Big Data Lab: Hadoop Command Line Filesystem Operations

Author: Walid Belal, September 2020

Revisions by Michael McKee, September 2020

Tutorial summary

In this tutorial you will become familiar with command line operations on a Hadoop cluster by performing the following steps:

1. Lab preparation
2. Start a Hadoop cluster on the Sheridan private cloud
3. Use Hadoop command lines to create folders in HDFS
4. Copy a text file from your Hadoop namenode's local file system to HDFS
5. Copy a text file from HDFS to your namenode's local file system
6. Clean up by deleting the files just copied to HDFS

Lab Pre-requisites

You must be able to connect to your **Prod** Hadoop cluster in the Sheridan private cloud. See the document **Software Required for the Course** and the video: Connect to Your Hadoop Cluster on Sheridan Cloud.

Note: Your **Prod** cluster is specified since it has Hadoop already installed and configured. Instead of your **Prod** cluster, you may use your **Dev** cluster if you have successfully installed and configured Hadoop on it.

1. Lab preparation

Connect to the Sheridan VPN

Open Cisco AnyConnect Secure Mobility Client.

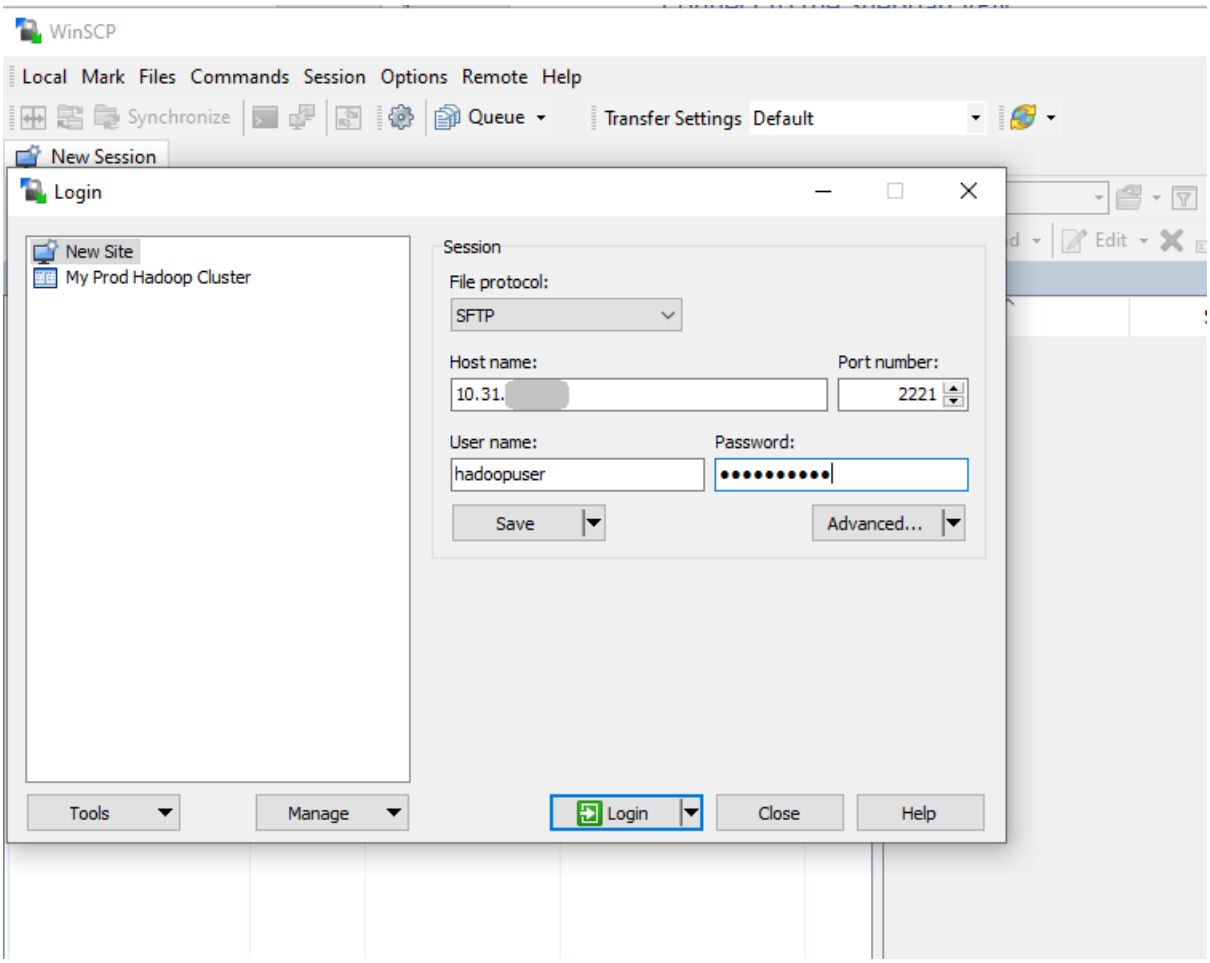
Specify the VPN *vpn.sheridancollege.ca* and click **Connect**.

Provide your Sheridan credentials and click OK.

