Real-time monitoring of volatile, retail investor-driven price fluctuations in the equity market

Purpose of the document

This report defines the scope of the service that provides. The investors use this platform to have a better understanding of the trend of the stock market and understand the public view on social media. This document will clarify the scope and services for future deployment purposes.

Inception Report

Project Name: Real-time monitoring of volatile, retail investor-driven price fluctuations in the equity market	
Date: Feb 19, 2023	Revision Number:1.0

Authors

This document was prepared by:

Name: Layne Pu Title: Team Leader, Software Developer Organization: Virginia Tech Email address: laynepu@vt.edu	Name: Huai-Che Chang Title: Project Manager, Software Developer Organization: Virginia Tech Email address: huaiche@vt.edu
Name: Wen-Fang Lu Title: Scrum master, Software Developer Organization: Virginia Tech Email address: wenfang@vt.edu	Name: Wan-Yi Mao Title: DevOps, Software Developer Organization: Virginia Tech Email address: wanyi@vt.edu

Date	Version	Document Revision History	Document Author
Feb 19, 2023	1.0	Initial version	ALL

I. Background

Spoofing is a technique involving placing a large number of fake orders in the market with the intention of tricking other traders into thinking there is more demand or supply than there actually is. This can cause prices to move in the desired direction, allowing market companies to profit from the manipulation.

II. Project Objective

The objective of this project is to provide an objective view of the stock market. When investors are making a decision if they should buy or sell a specific stock, it often relates to the incident that happened recently, the financial report revealed or views from professional financial analysts. To save some time for investors, we provide a platform that collects information about public views for the stock market and uses visualization tools to make it clear of the trend of a specific stock. In that way, the investors would save time in browsing related information about the stock and can be more objective when making a decision.

III. Project Scope

40% data engineering - data streaming of reddit forum
40% data science - sentiment analysis
20% web development - full stack development

IV. Project Deliverables

1. Show daily stock price with a graph for selected companies.
2. Filter posts with date, subreddit and keywords.
3. Do sentiment analysis on listed posts and show the percentage of bullish.

V. Project Assumptions and Constraints

Assumptions -
1. Price of meme stock is fully retail investor driven.
2. Meme stock is highly affected by the public opinion on social media (number of traffic, number of posts).

Constraints -
1. Twitter API is not free now.
2. Opinions in the same subreddit may be unanimous.

VI. Team Organization

A. Huai-Che Chang : Project manager, Software engineer
B. Wen-Fang Lu : Scrum master, Software engineer
C. Layne Pu : Team lead, Software engineer
D. Wan-Yi Mao : DevOps, Software engineer

VII. Project Management Approach

A project management methodology is a set of principles and practices that guide us in organizing our projects to ensure our optimum performance. For this capstone project, we would like to adopt the agile approach to manage our project.

Agile is an iterative approach to project management and software development that helps teams deliver value to their customers faster and with fewer headaches. We would like to keep track of our progress by continuously evaluating the result and responding to the problems we face quickly. We would like to use the Jira board to track our working progress.

A. Jira board

Jira Software is a mission-critical tool to plan, track, release, and report on work. We would run two-week sprints and create tickets for each member of the team. We would have stand-up meetings every week to learn the current progress of every team member that works on Scrum tasks

VIII. Methodology

A. Data Collections

PushShift is an API that provides access to historical Reddit data. First, we determine the scope of the data. For this project, we focus on subreddits that are related to the stock market, for example, r/Stocks, r/Investing, r/WallStreetBets etc. Second, we will construct an API with different requests and key words. After collecting the data, we will process it with the pre-trained model to do sentiment analysis.

B. Sentiment Analysis

Sentiment Analysis is a natural language processing (NLP) technique that involves identifying and extracting subjective information from text data, such as opinions, emotions and attitudes. The goal of sentiment analysis for this project is to have a general understanding of how public or social media opinions affect the stock market. In this process, we would train an existing pre-trained model on the large dataset we collect with labeled text data. Label each text as positive, negative or neutral. Also, we would clean and preprocess the data by removing special characters, punctuation, and stop words, and then tokenize the text. Then we can feed the data into the pre-trained model and fine-tune the hyperparameters so we can find the best results for the specific dataset.

