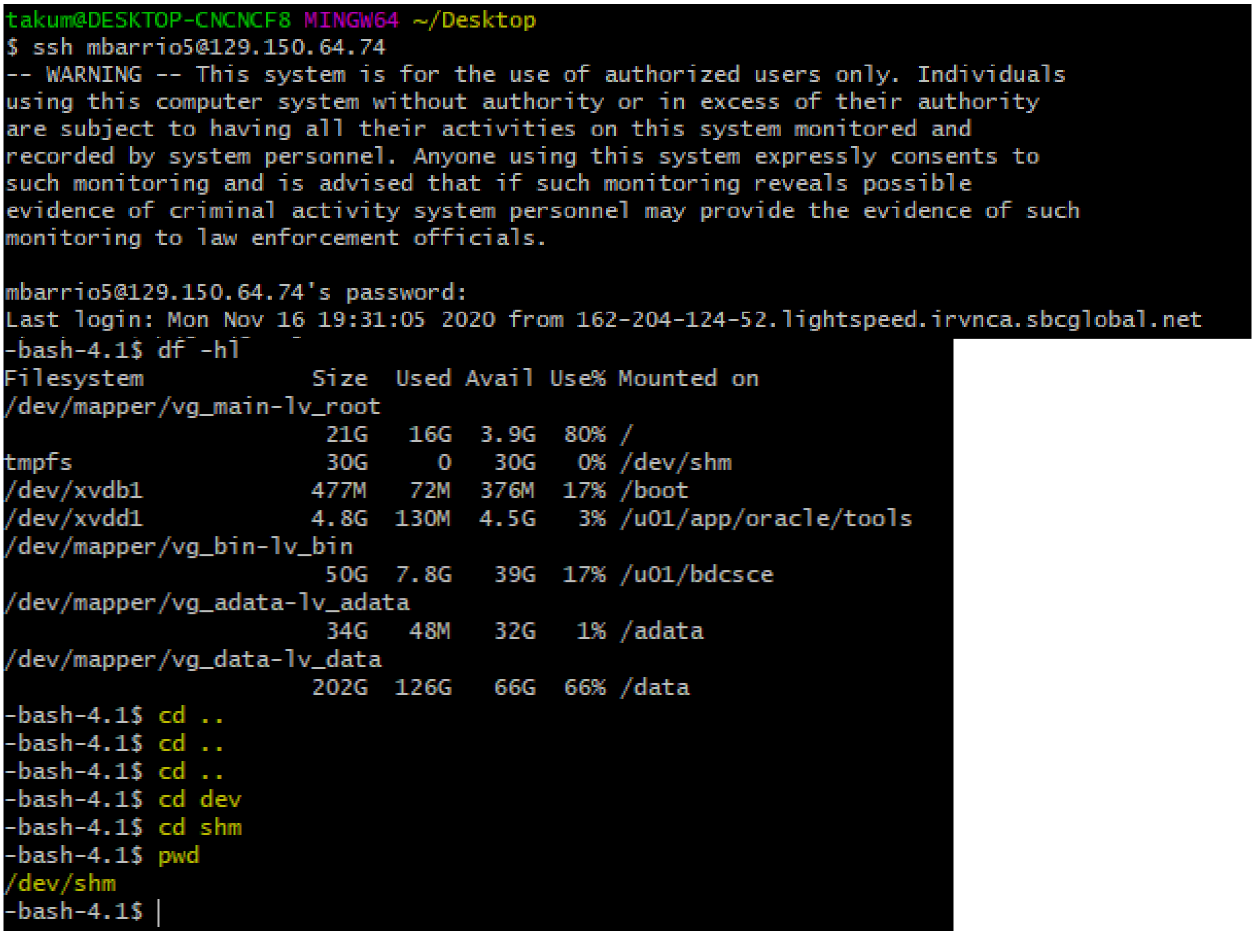
In this tutorial, we will download and analyze the 1996 to 2000 stock market data. The files span over 100,000+ directories and include stocks from around the world. The data does not include the location in the excel file so we will add it after the data is obtained.

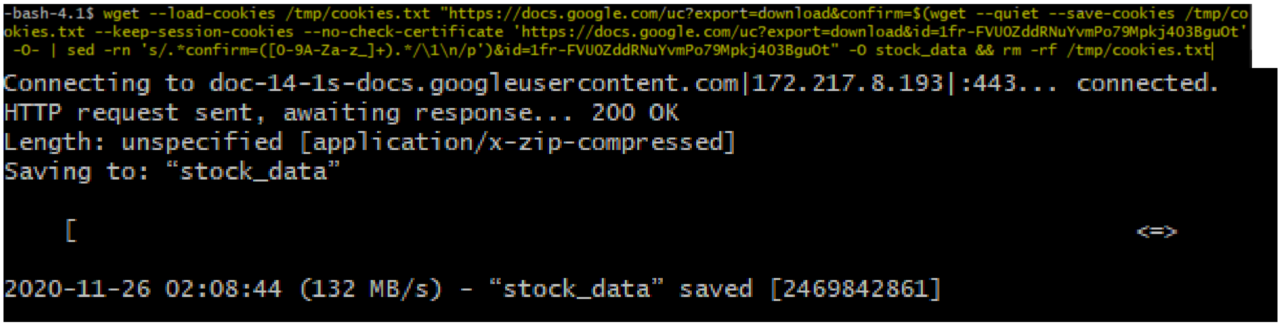
1. Open a Git Bash terminal and SSH



```
takum@DESKTOP-CNCNCF8 MINGW64 ~/Desktop
$ ssh mbarrio5@129.150.64.74
-- WARNING -- This system is for the use of authorized users only. Individuals
using this computer system without authority or in excess of their authority
are subject to having all their activities on this system monitored and
recorded by system personnel. Anyone using this system expressly consents to
such monitoring and is advised that if such monitoring reveals possible
evidence of criminal activity system personnel may provide the evidence of such
monitoring to law enforcement officials.

mbarrio5@129.150.64.74's password:
Last login: Mon Nov 16 19:31:05 2020 from 162-204-124-52.lightspeed.irvnca.sbcglobal.net
-bash-4.1$ df -hl
Filesystem            Size  Used Avail Use% Mounted on
/dev/mapper/vg_main-lv_root
                       21G   16G  3.9G  80% /
tmpfs                  30G     0   30G   0% /dev/shm
/dev/xvdb1            477M   72M  376M  17% /boot
/dev/xvdd1            4.8G  130M  4.5G   3% /u01/app/oracle/tools
/dev/mapper/vg_bin-lv_bin
                       50G  7.8G   39G  17% /u01/bdcsce
/dev/mapper/vg_adata-lv_adata
                       34G   48M   32G   1% /adata
/dev/mapper/vg_data-lv_data
                      202G  126G   66G  66% /data
-bash-4.1$ cd ..
-bash-4.1$ cd ..
-bash-4.1$ cd ..
-bash-4.1$ cd dev
-bash-4.1$ cd shm
-bash-4.1$ pwd
/dev/shm
-bash-4.1$ |
```

```
$ ssh [yourusernamehere]@129.150.64.74

$ cd ..

$ cd ..

$ cd ..

$ cd dev

$ cd shm

$ pwd
```

2. Download the dataset to local Linux system

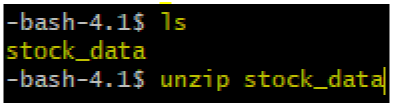


```
-bash-4.1$ wget --load-cookies /tmp/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/co
okies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=1fr-FVU0ZddRNuYvmPo79Mpkj403BguOt'
 -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1fr-FVU0ZddRNuYvmPo79Mpkj403BguOt" -O stock_data && rm -rf /tmp/cookies.txt
Connecting to doc-14-1s-docs.googleusercontent.com|172.217.8.193|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/x-zip-compressed]
Saving to: “stock_data”

    [                                                                    <=>

2020-11-26 02:08:44 (132 MB/s) - “stock_data” saved [2469842861]
```

$ wget --load-cookies /tmp/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=1fr-FVU0ZddRNuYvmPo79Mpkj403BguOt' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1fr-FVU0ZddRNuYvmPo79Mpkj403BguOt" -O stock_data && rm -rf /tmp/cookies.txt

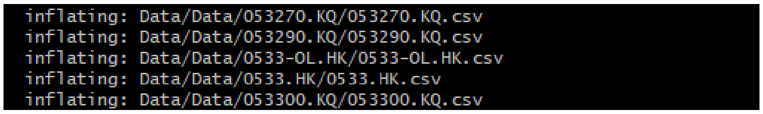
Note: “stock_data” saved will appear if the download is successful

3. Unzipping dataset (takes approximately 5 minutes)



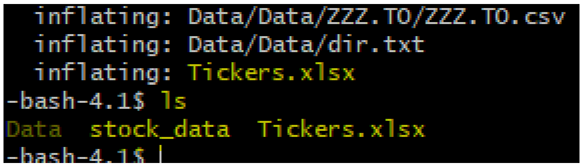
```
-bash-4.1$ ls
stock_data
-bash-4.1$ unzip stock_data
```

-Unzip in progress



```
  inflating: Data/Data/053270.KQ/053270.KQ.csv
  inflating: Data/Data/053290.KQ/053290.KQ.csv
  inflating: Data/Data/0533-OL.HK/0533-OL.HK.csv
  inflating: Data/Data/0533.HK/0533.HK.csv
  inflating: Data/Data/053300.KQ/053300.KQ.csv
```

