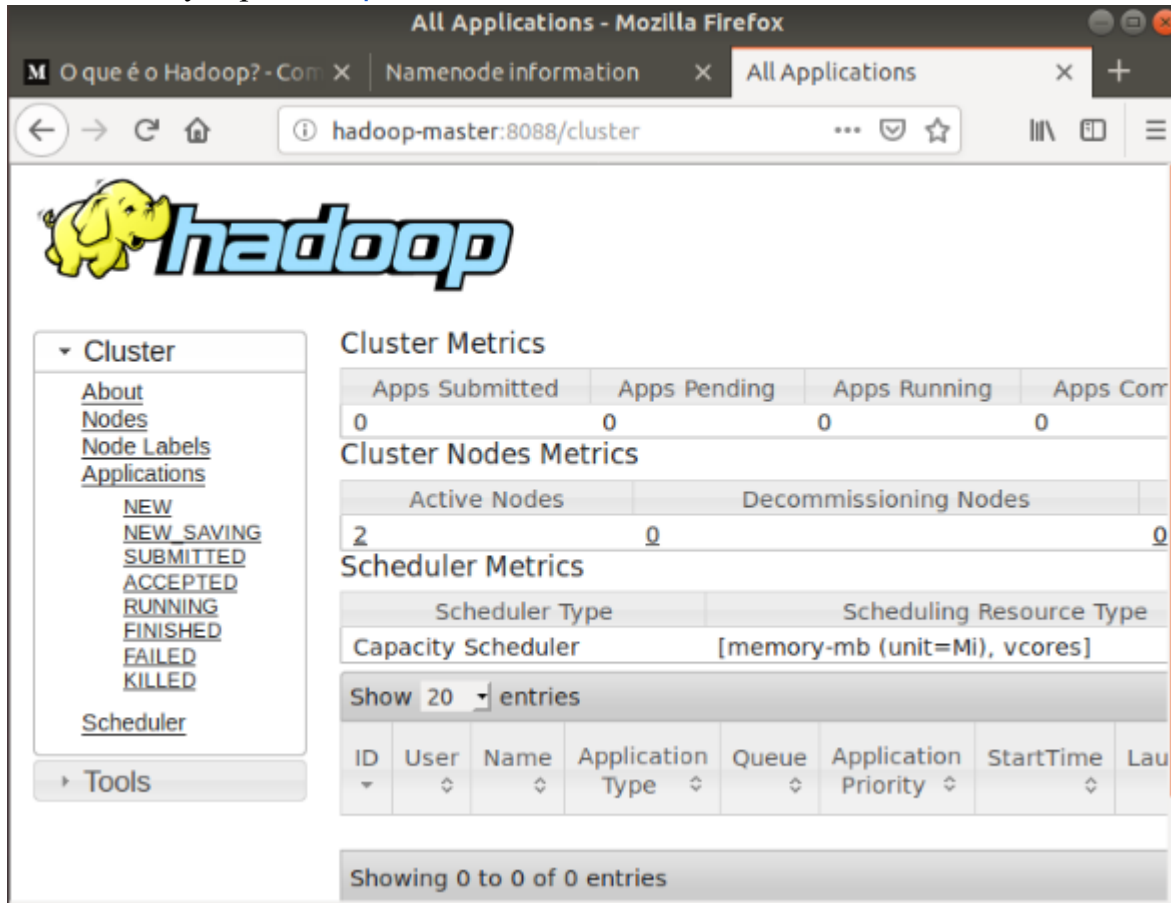
II. Trên máy master

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



The screenshot shows the Hadoop web interface in a Mozilla Firefox browser window. The address bar displays `hadoop-master:8088/cluster`. The interface includes the Hadoop logo and a sidebar with navigation links: Cluster, About, Nodes, Node Labels, Applications, NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler, and Tools. The main content area displays several metrics:

