Access HDFS with Command Line

Set Up Your Environment

1. Before starting the assignment, be sure you have run the course setup script in a terminal window. (You only need to run this script once; if you ran it earlier, you do not need to run it again.)

```
$ $DEV1/scripts/training_setup_dev1.sh
```

Explore the HDFS Command Line Interface

HDFS is already installed, configured, and running on your virtual machine.

The simplest way to interact with HDFS is by using the `hdfs` command. To execute file system commands within HDFS, use the `hdfs dfs` command.

1. Open a terminal window (if one is not already open) by double-clicking the Terminal icon on the desktop.

2. Enter:

```
$ hdfs dfs -ls /
```

This shows you the contents of the root directory in HDFS. There will be multiple entries, one of which is `/user`. Individual users have a "home" directory under this directory, named after their username; your username in this course is `training`, therefore your home directory is `/user/training`.

3. Try viewing the contents of the `/user` directory by running:

```
$ hdfs dfs -ls /user
```

You will see your home directory in the directory listing.

4. List the contents of your home directory by running:

```
$ hdfs dfs -ls /user/training
```

There are no files yet, so the command silently exits. This is different than if you ran `hdfs dfs -ls /foo`, which refers to a directory that doesn't exist and which would display an error message.

Note that the directory structure in HDFS has nothing to do with the directory structure of the local filesystem; they are completely separate namespaces.

Upload Files to HDFS

Besides browsing the existing filesystem, another important thing you can do with the HDFS command line interface is to upload new data into HDFS.

5. Start by creating a new top level directory for homework assignments. You will use this directory throughout the rest of the course.

```
$ hdfs dfs -mkdir /loudacre
```

6. Change directories to the local filesystem directory containing the sample data we will be using in the course.

```
$ cd $DEV1DATA
```

If you perform a regular Linux `ls` command in this directory, you will see several files and directories used in this class. One of the data directories is `kb`. This directory holds Knowledge Base articles that are part of Loudacre's customer service website.

7. Insert this directory into HDFS:

```
$ hdfs dfs -put kb /loudacre/
```

This copies the local `kb` directory and its contents into a remote HDFS directory named `/loudacre/kb`.

8. List the contents of the new HDFS directory now:

```
$ hdfs dfs -ls /loudacre/kb
```

You should see the KB articles that were in the local directory.

9. Practice uploading a directory, then remove it, as it is not actually needed for the exercises.

```
$ hdfs dfs -put $DEV1DATA/calllogs /loudacre/
$ hdfs dfs -rm -r /loudacre/calllogs
```

View HDFS files

Now view some of the data you just copied into HDFS.

10. Enter:

```
$ hdfs dfs -cat /loudacre/kb/KBDOC-00289.html | tail \
-n 20
```

This prints the last 20 lines of the article to your terminal. This command is handy for viewing HDFS data. An individual file is often very large, making it inconvenient to view the entire file in the terminal. For this reason, it's often a good idea to pipe the output of the `fs -cat` command into `head`, `tail`, `more`, or `less`.

11. You can use the `hdfs dfs -get` command to retrieve a local copy of a file or directory from HDFS. This command takes two arguments: an HDFS path and a local path. It copies the HDFS contents into the local filesystem:

```
$ hdfs dfs -get \
/loudacre/kb/KBDOC-00289.html ~/article.html
$ less ~/article.html
```

12. There are several other operations available with the `hdfs dfs` command to perform most common filesystem manipulations: `mv`, `cp`, `mkdir`, etc.

    In the terminal window, enter:

```
$ hdfs dfs
```

    You see a help message describing all the file system commands provided by HDFS.

    Try playing around with a few of these commands if you like.