## **Installation Instructions (Hadoop, Spark, Hive, Derby)**

## **1. Installation of Hadoop 3.1.3 in ubuntu 18.04/19.04/19.10**

**Step 1: Installation of openJDK-8 (In loclt user)**

**$ sudo apt update**

$ sudo apt install openjdk-8-jdk openjdk-8-jre -y (tải java, javac)

$ java -version # kiểm tra version java

$ javac -version

**Step 2: Adding the Jdk path to the path variable (Ở loclt User)**

Open ~/.bashrc and add

$ sudo nano ~/.bashrc

# go to the last line and add the following

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export PATH=$PATH:$JAVA\_HOME

## save and exit

**Inform the OS about the modification**

**$ source ~/.bashrc**

Type

$ echo $JAVA\_HOME

$ echo $PATH

**Step 3: Add a dedicated user for the HADOOP**

$ sudo adduser hadoop

$ sudo usermod -aG sudo hadoop

**Step 4: Once the user is added, login to the user “Hadoop” to generate the ssh key for passwordless login ( hadoop@machinename) – Làm hết trong Hadoop User luôn (hadoop)**

$ sudo su - hadoop

$ sudo apt install openssh-server openssh-client (cài đặt openssh để chạy ssh – lệnh này để có thể chuyển tiếp tới máy khác và thao tác mọi thứ trên đó, shh này khi tải thì sẽ dùng được trên toàn bộ hệ điều hành)

$ ssh-keygen -t rsa

$ cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

$ chmod 0600 ~/.ssh/authorized\_keys

Check the login to localhost using ssh is valid

$ ssh localhost

**IMPORTANT**

**Once the connection is made, logout from ssh**

$ exit

**Step 5: Download the latest binary from Hadoop site**

**$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz**

$ tar -xvzf **hadoop-3.3.6.tar.gz**

$ sudo mv **hadoop-3.3.6 /usr/local/hadoop (lệnh sudo để có permissioned)**

**Step 6: Setup the path variables for hadoop**

**$ sudo nano** /etc/profile.d/hadoop\_java.sh

Add the following lines to it

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export HADOOP\_HOME=/usr/local/hadoop

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

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

export HADOOP\_OPTS="$HADOOP\_OPTS -Djava.library.path=$HADOOP\_HOME/lib/native"

**Save and exit. Then source the file**

$ source /etc/profile.d/hadoop\_java.sh

**Confirm your hadoop and hdfs version**

$ hadoop version

$ hdfs version

**Step 7: Configuring Hadoop**

**- Give the permission for the hadoop folder to hadoop user**

**$ cd $HADOOP\_HOME**

$ sudo chown -R hadoop:hadoop /usr/local/hadoop

- Navigate to /usr/local/hadoop/etc/hadoop and type ls

$ cd /usr/local/hadoop/etc/hadoop (chuyển tới thư mục chứa các file .sh để config Hadoop)

$ hadoop@machine: /usr/local/hadoop/etc/hadoop: ls (ls để xem các file cần config)

**Step 7a: Specify JAVA\_HOME in hadoop-env.sh (/usr/local/hadoop/etc/hadoop)**

$ nano hadoop-env.sh

=> chính xác là: sudo nano /usr/local/hadoop/etc/hadoop/hadoop\_env.sh

Add the following line in java implementation

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64 (54 line)

Save and exit

**Step 7b: Modify core-site.xml to setup web portal for hadoop**

$ nano core-site.xml

Add the following lines to it

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9000</value>

<description>The default file system URI</description>

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>/usr/local/hadoop/htemp</value>

</property>

</configuration>

**Step 7c: Modify hdfs-site.xml to setup namenode and datanode path and replication factor**

**Create a folder for namenode and datanode usage**

**$ cd $HADOOP\_HOME**

$ sudo mkdir -p htemp

$ sudo mkdir -p hdfs/{namenode,datanode}

**Give the permission for the hdfs and htemp folder to hadoop user**

$ sudo chown -R hadoop:hadoop /usr/local/hadoop/hdfs

$ sudo chown -R hadoop:hadoop /usr/local/hadoop/htemp

**Modify hdfs-site.xml and add the following lines inside**

$ nano hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.name.dir</name>

<value>file:/usr/local/hadoop/hdfs/namenode</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>file:/usr/local/hadoop/hdfs/datanode</value>

