SCALA PROJECT

Hedge Fund Application:
Real Time Risk Analysis

Team 11:
Amit Pingale
Mayank Gangrade

Goal

Building reactive application for portfolio management and risk analysis

Leveraging:

- Kafka + Spark streaming
- Spark Analysis engine
- MongoDB for maintaining historic records + batch processing

Why Real-time Big Data Pipeline Important

It is estimated that by 2020 approximately 1.7 megabytes of data will be created every second. This results in an increasing demand for real-time and streaming data analysis. For historical data analysis descriptive, prescriptive, and predictive analysis techniques are used. On the other hand, for real-time data analysis, streaming data analysis is the choice. The main benefit of real-time analysis is one can analyze and visualize the report on a real-time basis.

Technology and Tools:

1. Alpha Vantage API for Real Time Data
2. Kafka for capturing Real Time Data
3. Spark Streaming For Consuming Data
4. Spark + Scala for performing Analysis on the data
5. Spark Machine Learning Library
6. MongoDb*
7. Tableau
8. Jupyter Notebook - Scala kernel

* Dumping historic and predicted data on NoSQL database like MongoDB

Risk Analysis Process

1. Fetch real-time stock data from Alpha Vantage API
2. Perform cointegration test for pairs trading strategy
3. Develop ensemble of machine learning models to predict the momentum of the asset
4. Analyse stock and predicted value to make a decision
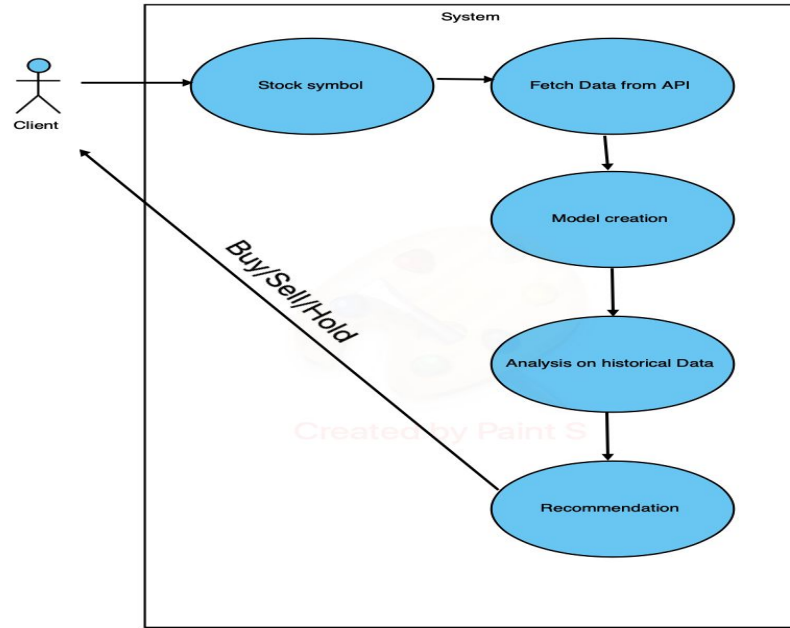5. Calibrate portfolio to minimise risk

