LAB-4

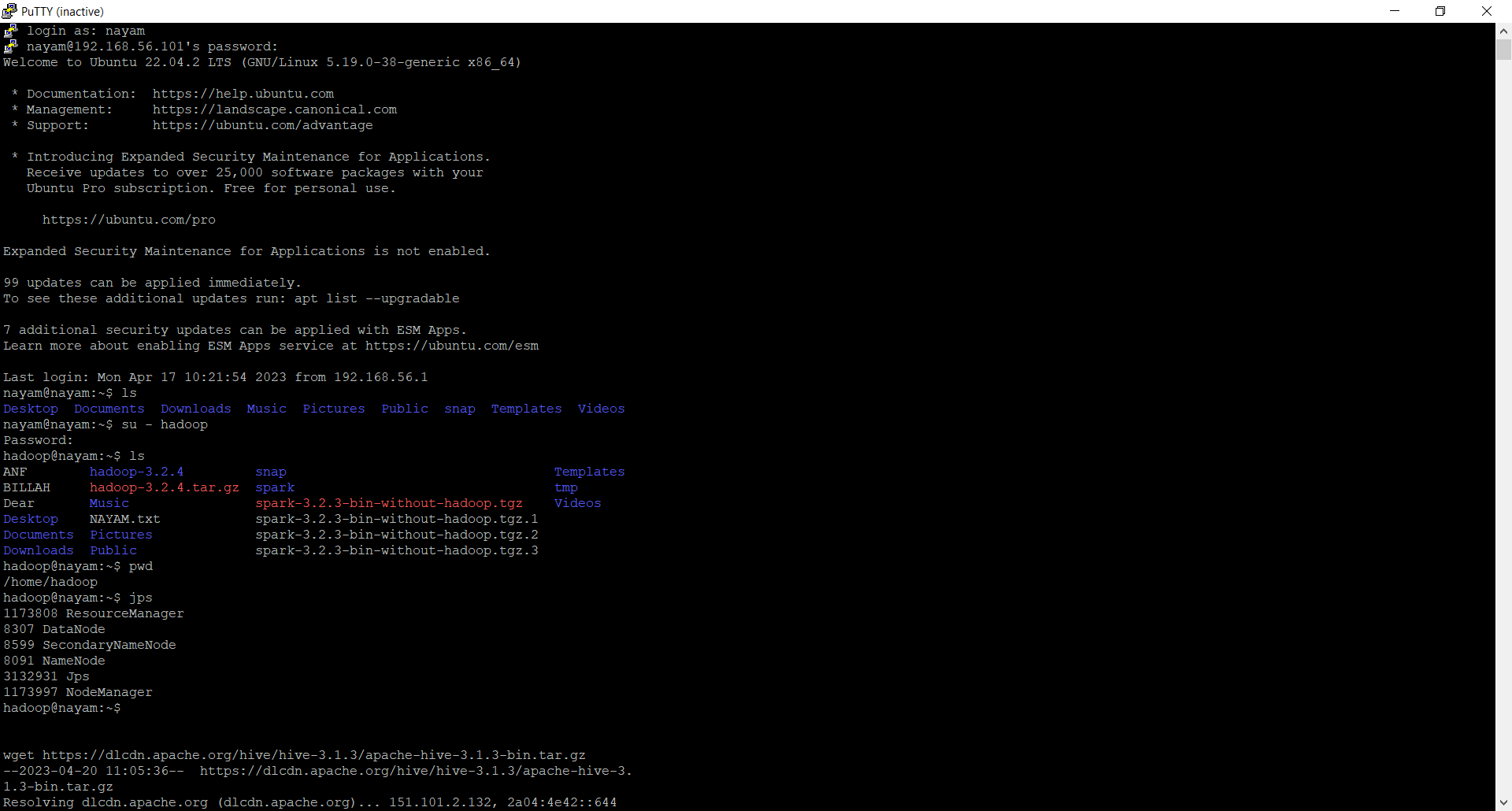
In this case, we are going to work with Apache Hive. Apache Hive is an open-source data warehouse software for reading, writing and managing large data set files that are stored directly in either the Apache Hadoop Distributed File System (HDFS) or other data storage systems such as Apache HBase.

A machine with Ubuntu 20.04 LTS operating system installed as well as Apache Hadoop 3.2.4 pre-installed.