</property>

</configuration>

**Step 7d: Configure the mapreduce framework by editing the mapred-site.xml**

$ nano mapred-site.xml

**Modify the mapred-site.xml and add the following lines**

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>mapreduce.application.classpath</name> <value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>

</property>

</configuration>

**Step 7e: Configure the YARN resource manager by editing the yarn-site.xml**

$ nano yarn-site.xml

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

</configuration>

**Step 8: Format the namenode using the command (Để chạy Hadoop phải chạy lại từ đây)**

**(Vẫn là ở cd $HADOOP\_HOME)**

$ hdfs namenode -format

**Test HDFS configuration (/usr/local/hadoop/sbin/)**

**$ cd /usr/local/hadoop/sbin**

**$ ./start-dfs.sh**

**$ ./start-yarn.sh**

**$ ./start-all.sh**

**Check the availability of all the nodes by typing**

**$ jps**

12293 Jps

9877 NameNode

10085 DataNode

10953 NodeManager

10590 ResourceManager

10335 SecondaryNameNode

**Step 9: Access the Web portal for hadoop management by typing in the following IP address in the browser**

<http://localhost:9870>

**Step 10: Check the hadoop cluster overview at**

http://localhost:8088

Execute $HADOOP\_HOME/sbin - ./stop-all.sh

**2. Install and Setup Apache Hive 3.1.3**

[How to Install Apache Hive on Ubuntu {Step-by-Step Guide} (phoenixnap.com)](https://phoenixnap.com/kb/install-hive-on-ubuntu)

<https://downloads.apache.org/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz>

[Hướng dẫn cài đặt Hive (viblo.asia)](https://viblo.asia/p/huong-dan-cai-dat-hive-vyDZOw8OZwj)

*1. Install Apache Hive From Offical Hive Website*

loclt@loclt-bigdata-master:~$ wget <https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz>

*2. Unzip file*

$ tar xzf apache-hive-3.1.2-bin.tar.gz

$ sudo mv apache-hive-3.1.2-bin **/usr/local/hive**

*3. nano ~/.bashrc* (Từ loclt user)

export HADOOP\_HOME=/usr/local/hadoop

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

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

export HADOOP\_MAPRED\_HOME=${HADOOP\_HOME}

export HADOOP\_COMMON\_HOME=${HADOOP\_HOME}

export HADOOP\_HDFS\_HOME=${HADOOP\_HOME}

export YARN\_HOME=${HADOOP\_HOME}

export HIVE\_HOME=/usr/local/hive

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

export CLASSPATH=$CLASSPATH:$HADOOP\_HOME/lib/\*:$HIVE\_HOME/lib/\*

**cd $HIVE\_HOME/bin/** (Vẫn ở loclt user)

*4. nano hive-config.sh*

export HIVE\_CONF\_DIR=/usr/local/hive/conf

export HADOOP\_HOME=/usr/local/hadoop

*5. Edit Hive Configurations with configure hive-site.xml (cd $HIVE\_HOME/conf) (Vẫn từ loclt user)*

sudo cp hive-default.xml.template hive-site.xml

sudo gedit hive-site.xml

- Replace all occurrences of ${system:java.io.tmpdir} to /tmp/hive => This is the location Hive stores all it’s temporary files

- Replace all occurrences of ${system:user.name} to username (loclt), the username should be the one you log in with

=> After replace above two properties, you should have something like below for the properties you updated (sau khi chỉnh sửa thì các đoạn sau trong file hive-site.xml sẽ trở thành như này):

<property>

<name>hive.exec.local.scratchdir</name>

<value>/tmp/hive/loclt </value>

<description>Local scratch space for Hive jobs</description>

</property>

<property>

<name>hive.downloaded.resources.dir</name>

<value>/tmp/hive\_io/${hive.session.id}\_resources</value>

<description>Temporary local directory for added resources in the remote file system.</description>

