

LAB 1

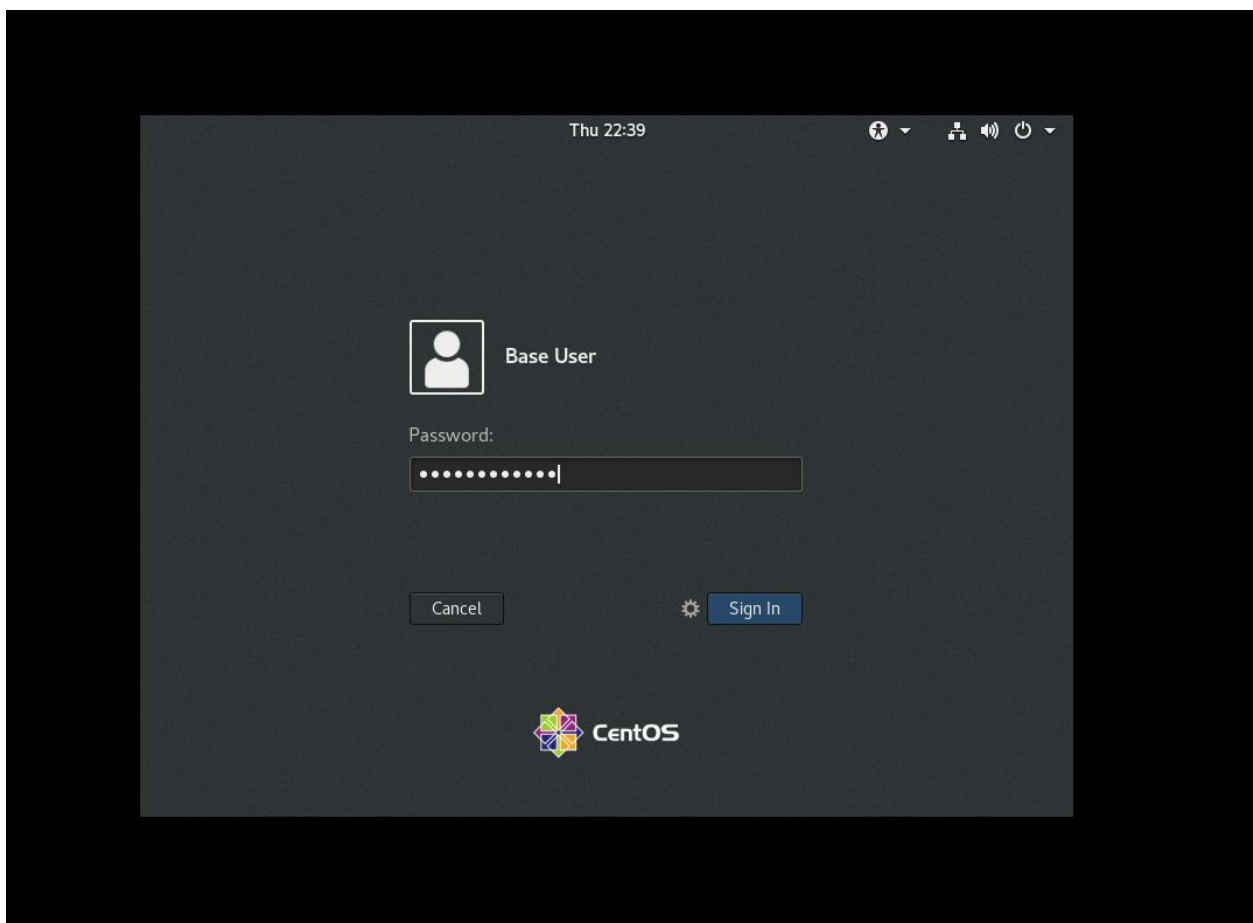
Name: Karim Gamal Mahmoud Mohamed

E-Mail: 21KGMM@queensu.ca

- 1) The virtual machine login (assuming you have a Cloudera VM installed at your location)

