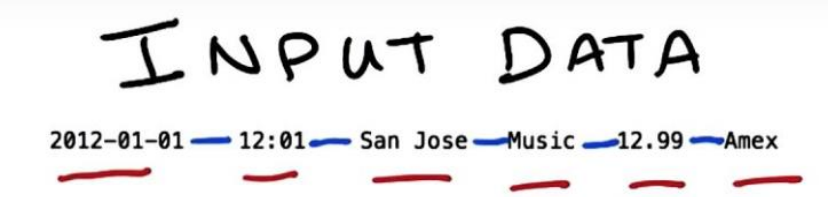
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

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



I. Trên 3 máy master, slave1, slave2

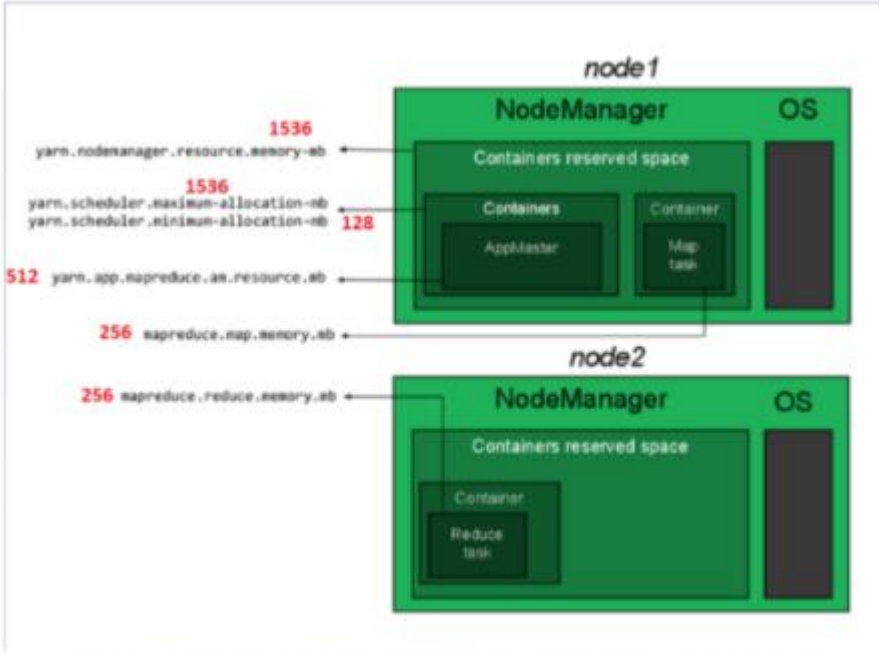
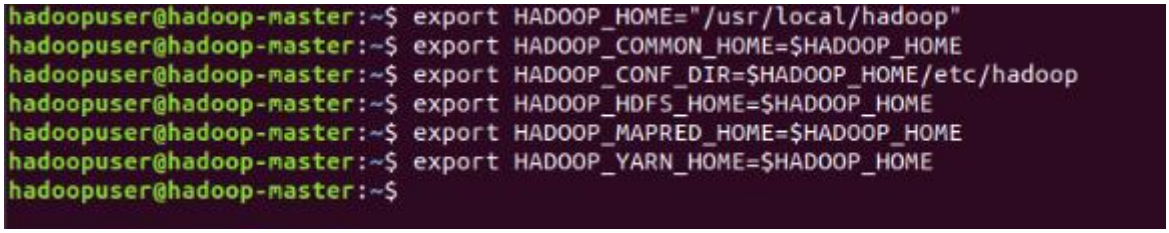


Figure 8: Sample config for 2GB Nodes (will change in next lab)

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
```



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>
```