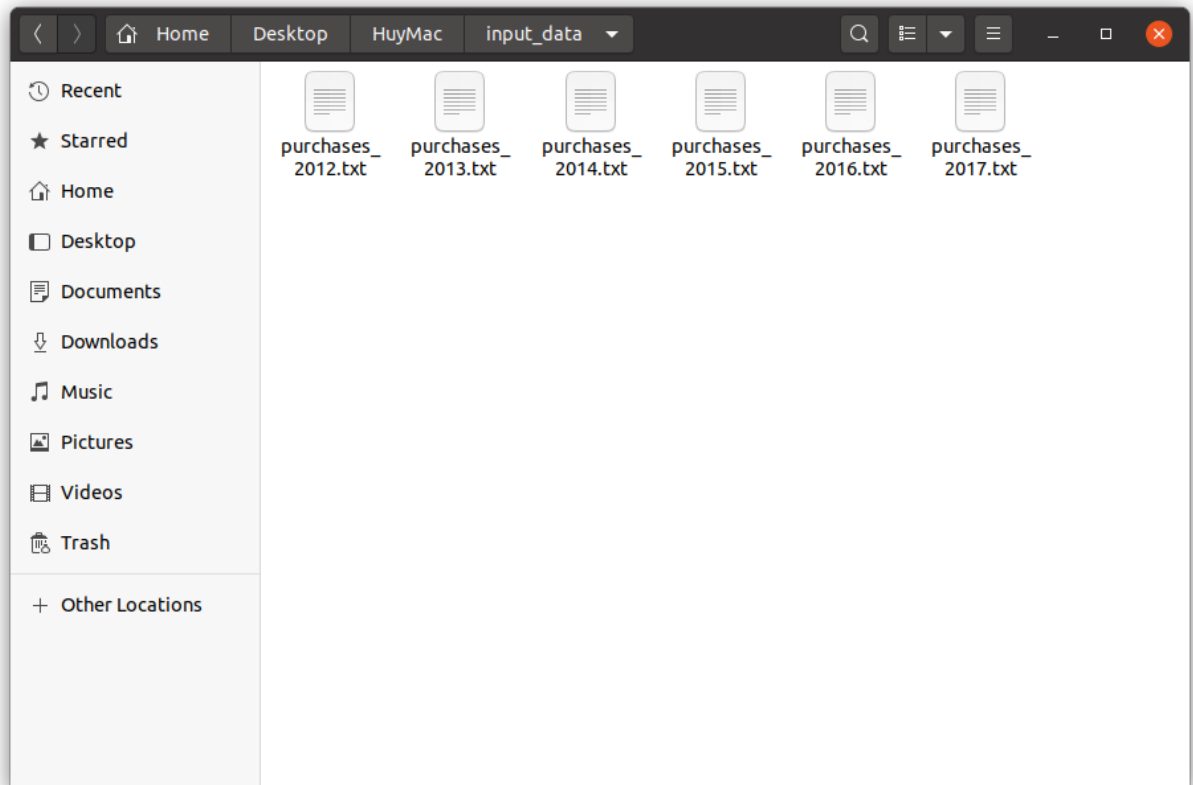
- Cluster Metrics:** A table with 4 columns: Apps Submitted (0), Apps Pending (0), Apps Running (0), and Apps Completed (0).
- Cluster Nodes Metrics:** A table with 2 columns: Active Nodes (2) and Decommissioning Nodes (0).
- Scheduler Metrics:** A table with 2 columns: Scheduler Type (Capacity Scheduler) and Scheduling Resource Type ([memory-mb (unit=Mi), vcores]).

Below these metrics is a table for applications, with a 'Show 20 entries' dropdown. The table has 8 columns: ID, User, Name, Application Type, Queue, Application Priority, StartTime, and Last Update Time. The status at the bottom indicates 'Showing 0 to 0 of 0 entries'.

2nd Step: Tạo MapReduce Job

- Chuẩn bị data: Tài trên https://github.com/MacHuy/HDFS-MultiNode/blob/main/HuyMac/input_data/test.txt

