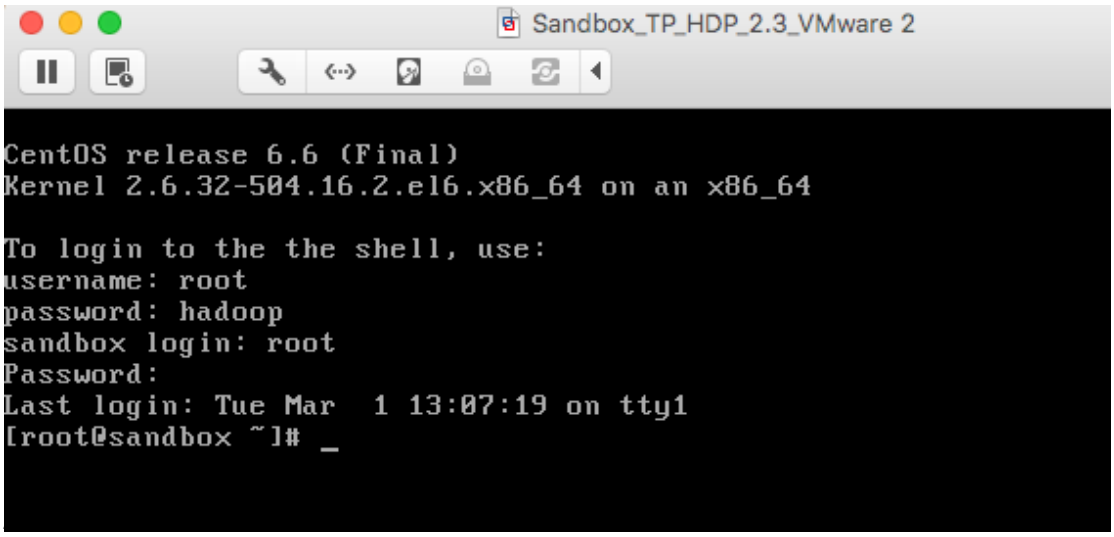
R·I·T

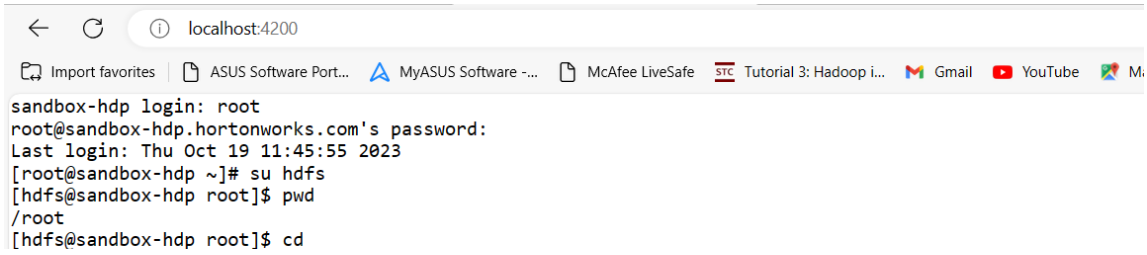
Lab: MapReduce Practical Exercise using Hadoop

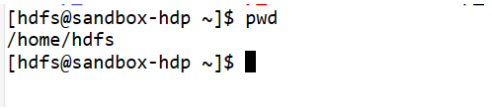
Introduction

Start Hortonworks Sandbox VM – Start it fresh after reboot

Open Terminal **localhost:4200**



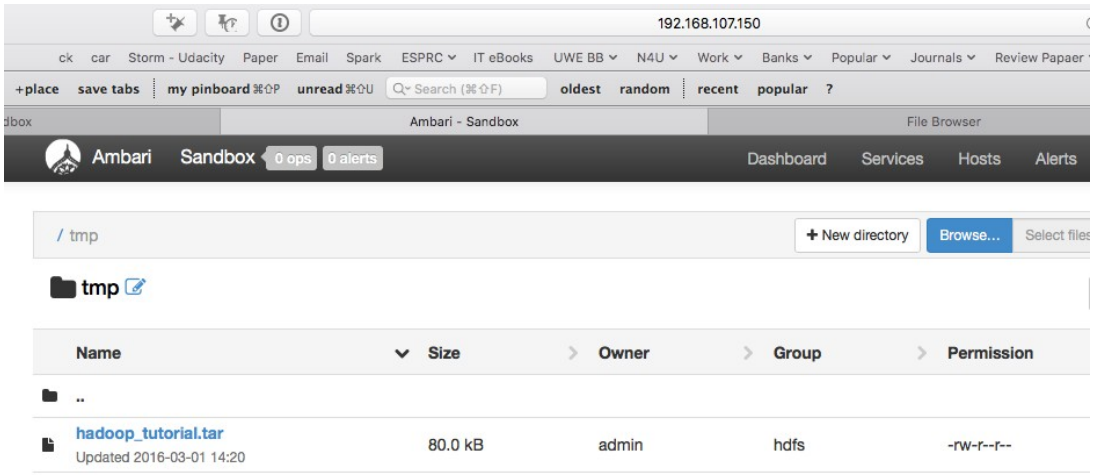




Activity 1 – Downloading the data

You can download the data files from MyCourses.
Download the tar file **hadoop_tutorial.tar**

