

ASSIGNMENT 3

PART 2 - MongoDB indexing

Most of the time, you'll want to declare your indexes before putting your application into production. This allows indexes to be built incrementally, as the data is inserted. But there are two cases where you might choose to build an index after the fact. The first case occurs when you need to import a lot of data before switching into production. For instance, you might be migrating an application to MongoDB and need to seed the database with user information from a data warehouse. You could create the indexes on your user data in advance but doing so after you have imported the data will ensure an ideally balanced and compacted index from the start. This will also minimize the net time to build the index. Use the NYSE dataset to declare your indexes before putting your application into production.

ANS-

Created new DB- Assignment_4

Then created a collections nyse_new

Created index on key: stock_volume in ascending order

```
> use assignment_4
switched to db assignment_4
>
>
> db.nyse_new.find()
>
> db.nyse_new.createIndex({"stock_volume":1})
{
  "createdCollectionAutomatically" : false,
  "numIndexesBefore" : 1,
  "numIndexesAfter" : 2,
  "ok" : 1
}
>
>
```

Then imported the nyse data to this collection using bat file

```
@echo off
FOR %%G IN (A B C D E F G H I J K L M N O P Q R S T U V W X Y Z) DO mongoimport --db assignment_4 --type csv
--collection nyse_new --headerline --file C:\temp\unzip\NYSE\NYSE_daily_prices_%%G.csv
pause
```

The indexes are correctly created.

```
>
> db.nyse_new.getIndexes()
[
  {
    "v" : 2,
    "key" : {
      "_id" : 1
    },
    "name" : "_id_",
    "ns" : "assignment_4.nyse_new"
  },
  {
    "v" : 2,
    "key" : {
      "stock_volume" : 1
    },
    "name" : "stock_volume_1",
    "ns" : "assignment_4.nyse_new"
  }
]
>
```

PART 3 - MongoDB Indexing

Insert the NYSE dataset into a new database. You may use the existing NYSE database created before.

Now, create indexes on existing data sets.

Ans—

Imported all the nyse data to new collection stocks

Then after that created index on key: stock_symbol in ascending order

```
>
> db.stocks.createIndex({"stock_symbol":1})
{
  "createdCollectionAutomatically" : false,
  "numIndexesBefore" : 1,
  "numIndexesAfter" : 2,
  "ok" : 1
}
>
> db.stocks.getIndexes()
[
  {
    "v" : 2,
    "key" : {
      "_id" : 1
    },
    "name" : "_id_",
    "ns" : "assignment_4.stocks"
  },
  {
    "v" : 2,
    "key" : {
      "stock_symbol" : 1
    },
    "name" : "stock_symbol_1",
    "ns" : "assignment_4.stocks"
  }
]
>
```

PART 4 – Programming Assignment

All hadoop commands are invoked by the bin/hadoop script. Running the hadoop script without any arguments prints the description for all commands.

Usage: hadoop [--config confdir] [--loglevel loglevel] [COMMAND] [GENERIC_OPTIONS] [COMMAND_OPTIONS]

Execute each hadoop command once and place the screenshots into a word file. If a command cannot be executed for any reason (such as, a distributed environment is needed), you may write the definition of the command, and skip execution.

<http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>

Ans:

- appendToFile
- cat

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -appendToFile ~/Downloads/ebook/fileaa /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /myEbooks/kvf.txt
A
1
B
2
C
3
d
7
A
1
B
2
C
3
d
7
A
1
B
2
C
3
d
7
The Project Gutenberg EBook of Studies of Trees, by Jacob Joshua Levison
This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever. You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.net
Title: Studies of Trees
Author: Jacob Joshua Levison
Release Date: June 23, 2005 [EBook #16116]
Language: English
```

- [checksum](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -checksum /myEbooks/kvf.txt
/myEbooks/kvf.txt MD5-of-0MD5-of-512CRC32C 00000200000000000000000068e3660cac1d9110c04d77ffcf11c119
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [chgrp](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -chgrp -R user /myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 2 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [chmod](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -chmod 777 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks/kvf.txt
-rwxrwxrwx 1 ankit user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [chown](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks/kvf.txt
-rwxrwxrwx 1 ankit user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -chown NewUser /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks/kvf.txt
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [copyFromLocal](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 2 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyFromLocal ~/Downloads/ebook/fileaa /myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 3 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [copyToLocal](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyToLocal /myEbooks/kvf.txt ~/Downloads/test.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$

ankit@ankit-VirtualBox: ~/Downloads$ ls -lr
total 29136
-rw-r--r-- 1 ankit ankit 10051 Jun 12 15:02 test.txt
-rw-rw-r-- 1 ankit ankit 5407 Jun 11 17:33 Q7.jar
-rw-rw-r-- 1 ankit ankit 24713 Jun 11 22:33 Q73.jar
-rw-rw-r-- 1 ankit ankit 24517 Jun 11 22:52 Q731.jar
-rw-rw-r-- 1 ankit ankit 17117 Jun 11 21:28 Q72.jar
-rw-rw-r-- 1 ankit ankit 10190 Jun 11 19:03 Q71.jar
-rw-rw-r-- 1 ankit ankit 5233 Jun 11 17:01 Q62.jar
-rw-rw-r-- 1 ankit ankit 4637 Jun 11 15:52 Q41.jar
-rw-rw-r-- 1 ankit ankit 38050 Jun 7 15:26 petrified.txt
-rw-rw-r-- 1 ankit ankit 77367 Jun 11 15:58 part-r-00000
drwxr-xr-x 2 ankit ankit 4096 Jun 11 16:22 output
drwxrwxr-x 2 ankit ankit 4096 Jun 11 17:01 NYSE
-rw-rw-r-- 1 ankit ankit 5837 Jun 7 17:46 mrwordcount.jar
-rw-rw-r-- 1 ankit ankit 5263 Jun 8 09:18 mrwordcount-2.jar
-rw-rw-r-- 1 ankit ankit 14777349 Jun 11 15:09 logs
drwxrwxr-x 8 ankit ankit 4096 Jun 8 09:54 idea-IC-191.7479.19
-rw-rw-r-- 1 ankit ankit 14777349 Jun 11 15:03 http_access.log
drwxr-xr-x 7 ankit ankit 4096 Jun 11 14:58 Engg-Of-Big-Data
drwxr-xr-x 2 ankit ankit 4096 Jun 11 22:46 ebook
ankit@ankit-VirtualBox: ~/Downloads$ cat test.txt | head
A 1
B 2
C 3
d 7
A 1
B 2
C 3
d 7
A 1
B 2
ankit@ankit-VirtualBox: ~/Downloads$
```

