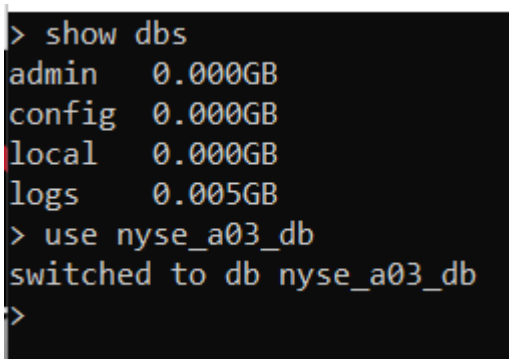
Aniruddha_Tambe_002101113_a03.md

PART 3 - MongoDB indexing

1. Creating the NYSE dataset

```
show dbs
use nyse_a03_db
```

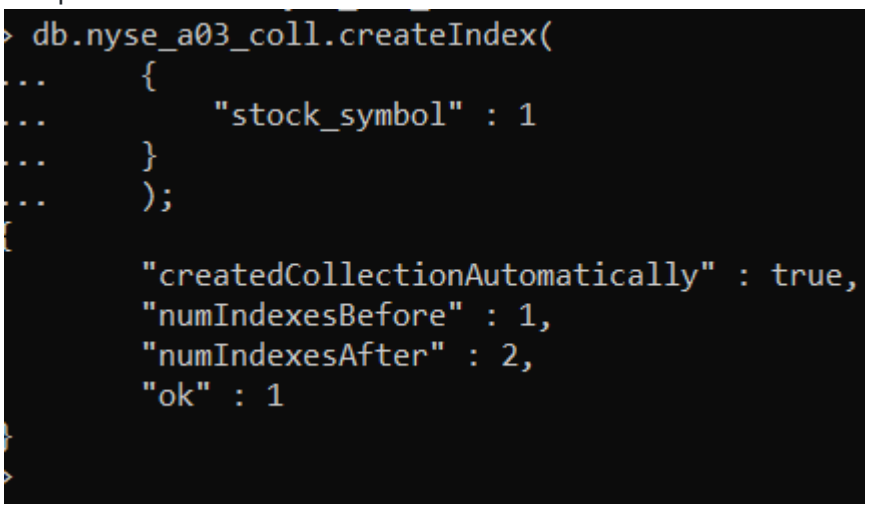
Output:



2. Creating an index on `stock_symbol` attribute

```
db.nyse_a03_coll.createIndex(
    {
        "stock_symbol" : 1
    }
    );
```

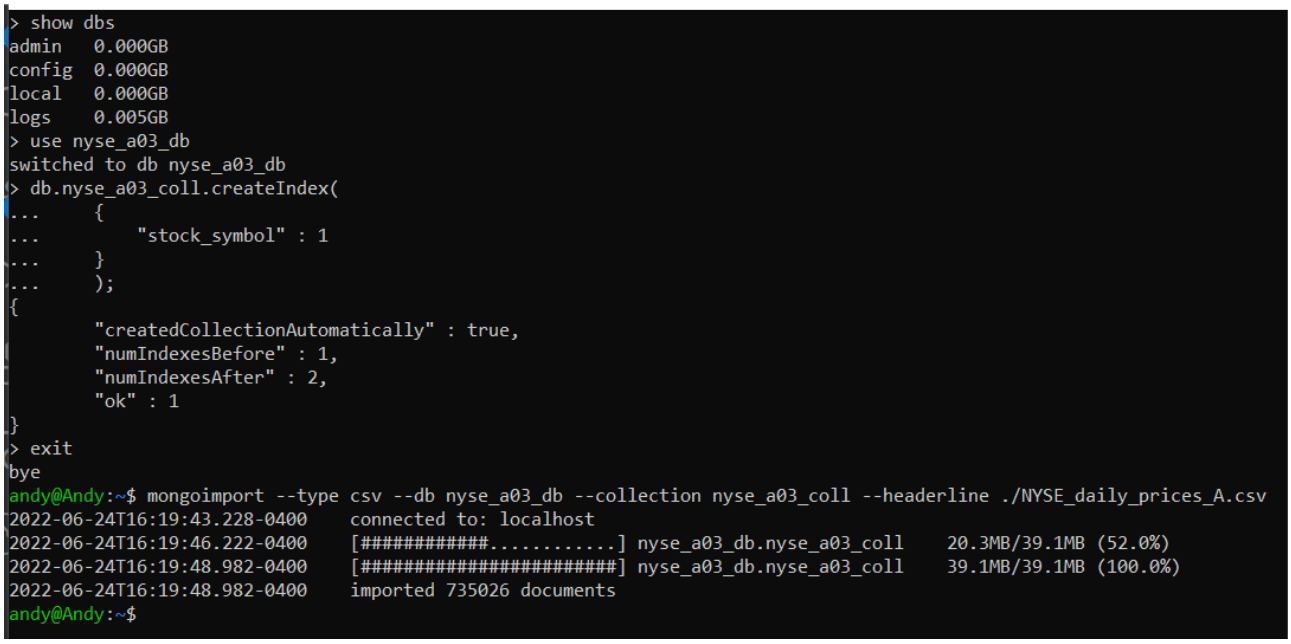
Output:



3. Importing the dataset

```
mongoimport --type csv --db nyse_a03_db --collection nyse_a03_coll --headerline ./NY
```

Output:



4. Check for created indexes

```
use nyse_a03_db
db.nyse_a03_coll.getIndexes();
```

Output:

```
> db.nyse_a03_coll.getIndexes();
[
        {
                "v" : 2,
                "key" : {
                        "_id" : 1
                },
                "name" : "_id_",
                "ns" : "nyse_a03_db.nyse_a03_coll"
        },
        {
                "v" : 2,
                "key" : {
                        "stock_symbol" : 1
                },
                "name" : "stock_symbol_1",
                "ns" : "nyse_a03_db.nyse_a03_coll"
        }
]
>
```

PART 4 - MongoDB Indexing

1. Importing the dataset in another database

Output:

```
andy@Andy:~$ mongoimport --type csv --db nyse_a03_db_copy --collection nyse_a03_coll_copy --headerline ./NYSE_daily_prices_A.csv
2022-06-24T17:22:32.402-0400    connected to: localhost
2022-06-24T17:22:35.400-0400    [############............] nyse_a03_db_copy.nyse_a03_coll_copy  21.1MB/39.1MB (53.9%)
2022-06-24T17:22:37.544-0400    [########################] nyse_a03_db_copy.nyse_a03_coll_copy  39.1MB/39.1MB (100.0%)
2022-06-24T17:22:37.544-0400    imported 735026 documents
andy@Andy:~$
```

2. Creating an index on this database

Output:

```
> use nyse_a03_db_copy
switched to db nyse_a03_db_copy
> db.nyse_a03_coll_copy.createIndex({"stock_symbol":1});
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
>
```

3. Checking for created index

Output:

```
> db.nyse_a03_coll_copy.getIndexes();
[
        {
                "v" : 2,
                "key" : {
                        "_id" : 1
                },
                "name" : "_id_",
                "ns" : "nyse_a03_db_copy.nyse_a03_coll_copy"
        },
        {
                "v" : 2,
                "key" : {
                        "stock_symbol" : 1
                },
                "name" : "stock_symbol_1",
                "ns" : "nyse_a03_db_copy.nyse_a03_coll_copy"
        }
]
>
```

PART 5 - MongoDB Text search (refer to chapter 9).

Q. Write and execute one query to perform each of the followings on any collection of your choice.

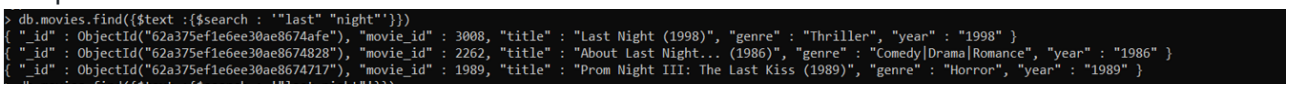
1. Specify and word matches instead of or word matches.

