DS8003 Final Project: Stock market analysis

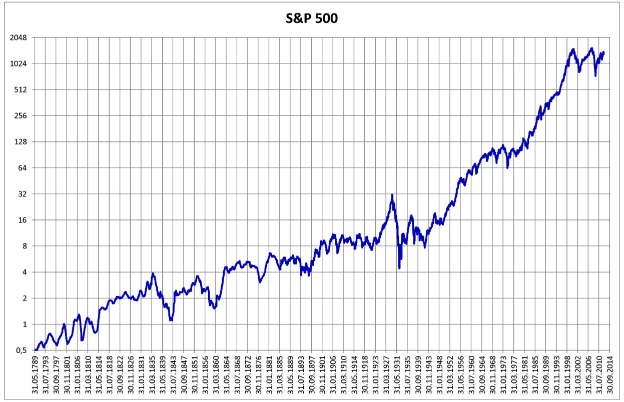
Adam Azoulay
Hee Kyoung Nam
Peter Yu

The Problem

- Stock market → Highly researched complicated data ecosystem

- Data → constantly changing and we want to be up to date with market

- Existing tools/sites exist → Costly, doesn’t meet our needs



Work Distribution

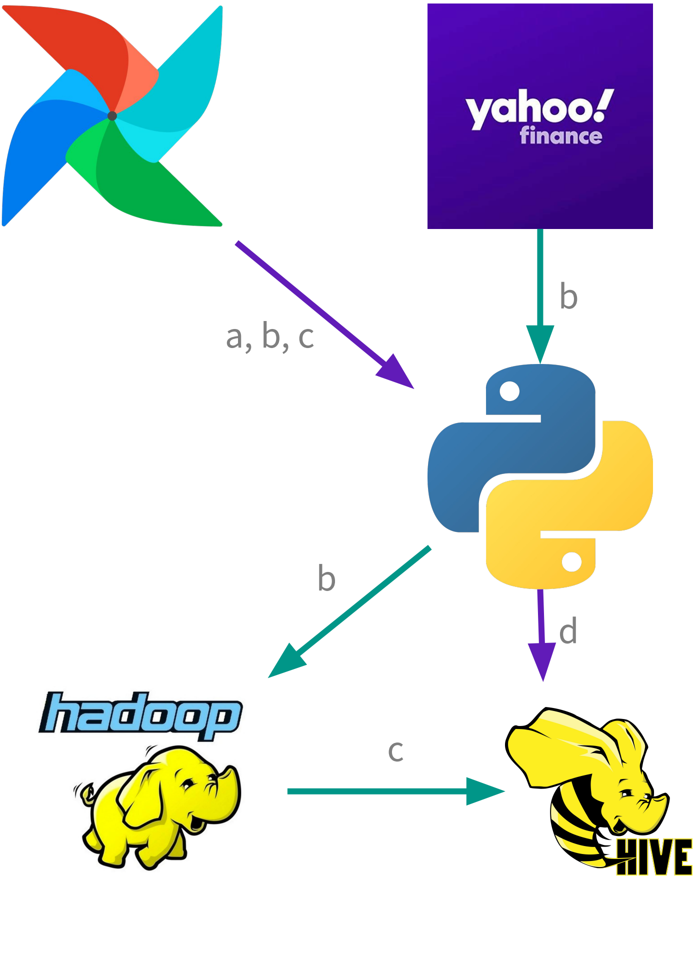
- Adam
    - Infrastructure, ETL (Airflow → Hadoop → Hive → Kibana + Spark)
- Hee
    - Data analysis using Spark and Spark SQL
- Peter
    - Data visualization with Kibana

Data Sources

- NASDAQ Stock Screener
  - Ticker name and sector information, along with volumes and market cap data
- Yahoo Finance API
  - Chart data to track price movements (/v8/finance/chart)
  - Options information (/v7/finance/options)
  - Recent news (/v1/finance/search)
  - Sustainability report (scraped from yahoo sustainability page)
  - Recommendation information

Processing & Tools

- The data from the endpoints goes through a few steps before it end up in our Hive DB
  a. Airflow runs a python script to download the desired data from Yahoo
  b. From memory we format the data and write it to our hdfs in a staging folder
  c. We execute a command from Airflow which loads the staged file into a hive table, moving the data to the final location
  d. Airflow executes any processing in the form of HiveQL queries to refresh our analysis tables
  e. If desired, we can have Airflow notify us via message or email about the status and a summary of the results