C. Kafka

The main function of Apache Kafka is to act as a distributed messaging system for real-time data streaming. It allows producers to publish data to one or more topics, which are then consumed by one or more consumers. Producers and consumers can be distributed across multiple machines, making it possible to process and transport large volumes of data across a network.

Kafka uses a publish-subscribe model, where data is published to one or more topics and consumed by one or more consumers. Data is organized into topics, which can be divided into partitions for scalability and fault tolerance. Each partition is replicated across multiple machines to ensure high availability. Kafka also provides features such as message retention, which allows data to be stored for a specified period of time, and message replay, which allows consumers to replay messages from a specified point in time. These features make it possible to process and analyze historical data, as well as real-time data.

D. Elasticsearch

Elasticsearch is a widely-used open-source search and analytics engine that offers a range of features. With its full-text search capabilities, it enables our project to efficiently search and retrieve data. Additionally, as an OLAP database, Elasticsearch facilitates statistical analysis by quickly processing large volumes of data. These combined features make Elasticsearch a powerful tool for us to manage and analyze data effectively.

In Elasticsearch, a data structure called an “inverted index” is used to efficiently search for text-based data. It is created by analyzing the text in each document and creating an index that maps each unique word or term to the documents in which it appears. Instead of mapping from documents to terms, it maps from terms to documents. This means that when a search query is performed, Elasticsearch can quickly look up the terms in the index and retrieve the documents that contain those terms, rather than searching through the entire collection of documents. In the context of our project, a document refers to a Reddit post or comment, and the terms are the keywords that we search for. With that being said, Elasticsearch makes it possible to efficiently search through vast amounts of text data on Reddit and perform accurate statistical analysis.

E. Spark streaming

Apache Spark is a data processing framework that can quickly process operations on very large data sets. Live data streams can be processed in a scalable, high-throughput using Spark Streaming API. The data input can be ingested from a variety of sources, including Kafka, Flume, Twitter, etc. After processing, processed data can be pushed out to databases and live dash-boards.

F. Web application for Data visualization

Spring Boot offers a fast way to build applications. It looks at our classpath and at the beans we have configured, makes reasonable assumptions about what we are missing, and adds those items. With Spring Boot, we can focus more on business features and less on infrastructure. As a result, we would like to use Spring boot to build a web application to visualize the data we are interested in from Reddit.

IX. Risks Management

	Data quality	Database Scalability	Developmental cost
Risk Description	The collected data may not meet the required standards for accuracy and reliability	Given the scale and complexity of real-time data, there may be limitations to our database's capacity to store large amounts of this data	The development of our project requires a variety of tools and expertise, which may present time constraints and challenges
Possible impact	The final statistics displayed on our website do not accurately reflect the true reality of the situation	Our database has reached its capacity limit and is unable to store all of the data	There are unfinished components of our project, which is impacting our ability to deliver a satisfactory presentation
Corrective action	Filter out poor-quality data	Optimize the database by removing outdated or duplicate data, or prioritizing the most representative or essential data to be stored in the database	Throughout the development process, it is important to prioritize the most critical elements and focus on the most essential requirements

X. Detailed work plan

A. Major milestones

1. Project discussion
2. Data schema definition
3. Data streaming (produce, consume), false tolerance, caching
4. Web development
5. Deployment

B. Project Schedule

	Week 1-4	Week 5-6	Week 7-8	Week 9-10	Week 11-12	Week 13-14	Week 15-16
Project ideation	🟥						
Schema definition	🟧	🟧					
Data streaming		🟨	🟨	🟨			
Web development			🟩	🟩	🟩	🟩	
Deployment						🟦	🟦