BA-BEAD Big Data Engineering for Analytics

Workshop Instructions

Access HDFS with Command Line and Hue

In Academic Partnership with:

cloudera®

ACADEMIC PARTNER

Contents

Homework

In this homework assignment you will practice working with HDFS, the Hadoop Distributed File System. The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and written entirely in Java. You will use the HDFS command line tool and the Hue File Browser web-based interface to manipulate files in HDFS.

Prerequisite

Basic familiarity with the Linux command line is assumed. Basic knowledge of SQL is helpful; prior knowledge of Hadoop is not required.

Running the VMware Image

Start the virtual Image and you should be able to work with the CentOS virtual machine.

Set Up Your Environment

Before starting the assignment, be sure you have run the cloudera launch script from the desktop window. (You only need to run this script once; if you ran it earlier, you do not need to run it again.)

Working with the Virtual Machine

1. The VM is set to automatically log in as the user `cloudera`. Should you log out at any time, you can log back in as the user `cloudera` with the password training.
2. Should you need it, the root password is `cloudera`. You may be prompted for this if, for example, you want to change the keyboard layout. In general, you should not need this password since the `cloudera` user has unlimited sudo privileges.

You can retrieve the data needed from

Explore the HDFS Command Line Interface

3. Cluster-wide block size is controlled by the `dfs.blocksize` configuration property in the `hdfs-site.xml` file. The `dfs.blocksize` configuration property applies for files that are created without a block size specification.
4. 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.
5. Open a terminal window (if one is not already open) by double-clicking the Terminal icon on the desktop. Enter:

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

6. 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 `cloudera`, therefore your home directory is `/user/cloudera`.
7. Try viewing the contents of the `/user` directory by running (You will see your home directory in the directory listing):

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

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

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

   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 file system; 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.

9. 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
   ```

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

    ```
    $ cd <data directory>
    ```

    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.

11. 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`.

12. 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.

13. Practice uploading a directory, then remove it.

Relative paths

In HDFS, any relative (non-absolute) paths are considered relative to your home directory. There is no concept of a "current" or "working" directory as there is in Linux and similar file systems.

View HDFS files

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

15. 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 file system:

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

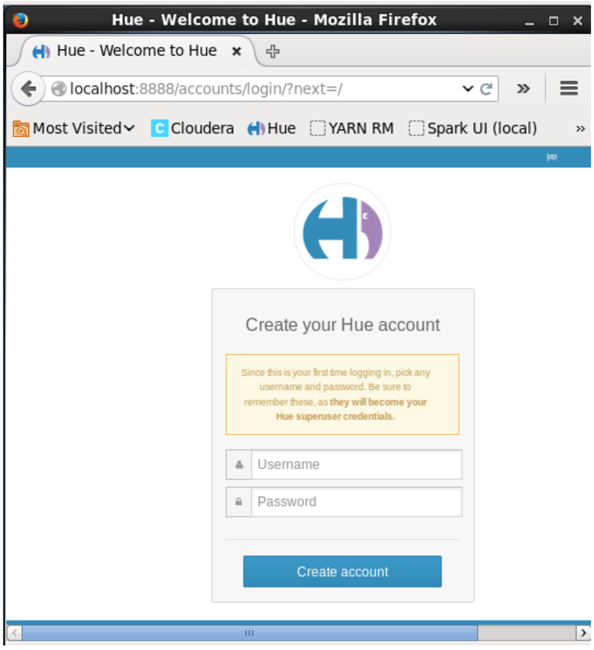
16. There are several other operations available with the `hdfs dfs` command to perform most common file system manipulations: `mv, cp, mkdir`, etc. In the terminal window, enter:

```
$ 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.

Use the Hue File Browser to browse, view and manage files

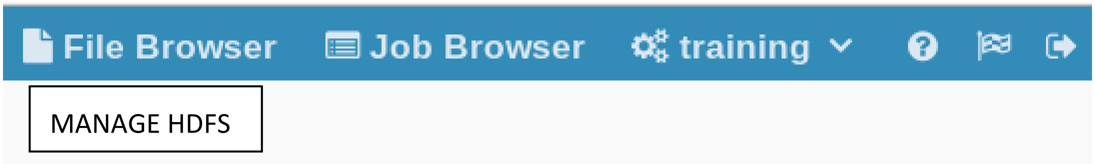
17. Start Firefox on the VM by clicking its icon on the main menu panel at the top of the screen.
18. Click the Hue bookmark, or visit http://localhost:8888



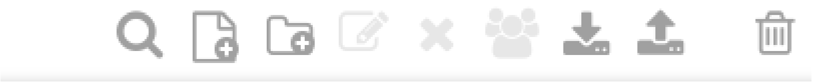
Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials.**

19. Because this is the first time anyone has logged into Hue on this server, you will be prompted to create a new user account. Enter username `cloudera` and password `cloudera`, then click Create Account. (If prompted you may click "Remember Password")

    Note: When you first log in to Hue you may see a misconfiguration warning. This is because not all the services Hue depends on are running on the course VM. You can disregard the message.
20. Hue has many useful features, many of which will be covered later in the course. For now, to access HDFS, click File Browser in the Hue menu bar. (The mouse over text is "Manage HDFS").

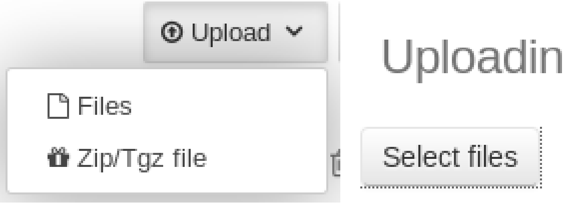


Note: If your Firefox window is too small to display the full menu names, you will see just the icons instead.



21. By default, the contents of your HDFS home directory (`/user/cloudera`) display. In the directory path name, click the leading slash (/) to view the HDFS root directory.
22. The contents of the root directory display, including the `loudacre` directory you created earlier. Click on that directory to see the contents.
23. Click on the name of the **kb** directory to see the knowledge base articles you uploaded.
24. View one of the files by clicking on the name of any one of the articles.

25. In the file viewer, the contents of the file are displayed on the right. In this case, the file is fairly small, but typical files in HDFS are very large, so rather than displaying the entire contents on one screen, Hue provides buttons to move between pages.
26. Return to the directory review by clicking View file location in the Action panel on the left. Click the up arrow (⇧) to return to the `/loudacre` base directory.
27. To upload a file, click the Upload button. You can choose to upload a plain file, or to upload a zipped file (which will be automatically unzipped after upload). In this case, select Files, then click Select Files.



28. A Linux file browser appears.
29. Browse to `/home/cloudera/data`
30. Choose `base_stations.tsv` and click the Open button.
31. When the file has uploaded, it will be displayed in the directory. Click the checkbox next to the file’s icon, then click the Actions button to see a list of actions that can be performed on the selected file(s).
    Optional: Explore the various file actions available. When you’ve finished, select any unneeded files you have uploaded and click the Move to trash button to delete.

Switching off the Virtual Image.

When you have finished working with the image you can properly close the machine by choosing the Shut Down option under the System Menu of the Cent OS Linux Virtual Image.

---------END OF DOCUMENT----------