- [count](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -count -q /myEbooks
none      inf      none      inf      1      3      308377 /myEbooks
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -count -q /myEbooks/kvf.txt
none      inf      none      inf      0      1      10051 /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -count -u /myEbooks/kvf.txt
none      inf      none      inf      /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -count -h /myEbooks/kvf.txt
0      1      9.8 K /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -count -h /myEbooks
1      3      301.1 K /myEbooks
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [cp](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /testdir
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cp /myEbooks/kvf.txt /testdir
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /testdir
Found 1 items
-rw-r--r-- 1 ankit supergroup 10051 2019-06-12 15:08 /testdir/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [createSnapshot](#)

Create a snapshot of a snapshottable directory. This operation requires owner privilege of the snapshottable directory.

Command:

hdfs dfs -createSnapshot <path> [<snapshotName>]

- [deleteSnapshot](#)

Delete a snapshot of from a snapshottable directory. This operation requires owner privilege of the snapshottable directory.

Command:

hdfs dfs -deleteSnapshot <path> <snapshotName>

- [df](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -df /
Filesystem      Size      Used      Available Use%
hdfs://localhost:9000 52414619648 1051281306 33708437504 2%
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -df -h /
Filesystem      Size      Used      Available Use%
hdfs://localhost:9000 48.8 G 1002.6 M 31.4 G 2%
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [du](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -du -h /
0      /31May
169.0 K /Assignment3_Q7Output
75.6 K /IpCountMR
281.6 K /ebooks
37.2 K /guttenberg
20.2 K /guttenbergOutput1
20.2 K /guttenbergOutput2
14.1 M /logs
75.6 K /logsOutput
301.1 K /myEbooks
487.4 M /nyse
27.3 K /nyseOutput
487.4 M /nyse_merged
9.8 K /testdir
4.1 M /tmp
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [dus](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -dus /
dus: DEPRECATED: Please use 'du -s' instead.
1042254641 /
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -dus -h /
dus: DEPRECATED: Please use 'du -s' instead.
994.0 M /
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [expunge](#)

Usage: `hadoop fs -expunge`

Permanently delete files in checkpoints older than the retention threshold from trash directory and create new checkpoint.

When checkpoint is created, recently deleted files in trash are moved under the checkpoint. Files in checkpoints older than `fs.trash.interval` will be permanently deleted on the next invocation of `-expunge` command.

- [find](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -find /myEbooks -name kvf.txt
/myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -find / -name kvf.txt
/myEbooks/kvf.txt
/testdir/kvf.txt
```

- [get](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -get /myEbooks/kvf.txt ~/Downloads/copy/test
```

```
ankit@ankit-VirtualBox:~/Downloads/copy$ ls -lr
total 12
-rw-r--r-- 1 ankit ankit 10051 Jun 12 15:22 test
ankit@ankit-VirtualBox:~/Downloads/copy$
```

- [getfacl](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getfacl /myEbooks/kvf.txt
# file: /myEbooks/kvf.txt
# owner: NewUser
# group: user
user::rwx
group::rwx
other::rwx

ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getfacl /
# file: /
# owner: ankit
# group: supergroup
user::rwx
group::r-x
other::r-x
```

- [getfattr](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getfattr -d /
# file: /
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getfattr -d /myEbooks/kvf.txt
# file: /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [getmerge](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getmerge /ebooks ~/Downloads/copy/output.txt
ankit@ankit-VirtualBox:~/Downloads/copy$ ls -lr -h
total 296K
-rw-r--r-- 1 ankit ankit 9.9K Jun 12 15:22 test
-rw-r--r-- 1 ankit ankit 282K Jun 12 15:28 output.txt
ankit@ankit-VirtualBox:~/Downloads/copy$
```

- [help](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -help
Usage: hadoop fs [generic options]
[-appendToFile <localsrc> ... <dst>]
[-cat [-ignoreCrc] <src> ...]
[-checksum <src> ...]
[-chgrp [-R] GROUP PATH...]
[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]
[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
[-createSnapshot <snapshotDir> [<snapshotName>]]
[-deleteSnapshot <snapshotDir> <snapshotName>]
[-df [-h] [<path> ...]]
[-du [-s] [-h] [-x] <path> ...]
[-expunge]
[-find <path> ... <expression> ...]
[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-getfacl [-R] <path>]
[-getfattr [-R] {-n name | -d} [-e en] <path>]
[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
[-help [cmd ...]]
[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
[-mkdir [-p] <path> ...]
[-moveFromLocal <localsrc> ... <dst>]
```

- [ls](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /
Found 15 items
drwxr-xr-x - ankit supergroup 0 2019-06-06 05:34 /31May
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:53 /Assignment3_Q7Output
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:00 /IpCountMR
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:48 /ebooks
drwxr-xr-x - ankit supergroup 0 2019-06-07 15:32 /guttenberg
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:10 /guttenbergOutput1
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:29 /guttenbergOutput2
-rw-r--r-- 1 ankit supergroup 14777349 2019-06-11 15:59 /logs
drwxr-xr-x - ankit supergroup 0 2019-06-11 15:55 /logsOutput
drwxr-xr-x - ankit user 0 2019-06-12 14:59 /myEbooks
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:14 /nyse
drwxr-xr-x - ankit supergroup 0 2019-06-11 17:03 /nyseOutput
-rw-r--r-- 1 ankit supergroup 511085653 2019-06-11 16:23 /nyse_merged
drwxr-xr-x - ankit supergroup 0 2019-06-12 15:08 /testdir
drwx----- - ankit supergroup 0 2019-06-07 17:48 /tmp
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```


- [lsr](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -lsr /
lsr: DEPRECATED: Please use 'ls -R' instead.
drwxr-xr-x - ankit supergroup 0 2019-06-06 05:34 /31May
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:53 /Assignment3_Q70Output
drwxr-xr-x - ankit supergroup 0 2019-06-11 17:38 /Assignment3_Q70Output/A_TextInputFormat
-rw-r--r-- 1 ankit supergroup 0 2019-06-11 17:38 /Assignment3_Q70Output/A_TextInputFormat/_SUCCESS
-rw-r--r-- 1 ankit supergroup 86505 2019-06-11 17:38 /Assignment3_Q70Output/A_TextInputFormat/part-r-00000
drwxr-xr-x - ankit supergroup 0 2019-06-11 19:05 /Assignment3_Q70Output/B_KeyValueTextInputFormat
-rw-r--r-- 1 ankit supergroup 0 2019-06-11 19:05 /Assignment3_Q70Output/B_KeyValueTextInputFormat/_SUCCESS
-rw-r--r-- 1 ankit supergroup 16 2019-06-11 19:05 /Assignment3_Q70Output/B_KeyValueTextInputFormat/part-r-00000
drwxr-xr-x - ankit supergroup 0 2019-06-11 21:37 /Assignment3_Q70Output/C_NLineInputFormat
-rw-r--r-- 1 ankit supergroup 0 2019-06-11 21:37 /Assignment3_Q70Output/C_NLineInputFormat/_SUCCESS
-rw-r--r-- 1 ankit supergroup 86505 2019-06-11 21:37 /Assignment3_Q70Output/C_NLineInputFormat/part-r-00000
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:54 /Assignment3_Q70Output/D_CombineInputFileFormat1
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:00 /IpCountMR
-rw-r--r-- 1 ankit supergroup 0 2019-06-11 16:00 /IpCountMR/_SUCCESS
-rw-r--r-- 1 ankit supergroup 77757 2019-06-11 16:00 /IpCountMR/part-r-00000
```

