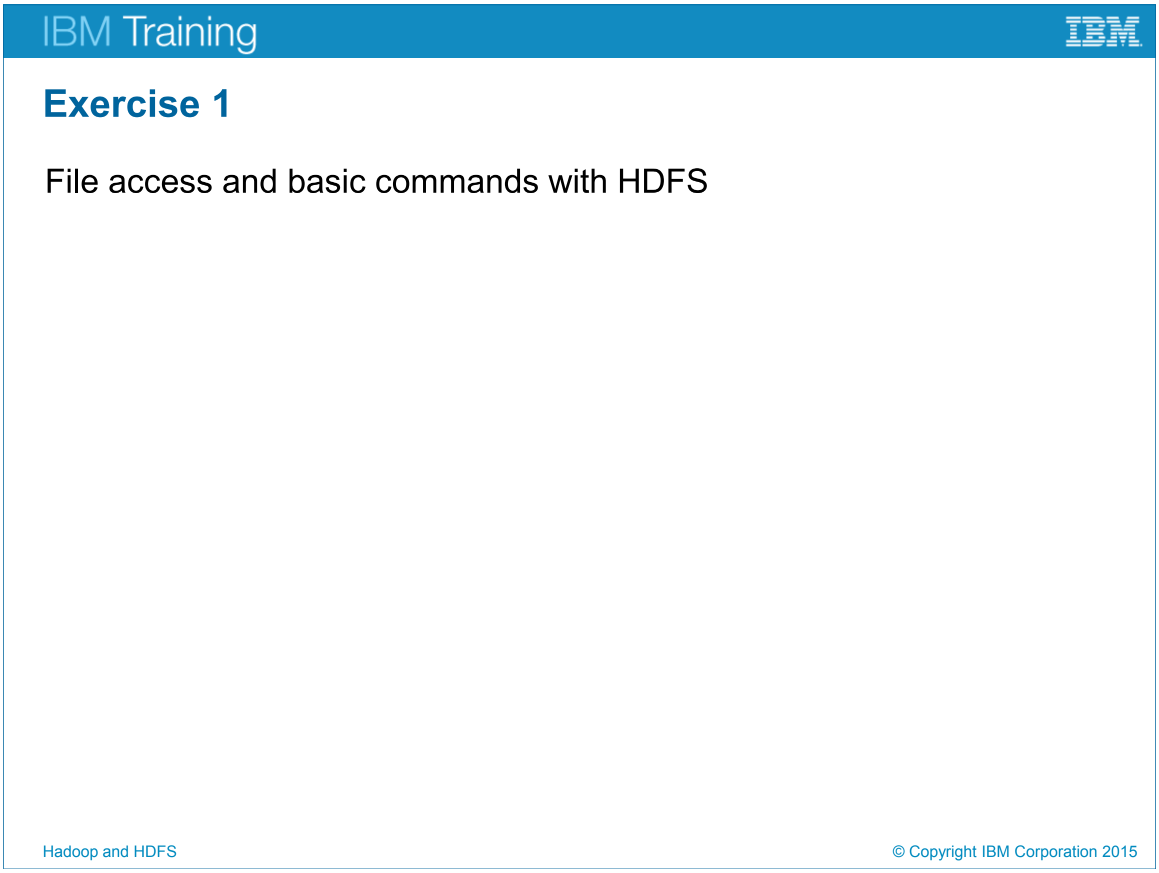

IBM Training

Exercise 1

File access and basic commands with HDFS

Hadoop and HDFS

© Copyright IBM Corporation 2015

Exercise 1: File access and basic commands with HDFS

Exercise 1:
File access and basic commands with HDFS

Purpose:
This exercise is intended to provide you with experience in using the Hadoop Distributed File System (HDFS). The basic HDFS file system commands learned here will be used throughout the remainder of the course.

You will also be moving some data into HDFS that will be used in later units of this course. The files that you will need are stored in the Linux directory /home/labfiles.

VM Hostname: **http://ibmclass.localdomain**
User/Password: **biadmin / biadmin**
root/dalvm3

Task 1. Learn some of the basic HDFS file system commands.