Ensemble Models

1. Linear Regression
2. Decision Tree Regression
3. Random forest Regression
4. Gradient Boosting Regression

Data features: Stocks

1. <u>Open</u> (Independent feature)
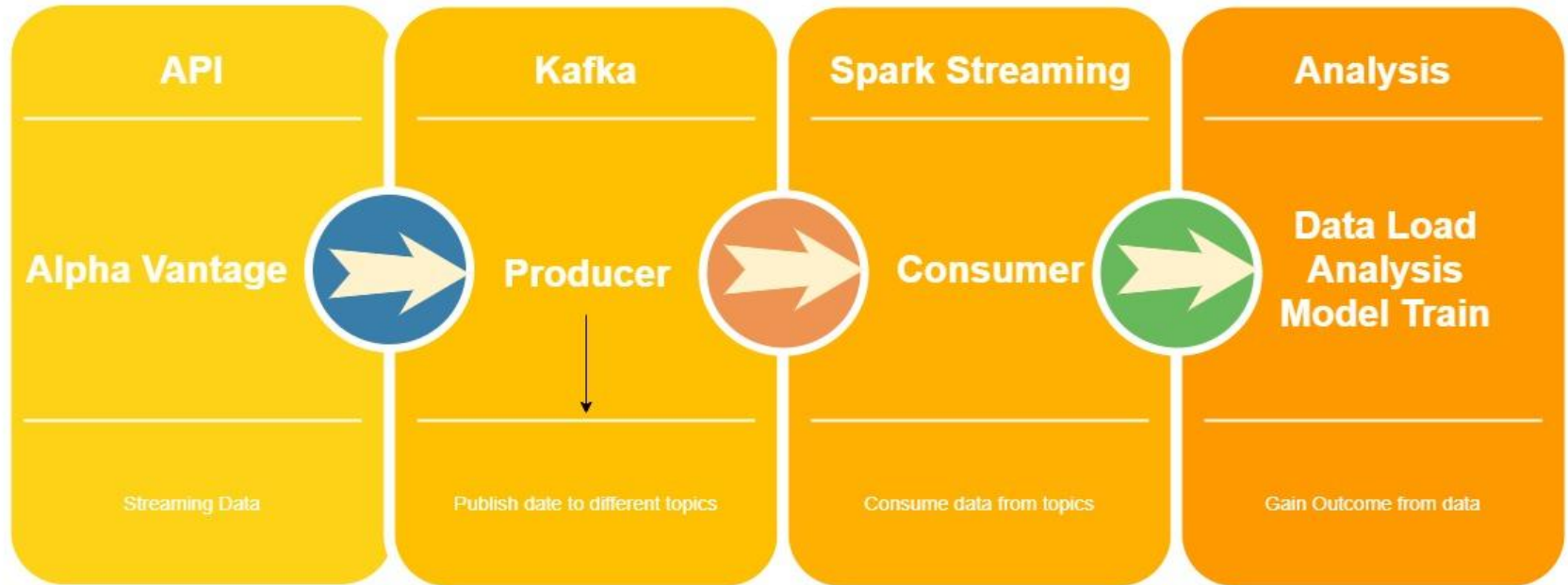2. High
3. Close
4. Adj Open
5. Volume
6. Low

Use Case Diagram

High Level Architecture

API	Kafka	Spark Streaming	Analysis
Alpha Vantage	Producer	Consumer	Data Load Analysis Model Train
Streaming Data	Publish date to different topics	Consume data from topics	Gain Outcome from data

Project Plan

Week 1: (23rd March - 29th March)

- Setting up new git repo
- Setting up whole ecosystem (Spark + Scala + Kafka + API Keys)

Week 2: (30th March - 5th April)

- Improve the code base
- Handle Exceptions and Errors

Project Plan

Week 3: (6th April - 12th April)

- Create test cases and verify all the functions
- Make the code more modular and remove repetition

Week 4: (13 April - 15th April)

- Check the robustness the code by running on longer time
- Visualize the data using a visualization tool

Project subdivision:

- Data Engineering
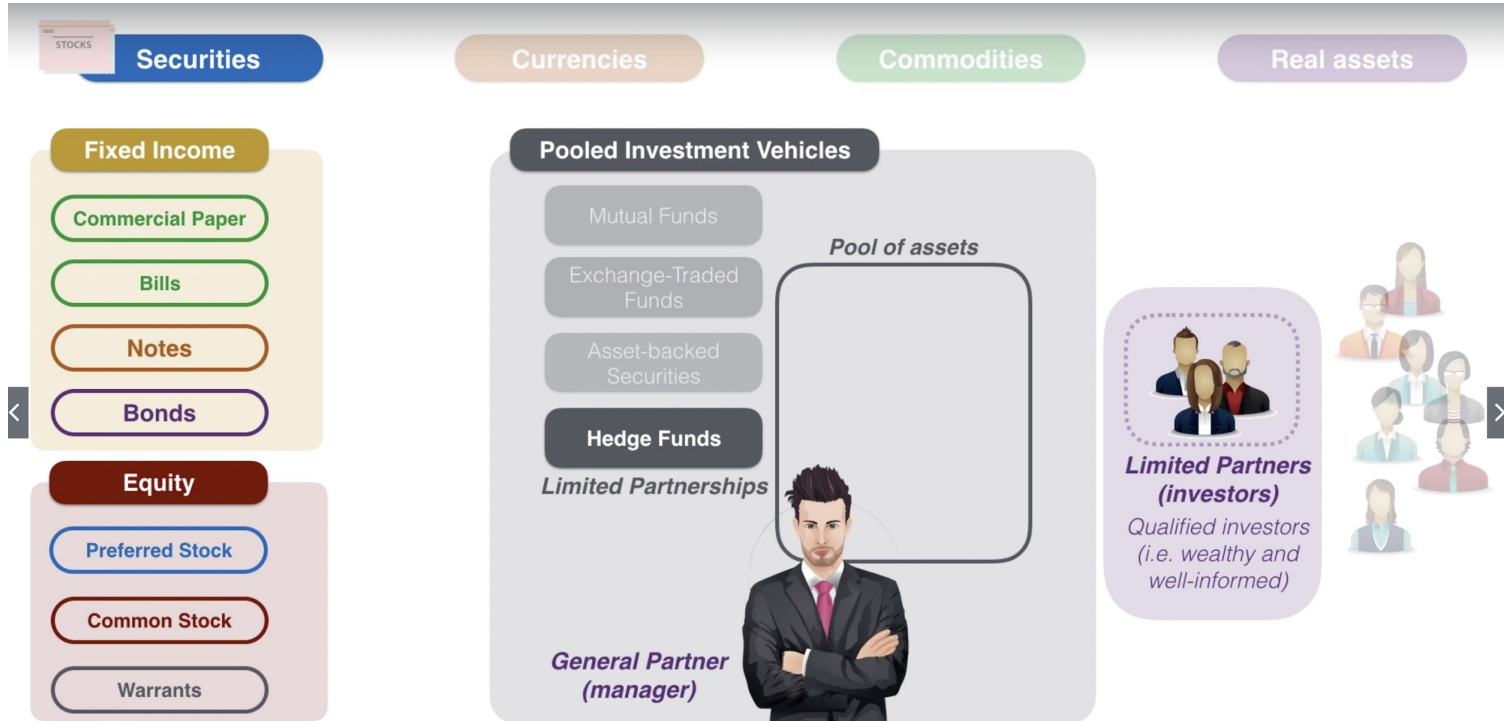- Machine Learning
- Data Analytics

