EXPERIMENT NO:1

AIM: installation of Hadoop and experiment on HDFS commands.

Theory:

Cloudera is a company that provides a platform for big data processing and analytics, based on the Hadoop ecosystem. Hadoop is an open-source framework for distributed storage and processing of large datasets across clusters of computers. Hadoop Distributed File System (HDFS) is the primary storage system used in Hadoop. Here's a theoretical overview of Cloudera, Hadoop, and some common HDFS commands:

1. Cloudera: Cloudera is a software company that offers a comprehensive platform for big data management, analytics, and machine learning. Cloudera's platform is built on open-source Hadoop technologies and includes a range of tools and services to help organizations collect, store, process, and analyse vast amounts of data.
2. Hadoop: Hadoop is an open-source framework designed for distributed storage and processing of large datasets. It's widely used in various industries to handle big data challenges. The core components of Hadoop include HDFS (Hadoop Distributed File System) and the MapReduce processing engine. Hadoop is known for its scalability, fault tolerance, and flexibility.
3. Hadoop Distributed File System (HDFS): HDFS is the primary storage system in Hadoop. It's designed to store large files across a distributed cluster of commodity hardware. HDFS is based on a few key principles:
    - **Blocks**: Files are divided into fixed-size blocks (typically 128MB or 256MB) and are distributed across the cluster.
    - **Replication**: Each block is replicated across multiple nodes (usually three by default) to ensure data durability and fault tolerance.
    - **Master-Slave Architecture**: HDFS has a master node called the NameNode, which manages metadata, and multiple DataNodes that store the actual data blocks.

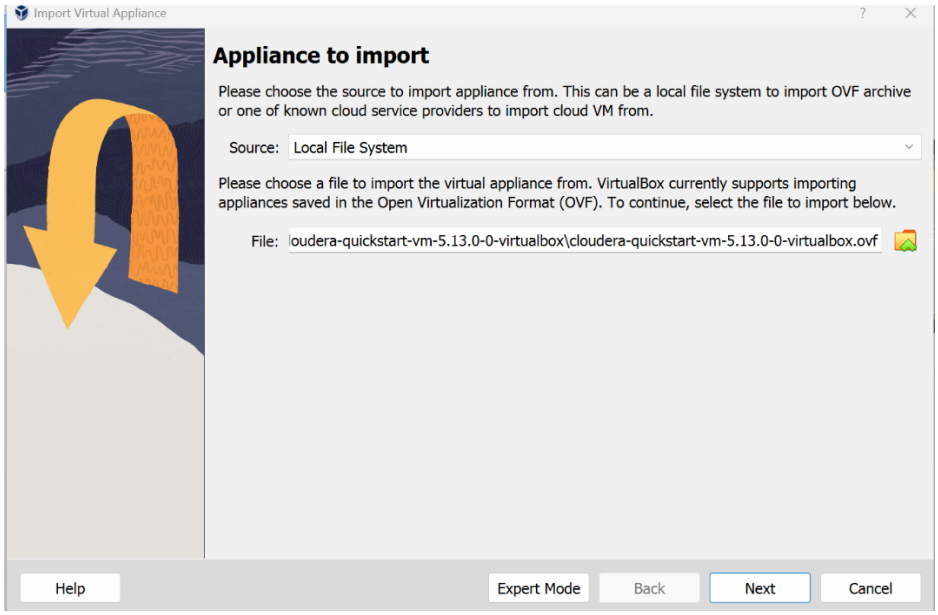
Hadoop Distributed File System (HDFS) commands are essential for managing data in a Hadoop cluster. Common commands include "ls" to list files and directories, "mkdir" to create directories, "copyFromLocal" to upload files.

Step 1:

Installing virtual box and Cloudera software.