- [mkdir](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /test
ls: '/test': No such file or directory
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -mkdir /test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /
Found 16 items
drwxr-xr-x - ankit supergroup 0 2019-06-06 05:34 /31May
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:53 /Assignment3_Q70Output
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:00 /IpCountMR
drwxr-xr-x - ankit supergroup 0 2019-06-11 22:48 /ebooks
drwxr-xr-x - ankit supergroup 0 2019-06-07 15:32 /gutenberg
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:10 /gutenbergOutput1
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:29 /gutenbergOutput2
-rw-r--r-- 1 ankit supergroup 14777349 2019-06-11 15:59 /logs
drwxr-xr-x - ankit supergroup 0 2019-06-11 15:55 /logsOutput
drwxr-xr-x - ankit user 0 2019-06-12 14:59 /myEbooks
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:14 /nyse
drwxr-xr-x - ankit supergroup 0 2019-06-11 17:03 /nyseOutput
-rw-r--r-- 1 ankit supergroup 511085653 2019-06-11 16:23 /nyse_merged
drwxr-xr-x - ankit supergroup 0 2019-06-12 15:32 /test
```

- [moveFromLocal](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 3 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -moveFromLocal ~/Downloads/copy/output.txt /myEbooks
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 4 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 288326 2019-06-12 15:35 /myEbooks/output.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

```
ankit@ankit-VirtualBox:~/Downloads/copy$ ls -lr -h
total 296K
-rw-r--r-- 1 ankit ankit 9.9K Jun 12 15:22 test
-rw-r--r-- 1 ankit ankit 282K Jun 12 15:28 output.txt
ankit@ankit-VirtualBox:~/Downloads/copy$ ls -lr -h
total 12K
-rw-r--r-- 1 ankit ankit 9.9K Jun 12 15:22 test
ankit@ankit-VirtualBox:~/Downloads/copy$
```

- [moveToLocal](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -moveToLocal ~/Downloads/copy/output.txt /myEbooks
moveToLocal: Option '-moveToLocal' is not implemented yet.
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [mv](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 4 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 288326 2019-06-12 15:35 /myEbooks/output.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks/output.txt /test
-rw-r--r-- 1 ankit user 288326 2019-06-12 15:35 /myEbooks/output.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -mv /myEbooks/output.txt /test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /test
Found 1 items
-rw-r--r-- 1 ankit user 288326 2019-06-12 15:35 /test/output.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
ls: '/myEbooks': No such file or directory
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 3 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [put](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 4 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -put ~/Downloads/copy/petrified.txt /myEbooks
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 5 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [renameSnapshot](#)

Rename a snapshot. This operation requires owner privilege of the snapshottable directory.

Command:

hdfs dfs -renameSnapshot <path> <oldName> <newName>

- [rm](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 5 items
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rw-r--r-- 1 ankit user 10000 2019-06-12 14:59 /myEbooks/fileaa
-rwxrwxrwx 1 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -rm /myEbooks/fileaa
Deleted /myEbooks/fileaa
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [rmdir](#)

