Part – 3: MongoDB Indexing

Use the NYSE dataset to declare your indexes before putting your application into production.

```
nyse_hw3> use nysedb
switched to db nysedb
nysedb> db.createCollection("nysecoll")
{ ok: 1 }
nysedb> db.nysecoll.createIndex({"stock_symbol":1})
stock_symbol_1
nysedb> db.nysecoll.getIndexes()
[
  { v: 2, key: { _id: 1 }, name: '_id_' },
  { v: 2, key: { stock_symbol: 1 }, name: 'stock_symbol_1' }
]
nysedb>
```

Bat file to import:

```
#!/bin/bash

for dataFileSuffix in {A..Z}

do

file_loc="C:/Program Files/MongoDB/Server/5.0/NYSE/NYSE_daily_prices_$dataFileSuffix.csv"

        echo $file_loc

        mongoimport --db=nysedb --type=csv --collection=nysecoll --headerline --file="$file_loc"

done
```

Import using bat file:

```
MINGW64:/c/Users/zeeni/OneDrive/Documents/NEU/Sem 3/Big Data/HW3

zeeni@Zeenia MINGW64 ~/OneDrive/Documents/NEU/Sem 3/Big Data/HW3
$ bash import_hw3.bat
C:/Program Files/MongoDB/Server/5.0/NYSE/NYSE_daily_prices_A.csv
2022-10-18T11:02:24.096-0400    connected to: mongodb://localhost/
2022-10-18T11:02:27.097-0400    [######..................] nysedb.nysecoll     1
0.6MB/39.1MB (27.0%)
2022-10-18T11:02:30.097-0400    [#############...........] nysedb.nysecoll     2
1.8MB/39.1MB (55.8%)
2022-10-18T11:02:33.096-0400    [####################....] nysedb.nysecoll     3
3.5MB/39.1MB (85.8%)
2022-10-18T11:02:34.535-0400    [########################] nysedb.nysecoll     3
9.1MB/39.1MB (100.0%)
2022-10-18T11:02:34.535-0400    735026 document(s) imported successfully. 0 docu
ment(s) failed to import.
```

```
nysedb> db.nysecoll.getIndexes()
[
  { v: 2, key: { _id: 1 }, name: '_id_' },
  { v: 2, key: { stock_symbol: 1 }, name: 'stock_symbol_1' }
]
```

Part-4: 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.

```
nysedb> db.nysecoll.createIndex({"date":1})
date_1
nysedb> db.nysecoll.getIndexes()
[
  { v: 2, key: { _id: 1 }, name: '_id_' },
  { v: 2, key: { stock_symbol: 1 }, name: 'stock_symbol_1' },
  { v: 2, key: { date: 1 }, name: 'date_1' }
]
nysedb>
```

Part-5: MongoDB Text Search

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

```
books> db.createCollection("bookscoll")
{ ok: 1 }
books> db.bookscoll.insertMany(
... [
... {_id:1, name:"You Can Win", desc:"Win and Success"},
... {_id:2, name:"Playing to Win", desc:"Win and Strategy"},
... {_id:3, name:"Hooponopono", desc:"Hawaiian Philosophy"},
... {_id:4, name:"Ikigai", desc:"Japanese Philosophy"},
... {_id:5, name:"Supee-Accelerated Living", desc:"Higher Living"}])
{
  acknowledged: true,
  insertedIds: { '0': 1, '1': 2, '2': 3, '3': 4, '4': 5 }
}
books> db.bookscoll.createIndex({name:"text", desc:"text"})
name_text_desc_text
```

```
books> db.bookscoll.find()
[
  { _id: 1, name: 'You Can Win', desc: 'Win and Success' },
  { _id: 2, name: 'Playing to Win', desc: 'Win and Strategy' },
  { _id: 3, name: 'Hooponopono', desc: 'Hawaiian Philosophy' },
  { _id: 4, name: 'Ikigai', desc: 'Japanese Philosophy' },
  { _id: 5, name: 'Supee-Accelerated Living', desc: 'Higher Living' },
  { _id: 6, name: 'How to Win', desc: 'Win' },
  { _id: 7, name: 'What Drives Winning', desc: 'Strategy' },
  {
    _id: 8,
    name: 'What Dives Winning Environments',
    desc: 'Strategy and Win'
  },
  { _id: 9, name: 'Drives Winning', desc: 'Strategy drive' },
  { _id: 10, name: 'What Drives Winning Teams', desc: 'Strategy Win' }
]
```

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

```
books> db.bookscoll.find({$text:{$search:"Win"}})
[
  { _id: 2, name: 'Playing to Win', desc: 'Win and Strategy' },
  { _id: 1, name: 'You Can Win', desc: 'Win and Success' }
]
```

2. Perform exact phrase matches.

```
books> db.bookscoll.find({$text:{$search:"\"to Win\""}})
[
  { _id: 6, name: 'How to Win', desc: 'Win' },
  { _id: 2, name: 'Playing to Win', desc: 'Win and Strategy' }
]
```

3. Exclude documents with certain words.

```
books> db.bookscoll.find({$text:{$search:'"What Drives"'}})
[
  { _id: 7, name: 'What Drives Winning', desc: 'Strategy' },
  { _id: 10, name: 'What Drives Winning Teams', desc: 'Strategy Win' }
]
```

4. Exclude documents with certain phrases.

```
books> db.bookscoll.find({$text:{$search:"Drives -\"What Drives\""}}, {_id:0})
[ { name: 'Drives Winning', desc: 'Strategy drive' } ]
```

Part-6: Programming Assignment

Hadoop Commands

- **mkdir:**

```
[cloudera@quickstart ~]$ cd Desktop/
[cloudera@quickstart Desktop]$ mkdir shared
[cloudera@quickstart Desktop]$ su
```

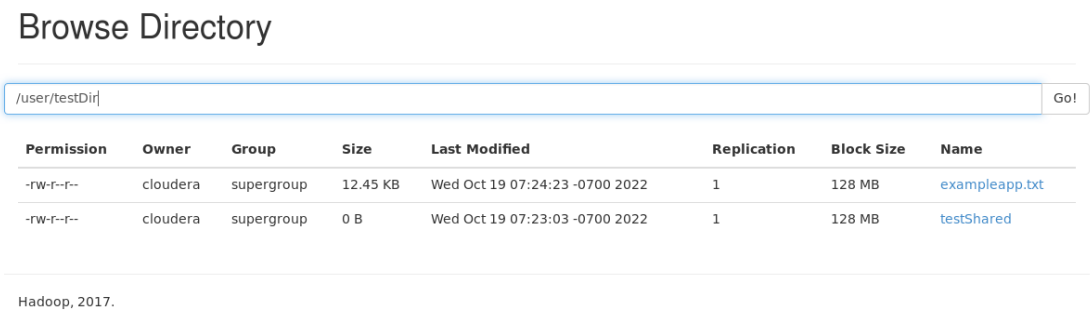
- **copyFromLocal:**

```
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/shared
/user/cloudera/shared1
[cloudera@quickstart ~]$ hadoop fs -ls ^C
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/shared1
Found 2 items
-rw-r--r--   1 cloudera cloudera   40990992 2022-10-19 07:15 /user/cloudera/shar
ed1/NYSE_daily_prices_A.csv
-rw-r--r--   1 cloudera cloudera      12753 2022-10-19 07:15 /user/cloudera/shar
ed1/example.docx
```

- **appendToFile:**

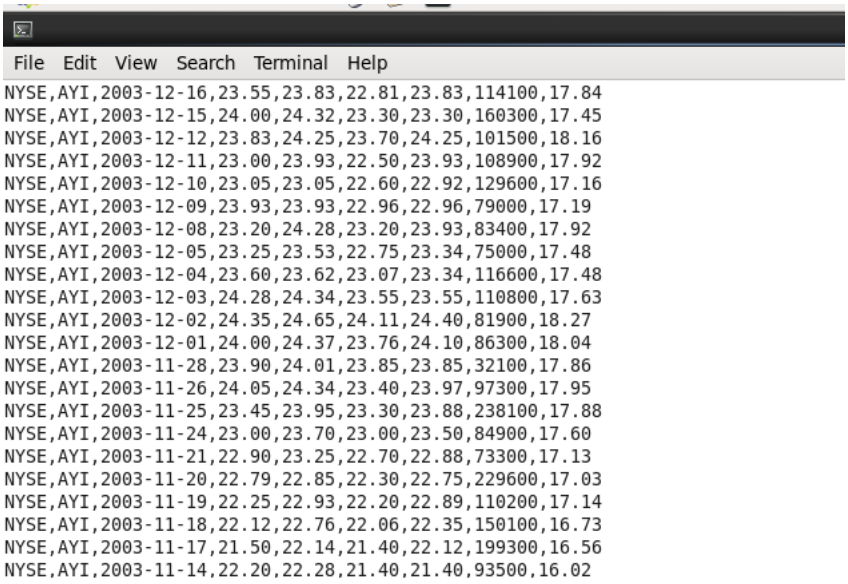
Append single src, or multiple srcs from local file system to the destination file system. Also reads input from stdin and appends to destination file system.

```
[cloudera@quickstart ~]$ hadoop fs -appendToFile /home/cloudera/Desktop/shared/example.docx  /user/testDir/exampleapp.txt
[cloudera@quickstart ~]$ hadoop fs -ls /user/testDir
Found 2 items
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:24 /user/testDir/exampleapp.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
[cloudera@quickstart ~]$
```



- **cat:** Copies source paths to stdout.

Hadoop fs -cat /testDrive/test.csv



- **checksum:** Returns the checksum information of a file.

```
.[cloudera@quickstart ~]$ hadoop fs -checksum /user/testDir/exampleapp.txt
/user/testDir/exampleapp.txt    MD5-of-0MD5-of-512CRC32C        000002000000000000000000fce857b8ad3482fc7d7d1a5a83d6226
[cloudera@quickstart ~]$
```

- **chgrp:** Change group association of files. The user must be the owner of files, or else a super-user.

```
[root@quickstart workspace]# hadoop fs -ls /testDrive
Found 2 items
-rw-r--r--   1 root supergroup          0 2022-10-18 11:00 /testDrive/sample.txt
-rw-r--r--   1 root supergroup   40990992 2022-10-18 10:50 /testDrive/test.csv
[root@quickstart workspace]#
```

```
[root@quickstart workspace]# hadoop fs -chgrp root /testDrive/sample.txt
[root@quickstart workspace]# hadoop fs -ls /testDrive
Found 2 items
-rw-r--r--   1 root root                0 2022-10-18 11:00 /testDrive/sample.txt
-rw-r--r--   1 root supergroup   40990992 2022-10-18 10:50 /testDrive/test.csv
```

- **chmod:** Change the permissions of files. With -R, make the change recursively through the directory structure. The user must be the owner of the file, or else a super-user.

```
[cloudera@quickstart ~]$ hadoop fs -chmod 777 /user/cloudera/shared1/example.docx
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/shared1/
Found 2 items
-rw-r--r--   1 cloudera cloudera   40990992 2022-10-19 07:15 /user/cloudera/shared1/NYSE_daily_prices_A.csv
-rwxrwxrwx   1 cloudera cloudera      12753 2022-10-19 07:15 /user/cloudera/shared1/example.docx
[cloudera@quickstart ~]$
```

Browsing HDFS... | cloudera | [cloudera-mana... | cloudera@quic... | cloudera@quic...

- **chown:** Change the owner of files. The user must be a super-user.
  Hadoop fs -chown 777 /user/cloudera/shared1/example.docx

```
[cloudera@quickstart ~]$ hadoop fs -chown 777 /user/cloudera/shared1/example.docx
```

- **ls:**

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/shared1/
Found 2 items
-rw-r--r--   1 cloudera cloudera   40990992 2022-10-19 07:15 /user/cloudera/shared1/NYSE_daily_prices_A.csv
-rwxrwxrwx   1 cloudera cloudera      12753 2022-10-19 07:15 /user/cloudera/shared1/example.docx
```

- **count:** Count the number of directories, files and bytes under the paths that match the specified file pattern.

```
[cloudera@quickstart ~]$ hadoop fs -count /user/cloudera/shared1
           1            2           41003745 /user/cloudera/shared1
[cloudera@quickstart ~]$
```

- **touchz:** Create a file of zero length. An error is returned if the file exists with non-zero length.

```
[cloudera@quickstart ~]$ hadoop fs -touchz /user/testDir/test.txt
[cloudera@quickstart ~]$ hadoop fs -ls /user/testDir
Found 3 items
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:24 /user/testDir/exampleapp.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:48 /user/testDir/test.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
[cloudera@quickstart ~]$
```

- **cp:** Copy files from source to destination. This command allows multiple sources as well in which case the destination must be a directory.

```
[cloudera@quickstart ~]$ hadoop fs -cp /user/testDir/exampleapp.txt /user/testDir/testcopy.txt
[cloudera@quickstart ~]$ hadoop fs -ls /user/testDir
Found 4 items
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:24 /user/testDir/exampleapp.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:48 /user/testDir/test.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:50 /user/testDir/testcopy.txt
[cloudera@quickstart ~]$
```

- **rm:** Delete files specified as args. If trash is enabled, file system instead moves the deleted file to a trash directory

```
[cloudera@quickstart ~]$ hadoop fs -rm /user/testDir/testcopy.txt
Deleted /user/testDir/testcopy.txt
[cloudera@quickstart ~]$ hadoop fs -ls /user/testDir
Found 3 items
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:24 /user/testDir/exampleapp.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:48 /user/testDir/test.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
[cloudera@quickstart ~]$
```

- **mv:** Moves files from source to destination. This command allows multiple sources as well in which case the destination needs to be a directory. Moving files across file systems is not permitted.

```
[cloudera@quickstart ~]$ hadoop fs -mv /user/cloudera/shared1/example.docx /user/testDir/testmove.txt
[cloudera@quickstart ~]$ hadoop fs -ls /user/testDir
Found 4 items
-rw-r--r--   1 cloudera supergroup      12753 2022-10-19 07:24 /user/testDir/exampleapp.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:48 /user/testDir/test.txt
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
-rwxrwxrwx   1 cloudera cloudera        12753 2022-10-19 07:15 /user/testDir/testmove.txt
[cloudera@quickstart ~]$
```

- **du:** Displays sizes of files and directories contained in the given directory or the length of a file in case its just a file.

```
[cloudera@quickstart ~]$ hadoop fs -du /user/testDir
12753  12753  /user/testDir/exampleapp.txt
0      0      /user/testDir/test.txt
0      0      /user/testDir/testShared
12753  12753  /user/testDir/testmove.txt
[cloudera@quickstart ~]$
```

- **dus:** Displays a summary of file lengths.

```
[cloudera@quickstart ~]$ hadoop fs -dus /user/testDir
dus: DEPRECATED: Please use 'du -s' instead.
25506  25506  /user/testDir
[cloudera@quickstart ~]$
```

- **help:** Return usage output.

```
[cloudera@quickstart ~]$ 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] [-v] [-x] <path> ...]
        [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
```

- **stat:** Print statistics about the file/directory at <path> in the specified format.

```
[cloudera@quickstart ~]$ hadoop fs -stat /user/testDir
2022-10-19 14:53:51
[cloudera@quickstart ~]$
```

- **tail:** Displays last kilobyte of the file to stdout.

```
-rw-r--r--   1 cloudera supergroup          0 2022-10-19 07:23 /user/testDir/testShared
-rwxrwxrwx   1 cloudera cloudera        12753 2022-10-19 07:15 /user/testDir/testmove.txt
[cloudera@quickstart ~]$ hadoop fs -tail /user/testDir/testmove.txt
```

- **text:** Takes a source file and outputs the file in text format. The allowed formats are zip and TextRecordInputStream.

```
[cloudera@quickstart ~]$ hadoop fs -text /user/testDir/testmove.txt
```

- **truncate:** Truncate all files that match the specified file pattern to the specified length.

```
[cloudera@quickstart ~]$ ./hadoop fs -truncate 1 /user/testDir/exampleapp.txt
```

- **test:**

```
-concat: Unknown command
[cloudera@quickstart ~]$ hadoop fs -test -d /user/testDir
```

- **expunge:** Permanently delete files in checkpoints older than the retention threshold from trash directory, and create new checkpoint.

  Command: hadoop fs -expunge

- **find:** Finds all files that match the specified expression and applies selected actions to them. If no *path* is specified then defaults to the current working directory. If no expression is specified then defaults to -print.

  Command: hadoop fs -find / -name test -print

- **get:** Copy files to the local file system.

```
[cloudera@quickstart ~]$ hadoop fs -get /user/testDir/
[cloudera@quickstart ~]$ hadoop fs -get /user/testDir/
get: `testDir/exampleapp.txt': File exists
get: `testDir/test.txt': File exists
get: `testDir/testShared': File exists
get: `testDir/testmove.txt': File exists
[cloudera@quickstart ~]$
```

- **getfacl:** Displays the Access Control Lists (ACLs) of files and directories.

```
[cloudera@quickstart ~]$ hadoop fs -getfacl -R /user/testDir
# file: /user/testDir
# owner: cloudera
# group: supergroup
getfacl: The ACL operation has been rejected.  Support for ACLs has been disabled by setting dfs.namenode.acls.enabled to false.
[cloudera@quickstart ~]$
```

Browsing HDFS... | cloudera | [cloudera-mana... | cloudera@quic... | cloudera@quic...

- **getfattr:** Displays the extended attribute names and values (if any) for a file or directory.

```
[cloudera@quickstart ~]$ hadoop fs -getfattr -d /user/testDir
# file: /user/testDir
[cloudera@quickstart ~]$
```

Browsing HDFS... | cloudera | [cloudera-mana... | cloudera@quic... | cloudera@quic...

Part-7: Programming Assignment

Copy all the files from (A to Z) from NYSE dataset into HDFS.

Step 1: Created local directory shared_nyse and used ‘mount’ command

```
[root@quickstart Desktop]# mkdir shared_nyse
[root@quickstart Desktop]# su
[root@quickstart Desktop]# mount -t vboxsf shared_nyse shared_nyse
[root@quickstart Desktop]#
```

Step 2: Used ‘copyFromLocal’ command to copy files

```
/shar[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/shared_nyse /user/cloudera/shared_nyse
[cloudera@quickstart ~]$ hadoop fs ls /user/cloudera/shared_nyse
ls: Unknown command
Did you mean -ls?  This command begins with a dash.
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/shared_nyse
Found 26 items
-rw-r--r--   1 cloudera cloudera   40990992 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_A.csv
-rw-r--r--   1 cloudera cloudera   32034760 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_B.csv
-rw-r--r--   1 cloudera cloudera   45790256 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_C.csv
-rw-r--r--   1 cloudera cloudera   19234471 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_D.csv
-rw-r--r--   1 cloudera cloudera   22104043 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_E.csv
-rw-r--r--   1 cloudera cloudera   17387253 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_F.csv
-rw-r--r--   1 cloudera cloudera   22608522 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_G.csv
-rw-r--r--   1 cloudera cloudera   23127143 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_H.csv
-rw-r--r--   1 cloudera cloudera   20680033 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_I.csv
-rw-r--r--   1 cloudera cloudera    9537527 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_J.csv
-rw-r--r--   1 cloudera cloudera   14782892 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_K.csv
-rw-r--r--   1 cloudera cloudera   12958785 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_L.csv
-rw-r--r--   1 cloudera cloudera   38124545 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_M.csv
-rw-r--r--   1 cloudera cloudera   31488945 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_N.csv
-rw-r--r--   1 cloudera cloudera    8865718 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_O.csv
-rw-r--r--   1 cloudera cloudera   31943478 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_P.csv
-rw-r--r--   1 cloudera cloudera     190989 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_Q.csv
-rw-r--r--   1 cloudera cloudera   16808595 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_R.csv
-rw-r--r--   1 cloudera cloudera   31852353 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_S.csv
-rw-r--r--   1 cloudera cloudera   28754690 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_T.csv
-rw-r--r--   1 cloudera cloudera    9951590 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_U.csv
-rw-r--r--   1 cloudera cloudera    9503196 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_V.csv
-rw-r--r--   1 cloudera cloudera   15972013 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_W.csv
-rw-r--r--   1 cloudera cloudera    3613198 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_X.csv
-rw-r--r--   1 cloudera cloudera     686216 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_Y.csv
-rw-r--r--   1 cloudera cloudera    2093424 2022-10-19 09:26 /user/cloudera/shared_nyse/NYSE_daily_prices_Z.csv
[cloudera@quickstart ~]$
```