Using HDFS with Ambari

Objective: Begin to get acquainted with Hadoops file system. And manipulate files in HDFS, the Hadoop Distributed File System.

Exercise directory: `~/data/iot`

In this section, you will download the sensor data and load that into HDFS using Ambari User Views. You will get introduced to the Ambari Files User View to manage files. You can perform tasks like create directories, navigate file systems and upload files to HDFS. In addition, you'll perform a few other file-related tasks as well. Once you get the basics, you will create two directories and then load two files into HDFS using the Ambari Files User View.

HDFS for IoT

A single physical machine gets saturated with its storage capacity as the data grows. This growth drives the need to partition your data across separate machines. This type of File system that manages storage of data across a network of machines is called Distributed File Systems. HDFS is a core component of Hadoop and is designed to store large files with streaming data access patterns, running on clusters of commodity hardware. With HDP, HDFS is now expanded to support heterogeneous storage media within the HDFS cluster.

Download and Extract Sensor Data Files

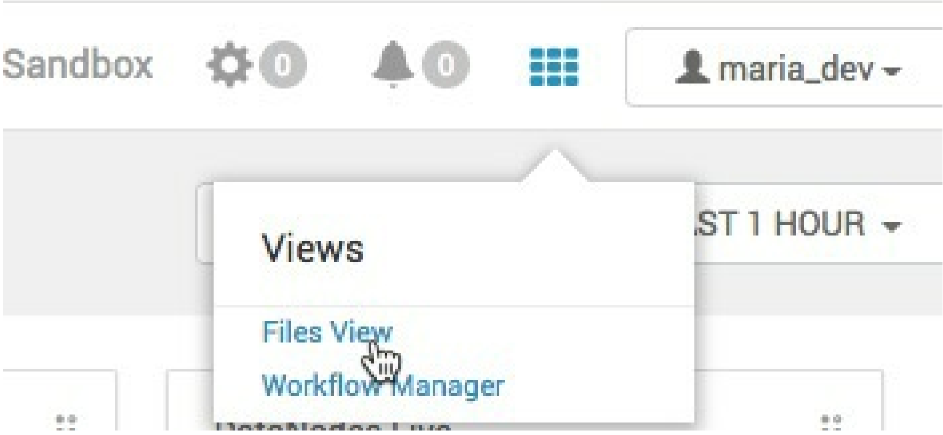
Download the sample sensor data either local or contained in a compressed (.zip) folder here: Geolocation.zip

Save the zip file to your computer, then extract the files. You should see a Geolocation folder that contains the following files:
* geolocation.csv – collected geolocation data from trucks. It contains records showing truck location, date, time, type of event, speed, etc.
* trucks.csv – exported from a relational database and it shows information on truck models, driverid, truckid, and aggregated mileage info.