-Unzip complete



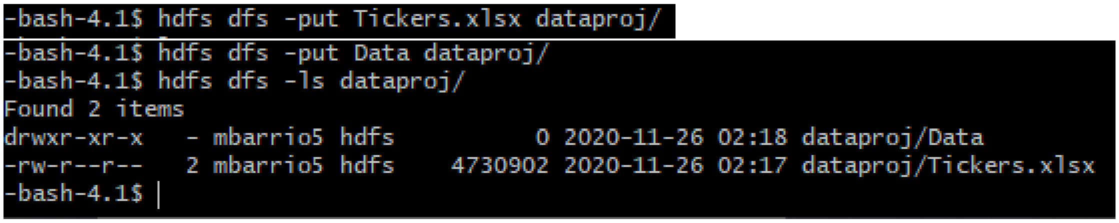
```
  inflating: Data/Data/ZZZ.TO/ZZZ.TO.csv
  inflating: Data/Data/dir.txt
  inflating: Tickers.xlsx
-bash-4.1$ ls
Data  stock_data  Tickers.xlsx
-bash-4.1$
```

$ ls

$ unzip stock_data

$ hdfs dfs -mkdir dataproj

4. Move files from local linux to HDFS (Moving “Data” to dataproj/ takes 1.1 hours)



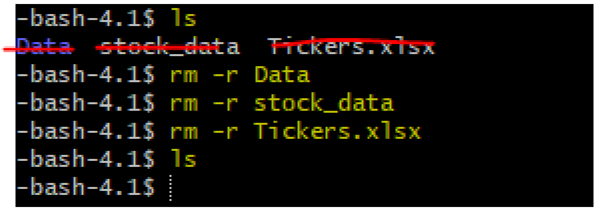
```
-bash-4.1$ hdfs dfs -put Tickers.xlsx dataproj/
-bash-4.1$ hdfs dfs -put Data dataproj/
-bash-4.1$ hdfs dfs -ls dataproj/
Found 2 items
drwxr-xr-x   - mbarrio5 hdfs          0 2020-11-26 02:18 dataproj/Data
-rw-r--r--   2 mbarrio5 hdfs    4730902 2020-11-26 02:17 dataproj/Tickers.xlsx
-bash-4.1$ |
```

$ hdfs dfs -put Tickers.xlsx dataproj/

$ hdfs dfs -put Data dataproj/

$ hdfs dfs -ls dataproj/

Note: remove files to save storage on system



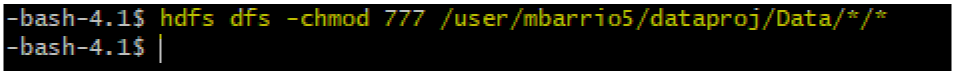
```
-bash-4.1$ ls
~~Data~~  ~~stock_data~~  ~~Tickers.xlsx~~
-bash-4.1$ rm -r Data
-bash-4.1$ rm -r stock_data
-bash-4.1$ rm -r Tickers.xlsx
-bash-4.1$ ls
-bash-4.1$
```

$ rm -r Data

$ rm -r stock_data

$ rm -r Tickers.xlsx

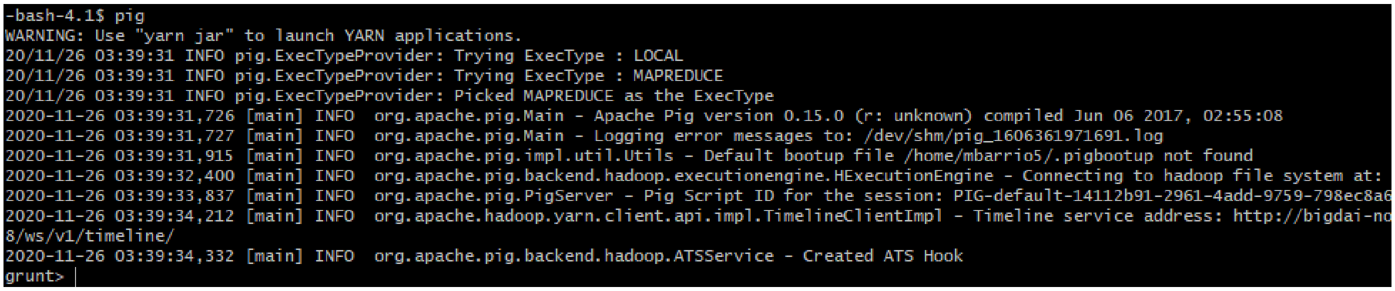
5. Change file permissions of dataset



```
-bash-4.1$ hdfs dfs -chmod 777 /user/mbarrio5/dataproj/Data/*/*
-bash-4.1$ |
```

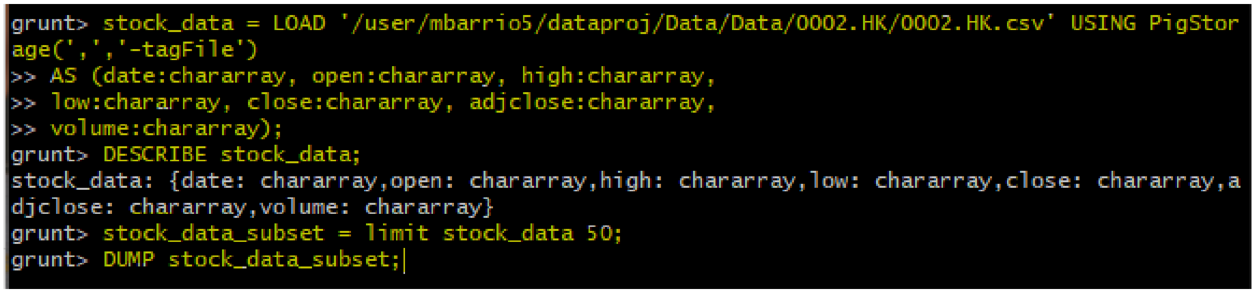
$ hdfs dfs -chmod 777 /user/mbarrio5/dataproj/Data/Data/*/*

6. Using Pig to consolidate all 101,000 directories and files (In the step we create a test schema)



```
-bash-4.1$ pig
WARNING: Use "yarn jar" to launch YARN applications.
20/11/26 03:39:31 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/11/26 03:39:31 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
20/11/26 03:39:31 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2020-11-26 03:39:31,726 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0 (r: unknown) compiled Jun 06 2017, 02:55:08
2020-11-26 03:39:31,727 [main] INFO  org.apache.pig.Main - Logging error messages to: /dev/shm/pig_1606361971691.log
2020-11-26 03:39:31,915 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/mbarrio5/.pigbootup not found
2020-11-26 03:39:32,400 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at:
2020-11-26 03:39:33,837 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-14112b91-2961-4add-9759-798ec8a6
2020-11-26 03:39:34,212 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://bigdai-no
8/ws/v1/timeline/
2020-11-26 03:39:34,332 [main] INFO  org.apache.pig.backend.hadoop.ATSService - Created ATS Hook
grunt> 
```

$ pig



```
grunt> stock_data = LOAD '/user/mbarrio5/dataproj/Data/Data/0002.HK/0002.HK.csv' USING PigStor
age(',','-tagFile')
>> AS (date:chararray, open:chararray, high:chararray,
>> low:chararray, close:chararray, adjclose:chararray,
>> volume:chararray);
grunt> DESCRIBE stock_data;
stock_data: {date: chararray,open: chararray,high: chararray,low: chararray,close: chararray,a
djclose: chararray,volume: chararray}
grunt> stock_data_subset = limit stock_data 50;
grunt> DUMP stock_data_subset;
```

$ stock_data = LOAD '/user/mbarrio5/dataproj/Data/Data/0002.HK/0002.HK.csv' USING PigStorage(',','-tagFile')

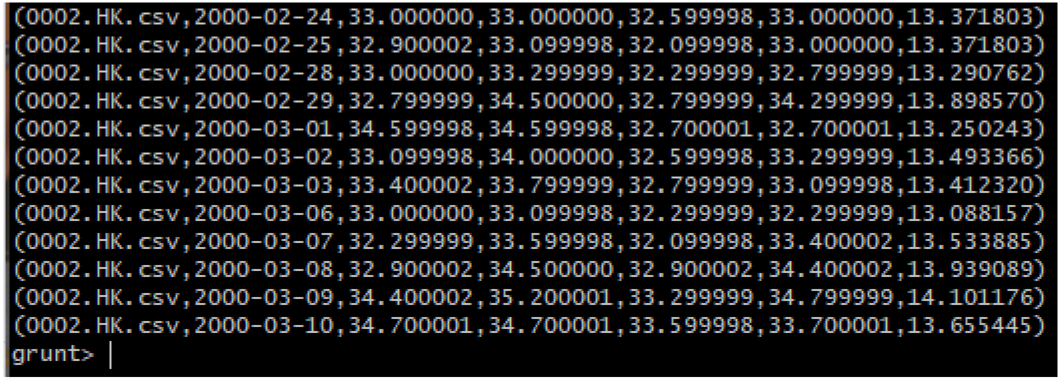
```
        AS (date:chararray, open:chararray, high:chararray,
        low:chararray, close:chararray, adjclose:chararray,
        volume:chararray);
```

$ Describe stock_data;

$ stock_data_subset = limit stock_data 50;

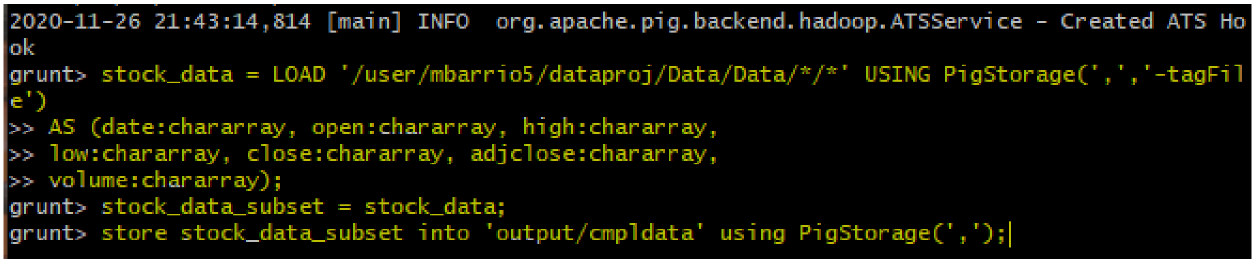
$ DUMP stock_data_subset;

Output of “DUMP stock_data_subset;”:



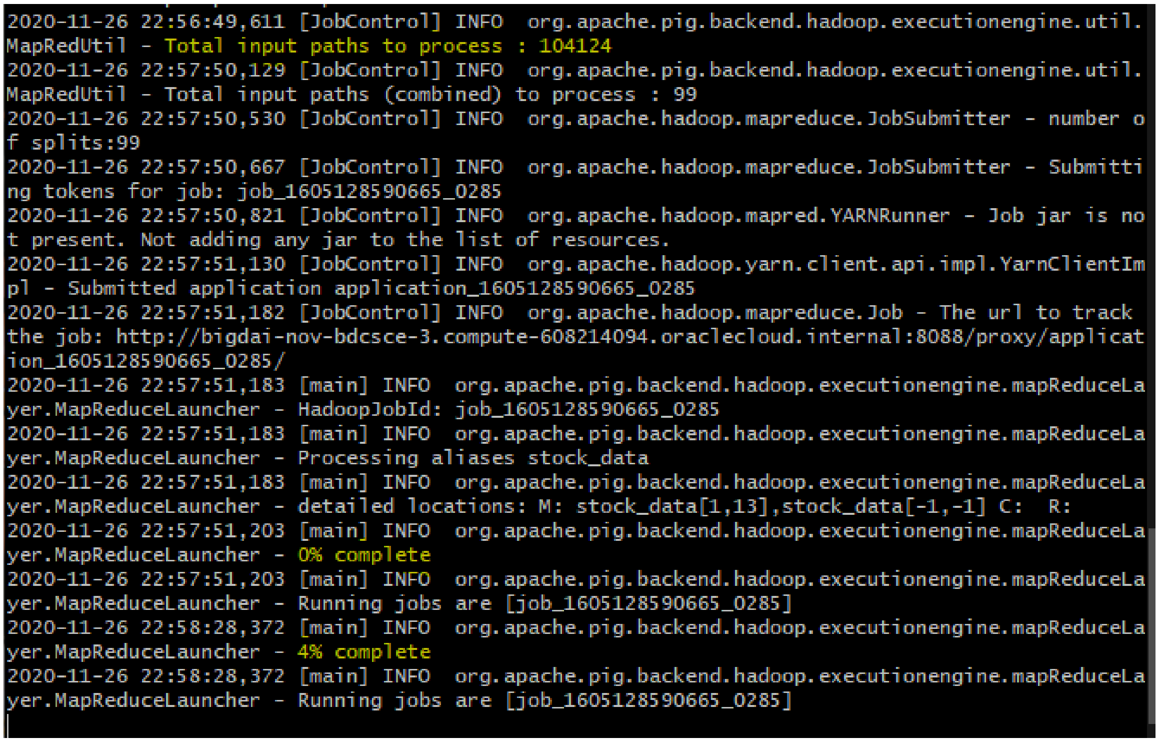
```
(0002.HK.csv,2000-02-24,33.000000,33.000000,32.599998,33.000000,13.371803)
(0002.HK.csv,2000-02-25,32.900002,33.099998,32.099998,33.000000,13.371803)
(0002.HK.csv,2000-02-28,33.000000,33.299999,32.299999,32.799999,13.290762)
(0002.HK.csv,2000-02-29,32.799999,34.500000,32.799999,34.299999,13.898570)
(0002.HK.csv,2000-03-01,34.599998,34.599998,32.700001,32.700001,13.250243)
(0002.HK.csv,2000-03-02,33.099998,34.000000,32.599998,33.299999,13.493366)
(0002.HK.csv,2000-03-03,33.400002,33.799999,32.799999,33.099998,13.412320)
(0002.HK.csv,2000-03-06,33.000000,33.099998,32.299999,32.299999,13.088157)
(0002.HK.csv,2000-03-07,32.299999,33.599998,32.099998,33.400002,13.533885)
(0002.HK.csv,2000-03-08,32.900002,34.500000,32.900002,34.400002,13.939089)
(0002.HK.csv,2000-03-09,34.400002,35.200001,33.299999,34.799999,14.101176)
(0002.HK.csv,2000-03-10,34.700001,34.700001,33.599998,33.700001,13.655445)
grunt> 
```

7. Complete schema to map reduce ALL of the data



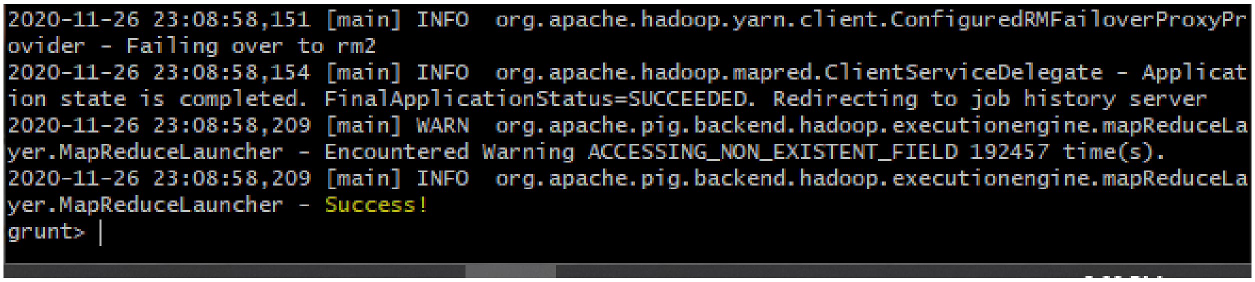
```
2020-11-26 21:43:14,814 [main] INFO  org.apache.pig.backend.hadoop.ATSService - Created ATS Ho
ok
grunt> stock_data = LOAD '/user/mbarrio5/dataproj/Data/Data/*/*' USING PigStorage(',','-tagFil
e')
>> AS (date:chararray, open:chararray, high:chararray,
>> low:chararray, close:chararray, adjclose:chararray,
>> volume:chararray);
grunt> stock_data_subset = stock_data;
grunt> store stock_data_subset into 'output/cmpldata' using PigStorage(',');
```

Store function in progress:



```
2020-11-26 22:56:49,611 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.
MapRedUtil - Total input paths to process : 104124
2020-11-26 22:57:50,129 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.
MapRedUtil - Total input paths (combined) to process : 99
2020-11-26 22:57:50,530 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number o
f splits:99
2020-11-26 22:57:50,667 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitti
ng tokens for job: job_1605128590665_0285
2020-11-26 22:57:50,821 [JobControl] INFO  org.apache.hadoop.mapred.YARNRunner - Job jar is no
t present. Not adding any jar to the list of resources.
2020-11-26 22:57:51,130 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientIm
pl - Submitted application application_1605128590665_0285
2020-11-26 22:57:51,182 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track
the job: http://bigdai-nov-bdcsce-3.compute-608214094.oraclecloud.internal:8088/proxy/applicat
ion_1605128590665_0285/
2020-11-26 22:57:51,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - HadoopJobId: job_1605128590665_0285
2020-11-26 22:57:51,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - Processing aliases stock_data
2020-11-26 22:57:51,183 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - detailed locations: M: stock_data[1,13],stock_data[-1,-1] C:  R:
2020-11-26 22:57:51,203 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - 0% complete
2020-11-26 22:57:51,203 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - Running jobs are [job_1605128590665_0285]
2020-11-26 22:58:28,372 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - 4% complete
2020-11-26 22:58:28,372 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - Running jobs are [job_1605128590665_0285]
```

```
$ stock_data = LOAD '/user/mbarrio5/dataproj/Data/Data/*/*' USING PigStorage(',','-tagFile')
        AS (date:chararray, open:chararray, high:chararray,

        low:chararray, close:chararray, adjclose:chararray,

        volume:chararray);

$ stock_data_subset = stock_data;

$ store stock_data_subset into 'output/cmpldata' using PigStorage(',');
```



```
2020-11-26 23:08:58,151 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyPr
ovider - Failing over to rm2
2020-11-26 23:08:58,154 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Applicat
ion state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-11-26 23:08:58,209 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 192457 time(s).
2020-11-26 23:08:58,209 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLa
yer.MapReduceLauncher - Success!
grunt>
```