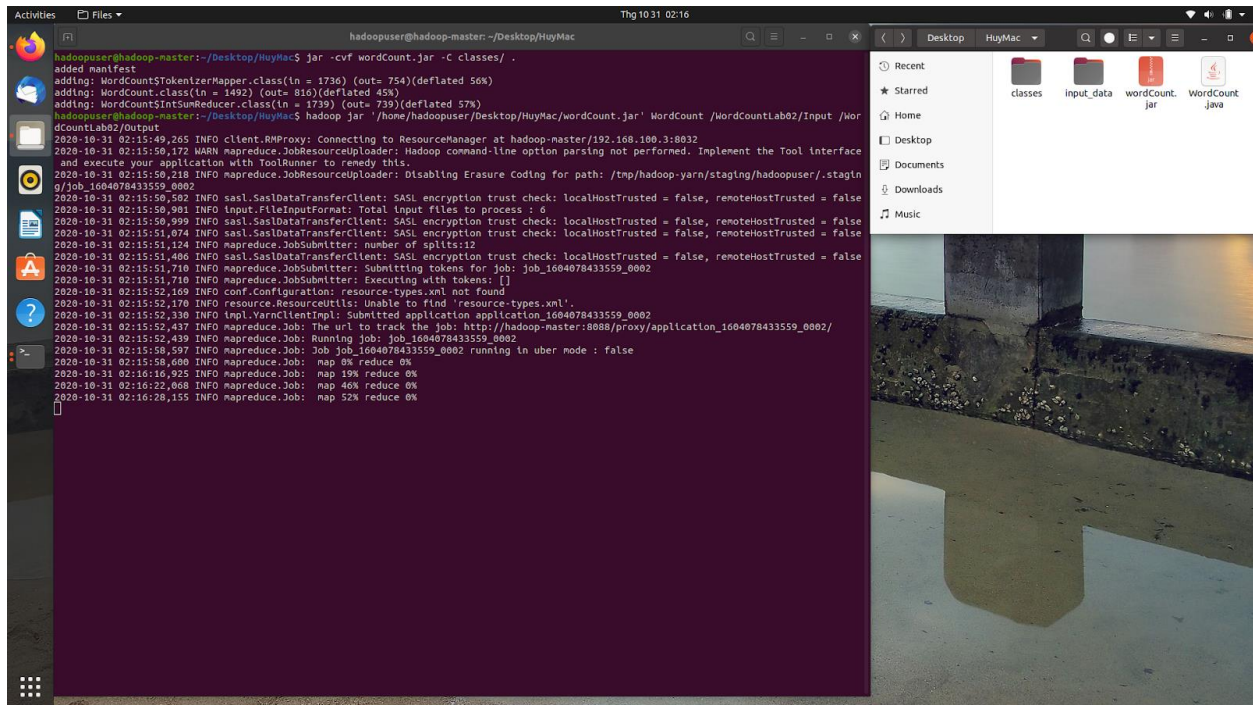
- Hoặc có thể toàn bộ folder HuyMac và move nó để ở Desktop:
<https://github.com/MacHuy/HDFS-MultiNode>



