LAB 2 INSTRUCTIONS

DS8003 – MGT OF BIG DATA AND TOOLS

Ryerson University

Instructor: Kanchana Padmanabhan

Lab & Assignments

- **HDP Sandbox (VIRTUAL BOX LOGIN)**
  - SSH using Putty (host: localhost, port: 2222)
  - Username: root
  - Password: depends on what you set it
- Instructions for setup are in Lab 1

Download file

- Download the “shakespere_100.txt” file from Datasets & Scripts

Make a new directory on Virtual Machine

- SSH using Putty and Log-in to your Virtual Machine.
- Type the following commands

kpadmanabhan — root@sandbox:~/lab — ssh — 80×29

```
[root@sandbox ~]# mkdir lab
[root@sandbox ~]# cd lab/
```

- This will make a new directory called “lab.”
- Use “cd” command to navigate into the lab folder
- Move the file “shakespere_100.txt” into the “lab” folder on VirtualBox using FileZilla (See Lab 1 for instructions)

Today's lab – HDFS Shell Commands

- Step 1: Create a directory in HDFS,
- Step 2: Upload a file & List files in Directory
- Step 3: Try few basic Linux commands
- Step 4: Download Files From HDFS to Local File System
- Step 5: Find Out Space Utilization in a HDFS Directory
- Step 6: Explore Three Advanced Features
- Step 7: Use Help Command

Hadoop FileSystem Shell

Hadoop Filesystem Command	Description
hadoop fs -mkdir	create a new directory in hdfs
hadoop fs -ls	list files in a directory
hadoop fs -put	to copy file from local to hdfs
hadoop fs -cat	to preview the content of an hdfs file
hadoop fs -get	to move file from hdfs to local
hadoop fs -rmdir	to delete a directory
hadoop fs -cp	to make a copy of an hdfs file
hadoop fs –du	to display the size of an hdfs file
hadoop fs –mv	to move hdfs files from source to destination
hadoop fs –tail	to print the last few lines of an hdfs file/directory
hadoop fs –head	to print the first few lines of an hdfs file
hadoop fs –getmerge	to merge several hdfs files into one single file and copy to local

To learn more about Hadoop shell commands with Example Usage, check out the documentations
http://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-common/FileSystemShell.html
https://hadoop.apache.org/docs/r1.0.4/file_system_shell.html

Step 1: Create a folder on HDFS that can be accessed by the user "root"

- Let's create a folder on the HDFS that can be accessed as "root" user
- "hadoop fs" command let's us access the hdfs
- "-mkdir" creates a directory.
- Here we are creating a folder in the distributed file system
- "-chown" let's us change the owner of the folder from "hdfs" to "root"

kpadmanabhan — root@sandbox:~ — ssh — 80×44

```
[root@sandbox ~]# sudo -u hdfs hadoop fs -ls /user
Found 11 items
drwxrwx---   - ambari-qa hdfs          0 2015-10-27 12:39 /user/ambari-qa
drwxr-xr-x   - guest     guest         0 2015-10-27 12:55 /user/guest
drwxr-xr-x   - hcat      hdfs          0 2015-10-27 12:43 /user/hcat
drwx------   - hdfs      hdfs          0 2016-01-14 21:51 /user/hdfs
drwx------   - hive      hdfs          0 2016-01-15 19:51 /user/hive
drwxrwxrwx   - hue       hdfs          0 2015-10-27 12:55 /user/hue
drwxrwxr-x   - oozie     hdfs          0 2015-10-27 12:44 /user/oozie
drwxr-xr-x   - solr      hdfs          0 2015-10-27 12:48 /user/solr
drwxrwxr-x   - spark     hdfs          0 2015-10-27 12:41 /user/spark
drwxr-xr-x   - unit      hdfs          0 2015-10-27 12:46 /user/unit
drwxr-xr-x   - zeppelin  zeppelin      0 2015-10-27 12:10 /user/zeppelin
[root@sandbox ~]# sudo -u hdfs hadoop fs -mkdir /user/root
[root@sandbox ~]# sudo -u hdfs hadoop fs -ls /user
Found 12 items
drwxrwx---   - ambari-qa hdfs          0 2015-10-27 12:39 /user/ambari-qa
drwxr-xr-x   - guest     guest         0 2015-10-27 12:55 /user/guest
drwxr-xr-x   - hcat      hdfs          0 2015-10-27 12:43 /user/hcat
drwx------   - hdfs      hdfs          0 2016-01-14 21:51 /user/hdfs
drwx------   - hive      hdfs          0 2016-01-15 19:51 /user/hive
drwxrwxrwx   - hue       hdfs          0 2015-10-27 12:55 /user/hue
drwxrwxr-x   - oozie     hdfs          0 2015-10-27 12:44 /user/oozie
drwxr-xr-x   - hdfs      hdfs          0 2016-01-22 01:53 /user/root
drwxr-xr-x   - solr      hdfs          0 2015-10-27 12:48 /user/solr
drwxrwxr-x   - spark     hdfs          0 2015-10-27 12:41 /user/spark
drwxr-xr-x   - unit      hdfs          0 2015-10-27 12:46 /user/unit
drwxr-xr-x   - zeppelin  zeppelin      0 2015-10-27 12:10 /user/zeppelin
[root@sandbox ~]# sudo -u hdfs hadoop fs -chown root:root /user/root
[root@sandbox ~]# sudo -u hdfs hadoop fs -ls /user
Found 12 items
drwxrwx---   - ambari-qa hdfs          0 2015-10-27 12:39 /user/ambari-qa
drwxr-xr-x   - guest     guest         0 2015-10-27 12:55 /user/guest
drwxr-xr-x   - hcat      hdfs          0 2015-10-27 12:43 /user/hcat
drwx------   - hdfs      hdfs          0 2016-01-14 21:51 /user/hdfs
drwx------   - hive      hdfs          0 2016-01-15 19:51 /user/hive
drwxrwxrwx   - hue       hdfs          0 2015-10-27 12:55 /user/hue
drwxrwxr-x   - oozie     hdfs          0 2015-10-27 12:44 /user/oozie
drwxr-xr-x   - root      root          0 2016-01-22 01:53 /user/root
```