Note: Make sure to "quit" after completing the "store" line

$ quit

8. Change permissions and merge all 100 Pig outputs into 1 CSV file

**INSERT SCREENSHOT OF CODE EXECUTION HERE

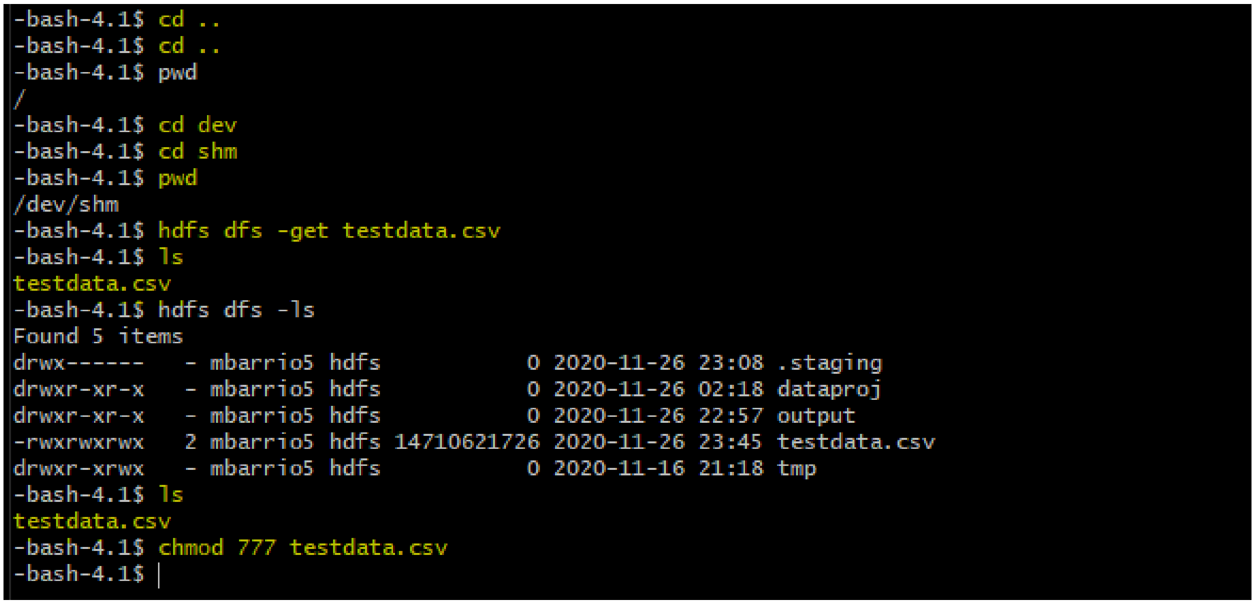
$ hdfs dfs -chmod 777 output/*

$ hdfs dfs -cat output/cmpldata/* | hadoop fs -put - testdata.csv

$ hdfs dfs -chmod 777 testdata.csv

Continue to next page

9. Download CSV to temp file system due to size restrictions



```
-bash-4.1$ cd ..
-bash-4.1$ cd ..
-bash-4.1$ pwd
/
-bash-4.1$ cd dev
-bash-4.1$ cd shm
-bash-4.1$ pwd
/dev/shm
-bash-4.1$ hdfs dfs -get testdata.csv
-bash-4.1$ ls
testdata.csv
-bash-4.1$ hdfs dfs -ls
Found 5 items
drwx------   - mbarrio5 hdfs           0 2020-11-26 23:08 .staging
drwxr-xr-x   - mbarrio5 hdfs           0 2020-11-26 02:18 dataproj
drwxr-xr-x   - mbarrio5 hdfs           0 2020-11-26 22:57 output
-rwxrwxrwx   2 mbarrio5 hdfs 14710621726 2020-11-26 23:45 testdata.csv
drwxr-xrwx   - mbarrio5 hdfs           0 2020-11-16 21:18 tmp
-bash-4.1$ ls
testdata.csv
-bash-4.1$ chmod 777 testdata.csv
-bash-4.1$ |
```

$ cd ..

$ cd ..

$ cd dev

$ cd shm

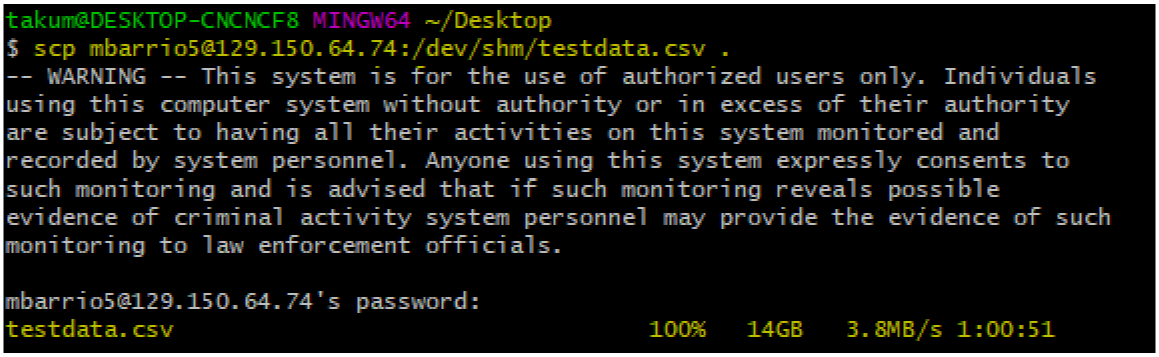
$ ls

$ hdfs dfs -get testdata.csv

$ ls

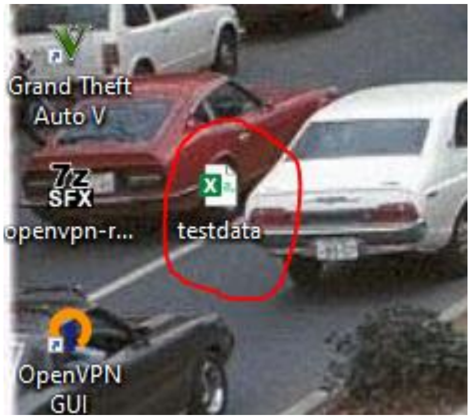
10. SCP to retrieve testdata.csv (note the “ . “ at the end of csv)

$ scp mbarrio5@129.150.64.74:/home/mbarrio5/testdata.csv .

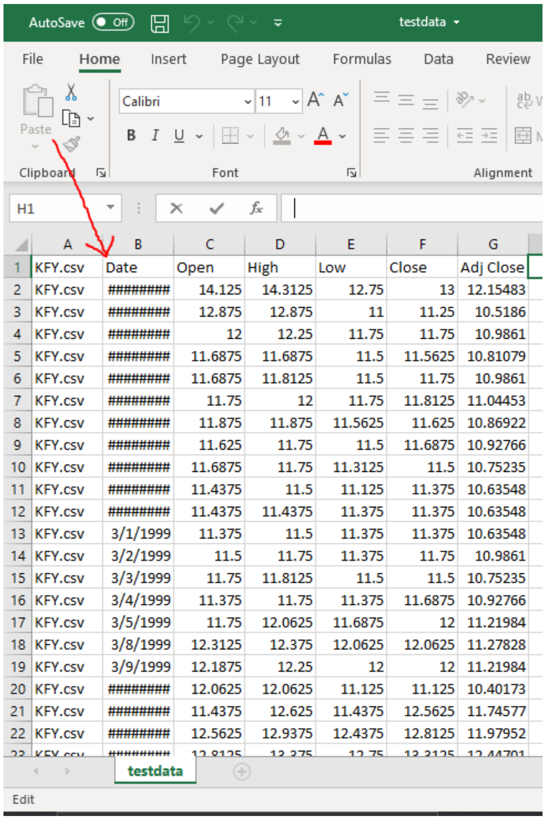


```
takum@DESKTOP-CNCNCF8 MINGW64 ~/Desktop
$ scp mbarrio5@129.150.64.74:/dev/shm/testdata.csv .
-- WARNING -- This system is for the use of authorized users only. Individuals
using this computer system without authority or in excess of their authority
are subject to having all their activities on this system monitored and
recorded by system personnel. Anyone using this system expressly consents to
such monitoring and is advised that if such monitoring reveals possible
evidence of criminal activity system personnel may provide the evidence of such
monitoring to law enforcement officials.

mbarrio5@129.150.64.74's password:
testdata.csv                                  100%   14GB   3.8MB/s 1:00:51
```

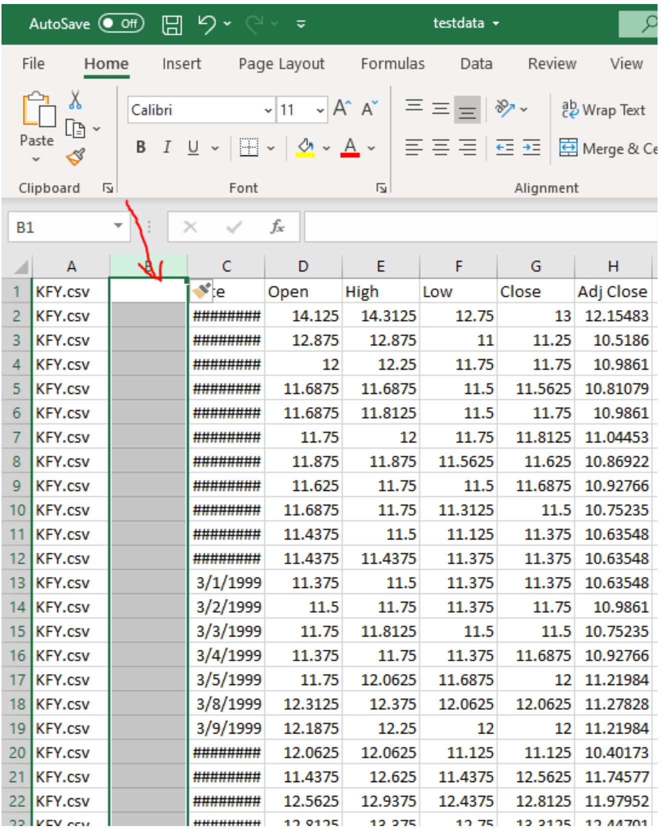
11. Note: a successful file SCP will show the file at your desktop



