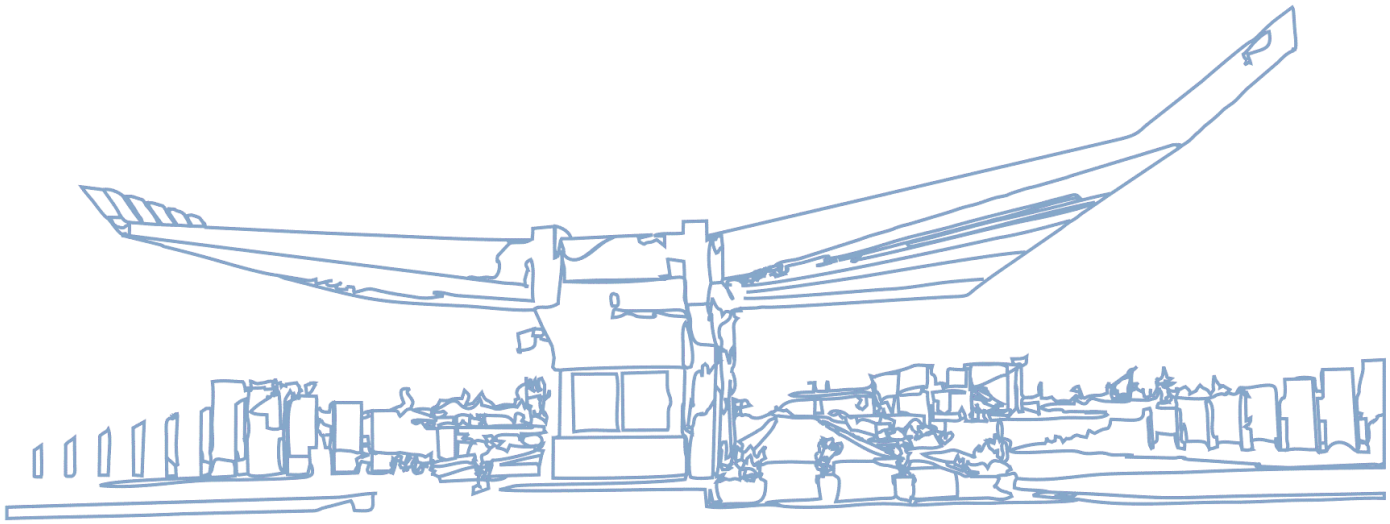
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Assignment 01



PREPARED:
Baftjar TABAKU

12.04.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

Assignment tasks and notes

1. Install VirtualBox and Setup Cloudera QuickStart VM (50)
2. Add Ubuntu in VirtualBox and Setup a One Node Hadoop Cluster (50)
3. (optional/extra credit): Setup a Four Node Hadoop Cluster using AWS (20)

For each of the above, submit a word document that includes a list of commands accompanied by screenshots. Each document must look like a tutorial.

Hadoop and Cloudera

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking. Hadoop's ability to process and store different types of data makes it a particularly good fit for big data environments. They typically involve not only large amounts of data, but also a mix of structured transaction data and semi structured and unstructured information, such as internet clickstream records, web server and mobile application logs, social media posts, customer emails and sensor data from the internet of things.

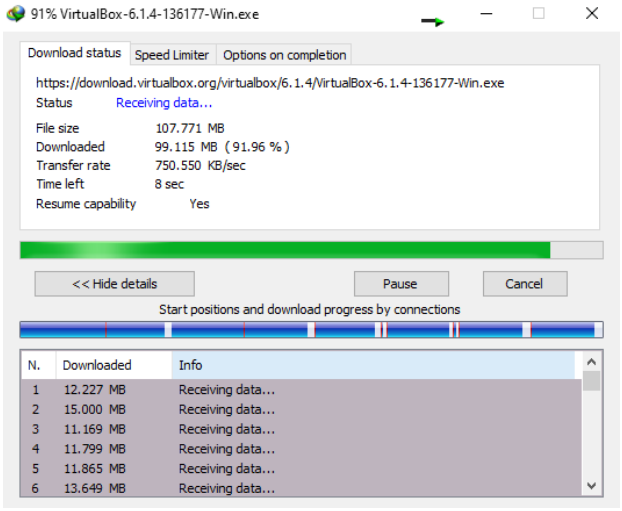
Ref: https://searchdatamanagement.techtarget.com

The Cloudera delivers the modern platform for machine learning and advanced analytics built on the latest open source technologies. The world’s leading organizations trust Cloudera to help solve their most challenging business problems by efficiently capturing, storing, processing and analyzing vast amounts of data.

