Manage Files on HDFS via Cli/Ambari Files View

Manage Files on HDFS with the Command Line

Introduction

In this tutorial, we will walk through many of the common of the basic Hadoop Distributed File System (HDFS) commands you will need to manage files on HDFS. The particular datasets we will utilize to learn HDFS file management is truck drivers statistics.

Prerequisites

- Downloaded and deployed the Hortonworks Data Platform (HDP) Sandbox
- Learning the Ropes of the HDP Sandbox

Outline

Download the Drivers Related Datasets

We will download **geolocation.csv** and **trucks.csv** data onto our local filesystems of the sandbox. The commands are tailored for mac and linux users.

Then, we will download **geolocation.csv** and **trucks.csv** data onto our local filesystems of the sandbox. The commands are tailored for mac and linux users.

1.Open a terminal on your local machine, SSH into the sandbox:

```
ssh root@sandbox-hdp.hortonworks.com -p 2222
```

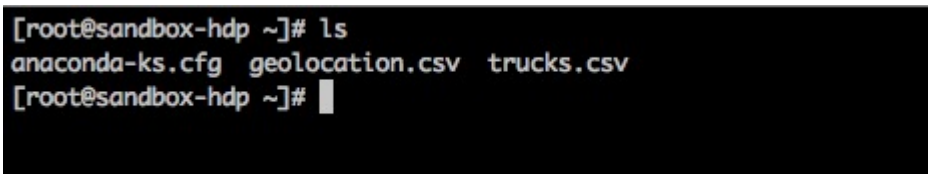
> Note: If you're on VMware or Docker, ensure that you map the sandbox IP to the correct hostname in the hosts file. Map your Sandbox IP

2.Copy and paste the commands to download the **geolocation.csv** and **trucks.csv** files. We will use them while we learn file management operations.

```
#Download geolocation.csv
wget https://github.com/hortonworks/data-
```

```
tutorials/raw/master/tutorials/hdp/manage-files-on-hdfs-via-cli-ambari-files-view/assets/drivers-datasets/geolocation.csv

#Download trucks.csv
wget https://github.com/hortonworks/data-tutorials/raw/master/tutorials/hdp/manage-files-on-hdfs-via-cli-ambari-files-view/assets/drivers-datasets/trucks.csv
```



Create a Directory in HDFS, Upload a File and List Contents

1. Let's learn by writing the syntax. You will be able to copy and paste the following example commands into your terminal. Login under **hdfs** user, so we can give root user permission to perform file operations:

```
#Login under hdfs user
su hdfs
cd
```

2. We will use the following command to run filesystem commands on the file system of Hadoop:

```
hdfs dfs [command_operation]
```

> Refer to the File System Shell Guide to view various command_operations.

hdfs dfs -chmod:

The command **chmod** affects the permissions of the folder or file. It controls who has read/write/execute privileges.

1. We will give root access to read and write to the user directory. Later we will perform an operation in which we send a file from our local filesystem to hdfs.

```
hdfs dfs -chmod 777 /user
```

> Warning in production environments, setting the folder with the permissions above is not a good idea because anyone can read/write/execute files or folders.

2. Type the following command, so we can switch back to the root user. We can perform the remaining file operations under the **user** folder since the permissions were changed.

```
exit
```

hdfs dfs -mkdir:

The command **mkdir** takes the path URI's as an argument and creates a directory or multiple directories. The full syntax of how to create a directory is below:

```
#Syntax to create directory in HDFS
hdfs dfs -mkdir <paths>
```

1. Let's create the directory for the driver dataset by entering the following commands into your terminal:

```
#Creates a directory called hadoop under users
hdfs dfs -mkdir /user/hadoop

#Creates two directories geolocation.csv and trucks.csv under the
directory hadoop
hdfs dfs -mkdir /user/hadoop/geolocation /user/hadoop/trucks
```

hdfs dfs -put:

The command **put** copies single src file or multiple src files from local file system to the Hadoop Distributed File System.

```
#Syntax to copy file(s) from local to HDFS
hdfs dfs -put <local-src> ... <HDFS_dest_path>
```

1. Now let's copy both source files from your local file system to the Hadoop Distributed File System by entering the following commands into your terminal:

```
#Copy the geolocation.csv file to HDFS
hdfs dfs -put geolocation.csv /user/hadoop/geolocation

#Copy the trucks.csv file to HDFS
hdfs dfs -put trucks.csv /user/hadoop/trucks
```

hdfs dfs -ls:

The command **ls** lists the contents of a directory. For a file, it returns stats of a file. The full syntax is below:

```