12. Open testdata and create a new column like so. Then copy ALL of column "A"

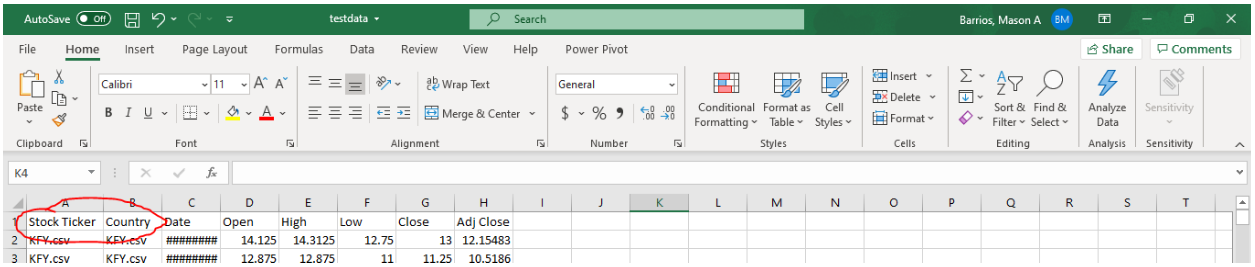


	A	B	C	D	E	F	G
1	KFY.csv	Date	Open	High	Low	Close	Adj Close
2	KFY.csv	#########	14.125	14.3125	12.75	13	12.15483
3	KFY.csv	#########	12.875	12.875	11	11.25	10.5186
4	KFY.csv	#########	12	12.25	11.75	11.75	10.9861
5	KFY.csv	#########	11.6875	11.6875	11.5	11.5625	10.81079
6	KFY.csv	#########	11.6875	11.8125	11.5	11.75	10.9861
7	KFY.csv	#########	11.75	12	11.75	11.8125	11.04453
8	KFY.csv	#########	11.875	11.875	11.5625	11.625	10.86922
9	KFY.csv	#########	11.625	11.75	11.5	11.6875	10.92766
10	KFY.csv	#########	11.6875	11.75	11.3125	11.5	10.75235
11	KFY.csv	#########	11.4375	11.5	11.125	11.375	10.63548
12	KFY.csv	#########	11.4375	11.4375	11.375	11.375	10.63548
13	KFY.csv	3/1/1999	11.375	11.5	11.375	11.375	10.63548
14	KFY.csv	3/2/1999	11.5	11.75	11.375	11.75	10.9861
15	KFY.csv	3/3/1999	11.75	11.8125	11.5	11.5	10.75235
16	KFY.csv	3/4/1999	11.375	11.75	11.375	11.6875	10.92766
17	KFY.csv	3/5/1999	11.75	12.0625	11.6875	12	11.21984
18	KFY.csv	3/8/1999	12.3125	12.375	12.0625	12.0625	11.27828
19	KFY.csv	3/9/1999	12.1875	12.25	12	12	11.21984
20	KFY.csv	#########	12.0625	12.0625	11.125	11.125	10.40173
21	KFY.csv	#########	11.4375	12.625	11.4375	12.5625	11.74577
22	KFY.csv	#########	12.5625	12.9375	12.4375	12.8125	11.97952
23	KFY.csv	#########	12.8125	13.375	12.75	13.3125	12.44701



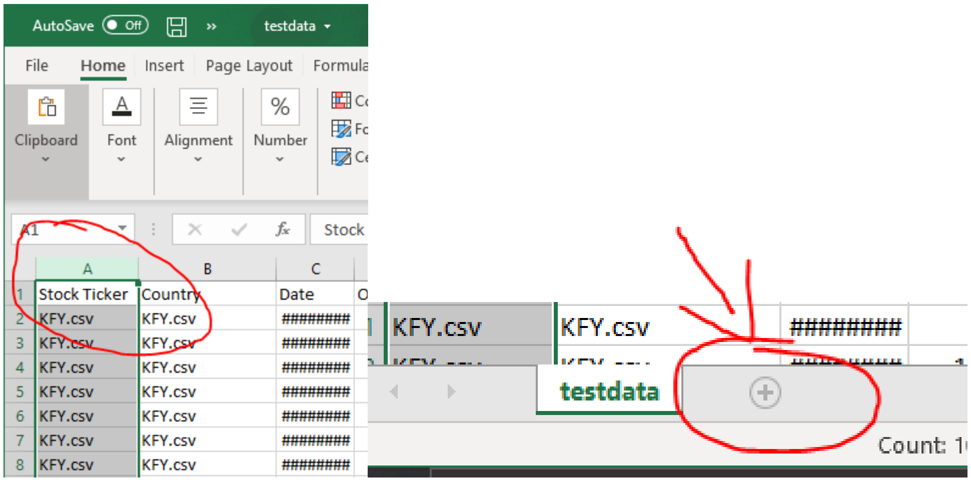
	A	B	C	D	E	F	G	H
1	KFY.csv		te	Open	High	Low	Close	Adj Close
2	KFY.csv		#########	14.125	14.3125	12.75	13	12.15483
3	KFY.csv		#########	12.875	12.875	11	11.25	10.5186
4	KFY.csv		#########	12	12.25	11.75	11.75	10.9861
5	KFY.csv		#########	11.6875	11.6875	11.5	11.5625	10.81079
6	KFY.csv		#########	11.6875	11.8125	11.5	11.75	10.9861
7	KFY.csv		#########	11.75	12	11.75	11.8125	11.04453
8	KFY.csv		#########	11.875	11.875	11.5625	11.625	10.86922
9	KFY.csv		#########	11.625	11.75	11.5	11.6875	10.92766
10	KFY.csv		#########	11.6875	11.75	11.3125	11.5	10.75235
11	KFY.csv		#########	11.4375	11.5	11.125	11.375	10.63548
12	KFY.csv		#########	11.4375	11.4375	11.375	11.375	10.63548
13	KFY.csv		3/1/1999	11.375	11.5	11.375	11.375	10.63548
14	KFY.csv		3/2/1999	11.5	11.75	11.375	11.75	10.9861
15	KFY.csv		3/3/1999	11.75	11.8125	11.5	11.5	10.75235
16	KFY.csv		3/4/1999	11.375	11.75	11.375	11.6875	10.92766
17	KFY.csv		3/5/1999	11.75	12.0625	11.6875	12	11.21984
18	KFY.csv		3/8/1999	12.3125	12.375	12.0625	12.0625	11.27828
19	KFY.csv		3/9/1999	12.1875	12.25	12	12	11.21984
20	KFY.csv		#########	12.0625	12.0625	11.125	11.125	10.40173
21	KFY.csv		#########	11.4375	12.625	11.4375	12.5625	11.74577
22	KFY.csv		#########	12.5625	12.9375	12.4375	12.8125	11.97952
23	KFY.csv		#########	12.8125	13.375	12.75	13.3125	12.44701

13. Paste into the NEW column then rename column A and B like so (Stock Ticker and Country)

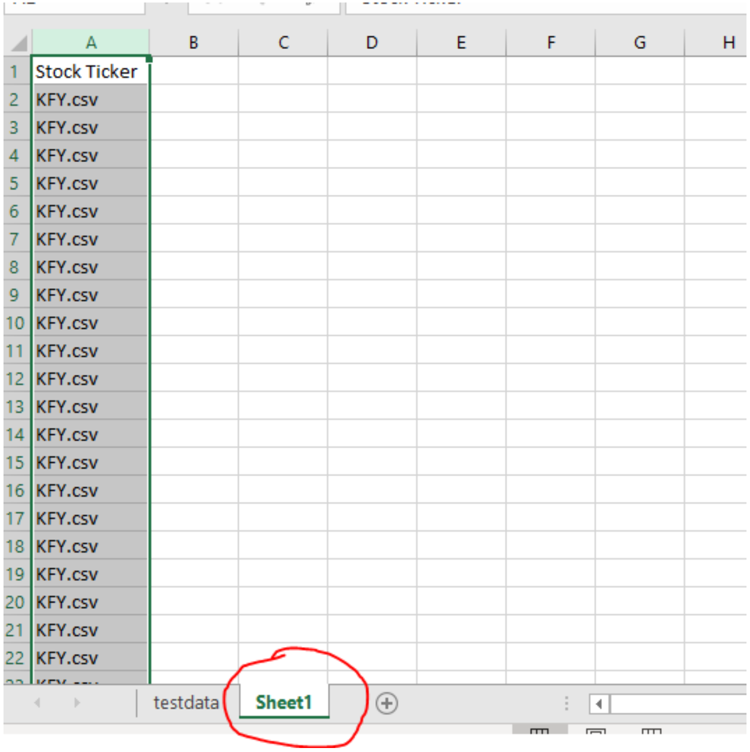


	A	B	C	D	E	F	G	H
1	Stock Ticker	Country	Date	Open	High	Low	Close	Adj Close
2	KFY.csv	KFY.csv	########	14.125	14.3125	12.75	13	12.15483
3	KFY.csv	KFY.csv	########	12.875	12.875	11	11.25	10.5186

14. Now copy column A again, make a new sheet, and paste it into the new sheet. This is to ensure that we did not lose the data when we modify the sheet using the Microsoft Excel Replace tool.



	A	B	C
1	Stock Ticker	Country	Date
2	KFY.csv	KFY.csv	########
3	KFY.csv	KFY.csv	########
4	KFY.csv	KFY.csv	########
5	KFY.csv	KFY.csv	########
6	KFY.csv	KFY.csv	########
7	KFY.csv	KFY.csv	########
8	KFY.csv	KFY.csv	########

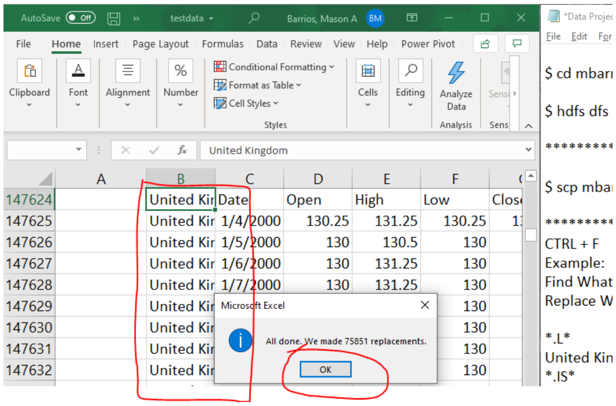


	A
1	Stock Ticker
2	KFY.csv
3	KFY.csv
4	KFY.csv
5	KFY.csv
6	KFY.csv
7	KFY.csv
8	KFY.csv
9	KFY.csv
10	KFY.csv
11	KFY.csv
12	KFY.csv
13	KFY.csv
14	KFY.csv
15	KFY.csv
16	KFY.csv
17	KFY.csv
18	KFY.csv
19	KFY.csv
20	KFY.csv
21	KFY.csv
22	KFY.csv

15. Now we begin the find and replace section. This is to add a location to each stock ticker. This is due to the fact that the dataset did not come with a location dimension. All of the tickers have an extension such as: HK, F, DE which are codenames for countries. We will use the replace tool to find these extensions and replace all matching cells with the proper country (ex: .HK = Hong Kong, .PA = France, .DE = Germany). The following replace entries **MUST** be done in the order specified. **Note:** all us tickers do NOT have an extension so the remaining .csv are tickers in the United States.