Step 2: Let's copy a file to the HDFS

- "-put" command let's you copy a file from local file system to the HDFS
- "-ls" command let's you copy a list all files in directory

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -put test.txt /user/root
[root@sandbox lab]# hadoop fs -ls /user/root
Found 1 items
-rw-r--r--   3 root root        318 2016-09-05 17:43 /user/root/test.txt
[root@sandbox lab]#
```

Architecture

Lab computer
Your PC/Mac

NOTE: Your client linux file system and hadoop filesystem (HDFS) are separate environment. First, you need to learn how to move files between the two file systems

SSH 127.0.0.1 –p 2222

You Are Here!

access node

$ hadoop fs –put test.txt /user/root/

$ hadoop fs –get /user/root/test.txt

HDP Hadoop (peudo-distributed mode)

Step 3: Try a few basic Linux commands

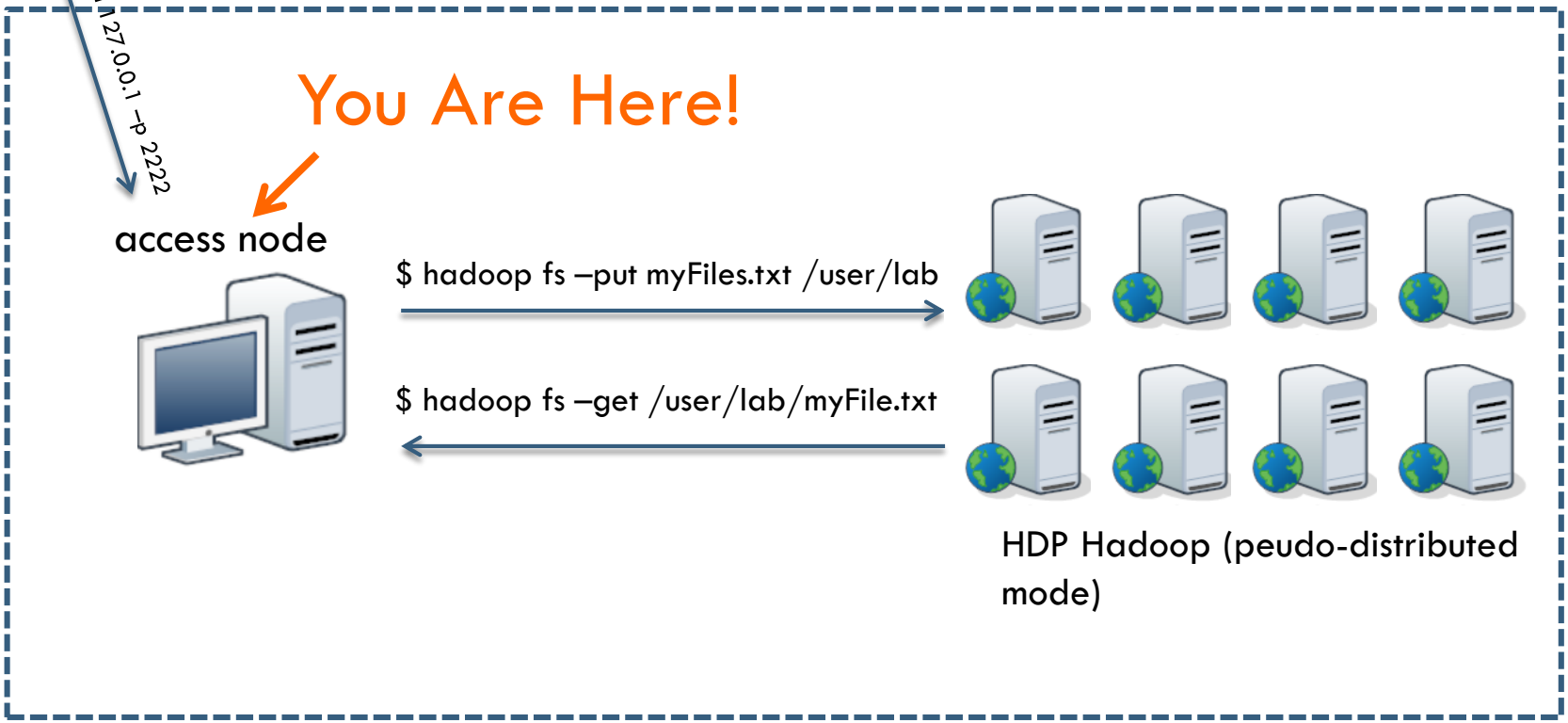
- "-cat" concatenates the contents of the file to the screen
- "head" to print the first few lines of an hdfs file

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -cat test.txt | head -10
Year
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
Rank 1
Rank 2
Rank 3
Rank 4
[root@sandbox lab]#
```