Note: The user : asosboxes

Password : BaseUser@123



2) List all the files and directories under /user/osboxes in HDFS

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2020-06-02 19:18 /user/hdfs
drwxrwxrwx - mapred hadoop 0 2020-06-02 18:10 /user/history
drwxrwxrwx - hive hive 0 2020-06-02 18:10 /user/hive
drwxrwxrwx - hue hue 0 2020-06-02 18:15 /user/hue
drwxrwxrwx - impala impala 0 2020-06-02 18:08 /user/impala
drwxrwxrwx - osboxes osboxes 0 2022-06-16 22:16 /user/osboxes
drwxrwxrwx - spark spark 0 2020-06-02 18:08 /user/spark
drwxrwxrwx - hdfs supergroup 0 2020-06-02 18:08 /user/yarn
```

3) Create a directory with name 'inputdata' in HDFS






(command: `hdfs dfs -mkdir directoryname`)

&

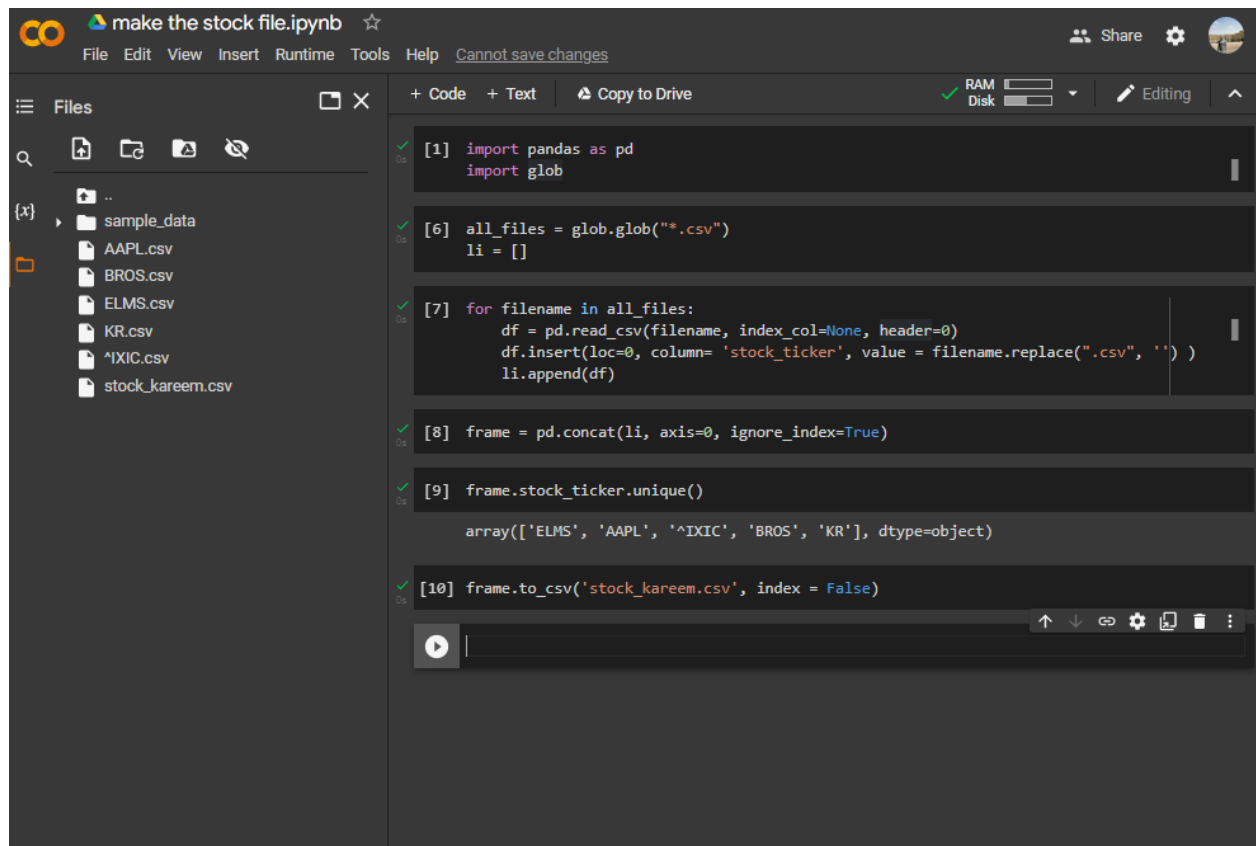
4) List the folders in /user/osboxes within hdfs

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -mkdir /user/osboxes/inputdata
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes
Found 1 items
drwxr-xr-x  - osboxes osboxes  _      0 2022-06-16 22:27 /user/osboxes/inputdata
```

5) Now, from your local machine, download historical stock data (csv format) of at least 5 stock quotes (random) from yahoo financial (finance.yahoo.com) and put them in a local folder of your machine.

 ^IXIC.csv	2022-06-16 7:32 PM	Microsoft Excel C...	22 KB
 AAPL.csv	2022-06-16 7:32 PM	Microsoft Excel C...	19 KB
 BROS.csv	2022-06-16 7:32 PM	Microsoft Excel C...	13 KB
 ELMS.csv	2022-06-16 7:32 PM	Microsoft Excel C...	16 KB
 KR.csv	2022-06-16 7:32 PM	Microsoft Excel C...	18 KB

6) In each of the files put the stock ticker (e.g. TSLA) in the first column of the file. Similarly do the same for others and finally merge them in one file named 'stock.csv'.