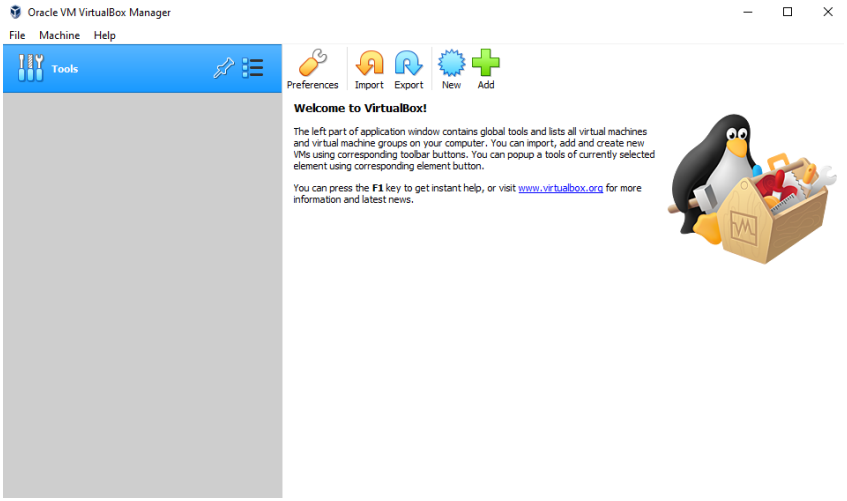
Ref: https://cloudian.com

Installing VirtualBox and Setting up Cloudera QuickStart VM (50)

I downloaded VirtualBox and set up, as the screenshots shows,

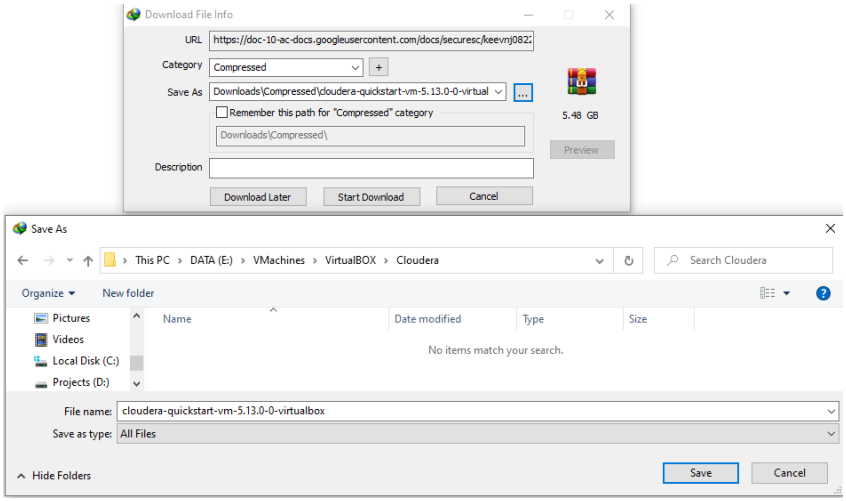


After installing



Next step was to download the Cloudera QuickStart VM, provided in the

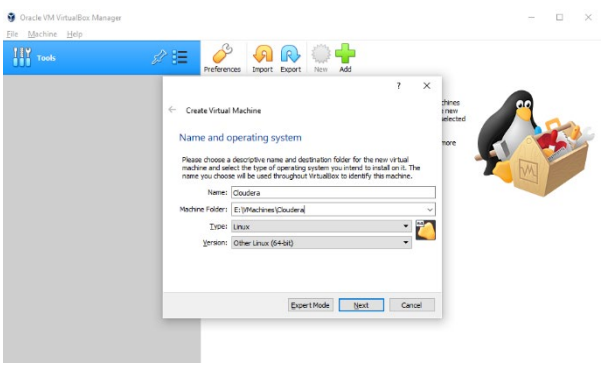
link: https://drive.google.com/file/d/1IK-ZfiKfKaY8LrU0oDk--HH4ijgVpFrB/view