</property>

<property>

<name>hive.querylog.location</name>

<value>/tmp/hive/loclt </value>

<description>Location of Hive run time structured log file</description>

</property>

<property>

<name>hive.server2.logging.operation.log.location</name>

<value>/tmp/hive/loclt /operation\_logs</value>

<description>Top level directory where operation logs are stored if logging functionality is enabled</description>

</property>

<property>

<name>hive.metastore.warehouse.dir</name>

<value>/user/hive/warehouse</value>

<description>location of default database for the warehouse</description>

</property>

<property>

<name>javax.jdo.option.ConnectionURL</name>

<value>jdbc:derby:;databaseName=metastore\_db;create=true</value>

<description>

JDBC connect string for a JDBC metastore.

To use SSL to encrypt/authenticate the connection, provide database-specific SSL flag in the connection URL.

For example, jdbc:postgresql://myhost/db?ssl=true for postgres database.

</description>

</property>

NOTE: dùng hàm getdit để in ra toàn bộ nội dung của 1 file ra text để dễ chỉnh sửa và thấy được số ở dòng (ở dòng 3215 thì có mấy ký tự bị lỗi, xóa đi là được)

6. Create Hive Warehouse Directories

cd $HADOOP\_HOME/sbin (In hadoop user)

**- Create tmp Directory**

hdfs dfs -mkdir -p /tmp

hdfs dfs -chmod g+w /tmp

hdfs dfs -ls /

**- Create warehouse Directory**

hdfs dfs -mkdir -p /user/hive/warehouse

hdfs dfs -chmod g+w /user/hive/warehouse

hdfs dfs -ls /user/hive

7. Create Hive Metastore Derby Database

- Cấp quyền đọc ghi để tạo folder metastore\_db trong /usr/local/hive/bin

$ sudo chmod a+rwx /usr/local/hive/bin/

$ schematool -initSchema -dbType derby (khởi tạo derby – một framework của apache để làm metastore database cho Hive)

8. Cuối dùng

cd $HIVE\_HOME/bin (Ở hadoop user)

$ hive –version

$ hive

9. Download and setup Derby Database for Hive Metastore

<https://www.youtube.com/watch?v=1PgCOZ2837g&ab_channel=JoeyBlue>

**3. Install Spark and Pyspark Single Node**

[How to Install Spark on Ubuntu {Instructional guide} (phoenixnap.com)](https://phoenixnap.com/kb/install-spark-on-ubuntu)

[(17) How to Setup / Install an Apache Spark 3.1.1 Cluster on Ubuntu | LinkedIn](https://www.linkedin.com/pulse/how-setup-install-apache-spark-311-cluster-ubuntu-shrivastava/)

[(1522) Install Apache Spark and PySpark on Ubuntu 20.04 Linux Debian, Python 3.7 - Part 1a - YouTube](https://www.youtube.com/watch?v=7tDOUrl7Aoc&ab_channel=datyrlab)

*1. Install Scala*

$ sudo apt-get install scala -y

$ scala -version

*2. Install Spark and unzip then move that folder to /usr/local/spark*

$ wget <https://dlcdn.apache.org/spark/spark-3.4.1/spark-3.4.1-bin-hadoop3.tgz>

$ tar -xvzf spark-3.4.1-bin-hadoop3.tgz

$ sudo mv spark-3.4.1-bin-hadoop3 **/usr/local/spark**

*3. Setup the enviroment for Spark*

$ nano ~/.bashrc

Add these lines:

export SPARK\_HOME=**/usr/local/spark**

export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

export PYSPARK\_HOME=/usr/bin/python3.8

export SPARK\_DRIVER\_PYTHON=/usr/bin/python3.8

*4. Install Pip*

$ sudo apt install python3-pip

*5. Install Pyspark*

sudo pip3 install pyspark

6. Run Spark and Pyspark

Check validate Pyspark is working via the Spark-shell

$ cd /usr/local/spark

$ ls -al

$ cd bin

$ ls -al

$ spark-shell

$ pyspark

Access to SparkUI

7. Setup Spark Warehouse Location (IMPORTANT)

<https://medium.com/@madtopcoder/use-hadoop-hdfs-as-spark-warehouse-directory-without-installing-hive-4ea3af86dfae>