Recall

Lab computer
Your PC/Mac

NOTE: Your client linux file system and hadoop filesystem (HDFS) are separate environment. First, you need to learn how to move files between the two file systems



Step 4: Download Files From HDFS to Local File System

- "-get" Copies/Downloads files from HDFS to the local file system

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -get /user/root/test.txt test_copy.txt
[root@sandbox lab]# ls
test_copy.txt  test.txt
[root@sandbox lab]# head -10 test_copy.txt
Year
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
Rank 1
Rank 2
Rank 3
Rank 4
[root@sandbox lab]#
```

Step 5: Find Out Space Utilization in a HDFS Directory

- -du Displays size of files and directories contained in the given directory or the size of a file if its just a file

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -du /user/root
318  /user/root/test.txt
[root@sandbox lab]#
```

Step 6: Advanced HDFS feature

- "-getmerge" Takes a source directory file or files as input and concatenates files in src into the local destination file

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -put test_copy.txt /user/root
[root@sandbox lab]# hadoop fs -ls /user/root
Found 3 items
drwx------   - root root          0 2016-09-05 17:54 /user/root/.staging
-rw-r--r--   3 root root        318 2016-09-05 17:43 /user/root/test.txt
-rw-r--r--   3 root root        318 2016-09-05 17:54 /user/root/test_copy.txt
[root@sandbox lab]# hadoop fs -getmerge /user/root test_two_copies.txt
[root@sandbox lab]# ls
test_copy.txt  test_two_copies.txt  test.txt
[root@sandbox lab]# wc -l test.txt
43 test.txt
[root@sandbox lab]# wc -l test_copy.txt
43 test_copy.txt
[root@sandbox lab]# wc -l test_two_copies.txt
86 test_two_copies.txt
[root@sandbox lab]#
```

Step 6: Advanced HDFS feature

- "-distcp"
- It is a tool used for large inter/intra-cluster copying
- It uses MapReduce to effect its distribution copy, error handling and recovery, and reporting

```
root@sandbox:~/lab
[root@sandbox lab]# hadoop fs -mkdir /user/test
[root@sandbox lab]# hadoop fs -chown root:root /user/test

[root@sandbox lab]# hadoop distcp /user/root/ /user/test
16/09/05 17:52:26 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=f
alse, syncFolder=false, deleteMissing=false, ignoreFailures=false, maxMaps=20, s
slConfigurationFile='null', copyStrategy='uniformsize', sourceFileListing=null,
sourcePaths=[/user/root], targetPath=/user/test, targetPathExists=true, preserve
RawXattrs=false}
16/09/05 17:52:28 INFO impl.TimelineClientImpl: Timeline service address: http:/
/sandbox.hortonworks.com:8188/ws/v1/timeline/
16/09/05 17:52:28 INFO client.RMProxy: Connecting to ResourceManager at sandbox.
hortonworks.com/10.0.2.15:8050
16/09/05 17:52:30 INFO impl.TimelineClientImpl: Timeline service address: http:/
/sandbox.hortonworks.com:8188/ws/v1/timeline/
16/09/05 17:52:30 INFO client.RMProxy: Connecting to ResourceManager at sandbox.
hortonworks.com/10.0.2.15:8050
16/09/05 17:52:34 INFO mapreduce.JobSubmitter: number of splits:2
16/09/05 17:52:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
73085803899_0001
```

Step 6: Advanced HDFS feature

- "hadoop dfsadmin -report"
- To understand storage availability in the cluster
- Remember: user "hdfs" can view status of storage status of cluster

```
root@sandbox:~/lab
[root@sandbox lab]# sudo -u hdfs hadoop dfsadmin -report
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Configured Capacity: 44716605440 (41.65 GB)
Present Capacity: 31778260697 (29.60 GB)
DFS Remaining: 30638580736 (28.53 GB)
DFS Used: 1139679961 (1.06 GB)
DFS Used%: 3.59%
Under replicated blocks: 553
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (1):

Name: 10.0.2.15:50010 (sandbox.hortonworks.com)
Hostname: sandbox.hortonworks.com
Decommission Status : Normal
Configured Capacity: 44716605440 (41.65 GB)
```

Try yourself

- Print the last 2 lines of the Shakespere_100.txt file
- Create another HDFS directory called “labs_temp”
- Move the “Shakespere_100.txt” to labs_temp

Note: Look at the “Hadoop HDFS Shell Commands” link on the D2L