Output:

```
> db.movies.find({$text :{$search : "night"}})
{ "_id" : ObjectId("62a375ef1e6ee30ae8674db5"), "movie_id" : 3696, "title" : "Night of the Creeps (1986)", "genre" : "Comedy|Horror|Sci-Fi", "year" : "1986" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674c93"), "movie_id" : 3405, "title" : "Night to Remember- A (1958)", "genre" : "Action|Drama", "year" : "1958" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674c55"), "movie_id" : 3348, "title" : "Night Visitor- The (1970)", "genre" : "Crime|Thriller", "year" : "1970" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674bd6"), "movie_id" : 3220, "title" : "Night Tide (1961)", "genre" : "Drama", "year" : "1961" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674b63"), "movie_id" : 3112, "title" : "'Night Mother (1986)", "genre" : "Drama", "year" : "1986" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674afe"), "movie_id" : 3008, "title" : "Last Night (1998)", "genre" : "Thriller", "year" : "1998" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674a7e"), "movie_id" : 2878, "title" : "Hell Night (1981)", "genre" : "Horror", "year" : "1981" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674a75"), "movie_id" : 2867, "title" : "Fright Night (1985)", "genre" : "Comedy|Horror", "year" : "1985" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674978"), "movie_id" : 2613, "title" : "Night of the Comet (1984)", "genre" : "Action|Horror|Sci-Fi", "year" : "1984" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674917"), "movie_id" : 2518, "title" : "Night Shift (1982)", "genre" : "Comedy", "year" : "1982" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674841"), "movie_id" : 2296, "title" : "Night at the Roxbury- A (1998)", "genre" : "Comedy", "year" : "1998" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674828"), "movie_id" : 2262, "title" : "About Last Night... (1986)", "genre" : "Comedy|Drama|Romance", "year" : "1986" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674713"), "movie_id" : 1987, "title" : "Prom Night (1980)", "genre" : "Horror", "year" : "1980" }
{ "_id" : ObjectId("62a375ef1e6ee30ae86746ec"), "movie_id" : 1950, "title" : "In the Heat of the Night (1967)", "genre" : "Drama|Mystery", "year" : "1967" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674638"), "movie_id" : 1771, "title" : "Night Flier (1997)", "genre" : "Horror", "year" : "1997" }
{ "_id" : ObjectId("62a375ef1e6ee30ae86745e1"), "movie_id" : 1673, "title" : "Boogie Nights (1997)", "genre" : "Drama", "year" : "1997" }
{ "_id" : ObjectId("62a375ef1e6ee30ae867446c"), "movie_id" : 1279, "title" : "Night on Earth (1991)", "genre" : "Comedy|Drama", "year" : "1991" }
{ "_id" : ObjectId("62a375ef1e6ee30ae86743ff"), "movie_id" : 1152, "title" : "He Walked by Night (1948)", "genre" : "Crime|Film-Noir|Thriller", "year" : "1948" }
{ "_id" : ObjectId("62a375ef1e6ee30ae8674356"), "movie_id" : 994, "title" : "Big Night (1996)", "genre" : "Drama", "year" : "1996" }
{ "_id" : ObjectId("62a375ef1e6ee30ae86742f8"), "movie_id" : 893, "title" : "Mother Night (1996)", "genre" : "Drama", "year" : "1996" }
```

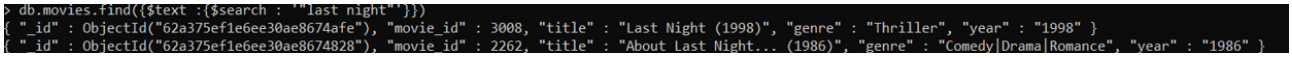
2. Perform exact phrase matches.

Output:



3. Exclude documents with certain words.

Output:



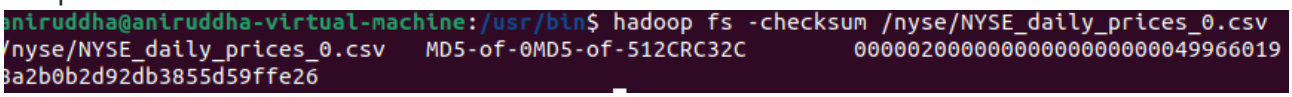
4. Exclude documents with certain phrases.