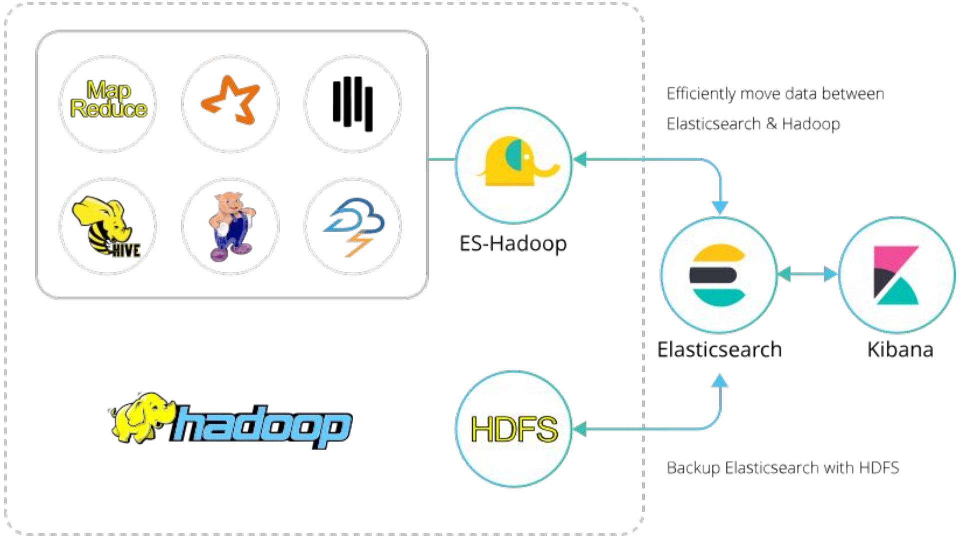
Processing & Tools

- The data in hdfs is organized by hive, and will sit in staging if the pipeline fails which is nice to have
- When we load into Hive, we make sure to only insert new rows, so we prevent duplicates!
- At the end of the ETL we can query all of our updated data in HiveQL!

Code available here

Processing & Tools

- Having loaded the data into Hive tables, we can create external Elasticsearch tables and insert our Hive data
- We can then run Kibana, define an index pattern and use the Kibana UI


Map Reduce
HIVE
hadoop
ES-Hadoop
HDFS
Elasticsearch
Kibana
Efficiently move data between Elasticsearch & Hadoop
Backup Elasticsearch with HDFS

Processing & Tools

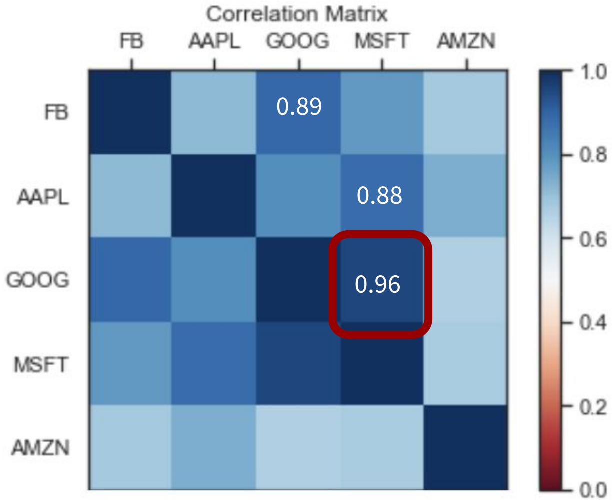
- **Spark**
  - Used to measure correlations between stocks with Machine learning library (MLlib)
  - Correlation is an important measure in stock market
    - Measure the amount of diversification among the stocks in a portfolio
    - Choosing assets with low correlation help to reduce the risk

- **Spark SQL**
  - Used to examine recommendation trends
  - Pulled top analysts recommendation data using“[ticker].recommendations”
  - Filter to count votes from analysts

Analysis & Insights

I want to sell one asset in my portfolio that is too closely related.