We use Apache Hive 3.1.3. Download it from <https://www.apache.org/dyn/closer.cgi/hive/>

In bash terminal, run the following command to download the package:

wget https://dlcdn.apache.org/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz

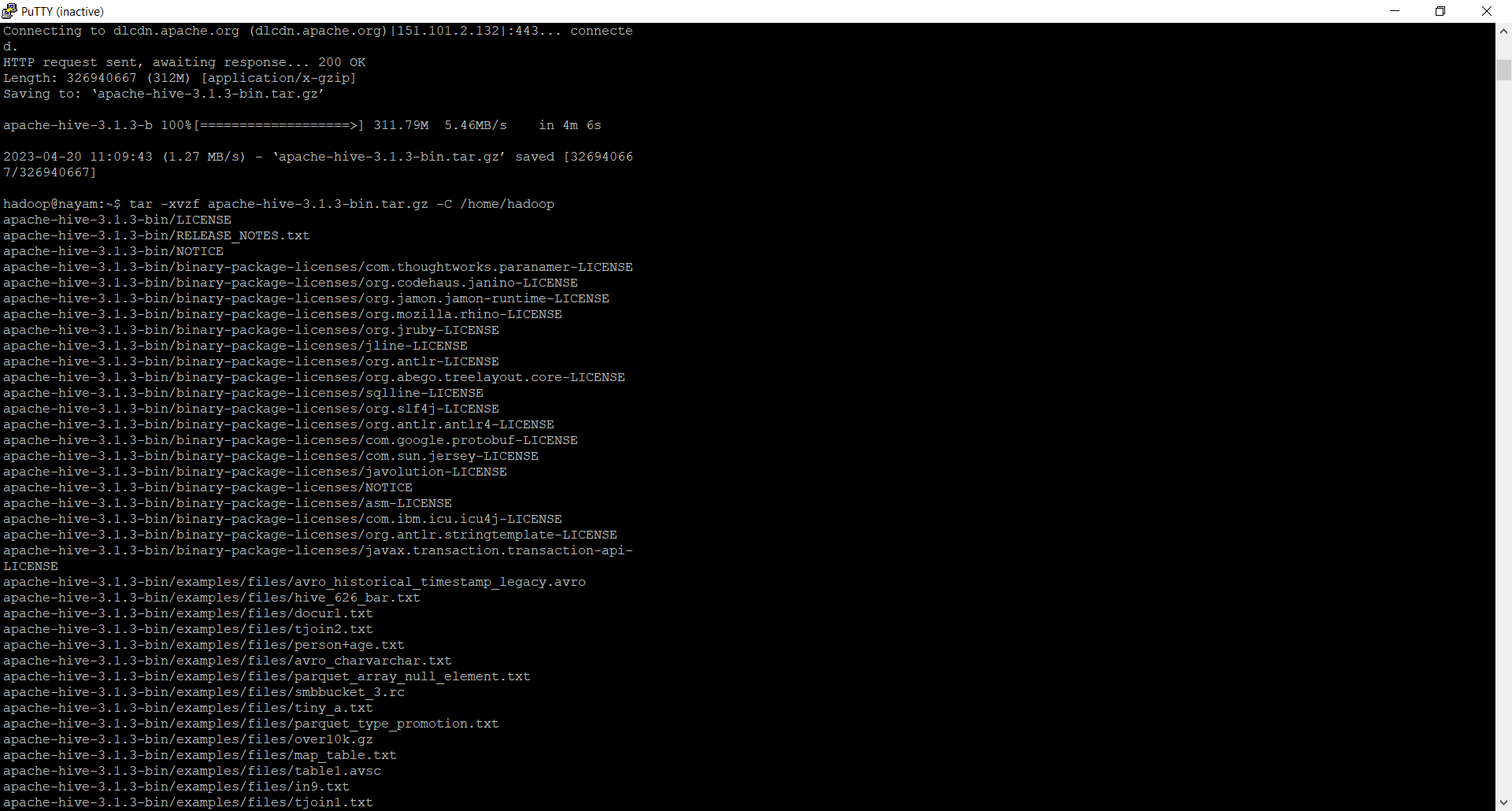


After downloading, unzip binary package

Ensure apache-hive-3.1.3-bin.tar.gz is in /home/hadoop.