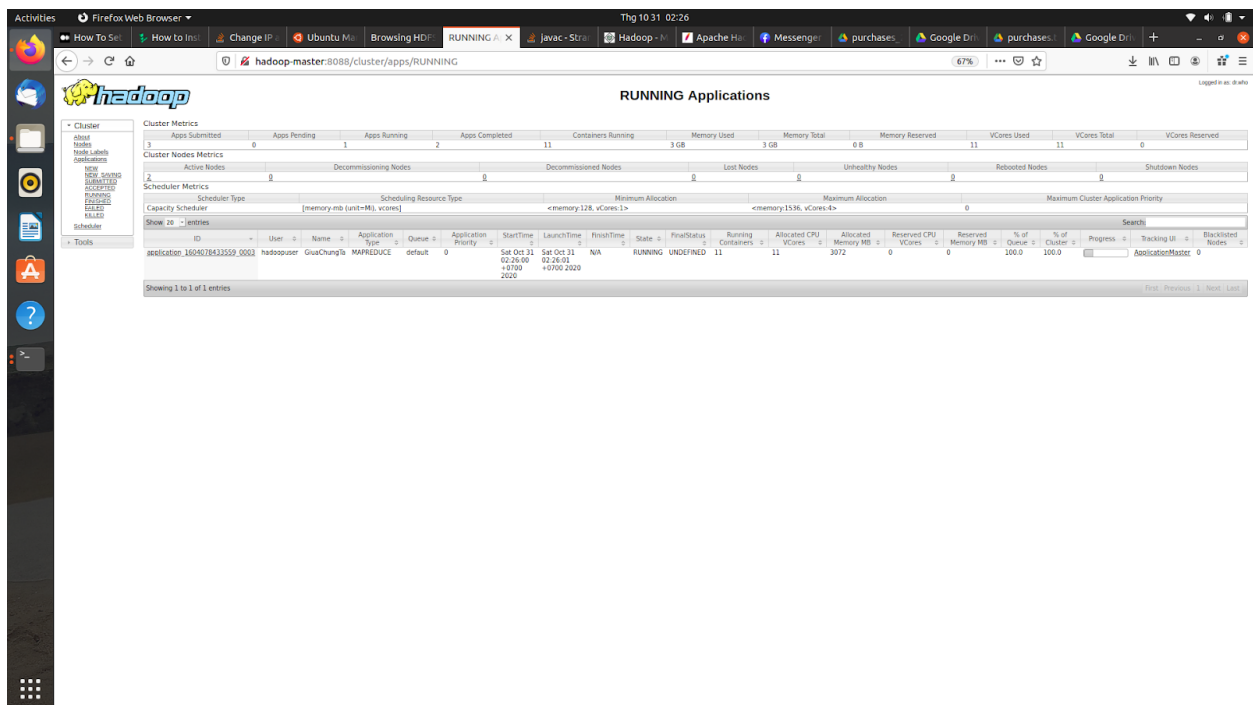
-
- Updata lên HDFS folder /WordCountLab02/Input
Thực thi file wordcount.java

```
hadoopuser@hadoop-master: ~/Desktop/HuyMac
hadoopuser@hadoop-master:~$ hadoop fs -mkdir /WordCountLab02
hadoopuser@hadoop-master:~$ hadoop fs -mkdir /WordCountLab02/Input
hadoopuser@hadoop-master:~$ hadoop fs -put '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2012.txt' '/home/hadoopuser/Desktop/HuyMac/in
put_data/purchases_2013.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2014.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purch
ases_2015.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2016.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2017.txt'
/WordCountLab02/Input
2020-10-31 01:51:19,744 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:52:25,849 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:53:03,786 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:54:14,978 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:54:55,215 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:56:03,635 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:56:41,432 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:57:51,258 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:58:29,327 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 01:59:42,324 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 02:00:24,388 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2020-10-31 02:01:55,751 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
hadoopuser@hadoop-master:~$ cd /home/hadoopuser/Desktop/HuyMac/
hadoopuser@hadoop-master:~/Desktop/HuyMac$ javac -classpath /usr/local/hadoop/share/hadoop/common/hadoop-common-3.2.1.jar:/usr/local/hadoop/sh
are/hadoop/mapreduce/hadoop-mapreduce-client-core-3.2.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar -d '/home/hadoopuser
/Desktop/HuyMac/classes' '/home/hadoopuser/Desktop/HuyMac/WordCount.java'
hadoopuser@hadoop-master:~/Desktop/HuyMac$
```

- Xuất hiện 3 file class trong folder /classes



- Lên web, check job đang chạy



- Xem kết quả


```
hadoopuser@hadoop-master:~$ hdfs dfs -cat /WordCountTutorial/Output/*

Nashville 239124
New 481668
Newark 243462
Norfolk 241386
North 240078
Oakland 238368
Oklahoma 242676
Omaha 241254
Orlando 241170
Orleans 239484
Paso 239292
Paul 240960
Pet 1375332
Petersburg 248558
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/HuyMac$
```

TH chạy lại job, nhớ đổi sang file output mới, folder output cũ đã có, HDFS ko cho phép overwrite

```
hadoopuser@hadoop-master:~/Desktop/HuyMac$ hdfs dfs -cat /WordCountTutorial/Output/*

Map Input records=24830856
Map output records=167897378
Map output bytes=193947024
Map output materialized bytes=7735584
Input split bytes=1560
Combine input records=170928474
Combine output records=3641556
Reduce input groups=52878
Reduce shuffle bytes=7755584
Reduce input records=610452
Reduce output records=52878
Spilled Records=625000
Shuffled Maps =12
Failed Shuffles=0
Merged Map outputs=12
CPU time elapsed (ms)=9980
GC time elapsed (ms)=351980
Physical memory (bytes) snapshot=3755573248
Virtual memory (bytes) snapshot=24765589632
Total committed heap usage (bytes)=2697463740
Peak Map Physical memory (bytes)=309309440
Peak Map Virtual memory (bytes)=1915801824
Peak Reduce Physical memory (bytes)=210483248
Peak Reduce Virtual memory (bytes)=1921073152

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=1267982120
File Output Format Counters
Bytes Written=578924

hadoopuser@hadoop-master:~/Desktop/HuyMac$ hdfs dfs -cat /WordCountTutorial/Output/*
2020-10-31 02:27:28,552 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.1.80:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://hadoop-master:9000/WordCountLab02/Output2 already exists
at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:164)
at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:277)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1570)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1567)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.mapreduce.Job.doAs(Job.java:1730)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1580)
at WordCount.main(WordCount.java:59)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:333)
at org.apache.hadoop.util.RunJar.main(RunJar.java:236)

hadoopuser@hadoop-master:~/Desktop/HuyMac$
```

