Hadoop Linux - Pseudo Distribution

Commands as following:

1. Install Java
   ```
   sudo apt-get update
   sudo apt-get install default-jdk
   java -version
   ```

2. Install ssh if doesn’t exist
   ```
   sudo apt-get install ssh
   which ssh
   ```

3. Generate rsa key pair
   ```
   ssh-keygen -t rsa -P ""
   ```

4. Copy rsa Public key in authorized key file
   ```
   cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
   ```

5. Test if your ssh is working fine
   ```
   ssh localhost
   ```

6. Download Hadoop binaries and extract it
   ```
   wget http://apache.mirror.rafal.ca/hadoop/common/hadoop-2.6.5/hadoop-2.6.5.tar.gz
   tar -xvzf hadoop-2.6.5.tar.gz
   ```

7. Move it to User Local folder (You can keep it anywhere)
   ```
   sudo mv hadoop-2.6.5 /usr/local/hadoop
   ```

8. Change the ownership of hadoop folder if need be
   ```
   sudo chown -R <ID:GROUP> /usr/local/hadoop
   ```

9. Modify .bashrc file and add following variables in it
   ```
   #HADOOP VARIABLES START
   export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
   export HADOOP_INSTALL=/usr/local/hadoop
   export PATH=$PATH:$HADOOP_INSTALL/bin
   export PATH=$PATH:$HADOOP_INSTALL/sbin
   export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
   export HADOOP_COMMON_HOME=$HADOOP_INSTALL
   export HADOOP_HDFS_HOME=$HADOOP_INSTALL
   export YARN_HOME=$HADOOP_INSTALL
   export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
   ```

```
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib/native"
export HADOOP_CONF_DIR=$HADOOP_INSTALL/etc/hadoop
#HADOOP VARIABLES END
```

10. Apply .bashrc file

```
source ~/.bashrc
```

11. Create directory structure for hadoop

```
sudo mkdir -p /app/hadoop/tmp
sudo chown <USER:GROUP> /app/hadoop/tmp

sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode
sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode
sudo chown -R <USER:GROUP> /usr/local/hadoop_store
```

12. Modify certain files specific to hadoop configurations (All files are at location **/usr/local/hadoop/etc/hadoop/**):

    a. hadoop-env.sh

    ```
    export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
    ```

    b. core-site.xml

    ```
    <configuration>
     <property>
                <name>hadoop.tmp.dir</name>
                <value>/app/hadoop/tmp</value>
                <description>A base for other temporary directories.</description>
     </property>

     <property>
                <name>fs.default.name</name>
                <value>hdfs://localhost:54310</value>
                <description>The name of the default file system.  A URI whose
                scheme and authority determine the FileSystem implementation.  The
                uri's scheme determines the config property (fs.SCHEME.impl) naming
                the FileSystem implementation class.  The uri's authority is used to
                determine the host, port, etc. for a filesystem.</description>
     </property>
    </configuration>
    ```

    c. hdfs-site.xml

    ```
    <configuration>
            <property>
                <name>dfs.replication</name>
                <value>1</value>
    ```

```
  <description>Default block replication.
  The actual number of replications can be specified when the file is
created.
  The default is used if replication is not specified in create time.
</description>
 </property>

 <property>
   <name>dfs.namenode.name.dir</name>
   <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
 </property>

 <property>
   <name>dfs.datanode.data.dir</name>
   <value>file:/usr/local/hadoop_store/hdfs/datanode</value>
 </property>
</configuration>
```

d. yarn-site.xml

```
<configuration>
        <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
        </property>
        <property>
        <name>yarn.nodemanager.aux-
        services.mapreduce.shuffle.class</name>
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>
        </property>
</configuration>
```

13. Format namenode

```
hadoop namenode -format
```

14. Start all services

```
cd /usr/local/hadoop/sbin
start-dfs.sh
start-yarn.sh
```

15. Check if all services are up and running

```
jps
```

Hadoop Windows - Pseudo Distribution

1. Extract hadoop-2.6.2 file on your local system where you want to install Hadoop
2. Copy java JDK in C: drive if already not there
3. Go to location **hadoop-2.6.2/etc/hadoop**
    a. Modify core-site.xml file to update hadoop.tmp.dir property
       Create temp directory somewhere and use that as value to mentioned property
    b. Modify Hadoop-env.cmd to update JAVA_HOME
       Should be the location of your JDK in C: directory
    c. Hdfs-site.xml:
       Create directories /hadoop-2.6.2/data/namenode and /hadoop-2.6.2/data/datanode
       Update dfs.namenode.name.dir and dfs.datanode.name.dir properties accordingly
    d. yarn-site.xml
       Update yarn.nodemanager.log-dirs property
4. set HADOOP_HOME environment variable to bin directory within hadoop-2.6.2 folder
5. open windows command prompt cmd and execute following commands
    a. hdfs namenode –format
    b. change directory to hadoop-2.6.2/sbin
    c. start-dfs.cmd
    d. start-yarn.cmd

Cloudera Quick Start VM:

https://www.cloudera.com/downloads/quickstart_vms/5-12.html

Select VMWare as platform

Reference

Download Ubuntu 16: https://www.ubuntu.com/download/desktop

Installing Hadoop Step by Step:

http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_16_04_single_node_cluster.php

Note: Above link is missing YARN configuration

Refer following link for YARN configuration

https://www.ibm.com/developerworks/community/blogs/d9a07ec3-11e2-467d-b758-6861c4cb1d44/entry/How_to_install_Hadoop_2_7_0_in_ubuntu_16_04?lang=en