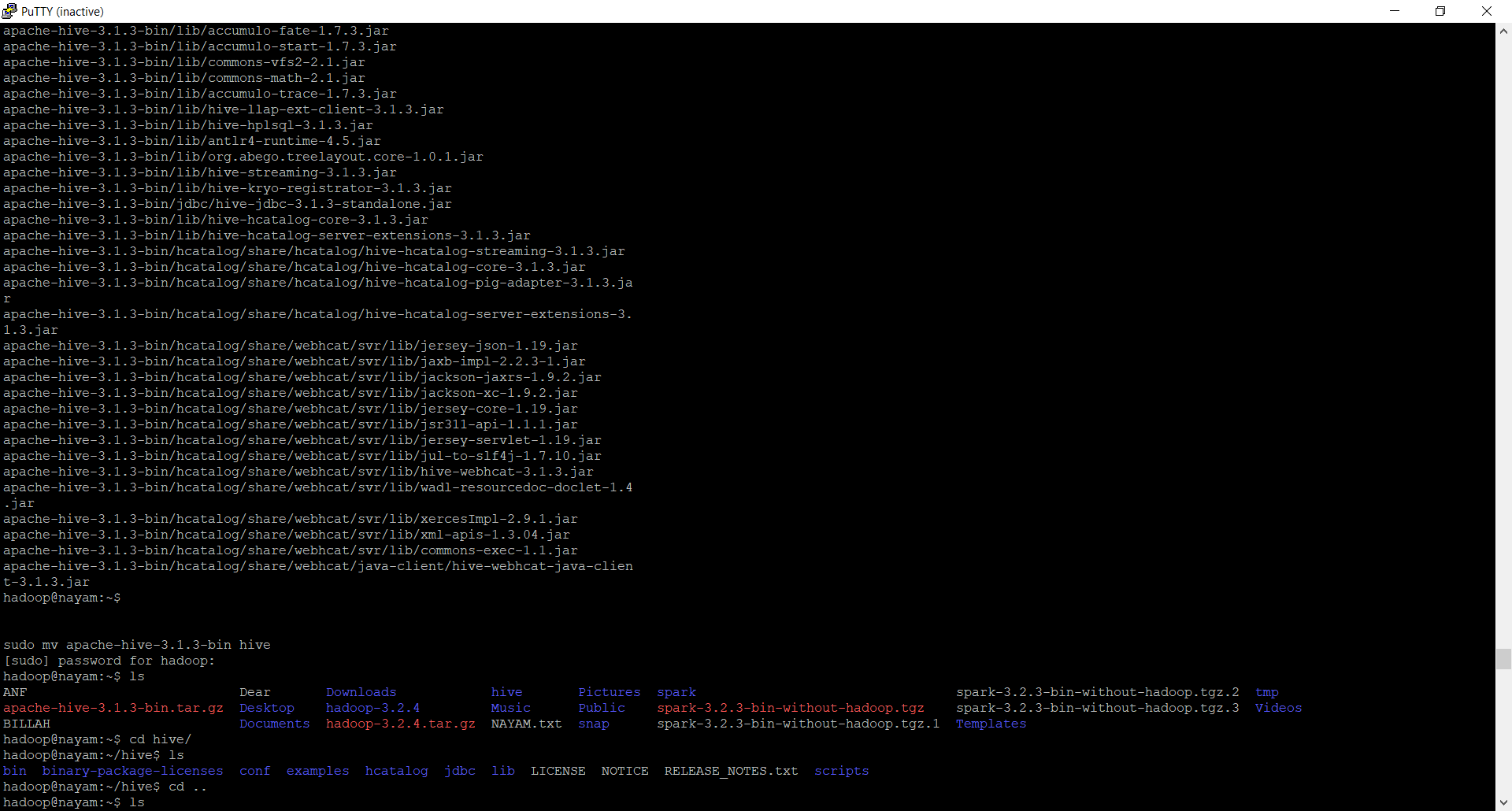
Now unzip Hive package using the following command:

tar -xvzf apache-hive-3.1.3-bin.tar.gz -C /home/Hadoop



sudo mv apache-hive-3.1.3-bin hive

now, if we check the current status of the system then we can see that there is a file which is Hive. We can also change the name if the Hive. Hive to Hive4 by using mv hive hive4 command. Finally, use ls command to check the current name.