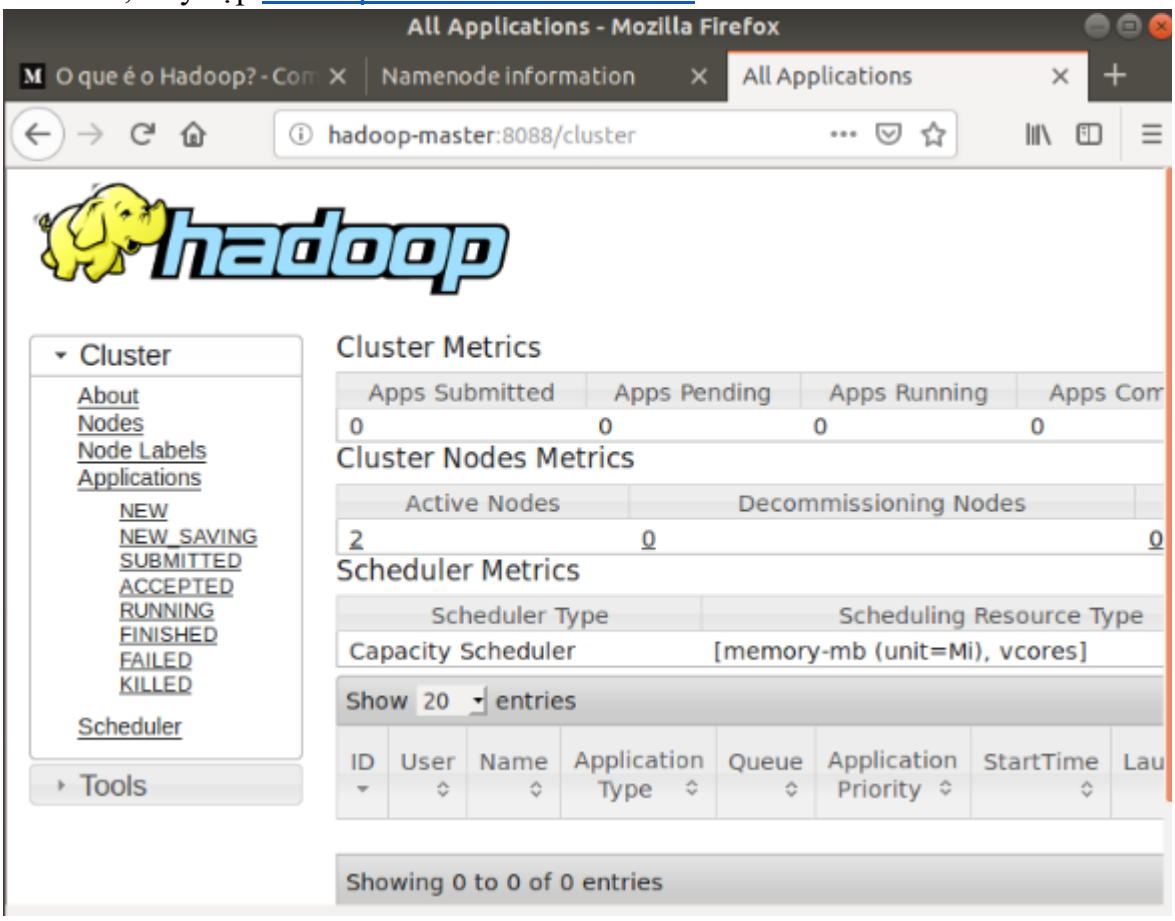
II. Trên máy master

1st Step: Start yarn

```
start-yarn.sh
```