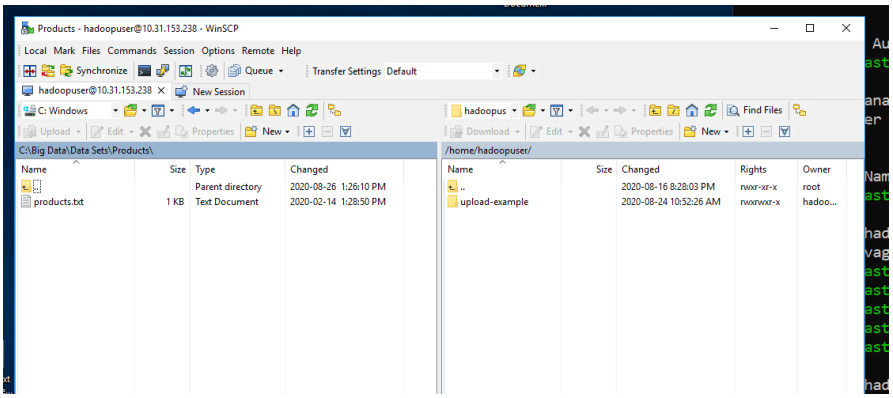
Download a sample file from Slate and move it to the Hadoop master node

Download the file from Slate – location will be provided by your instructor.

Open WinSCP. Specify your own cluster's IP address and port.



When connected, the left panel will be your laptop file system, and the right will be the master node's local file system.



Drag the products.txt file from your file system into the directory

/home/hadoopuser/upload-example

2. Start the Hadoop cluster

Connect to Hadoop cluster's namenode from your laptop using SSH

You should have already connected to the Sheridan VPN at this point. Please see above if you have not.

Open **cmd**

Type *ssh hadoopuser@<insert your IP address> -p 2221*

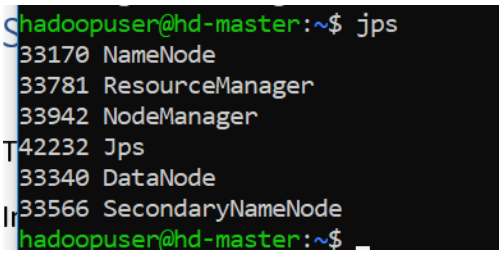
Note: replace *<insert your IP address>* with the IP address of your Sheridan cloud router.

You will be prompted to enter password. The default password is **Sher1dan**

Start Hadoop

Type *jps* to check if Hadoop is started.

You need to see the below processes



If you do not see the above, you need to start Hadoop

In your ssh connection type *start-all.sh*

The daemons should start

Type *jps* again and check that the processes started

3. Use Hadoop command lines to create folders in HDFS

From the ssh command prompt, type the following

hadoop fs -mkdir /products

This will create directory called **products** under the Hadoop root directory

You can list the directories by typing ***hadoop fs -ls /***

You should see something like the following

You can make a tree of folders exactly like any other file system.
Create two folders caller **prod1** and **prod2** inside products folder.

hadoop fs -mkdir /products/prod1

hadoop fs -mkdir /products/prod2

Now you can use **ls** command to see the folders

hadoop fs -ls /products

You should see a result similar to the below image

```
drwxr-xr-x   - hadoopuser supergroup          0 2020-02-25 15:21 /products/prod1
drwxr-xr-x   - hadoopuser supergroup          0 2020-02-25 15:22 /products/prod2
```

4. Copy a file from Hadoop namenode's local file system to HDFS

In this step you will copy the file **products.txt** from a local folder on master node (e.g. hadoop1 or hd-master) Ubuntu machine to the HDFS filesystem.

hadoop fs -copyFromLocal /home/hadoopuser/upload-example/products.txt /products/prod1

Check that the file exists in HDFS.

hadoop fs -ls /products/prod1

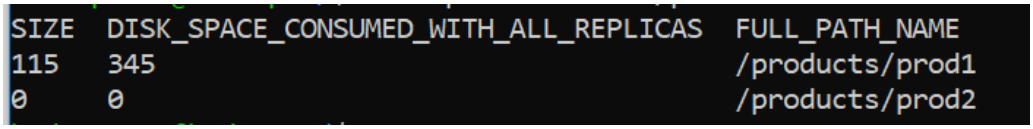
You should see an output like the following:

```
-rw-r--r--   3 hadoopuser supergroup        115 2020-02-25 15:39 /products/prod1/products.txt
```

You can check the size consumed by every folder:

hadoop fs -du -v /products

You should see the following:



The Prod1 folder will contain 115 bytes of data which is the size of products.txt table and in total 345 bytes because there are 3 replicas of the file on HDFS (remember that the replication factor was set to 3)

You can also use cat command to see the contents of the file:

hadoop fs -cat /products/prod1/products.txt

5. Copy file from HDFS to the local file system

hadoop fs -copyToLocal /products/prod1/products.txt /home/hadoopuser/upload-example/productsCopy.txt

You can use Linux cat command to see the content of the copied file:

cat /home/hadoopuser/upload-example/productsCopy.txt

6. Clean up: delete the products folder and its subfolders from HDFS

hadoop fs -rm -R /products