```

ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -rmdir /test
rmdir: `/test': Directory is not empty
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -rmdir /31stMay
rmdir: `/31stMay': No such file or directory
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /
Found 16 items
drwxr-xr-x - ankit supergroup          0 2019-06-06 05:34 /31May
drwxr-xr-x - ankit supergroup          0 2019-06-11 22:53 /Assignment3_Q70Output
drwxr-xr-x - ankit supergroup          0 2019-06-11 16:00 /IpCountMR
drwxr-xr-x - ankit supergroup          0 2019-06-11 22:48 /ebooks
drwxr-xr-x - ankit supergroup          0 2019-06-07 15:32 /gutemberg
drwxr-xr-x - ankit supergroup          0 2019-06-08 09:10 /gutembergOutput1
drwxr-xr-x - ankit supergroup          0 2019-06-08 09:29 /gutembergOutput2
-rw-r--r-- 1 ankit supergroup 14777349 2019-06-11 15:59 /logs
drwxr-xr-x - ankit supergroup          0 2019-06-11 15:55 /logsOutput
drwxr-xr-x - ankit user                0 2019-06-12 15:42 /myEbooks
drwxr-xr-x - ankit supergroup          0 2019-06-11 16:14 /nyse
drwxr-xr-x - ankit supergroup          0 2019-06-11 17:03 /nyseOutput
-rw-r--r-- 1 ankit supergroup 511085653 2019-06-11 16:23 /nyse_merged
drwxr-xr-x - ankit supergroup          0 2019-06-12 15:38 /test
drwxr-xr-x - ankit supergroup          0 2019-06-12 15:08 /testdir
drwx----- - ankit supergroup          0 2019-06-07 17:48 /tmp
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -rmdir /31May
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /
Found 15 items
drwxr-xr-x - ankit supergroup          0 2019-06-11 22:53 /Assignment3_Q70Output
drwxr-xr-x - ankit supergroup          0 2019-06-11 16:00 /IpCountMR
drwxr-xr-x - ankit supergroup          0 2019-06-11 22:48 /ebooks
drwxr-xr-x - ankit supergroup          0 2019-06-07 15:32 /gutemberg
drwxr-xr-x - ankit supergroup          0 2019-06-08 09:10 /gutembergOutput1
drwxr-xr-x - ankit supergroup          0 2019-06-08 09:29 /gutembergOutput2
-rw-r--r-- 1 ankit supergroup 14777349 2019-06-11 15:59 /logs
drwxr-xr-x - ankit supergroup          0 2019-06-11 15:55 /logsOutput
drwxr-xr-x - ankit user                0 2019-06-12 15:42 /myEbooks
drwxr-xr-x - ankit supergroup          0 2019-06-11 16:14 /nyse
drwxr-xr-x - ankit supergroup          0 2019-06-11 17:03 /nyseOutput
-rw-r--r-- 1 ankit supergroup 511085653 2019-06-11 16:23 /nyse_merged
drwxr-xr-x - ankit supergroup          0 2019-06-12 15:38 /test
drwxr-xr-x - ankit supergroup          0 2019-06-12 15:08 /testdir
drwx----- - ankit supergroup          0 2019-06-07 17:48 /tmp

```

- [rmr](#)

```

ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -rmr /test
rmr: DEPRECATED: Please use '-rm -r' instead.
Deleted /test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$

```

- [setfacl](#)

Usage: `hadoop fs -setfacl [-R] [-b |-k -m |-x <acl_spec> <path>] [--set <acl_spec> <path>]`

Sets Access Control Lists (ACLs) of files and directories.

- [setfattr](#)

```

ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -setfattr -n user.noValue /myEbooks/petrified.txt
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 4 items
-rw-r--r-- 1 ankit user      288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 1 NewUser user    10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user      38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user      10051 2019-06-12 15:40 /myEbooks/test
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$

```

- [setrep](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -setrep -R 3 /myEbooks/kvf.txt
Replication 3 set: /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [stat](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -stat "type:%F perm:%a %u:%g size:%b mtime:%y atime:%x name:%n" /myEbooks/kvf.txt
type:regular file perm:a NewUser:user size:10051 mtime:2019-06-12 18:48:05 atime:x name:kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -stat "type:%F perm:%a %u:%g size:%b mtime:%y atime:%x name:%n" /myEbooks
type:directory perm:a ankit:user size:0 mtime:2019-06-12 19:42:35 atime:x name:myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

- [tail](#)

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -tail /myEbooks/kvf.txt
ches or more in length. These needles
keep green throughout the entire year. This is characteristic of all
coniferous trees, except the larch and cypress, which shed their
leaves in winter.

[Illustration: FIG. 2.--Twig of the White Pine.]

The pines are widely distributed throughout the Northern Hemisphere,
and include about 80 distinct species with over 600 varieties. The
species enumerated here are especially common in the eastern part of
the United States, growing either native in the forest or under
cultivation in the parks. The pines form a very important class of
timber trees, and produce beautiful effects when planted in groups
in the parks.

How to tell them from each other: The pine needles are arranged in
_clusters_; see Fig. 1. Each species has a certain characteristic
number of needles to the cluster and this fact generally provides
the simplest and most direct way of distinguishing the different
```

- [test](#)

Usage: `hadoop fs -test [-defsz] URI`

Options:

- d: if the path is a directory, return 0.
- e: if the path exists, return 0.
- f: if the path is a file, return 0.
- s: if the path is not empty, return 0.
- r: if the path exists and read permission is granted, return 0.
- w: if the path exists and write permission is granted, return 0.
- z: if the file is zero length, return 0.

Example:

`hadoop fs -test -e filename`

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -test -e /myEbooks/kvf.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -test -e /myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```


- [text](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -text /myEbooks/test
```

```
A 1
B 2
C 3
d 7
A 1
B 2
C 3
d 7
A 1
B 2
C 3
d 7
```

The Project Gutenberg EBook of Studies of Trees, by Jacob Joshua Levison

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.net

- [touchz](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
```

Found 4 items

```
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 3 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 3 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -touchz /myEbooks/newFile
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
```

Found 5 items

```
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 3 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 0 2019-06-12 15:55 /myEbooks/newFile
-rw-r--r-- 3 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [truncate](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
```

Found 5 items

```
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 3 NewUser user 10051 2019-06-12 14:48 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 0 2019-06-12 15:55 /myEbooks/newFile
-rw-r--r-- 3 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -truncate 127 /myEbooks/kvf.txt
```

Truncating /myEbooks/kvf.txt to length: 127. Wait for block recovery to complete before further updating this file.

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
```

Found 5 items

```
-rw-r--r-- 1 ankit user 288326 2019-06-11 17:35 /myEbooks/ebook.txt
-rwxrwxrwx 3 NewUser user 127 2019-06-12 15:56 /myEbooks/kvf.txt
-rw-r--r-- 1 ankit user 0 2019-06-12 15:55 /myEbooks/newFile
-rw-r--r-- 3 ankit user 38050 2019-06-12 15:41 /myEbooks/petrified.txt
-rw-r--r-- 1 ankit user 10051 2019-06-12 15:40 /myEbooks/test
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

- [usage](#)

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -usage ls
```

Usage: hadoop fs [generic options] -ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -usage mv
```