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

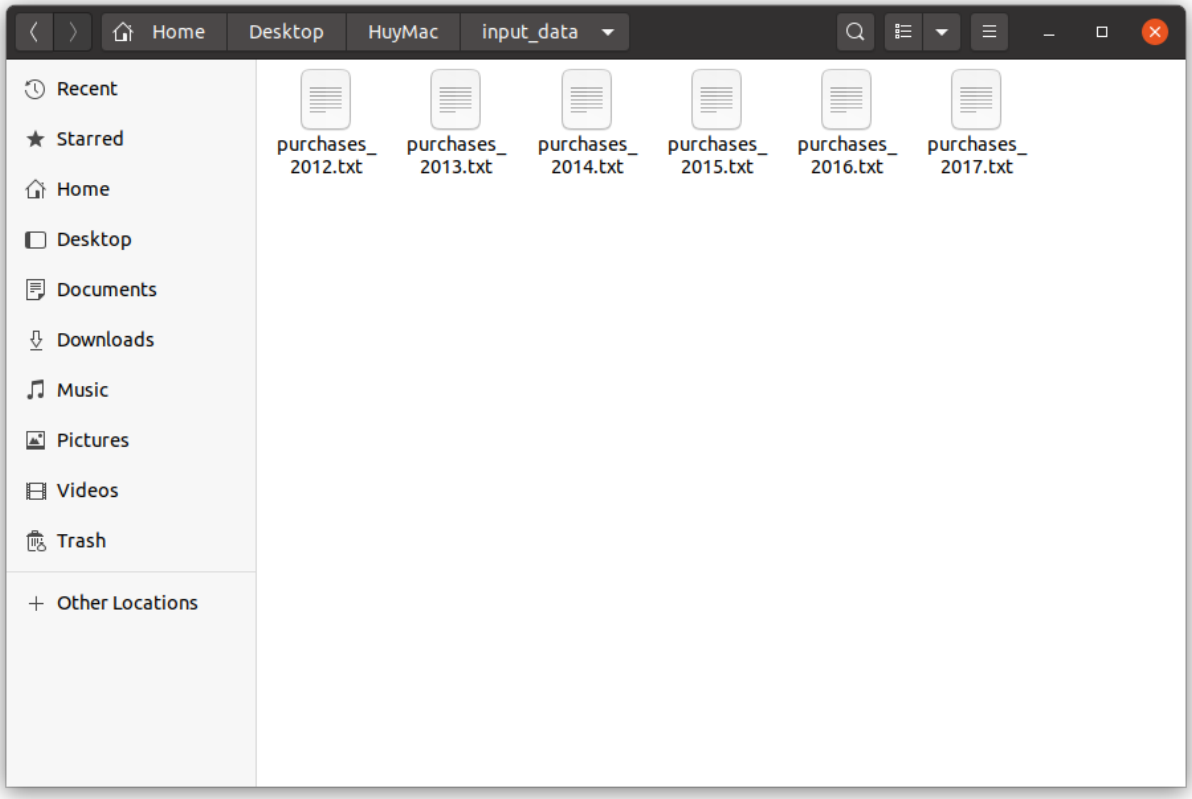
Lên web, truy cập hadoop-master:8088/cluster