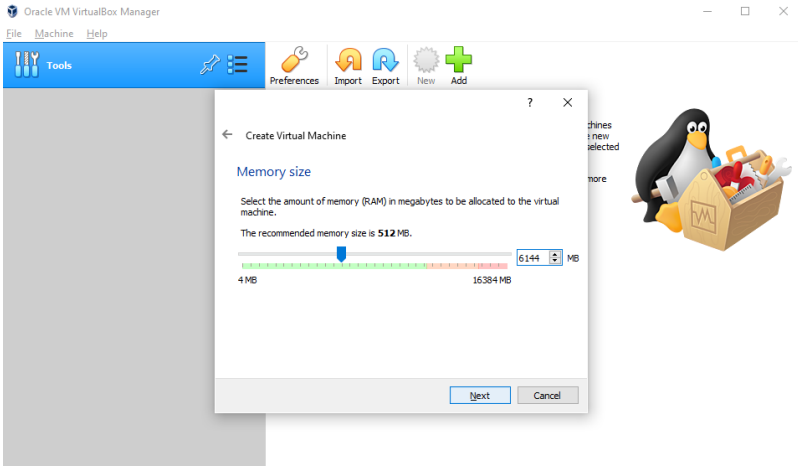


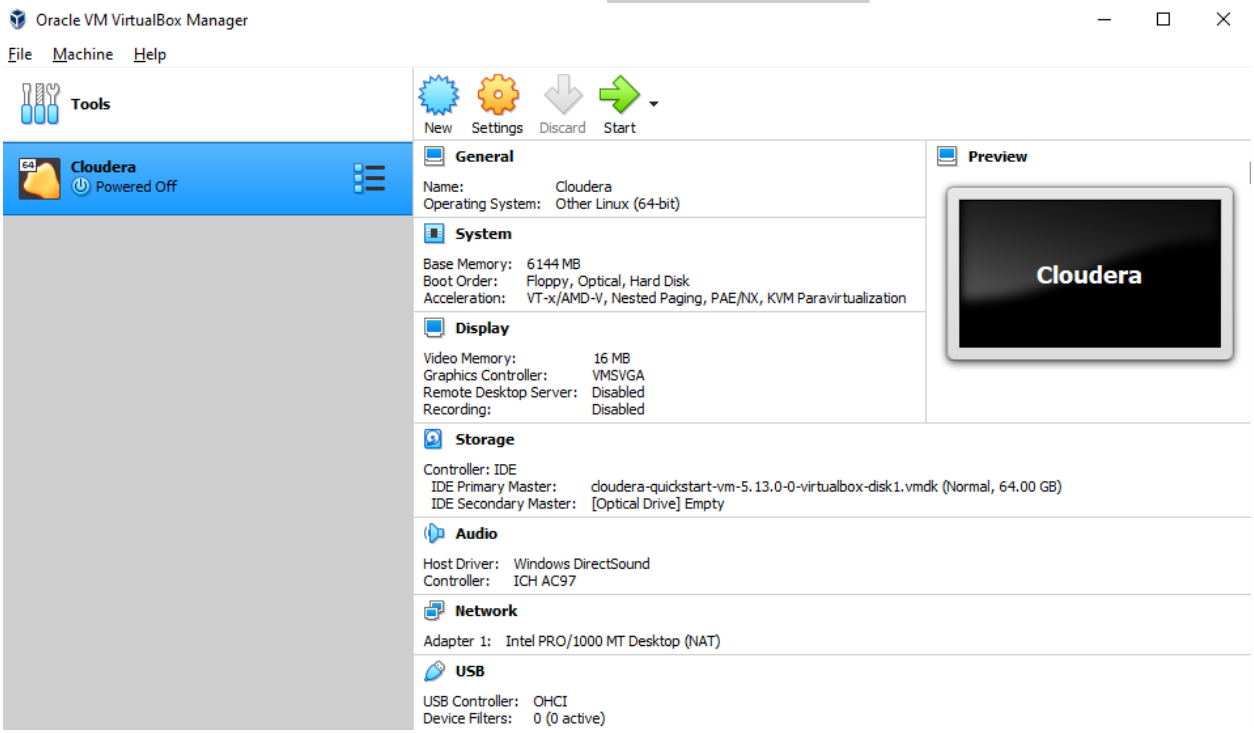
That took approximately 1h 15min.

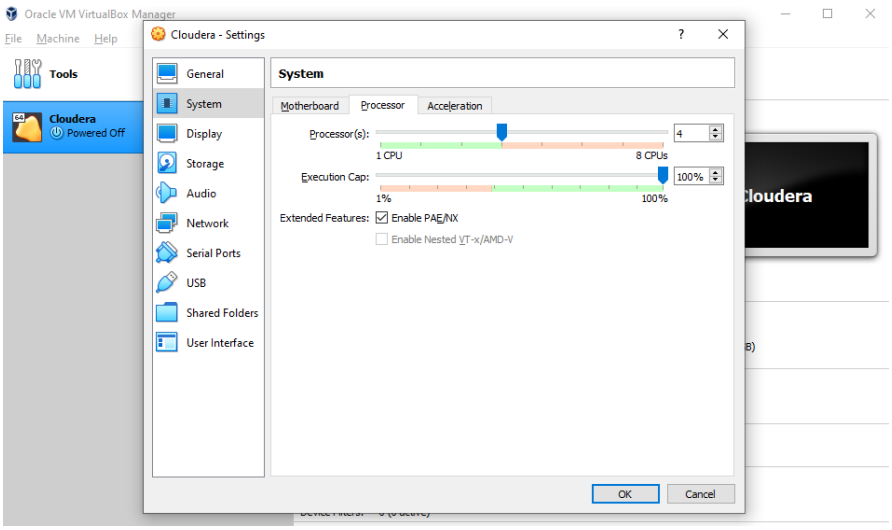
After downloading it, In set it up on VirtualBox.

Setting up the Cloudera QuickStart VM process is shown in the following screenshots,

