

Manage Files on HDFS via Cli/Ambari Files View

Manage Files on HDFS with Ambari Files View

Introduction

In the previous tutorial, we learned to manage files on the Hadoop Distributed File System (HDFS) with the command line. Now we will use Ambari Files View to perform many of the file management operations on HDFS that we learned with CLI, but through the web-based interface.

Prerequisites

- Downloaded and deployed the [Hortonworks Data Platform \(HDP\) Sandbox](#)
- [Learning the Ropes of the HDP Sandbox](#)

Outline

- [Download the Drivers Related Datasets](#)
- [Create a Directory in HDFS, Upload a file and List Contents](#)
- [Find Out Space Utilization in a HDFS Directory](#)
- [Download Files From HDFS to Local Machine](#)
- [Explore Two Advanced Features](#)
- [Summary](#)
- [Further Reading](#)

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. Open another terminal, change your current directory to Downloads then copy and paste the commands to download the **geolocation.csv** and **trucks.csv** files. We will use them while we learn file management operations.

```
#Change your current directory to Downloads  
cd Downloads
```

```
#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 directory for drivers-datasets
mkdir drivers-datasets

#Move the geolocation and trucks csv files to the directory
mv geolocation.csv trucks.csv drivers-datasets/
```

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


Create Directory Tree in User

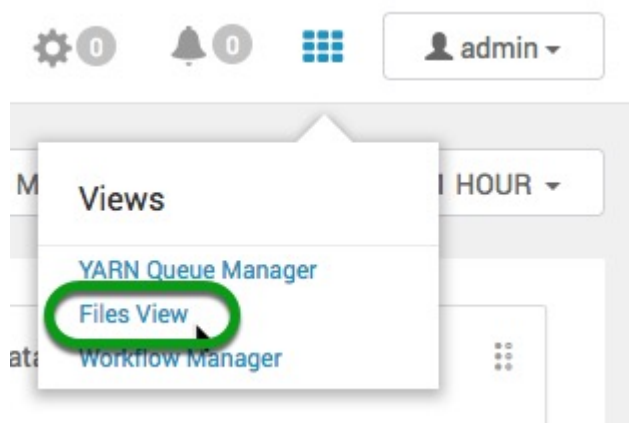
1. Login to Ambari Interface at sandbox-hdp.hortonworks.com:8080. Use the following login credentials in **Table 1**.

Table 1: Ambari Login credentials

Username	Password
admin	**setup process

[Setup Ambari Admin Password Manually](#)

2. Now that we have admin privileges, we can manage files on HDFS using Files View. Hover over the Ambari Selector Icon , enter the Files View web-interface.



The Files View Interface will appear with the following default folders.

Files View

Sandbox

admin

Total: 12 files or folders

+ Select All New Folder Upload

Search in current directory...

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
mapred	--	2018-09-06 23:50	mapred	hdfs	drwxr-xr-x		No
app-logs	--	2018-09-13 09:59	yarn	hadoop	drwxrwxrwx		No
ats	--	2018-09-06 23:50	yarn	hadoop	drwxr-xr-x		No
atsv2	--	2018-09-06 23:51	hdfs	hdfs	drwxr-xr-x		No
hdp	--	2018-09-06 23:51	hdfs	hdfs	drwxr-xr-x		No
livy2-recovery	--	2018-09-07 00:22	livy	hdfs	drwx-----		No
apps	--	2018-09-07 01:20	hdfs	hdfs	drwxr-xr-x		No
mr-history	--	2018-09-06 23:51	mapred	hadoop	drwxrwxrwx		No
spark2-history	--	2018-09-14 12:41	spark	hadoop	drwxrwxrwx		No
tmp	--	2018-09-13 10:16	hdfs	hdfs	drwxrwxrwx		No
user	--	2018-09-13 18:03	hdfs	hdfs	drwxrwxrwx		No
warehouse	--	2018-09-07 00:16	hdfs	hdfs	drwxr-xr-x		No

3. We will create 3 folders using the Files View web-interface. All *three folders*: **hadoop**, **geolocations** and **trucks** the last two which will reside in the **hadoop** folder, which resides in **user**.

Navigate into the **user** folder. Click the **new folder** button , an add new folder window appears and name the folder **hadoop**. Press **enter** or **+Add**

Add new folder

Name

hadoop

Cancel Add

4. Navigate into the **hadoop** folder. Create the *two folders*: **geolocation** and **trucks** following the process stated in the previous instruction.