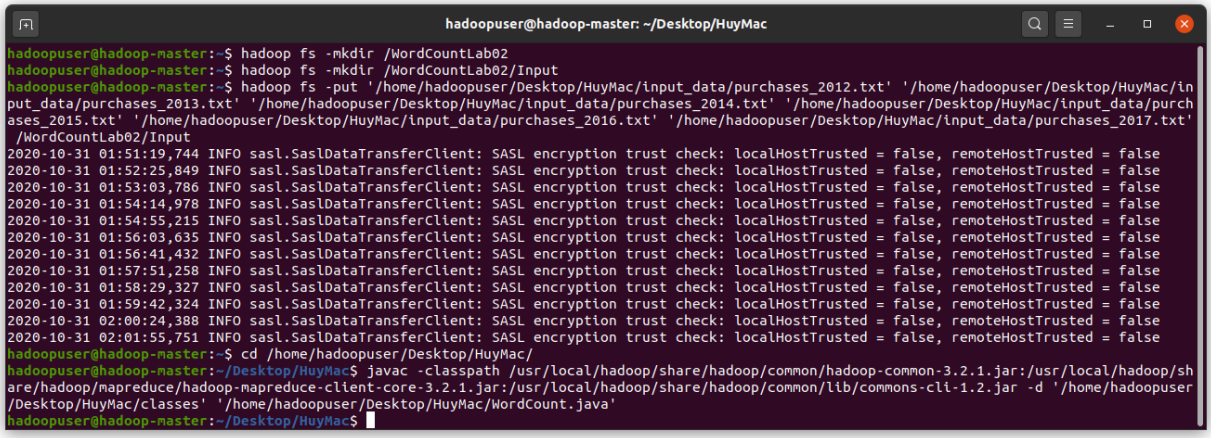
2nd Step: Tạo MapReduce Job

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

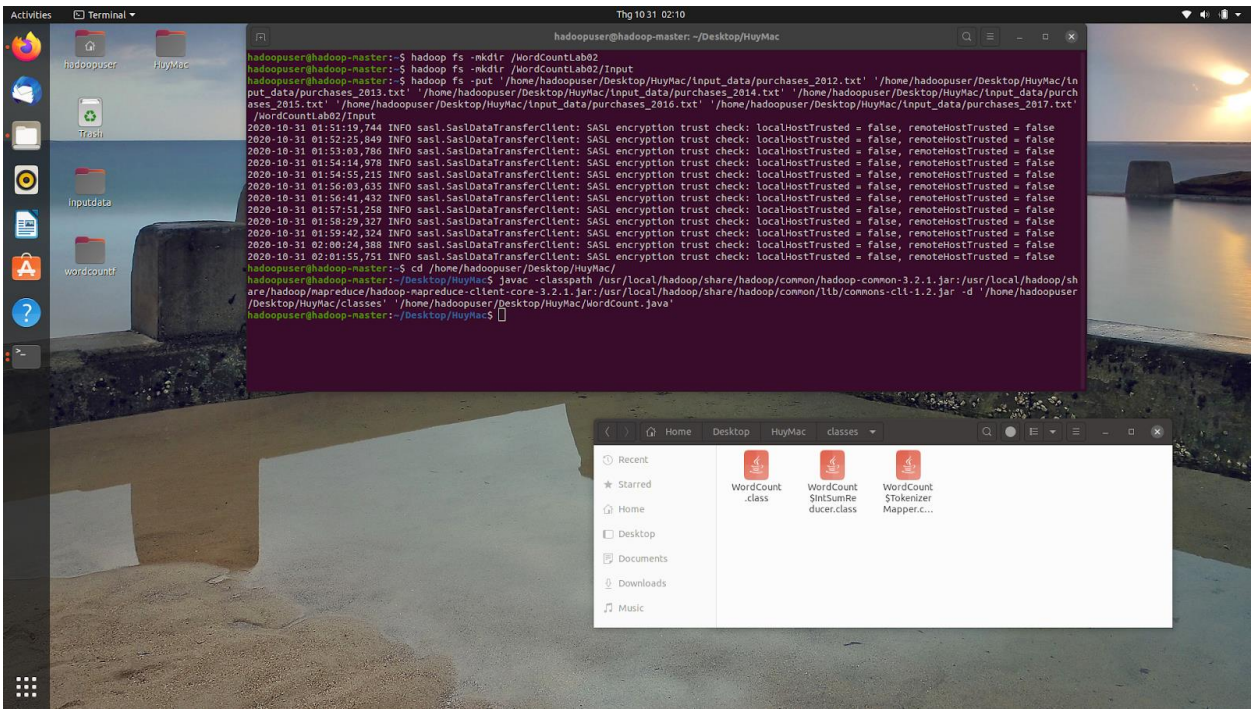
- Hoặc có thể toàn bộ folder HuyMac và move nó để ở Desktop: https://github.com/MacHuy/HDFS-MultiNode