Usage: hadoop fs [generic options] -mv <src> ... <dst>

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -usage truncate
```

Usage: hadoop fs [generic options] -truncate [-w] <length> <path> ...

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

PART 5 – Programming Assignment

Copy the attached 'http_access.log' file into HDFS under /logs directory.

Using the access.log file stored in HDFS, implement MapReduce in Hadoop to find the number of times each IP accessed the website.

Ans--

copy http_access.log to /logs

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$  
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyFromLocal /home/ankit/Downloads/http_access.log /logs  
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q41.jar MR.LogMainMR /logs/ /IpCountMR  
19/06/11 16:00:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
19/06/11 16:00:13 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.  
19/06/11 16:00:13 INFO InputFileInputFormat: Total input files to process : 1  
19/06/11 16:00:13 INFO mapreduce.JobSubmitter: number of splits:1  
19/06/11 16:00:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560279693514_0003  
19/06/11 16:00:14 INFO impl.VarnClientImpl: Submitted application application_1560279693514_0003  
19/06/11 16:00:14 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560279693514_0003/  
19/06/11 16:00:14 INFO mapreduce.Job: Running job: job_1560279693514_0003  
19/06/11 16:00:24 INFO mapreduce.Job: Job job_1560279693514_0003 running in uber mode : false  
19/06/11 16:00:24 INFO mapreduce.Job: map 0% reduce 0%
```

Also ran MAP reduce on the logs to count the number of times ip is used

The program folder is inside the Assignment folder

Output head is as follows

```
19/06/11 16:00:24 INFO mapreduce.Job: Job job_1560279693514_0003 running in uber mode : false  
19/06/11 16:00:24 INFO mapreduce.Job: map 0% reduce 0%  
19/06/11 16:00:44 INFO mapreduce.Job: map 100% reduce 0%  
19/06/11 16:00:53 INFO mapreduce.Job: map 100% reduce 100%  
19/06/11 16:00:54 INFO mapreduce.Job: Job job_1560279693514_0003 completed successfully  
19/06/11 16:00:55 INFO mapreduce.Job: Counter: 49  
File System Counters  
FILE: Number of bytes read=2996770  
FILE: Number of bytes written=6309919  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=14777440  
HDFS: Number of bytes written=77367  
HDFS: Number of read operations=6  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=2  
Job Counters  
Launched map tasks=1  
Launched reduce tasks=1  
Data-local map tasks=1  
Total time spent by all maps in occupied slots (ms)=17259  
Total time spent by all reduces in occupied slots (ms)=6057  
Total time spent by all map tasks (ms)=17259  
Total time spent by all reduce tasks (ms)=6057  
Total vcore-millisecods taken by all map tasks=17259  
Total vcore-millisecods taken by all reduce tasks=6057  
Total megabyte-millisecods taken by all map tasks=17673216  
Total megabyte-millisecods taken by all reduce tasks=6202368  
Map-Reduce Framework  
Map input records=148202  
Map output records=148202  
Map output bytes=2700360  
Map output materialized bytes=2996770  
Input split bytes=91  
Combine input records=0  
Combine output records=0  
Reduce input groups=4713  
Reduce shuffle bytes=2996770  
Reduce input records=148202  
Reduce output records=4713  
Spilled Records=296404  
Shuffled Maps =1  
Failed Shuffles=0  
Merged Map outputs=1  
GC time elapsed (ms)=487  
CPU time spent (ms)=5030
```

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /IpCountMR/part-r-000000 | head  
1  
1.170.44.84 164  
1.179.132.221 103  
1.186.119.102 4  
1.202.184.142 2  
1.202.184.145 2  
1.202.22.83 1  
1.202.89.134 3  
1.22.56.96 10  
1.23.162.141 1
```

PART 6 – Programming Assignment

Download and Copy all the files (<http://msis.neu.edu/nyse/>) (DailyPrices_A to DailyPrices_Z) to a folder in HDFS.

Write a MapReduce to find the Max price of stock_price_high for each stock. Capture the running time programmatically (or manually using a wristwatch or smartphone).

Ans--

Download and Copy all the files (<http://msis.neu.edu/nyse/>) (DailyPrices_A to DailyPrices_Z) to a folder in HDFS.

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -put ~/Downloads/NYSE /nyse
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /nyse
Found 26 items
-rw-r--r-- 1 ankit supergroup 40990992 2019-06-11 16:13 /nyse/NYSE_daily_prices_A.csv
-rw-r--r-- 1 ankit supergroup 32034760 2019-06-11 16:13 /nyse/NYSE_daily_prices_B.csv
-rw-r--r-- 1 ankit supergroup 45790256 2019-06-11 16:13 /nyse/NYSE_daily_prices_C.csv
-rw-r--r-- 1 ankit supergroup 19234471 2019-06-11 16:13 /nyse/NYSE_daily_prices_D.csv
-rw-r--r-- 1 ankit supergroup 22104043 2019-06-11 16:13 /nyse/NYSE_daily_prices_E.csv
-rw-r--r-- 1 ankit supergroup 17387253 2019-06-11 16:13 /nyse/NYSE_daily_prices_F.csv
-rw-r--r-- 1 ankit supergroup 22608522 2019-06-11 16:13 /nyse/NYSE_daily_prices_G.csv
-rw-r--r-- 1 ankit supergroup 23127143 2019-06-11 16:13 /nyse/NYSE_daily_prices_H.csv
-rw-r--r-- 1 ankit supergroup 20680033 2019-06-11 16:13 /nyse/NYSE_daily_prices_I.csv
-rw-r--r-- 1 ankit supergroup 9537527 2019-06-11 16:14 /nyse/NYSE_daily_prices_J.csv
-rw-r--r-- 1 ankit supergroup 14782892 2019-06-11 16:14 /nyse/NYSE_daily_prices_K.csv
-rw-r--r-- 1 ankit supergroup 12958785 2019-06-11 16:14 /nyse/NYSE_daily_prices_L.csv
-rw-r--r-- 1 ankit supergroup 38124545 2019-06-11 16:14 /nyse/NYSE_daily_prices_M.csv
-rw-r--r-- 1 ankit supergroup 31488945 2019-06-11 16:14 /nyse/NYSE_daily_prices_N.csv
-rw-r--r-- 1 ankit supergroup 8865718 2019-06-11 16:14 /nyse/NYSE_daily_prices_O.csv
-rw-r--r-- 1 ankit supergroup 31943478 2019-06-11 16:14 /nyse/NYSE_daily_prices_P.csv
-rw-r--r-- 1 ankit supergroup 190989 2019-06-11 16:14 /nyse/NYSE_daily_prices_Q.csv
-rw-r--r-- 1 ankit supergroup 16808595 2019-06-11 16:14 /nyse/NYSE_daily_prices_R.csv
-rw-r--r-- 1 ankit supergroup 31852353 2019-06-11 16:14 /nyse/NYSE_daily_prices_S.csv
-rw-r--r-- 1 ankit supergroup 28754690 2019-06-11 16:14 /nyse/NYSE_daily_prices_T.csv
-rw-r--r-- 1 ankit supergroup 9951590 2019-06-11 16:14 /nyse/NYSE_daily_prices_U.csv
-rw-r--r-- 1 ankit supergroup 9503196 2019-06-11 16:14 /nyse/NYSE_daily_prices_V.csv
-rw-r--r-- 1 ankit supergroup 15972013 2019-06-11 16:14 /nyse/NYSE_daily_prices_W.csv
-rw-r--r-- 1 ankit supergroup 3613198 2019-06-11 16:14 /nyse/NYSE_daily_prices_X.csv
-rw-r--r-- 1 ankit supergroup 686216 2019-06-11 16:14 /nyse/NYSE_daily_prices_Y.csv
-rw-r--r-- 1 ankit supergroup 2093424 2019-06-11 16:14 /nyse/NYSE_daily_prices_Z.csv
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

Merging all records in single file and saving in HDFS

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -getmerge -nl /nyse ~/Downloads/output/all_stocks.csv
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyFromLocal /home/ankit/Downloads/output/all_stocks.csv /nyse_merged
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /
Found 12 items
drwxr-xr-x - ankit supergroup 0 2019-06-06 05:34 /31May
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:00 /IpCountMR
drwxr-xr-x - ankit supergroup 0 2019-06-07 15:32 /gutenberg
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:10 /gutenbergOutput1
drwxr-xr-x - ankit supergroup 0 2019-06-08 09:29 /gutenbergOutput2
-rw-r--r-- 1 ankit supergroup 14777349 2019-06-11 15:59 /logs
drwxr-xr-x - ankit supergroup 0 2019-06-11 15:55 /logsOutput
drwxr-xr-x - ankit supergroup 0 2019-06-06 07:00 /myEbooks
drwxr-xr-x - ankit supergroup 0 2019-06-11 16:14 /nyse
-rw-r--r-- 1 ankit supergroup 511085653 2019-06-11 16:23 /nyse_merged
drwxr-xr-x - ankit supergroup 0 2019-06-05 11:25 /testdir
drwx----- - ankit supergroup 0 2019-06-07 17:48 /tmp
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /nyse_merged
```