Files View

Sandbox

admin

Folder Path

Total: 2 files or folders

+ Select All New Folder Upload

Search in current directory...

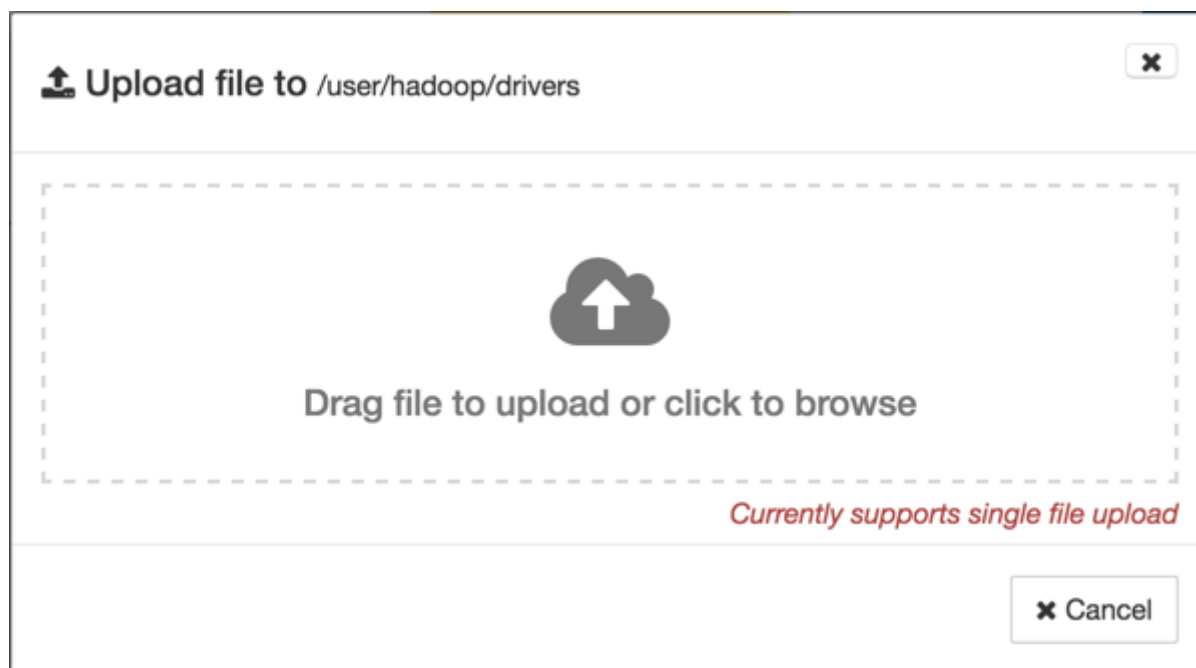
Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
Internal Folders							
geolocation	--	2018-09-14 13:06	admin	hdfs	drwxr-xr-x		No
trucks	--	2018-09-14 13:06	admin	hdfs	drwxr-xr-x		No

Upload Local Machine Files to HDFS

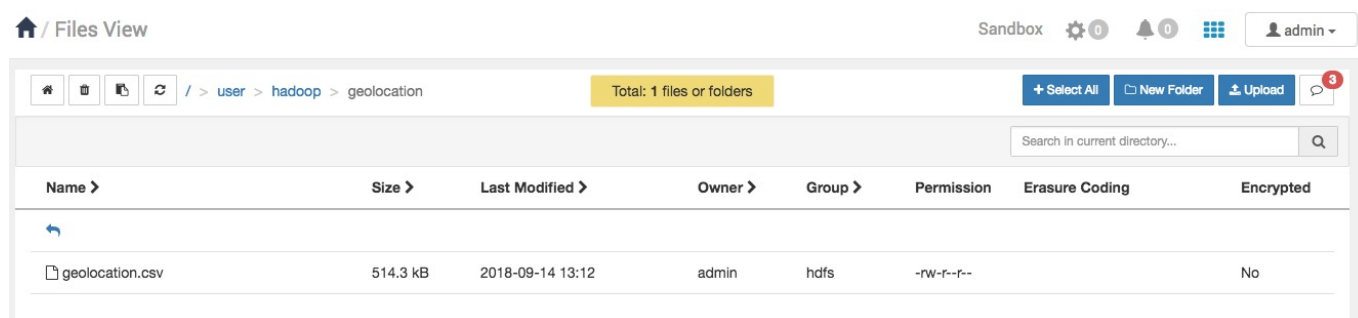
We will upload two files from our local machine: **geolocation.csv** and **trucks.csv** to appropriate HDFS directories.