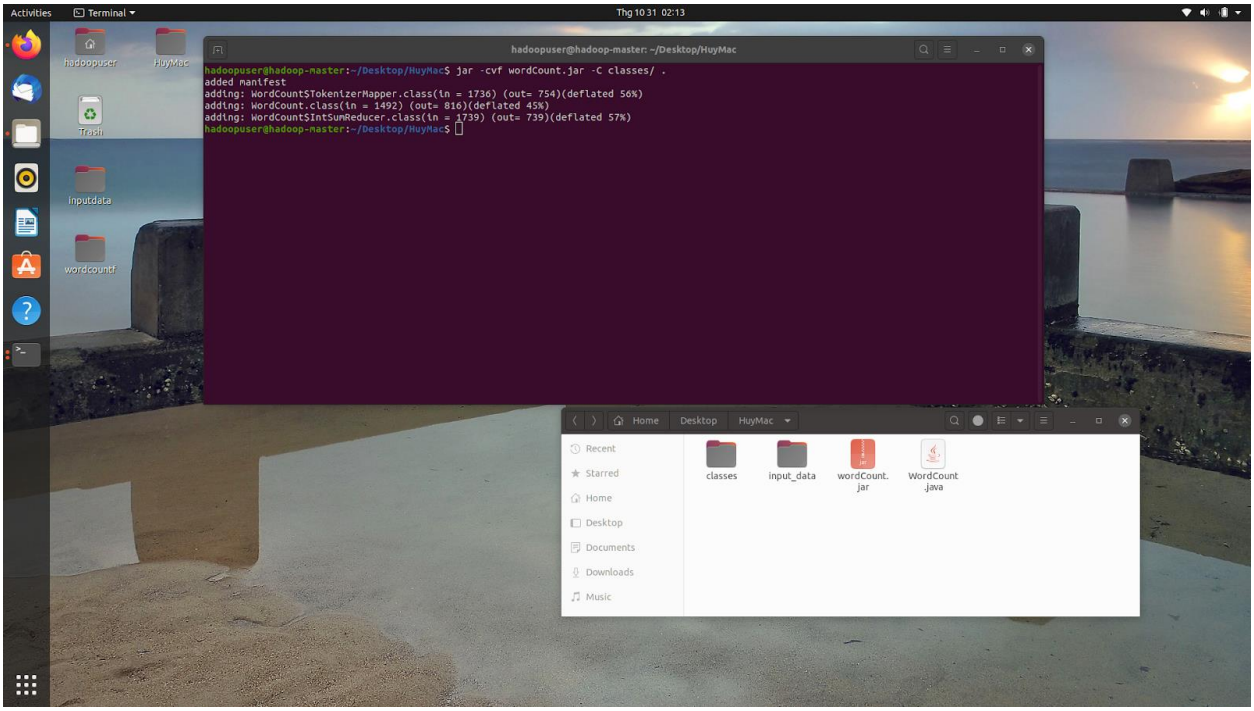
-
- Updata lên HDFS folder /WordCountLab02/Input
  Thực thi file wordcount.java



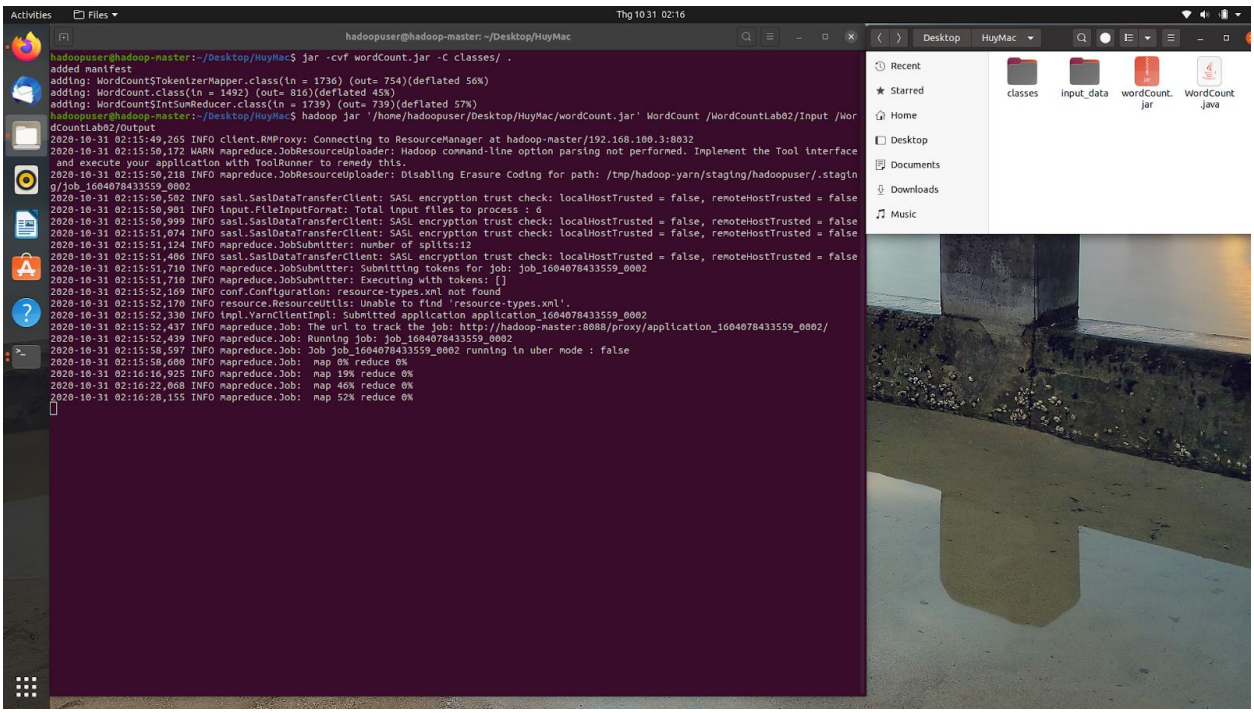
- Xuất hiện 3 file class trong folder /classes