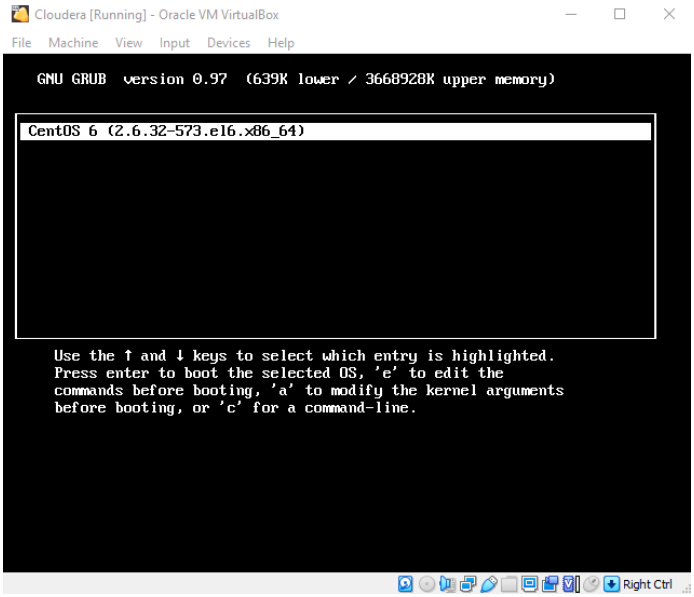




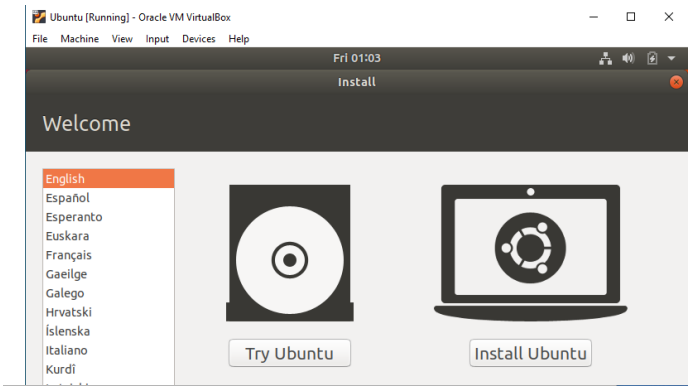


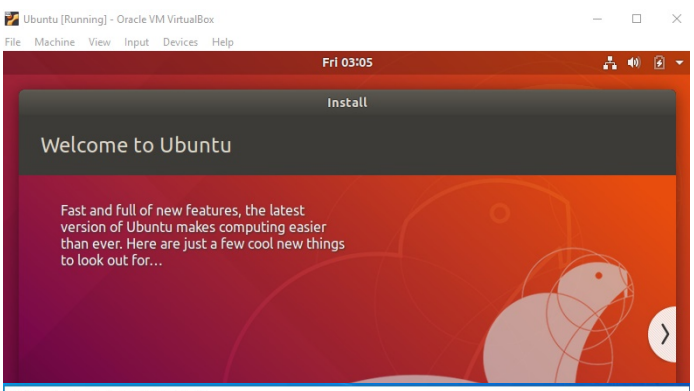


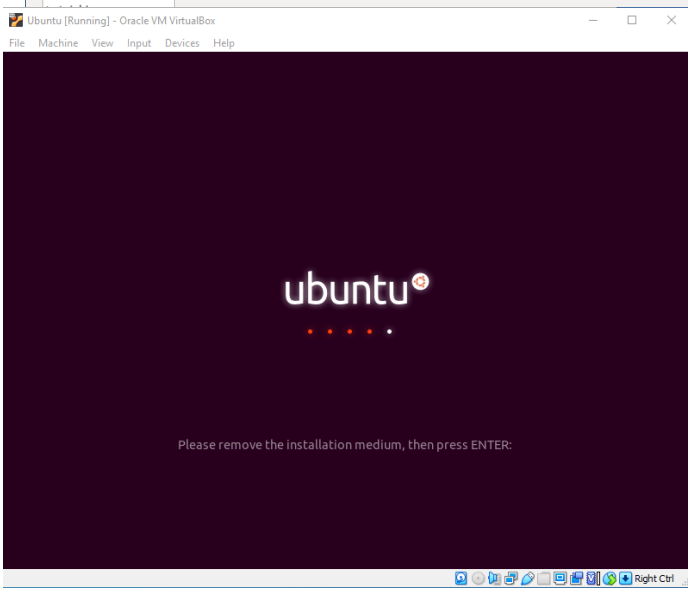
The Cloudera successfully set up on VirtualBox.

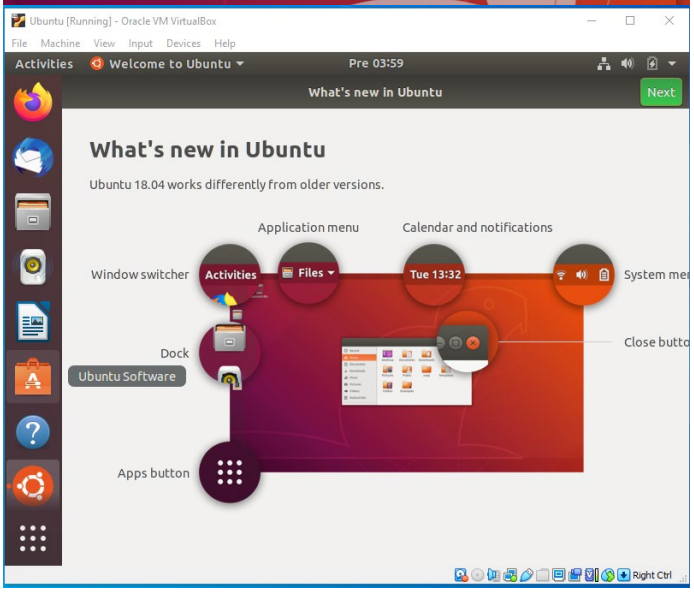


Adding Ubuntu in VirtualBox and setting up a One Node Hadoop Cluster.