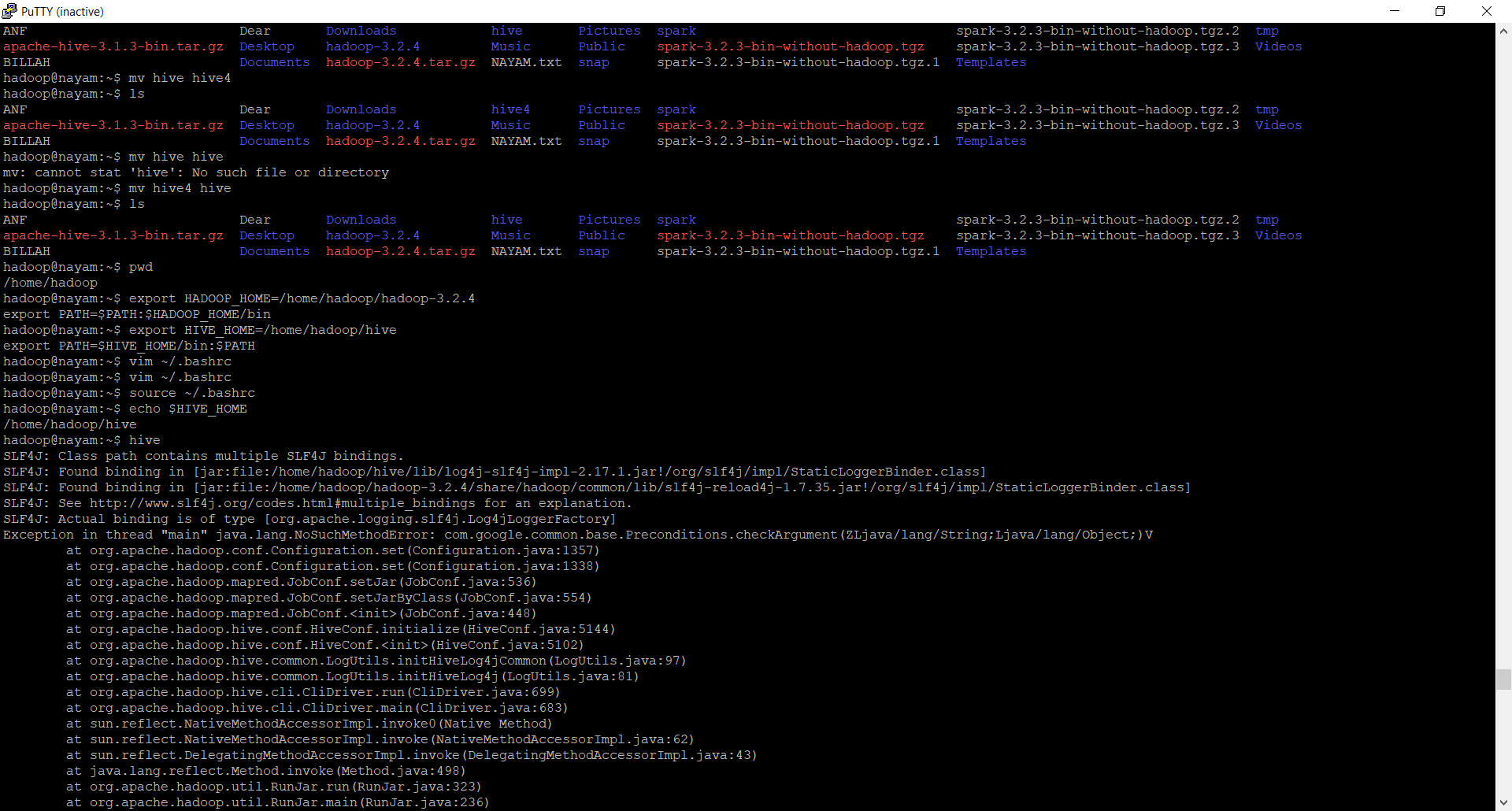
In the **hadoop**folder there are now two subfolders at least (one for Hadoop and another for Hive)