1. Navigate through the path **/user/hadoop/geolocation** or if you're already in **hadoop**, enter the **geolocation** folder. Click the upload button  to transfer **geolocation.csv** into HDFS.

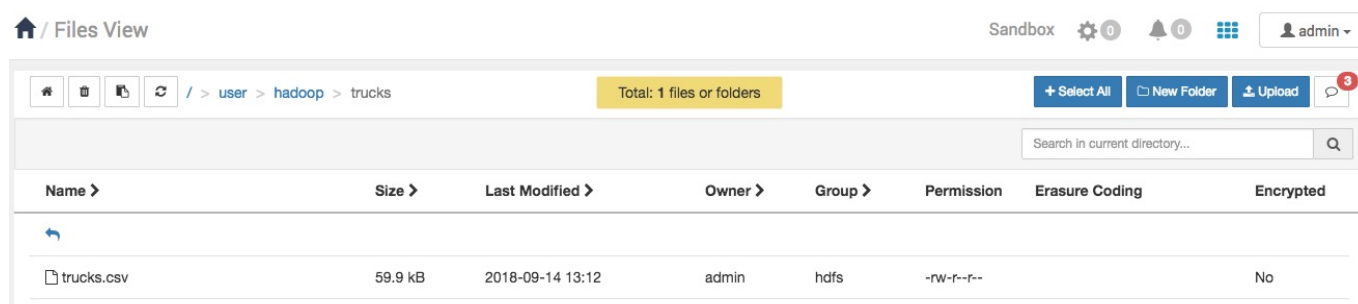
An Upload file window appears:



2. Click on the cloud with an arrow. A window with files from your local machine appears, find **geolocation.csv** in the **Downloads/drivers_datasets** folder, select it and then press **open** button.



3. In Files View, navigate to the **hadoop** folder and enter the **trucks** folder. Repeat the upload file process to upload **trucks.csv**.



View and Examine Directory Contents

Each time we open a directory, the Files View automatically lists the contents. Earlier we started in the **user** directory.

1. Let's navigate back to the **user** directory to examine the details given by the contents. Reference the image below while you read the Directory Contents Overview.

Directory Contents Overview of Columns

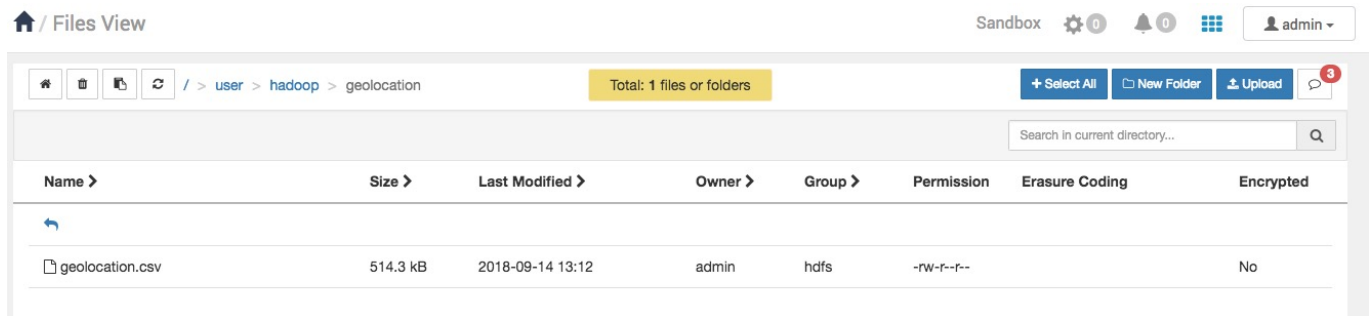
- **Name** are the files/folders
- **Size** contains bytes for the Contents
- **Last Modified** includes the date/time the content was created or Modified
- **Owner** is who owns that contents
- **Group** is who can make changes to the files/folders
- **Permissions** establishes who can read, write and execute data