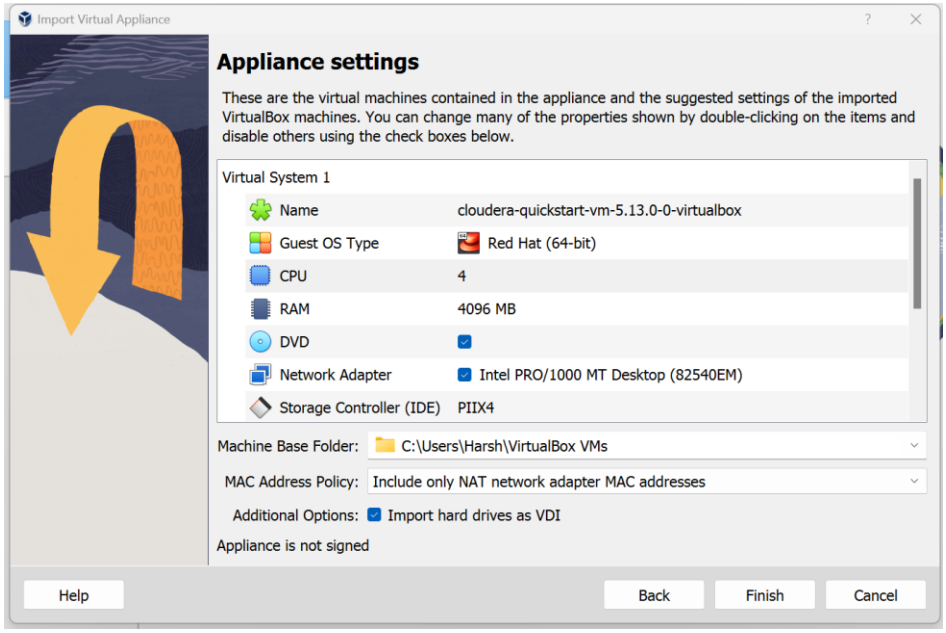
Step 2:

Unzipping the files and importing in the virtual machine. This is done by importing the link that we have downloaded.



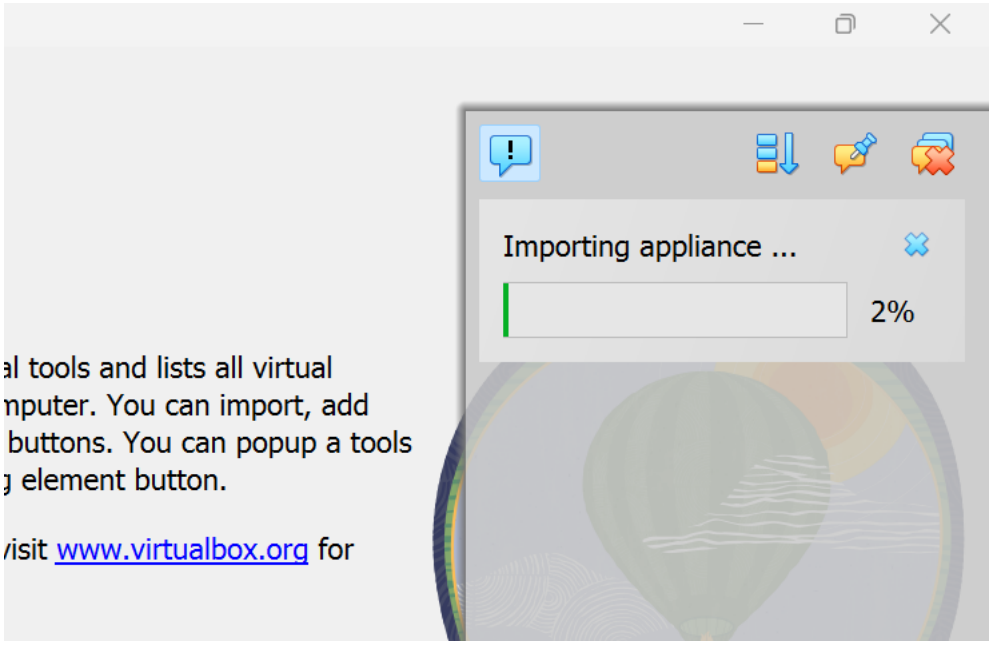
Step 3:

Selecting the import option. Also changing settings.



