# Multi-node Hadoop Cluster Setup steps

- step 1: Identify machines on which Hadoop cluster is to be built.
  - Recommended 1 master and 3 worker machines.
  - All these machines must be in same network and should be able to communicate (ping) with each other.
- step 2: On all machines -- Create new admin user "hduser".
  - You can do from GUI or using following commands.
  - terminal> sudo adduser hduser
  - terminal> sudo usermod -aG sudo hduser
- step 3: On all machines -- Login with "hduser" and install necessary softwares.
  - terminal> sudo apt update
  - terminal> sudo apt install vim ssh net-tools openjdk-8-jdk git
  - terminal> cd ~
  - terminal> wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.2/hadoop-3.3.2.tar.gz
  - terminal> tar xvf hadoop-3.3.2.tar.gz
- step 4: On respective machines -- change hostnames.
  - master terminal> sudo hostnamectl set-hostname master
  - worker1 terminal> sudo hostnamectl set-hostname worker1
  - worker2 terminal> sudo hostnamectl set-hostname worker2
  - worker3 terminal> sudo hostnamectl set-hostname worker3
- step 5: On all machines -- Modify /etc/hosts to enter IP address.
  - Assuming IP addresses -- master = 172.18.1.50, worker1 = 172.18.1.51, worker2 = 172.18.1.52, worker3 = 172.18.1.53 -- add following lines into /etc/hosts.
  - terminal> sudo vim /etc/hosts

```
172.18.1.50   master
172.18.1.51   worker1
172.18.1.52   worker2
172.18.1.53   worker3
```

- step 6: On master -- generate ssh id and copy to all nodes.
  - master terminal> ssh-keygen -t rsa -P ""
  - master terminal> ssh-copy-id hduser@master
  - master terminal> ssh-copy-id hduser@worker1
  - master terminal> ssh-copy-id hduser@worker2
  - master terminal> ssh-copy-id hduser@worker3
- step 7: On all machines -- Configure Hadoop.
  - terminal> vim ~/.bashrc

    ```
    export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
    export HADOOP_HOME=$HOME/hadoop-3.3.2
    export PATH=$HADOOP_HOME/sbin:$HADOOP_HOME/bin:$PATH
    ```

  - terminal> source ~/.bashrc
  - terminal> git clone https://github.com/nilesh-g/hadoop-cluster-install.git
  - master terminal> cp hadoop-cluster-install/master/* $HADOOP_HOME/etc/hadoop/
  - worker terminal> cp hadoop-cluster-install/worker/* $HADOOP_HOME/etc/hadoop/
- step 8: On master -- Examine the master configration.
  - terminal> cd $HADOOP_HOME/etc/hadoop/
  - terminal> vim hadoop-env.sh
  - terminal> vim core-site.sh
  - terminal> vim hdfs-site.xml
  - terminal> vim mapred-site.xml
  - terminal> vim yarn-site.xml
  - terminal> vim workers
- step 9: On workers -- Examine the workers configration.
  - terminal> cd $HADOOP_HOME/etc/hadoop/
  - terminal> vim hadoop-env.sh
  - terminal> vim core-site.sh
  - terminal> vim hdfs-site.xml

  - terminal> vim mapred-site.xml
  - terminal> vim yarn-site.xml
- step 10: Start Hadoop and verify.
  - master terminal> hdfs namenode -format
  - master terminal> start-dfs.sh
  - master terminal> start-yarn.sh
  - master terminal> jps
  - worker1 terminal> jps
  - worker2 terminal> jps
  - worker3 terminal> jps
- step 11: On master -- Upload a file on HDFS to verify if cluster and replication is working properly.
  - master terminal> hadoop fs -mkdir /user
  - master terminal> hadoop fs -mkdir /user/hduser
  - master terminal> hadoop fs -ls /user
  - master terminal> echo "Welcome to Hadoop cluster" > hello.txt
  - master terminal> hadoop fs -put hello.txt /user/hduser
  - master terminal> hadoop fs -head /user/hduser/hello.txt
  - Now you may contain multiple such files from ANY nodes in the cluster and upload them. Also you can see/download the files.
- step 12: To monitor Hadoop cluster from any node in the cluster.
  - Browser: http://master:9870
- step 13: On master -- Stop the Hadoop and verify.
  - master terminal> stop-yarn.sh
  - master terminal> stop-dfs.sh
- step 14: Shutdown all VMs.