For setup environment variables,

in the prerequisites sections, we’ve already configured some environment variables like the following:

export HADOOP\_HOME=/home/hadoop/hadoop-3.2.4

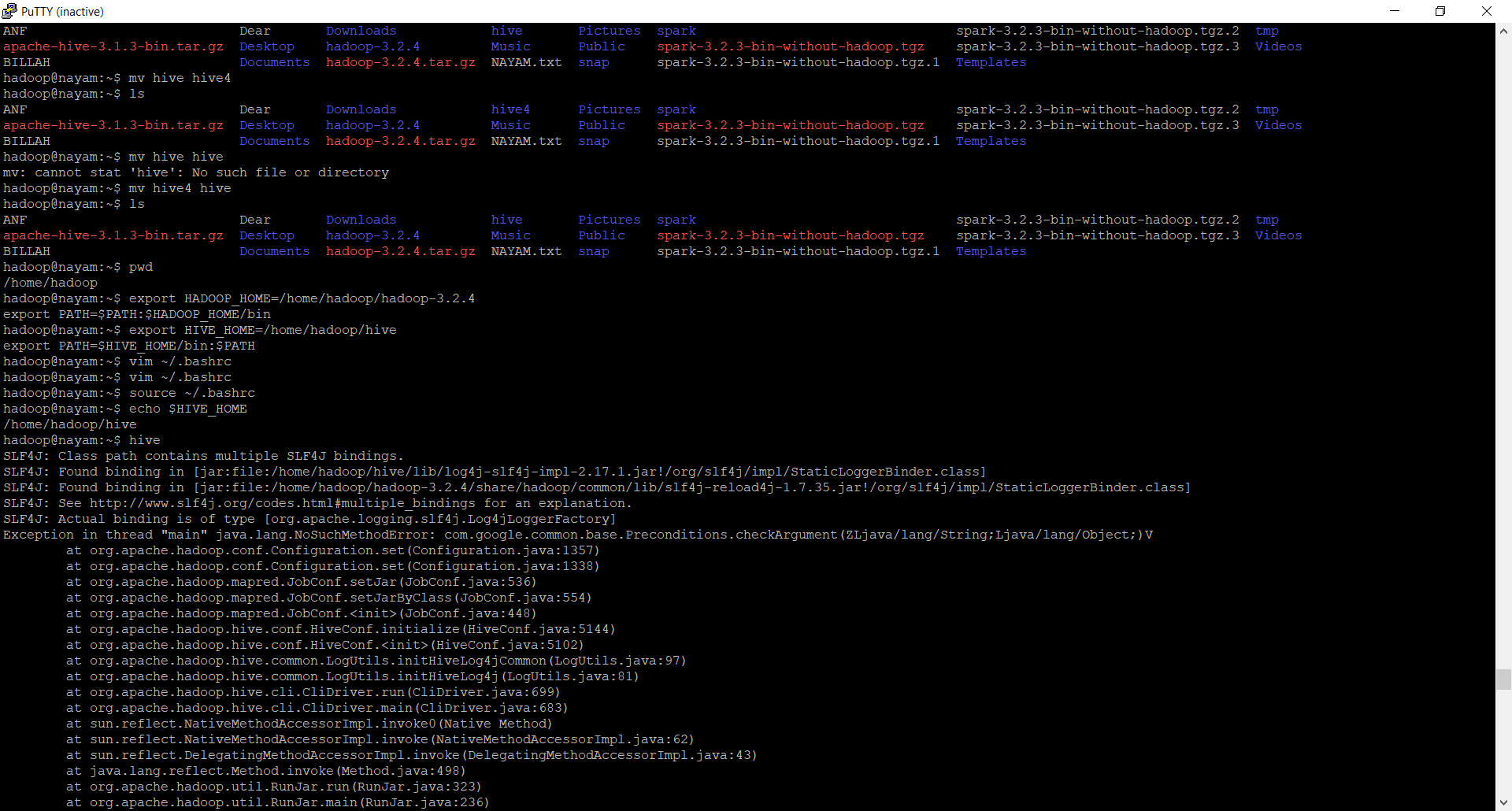
export PATH=$PATH:$HADOOP\_HOME/bin



Note: your user name can be different.

