**Clairvoyant**

**HDFS Exercise**

# **Hands-On Exercise: Using HDFS**

In this exercise you will begin to get acquainted with the Hadoop tools. You will manipulate files in **HDFS**, the Hadoop Distributed File System.

**Hadoop**

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

Most of your interaction with the system will be through a command-line wrapper called *hadoop.* To try this, run the following command:

The *hadoop* command is subdivided into several subsystems. For example, there is a subsystem for working with files in **HDFS** and another for launching and managing MapReduce processing jobs.

**Step 1: Exploring HDFS**

The subsystem associated with **HDFS** in the Hadoop wrapper program is called FsShell. This subsystem can be invoked with the command *hadoop fs*.

1. Open a terminal window by double clicking on the **Terminal** icon.
2. In the terminal window, enter:

$ hadoop fs

You should see a help message describing the commands associated with this subsystem.

1. Enter:

$ hadoop fs –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. The home directory for this workshop is going to be /user/cloudera. This is the directory under which you will be copying the files and running the MapReduce jobs.

1. Try running:

$ hadoop fs –ls /user/cloudera

The contents of path used is displayed. If you try to access a path that does not exist, it will display an error message.

**Note**: The directory structure in HDFS has nothing to do with the directory structure of the local file system.

**Step2: Uploading Files**

Using FsShell we can upload new data into **HDFS**. Following the steps listed below we will upload a few files that we will be using in our first MapReduce program.

1. Download and extract the *sampledata.zip* from the workshop site.
2. Unzip the sampledata.zip file

cd {path\_to\_dir\_with\_zip}

unzip sampledata.zip

1. Change directories to the directory containing the sample data for invertedindex problem.

cd sampledata

If you perform an ls command in this directory, you will see a few files numbered 1,2,3…

1. Lets create a new directory on **HDFS**.

$ hadoop fs -mkdir /user/cloudera/workshop/

$ hadoop fs -mkdir workshop/mapreduce/

$ hadoop fs -mkdir workshop/mapreduce/invertedIndex/

1. **HDFS** from the {path\_to\_dir\_with\_zip}/sampledata/invertedIndex/.

|  |
| --- |
| $ hadoop fs -put input/ /user/cloudera/workshop/mapreduce/invertedIndex/ |

6. Lets verify that the input files are copied to **HDFS** .

|  |
| --- |
| $ hadoop fs -ls /user/cloudera/workshop/mapreduce/invertedIndex/input/ |

This should list a number of files numbered 1,2,3…

|  |
| --- |
| **Relative paths**  If you pass any relative (non-absolute) paths to FsShell commands (or use relative paths in MapReduce programs), they are **always** considered relative to your home directory. |

**Step 3: Viewing and Manipulating Files**

1. Lets view a few files on **HDFS**.

|  |
| --- |
| $ hadoop fs –ls /user/cloudera/workshop/mapreduce/invertedIndex/input |

2. Lets print the last 50 lines of the file on to the console.

|  |
| --- |
| $ hadoop fs -cat /user/cloudera/workshop/mapreduce/invertedIndex/input/10 | tail -n 50 |

At times, this command can be useful for viewing the output of MapReduce programs. It is often a good idea to pipe the output of fs –cat command into head, tail more or less.  
  
**Note**: When output of fs –cat command is piped to a local UNIX command, the full contents of the file are still extracted from **HDFS** and sent to your local machine. Once on your local machine, the file contents are then modified before being displayed.

3. If we want to download a file from **HDFS**, we can use the following command:

|  |
| --- |
| $ hadoop fs –get /user/cloudera/workshop/mapreduce/invertedIndex/input/20 fromhdfs20 |

4. If we want to download the first 100 lines of a file from **HDFS**, we can use the following command:

|  |
| --- |
| $ hadoop fs –get /user/cloudera/workshop/mapreduce/invertedIndex/input/20 | head –n 100 20 |

5. We can delete the files and folders using the command

|  |
| --- |
| $ hadoop fs –rm -r /user/cloudera/workshop/mapreduce/invertedIndex/input |

6. The various other commands within *FsShell* subsystem get displayed by using the below command.

|  |
| --- |
| $ hadoop fs |

**Other Commands**

There are a number of other commands associated with the FsShell subsystem used to perform most common filesystem manipulations: mv,cp, rm

1. Try to delete a directory or file from **HDFS**.