Load the Sensor Data into HDFS

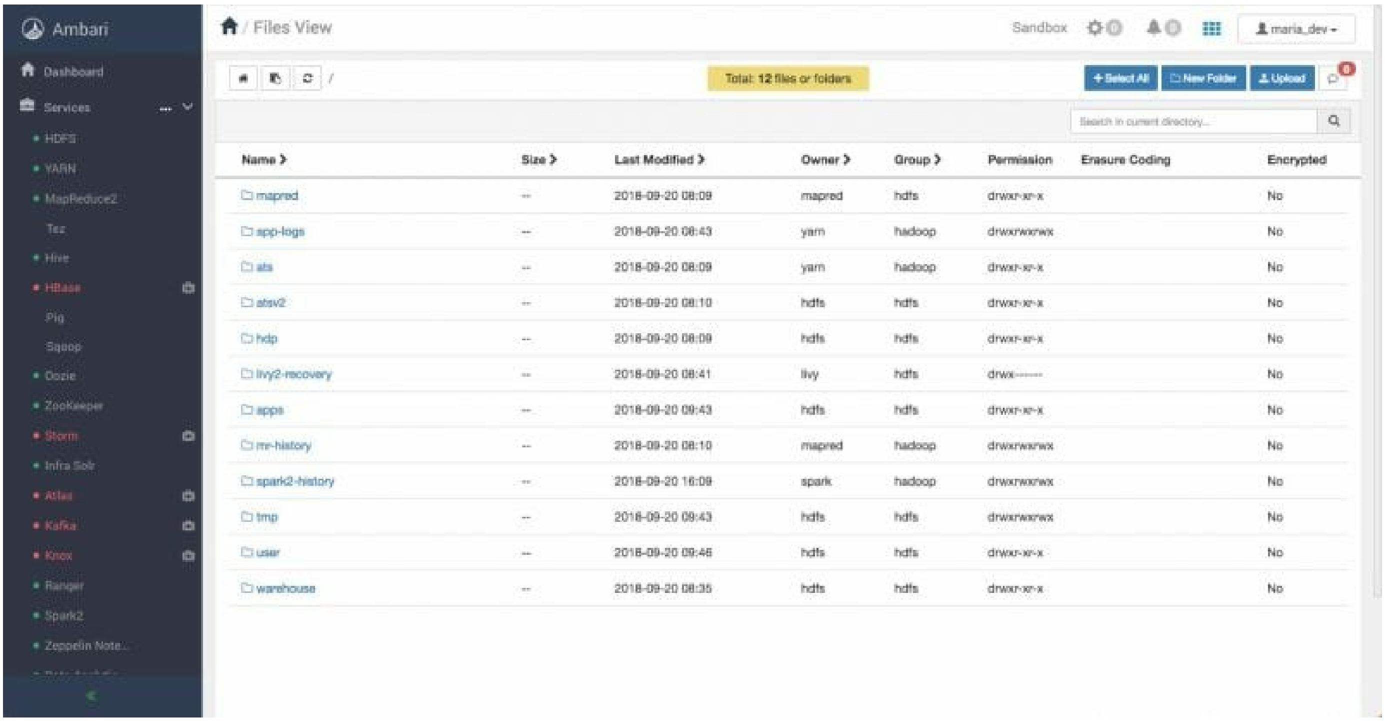
<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />
<h4>1. Logon to Ambari</h4>

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />
<h4>2. Go to Ambari Dashboard and open Files View.</h4>



<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />

<h4>3. Start from the top root of the HDFS file system, you will see all the files the logged in user has access to see:</h4>



Total: 12 files or folders

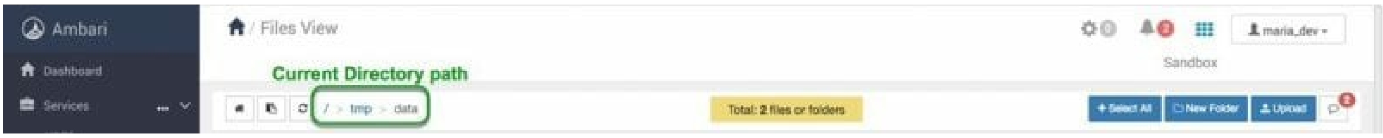
Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
mapred	--	2018-09-20 08:09	mapred	hdfs	drwxr-xr-x		No
app-logs	--	2018-09-20 08:43	yarn	hadoop	drwxrwxrwx		No
ats	--	2018-09-20 08:09	yarn	hadoop	drwxr-xr-x		No
atsv2	--	2018-09-20 08:10	hdfs	hdfs	drwxr-xr-x		No
hdp	--	2018-09-20 08:09	hdfs	hdfs	drwxr-xr-x		No
livy2-recovery	--	2018-09-20 08:41	livy	hdfs	drwx------		No
apps	--	2018-09-20 09:43	hdfs	hdfs	drwxr-xr-x		No
mr-history	--	2018-09-20 08:10	mapred	hadoop	drwxrwxrwx		No
spark2-history	--	2018-09-20 16:09	spark	hadoop	drwxrwxrwx		No
tmp	--	2018-09-20 09:43	hdfs	hdfs	drwxrwxrwx		No
user	--	2018-09-20 09:46	hdfs	hdfs	drwxr-xr-x		No
warehouse	--	2018-09-20 08:35	hdfs	hdfs	drwxr-xr-x		No

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-

9178-7bde851ac7bd.png" align="left" width="50" height="50" title="ToDo Logo" />
<h4>4. Navigate to `/tmp/` directory by clicking on the directory links.
</h4>

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align="left" width="50" height="50" title="ToDo Logo" />
<h4>5. Create directory data.</h4>

Click the New Folder button to create that directory. Then navigate to it. The directory path you should see: `/tmp/data`