□ Meta
□ Apple
□ Google
□ Microsoft
□ Amazon

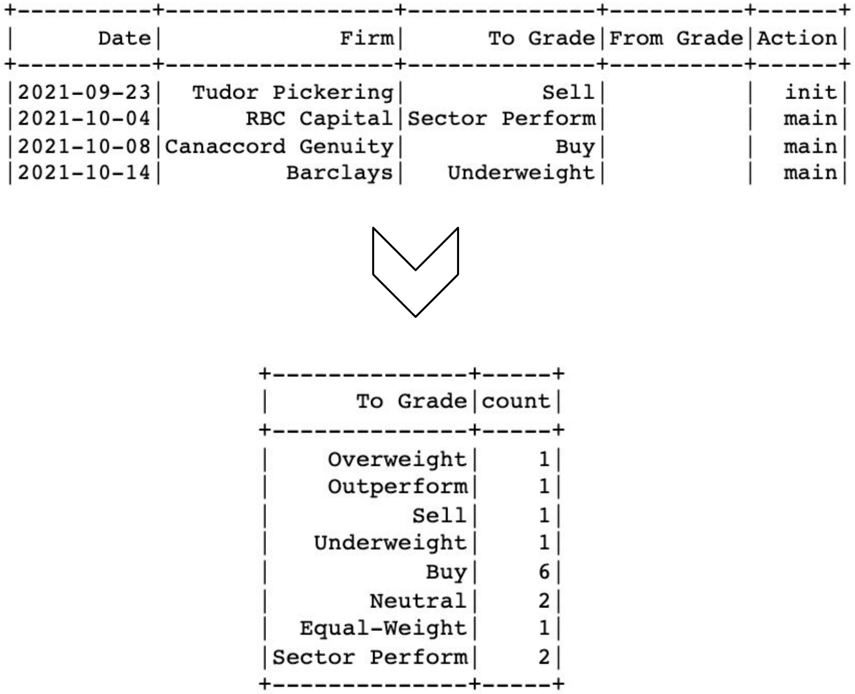


- Examined last 1 year adj close prices
- **Very high positive** correlation between **Microsoft** and **google**
- High positive correlation between Meta and Google & Apple and Microsoft

Analysis & Insights

Interested in buying. Any thoughts?

