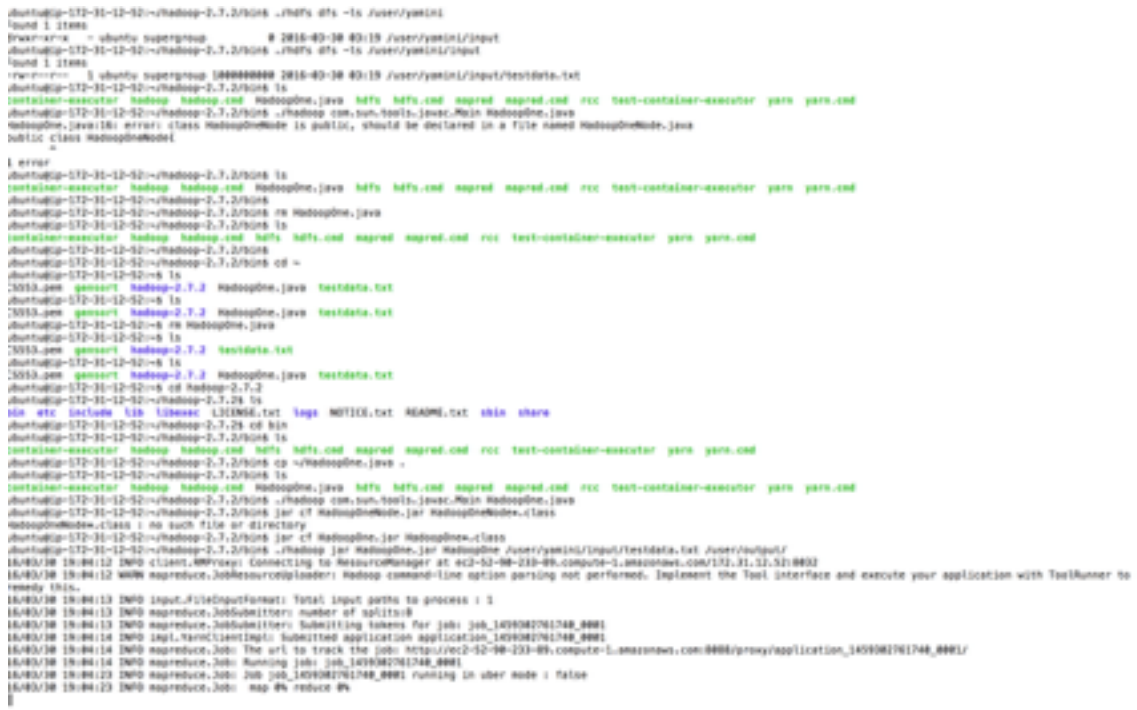
CS553 Programming Assignment 2

Hadoop:(1 node-1 GB data)

1.Hadoop has been installed on 1 node(c3.large instance).The following config files have been modified:
conf/core-site.xml
conf/hdfs-site.xml
conf/mapred-site.xml
conf/yarnsite.xml
conf/slaves
conf/hadoop-env.sh

The sort program on 1GB of data execution is shown below:



Hadoop cluster:(Screenshots attached)

Hadoop:

1) What is a Master node? What is a Slaves node?
Master node manages the process of mission partitioning and task delegation. Slave nodes are workers that execute the tasks that are assigned to them by the master node.

2) Why do we need to set unique available ports to those configuration files on a shared environment?
What errors or side-effects will show if we use same port number for each user?
Different ports have different functionalities (example - HDFS on port 1, 9002 and Master node on port 0, 9001). Consequently, port collision becomes a possibility if the same port number is used for each user.

3) How can we change the number of mappers and reducers from the configuration file?
The number of map tasks can be increased manually using the jobConf's conf.setNumMapTasks(int num). It must be noted, however, that this will not set the number below what Hadoop determines by splitting the input data. Similarly, the number of reduce tasks may be increased using the JobConf's con.setNumReduceTasks(int num).

Steps to launch a Spark cluster and run the sort program on it:

A 17 node spark cluster has been setup and the sort program was run on it with a 100 GB dates generated through gensort .

Versions used:

1.OS used-ubuntu
2.Java version-1.7
3.Spark-1.6.0-bin-hadoop2.6
4.Code is written in python(2.7)
5.Instance type-c3 large

Steps :

1.Login to instance and install java
2.Download and unzip spark
3.Install scala
4.modify .bashrc
5.Export Access key and secret access key downloaded from AWS
6.In the spark ec2 folder,run the command to launch a master with 16 slaves (ebs volume 400 added) using spot instances.(mentioned the instance type in the command)
7.1 master and 16 slave instances get created.
8.ssh into the master and created a folder name ‘knn' in it
9.Uploaded the code and data files in inn
10.Using the command below uploaded the code to all nodes in the cluster
    . /spark-ec2/copy-dir knn
11.Upload the data file in hfs
12.Run the spark submit command in /spark/bin/folder
13.This initiates the executors on all nodes and runs the tasks(screenshots) provided
14.Output captured in ‘output’ folder
15.Validate the output.(First part and last part screenshots provided).