Example:
CTRL + F
Example:
Find What = *.L*
Replace With = United Kingdom

Entries for the Replace tool (complete in this order, do not include "|"):
.L United Kingdom | *.SW* Switzerland | *.ST* Sweden | *.MC* Spain | *.KQ* South Korea *.OL* Norway | *.AS* Netherlands | *.TA* Israel | *.JK* Indonesia | *.NS* India | *.BO* India | *.AT* Greece | *.DE* Germany | *.MU* Germany | *.BE* Germany | *.PA* France | *.F* France | *.CO* Denmark | *.SS* China | *.TO* Canada | *.V* Canada | *.SA* Brazil | *.BR* Belgium | *.AX* Australia | *.CSV* United States



United Kingdom

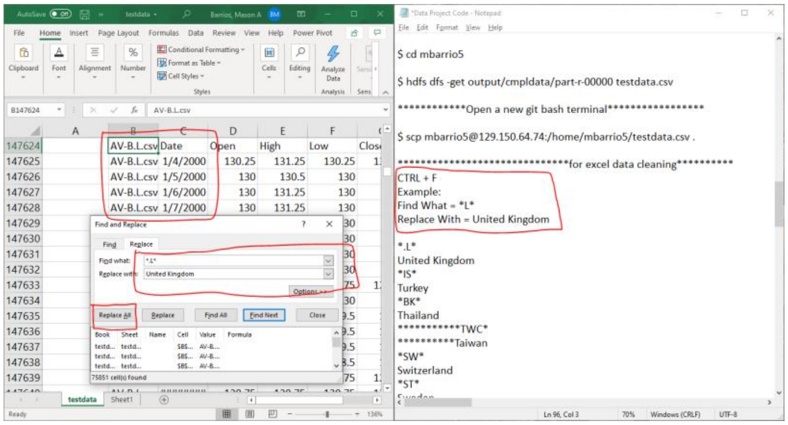
	A	B	C	D	E	F	(
147624		United Kir	Date	Open	High	Low	Clos
147625		United Kir	1/4/2000	130.25	131.25	130.25	1:
147626		United Kir	1/5/2000	130	130.5	130	
147627		United Kir	1/6/2000	130	131.25	130	
147628		United Kir	1/7/2000	130	131.25	130	
147629		United Kir				130	
147630		United Kir				130	
147631		United Kir				130	
147632		United Kir				130	

Microsoft Excel

All done. We made 75851 replacements.

OK

```
$ cd mbari
$ hdfs dfs
*********
$ scp mbai
*********
CTRL + F
Example:
Find What
Replace W

*.L*
United Kin
*.IS*
```



B147624 AV-B.L.csv

	A	B	C	D	E	F	(
147624		AV-B.L.csv	Date	Open	High	Low	Clos
147625		AV-B.L.csv	1/4/2000	130.25	131.25	130.25	1:
147626		AV-B.L.csv	1/5/2000	130	130.5	130	
147627		AV-B.L.csv	1/6/2000	130	131.25	130	
147628		AV-B.L.csv	1/7/2000	130	131.25	130	

Find and Replace

Find | Replace

Find what: *.L*

Replace with: United Kingdom

Options >>

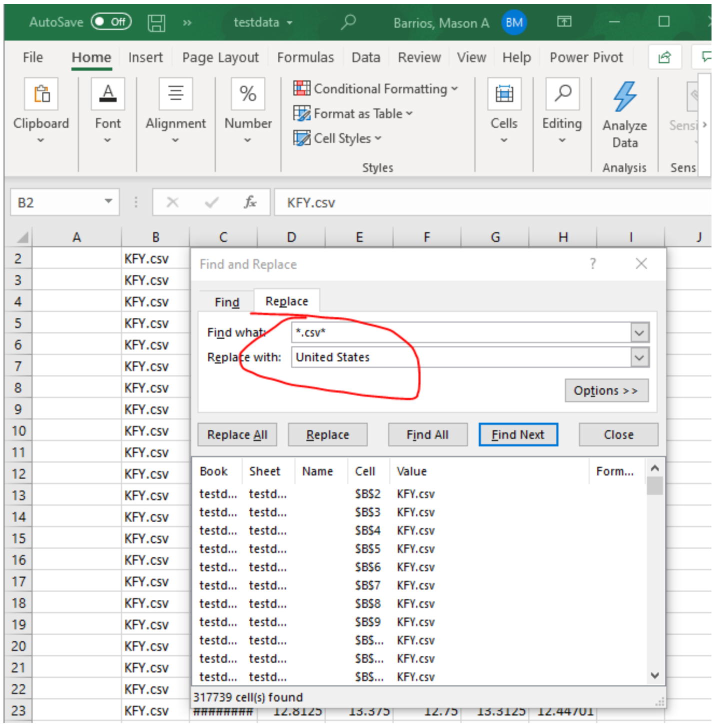
Replace All | Replace | Find All | Find Next | Close

Book	Sheet	Name	Cell	Value	Formula
testd...	testd...		B...	AV-B...	
testd...	testd...		B...	AV-B...	
testd...	testd...		B...	AV-B...	

75851 cell(s) found

```
$ cd mbarrio5

$ hdfs dfs -get output/cmpldata/part-r-00000 testdata.csv

************Open a new git bash terminal*****************

$ scp mbarrio5@129.150.64.74:/home/mbarrio5/testdata.csv .

******************************for excel data cleaning**********
CTRL + F
Example:
Find What = *L*
Replace With = United Kingdom

*.L*
United Kingdom
*IS*
Turkey
*BK*
Thailand
***********TWC*
**********Taiwan
*SW*
Switzerland
*ST*
```

13. Here is the last replace you will do, for the US: Now go to the other sheet in the excel document to copy column A in sheet1 to the empty column (column A) into the sheet test data. Your sheet should like the image on the left.



B2 KFY.csv

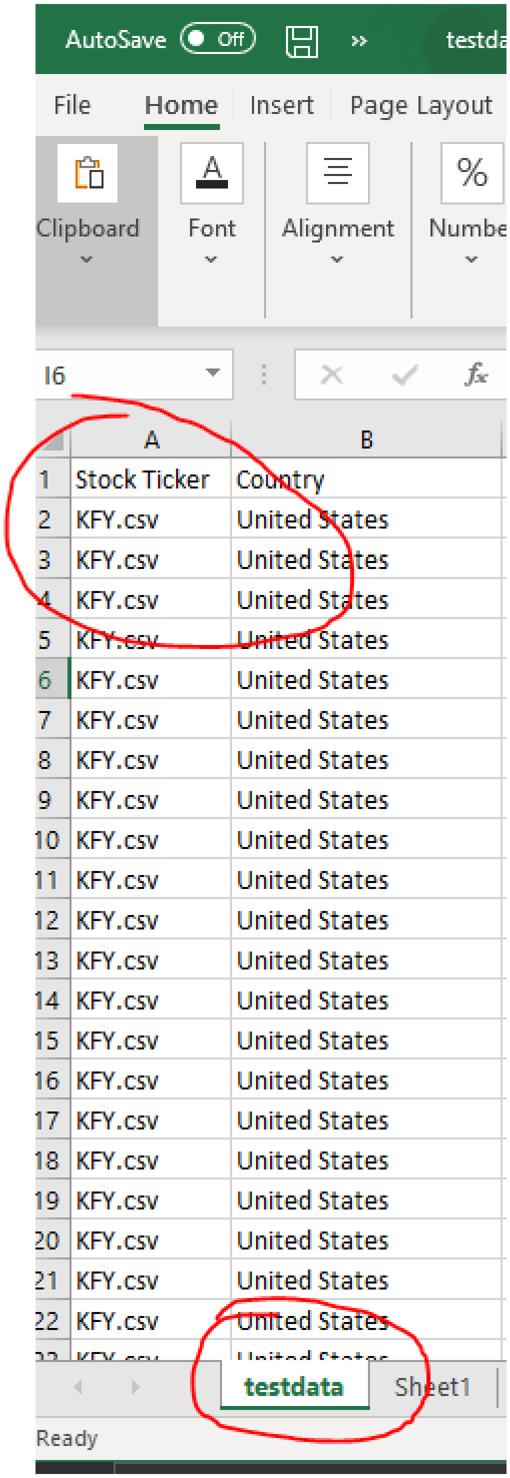
Find what: *.csv*

Replace with: United States

Book	Sheet	Name	Cell	Value	Form...
testd...	testd...		B2	KFY.csv	
testd...	testd...		B3	KFY.csv	
testd...	testd...		B4	KFY.csv	
testd...	testd...		B5	KFY.csv	
testd...	testd...		B6	KFY.csv	
testd...	testd...		B7	KFY.csv	
testd...	testd...		B8	KFY.csv	
testd...	testd...		B9	KFY.csv	
testd...	testd...		B...	KFY.csv	
testd...	testd...		B...	KFY.csv	
testd...	testd...		B...	KFY.csv	