Current Directory path

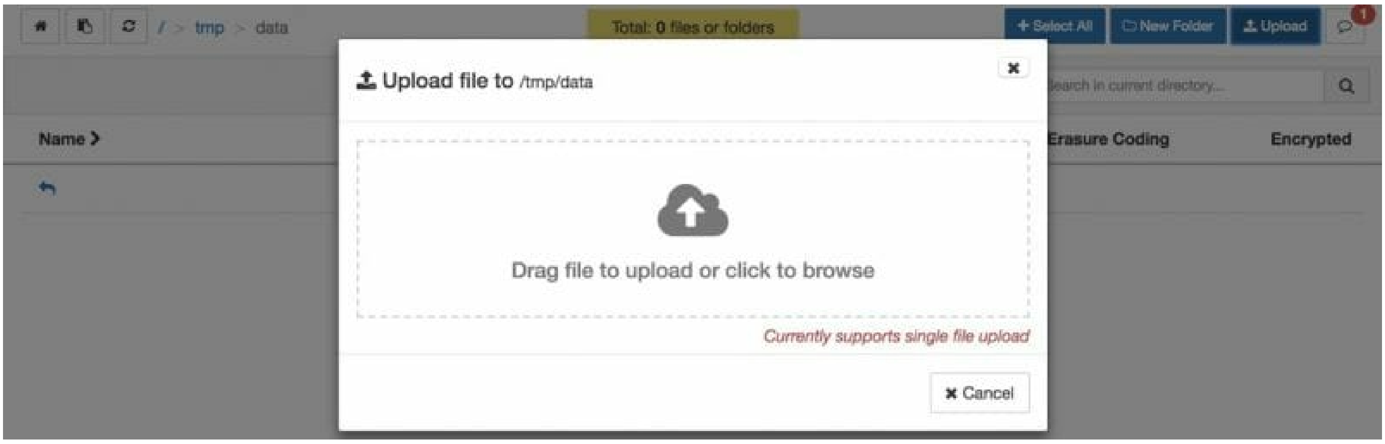
Upload Geolocation and Trucks CSV Files to data Folder

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align="left" width="50" height="50" title="ToDo Logo" />
<h4>1. If you're not already in your newly created directory path `/tmp/data`, go to the data folder. </h4>

Then click on the

Upload button to upload the corresponding geolocation.csv and trucks.csv files into it.

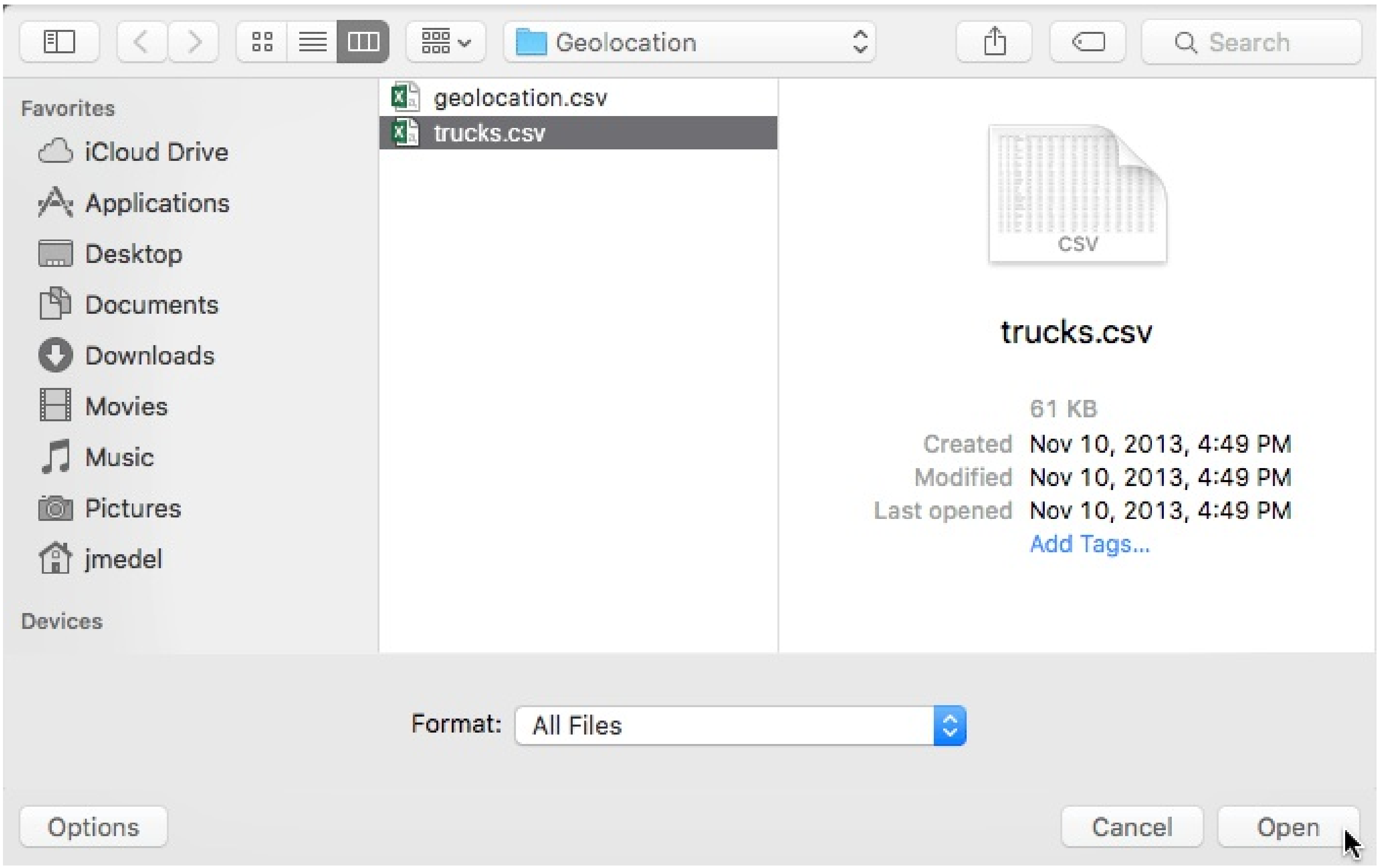
<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />
<h4>2. An Upload file window will appear, click on the cloud symbol.
</h4>