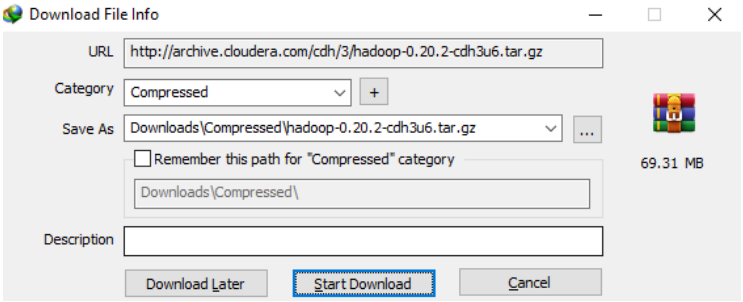




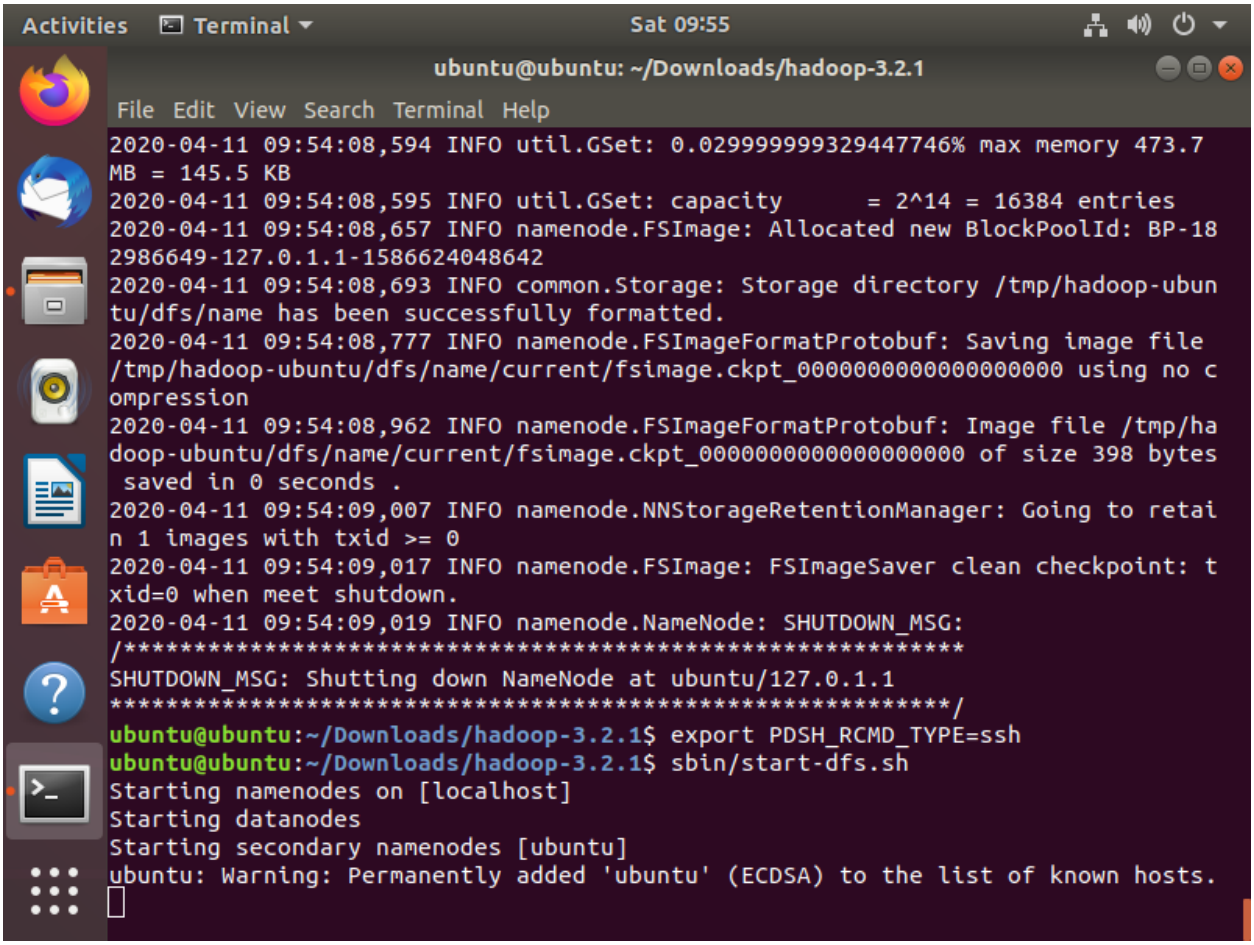


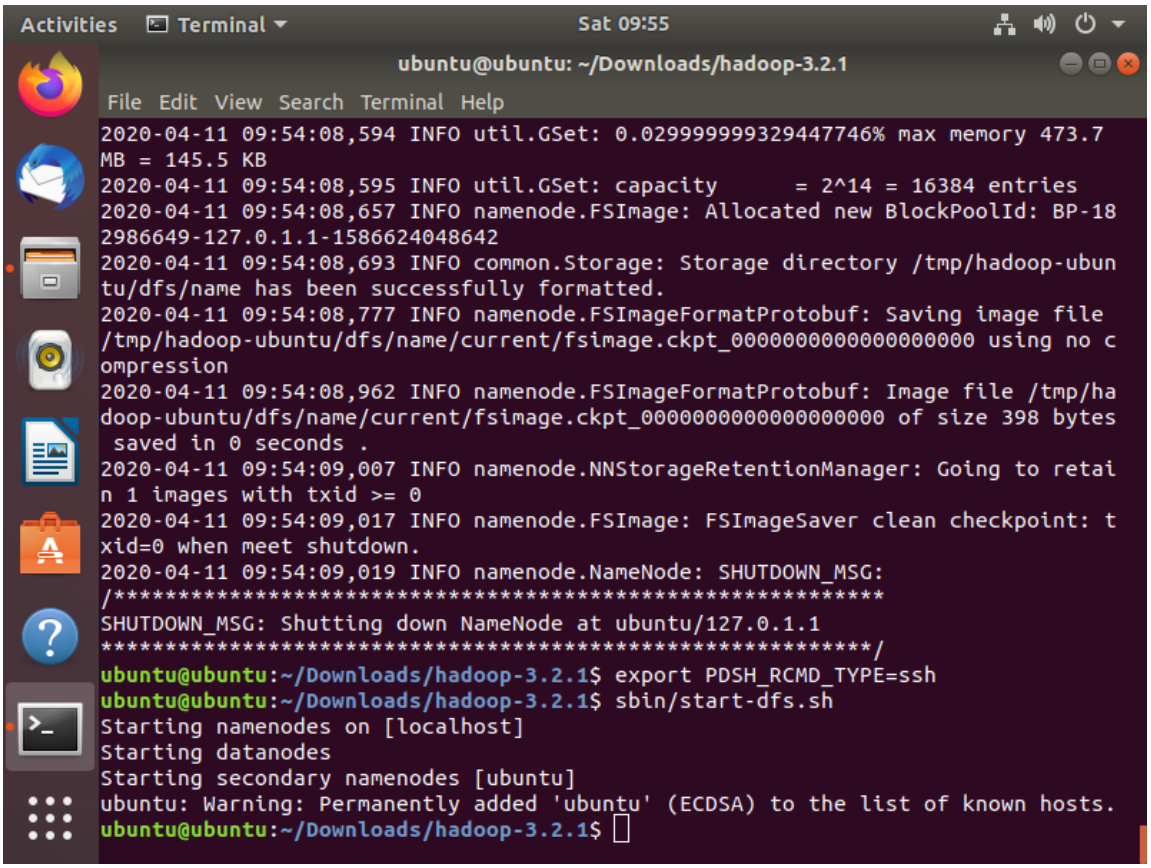
Installing ubuntu in VirtualBox