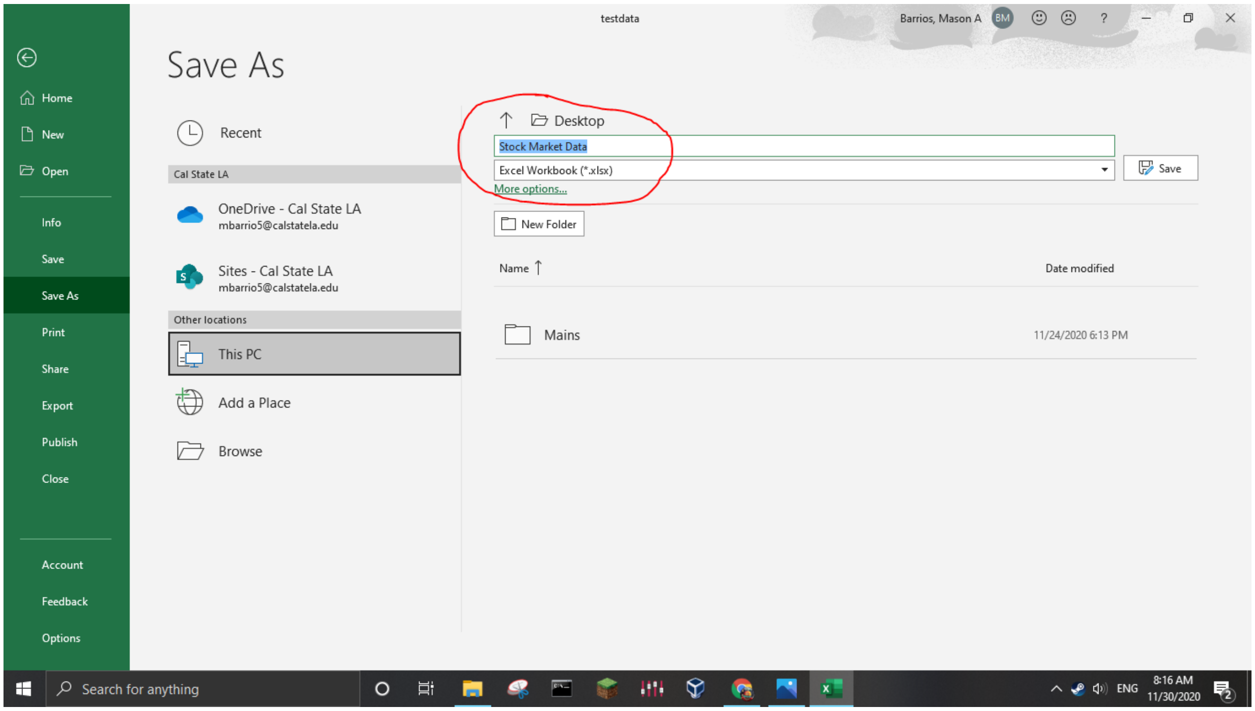
317739 cell(s) found

	A	B	C	D	E	F	G	H
2		KFY.csv						
3		KFY.csv						
4		KFY.csv						
5		KFY.csv						
6		KFY.csv						
7		KFY.csv						
8		KFY.csv						
9		KFY.csv						
10		KFY.csv						
11		KFY.csv						
12		KFY.csv						
13		KFY.csv						
14		KFY.csv						
15		KFY.csv						
16		KFY.csv						
17		KFY.csv						
18		KFY.csv						
19		KFY.csv						
20		KFY.csv						
21		KFY.csv						
22		KFY.csv						
23		KFY.csv	########	12.8125	13.375	12.75	13.3125	12.44701



	A	B
1	Stock Ticker	Country
2	KFY.csv	United States
3	KFY.csv	United States
4	KFY.csv	United States
5	KFY.csv	United States
6	KFY.csv	United States
7	KFY.csv	United States
8	KFY.csv	United States
9	KFY.csv	United States
10	KFY.csv	United States
11	KFY.csv	United States
12	KFY.csv	United States
13	KFY.csv	United States
14	KFY.csv	United States
15	KFY.csv	United States
16	KFY.csv	United States
17	KFY.csv	United States
18	KFY.csv	United States
19	KFY.csv	United States
20	KFY.csv	United States
21	KFY.csv	United States
22	KFY.csv	United States
23	KFY.csv	United States

15. Now go to save as and save the document as “Stock Market Data” and the .xlsx extension. It is necessary to be saved in .xlsx to use power maps.



Save As

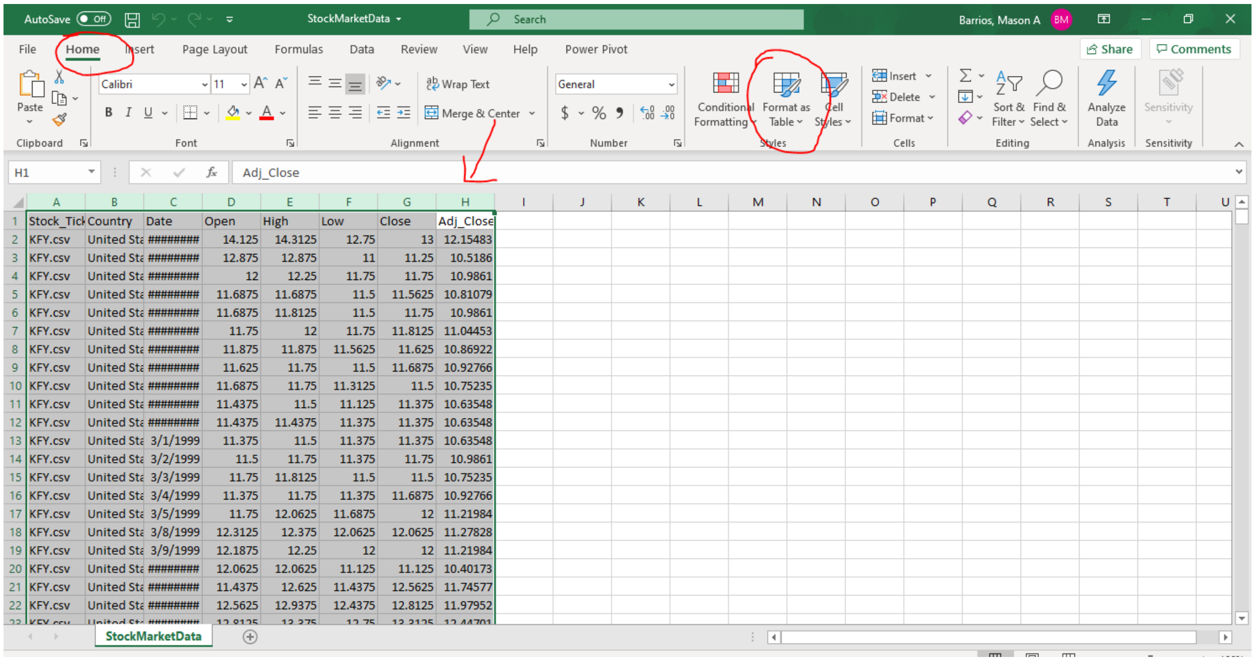
Desktop

Stock Market Data

Excel Workbook (*.xlsx)

Name	Date modified
Mains	11/24/2020 6:13 PM

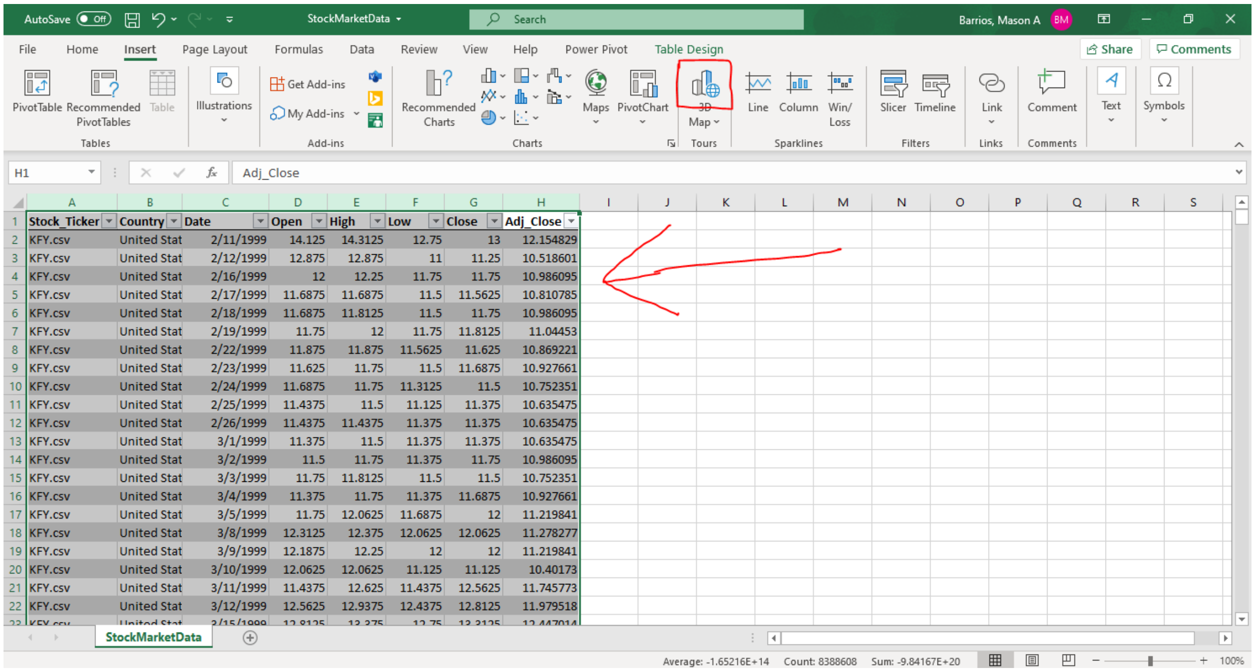
16. Select all then format as table via the home page. Use any design.