❑ Tesla

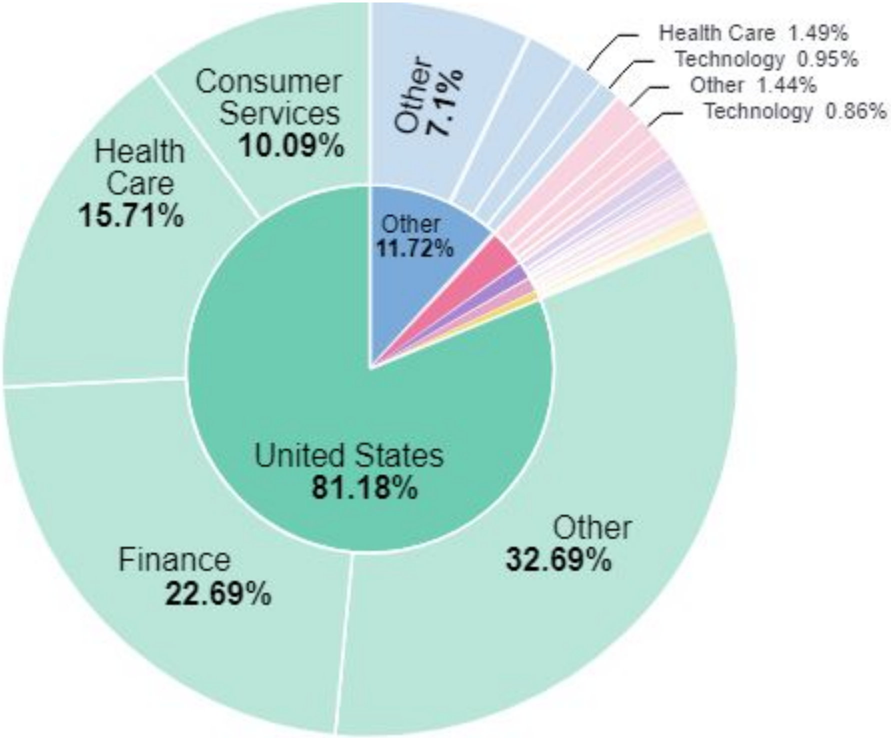


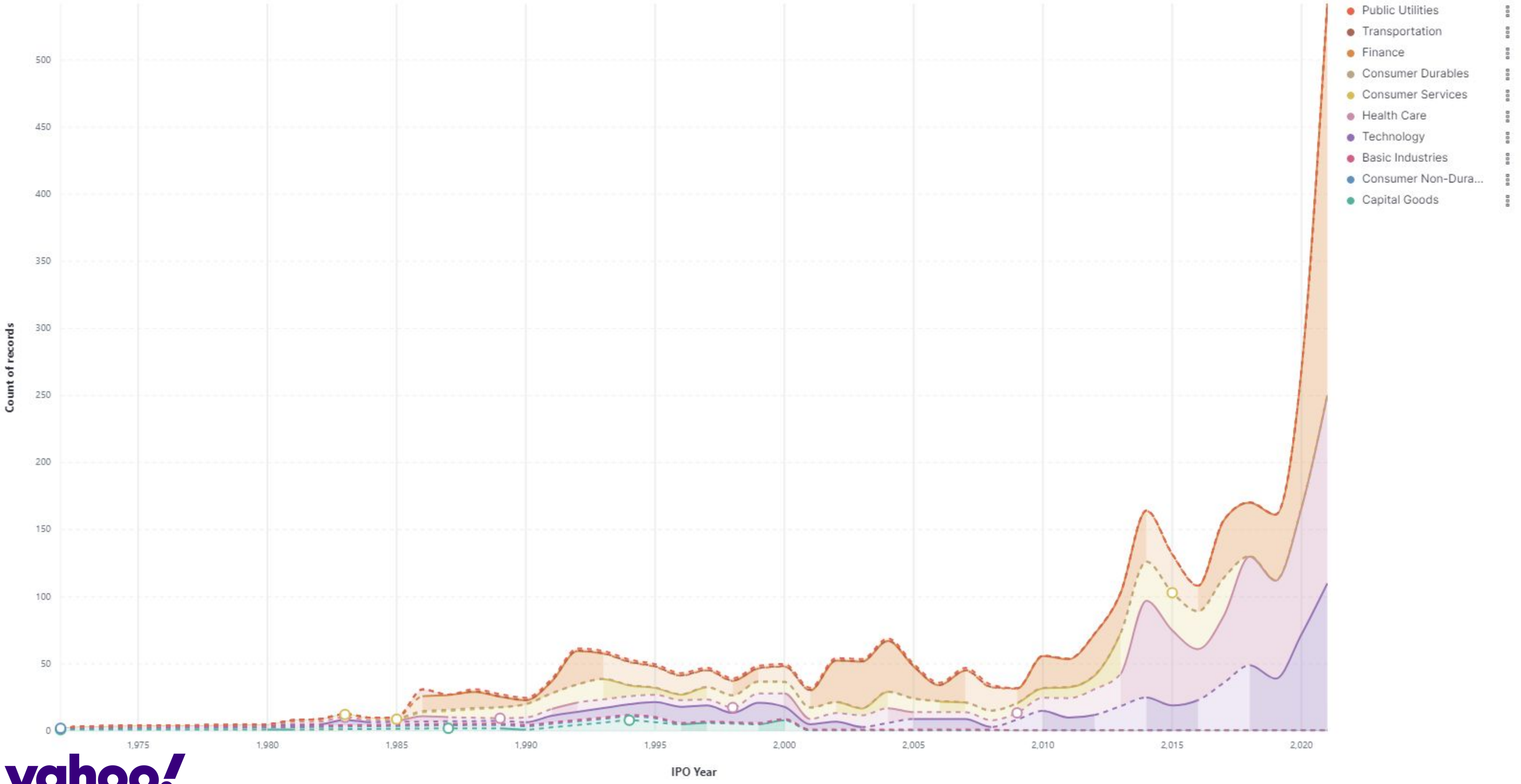
```
+----------+----------------+-------------+----------+------+
|      Date|            Firm|     To Grade|From Grade|Action|
+----------+----------------+-------------+----------+------+
|2021-09-23| Tudor Pickering|         Sell|          |  init|
|2021-10-04|     RBC Capital|Sector Perform|         |  main|
|2021-10-08|Canaccord Genuity|         Buy|          |  main|
|2021-10-14|        Barclays|  Underweight|          |  main|
```

```
+--------------+-----+
|      To Grade|count|
+--------------+-----+
|    Overweight|    1|
|    Outperform|    1|
|          Sell|    1|
|   Underweight|    1|
|           Buy|    6|
|       Neutral|    2|
|  Equal-Weight|    1|
|Sector Perform|    2|
+--------------+-----+
```

- Pulled last 3 months recommendations
- Most votes for “Buy” from top analysts

Analysis & Insights





Count of records
500
450
400
350
300
250
200
150
100
50
0
1,975
1,980
1,985
1,990
1,995
2,000
2,005
2,010
2,015
2,020
IPO Year
● Public Utilities
● Transportation
● Finance
● Consumer Durables
● Consumer Services
● Health Care
● Technology
● Basic Industries
● Consumer Non-Dura...
● Capital Goods

Lessons Learned

- The infrastructure is by far the hardest part of the setup
- There is a lot of applicability in the big data tools taught in class
  - The tools we used were very powerful tools for any data-enthusiast
- ETL pipelines → Powerful but maintenance increases with additional infrastructure added