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
mapred	--	2018-09-06 23:50	mapred	hdfs	drwxr-xr-x		No
app-logs	--	2018-09-13 09:59	yarn	hadoop	drwxrwxrwx		No
ats	--	2018-09-06 23:50	yarn	hadoop	drwxr-xr-x		No
atsv2	--	2018-09-06 23:51	hdfs	hdfs	drwxr-xr-x		No
hdp	--	2018-09-06 23:51	hdfs	hdfs	drwxr-xr-x		No
livy2-recovery	--	2018-09-07 00:22	livy	hdfs	drwx-----		No
apps	--	2018-09-07 01:20	hdfs	hdfs	drwxr-xr-x		No
mr-history	--	2018-09-06 23:51	mapred	hadoop	drwxrwxrwx		No
spark2-history	--	2018-09-14 12:41	spark	hadoop	drwxrwxrwx		No
tmp	--	2018-09-13 10:16	hdfs	hdfs	drwxrwxrwx		No
user	--	2018-09-13 18:03	hdfs	hdfs	drwxrwxrwx		No
warehouse	--	2018-09-07 00:16	hdfs	hdfs	drwxr-xr-x		No

Find Out Space Utilization in a HDFS Directory

In the command line when the directories and files are listed with the `hadoop fs -du /user/hadoop/`, the size of the directory and file is shown. In Files View, we must navigate to the file to see the size, we are not able to see the **size** of the directory even if it contains files.

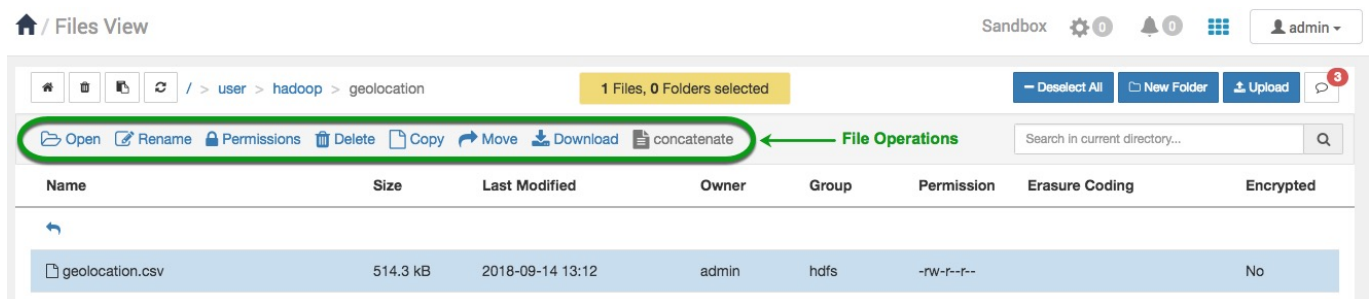
Let's view the size of **geolocation.csv** file. Navigate through `/user/hadoop/geolocation`. How much space has the file utilized? Files View shows **514.3 KB** for **geolocation.csv**.



Download File From HDFS to Local Machine

Files View enables users to download files and folders to their local machine with ease.

Let's download the **geolocation.csv** file to our computer. Click on the file's row, the row's color becomes blue, a group of file operations will appear, select the Download button. The default directory the file downloads to is our **Download** folder on our local machine.