The major reference for the HDFS File System Commands can be found on the Apache Hadoop website. Additional commands can be found there. The URL for Hadoop 2.7.0 is:

http://hadoop.apache.org/docs/r2.7.0/hadoop-project-dist/hadoop-common/FileSystemShell.html

Use the documentation that corresponds to your current release of HDFS (substitute the appropriate value for r2.7.0, if different).

The HDFS File System Commands are generally prefixed by:

```
hadoop fs
```

but can also be executed with `dfs` instead of `fs` and also in the form:

```
hdfs fs
```

Other HDFS commands - generally administrative - use parameters other than `fs` (for example, fsck, rebalance, …).

1. Connect to your classroom VMware image and login as **biadmin**.
2. In a new terminal window, type `cd` to change to your home directory.
3. Type `pwd` to verify that you are in biadmin's home directory.

4. To verify that, as the user biadmin, you have a file system directory in HDFS, type **`hadoop fs -ls .`**.

   Note: At the end of this command there is a period (it means the current directory in HDFS).

   If you completed the first section of this course, running this command will give four results. Otherwise your results appear as shown below:

```
[biadmin@ibmclass ~]$ hadoop fs -ls .
[biadmin@ibmclass ~]$
```

   In Linux, there is no response to the command; this indicates success. If you get this response, go to step 8.

   If, however, you receive an error response such as:

```
[biadmin@ibmclass ~]$ hadoop fs -ls .
ls: '.': No such file or directory
[biadmin@ibmclass ~]$
```

   then you do not have a home directory in HDFS. You will need to have one created for you. The ability to create a home directory for you in HDFS is not something that you can do; it must be done as the hdfs user. You will do this in steps 5 through 7.

5. To login as the hdfs user, use **`sudo su - hdfs`**

```
[biadmin@ibmclass ~]$ sudo su - hdfs
[sudo] password for biadmin:
[hdfs@ibmclass ~]$
```

   or **`su -`** (the root password is **`dalvm3`**)

```
[biadmin@ibmclass ~]$ su -
Password:
[root@ibmclass ~]# su - hdfs
[hdfs@ibmclass ~]$
```

6. Execute the following three commands:

```
hadoop fs -mkdir /user/biadmin
hadoop fs -chown biadmin /user/biadmin
hadoop fs -chgrp biadmin /user/biadmin
```

```
[hdfs@ibmclass ~]$ hadoop fs -mkdir /user/biadmin
[hdfs@ibmclass ~]$ hadoop fs -ls /
Found 8 items
drwxrwxrwx   - yarn   hadoop          0 2015-04-15 12:21 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2015-04-15 12:19 /apps
drwxrwxr-x   - hdfs   hadoop          0 2015-05-19 17:29 /biginsights
drwxr-xr-x   - hdfs   hdfs            0 2015-04-15 12:13 /iop
drwxr-xr-x   - mapred hdfs            0 2015-04-15 12:12 /mapred
drwxr-xr-x   - hdfs   hdfs            0 2015-04-15 12:12 /mr-history
drwxrwxrwx   - hdfs   hdfs            0 2015-05-20 18:17 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2015-06-04 09:02 /user
[hdfs@ibmclass ~]$ hadoop fs -ls /user
Found 8 items
drwxrwx---   - ambari-qa hdfs            0 2015-04-15 12:25 /user/ambari-qa
drwxr-xr-x   - hdfs      hdfs            0 2015-06-04 09:02 /user/biadmin
drwxrwxrwx   - bigr      hdfs            0 2015-05-20 18:17 /user/bigr
drwxr-xr-x   - hcat      hdfs            0 2015-04-15 12:23 /user/hcat
drwx------   - hive      hdfs            0 2015-04-15 12:19 /user/hive
drwxrwxr-x   - oozie     hdfs            0 2015-04-15 12:20 /user/oozie
drwxr-xr-x   - spark     hadoop          0 2015-04-15 12:14 /user/spark
drwxrwxr-x   - tauser    hdfs            0 2015-05-20 12:18 /user/tauser
[hdfs@ibmclass ~]$ hadoop fs -chown biadmin /user/biadmin
[hdfs@ibmclass ~]$ hadoop fs -chgrp biadmin /user/biadmin
[hdfs@ibmclass ~]$ hadoop fs -ls /user
Found 8 items
drwxrwx---   - ambari-qa hdfs             0 2015-04-15 12:25 /user/ambari-qa
drwxr-xr-x   - biadmin   biadmin          0 2015-06-04 09:02 /user/biadmin
drwxrwxrwx   - bigr      hdfs             0 2015-05-20 18:17 /user/bigr
drwxr-xr-x   - hcat      hdfs             0 2015-04-15 12:23 /user/hcat
drwx------   - hive      hdfs             0 2015-04-15 12:19 /user/hive
drwxrwxr-x   - oozie     hdfs             0 2015-04-15 12:20 /user/oozie
drwxr-xr-x   - spark     hadoop           0 2015-04-15 12:14 /user/spark
drwxrwxr-x   - tauser    hdfs             0 2015-05-20 12:18 /user/tauser
```