Let’s run the following command to add Hive required environment variables into **.bashrc** file too:

vim ~/.bashrc

 Add the following lines to the end of the file:

export HIVE\_HOME=/home/hadoop/hive

export PATH=$HIVE\_HOME/bin:$PATH

Run the following command to source the variables:

source ~/.bashrc

Verify the environment variables:

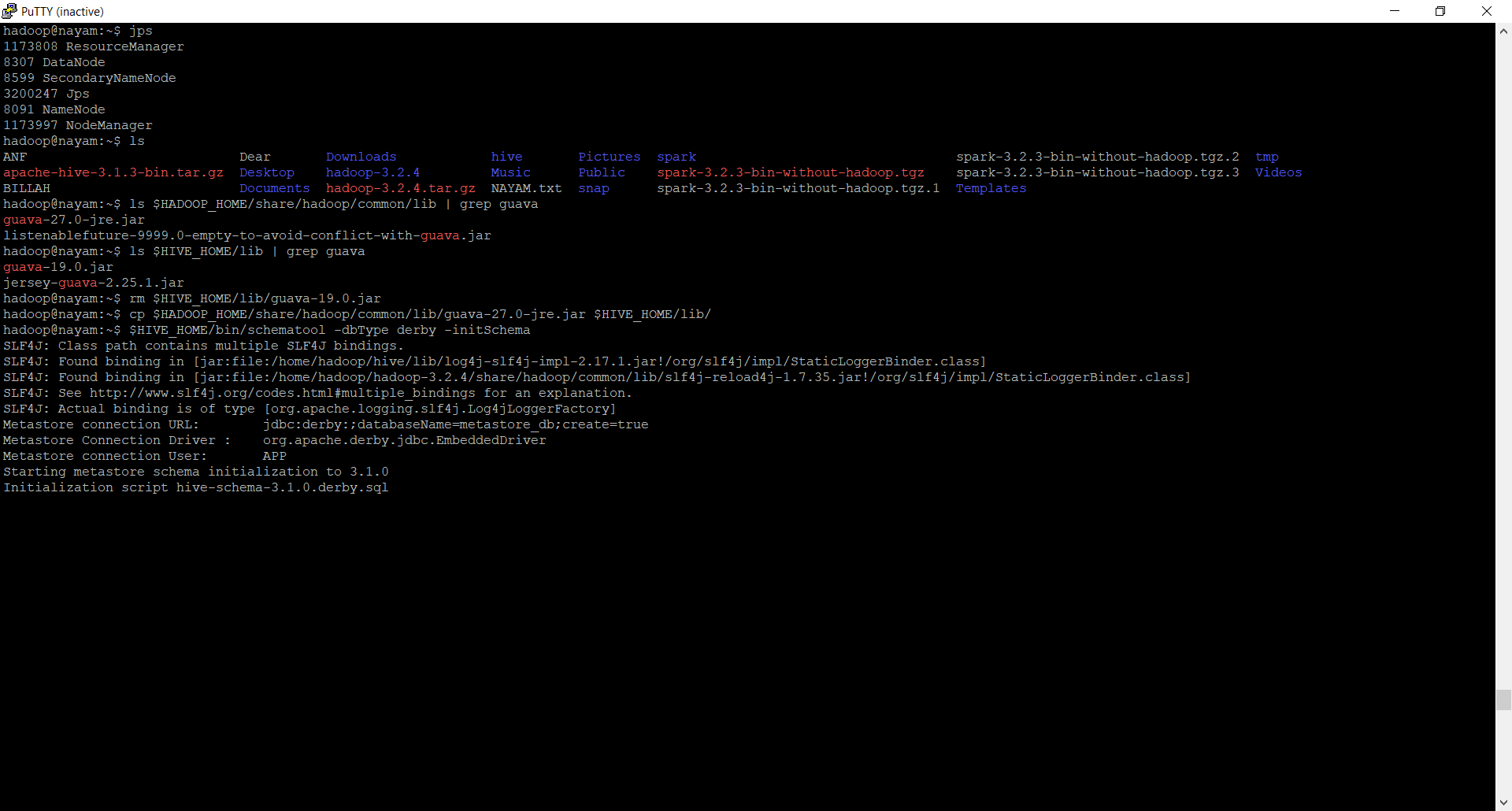
echo $HIVE\_HOME

/home/hadoop/hive

After all, we don’t need to test Hadoop, because, we have already done it. We can also run jps command to make sure all the services are running successfully.

Let’s move on there…

guava lib issue,



 let's ensure guava library version is consistent between Hive and Hadoop.