H1 Adj_Close

	A	B	C	D	E	F	G	H
1	Stock_Tick	Country	Date	Open	High	Low	Close	Adj_Close
2	KFY.csv	United Sta	########	14.125	14.3125	12.75	13	12.15483
3	KFY.csv	United Sta	########	12.875	12.875	11	11.25	10.5186
4	KFY.csv	United Sta	########	12	12.25	11.75	11.75	10.9861
5	KFY.csv	United Sta	########	11.6875	11.6875	11.5	11.5625	10.81079
6	KFY.csv	United Sta	########	11.6875	11.8125	11.5	11.75	10.9861
7	KFY.csv	United Sta	########	11.75	12	11.75	11.8125	11.04453
8	KFY.csv	United Sta	########	11.875	11.875	11.5625	11.625	10.86922
9	KFY.csv	United Sta	########	11.625	11.75	11.5	11.6875	10.92766
10	KFY.csv	United Sta	########	11.6875	11.75	11.3125	11.5	10.75235
11	KFY.csv	United Sta	########	11.4375	11.5	11.125	11.375	10.63548
12	KFY.csv	United Sta	########	11.4375	11.4375	11.375	11.375	10.63548
13	KFY.csv	United Sta	3/1/1999	11.375	11.5	11.375	11.375	10.63548
14	KFY.csv	United Sta	3/2/1999	11.5	11.75	11.375	11.75	10.9861
15	KFY.csv	United Sta	3/3/1999	11.75	11.8125	11.5	11.5	10.75235
16	KFY.csv	United Sta	3/4/1999	11.375	11.75	11.375	11.6875	10.92766
17	KFY.csv	United Sta	3/5/1999	11.75	12.0625	11.6875	12	11.21984
18	KFY.csv	United Sta	3/8/1999	12.3125	12.375	12.0625	12.0625	11.27828
19	KFY.csv	United Sta	3/9/1999	12.1875	12.25	12	12	11.21984
20	KFY.csv	United Sta	########	12.0625	12.0625	11.125	11.125	10.40173
21	KFY.csv	United Sta	########	11.4375	12.625	11.4375	12.5625	11.74577
22	KFY.csv	United Sta	########	12.5625	12.9375	12.4375	12.8125	11.97952
23	KFY.csv	United Sta	########	12.8125	13.375	12.75	13.3125	12.44701

17. Now select all again and click on 3D map

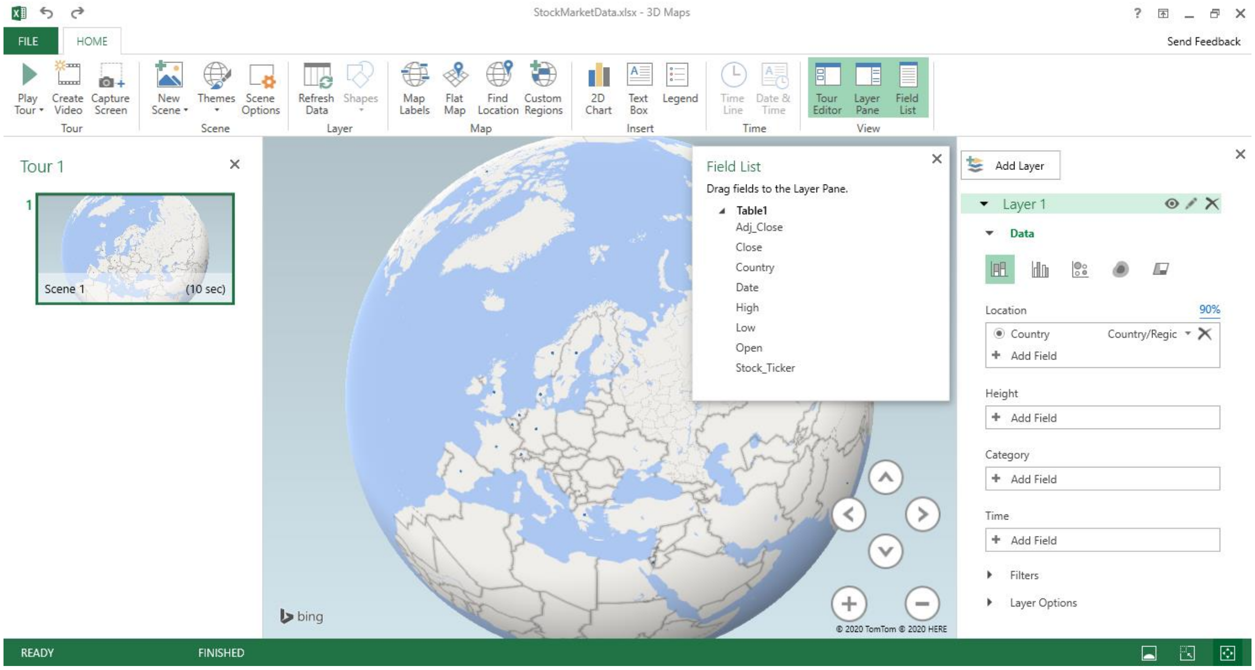


H1 Adj_Close

	A	B	C	D	E	F	G	H
1	Stock_Ticker	Country	Date	Open	High	Low	Close	Adj_Close
2	KFY.csv	United Stat	2/11/1999	14.125	14.3125	12.75	13	12.154829
3	KFY.csv	United Stat	2/12/1999	12.875	12.875	11	11.25	10.518601
4	KFY.csv	United Stat	2/16/1999	12	12.25	11.75	11.75	10.986095
5	KFY.csv	United Stat	2/17/1999	11.6875	11.6875	11.5	11.5625	10.810785
6	KFY.csv	United Stat	2/18/1999	11.6875	11.8125	11.5	11.75	10.986095
7	KFY.csv	United Stat	2/19/1999	11.75	12	11.75	11.8125	11.04453
8	KFY.csv	United Stat	2/22/1999	11.875	11.875	11.5625	11.625	10.869221
9	KFY.csv	United Stat	2/23/1999	11.625	11.75	11.5	11.6875	10.927661
10	KFY.csv	United Stat	2/24/1999	11.6875	11.75	11.3125	11.5	10.752351
11	KFY.csv	United Stat	2/25/1999	11.4375	11.5	11.125	11.375	10.635475
12	KFY.csv	United Stat	2/26/1999	11.4375	11.4375	11.375	11.375	10.635475
13	KFY.csv	United Stat	3/1/1999	11.375	11.5	11.375	11.375	10.635475
14	KFY.csv	United Stat	3/2/1999	11.5	11.75	11.375	11.75	10.986095
15	KFY.csv	United Stat	3/3/1999	11.75	11.8125	11.5	11.5	10.752351
16	KFY.csv	United Stat	3/4/1999	11.375	11.75	11.375	11.6875	10.927661
17	KFY.csv	United Stat	3/5/1999	11.75	12.0625	11.6875	12	11.219841
18	KFY.csv	United Stat	3/8/1999	12.3125	12.375	12.0625	12.0625	11.278277
19	KFY.csv	United Stat	3/9/1999	12.1875	12.25	12	12	11.219841
20	KFY.csv	United Stat	3/10/1999	12.0625	12.0625	11.125	11.125	10.40173
21	KFY.csv	United Stat	3/11/1999	11.4375	12.625	11.4375	12.5625	11.745773
22	KFY.csv	United Stat	3/12/1999	12.5625	12.9375	12.4375	12.8125	11.979518
23	KFY.csv	United Stat	3/15/1999	12.8125	13.375	12.75	13.3125	12.447014

Average: -1.65216E+14 Count: 8388608 Sum: -9.84167E+20 100%

Here is how the 3D map looks like upon opening



Field List

Drag fields to the Layer Pane.

- Table1
  - Adj_Close
  - Close
  - Country
  - Date
  - High
  - Low
  - Open
  - Stock_Ticker

Add Layer

Layer 1

Data

Location 90%

Country Country/Regic

+ Add Field

Height

+ Add Field

Category

+ Add Field

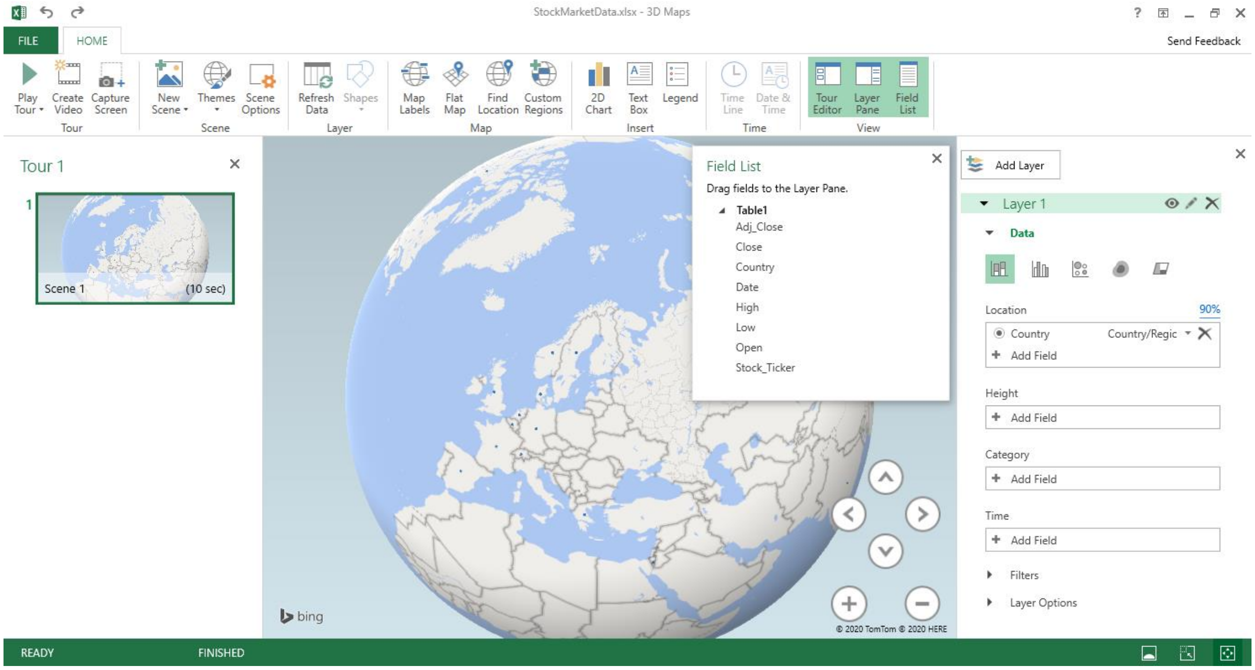
Time

+ Add Field

Filters

Layer Options

16. Now we can select “stock ticker” as the category, adj close as the size, and use the pie chart for visualization. We can click on the any section of the chart in a region and find the various SUM of the various stock tickers.



Field List
Drag fields to the Layer Pane.
Table1
Adj_Close
Close
Country
Date
High
Low
Open
Stock_Ticker

Add Layer
Layer 1
Data
Location 90%
Country Country/Regic
+ Add Field
Height
+ Add Field
Category
+ Add Field
Time
+ Add Field
Filters
Layer Options