Output:

PART 6 - Programming Assignment

1. checksum

Output:



2. help

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/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] <localsrc> ... <dst>]
        [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] <path> ...]
        [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] <path> ...]
        [-expunge]
```

3. ls

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - aniruddha supergroup          0 2022-06-21 15:10 /june21
drwxrwxrwx   - aniruddha supergroup          0 2022-06-30 22:06 /nyse
```

4. mkdir

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -mkdir /test
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

5. count

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -count /test
           1            0                  0 /test
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

6. touchz

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -touchz /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

7. cat

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -cat /nyse/NYSE_daily_prices_0.csv
exchange,stock_symbol,date,stock_price_open,stock_price_high,stock_price_low,stock_price_close,s
tock_volume,stock_price_adj_close
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

8. cp

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -cp /test/test_file.txt /nyse/test_file.
txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

9. rm

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -rm /test/test_file.txt
22/06/30 22:46:53 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval =
0 minutes, Emptier interval = 0 minutes.
Deleted /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

10. mv

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -mv /nyse/test_file.txt /test/test_file.
txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

11. du

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -du /nyse
130        /nyse/NYSE_daily_prices_0.csv
130        /nyse/NYSE_daily_prices_1.csv
130        /nyse/NYSE_daily_prices_2.csv
130        /nyse/NYSE_daily_prices_3.csv
130        /nyse/NYSE_daily_prices_4.csv
130        /nyse/NYSE_daily_prices_5.csv
130        /nyse/NYSE_daily_prices_6.csv
```

12. dus

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -dus /nyse
dus: DEPRECATED: Please use 'du -s' instead.
511086927  /nyse
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

13. stat

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -stat /nyse
2022-07-01 02:48:07
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

14. setrep

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -setrep -R 2 /test
Replication 2 set: /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

15. copyFromLocal

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -copyFromLocal /home/aniruddha/Desktop/p
ups.txt /test
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

16. moveFromLocal

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -moveFromLocal /home/aniruddha/Desktop/p
ups.txt /test_dir
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

17. moveToLocal

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -moveToLocal /test_dir/pups.txt /home/an
iruddha/Downloads/demo/
moveToLocal: Option '-moveToLocal' is not implemented yet.
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

Definition: The Hadoop fs shell command moveToLocal moves the file or directory from the Hadoop filesystem to the destination in the local filesystem.

18. -ls -R

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -ls -R /test_dir
-rw-r--r--   1 aniruddha supergroup      27153 2022-06-30 22:55 /test_dir
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

19. df

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -df -h /test_dir
Filesystem               Size     Used  Available  Use%
hdfs://localhost:9000  19.0 G  491.5 M      4.1 G    3%
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

20. get

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -get /test/test_file.txt /home/aniruddha
/Downloads/
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

21. expunge

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -expunge
22/06/30 23:03:19 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval =
0 minutes, Emptier interval = 0 minutes.
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

22. chmod

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -chmod 777 /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

23. usage

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -usage ls
Usage: hadoop fs [generic options] -ls [-d] [-h] [-R] [<path> ...]
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

24. tail

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -tail /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

25. head

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -head /test/test_file.txt
-head: Unknown command
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

Displays first kilobyte of the file to stdout.

26. rmdir

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -mkdir /dir
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -rmdir /dir
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

27. rm -r

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -touchz /test/test_file.txt
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -rm -r /test
```

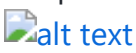
28. rm -r

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -cp /test/test_file.txt /nyse/test_file.
txt
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

29. cp

Output:



30. test

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -test -e /test
aniruddha@aniruddha-virtual-machine:/usr/bin$
```

31. appendToFile

Output:

```
aniruddha@aniruddha-virtual-machine:/usr/bin$ hadoop fs -appendToFile /test/test_file.txt /test_
dir/test_file.txt
```

32. getMerge

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -getmerge /test ~/Mergefile
aniruddha@aniruddha-virtual-machine:~$ cat Mergefile
The Project Gutenberg eBook of My twin puppies, by Edna Groff Deihl

This eBook is for the use of anyone anywhere in the United States and
most other parts of the world 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.org. If you are not located in the United States, you
will have to check the laws of the country where you are located before
using this eBook.

Title: My twin puppies

Author: Edna Groff Deihl

Release Date: June 6, 2022 [eBook #68254]

Language: English
```

33. text

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -text /test/test_file.txt
```

34. getfacl

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -getfacl /test/test_file.txt
# file: /test/test_file.txt
# owner: aniruddha
# group: supergroup
getfacl: The ACL operation has been rejected.  Support for ACLs has been disabled by setting dfs
.namenode.acls.enabled to false.
aniruddha@aniruddha-virtual-machine:~$
```