* Hadoop 3.2.4 -

$ ls $HADOOP\_HOME/share/hadoop/common/lib | grep guava

guava-27.0-jre.jar

* Hive 3.1.3 -

$ ls $HIVE\_HOME/lib | grep guava

guava-19.0.jar

jersey-guava-2.25.1.jar

* Copy newer version from Hadoop to Hive:

rm $HIVE\_HOME/lib/guava-19.0.jar

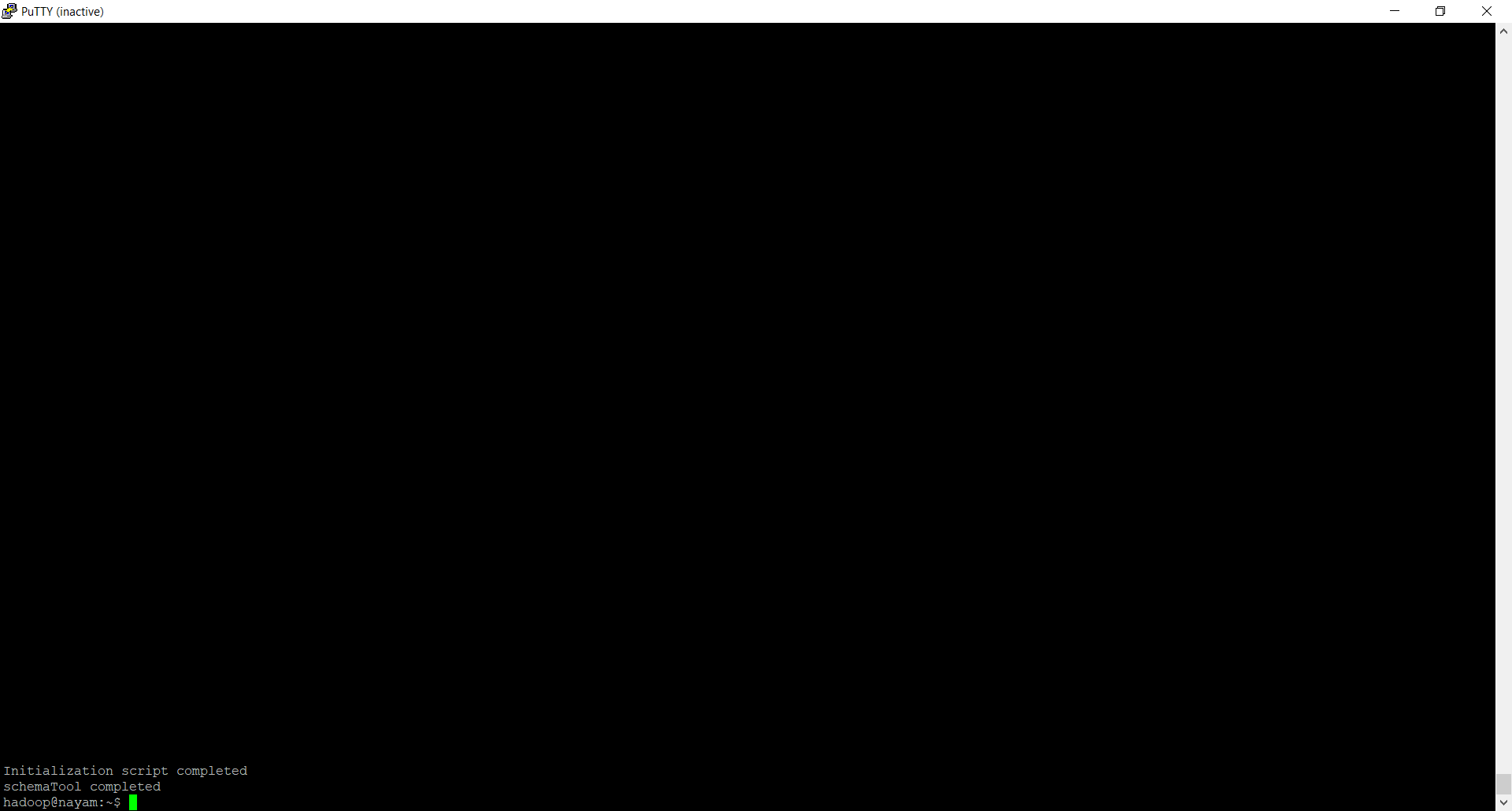
cp $HADOOP\_HOME/share/hadoop/common/lib/guava-27.0-jre.jar $HIVE\_HOME/lib/

ls $HADOOP\_HOME/share/hadoop/common/lib | grep guava

Initialize database structure

Now we need to run schematool to setup metastore for Hive. The command syntax looks like the following:

$HIVE\_HOME/bin/schematool -dbType derby -initSchema



Finally, we got the initialization script completed.