

CM2606 Data Engineering

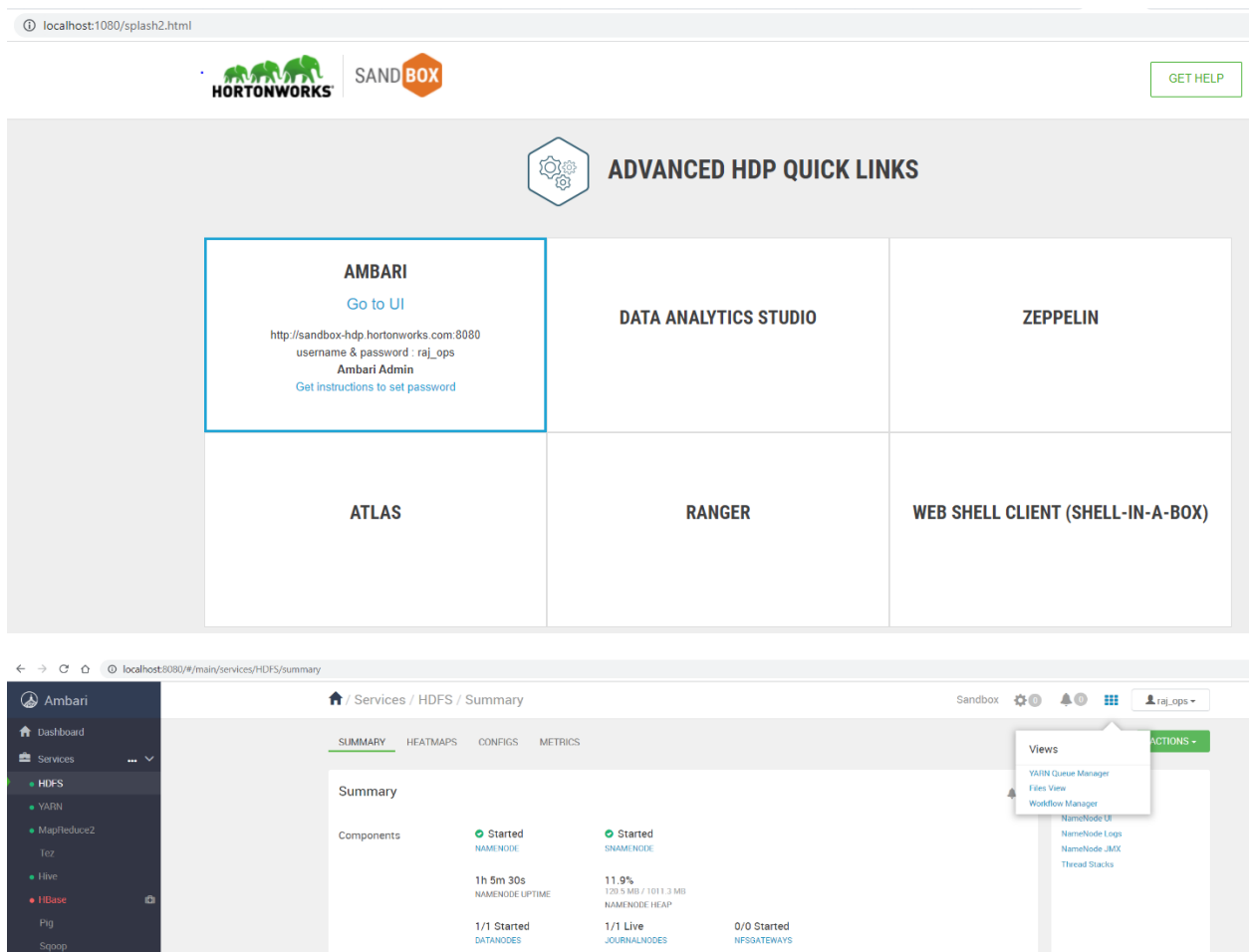
Tutorial 03

The aim of this tutorial is to:

- Give students hands-on experience with working on top of Hadoop Distributed File System (HDFS).
- 1. If you are using the sandbox for Hadoop Log into the sandbox VM as the root using the link <http://localhost:1080> Or else, you can use your local machine with Hadoop installed.
- 2. Download the below or any other sample data files in to the VMs file system or to your local machine.
#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>
- 3. Use ls command to see the downloaded files in VM
- 4. Now switch to hdfs user from root user and then go one directory up from root directory to home/hdfs directory level. Use <su hdfs> followed by <cd>
- 5. You can use <pwd> command to check the current directory path
- 6. Give the access to root user to make changes to the user folder on hdfs
<hdfs dfs -chmod 777 /user>
- 7. Now we can exit the hadoop user and make changes in hdfs by appending <hdfs dfs> when referring to hadoop file system

8. Possible operations that could be done on top of hdfs could be found [here](#).
9. Use mkdir command to create two directories to store the two csv files
10. Use put command to move the two files from VM's local file system to hdfs.
11. Use ls command to check whether files are correctly copied and du command to check out the file sizes.
12. Now use get command to move files from hdfs to local file system
13. Finally use the cp command to copy files from one hdfs location to another
14. Now try to repeat the steps from 8 to 12 by using the visual controls available in Ambari dashboard >> Files View to create folders, upload files etc,



The screenshot displays the Ambari dashboard interface. At the top, there's a navigation bar with the Hortonworks logo, a 'SAND BOX' label, and a 'GET HELP' button. Below this is the 'ADVANCED HDP QUICK LINKS' section, which contains a grid of links to various services: AMBARI (with a 'Go to UI' link and login details), DATA ANALYTICS STUDIO, ZEPPELIN, ATLAS, RANGER, and WEB SHELL CLIENT (SHELL-IN-A-BOX). The AMBARI link is highlighted with a blue border.

Below the quick links, the 'HDFS Summary' page is shown. It features a sidebar with navigation options like Dashboard, Services, and HDFS. The main content area displays the 'Summary' tab for HDFS, showing components like NAME NODE, SNAMENODE, DATANODES, JOURNALNODES, and NFS GATEWAYS with their respective status and metrics. A 'Views' dropdown menu is open on the right, showing options like YARN Queue Manager, Files View, Workflow Manager, NameNode UI, NameNode Logs, NameNode JMX, and Thread Stacks.

15. You may refer to [this cloudera tutorial](#) if you get stuck in any of the steps.