Upload file to /tmp/data

Drag file to upload or click to browse

Currently supports single file upload

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />
<h4>3. Another window will appear, navigate to the destination the two csv files were downloaded. Click on one at a time, press open to complete the upload. Repeat the process until both files are uploaded.
</h4>



trucks.csv

61 KB

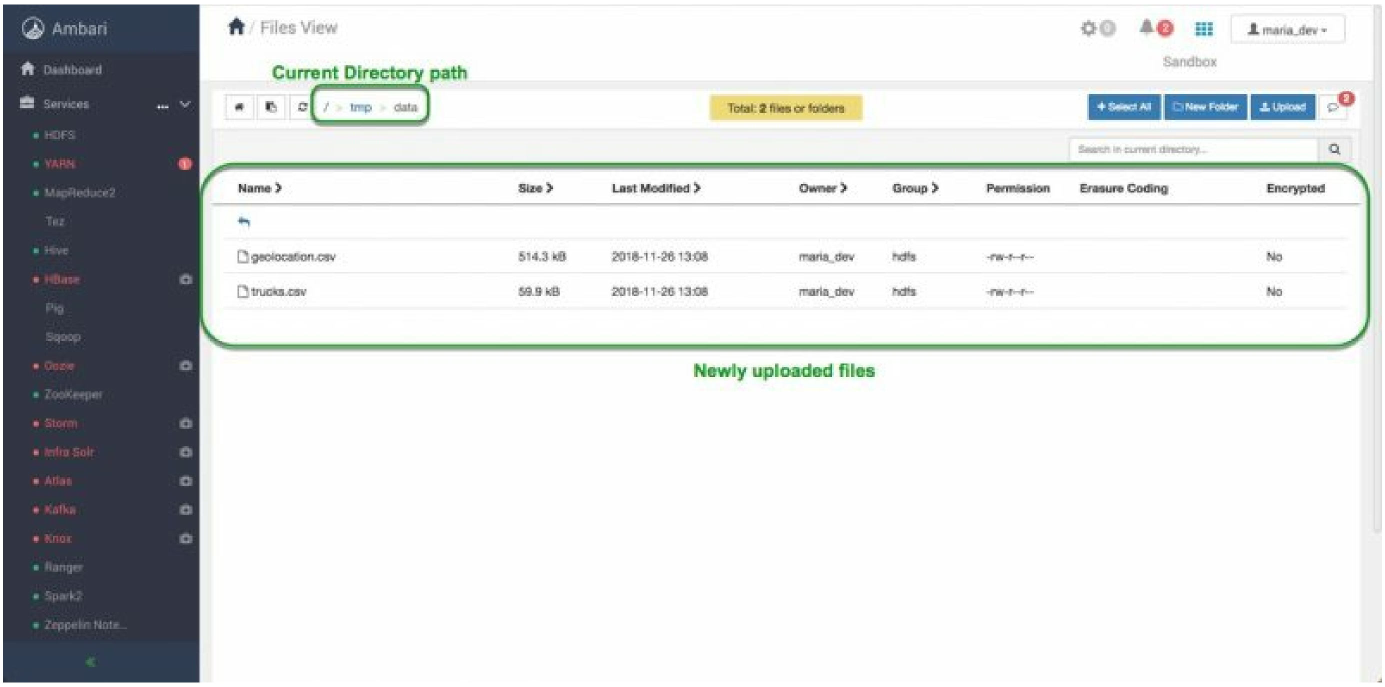
Created Nov 10, 2013, 4:49 PM

Modified Nov 10, 2013, 4:49 PM

Last opened Nov 10, 2013, 4:49 PM

Format: All Files

Both files are uploaded to HDFS as shown in the Files View UI:

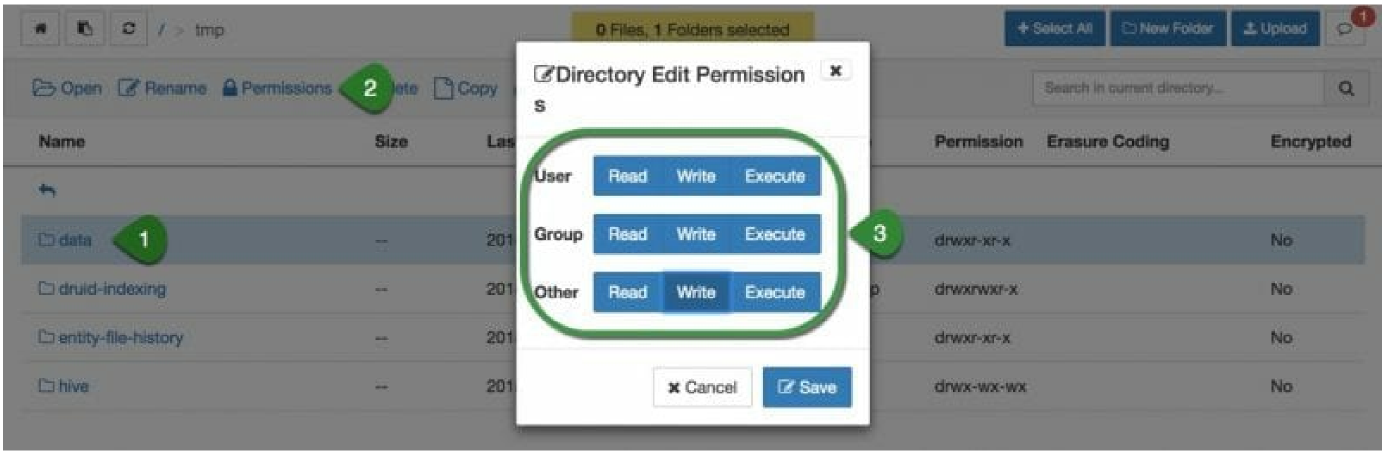


Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
geolocation.csv	514.3 kB	2018-11-26 13:08	maria_dev	hdfs	-rw-r--r--		No
trucks.csv	59.9 kB	2018-11-26 13:08	maria_dev	hdfs	-rw-r--r--		No

You can also perform the following operations on a file or folder by clicking on the entity’s row: `Open, Rename, Permissions, Delete,` `Copy, Move, Download and Concatenate`.

<img src="https://user-images.githubusercontent.com/558905/40613898-7a6c70d6-624e-11e8-9178-7bde851ac7bd.png" align=“left” width=“50” height=“50” title=“ToDo Logo” />
<h4>Set Write Permissions to Write to data Folder</h4>

<h4>1. click on the data folder’s row, which is contained within the directory path `/tmp/` </h4>
<h4>2. Click Permissions</h4>
<h4>3. Make sure that the background of all the write boxes are checked (blue)</h4>



Name	Size	Las	Permission	Erasure Coding	Encrypted
data	--	201	drwxr-xr-x		No
druid-indexing	--	201	drwxrwxr-x		No
entity-file-history	--	201	drwxr-xr-x		No
hive	--	201	drwx-wx-wx		No

Results

Congratulations! Let’s summarize the skills and knowledge we acquired from this tutorial. We learned Hadoop Distributed File System (HDFS) was built to manage storing data across multiple machines. Now we can upload data into the HDFS using Ambari’s HDFS Files view.

<button type=“button”><a href="https://virtuant.github.io/hadoop-overview-spark-hwx/">Go Back</a></button>

<br>

<br>