The java program is in the assignment folder

Output of MapReduce for maximum stock_price_high for each stock is as follows:

```

ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q61.jar com.hadoop.assignment3_q6.NyseMainMR /nyse_merged/ /nyseOutput
19/06/11 17:02:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/11 17:02:32 WARN mapreduce.jobresourceuploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/11 17:02:32 INFO Input.FileInputFormat: Total input files to process : 1
19/06/11 17:02:33 INFO mapreduce.JobSubmitter: number of splits:4
19/06/11 17:02:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560279693514_0006
19/06/11 17:02:34 INFO ImplVarClientImpl: Submitted application application_1560279693514_0006
19/06/11 17:02:34 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560279693514_0006/
19/06/11 17:02:34 INFO mapreduce.Job: Running job: job_1560279693514_0006
19/06/11 17:02:43 INFO mapreduce.Job: Job job_1560279693514_0006 running in uber mode : false
19/06/11 17:02:43 INFO mapreduce.Job: map 0% reduce 0%
19/06/11 17:03:27 INFO mapreduce.Job: map 60% reduce 0%
19/06/11 17:03:33 INFO mapreduce.Job: map 67% reduce 0%
19/06/11 17:03:34 INFO mapreduce.Job: map 75% reduce 0%
19/06/11 17:03:35 INFO mapreduce.Job: map 92% reduce 0%
19/06/11 17:03:36 INFO mapreduce.Job: map 100% reduce 0%
19/06/11 17:03:48 INFO mapreduce.Job: map 100% reduce 100%
19/06/11 17:03:49 INFO mapreduce.Job: Job job_1560279693514_0006 completed successfully
19/06/11 17:03:49 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=128173371
    FILE: Number of bytes written=257138082
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=511098333
    HDFS: Number of bytes written=27922
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2

```

```

ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /nyseOutput
Found 2 items
-rw-r--r-- 1 ankit supergroup 0 2019-06-11 17:03 /nyseOutput/_SUCCESS
-rw-r--r-- 1 ankit supergroup 27922 2019-06-11 17:03 /nyseOutput/part-r-00000
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /nyseOutput/part-r-00000
AA 94.62
AAI 57.88
AAN 35.21
AAP 83.65
AAR 25.25
AAV 24.78
AB 94.94
ABA 27.94
ABB 33.39
ABC 84.35
ABD 28.58
ABG 30.06
ABK 96.1
ABM 41.63
ABR 34.45
ABT 93.37
ABV 107.5
ABVT 100.0
ABX 54.74
ACC 37.0
ACE 104.0
ACF 64.9
ACG 12.63
ACH 111.6
ACI 112.89
ACL 178.56
ACM 38.25
ACN 44.03
ACO 42.7
ACS 109.55
ACV 65.32
ADC 37.7
ADI 185.5
ADM 48.95
ADP 84.31
ADS 80.79
ADX 40.56
ADY 44.0
AEA 23.94

```


PART 7 – Programming Assignment

Write one MapReduce program using each of the classes that extend `FileInputFormat<k,v>`

(`CombineFileInputFormat`, `FixedLengthInputFormat`, `KeyValueTextInputFormat`, `NLineInputFormat`, `SequenceFileInputFormat`, `TextInputFormat`)

<http://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/lib/input/FileInputFormat.html>

You could use any input file of your choice. The size of the input files is not important. The MR programs could simply do counting, or any other analysis you choose.