35. put

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -put /home/aniruddha/Desktop/pups.txt /test/tes
t_files.txt
```

36. getfattr

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -getfattr -d /
# file: /
aniruddha@aniruddha-virtual-machine:~$
```

37. find

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -find / -name test_file.txt
/nyse/test_file.txt
/test/test_file.txt
aniruddha@aniruddha-virtual-machine:~$
```

38. chgrp

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -chgrp -R user /test
aniruddha@aniruddha-virtual-machine:~$
```

39. chown

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -chown 777 /test/test_file.txt
aniruddha@aniruddha-virtual-machine:~$
```

40. setfacl

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -setfacl -b /test/test_file.txt
setfacl: The ACL operation has been rejected.  Support for ACLs has been disabled by setting dfs
.namenode.acls.enabled to false.
aniruddha@aniruddha-virtual-machine:~$
```

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

41. setfattr

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -setfattr -n user.noValue /test/test_file.txt
aniruddha@aniruddha-virtual-machine:~$
```

42. touch

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -touch -a /test/test_file.txt
-touch: Unknown command
aniruddha@aniruddha-virtual-machine:~$
```

Updates the access and modification times of the file specified by the URI to the current time. If the file does not exist, then a zero length file is created at URI with current time as the timestamp of that URI.

43. chgrp

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hadoop fs -concat /test/pups.txt /test/test_file.txt
-concat: Unknown command
aniruddha@aniruddha-virtual-machine:~$
```

44. concat

Output:

Concatenate existing source files into the target file. Target file and source files should be in the same directory.

45. createSnapshot

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hdfs dfsadmin -allowSnapshot /test
hdfs dfs -createSnapshot /test createsnapshot
Allowing snaphot on /test succeeded
Created snapshot /test/.snapshot/createsnapshot
aniruddha@aniruddha-virtual-machine:~$
```

46. renameSnapshot

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hdfs dfs -ls /test/.snapshot
Found 1 items
drwxr-xr-x   - aniruddha user          0 2022-06-30 23:36 /test/.snapshot/renamesnapshot
aniruddha@aniruddha-virtual-machine:~$
```

47. deleteSnapshot

Output:

```
aniruddha@aniruddha-virtual-machine:~$ hdfs dfs -deleteSnapshot /test renamesnapshot
aniruddha@aniruddha-virtual-machine:~$ hdfs dfs -ls /test/.snapshot
aniruddha@aniruddha-virtual-machine:~$
```

PART 7 - HDFS Command Assignment

Code:

```
package com.info7250.mongodb.assignment;

import java.io.IOException;
import java.net.URI;
import java.net.URISyntaxException;
```

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FSDataInputStream;
import org.apache.hadoop.fs.FSDataOutputStream;
import org.apache.hadoop.fs.FileStatus;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;

public class PutMerge {

public static void main(String[] args) throws IOException, URISyntaxException {

    Configuration conf = new Configuration();
    FileSystem hdfs = FileSystem.get(new URI("hdfs://localhost:9000"),conf);
    //FileSystem hdfs = FileSystem.get(new URI("hdfs://localhost:9000"),conf)
    FileSystem local = FileSystem.getLocal(conf);

    Path inputDir = new Path("/home/aniruddha/Downloads/nyse/NYSE/");
    Path hdfsFile = new Path("/nyse/");

    //FileSystem fs = FileSystem.get(new URI(<url:port>), configuration);
    //Path filePath = new Path(<path/to/file>);

    //System.out.println(inputDir);
    //System.out.println(hdfsFile);

    try {
        FileStatus[] inputFiles = local.listStatus(inputDir);
        //System.out.println(inputFiles);
        FSDataOutputStream out = hdfs.create(hdfsFile);
        //System.out.println(out);

        for (int i = 0; i < inputFiles.length; i++) {
            System.out.println(inputFiles[i].getPath().getName());
            FSDataInputStream in = local.open(inputFiles[i].getPath());
            byte buffer[] = new byte[256];
            int bytesRead = 0;
            while ((bytesRead = in.read(buffer)) > 0) {
                out.write(buffer, 0, bytesRead);
            }
            in.close();
        }
        out.close();
    } catch (Exception e) {
        //e.printStackTrace();

    }
}
}
```

Command:

```
mvn compile exec:java -Dexec.mainClass="com.info7250.mongodb.assignment.PutMerge" -D
```

Output: