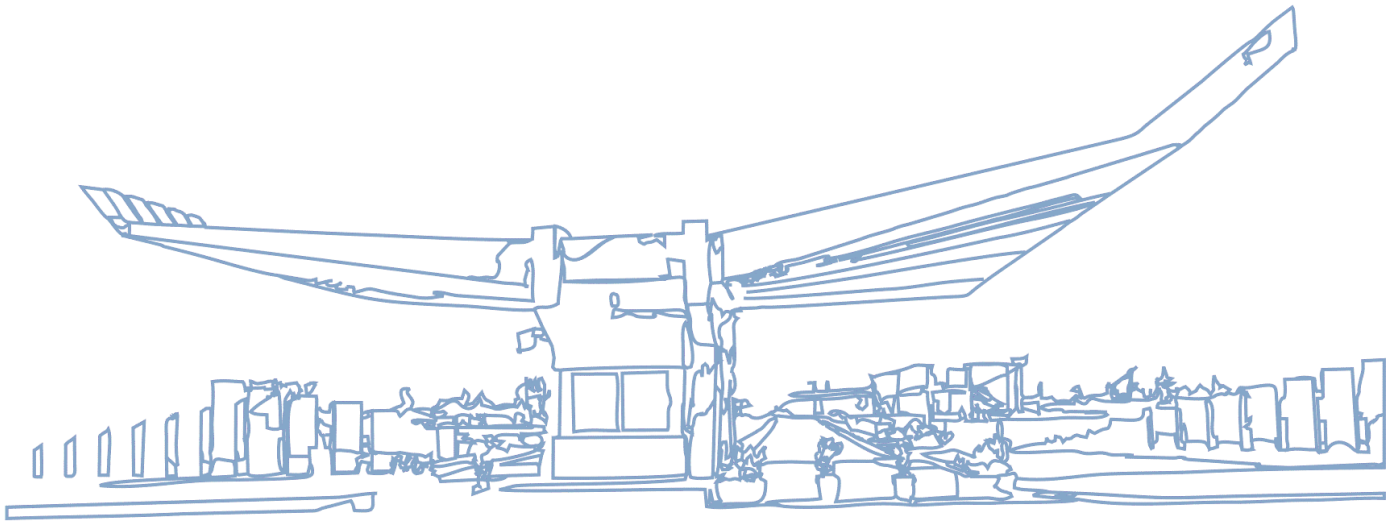
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Assignment 3: MapReduce



PREPARED:
Baftjar TABAKU

26.04.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

Assignment tasks and notes

Use the above file (NYSE.csv) as a source and complete these tasks. Consider only records for the last 10 years and with a volume greater than 250,000:

1. Modify an existing or write a new MapReduce code
2. Compile the program as a jar file
3. Upload the input file(s) into the Hadoop AWS cluster
4. Execute the program and display the results

The deliverables for each task are:

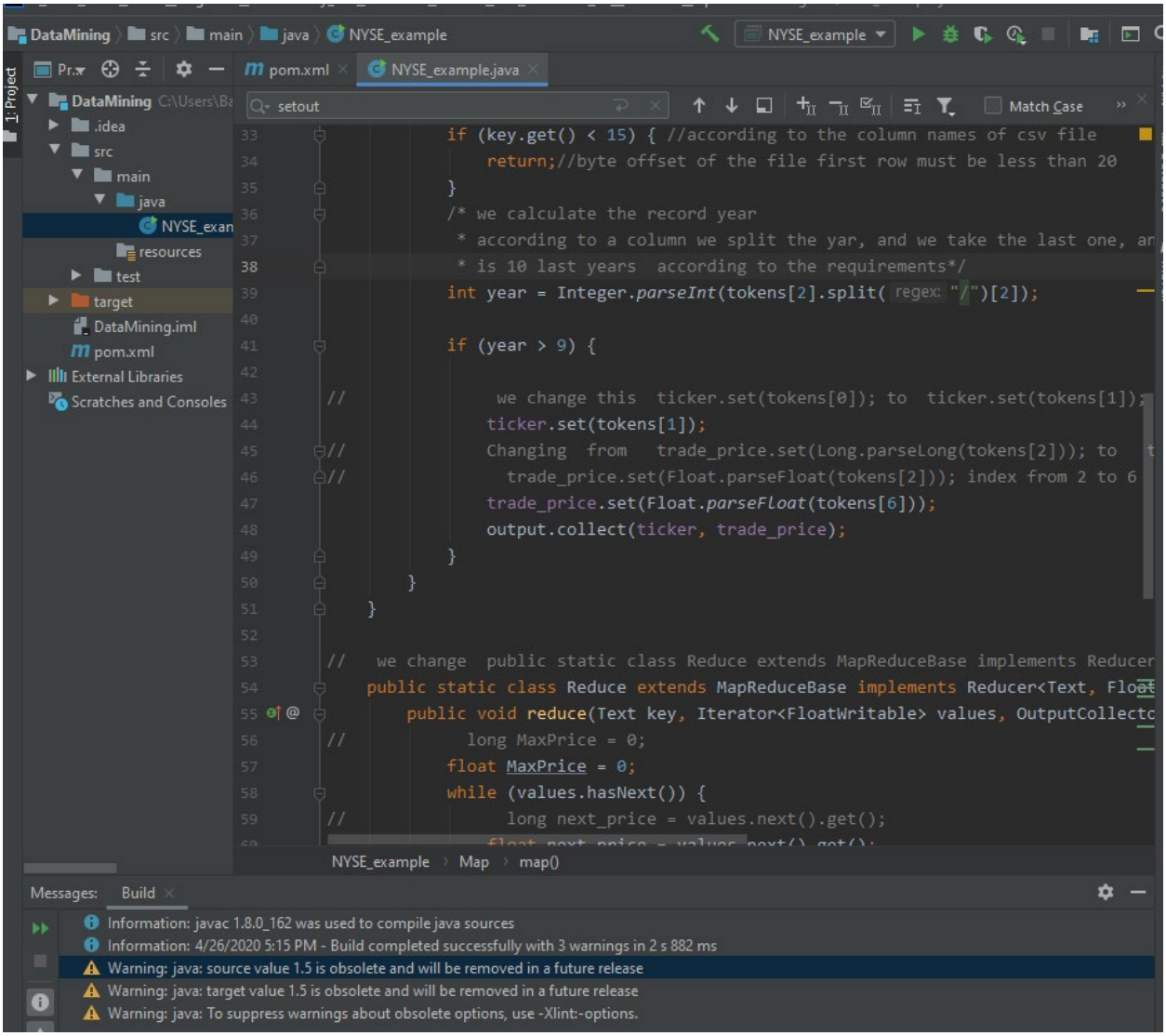
• Modified Java code
• List of Linux and HDFS commands that are used to execute the project
• Output results in a text file

Write a Map-Reduce program to find out maximum trading price for each stock

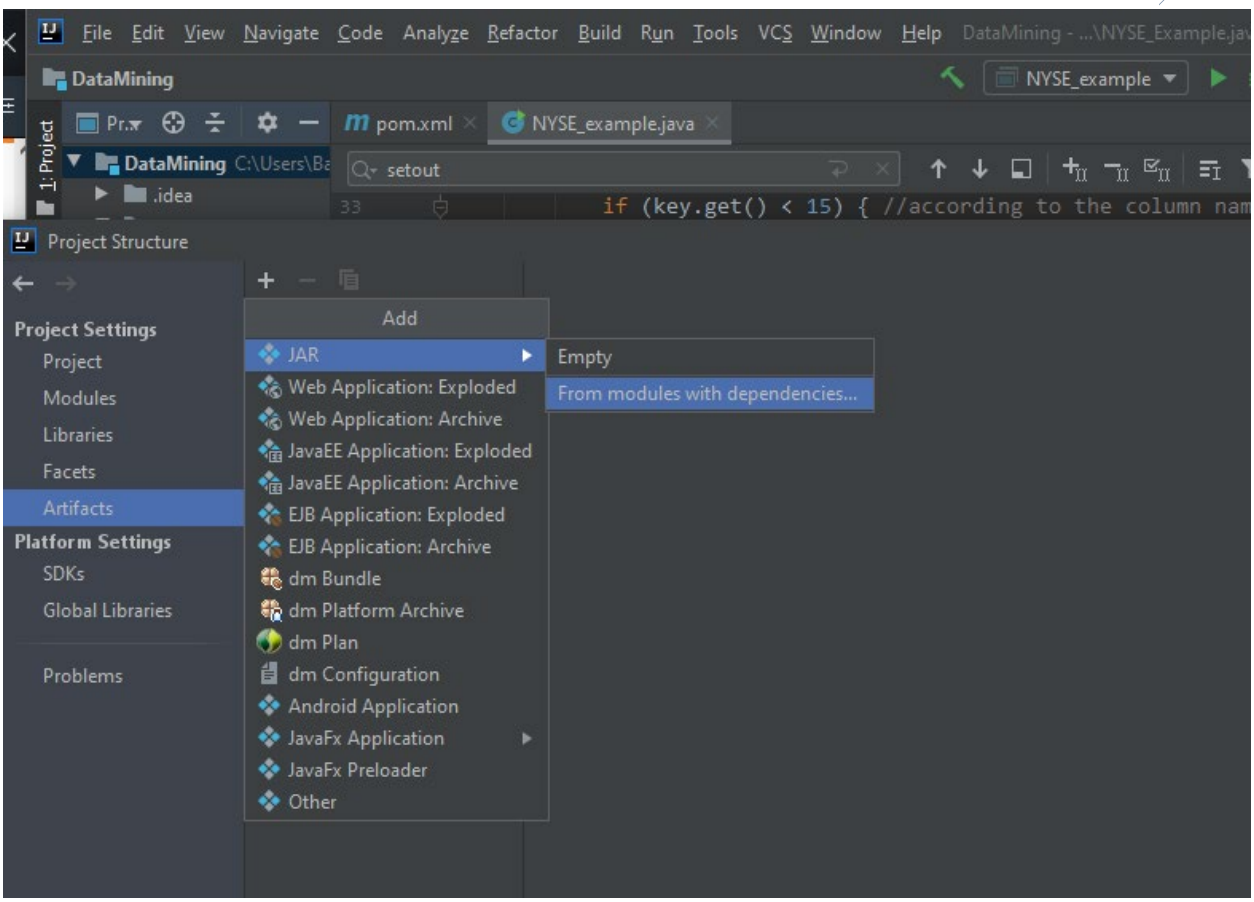
Modify the MapReduce Java code to find out highest price for each stock

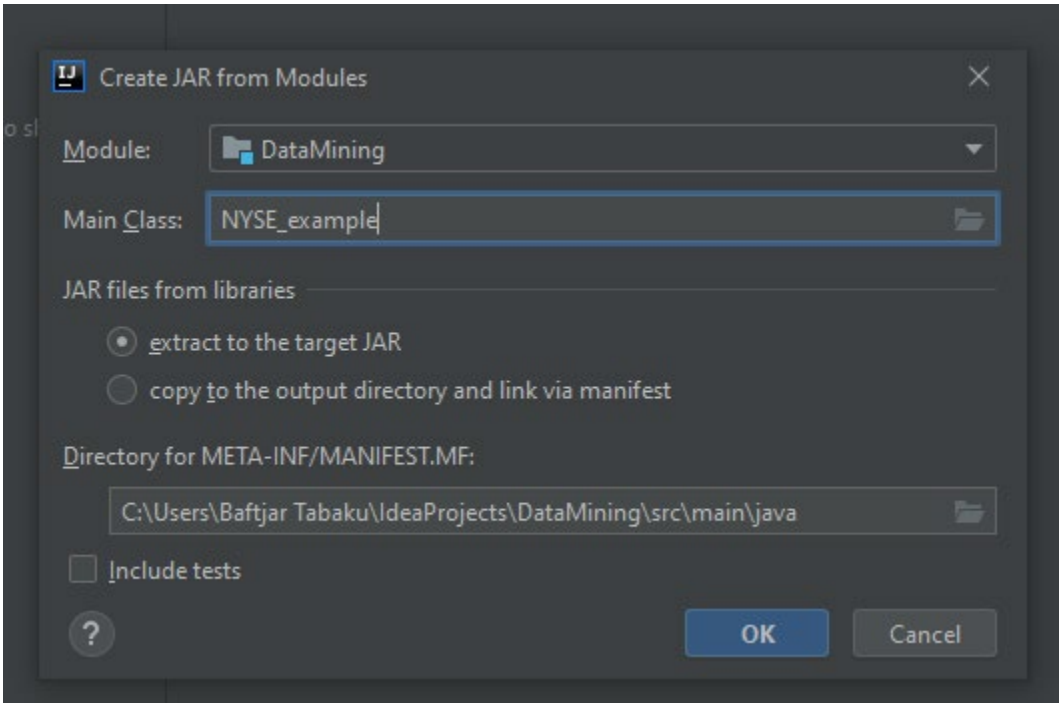
Assignment tasks and notes

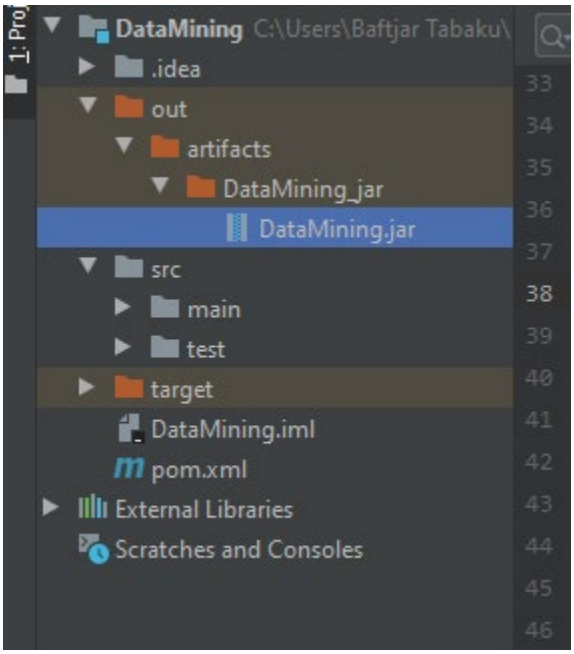
After modifying the following java code, that also is included and every change is commented, build was successfully.



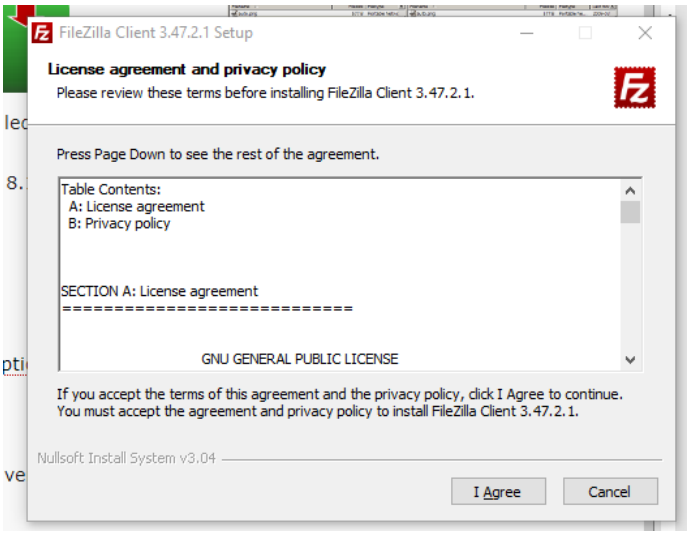
After this compiling the program as a jar file, as shown below

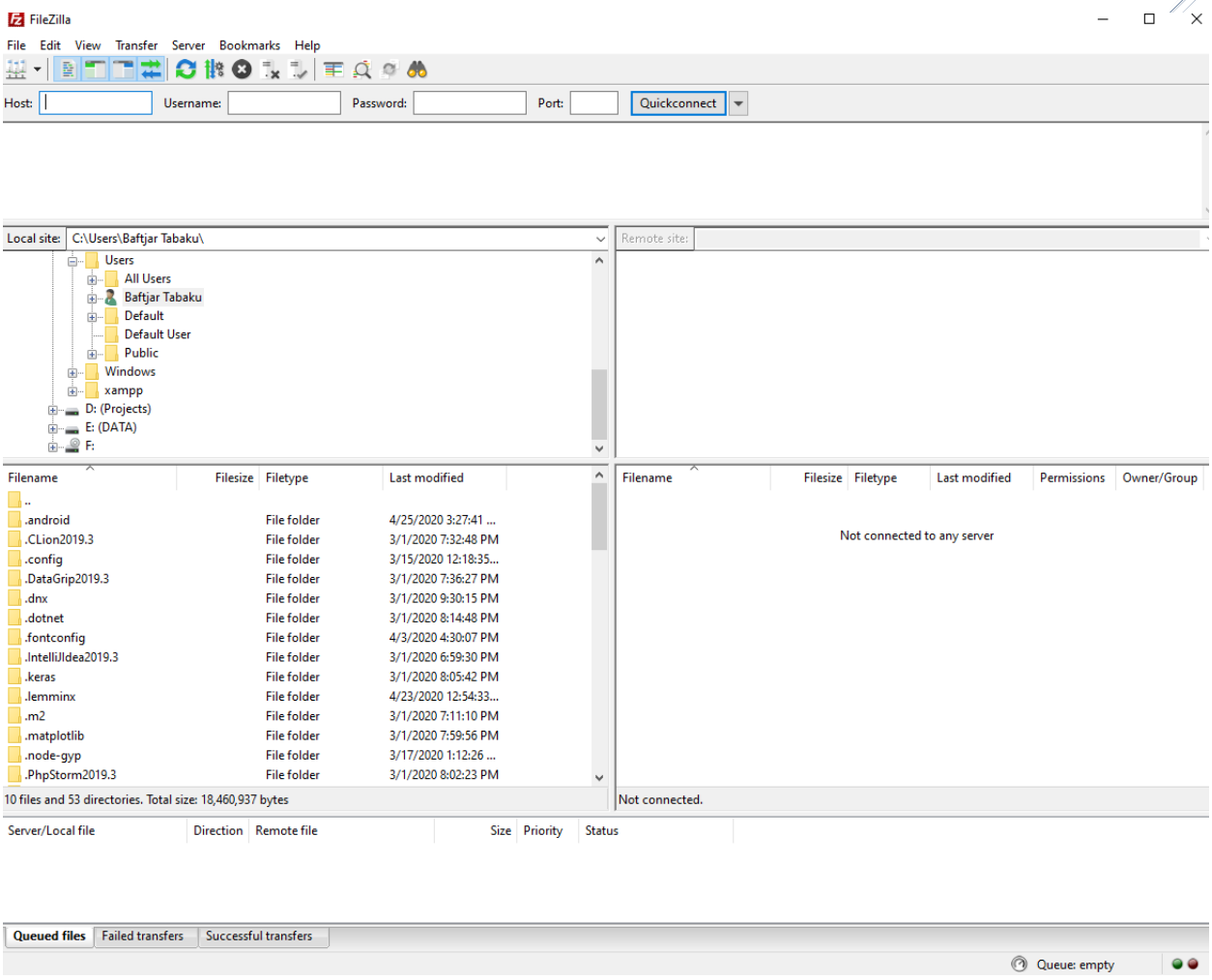




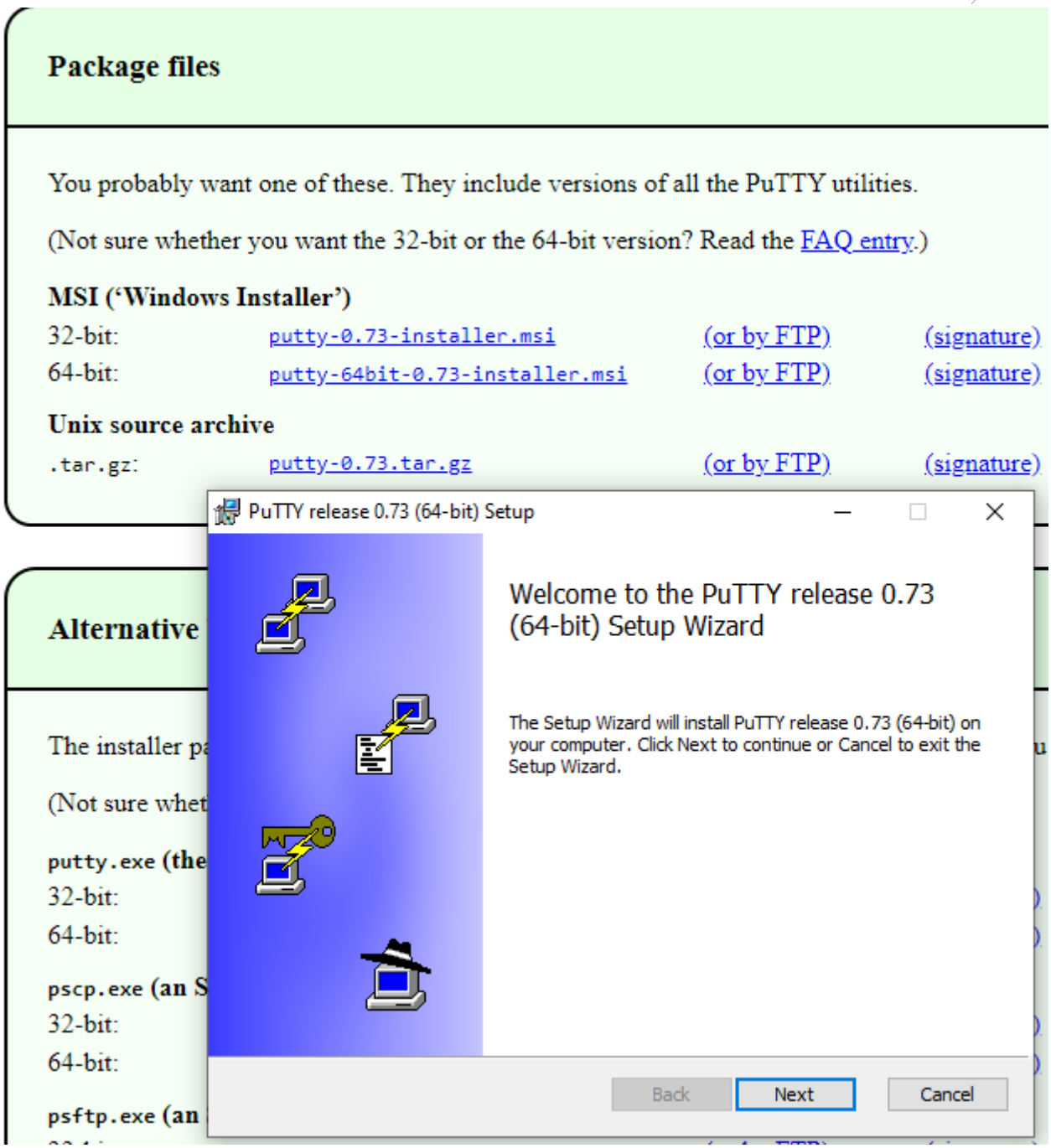


And now, to interact with Hadoop we need to download FileZilla and putty as shown below,

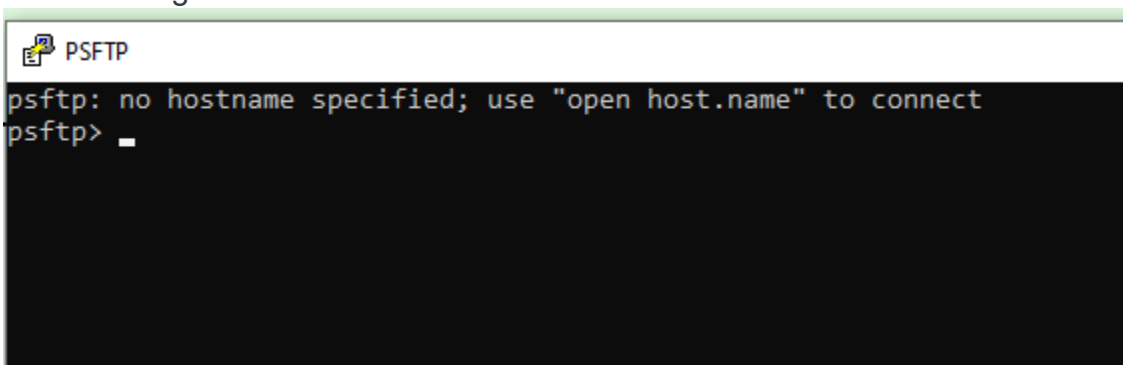



FileZilla
File Edit View Transfer Server Bookmarks Help
Host: Username: Password: Port: Quickconnect
Local site: C:\Users\Baftjar Tabaku\
Users
All Users
Baftjar Tabaku
Default
Default User
Public
Windows
xampp
D: (Projects)
E: (DATA)
F:
Remote site:
Filename Filesize Filetype Last modified
.android File folder 4/25/2020 3:27:41 ...
.CLion2019.3 File folder 3/1/2020 7:32:48 PM
.config File folder 3/15/2020 12:18:35...
.DataGrip2019.3 File folder 3/1/2020 7:36:27 PM
.dnx File folder 3/1/2020 9:30:15 PM
.dotnet File folder 3/1/2020 8:14:48 PM
.fontconfig File folder 4/3/2020 4:30:07 PM
.IntelliJIdea2019.3 File folder 3/1/2020 6:59:30 PM
.keras File folder 3/1/2020 8:05:42 PM
.lemminx File folder 4/23/2020 12:54:33...
.m2 File folder 3/1/2020 7:11:10 PM
.matplotlib File folder 3/1/2020 7:59:56 PM
.node-gyp File folder 3/17/2020 1:12:26 ...
.PhpStorm2019.3 File folder 3/1/2020 8:02:23 PM
10 files and 53 directories. Total size: 18,460,937 bytes
Filename Filesize Filetype Last modified Permissions Owner/Group
Not connected to any server
Not connected.
Server/Local file Direction Remote file Size Priority Status
Queued files Failed transfers Successful transfers
Queue: empty

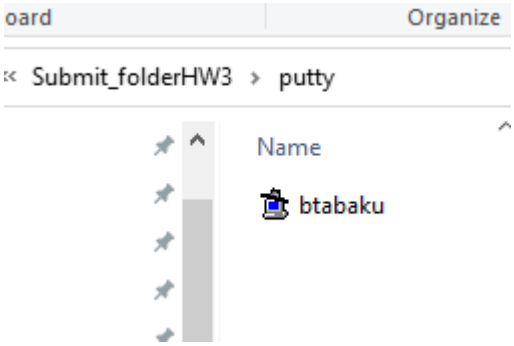
The next step was to install the 'putty' for windows,

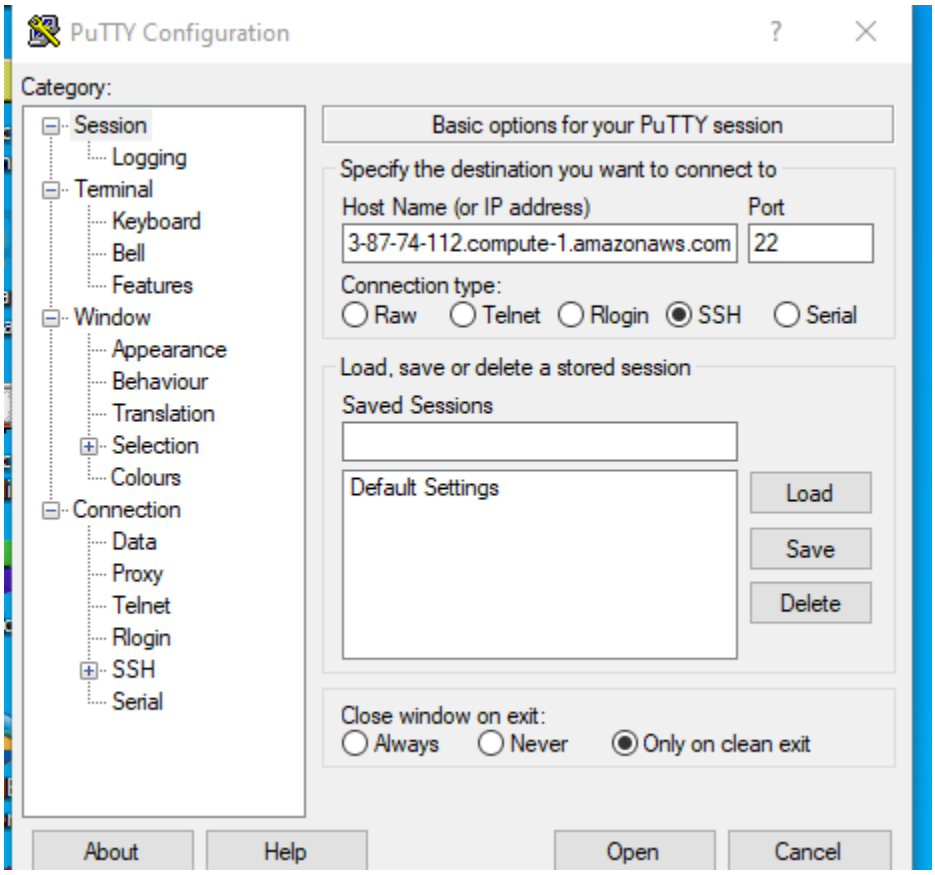


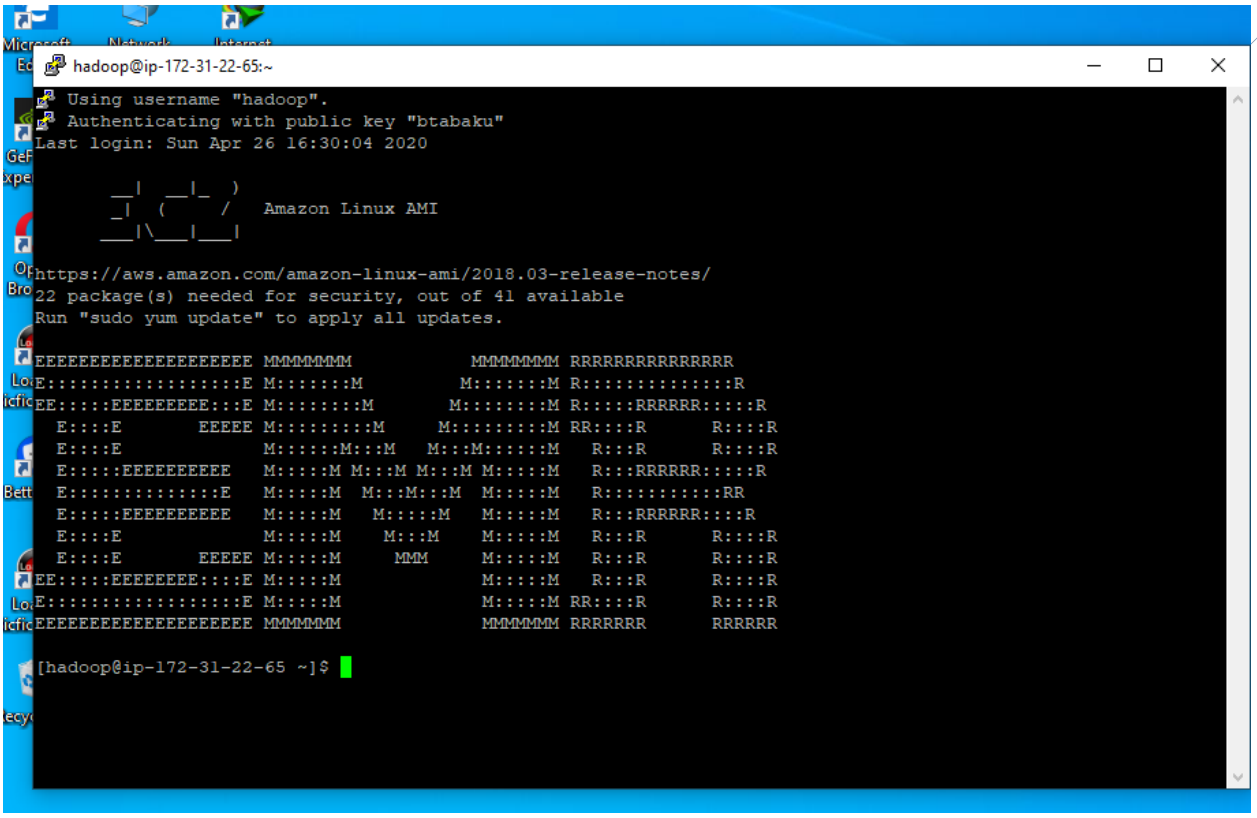
Then starting it

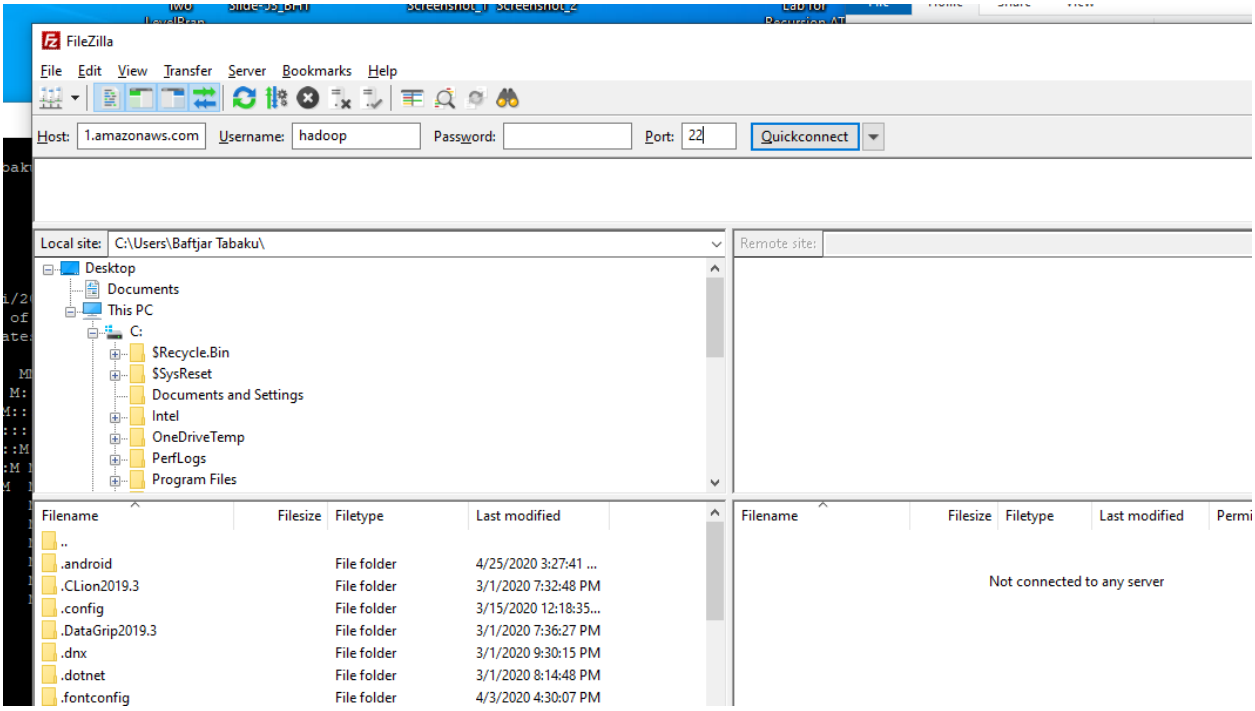


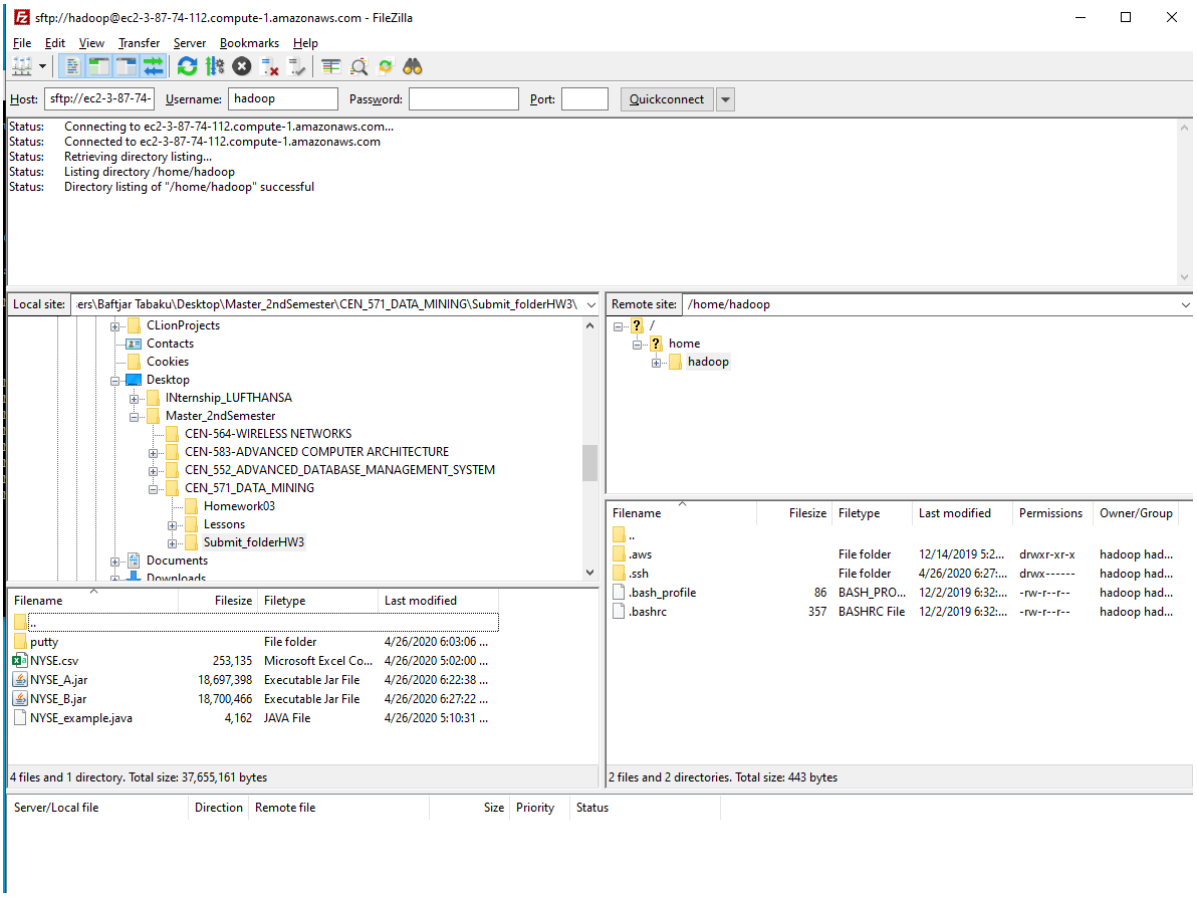
According to the file '.ppk'