Ans--

1> TextInputFormat

Code in package `textInputFormat`

Output

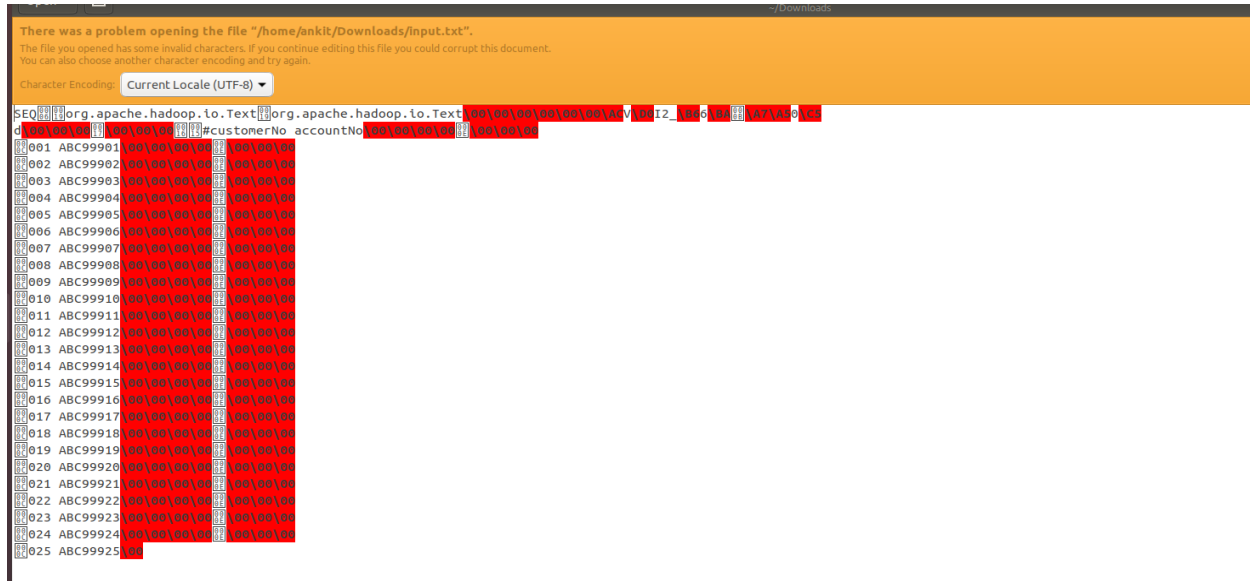
```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyFromLocal ~/Downloads/ebook.txt /myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 1 items
-rw-r--r-- 1 ankit supergroup 288326 2019-06-11 17:35 /myEbooks/ebook.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q7.jar con.hadoop.assignment3_q7.textInputFormat.TextMainMR /myEbooks/ebook.txt /Assignment3_Q7Output/A_TextInputFo
rmat
19/06/11 17:37:21 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/11 17:37:22 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/11 17:37:23 INFO input.FileInputFormat: Total input files to process : 1
19/06/11 17:37:23 INFO mapreduce.JobSubmitter: number of splits:1
19/06/11 17:37:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560279693514_0007
19/06/11 17:37:23 INFO Impl.VarnClientImpl: Submitted application application_1560279693514_0007
19/06/11 17:37:24 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560279693514_0007/
19/06/11 17:37:24 INFO mapreduce.Job: Running job: job_1560279693514_0007
19/06/11 17:37:35 INFO mapreduce.Job: Job job_1560279693514_0007 running in uber mode : false
19/06/11 17:37:35 INFO mapreduce.Job: map 0% reduce 0%
19/06/11 17:38:00 INFO mapreduce.Job: map 100% reduce 0%
19/06/11 17:38:08 INFO mapreduce.Job: map 100% reduce 100%
19/06/11 17:38:11 INFO mapreduce.Job: Job job_1560279693514_0007 completed successfully
19/06/11 17:38:11 INFO mapreduce.Job: Counters: 49
```

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -copyFromLocal ~/Downloads/ebook.txt /myEbooks
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /myEbooks
Found 1 items
-rw-r--r-- 1 ankit supergroup 288326 2019-06-11 17:35 /myEbooks/ebook.txt
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q7.jar con.hadoop.assignment3_q7.textInputFormat.TextMainMR /myEbooks/ebook.txt /Assignment3_Q7Output/A_TextInputFo
rmat
19/06/11 17:37:21 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/11 17:37:22 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/11 17:37:23 INFO input.FileInputFormat: Total input files to process : 1
19/06/11 17:37:23 INFO mapreduce.JobSubmitter: number of splits:1
19/06/11 17:37:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560279693514_0007
19/06/11 17:37:23 INFO Impl.VarnClientImpl: Submitted application application_1560279693514_0007
19/06/11 17:37:24 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560279693514_0007/
19/06/11 17:37:24 INFO mapreduce.Job: Running job: job_1560279693514_0007
19/06/11 17:37:35 INFO mapreduce.Job: Job job_1560279693514_0007 running in uber mode : false
19/06/11 17:37:35 INFO mapreduce.Job: map 0% reduce 0%
19/06/11 17:38:00 INFO mapreduce.Job: map 100% reduce 0%
19/06/11 17:38:08 INFO mapreduce.Job: map 100% reduce 100%
19/06/11 17:38:11 INFO mapreduce.Job: Job job_1560279693514_0007 completed successfully
19/06/11 17:38:11 INFO mapreduce.Job: Counters: 49
```

2> SequenceFileInputFormat

The code is in package seqFileInputFormat

Input Sequence File-



Output

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q76.jar com.hadoop.assignment3_q7.sequencefileinputformat.SeqMainMR /myEbooks/input.txt /Assignment3_Q7Output/E_SeqLenInputFormat
19/06/12 17:22:59 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/12 17:22:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/12 17:23:00 INFO InputFileInputFormat: Total input files to process : 1
19/06/12 17:23:00 INFO mapreduce.JobSubmitter: Number of splits:1
19/06/12 17:23:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560374278640_0002
19/06/12 17:23:00 INFO Impl.YarnClientImpl: Submitted application application_1560374278640_0002
19/06/12 17:23:00 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560374278640_0002/
19/06/12 17:23:00 INFO mapreduce.Job: Running job: job_1560374278640_0002
19/06/12 17:23:07 INFO mapreduce.Job: Job job_1560374278640_0002 running in uber mode : false
19/06/12 17:23:07 INFO mapreduce.Job:  map 0% reduce 0%
19/06/12 17:23:13 INFO mapreduce.Job:  map 100% reduce 0%
19/06/12 17:23:20 INFO mapreduce.Job:  map 100% reduce 100%
19/06/12 17:23:20 INFO mapreduce.Job: Job job_1560374278640_0002 completed successfully
19/06/12 17:23:20 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=188
    FILE: Number of bytes written=317049
    FILE: Number of read operations=0
```

```
      WRONG_MAP=0
      WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=659
  File Output Format Counters
    Bytes Written=4
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /Assignment3_Q7Output/E_SeqLenInputFormat/part-r-00000
26
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```

3>KeyValueTextInputFormat

Operation- Counting the number of keys

code in package KeyValueInputFormat

Output

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q71.jar com.hadoop.assignment3_q7.KeyValueTextInputFormat.KeyMainMR /myEbooks/kvf.txt /Assignment3_Q70Output/B_KeyVa
lueTextInputFormat
19/06/11 19:04:54 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/11 19:04:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/11 19:04:55 INFO Input.FileInputFormat: Total input files to process : 1
19/06/11 19:04:56 INFO mapreduce.JobSubmitter: number of splits:1
19/06/11 19:04:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560294229318_0001
19/06/11 19:04:57 INFO Impl.YarnClientImpl: Submitted application application_1560294229318_0001
19/06/11 19:04:57 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560294229318_0001/
19/06/11 19:04:57 INFO mapreduce.Job: Running job: job_1560294229318_0001
19/06/11 19:05:05 INFO mapreduce.Job: Job job_1560294229318_0001 running in uber mode : false
19/06/11 19:05:05 INFO mapreduce.Job:  map 0% reduce 0%
19/06/11 19:05:11 INFO mapreduce.Job:  map 100% reduce 0%
19/06/11 19:05:17 INFO mapreduce.Job:  map 100% reduce 100%
19/06/11 19:05:18 INFO mapreduce.Job: Job job_1560294229318_0001 completed successfully
19/06/11 19:05:18 INFO mapreduce.Job: Counters: 49
```

```
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -ls /Assignment3_Q70Output/B_KeyValueTextInputFormat
Found 2 items
-rw-r--r--  1 ankit supergroup          0 2019-06-11 19:05 /Assignment3_Q70Output/B_KeyValueTextInputFormat/_SUCCESS
-rw-r--r--  1 ankit supergroup        16 2019-06-11 19:05 /Assignment3_Q70Output/B_KeyValueTextInputFormat/part-r-000000
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /Assignment3_Q70Output/B_KeyValueTextInputFormat/part-r-000000
A      3
B      3
C      3
d      3
ankit@ankit-VirtualBox:/usr/local/bin/hadoop-2.8.5/bin$
```

