Database system design:

1. Part of designing the data includes collection of historic and real time reddit posts from subreddits such as “/r/wallstreetbets”. A raw format of the posts and comments will be stored on the MongoDB collections.
2. We are expecting to have a mix of useful and useless posts and comments under the subreddits we will be choosing for this project. The ETL phase should enable us to collect meaningful data to achieve better predictions.
3. NLP processing phase will include steps such as Feature Engineering, Entity Extraction, Sentiment Analysis which will produce data in the desired format, there by setting us up for creating our regression models.
4. The data generated after the NLP processing phase is again stored on newer MongoDB collections to assist various Visualization components.

Historical Reddit Posts

Real Time Reddit Posts from python server

ETL

(python/R scripts)

NLP Processing

(Entity Extraction, Sentiment Analysis)

MongoDB database

Web Server:

1. This webserver will be designed to handle both REST API and socket.IO based API which enables real time communication.
2. The REST API will handle requests which are associated to historical data such as,
   1. Reddit posts on $AMD over the last 6 months,
   2. Number of posts made about Micron ($MU) for the last 30 days,
   3. Sentiment score for AAPL over the last 60 days.
3. The Realtime API will perform the following tasks in sequence.
   1. Open a stream connection to a subreddit and monitor for new posts.
   2. When a new post is made on reddit, the server pushes that post to the MongoDB collection in the raw format.
   3. Perform, NLP processes such as Entity Extraction, Sentiment analysis and upload the information into the database.
   4. Push the data to the Web UI which is enabled to handle socket.io responses.

NLP Processing

MongoDB database

Python server

Realtime Reddit Streams

Web UI

Web UI:

1. The WEB UI will be enabled to showcase both
   1. “Ticker information” that is awfully important for any application which discuss the stock market and
   2. “Social sentiment” which we think will provide an innovative view on how we can visualize social media data related to stocks.
2. The UI will enable us to search a stock using the ticker information.
3. Based on the selected stock, the UI display
   1. Price history, sentiment score, volume of posts as a line graph
   2. Historical price as a table
   3. Fundamentals information
   4. Realtime Reddit posts/comments color coded based on the sentiment they represent
   5. Reddit post Statistics
4. A basic mockup of the UI is attached below. While we are not close it finalizing this yet, we still believe we have a good launch pad to create complete product.