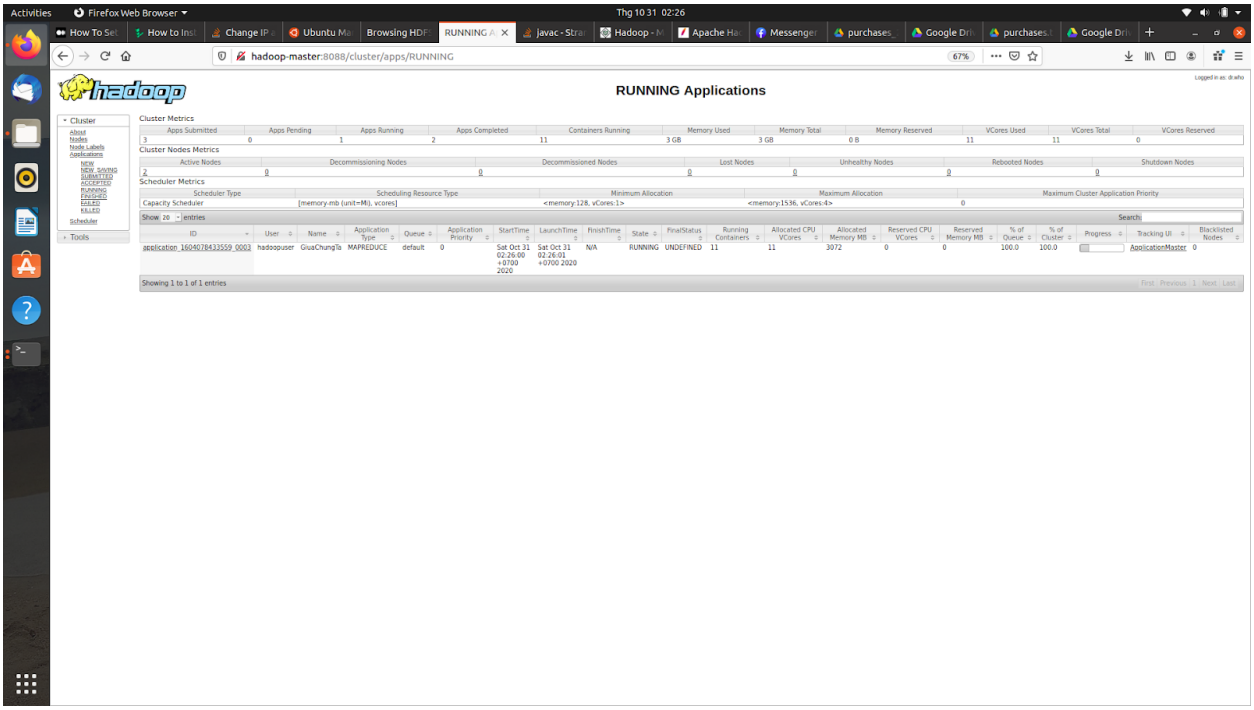
- Tạo file jar



- Sử dụng file jar thực thi chạy job mapreduce



- Lên web, check job đang chạy



- Xem kết quả

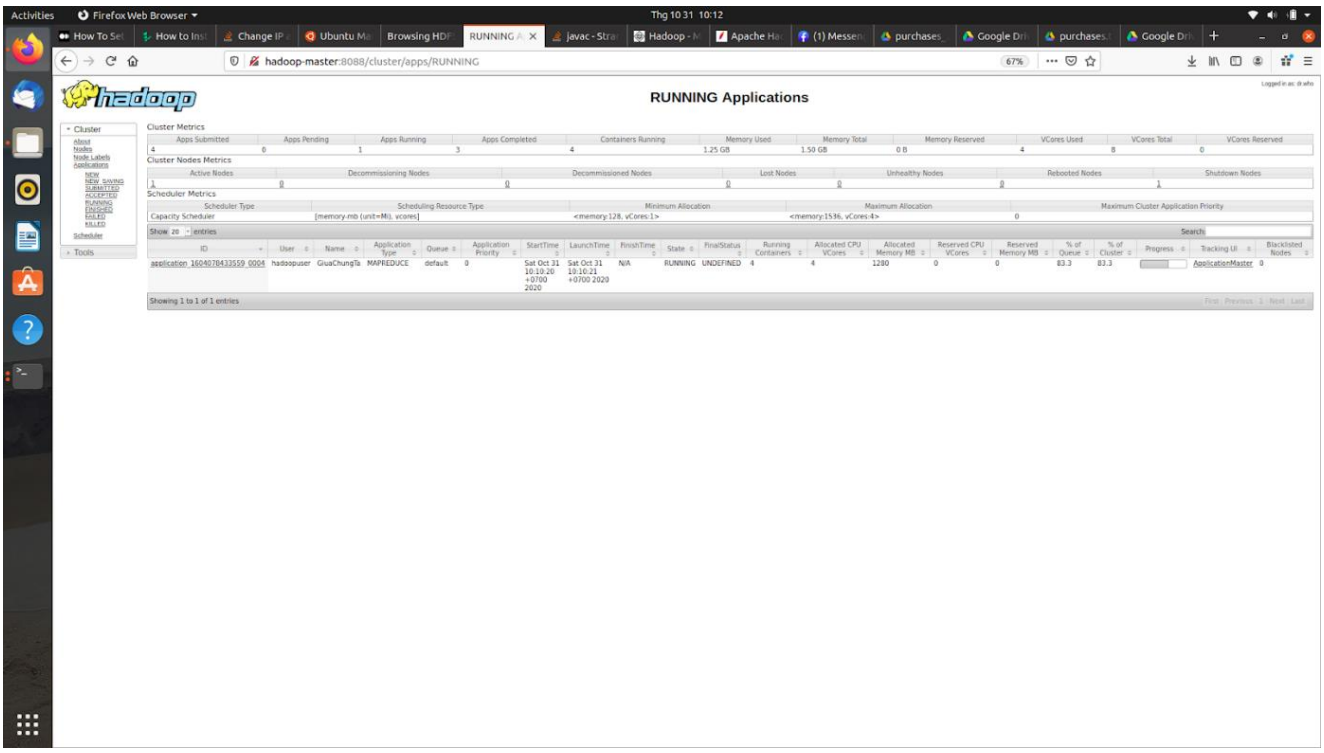
```
hadoopuser@hadoop-master:~$ hdfs dfs -cat /WordCountTutorial/Output/*
```