5> NLineInputFormat

The code is in package NLineInputFormat

Output

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop jar /home/ankit/Downloads/Q72.jar com.hadoop.assignment3_q7.NLineInputFormat.NLineMainMR /myEbooks/ebook.txt /Assignment3_Q7Output/C_NLineInputFormat
19/06/11 21:31:19 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/06/11 21:31:23 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/06/11 21:31:24 INFO input.FileInputFormat: Total input files to process : 1
19/06/11 21:31:26 INFO mapreduce.JobSubmitter: number of splits:65
19/06/11 21:31:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560294229318_0002
19/06/11 21:31:29 INFO impl.VarnClientImpl: Submitted application application_1560294229318_0002
19/06/11 21:31:31 INFO mapreduce.Job: The url to track the job: http://ankit-VirtualBox:8088/proxy/application_1560294229318_0002/
19/06/11 21:31:31 INFO mapreduce.Job: Running job: job_1560294229318_0002
19/06/11 21:32:12 INFO mapreduce.Job: Job job_1560294229318_0002 running in uber mode : false
19/06/11 21:32:12 INFO mapreduce.Job: map 0% reduce 0%
```

```
19/06/11 21:37:28 INFO mapreduce.Job: map 89% reduce 29%
19/06/11 21:37:34 INFO mapreduce.Job: map 91% reduce 30%
19/06/11 21:37:36 INFO mapreduce.Job: map 92% reduce 30%
19/06/11 21:37:37 INFO mapreduce.Job: map 94% reduce 30%
19/06/11 21:37:40 INFO mapreduce.Job: map 95% reduce 31%
19/06/11 21:37:41 INFO mapreduce.Job: map 97% reduce 31%
19/06/11 21:37:44 INFO mapreduce.Job: map 98% reduce 31%
19/06/11 21:37:45 INFO mapreduce.Job: map 100% reduce 31%
19/06/11 21:37:46 INFO mapreduce.Job: map 100% reduce 100%
19/06/11 21:37:48 INFO mapreduce.Job: Job job_1560294229318_0002 completed successfully
19/06/11 21:37:49 INFO mapreduce.Job: Counters: 51
  File System Counters
    FILE: Number of bytes read=630848
    FILE: Number of bytes written=11701564
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=555980
    HDFS: Number of bytes written=86505
    HDFS: Number of read operations=198
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Failed map tasks=14
    Killed map tasks=1
    Launched map tasks=79
    Launched reduce tasks=1
    Other local map tasks=79
    Total time spent by all maps in occupied slots (ms)=1681823
    Total time spent by all reduces in occupied slots (ms)=161950
    Total time spent by all map tasks (ms)=1681823
    Total time spent by all reduce tasks (ms)=161950
    Total vcore-milliseconds taken by all map tasks=1681823
    Total vcore-milliseconds taken by all reduce tasks=161950
    Total megabyte-milliseconds taken by all map tasks=1722186752
    Total megabyte-milliseconds taken by all reduce tasks=165836800
  Map-Reduce Framework
```

```
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /Assignment3_Q7Output/C_NLineInputFormat/part-r-00000 | head
12810
"Come 1
"Defects," 1
"Economic 1
"HARDWOODS" 1
"Hardwoods," 1
"Information 1
"Knees.]" 1
"Our 1
"Plain 2
cat: Unable to write to output stream.
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$ hadoop fs -cat /Assignment3_Q7Output/C_NLineInputFormat/part-r-00000 | tail
you 55
you!) 1
young 32
younger 2
your 11
youth 1
yucca 1
% 1
%-inch 1
% 1
ankit@ankit-VirtualBox: /usr/local/bin/hadoop-2.8.5/bin$
```