Stock Tickers with the Highest Increase: 1996 to 2020

Mason Barrios, Joaquin Flores, Jasmine Gomez, Dylan Mora

Department of Information Systems, California State University

Los Angeles

Tel. 323-343-2916, Fax. 323-343-5209

e-mail: masonxbarrios@gmail.com

**Abstract:** Our research takes a look at the stock market data from 1996 to 2020. We use the data set to find which stocks per country have the highest increases in closing prices.

**Introduction:** Our research takes a look at the stock market data from 1996 to 2020. We use the data set to find which stocks per country have the highest increases in closing prices. Getting our data begins with a google search. We make our way to Kaggle.com, a website that has many datasets for anyone to use. These range from 1 megabyte to 20 gigabytes in size. Our research revolves around a dataset that includes a portion of stock tickers and stock data dating from 1996 to 2020.

While this dataset (provided by Dip Modi from Kaggle) does not have every single stock ticker in existence, it has a large enough sample size to get an idea of the types of industries that flourish in each country. We must then use tools to analyze the data. This is done by storing the data onto a Hadoop File System then using Pig scripts to compile all 12 gigabytes (or the 104,123 individual .csv files which contain records of each stock ticket). The Pig script then writes all 104,123 stock tickers to a single a .csv file. From there we perform some data cleanup in excel which allows us to use the 3D Maps function. This presents the data geo-spatially and is interactive. Once we visualize the data, we can select any given stock and see its SUM value compared to other stock tickers. The higher the SUM value, the greater the change over the course of 24 years.

**1.0 Specifications**

The cluster we utilize is virtual linux machine in conjuction with the hadoop file system. THe hadoop cluster version is 2.7.1.2.4.2.0-258. The central processing unit is multiple Intel(R) Xeon(R) E5-2699C v4s with a clock speed of 2.20GHz.

**1.1 Getting the Data**

Getting our data begins with a google search. We make our way to Kaggle.com, a website that has many datasets for anyone to use. These range from 1 megabyte to 20 gigabytes in size. Our research revolves around a dataset that includes a portion of stock tickers and stock data dating from 1996 to 2020.

Now we open a terminal to SSH to our Linux machine and Hadoop File System. Once we SSH in, we need to retrieve the data set. The data set is quite large and in some systems, you are unable to simply download the data. In our case, we need to navigate to the temporary file system in Linux. In the temporary file system, there is 30 gigabytes of storage which is plentiful.

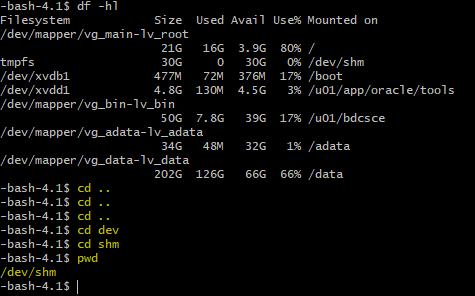


Figure 1. Temporary File System

Once there, we can wget the dataset to the linux side which takes around 30 minutes. After the download is complete, we can then unzip the folder. Upon unzipping, are left with a folder labled “data”. Once unzipped, the data is moved from linux into the hadoop file system. Now that the data is on the hadoop file system, we can progress to the next step which is data translation.

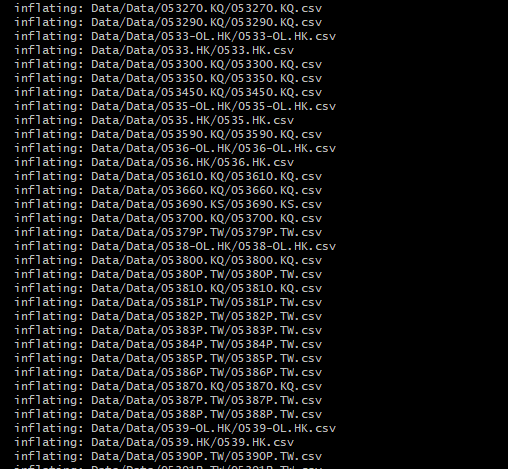


Figure 2. Unzipping the data to the linux machine

After moving the data to the hadoop file system, we can use the rm –r funtion to delete the files off of the linux machine. In our case, this is necessary as the machine is shared with other classmates.