7. Logout from hdfs (and root, if necessary) by using `exit` or **Ctrl-D**.

   Alternatively, you can close your current terminal window and open a new one.

   You have now have a home directory in HDFS. This home directory is /user/biadmin (note that this is "user" and not "usr" and not "home").

8. To validate that you now have a home directory in HDFS, run the following commands:

```
        hadoop fs -ls
        hadoop fs -ls .
        hadoop fs -ls /user
```

```
[biadmin@ibmclass ~]$ hadoop fs -ls
[biadmin@ibmclass ~]$

 [biadmin@ibmclass ~]$ hadoop fs -ls .
[biadmin@ibmclass ~]$

[hdfs@ibmclass ~]$ hadoop fs -ls /user
Found 8 items
drwxrwx---   - ambari-qa hdfs            0 2015-04-15 12:25 /user/ambari-qa
drwxr-xr-x   - biadmin   biadmin         0 2015-06-04 09:02 /user/biadmin
drwxrwxrwx   - bigr      hdfs            0 2015-05-20 18:17 /user/bigr
drwxr-xr-x   - hcat      hdfs            0 2015-04-15 12:23 /user/hcat
drwx------   - hive      hdfs            0 2015-04-15 12:19 /user/hive
drwxrwxr-x   - oozie     hdfs            0 2015-04-15 12:20 /user/oozie
drwxr-xr-x   - spark     hadoop          0 2015-04-15 12:14 /user/spark
drwxrwxr-x   - tauser    hdfs            0 2015-05-20 12:18 /user/tauser
```

If you completed the first section of this course, your results may differ slightly.

You are now ready to do some exploration of the HDFS file system and your Hadoop ecosystem.

9. To create a subdirectory called **Gutenberg** in your HDFS user directory, type **`hadoop fs -mkdir Gutenberg`**.

   Note that the directory you just created, defaulted to your home directory. This is how Linux handles this command. If you wanted the directory elsewhere, then you need to fully qualify it.

10. Type **`hadoop fs -ls`** to list your directory.

11. Use the recursive option (**`-R`**) of **`ls`** to see if there are any files in your Gutenberg directory (there are none at the moment): **`hadoop fs -ls -R`**

```
[biadmin@ibmclass ~]$ hadoop fs -ls
[biadmin@ibmclass ~]$ hadoop fs -mkdir Gutenberg
[biadmin@ibmclass ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - biadmin biadmin          0 2015-06-04 10:41 Gutenberg
[biadmin@ibmclass ~]$ hadoop fs -ls -R
drwxr-xr-x   - biadmin biadmin          0 2015-06-04 10:41 Gutenberg
[biadmin@ibmclass ~]$
```

12. Use the **`hadoop fs -put`** command to move files from the Linux /home/labfiles directory into your HDFS directory, Gutenberg:

```
    hadoop fs -put /home/biadmin/labfiles/*.txt Gutenberg
```

   This command can also be executed as:

```
    hadoop fs -copyFromLocal /home/biadmin/labfiles/*.txt Gutenberg
```