The screenshot shows a Jupyter Notebook titled "make the stock file.ipynb". The left sidebar displays a file explorer with a directory structure including "sample_data" and several CSV files: "AAPL.csv", "BROS.csv", "ELMS.csv", "KR.csv", "IXIC.csv", and "stock_kareem.csv". The main area contains a series of code cells, each with a green checkmark and a "0s" runtime indicator. The code performs the following steps:

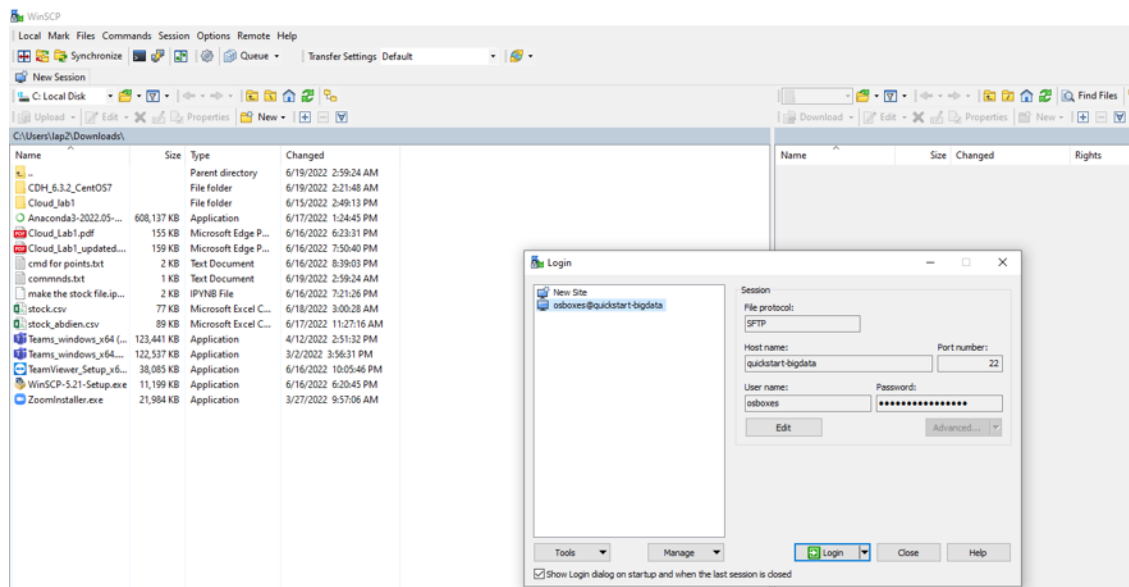
- Imports pandas as `pd` and glob as `glob`.
- Uses `glob.glob("*.csv")` to find all CSV files and initializes an empty list `li = []`.
- Iterates over each filename, reads the CSV file using `pd.read_csv`, inserts a new column named "stock_ticker" with the filename (replacing ".csv" with an empty string), and appends the resulting DataFrame to the list `li`.
- Concatenates all DataFrames in the list `li` along the columns axis using `pd.concat`, ignoring the index.
- Extracts the unique stock tickers from the "stock_ticker" column using `frame.stock_ticker.unique()`, resulting in an array: `array(['ELMS', 'AAPL', '^IXIC', 'BROS', 'KR'], dtype=object)`.
- Saves the final concatenated DataFrame to a new CSV file named "stock_kareem.csv" using `frame.to_csv`, with the index set to `False`.

At the bottom of the code cells, there is a play button icon and a progress bar.

7) Install Winscp and connect to VM to transfer the stock.csv file from your local machine to the VM (in a new folder called 'data' under /home/osboxes).

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -mkdir /user/osboxes/data
[osboxes@quickstart-bigdata ~]$
```

```
[osboxes@quickstart-bigdata root]$ hdfs dfs -ls /user/osboxes
Found 3 items
drwx-----  - osboxes osboxes          0 2022-06-19 06:30 /user/osboxes/.Trash
drwxr-xr-x  - osboxes osboxes          0 2022-06-19 06:41 /user/osboxes/data
drwxr-xr-x  - osboxes osboxes          0 2022-06-19 06:36 /user/osboxes/inputdata
[osboxes@quickstart-bigdata root]$
```



8) Within VM, copy the file stock.csv from /home/osboxes/data (local file system) to 'inputdata' directory in HDFS that are created earlier

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -copyFromLocal /home/osboxes/data/stock_kareem.csv /user/osboxes/inputdata
[osboxes@quickstart-bigdata ~]$ hdfs dfs -copyFromLocal /home/osboxes/data/stock_kareem.csv /user/osboxes/inputdata
copyFromLocal: '/user/osboxes/inputdata/stock_kareem.csv': File exists
```


9) Show the content of the last few records in the stock.csv file in HDFS.