Trường hợp chỉ có 1 datanode slave1

Activities Firefox Web Browser Thg 10 31 10:12

hadoop-master:8088/cluster/apps/RUNNING

RUNNING Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	V-Cores Used	V-Cores Total	V-Cores Reserved
4	0	1	3	4	1.25 GB	1.50 GB	0 B	4	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	1	0

Scheduler Metrics

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

Showing 1 to 1 of 1 entries

- Check các job finished

Activities Firefox Web Browser Thg 10 31 10:13

hadoop-master:8088/cluster/apps/FINISHED

FINISHED Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	V-Cores Used	V-Cores Total	V-Cores Reserved
4	0	0	4	0	0 B	1.50 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	1	0

Scheduler Metrics

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

Showing 1 to 4 of 4 entries

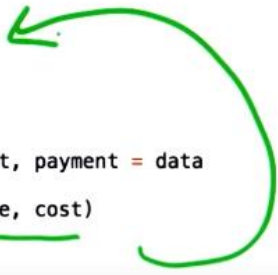
ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU V-Cores	Allocated Memory MB	Reserved CPU V-Cores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1604078433559_0004	hadoopuser	GiusChung% MAPREDUCE	MAPREDUCE	default	0	Sat Oct 31 10:10:20 +0700 2020	Sat Oct 31 10:10:21 +0700 2020	Sat Oct 31 10:12:03 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1604078433559_0003	hadoopuser	GiusChung% MAPREDUCE	MAPREDUCE	default	0	Sat Oct 31 02:16:00 +0700 2020	Sat Oct 31 02:16:01 +0700 2020	Sat Oct 31 02:16:58 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1604078433559_0002	hadoopuser	GiusChung% MAPREDUCE	MAPREDUCE	default	0	Sat Oct 31 02:15:12 +0700 2020	Sat Oct 31 02:15:13 +0700 2020	Sat Oct 31 02:16:51 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1604078433559_0001	hadoopuser	word count MAPREDUCE	MAPREDUCE	default	0	Sat Oct 31 00:57:07 +0700 2020	Sat Oct 31 00:57:09 +0700 2020	Sat Oct 31 00:57:25 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	

Part2: MapReduce cho job “Tính tổng Sales/Store”

⇒ Với dữ liệu đầu vào phần trước chọn cặp (key,value) = (store name, cost)
Sử dụng Hadoop streaming để code bằng Python

- Mapper Code:

```
def mapper():  
    for line in sys.stdin:  
        data = line.strip().split("\t")  
  
        if len(data) == 6:  
            date, time, store, item, cost, payment = data  
            print "{0}\t{1}".format(store, cost)
```



- Giữa Mapper và Reducer: Sử dụng Shuffle and Sort

- Reducer Code

```
def reducer():  
    salesTotal = 0  
    oldKey = None  
  
    for line in sys.stdin:  
        data = line.strip().split("\t")  
  
        if len(data) != 2:  
            continue  
  
        thisKey, thisSale = data  
  
        if oldKey and oldKey != thisKey:  
            print "{0}\t{1}".format(oldKey, salesTotal)  
            salesTotal = 0  
  
        oldKey = thisKey  
        salesTotal += float(thisSale)  
  
    if oldKey != None:  
        print "{0}\t{1}".format(oldKey, salesTotal)
```



- Output

```
Cleveland      10067835.84
Colorado Springs 10061105.87
Columbus       10035241.03
Corpus Christi  9976522.77
Dallas         10066548.45
Denver         10031534.87
Detroit        9979260.76
Durham         10153890.21
El Paso        10016409.97
Fort Wayne     10132594.02
Fort Worth     10120830.65
Fremont        10053242.36
Fresno         9976260.26
Garland        10071043.92
Gilbert        10062115.19
Glendale       10044493.97
Greensboro     10033781.39
Henderson      10053416.05
Hialeah        10047052.76
Honolulu       10006273.49
Houston        10042106.27
Indianapolis   10090272.77
Irvine         10084867.45
Irving         10133944.08
```

Toàn bộ về Hadoop 3.2.1 Multi-Node Cluster and Mapreduce Job:

<https://github.com/MacHuy/HDFS-MultiNode>

どうもありがとう、先生