How to setup Hadoop

1. Install packages (one all nodes)
    a. sudo apt-get update
    b. sudo apt-get install openjdk-11-jdk-headless
    c. sudo apt-get install net-tools

2. Download/setup hadoop (one all nodes)
    a. Download
        i. sudo wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
    b. Unpack
        i. tar -zxvf hadoop-3.3.4.tar.gz
    c. Set the JAVA-HOME (on master only)
        i. sudo nano .bashrc (add at the very end)
            1. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
            2. export PATH=$JAVA_HOME/bin:$PATH
        ii. source .bashrc

3. SSH setup (on all nodes)
    a. On all nodes:
        i. ssh-keygen -t rsa
            1. SSH file is created at .ssh/id_rsa.pub
    b. Display key
        i. cat .ssh/id_rsa.pub
            1. Copy all keys to a document
    c. Go to authorized_keys on all nodes and paste the keys
        i. nano .ssh/authorized_keys

4. Aliases
    a. sudo nano /etc/hosts
        i. 192.168.2.129 MasterNode
        ii. 192.168.2.33 WorkerNode
        iii. 192.168.2.121 WorkerNode2

5. Configure environment (on all nodes)
    a. sudo nano hadoop-3.3.4/etc/hadoop/hadoop-env.sh
    b. sudo nano hadoop-3.3.4/etc/hadoop/mapred-env.sh
    c. sudo nano hadoop-3.3.4/etc/hadoop/yarn-env.sh
        i. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
        ii. export HDFS_DATANODE_USER=ubuntu
        iii. export HDFS_NAMENODE_USER=ubuntu
        iv. export HDFS_SECONDARYNAMENODE_USER=ubuntu

v. export YARN_RESOURCEMANAGER_USER=ubuntu
vi. export YARN_NODEMANAGER_USER=ubuntu

6. Configuring hadoop (on all nodes)
    a. core-site.xml (ip entered here becomes master)
        i. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/core-site.xml
            1. <configuration>
            2. <property>
            3. <name>fs.default.name</name>
            4. <value>hdfs://MasterNode:50000</value>
            5. </property>
            6. </configuration>
    b. hdfs-site.xml
        i. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/hdfs-site.xml
            1. <configuration>
            2. <property>
            3. <name>dfs.namenode.name.dir</name>
            4. <value>/home/ubuntu/hadoop2-dir/namenode-dir</value>
            5. </property>
            6.
            7. <property>
            8. <name>dfs.datanode.data.dir</name>
            9. <value>/home/ubuntu/hadoop2-dir/datanode-dir</value>
            10. </property>
            11.
            12. <property>
            13. <name>dfs.replication</name>
            14. <value>2</value>
            15. </property>
            16.
            17. <property>
            18. <name>dfs.permission</name>
            19. <value>false</value>
            20. </property>
            21. </configuration>
    c. yarn-site.xml
        i. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/yarn-site.xml
            1. <configuration>
            2. <property>
            3. <name>yarn.nodemanager.aux-services</name>
               <value>mapreduce_shuffle</value>
            4. </property>
            5. <property>

6. <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
7. <value>org.apache.hadoop.mapred.ShuffleHandler</value>
8. </property>
9. <property>
10. <description>The hostname of the RM.</description>
11. <name>yarn.resourcemanager.hostname</name>
12. <value>MasterNode</value>
13. </property>
14. <property>
15. <description>The address of the applications manager interface in the RM.</description>
16. <name>yarn.resourcemanager.address</name>
17. <value>MasterNode:8032</value>
18. </property>
19. </configuration>

d. mapred-site.xml
   i. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/mapred-site.xml
      1. <configuration>
      2. <property>
      3. <name>mapreduce.framework.name</name>
      4. <value>yarn</value>
      5. </property>
      6. </configuration>

7. Configure DataNodes (on all nodes)
   a. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/workers
      i. WorkerNode
      ii. WorkerNode2
   b. sudo nano /home/ubuntu/hadoop-3.3.4/etc/hadoop/slaves
      i. WorkerNode
      ii. WorkerNode2

8. Format namenode (only on master node)
   a. hadoop-3.3.4/bin/hadoop namenode -format

9. Start/stop HDFS (on master node)
   a. sudo hadoop-3.3.4/sbin/start-all.sh
   b. sudo hadoop-3.3.4/sbin/stop-all.sh

10. Troubleshooting
    a. only one datanode is started
       i. Remove each node data directory

ii. Reformat namenode

11. Working with the HDFS
    a. Put a file
        i. hadoop-3.3.4/bin/hadoop fs -put /path/in/linux /hdfs/path
    b. Check HDFS files
        i. hadoop-3.3.4/bin/hadoop fs -ls /
        ii. hadoop-3.3.4/bin/hadoop fs -ls -R / | less
    c. Monitor active nodes
        i. hadoop-3.3.4/bin/hdfs dfsadmin -report

12. download dataset
    a. wget http://labrosa.ee.columbia.edu/~dpwe/tmp/millionsongsubset.tar.gz
    b. tar -zxvf millionsongsubset.tar.gz
    c. hadoop-3.3.4/bin/hadoop fs -mkdir /hdfs
    d. hadoop-3.3.4/bin/hadoop fs -mkdir /hdfs/data
    e. hadoop-3.3.4/bin/hadoop fs -put /home/ubuntu/MillionSongSubset /hdfs/data/
    f. hadoop-3.3.4/bin/hadoop fs -put /home/ubuntu/europarl-v7.de-en.en /hdfs/data/
    g. hadoop-3.3.4/bin/hadoop fs -put /home/ubuntu/SongCSV_A.csv /hdfs/data/
    h.