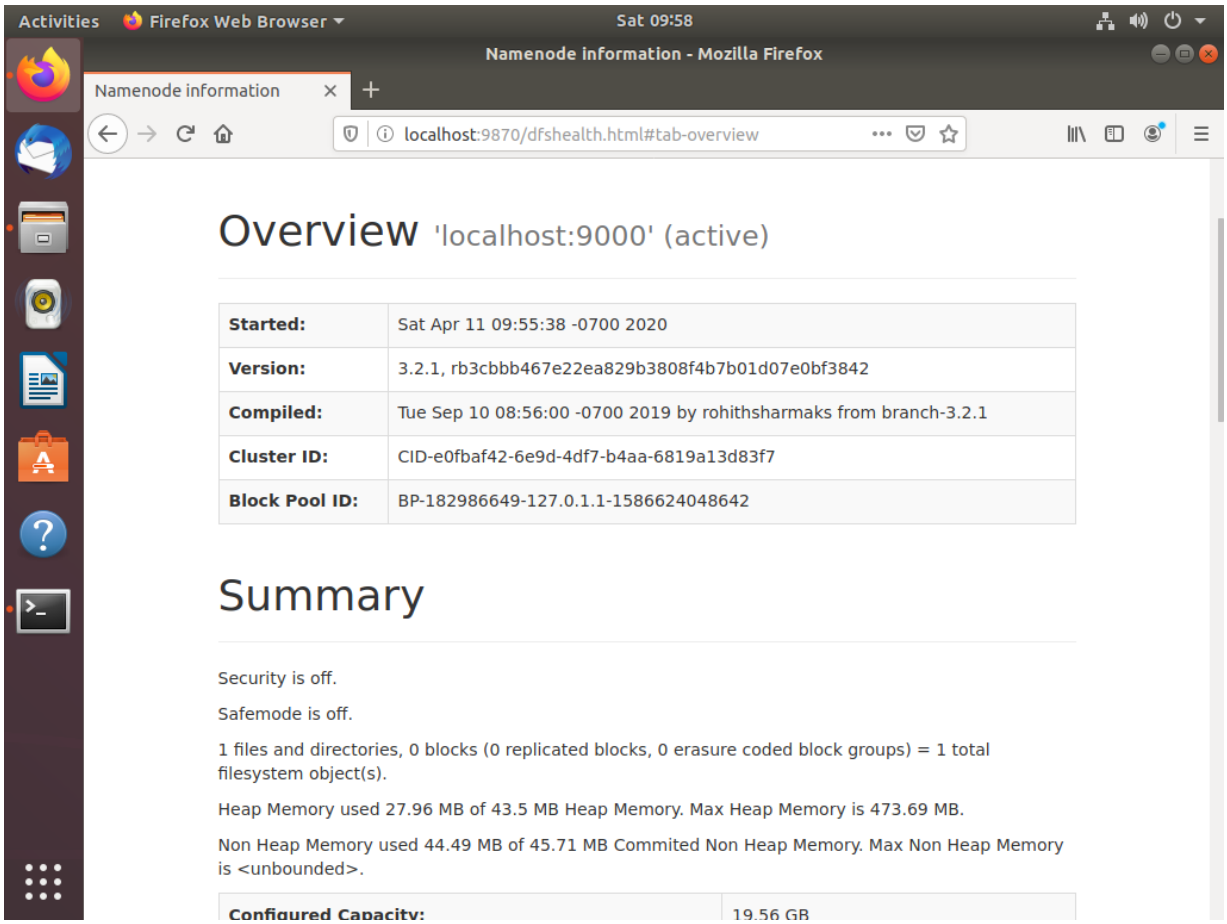
According to the provided manuals, I Installed the Hadoop as following screenshot shows.