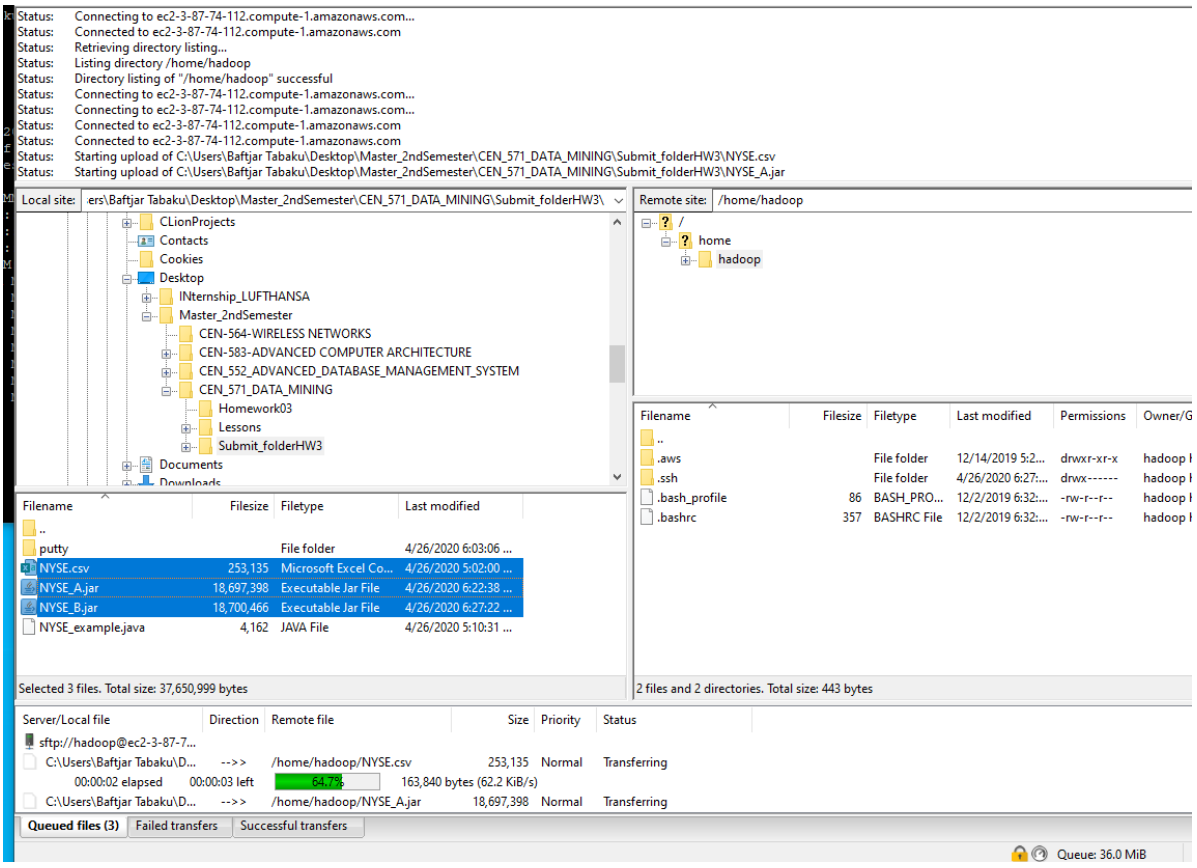


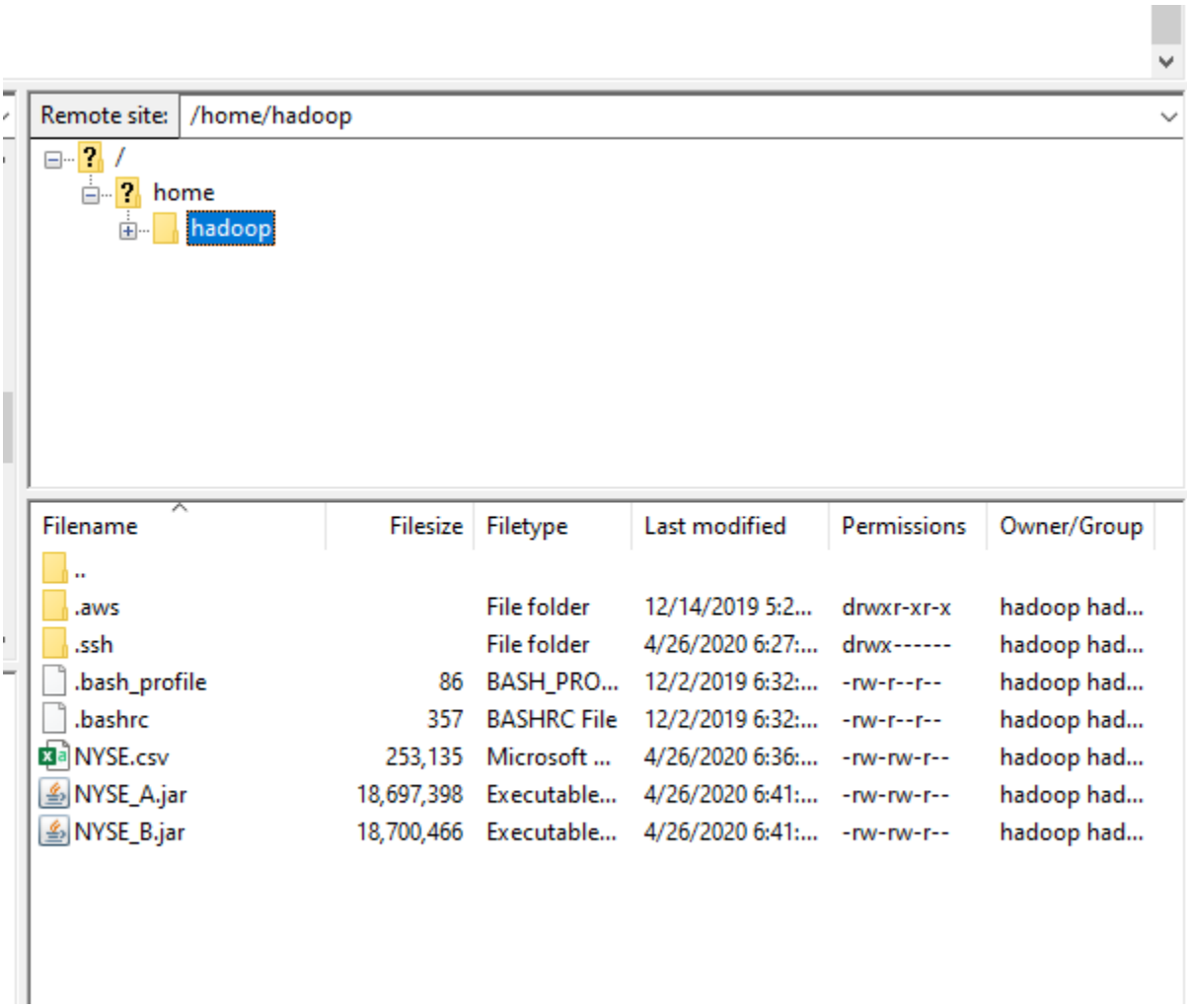



hadoop@ip-172-31-22-65:~
Using username "hadoop".
Authenticating with public key "btabaku"
Last login: Sun Apr 26 16:30:04 2020
Amazon Linux AMI
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
22 package(s) needed for security, out of 41 available
Run "sudo yum update" to apply all updates.
[hadoop@ip-172-31-22-65 ~]$


FileZilla
File Edit View Transfer Server Bookmarks Help
Host: 1.amazonaws.com Username: hadoop Password: Port: 22 Quickconnect
Local site: C:\Users\Baftjar Tabaku\
Remote site:
Desktop
Documents
This PC
C:
$Recycle.Bin
$SysReset
Documents and Settings
Intel
OneDriveTemp
PerfLogs
Program Files
Filename Filesize Filetype Last modified
.android File folder 4/25/2020 3:27:41 ...
.CLion2019.3 File folder 3/1/2020 7:32:48 PM
.config File folder 3/15/2020 12:18:35...
.DataGrip2019.3 File folder 3/1/2020 7:36:27 PM
.dnx File folder 3/1/2020 9:30:15 PM
.dotnet File folder 3/1/2020 8:14:48 PM
.fontconfig File folder 4/3/2020 4:30:07 PM
Not connected to any server


sftp://hadoop@ec2-3-87-74-112.compute-1.amazonaws.com - FileZilla
File Edit View Transfer Server Bookmarks Help
Host: sftp://ec2-3-87-74- Username: hadoop Password: Port: Quickconnect
Status: Connecting to ec2-3-87-74-112.compute-1.amazonaws.com...
Status: Connected to ec2-3-87-74-112.compute-1.amazonaws.com
Status: Retrieving directory listing...
Status: Listing directory /home/hadoop
Status: Directory listing of "/home/hadoop" successful
Local site: ers\Baftjar Tabaku\Desktop\Master_2ndSemester\CEN_571_DATA_MINING\Submit_folderHW3\
CLionProjects
Contacts
Cookies
Desktop
INternship_LUFTHANSA
Master_2ndSemester
CEN-564-WIRELESS NETWORKS
CEN-583-ADVANCED COMPUTER ARCHITECTURE
CEN_552_ADVANCED_DATABASE_MANAGEMENT_SYSTEM
CEN_571_DATA_MINING
Homework03
Lessons
Submit_folderHW3
Documents
Filename Filesize Filetype Last modified
putty File folder 4/26/2020 6:03:06 ...
NYSE.csv 253,135 Microsoft Excel Co... 4/26/2020 5:02:00 ...
NYSE_A.jar 18,697,398 Executable Jar File 4/26/2020 6:22:38 ...
NYSE_B.jar 18,700,466 Executable Jar File 4/26/2020 6:27:22 ...
NYSE_example.java 4,162 JAVA File 4/26/2020 5:10:31 ...
4 files and 1 directory. Total size: 37,655,161 bytes
Remote site: /home/hadoop
/
home
hadoop
Filename Filesize Filetype Last modified Permissions Owner/Group
.aws File folder 12/14/2019 5:2... drwxr-xr-x hadoop had...
.ssh File folder 4/26/2020 6:27:... drwx------ hadoop had...
.bash_profile 86 BASH_PRO... 12/2/2019 6:32:... -rw-r--r-- hadoop had...
.bashrc 357 BASHRC File 12/2/2019 6:32:... -rw-r--r-- hadoop had...
2 files and 2 directories. Total size: 443 bytes
Server/Local file Direction Remote file Size Priority Status


Status: Connecting to ec2-3-87-74-112.compute-1.amazonaws.com...
Status: Connected to ec2-3-87-74-112.compute-1.amazonaws.com
Status: Retrieving directory listing...
Status: Listing directory /home/hadoop
Status: Directory listing of "/home/hadoop" successful
Status: Connecting to ec2-3-87-74-112.compute-1.amazonaws.com...
Status: Connecting to ec2-3-87-74-112.compute-1.amazonaws.com...
Status: Connected to ec2-3-87-74-112.compute-1.amazonaws.com
Status: Connected to ec2-3-87-74-112.compute-1.amazonaws.com
Status: Starting upload of C:\Users\Baftjar Tabaku\Desktop\Master_2ndSemester\CEN_571_DATA_MINING\Submit_folderHW3\NYSE.csv
Status: Starting upload of C:\Users\Baftjar Tabaku\Desktop\Master_2ndSemester\CEN_571_DATA_MINING\Submit_folderHW3\NYSE_A.jar
Local site: ers\Baftjar Tabaku\Desktop\Master_2ndSemester\CEN_571_DATA_MINING\Submit_folderHW3\
Remote site: /home/hadoop
Selected 3 files. Total size: 37,650,999 bytes
2 files and 2 directories. Total size: 443 bytes
sftp://hadoop@ec2-3-87-7...
C:\Users\Baftjar Tabaku\D... --->> /home/hadoop/NYSE.csv 253,135 Normal Transferring
00:00:02 elapsed 00:00:03 left 64.7% 163,840 bytes (62.2 KiB/s)
C:\Users\Baftjar Tabaku\D... --->> /home/hadoop/NYSE_A.jar 18,697,398 Normal Transferring
Queued files (3) Failed transfers Successful transfers
Queue: 36.0 MiB



Using a simple command to list

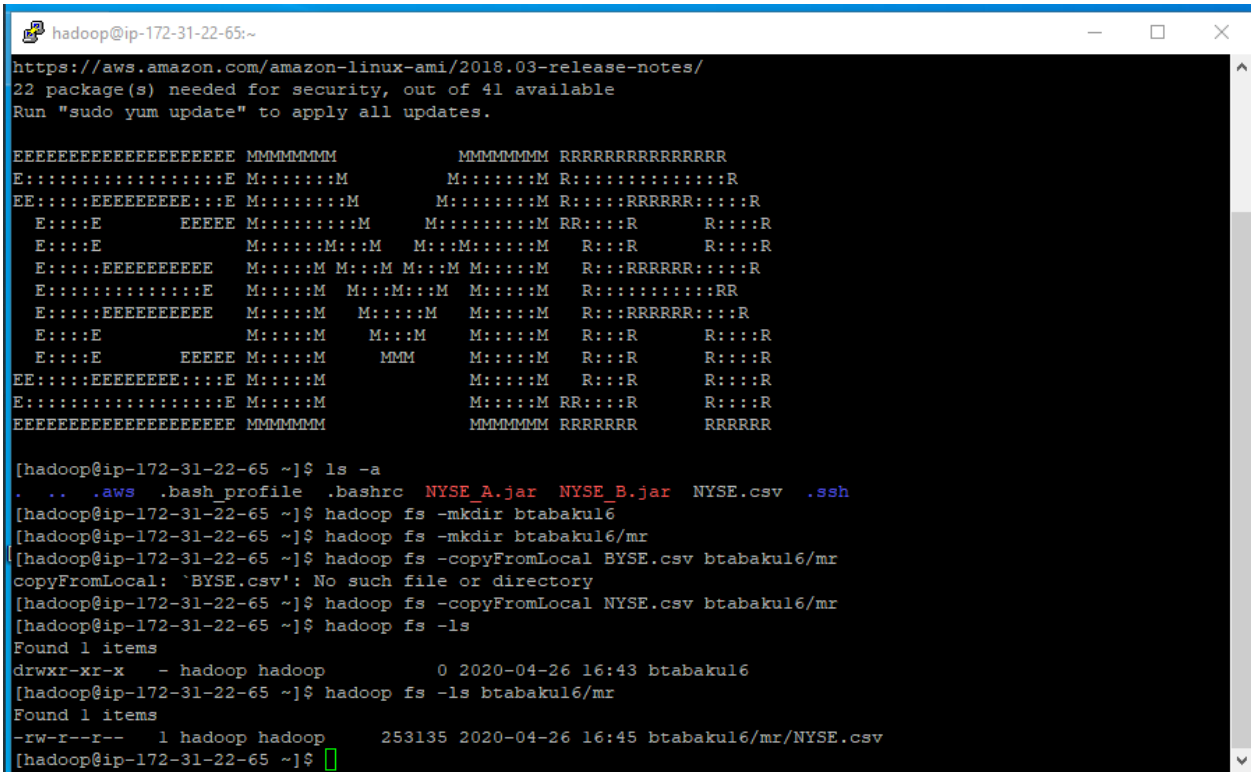
```
[hadoop@ip-172-31-22-65 ~]$ ls -a
.  ..  .aws  .bash_profile  .bashrc  NYSE_A.jar  NYSE_B.jar  NYSE.csv  .ssh
[hadoop@ip-172-31-22-65 ~]$
```

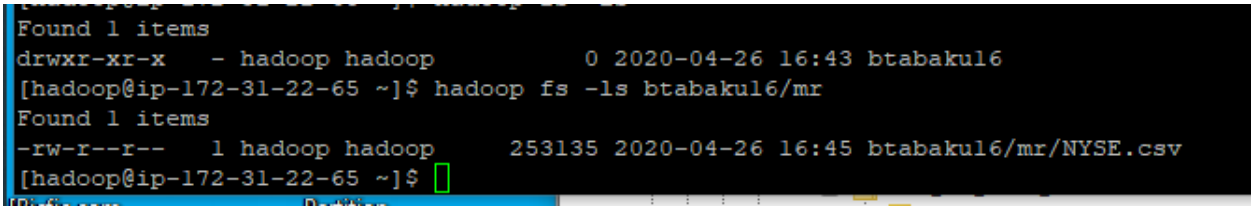
Making a folder btabaku16/mr

```
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -mkdir btabaku16
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -mkdir btabaku16/mr
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyFromLocal BYSE.csv btabaku16/mr
copyFromLocal: `BYSE.csv': No such file or directory
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyFromLocal NYSE.csv btabaku16/mr
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls
Found 1 items
```

We copy the input file NYSE.csv from local project folder to the AWS corresponding folder.

```
[[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyFromLocal NYSE.csv btabakul6/mr
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls
Found 1 items
```





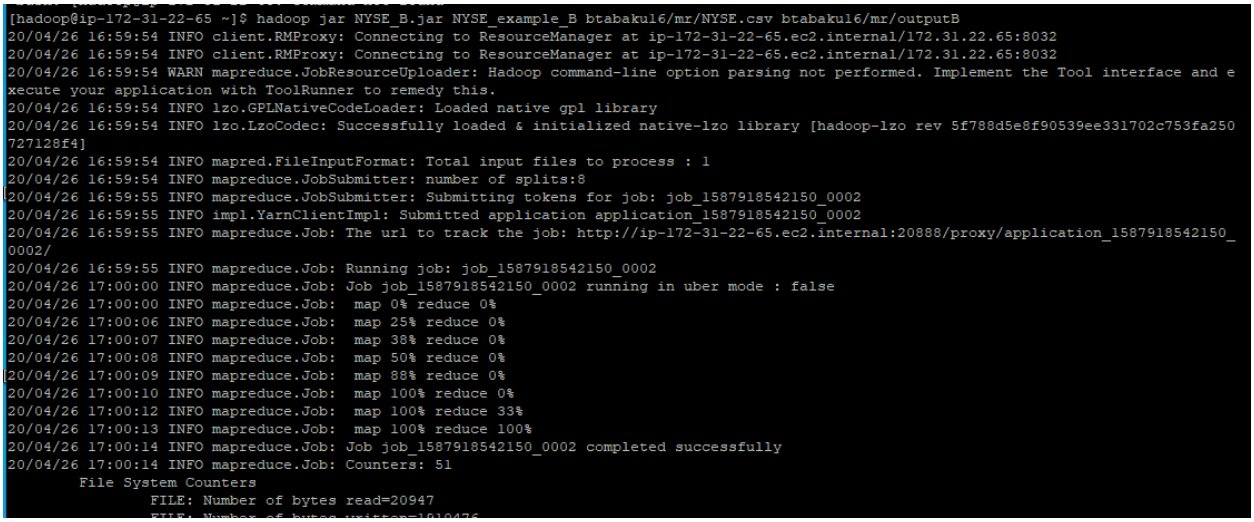
```
hadoop@ip-172-31-22-65:~
[hadoop@ip-172-31-22-65 ~]$ ls -a
.  ..  .aws  .bash_profile  .bashrc  NYSE_A.jar  NYSE_B.jar  NYSE.csv  .ssh
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -mkdir btabaku16
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -mkdir btabaku16/mr
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyFromLocal BYSE.csv btabaku16/mr
copyFromLocal: `BYSE.csv': No such file or directory
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyFromLocal NYSE.csv btabaku16/mr
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2020-04-26 16:43 btabaku16
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls btabaku16/mr
Found 1 items
-rw-r--r--   1 hadoop hadoop     253135 2020-04-26 16:45 btabaku16/mr/NYSE.csv
[hadoop@ip-172-31-22-65 ~]$ hadoop jar NYSE_A.jar NYSE_example btabaku16/mr/NYSE.csv btabaku16/mr/outputA
20/04/26 16:53:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-22-65.ec2.internal/172.31.2
2.65:8032
20/04/26 16:53:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-22-65.ec2.internal/172.31.2
2.65:8032
20/04/26 16:53:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Imple
ment the Tool interface and execute your application with ToolRunner to remedy this.
20/04/26 16:53:54 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
20/04/26 16:53:55 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 5f7
88d5e8f90539ee331702c753fa250727128f4]
20/04/26 16:53:55 INFO mapred.FileInputFormat: Total input files to process : 1
20/04/26 16:53:55 INFO mapreduce.JobSubmitter: number of splits:8
20/04/26 16:53:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1587918542150_0001
20/04/26 16:53:55 INFO impl.YarnClientImpl: Submitted application application_1587918542150_0001
20/04/26 16:53:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-22-65.ec2.internal:20888/pro
xy/application_1587918542150_0001/
20/04/26 16:53:55 INFO mapreduce.Job: Running job: job_1587918542150_0001
```

Running the first command and getting the output A, to the folder btabaku16/mr/outputA

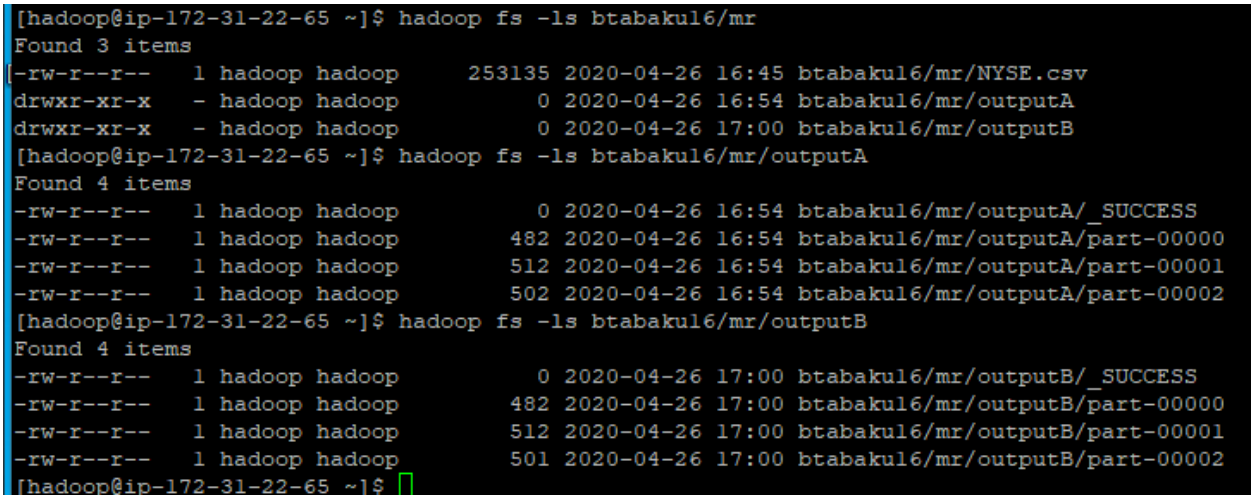
```
[hadoop@ip-172-31-22-65 ~]$ hadoop jar NYSE_A.jar NYSE_example btabaku16/mr/NYSE.csv btabaku16/mr/outputA
20/04/26 16:53:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-22-65.ec2.internal/172.31.2
2.65:8032
20/04/26 16:53:54 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-22-65.ec2.internal/172.31.2
2.65:8032
20/04/26 16:53:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Imple
ment the Tool interface and execute your application with ToolRunner to remedy this.
20/04/26 16:53:54 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
20/04/26 16:53:55 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 5f7
88d5e8f90539ee331702c753fa250727128f4]
20/04/26 16:53:55 INFO mapred.FileInputFormat: Total input files to process : 1
20/04/26 16:53:55 INFO mapreduce.JobSubmitter: number of splits:8
20/04/26 16:53:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1587918542150_0001
20/04/26 16:53:55 INFO impl.YarnClientImpl: Submitted application application_1587918542150_0001
20/04/26 16:53:55 INFO mapreduce.Job: The url to track the job: http://ip-172-31-22-65.ec2.internal:20888/pro
xy/application_1587918542150_0001/
20/04/26 16:53:55 INFO mapreduce.Job: Running job: job_1587918542150_0001
20/04/26 16:54:01 INFO mapreduce.Job: Job job_1587918542150_0001 running in uber mode : false
20/04/26 16:54:01 INFO mapreduce.Job:  map 0% reduce 0%
20/04/26 16:54:07 INFO mapreduce.Job:  map 25% reduce 0%
20/04/26 16:54:10 INFO mapreduce.Job:  map 63% reduce 0%
20/04/26 16:54:11 INFO mapreduce.Job:  map 88% reduce 0%
20/04/26 16:54:12 INFO mapreduce.Job:  map 100% reduce 0%
20/04/26 16:54:15 INFO mapreduce.Job:  map 100% reduce 100%
20/04/26 16:54:16 INFO mapreduce.Job: Job job_1587918542150_0001 completed successfully
20/04/26 16:54:17 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=2084
```

Which successfully worked and executed.

Then running the second command and getting the output B, to the folder btabaku16/mr/outputB



And after this, running a list command as shown in the following screenshot below.

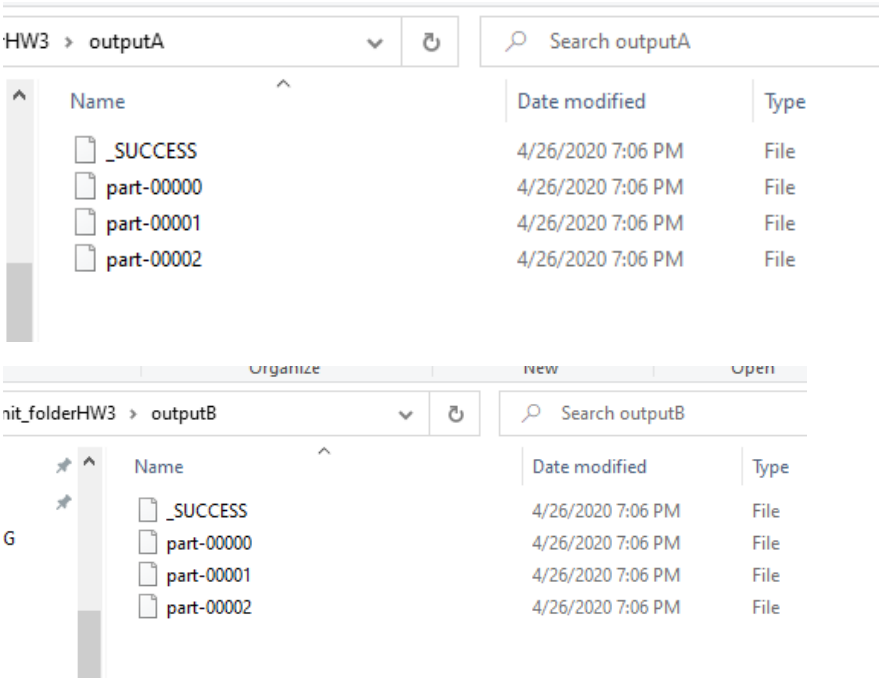


And after this, going to each output folder A and B as shown below and listing their corresponding contents.

```
                Bytes Written=1495
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls btabakul6/mr
Found 3 items
-rw-r--r--   1 hadoop hadoop     253135 2020-04-26 16:45 btabakul6/mr/NYSE.csv
drwxr-xr-x   - hadoop hadoop          0 2020-04-26 16:54 btabakul6/mr/outputA
drwxr-xr-x   - hadoop hadoop          0 2020-04-26 17:00 btabakul6/mr/outputB
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls btabakul6/mr/outputA
Found 4 items
-rw-r--r--   1 hadoop hadoop          0 2020-04-26 16:54 btabakul6/mr/outputA/_SUCCESS
-rw-r--r--   1 hadoop hadoop        482 2020-04-26 16:54 btabakul6/mr/outputA/part-00000
-rw-r--r--   1 hadoop hadoop        512 2020-04-26 16:54 btabakul6/mr/outputA/part-00001
-rw-r--r--   1 hadoop hadoop        502 2020-04-26 16:54 btabakul6/mr/outputA/part-00002
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -ls btabakul6/mr/outputB
Found 4 items
-rw-r--r--   1 hadoop hadoop          0 2020-04-26 17:00 btabakul6/mr/outputB/_SUCCESS
-rw-r--r--   1 hadoop hadoop        482 2020-04-26 17:00 btabakul6/mr/outputB/part-00000
-rw-r--r--   1 hadoop hadoop        512 2020-04-26 17:00 btabakul6/mr/outputB/part-00001
-rw-r--r--   1 hadoop hadoop        501 2020-04-26 17:00 btabakul6/mr/outputB/part-00002
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyToLocal btabakul6/mr/outputA ~
[hadoop@ip-172-31-22-65 ~]$ hadoop fs -copyToLocal btabakul6/mr/outputB ~
[hadoop@ip-172-31-22-65 ~]$ ls
NYSE_A.jar  NYSE_B.jar  NYSE.csv  outputA  outputB
[hadoop@ip-172-31-22-65 ~]$
```

And everything finished Successfully.

Note: there at the compilation of the jar files, are two classes, the 'NYSE_example.java' that stands for the task 1 and the 'NYSE_example_B.java' that stands for task 2 or B as shown. The output of folders outputA and outputB had like 3 files for each



And all this is splinted into two text files OutputA.txt and OutputB.txt.