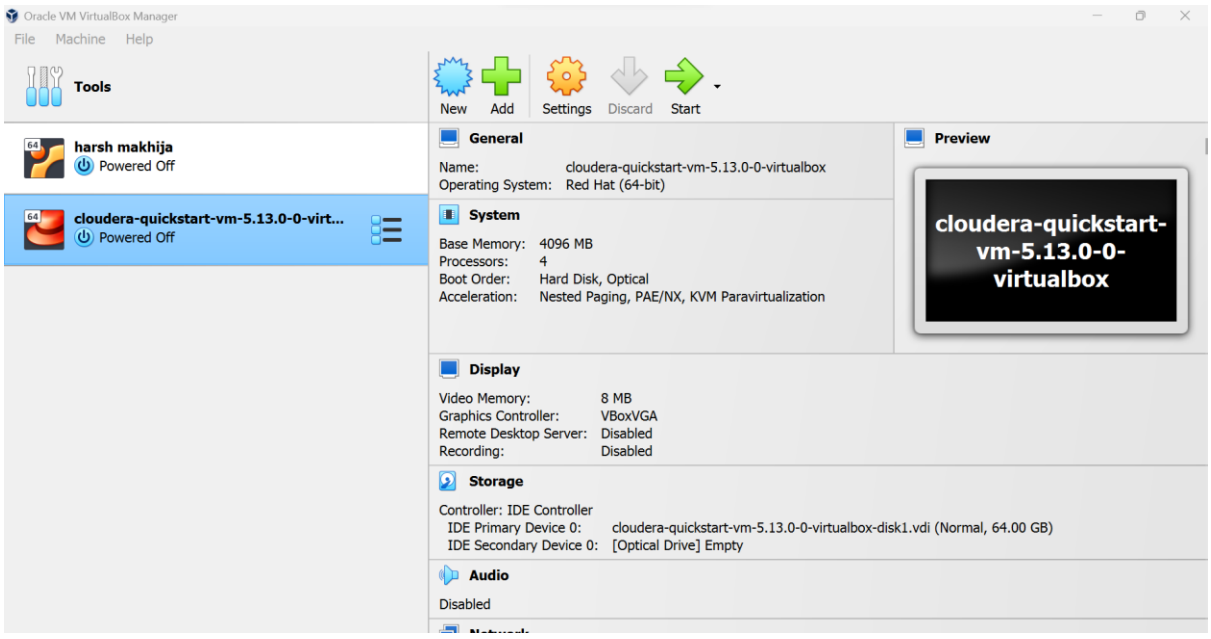
Step 4:

Importing the file and setting up.



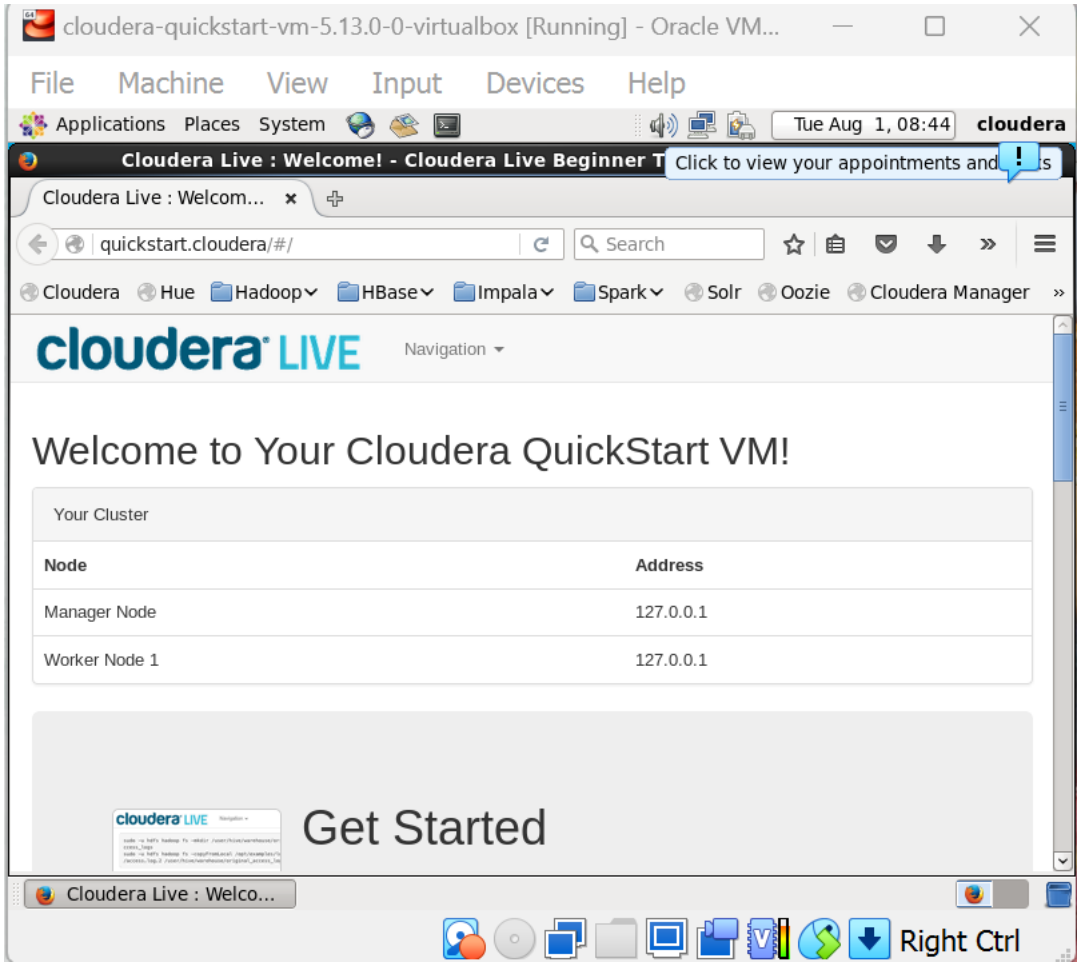
Step 5:

Import is successful, the screen will look something like this.



Step 6:

Starting the machine and loading Cloudera.

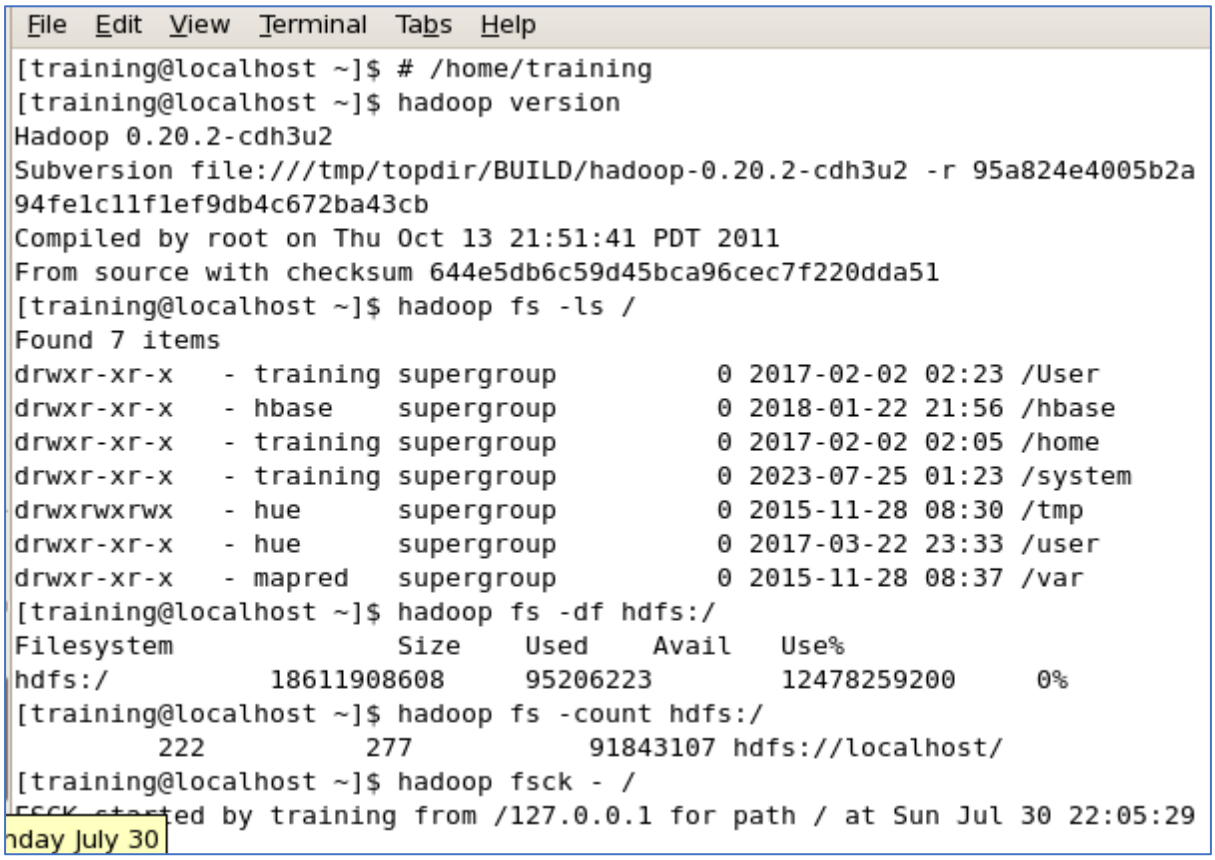


After successful completion the screen will look like this.

HDFS BASIC COMMANDS:

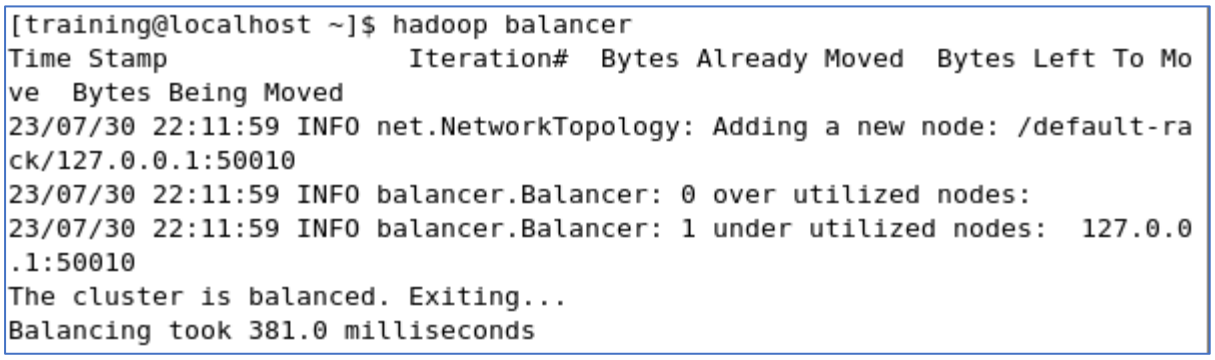
1. Hadoop version.

This command is used to check the installed version of the Hadoop framework on your system. It provides information about the Hadoop distribution, including its version number, build information, and other relevant details.



2. Hadoop balancer

The balancer command is used to balance data distribution across Hadoop HDFS (Hadoop Distributed File System) data nodes.



3. Making and removing directory.

These commands refer to basic file system operations in Hadoop HDFS. To make (create) a directory, you can use the **hadoop fs -mkdir** command followed by the directory path. To remove a directory, the **hadoop fs -rm -r** command is used, followed by the directory path.

```
[training@localhost ~]$ mkdir data/retail
[training@localhost ~]$ hadoop fs -put data/retail /user/training/hadoop
[training@localhost ~]$
```

```
[training@localhost ~]$ hadoop fs -rm hadoop/retail/customers
ter: muted
t remove hadoop/retail/customers: No such file or directory.
[training@localhost ~]$
```

4. Viewing directory retail

The **dus** command is used to determine the disk space usage (in bytes) of files and directories in Hadoop HDFS. It provides a summary of the storage consumed by each file and folder within the specified HDFS path, allowing administrators to monitor data storage utilization.

```
[training@localhost ~]$ hadoop fs -dus hadoop/retail
hdfs://localhost/user/training/hadoop/retail    0
training@localhost:~
```