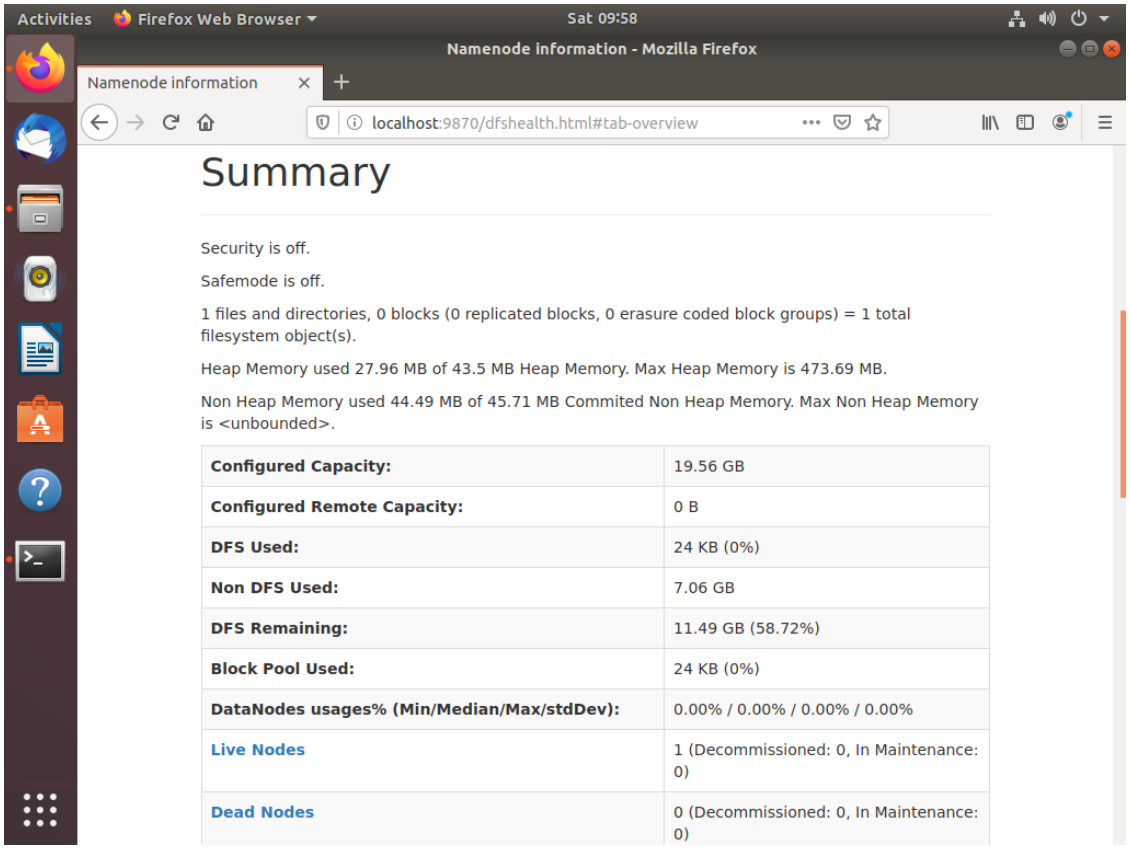


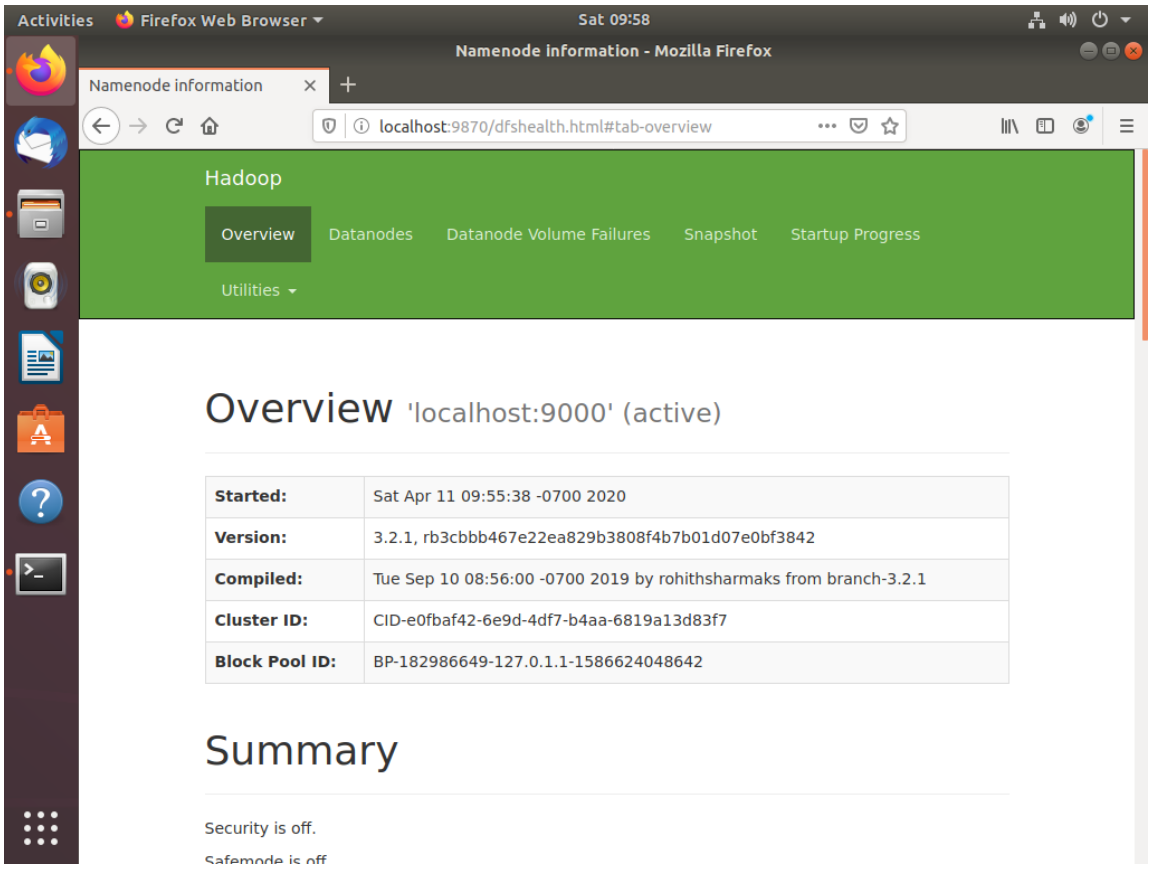
Then the Hadoop setup, downloaded from the cloudera archive, for more resource efficiency I added another kind of virtual machine like VMware, a light-weight version to use it with ubuntu installed and setting up Hadoop cluster single node.




Activities
Terminal
Sat 09:55
ubuntu@ubuntu: ~/Downloads/hadoop-3.2.1
File Edit View Search Terminal Help
2020-04-11 09:54:08,594 INFO util.GSet: 0.029999999329447746% max memory 473.7 MB = 145.5 KB
2020-04-11 09:54:08,595 INFO util.GSet: capacity = 2^14 = 16384 entries
2020-04-11 09:54:08,657 INFO namenode.FSImage: Allocated new BlockPoolId: BP-182986649-127.0.1.1-1586624048642
2020-04-11 09:54:08,693 INFO common.Storage: Storage directory /tmp/hadoop-ubuntu/dfs/name has been successfully formatted.
2020-04-11 09:54:08,777 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-ubuntu/dfs/name/current/fsimage.ckpt_0000000000000000000 using no compression
2020-04-11 09:54:08,962 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-ubuntu/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 398 bytes saved in 0 seconds .
2020-04-11 09:54:09,007 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2020-04-11 09:54:09,017 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2020-04-11 09:54:09,019 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at ubuntu/127.0.1.1
