DS8003 Project Proposal

Team: Adam Azoulay, Hee Kyoung Nam, Peter Yu

Problem definition:

We want to analyze data from the yahoo finance API (https://www.yahoofinanceapi.com/) to generate some insights into the data, but we don't want to be manually calling it to collect data, and we want it to be processed automatically so we can check the latest results any time we want them.

Solution:

The proposed pipeline would be set up as follows:
Apache Airflow (schedule api retrievals), HDFS (to store the files), Spark/Spark Streaming (to enrich/clean the data), elasticsearch/hive (to query the data on disk), kibana (to visualize the results).
This meets the requirements of having the data ready at all times (we can set the call frequency) and the final insights ready when we need them.

Dataset description:

We have many endpoints to collect data from, some interesting ones are the *chart* endpoint which allows us to retrieve historical data about symbols which leads to interesting analysis in terms of volume vs. day, comparison with similar symbols for performance, and comparison with the *options* endpoint data to check how the changing price affects the options available. The *trending* endpoint allows us to retrieve a list of trending stocks on a specific date and we can retrieve sector information (ex. technology, healthcare) of the trending companies from *insights* endpoint. Using the *trending* and *insights* endpoints of yahoo finance API, we can read sector trends and try to find anomalies (i.e. companies value falling while industry is trending up).