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

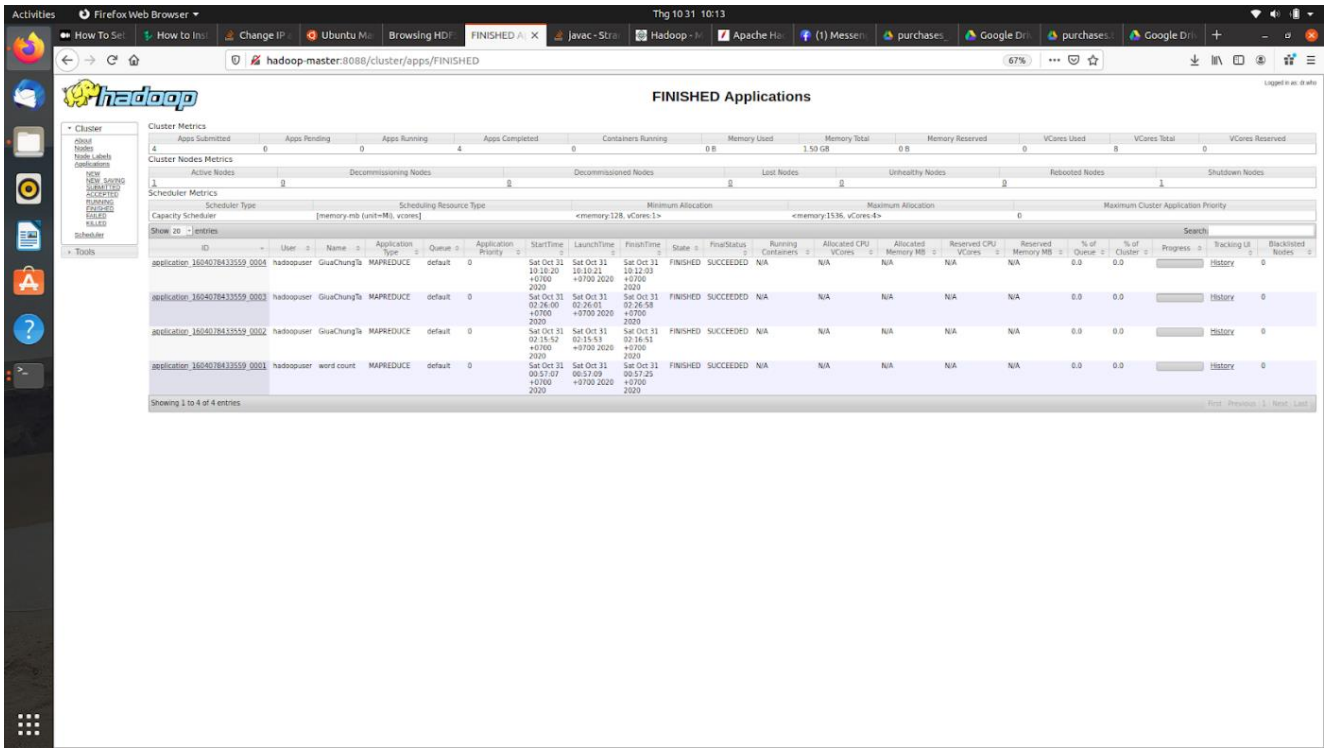
TH chạy lại job, nhớ đổi sang file output mới, folder output cũ đã có, HDFS ko cho phép overwrite

```
Activities   Terminal   Thg 10 31  02:27
hadoopuser@hadoop-master: ~/Desktop/HuyMac
                Map input records=24830856
                Map output records=167897378
                Map output bytes=1939467024
                Map output materialized bytes=7735584
                Input split bytes=1560
                Combine input records=170928474
                Combine output records=3641556
                Reduce input groups=52878
                Reduce shuffle bytes=7735584
                Reduce input records=610452
                Reduce output records=52878
                Spilled Records=4252008
                Shuffled Maps =12
                Failed Shuffles=0
                Merged Map outputs=12
                GC time elapsed (ms)=9980
                CPU time spent (ms)=351980
                Physical memory (bytes) snapshot=3755573248
                Virtual memory (bytes) snapshot=24765509632
                Total committed heap usage (bytes)=2697461760
                Peak Map Physical memory (bytes)=309309440
                Peak Map Virtual memory (bytes)=1915981824
                Peak Reduce Physical memory (bytes)=210485248
                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=1267902120
        File Output Format Counters
                Bytes Written=578924
hadoopuser@hadoop-master:~/Desktop/HuyMac$ hadoop jar '/home/hadoopuser/Desktop/HuyMac/wordCount.jar' WordCount /WordCountLab02/Input /WordCountLab02/Output2
2020-10-31 02:27:28,552 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/192.168.100.3:8032
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$11.run(Job.java:1567)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
        at WordCount.main(WordCount.java:59)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
hadoopuser@hadoop-master:~/Desktop/HuyMac$
```

Trường hợp chỉ có 1 datanode slave1



- Check các job finished



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