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -tail /user/osboxes/inputdata/stock_kareem.csv
2022-05-26,31.09,37.980999,30.950001,35.720001,35.720001,3251400
BROS,2022-05-27,35.75,39.990002,35.75,37.360001,37.360001,2109100
BROS,2022-05-31,37.5,38.736,36.360001,37.549999,37.549999,1584900
BROS,2022-06-01,37.470001,39.939999,36.022999,39.849998,39.849998,1454900
BROS,2022-06-02,39.490002,42.429001,39.196999,39.98,39.98,1560000
BROS,2022-06-03,39.66,41.959999,38.349998,41.709999,41.709999,1027200
BROS,2022-06-06,42.380001,43.490002,41.310001,41.700001,41.700001,1170900
BROS,2022-06-07,41.200001,42.66,40.060001,40.259998,40.259998,1233000
BROS,2022-06-08,38.330002,39.646999,37.119999,39.25,39.25,1747100
BROS,2022-06-09,38.32,40.59,37.611,39.16,39.16,1177700
BROS,2022-06-10,37.700001,37.950001,34.130001,34.400002,34.400002,1650400
BROS,2022-06-13,32.779999,33.84,30.280001,33.110001,33.110001,1747100
BROS,2022-06-14,33.150002,34.324001,32.380001,33.41,33.41,1094700
BROS,2022-06-15,33.950001,35.689999,32.650002,34.169998,34.169998,1501400
BROS,2022-06-16,32.369999,34.149899,30.780001,33.605,33.605,1591084
[osboxes@quickstart-bigdata ~]$
```

10) Make a copy of the stock.csv file as stock1.csv from the current hdfs location to another folder named 'testdata' in hdfs.

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls
Found 3 items
drwx-----  - osboxes osboxes      0 2022-06-19 06:30 .Trash
drwxr-xr-x  - osboxes osboxes      0 2022-06-19 07:40 data
drwxr-xr-x  - osboxes osboxes      0 2022-06-19 08:19 inputdata
[osboxes@quickstart-bigdata ~]$ hdfs dfs -mkdir /user/osboxes/testdata
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls
Found 4 items
drwx-----  - osboxes osboxes      0 2022-06-19 06:30 .Trash
drwxr-xr-x  - osboxes osboxes      0 2022-06-19 07:40 data
drwxr-xr-x  - osboxes osboxes      0 2022-06-19 08:19 inputdata
drwxr-xr-x  - osboxes osboxes      0 2022-06-19 08:27 testdata
```

```
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cp /user/osboxes/inputdata/stock_kareem.csv /user/osboxes/testdata/stock1_kareem.csv
[osboxes@quickstart-bigdata ~]$ hdfs dfs -tail /user/osboxes/testdata/stock1_kareem.csv
2022-05-26,31.09,37.980999,30.950001,35.720001,35.720001,3251400
BROS,2022-05-27,35.75,39.990002,35.75,37.360001,37.360001,2109100
BROS,2022-05-31,37.5,38.736,36.360001,37.549999,37.549999,1584900
BROS,2022-06-01,37.470001,39.939999,36.022999,39.849998,39.849998,1454900
BROS,2022-06-02,39.490002,42.429001,39.196999,39.98,39.98,1560000
BROS,2022-06-03,39.66,41.959999,38.349998,41.709999,41.709999,1027200
BROS,2022-06-06,42.380001,43.490002,41.310001,41.700001,41.700001,1170900
BROS,2022-06-07,41.200001,42.66,40.060001,40.259998,40.259998,1233000
BROS,2022-06-08,38.330002,39.646999,37.119999,39.25,39.25,1747100
BROS,2022-06-09,38.32,40.59,37.611,39.16,39.16,1177700
BROS,2022-06-10,37.700001,37.950001,34.130001,34.400002,34.400002,1650400
BROS,2022-06-13,32.779999,33.84,30.280001,33.110001,33.110001,1747100
BROS,2022-06-14,33.150002,34.324001,32.380001,33.41,33.41,1094700
BROS,2022-06-15,33.950001,35.689999,32.650002,34.169998,34.169998,1501400
BROS,2022-06-16,32.369999,34.149899,30.780001,33.605,33.605,1591004
[osboxes@quickstart-bigdata ~]$
```