**1.2 Data Analysis**

There are different methods of data analysis. In this course we were introduced to two tools. These tool include apache pig and hive. For this particular project, we utilized apache pig. Our data was quite simple in nature. Most datasets have multiple dimensions and columns but ours essentially had 2. This included the stock ticker as well as the adjusted closing price.

With the dataset being quite large at 12 gigabytes uncompressed, we wanted to test the schema against a small portion of the data. In this case, we declared the date, opening price, high price, low price, closing price, adjusted closing price, and volume columns.

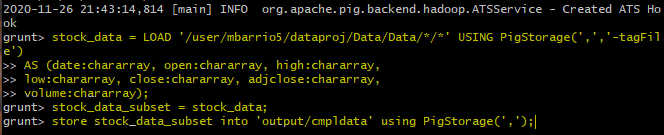


Figure 4. Data Schema for the stock tickers

On a side note, further into the project we realized that we could have omitted the other columns such as low price, high price, and volume. By omitting them, it would speed up the data analysis. The schema is tested against 0002.HK which outputs the declared columns as well as a column with the name of the stock. We use the –tagFile function in pig storage to do this. This works as the name of the file is the same as the name of the stock ticker, which is not included anywhere in the data set. Then we create the stock\_data\_subset to see the output.

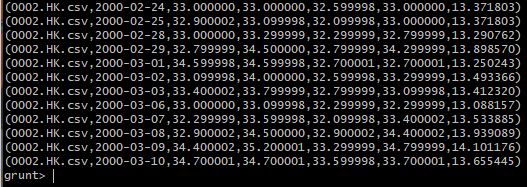


Figure 3. Example output from stock\_data\_subset

After the script checks out, we can finalize the pig script. This is done by adding astericks in the LOAD statement as to include all data files. From here, the output of the data analysis is stored into the directory folder cmpldata.

One issue we ran into after storing the data is the fact that the data was put into 99 different files. We used the –cat funtion to put the entire directory of cmpldata into a single .csv file. This made looking at our findings much easier as all of the data as placed in a single location.

We then do a –get command to download the data file to the linux system. In another terminal, we do an SCP command to download the file to the local machine. From this point on, we can use excel view our findings.

**1.3 Findings**

Before we dive into the findings, some data cleaning was required. In the data set we used, there was no column for location. However, each stock ticker had a tag on the end of it such as HK, PA, or DE. This was the country the stock ticker belonged to and the author of the data set provided a key so that anyone can identify where the stock ticker is from. We created a duplicate column of the stock tickers, this would be used for the country dimension of the data. Then the find and replace tool was utilized. Each extension was searched and each cell was replaced with its corresponding country. For example, if the tag was HK, then the country cell would be replaced as Hong Kong. Or if the the stock had the tag PA, then the country cell would be replaced as France.