5. Creating user inside directory

These commands involve administrative actions. To create a user in Hadoop, one would typically use the underlying operating system's user management tools to add a new user. Similarly, creating a sample file would involve using Hadoop's file system commands, such as **hadoop fs -touchz** to create an empty file in HDFS.

```
[training@localhost ~]$ hadoop fs -ls /user/training/user
Found 1 items
drwxr-xr-x   - training supergroup          0 2015-09-12 02:33 /user/train
ing/user/training
[training@localhost ~]$
```

```
[training@localhost ~]$ hadoop fs -mkdir /retail/user/training/hadoop
[training@localhost ~]$ hadoop fs -ls
Found 45 items
-rw-r--r--   1 training supergroup       1390 2015-10-02 20:20 /user/train
ing/Books
drwxr-xr-x   - training supergroup          0 2014-08-17 03:50 /user/train
ing/WeatherData
drwxr-xr-x   - training supergroup          0 2015-10-17 02:34 /user/train
ing/_sqoop
wse and run installed applications ergroup         23 2015-11-29 23:36 /user/train
ing/a
drwxr-xr-x   - training supergroup          0 2017-02-15 21:58 /user/train
ing/apache_hadoop
-rw-r--r--   1 training supergroup       2944 2015-09-26 02:37 /user/train
ing/bookinfo
drwxr-xr-x   - training supergroup          0 2015-09-20 01:38 /user/train
ing/class2009_dir1
-rw-r--r--   1 training supergroup         81 2015-09-18 21:05 /user/train
ing/deptinfo
drwxr-xr-x   - training supergroup          0 2014-08-17 01:59 /user/train
```

6. Making sample file

The **-put** command is a fundamental utility in Hadoop that is used to copy data from the local file system to the Hadoop Distributed File System (HDFS).

```
[training@localhost ~]$ hadoop fs -put data/sample.txt/user/training/hadoo
p
Usage: java FsShell [-put <localsrc> ... <dst>]
```

7. Hadoop count

Viewing a directory in Hadoop typically involves listing the contents of a directory within HDFS. The **Hadoop fs -ls** command is commonly used for this purpose. It provides a list of files and subdirectories, including their permissions, modification times, and file sizes, within the specified HDFS directory.

```
[training@localhost ~]$ hadoop fs -count hdfs:/
         228          278           91843853 hdfs://localhost/
[training@localhost ~]$
```

```
-rw-r--r--   1 training supergroup         44 2017-03-22 23:31 /user/training/i
nputWC.txt
-rw-r--r--   1 training supergroup         69 2015-10-10 02:35 /user/training/j
oin
-rw-r--r--   1 training supergroup        105 2015-10-10 02:38 /user/training/j
oin2
drwxr-xr-x   - training supergroup          0 2017-03-22 23:33 /user/training/o
utputWC.txt
-rw-r--r--   1 training supergroup         16 2015-10-04 01:23 /user/training/p
ig
-rw-r--r--   1 training supergroup         32 2015-10-04 02:04 /user/training/p
ig1
-rw-r--r--   1 training supergroup         90 2015-11-23 03:12 /user/training/p
oem
-rw-r--r--   1 training supergroup         90 2015-09-12 02:41 /user/training/p
oem912
drwxr-xr-x   - training supergroup          0 2015-10-10 03:53 /user/training/s
tr
drwxr-xr-x   - training supergroup          0 2015-10-17 01:26 /user/training/s
tud
drwxr-xr-x   - training supergroup          0 2015-10-17 01:42 /user/training/s
```

```
Found 46 items
-rw-r--r--   1 training supergroup       1390 2015-10-02 20:20 /user/training/B
ooks
drwxr-xr-x   - training supergroup          0 2014-08-17 03:50 /user/training/W
eatherData
drwxr-xr-x   - training supergroup          0 2015-10-17 02:34 /user/training/_
sqoop
-rw-r--r--   1 training supergroup         23 2015-11-29 23:36 /user/training/a
drwxr-xr-x   - training supergroup          0 2017-02-15 21:58 /user/training/a
pache_hadoop
-rw-r--r--   1 training supergroup       2944 2015-09-26 02:37 /user/training/b
ookinfo
drwxr-xr-x   - training supergroup          0 2015-09-20 01:38 /user/training/c
lass2009_dir1
-rw-r--r--   1 training supergroup         81 2015-09-18 21:05 /user/training/d
eptinfo
drwxr-xr-x   - training supergroup          0 2014-08-17 01:59 /user/training/d
ir1
drwxr-xr-x   - training supergroup          0 2015-09-12 02:42 /user/training/d
ir10
drwxr-xr-x   - training supergroup          0 2015-09-19 02:00 /user/training/d
```

```
[training@localhost ~]$ hadoop fs -df hdfs:/
Filesystem              Size    Used    Avail   Use%
hdfs:/          18611908608     95203328        12473774080     0%
[training@localhost ~]$
```