**/
ubuntu@ubuntu:~/Downloads/hadoop-3.2.1$ export PDSH_RCMD_TYPE=ssh
ubuntu@ubuntu:~/Downloads/hadoop-3.2.1$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
ubuntu: Warning: Permanently added 'ubuntu' (ECDSA) to the list of known hosts.
ubuntu@ubuntu:~/Downloads/hadoop-3.2.1$


Activities
Firefox Web Browser
Sat 09:58
Namenode information - Mozilla Firefox
Namenode information
localhost:9870/dfshealth.html#tab-overview
Overview 'localhost:9000' (active)
Started: Sat Apr 11 09:55:38 -0700 2020
Version: 3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled: Tue Sep 10 08:56:00 -0700 2019 by rohithsharmaks from branch-3.2.1
Cluster ID: CID-e0fbaf42-6e9d-4df7-b4aa-6819a13d83f7
Block Pool ID: BP-182986649-127.0.1.1-1586624048642
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 27.96 MB of 43.5 MB Heap Memory. Max Heap Memory is 473.69 MB.
Non Heap Memory used 44.49 MB of 45.71 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 19.56 GB





Note: *Screenshots shows some fractions at the cloudera and Hadoop setup installation.*

Setting up a Four Node Hadoop Cluster using AWS.

The idea was clear on what to do, and I would do it as was shown in the 7 parts with videos, the problem was that at the moment I don't have a credit card to sign up and creating a AWS account.