13. Repeat the list command with the recursive option (step 10).

```
[biadmin@ibmclass ~]$ hadoop fs -put /home/labfiles/*.txt Gutenberg
[biadmin@ibmclass ~]$ hadoop fs -ls .
Found 1 items
drwxr-xr-x   - biadmin biadmin          0 2015-06-04 11:01 Gutenberg
[biadmin@ibmclass ~]$ hadoop fs -ls -R
drwxr-xr-x   - biadmin biadmin          0 2015-06-04 11:01 Gutenberg
-rw-r--r--   3 biadmin biadmin     421504 2015-06-04 11:01 Gutenberg/Frankenstein.txt
-rw-r--r--   3 biadmin biadmin     697802 2015-06-04 11:01
Gutenberg/Pride_and_Prejudice.txt
-rw-r--r--   3 biadmin biadmin     757223 2015-06-04 11:01
Gutenberg/Tale_of_Two_Cities.txt
-rw-r--r--   3 biadmin biadmin     281398 2015-06-04 11:01 Gutenberg/The_Prince.txt
[biadmin@ibmclass ~]$
```

Note here in the listing of files the following for the last file:

```
-rw-r--r--   3 biadmin biadmin     281398
```

Here you will see the read-write (*rw*) permissions that you would find with a typical Linux file.

The "3" here is the typical replication factor for the blocks (or "splits") of the individual files. These files are too small (last file is 281KB) to have more than one block (the max block size is 128MB), but *each block* of *each file* is replicated three times.

You may see "1" instead of "3" in a single-node cluster (pseudo-distributed mode). That too is normal for a single-node cluster as it does not really make sense to replicate multiple copies on a single node.

Task 2. Explore one of the HDFS administrative commands.

There are a number of HDFS administration commands in addition to the HDFS file system commands. A reference for administration commands is: https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/CommandsManual.html.

We will look at just one of them, **`fsck`**. We will run it as: **`hdfs fsck`**

Administrative commands cannot be normally run as regular users. You will do the following as the **`hdfs`** user.

`fsck`

Runs a HDFS filesystem checking utility.

Usage: `hadoop fsck [GENERIC_OPTIONS] <path> [-move | -delete | -openforwrite] [-files [-blocks [-locations | -racks]]]`

COMMAND_OPTION	Description
path	Start checking from this path.
-move	Move corrupted files to /lost+found
-delete	Delete corrupted files.
-openforwrite	Print out files opened for write.
-files	Print out files being checked.
-blocks	Print out block report.
-locations	Print out locations for every block.
-racks	Print out network topology for data-node locations.

1. To log in as the **hdfs** user, type **`sudo su - hdfs`**.

```
[biadmin@ibmclass ~]$ sudo su - hdfs
[sudo] password for biadmin:
[hdfs@ibmclass ~]$
```

or **`su -`** (the root password is **`dalvm3`**)

```
[biadmin@ibmclass ~]$ su -
Password:
[root@ibmclass ~]# su - hdfs
[hdfs@ibmclass ~]$
```

2. Type the following command: **`hdfs fsck /`**

   In your lab environment, you may get a number of errors related to replication. Ignore those, as you are running a pseudo-distributed cluster and this replication is not standard.

   The results that you should see will be similar to the following:

```
. . .
Status: HEALTHY
 Total size:    713494586 B (Total open files size: 340 B)
 Total dirs:    12748
 Total files:   341
 Total symlinks:                0 (Files currently being written: 3)
 Total blocks (validated):      330 (avg. block size 2162104 B) (Total open file blocks
(not validated): 3)
 Minimally replicated blocks:   330 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       330 (100.0 %)
 Mis-replicated blocks:                 0 (0.0 %)
 Default replication factor:    3
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              660 (66.666664 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Thu Jun 04 11:11:43 GMT-05:00 2015 in 218 milliseconds


The filesystem under path '/' is HEALTHY
```

3. Close all open windows.

Results:
You used basic Hadoop Distributed File System (HDFS) file system commands, moving some data into HDFS that will be used in later units of this course.