```

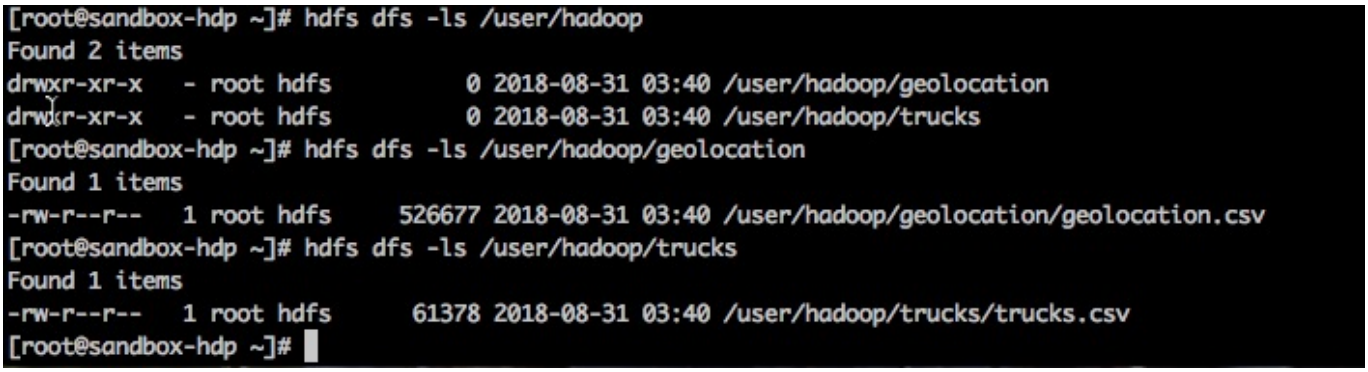
```
#Syntax for listing content on HDFS
hdfs dfs  -ls  <args>
```

1. Let's continue with our example, enter the commands below to list the content of the directories we just created:

```
#List the content of the hadoop directory
hdfs dfs -ls /user/hadoop

#List the content of the geolocation directory
hdfs dfs -ls /user/hadoop/geolocation

##List the content of the trucks directory
hdfs dfs -ls /user/hadoop/trucks
```



Find Out Space Utilization in a HDFS Directory

hdfs dfs -du:

The command **du** displays the size of files and directories contained in the given directory or the size of a file if its just a file.

```
#Syntax for displaying the size of a file and directory in HDFS
hdfs dfs -du URI
```

1. Continuing with our example, enter the commands below in your terminal to show the size of contents of the hadoop directory and the geolocation.csv file:

```
#Displays the size of the directories in the hadoop directory including
the geolocation.csv file
hdfs dfs -du  /user/hadoop/ /user/hadoop/geolocation/geolocation.csv
```

```
[root@sandbox-hdp ~]# hdfs dfs -du  /user/hadoop/ /user/hadoop/geolocation/geolocation.csv
526677  /user/hadoop/geolocation
61378   /user/hadoop/trucks
526677  /user/hadoop/geolocation/geolocation.csv
[root@sandbox-hdp ~]#
```

Download Files From HDFS to Local File System

hdfs dfs -get:

The command **get** Copies/Downloads files from HDFS to the local file system:

```
//Syntax to copy/download files from HDFS your local file system
hdfs dfs –get <hdfs_src> <localdst>
```

1. Let's enter the command below to copy the geolocation.csv file into your home directory:

```
#Copying geolocation.csv into your local file system directory
hdfs dfs –get /user/hadoop/geolocation/geolocation.csv /home/
```

Explore Two Advanced Features

hdfs dfs -cp:

The command **cp** copies a file or directories recursively, all the directory's files and subdirectories to the bottom of the directory tree are copied. The **cp** command is a tool used for large inter/intra-cluster copying.

```
#Syntax for copying a file recursively
hdfs dfs –cp <src-path> <dest-path>
```

1. Going back to our example, enter the following command in your terminal to copy the geolocation file into the trucks directory:

```
#Copies the content of geolocation and trucks
hdfs dfs –cp /user/hadoop/geolocation/ /user/hadoop/trucks/
```

2. Verify the files or directories successfully copied to the destination folder:

```
#Verify that the geolocation file was copied
hdfs dfs –ls /user/hadoop/geolocation
hdfs dfs –ls /user/hadoop/trucks
```

```
[root@sandbox-hdp ~]# hdfs dfs -ls /user/hadoop/geolocation
Found 1 items
-rw-r--r--   1 root hdfs     526677 2018-08-31 03:40 /user/hadoop/geolocation/geolocation.csv
[root@sandbox-hdp ~]# hdfs dfs -ls /user/hadoop/trucks
Found 2 items
drwxr-xr-x   - root hdfs          0 2018-09-04 02:23 /user/hadoop/trucks/geolocation
-rw-r--r--   1 root hdfs      61378 2018-08-31 03:40 /user/hadoop/trucks/trucks.csv
[root@sandbox-hdp ~]#
```

> Visual result of cp file operation. Notice that both src1 and src2 directories and their contents were copied to the dest directory.

hdfs dfs -getmerge

The **getmerge** command takes a source directory file or files as input and concatenates files in src into the local destination file. This command concatenates files in the same directory or from multiple directories as long as we specify their location and outputs them to the local file system, as can be seen in the Syntax below:

```
# Syntax for concatenating two files
hdfs dfs [-nl] -getmerge <src> <localdst>
hdfs dfs -getmerge <src1> <src2> <localdst>

# Option:
#nl: can be set to enable adding a newline on end of each file
```

Use Help Command to access Hadoop Command Manual

The **help** command opens the list of commands supported by Hadoop Data File System (HDFS):

```
#Syntax for the help command
hdfs dfs  -help
```

```
[root@sandbox ~]# hdfs dfs  -help
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
```

Summary

Congratulations! We just learned to use commands to manage our **geolocation.csv** and **trucks.csv** dataset files in HDFS. We learned to create, upload and list the the contents in our directories. We also acquired the skills to download files from HDFS to our local file system and explored a few advanced features of HDFS file management using the command line.

Further Reading

- HDFS Overview
- Hadoop File System Documentation