Now copy hadoop_tutorial.tar folder into /tmp using HORTON WORK WEB INTERFACE



Now we will move the hadoop tutorials to the /home/hdfs

HINT: NOW WE CAN ACCESS HADOOP HDFS USING HADOOP FS COMMANDS, Search for "apache hadoop fs commands"
OR see this
https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/FileSystemShell.html

[hdfs@sandbox-hdp ~]$ hadoop fs –copyToLocal /tmp/hadoop_tutorial.tar /home/hdfs

[hdfs@sandbox-hdp ~]$ cd /home/hdfs
[hdfs@sandbox-hdp ~]$ tar xvf hadoop_tutorial.tar

// If any problem with permission then use
#chown -R hdfs:hdfs hadoop_tutorial/

R·I·T

Notes: -

- We have to Move the text files (to count words in HADOOP HDFS)

HINT: WE CAN ACCESS HADOOP HDFS USING HADOOP FS COMMANDS, see this https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

// use hadoop fs command, make folders /user/cloudera/wordcount/input

[hdfs@sandbox-hdp ~]$ hadoop fs -mkdir /user/cloudera/
[hdfs@sandbox-hdp ~]$ hadoop fs -mkdir /user/cloudera/wordcount
[hdfs@sandbox-hdp ~]$ hadoop fs -mkdir /user/cloudera/wordcount/input

```
[hdfs@sandbox ~]$ hadoop fs -mkdir /user/cloudera
[hdfs@sandbox ~]$ hadoop fs -mkdir /user/cloudera/wordcount
[hdfs@sandbox ~]$ hadoop fs -mkdir /user/cloudera/wordcount/input
[hdfs@sandbox ~]$ _
```

$cd /home/hdfs/hadoop_tutorial/WordCount1

// Now copy text files to Hadoop File System
[hdfs@sandbox-hdp~]$hadoop fs -copyFromLocal file0 /user/cloudera/wordcount/input
[hdfs@sandbox-hdp~]$ hadoop fs -copyFromLocal file1 /user/cloudera/wordcount/input
[hdfs@sandbox-hdp~]$ hadoop fs -copyFromLocal file2 /user/cloudera/wordcount/input

// Remove a directory, if already exists (created every time we run program)
#hadoop fs -rm -r /user/cloudera/wordcount/output

```
[hdfs@sandbox ~]$ pwd
/home/hdfs
[hdfs@sandbox ~]$ cd hadoop_tutorial
[hdfs@sandbox hadoop_tutorial]$ cd WordCount1
[hdfs@sandbox WordCount1]$ ls
build  file0  file1  file2  Makefile  wordcount.jar  WordCount.java
[hdfs@sandbox WordCount1]$ hadoop fs -copyFromLocal file0 /user/cloudera/wordcou
nt/input
[hdfs@sandbox WordCount1]$ hadoop fs -copyFromLocal file1 /user/cloudera/wordcou
nt/input
[hdfs@sandbox WordCount1]$ hadoop fs -copyFromLocal file2 /user/cloudera/wordcou
nt/input
[hdfs@sandbox WordCount1]$ _
```

Finally #Now run Jar
#hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/output

```
WARNING: Use "yarn jar" to launch YARN applications.
16/03/01 17:06:29 INFO impl.TimelineClientImpl: Timeline service address: http:/
/sandbox.hortonworks.com:8188/ws/v1/timeline/
16/03/01 17:06:29 INFO client.RMProxy: Connecting to ResourceManager at sandbox.
hortonworks.com/192.168.107.150:8050
16/03/01 17:06:30 INFO input.FileInputFormat: Total input paths to process : 3
16/03/01 17:06:30 INFO mapreduce.JobSubmitter: number of splits:3
16/03/01 17:06:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
56837593876_0002
16/03/01 17:06:30 INFO impl.YarnClientImpl: Submitted application application_14
56837593876_0002
16/03/01 17:06:30 INFO mapreduce.Job: The url to track the job: http://sandbox.h
ortonworks.com:8088/proxy/application_1456837593876_0002/
16/03/01 17:06:30 INFO mapreduce.Job: Running job: job_1456837593876_0002
16/03/01 17:06:43 INFO mapreduce.Job: Job job_1456837593876_0002 running in uber
 mode : false
16/03/01 17:06:43 INFO mapreduce.Job:  map 0% reduce 0%
16/03/01 17:06:58 INFO mapreduce.Job:  map 33% reduce 0%
16/03/01 17:06:59 INFO mapreduce.Job:  map 100% reduce 0%
```

```
        File Input Format Counters
                Bytes Read=87
        File Output Format Counters
                Bytes Written=84
hdfs@sandbox WordCount1]$ _
```

Result: See the following OUTPUT FOLDER using web interface

Visit Ambari. Select File View and then navigate the path

/user/cloudera/wordcount/output/part-r-00000

Open the file and the output is:

```
.       1
Hadoop  3
Oh      1
a       1
an      1
as      2
be      1
can     1
elephant        1
fellow  1
is      3
what    1
yellow  2
```

Next: Do the same exercise for WordCount2 and WordCount3 programs and discuss/present the difference.

*********END OF LAB*******