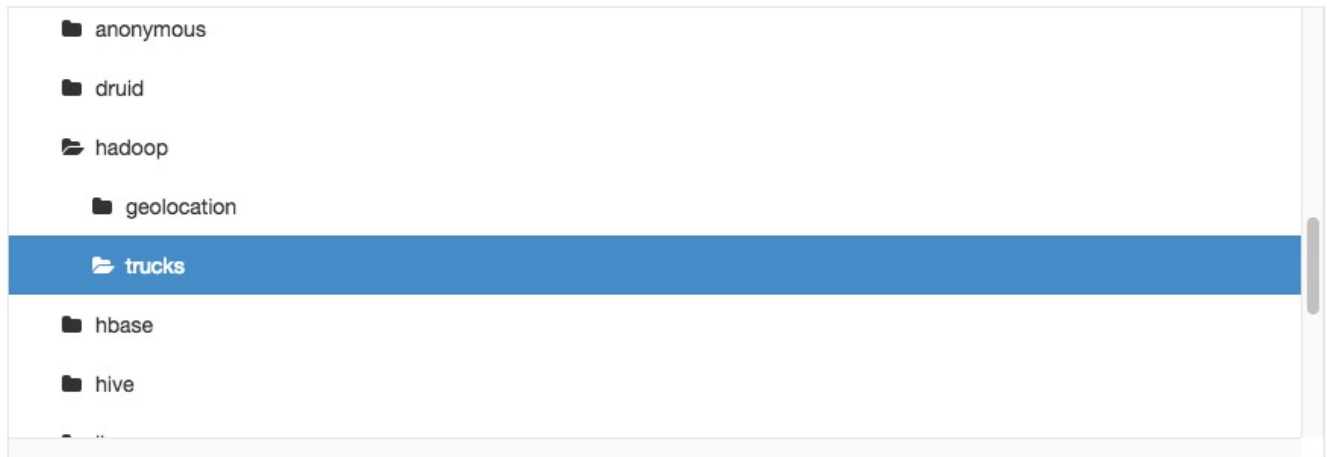
Explore Two Advanced Features

Concatenate Files

File Concatenation merges two files together. If we concatenate **trucks.csv** with **geolocation.csv**, the data from **geolocation.csv** will be appended to the end of **trucks.csv**. A typical use case for a user to use this feature is when they have similar large datasets that they want to merge together. The manual process to combine large datasets is inconvenient, so file concatenation was created to do the operation instantly.

1. Before we merge the csv files, we must place them in the same folder. Click on **geolocation.csv** row, it will highlight in blue, then press copy and in the copy window appears, select the **trucks** folder and press **Copy** to copy the csv file to it.

Copy to /user/hadoop/trucks



X Cancel

Copy

2. We will merge two large files together by selecting them both and performing concatenate operation. Navigate to the **trucks** folder. Select **geolocation.csv**, hold shift and click on **trucks.csv**. Click the concatenate button. The files will be downloaded into the **Download** folder on your local machine.

Home / Files View

Sandbox



admin

2 Files, 0 Folders selected

Open Rename Permissions Delete Copy Move Download concatenate Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
geolocation.csv	514.3 kB	2018-09-14 13:40	admin	hdfs	-rw-r--r--		No
trucks.csv	59.9 kB	2018-09-14 13:12	admin	hdfs	-rw-r--r--		No

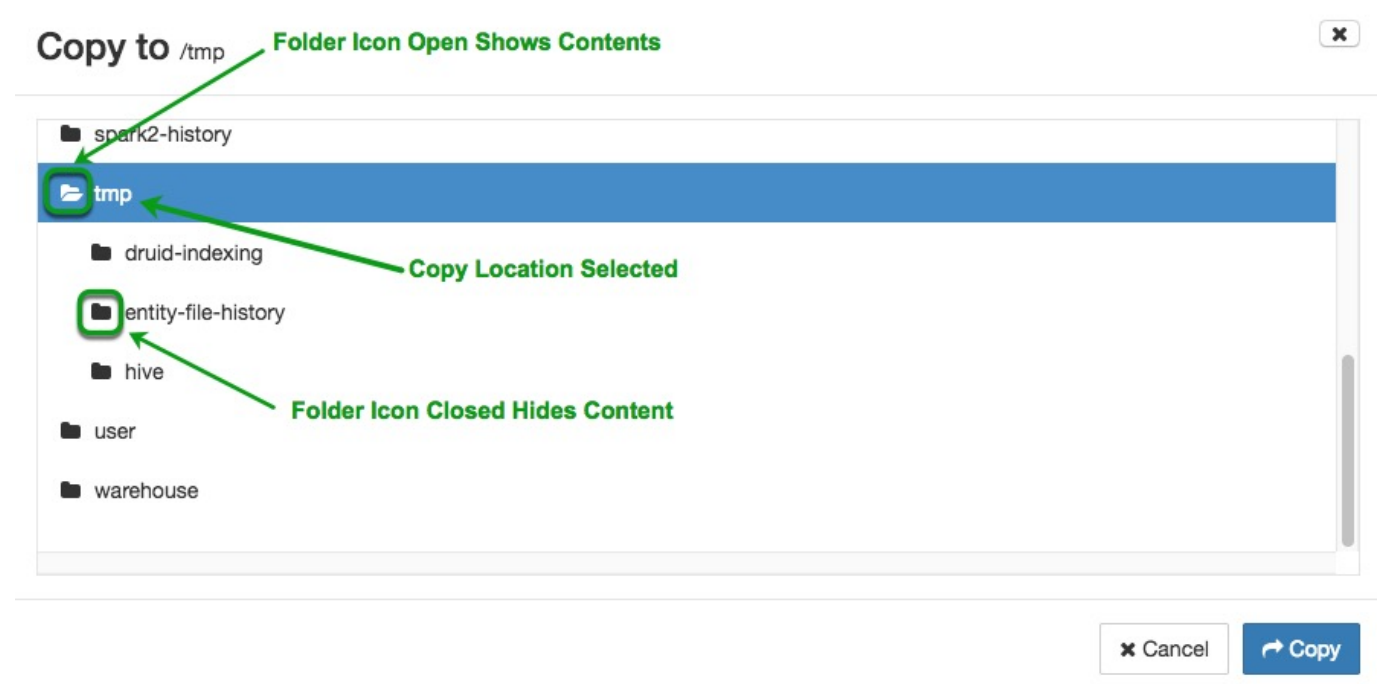
3. By default, Files View saves the merged files as a txt file, we can open the file and save it as a csv file. Then open the csv file and you will notice that all the content from geolocation is appended to the trucks file.

