Getting Started with HDP Sandbox

Loading Sensor Data into HDFS

Getting Started with HDP Sandbox

OVERVIEW

1. Concepts
2. Loading Sensor Data into HDFS
3. Hive - Data ETL
4. Spark - Risk Factor
5. Data Reporting With Zeppelin

NOTICE

As of January 31, 2021, this tutorial references legacy products that no longer represent Cloudera's current product offerings.

Please visit recommended tutorials:

- How to Create a CDP Private Cloud Base Development Cluster
- All Cloudera Data Platform (CDP) related tutorials

Introduction

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.

Prerequisites

The tutorial is a part of series of hands on tutorial to get you started on HDP using Hortonworks sandbox. Please ensure you complete the prerequisites before proceeding with this tutorial.

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

Outline

- HDFS backdrop
- Download and Extract Sensor Data Files
- Load the Sensor Data into HDFS
- Summary
- Further Reading

HDFS backdrop

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 Apache Hadoop and is designed to store large files with streaming data access patterns, running on clusters of commodity hardware. With Hortonworks Data Platform (HDP), HDFS is now expanded to support heterogeneous storage media within the HDFS cluster.

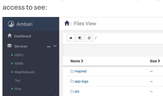
Download and Extract Sensor Data Files

1. Download the sample sensor data contained in a compressed (.zip) folder here: **Geolocation.zip**
2. Save the **Geolocation.zip** file to your computer, then extract the files. You should see a Geolocation folder that contains the following files:
   - **geolocation.csv** – This is the collected geolocation data from the trucks. It contains records showing *truck location, data, time, type of event, speed, etc.*
   - **trucks.csv** – This is data was exported from a relational database and it shows information on *truck models, driverid, truckid, and aggregated mileage info.*

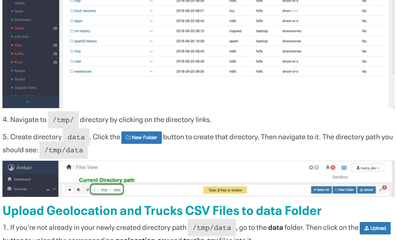
Load the Sensor Data into HDFS

1. Logon to Ambari using: **maria_dev/maria_dev**

2. Go to Ambari Dashboard and open **Files View**.

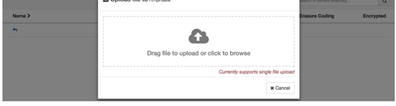


3. Start from the top root of the HDFS file system, you will see all the files the logged in user (**maria_dev** in this case) has access to see:



4. Navigate to `/tmp/` directory by clicking on the directory links.

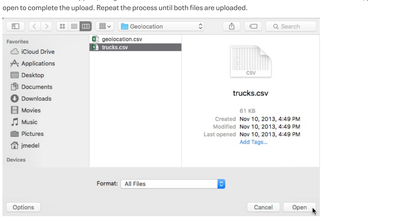
5. Create directory `data` . Click the **New Folder** button to create that directory. Then navigate to it. The directory path you should see: `/tmp/data`



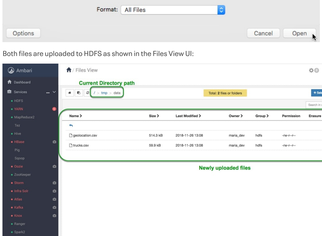
Upload Geolocation and Trucks CSV Files to data Folder

1. If you're not already in your newly created directory path `/tmp/data` , go to the **data** folder. Then click on the **Upload** button to upload the corresponding **geolocation.csv** and **trucks.csv** files into it.

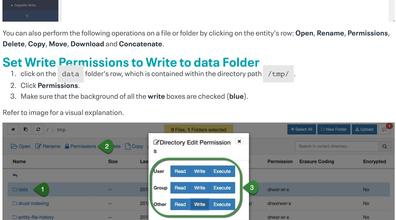
2. An **Upload file** window will appear, click on the cloud symbol.



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.



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

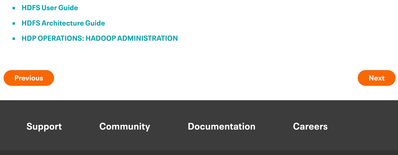


You can also perform the following operations on a file or folder by clicking on the entity's row: **Open**, **Rename**, **Permissions**, **Delete**, **Copy**, **Move**, **Download** and **Concatenate**.

Set Write Permissions to Write to data Folder

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

Refer to image for a visual explanation.



Summary

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**.

Further Reading

- HDFS
- HDFS User Guide
- HDFS Architecture Guide
- HDP OPERATIONS: HADOOP ADMINISTRATION

Previous Next

Contact Us
US: +1 888 789 1488
Outside the US: +1 650 362 0488

English