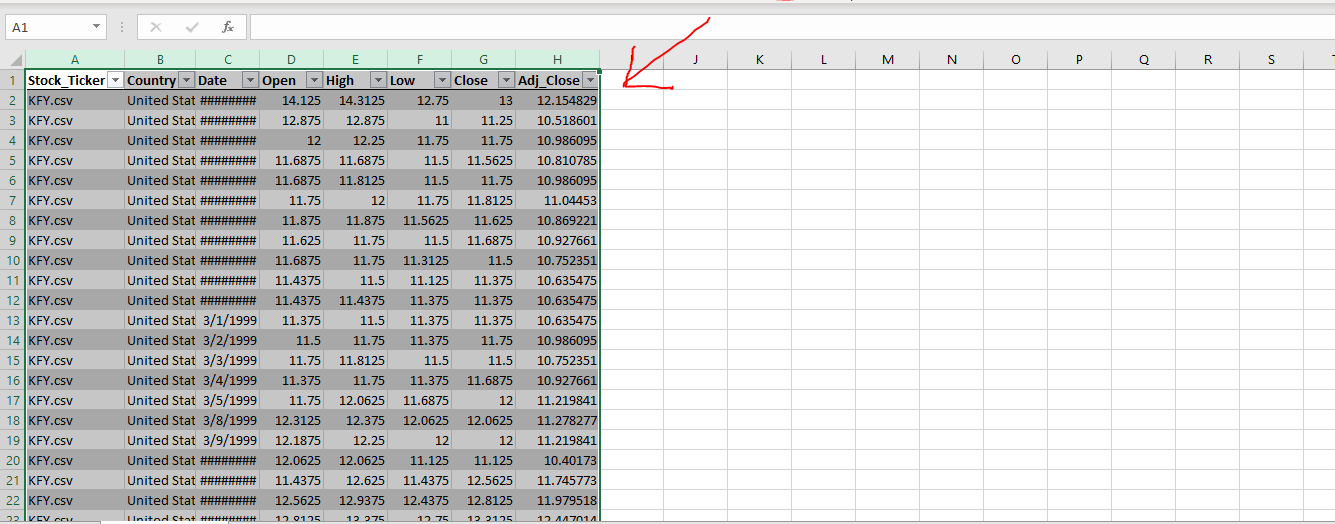


Figure 5. Dataset on excel

With the dataset, we wanted to find which stock showed the greatest increases over the course of 24 years. This would be visualized using the 3D powermaps tool in Excel.

Our findings are as follows. In the united states, global payment technologies a banking company, saw an increase of $6,194 over the course of 24 years. Altamir SCA a finance firm in France saw an increase of $6,080. Gowing asset management in Australia also a finance firm saw an increase of $6,273. JP Morgan, a banking company in the United Kingdom saw an increase of $6,278. In Canada, Petrol Shale an energy company, saw an increase of $6,113. Lastly, Jin Yang Pharmaceutical Co in Korea saw an increase in $5,035.

**1.4 Related Works**

The stock market is continuously changing daily. Stocks could be at an all-time high one hour and crash at another. Many methods are relevant to predicting the closing costs of stocks; in this section, we will mention some of the methods used to predict stock closing prices.

**Long short-term memory model (LSTM)**

LSTM is a technique subtle for processing and predicting important events. LSTM is an extended variant of RNN, a deep learning model that is good at processing time-series data. (Chen, Zhang & Lou, 2020) LSTM determines whether the input should be remembered or not, and initial inputs determine the results' forecasting. This model has been used in many data analyses because it is more accurate than other models.

Still, it is not the best at predicting nonlinear, unstable data, or complex data. With this model, it is prevalent to see other models being used. In the article "An Innovative Neutral Network Approach for Stock Market Predictions," we know that they use LSTM and the deep neural network to "forecast trading data of the Amazon stock and found that the effect of the deep neutral network was better than LSTM; prediction accuracy was 54%. (Pang, Zhou, Wang, Lin & Chang, 2018) The LSTM model needs to be paired with other models to predict more accurately stock market closing prices.

**Empirical Mode Decomposition (EMD)**

EMD is a signal analysis method; it can decompose a complex signal into a finite intrinsic mode function (IMF). (Jin, Yang & Liu, 2019) The data needs to have at least two extreme values; with this, the local time-domain can be determined. Because stock price data is so volatile, this is a great way to predict a change or price in the stock. By looking at this model, we can see a significant advantage; instead of using LSTM. EDM can also extract trend terms of complex stock pricing sequences by decomposing them into more predictable lines. Even though EMD was proposed for other disciplines such as biomedical engineering and structured health monitoring, it can be used for economics and predicting stock market changes.

Even though our data did not use any of these models, it shows the stock's sum over 24 years. The higher the sum is for that particular stock, the more the stock price rose. Using 3D maps in excel, we can see what store had a more significant increase in value and location. By using HDFS and Pig, we can organize the data to see how the prices rose or stayed the same.

**1.5 Conclusion**

Out of the countries sampled, financial institutions or finance related stock tickers saw the highest gains over the course of 24 years. For the US, France, UK, and Australia this was the case. For Korea it was a pharmaceutical company, and for Canada, an energy company.

### References

[1] Chen, Q., Zhang, W., & Lou, Y. (2020). Forecasting Stock Prices Using a Hybrid Deep Learning Model Integrating Attention Mechanism, Multi-Layer Perceptron, and Bidirectional Long-Short Term Memory Neural Network. *IEEE Access*, *8*, 117365-117376. doi: 10.1109/access.2020.3004284

[2] Jin, Z., Yang, Y., & Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing And Applications*, *32*(13), 9713-9729. doi: 10.1007/s00521-019-04504-2

[3] Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2018). An innovative neural network approach for stock market prediction. *The Journal Of Supercomputing*, *76*(3), 2098-2118. doi: 10.1007/s11227-017-2228-y