Copy Files or Directories recursively

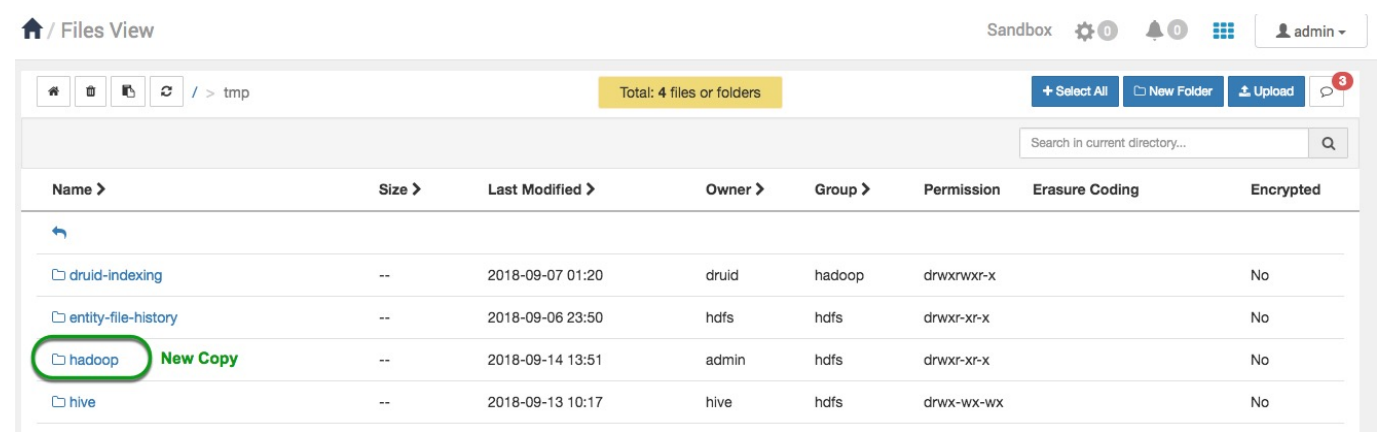
Copy file or directories recursively means all the directory's files and subdirectories to the bottom of the directory tree are copied. For instance, we will copy the **hadoop** directory and all of its contents to a new location within our hadoop cluster. In production, the copy operation is used to copy large datasets within the hadoop cluster or between 2 or more clusters.

1. Navigate to the **user** directory. Click on the row of the **hadoop** directory. Select the Copy button

2. The **Copy to** window will appear. Select the **tmp** folder, the row will turn blue. If you select the folder icon, the contents of **tmp** become visible. Make sure the row is highlighted blue to do the copy. Click the blue **Copy** button to copy the **hadoop** folder recursively to this new location.



3. A new copy of the **hadoop** folder and all of its contents can be found in the **tmp** folder. Navigate to **tmp** for verification. Check that all of the **hadoop** folder's contents copied successfully.



Summary

Congratulations! We just learned to use the Files View to manage our **geolocation.csv** and **trucks.csv** dataset files in HDFS. We learned to create, upload and list 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.

Further Reading

- [HDFS Overview](#)