Data Engineering

Data Engineering:
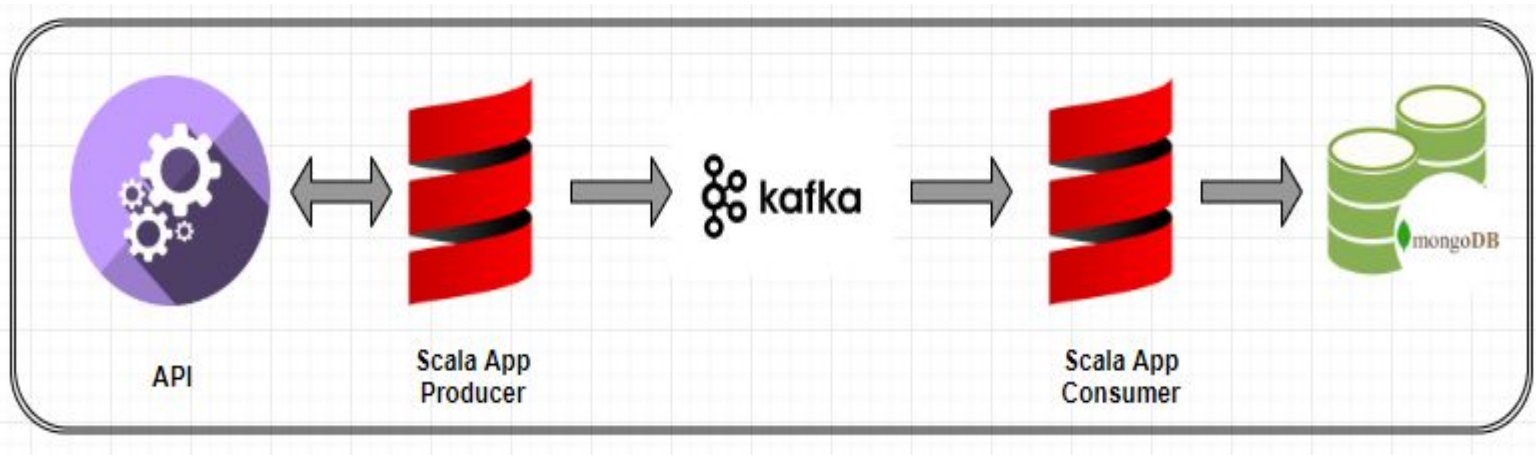
Machine Learning

What is Hedge Fund?

Assets, Contracts and Markets
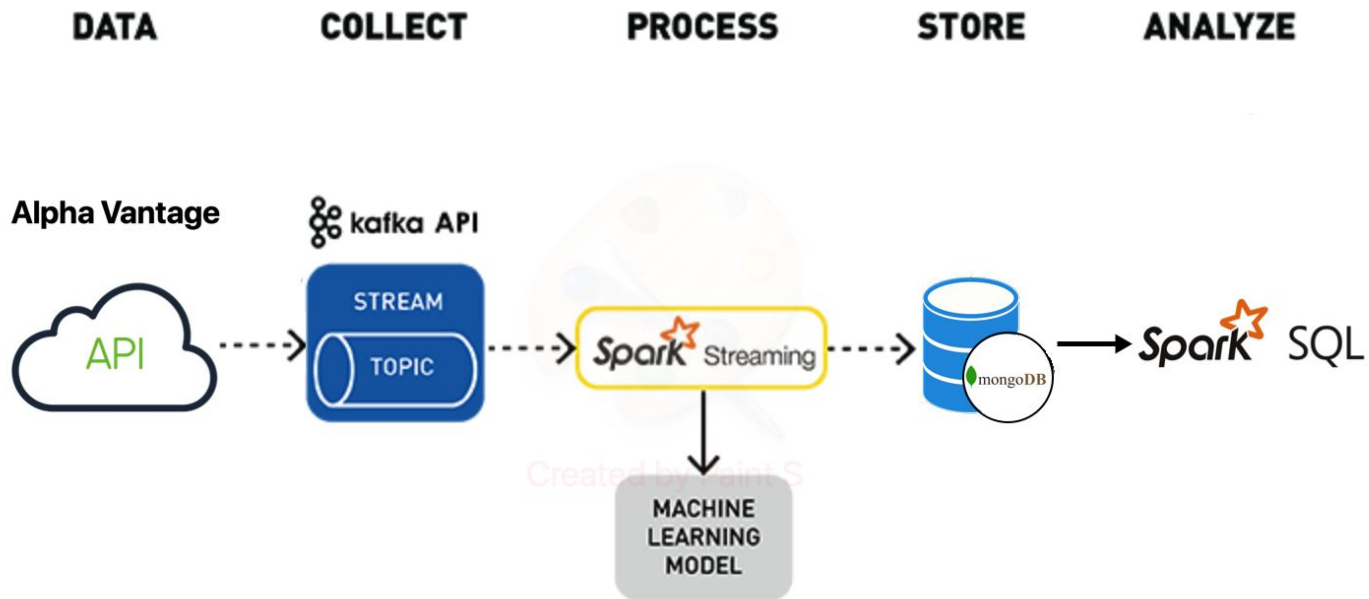
Risk Management in Portfolio Construction

Methods Used in Hedge Fund Company:

- Inverse correlation of the securities
- Protected Put (Long Position on a Asset + Long Put Position on the same Option Market Asset)
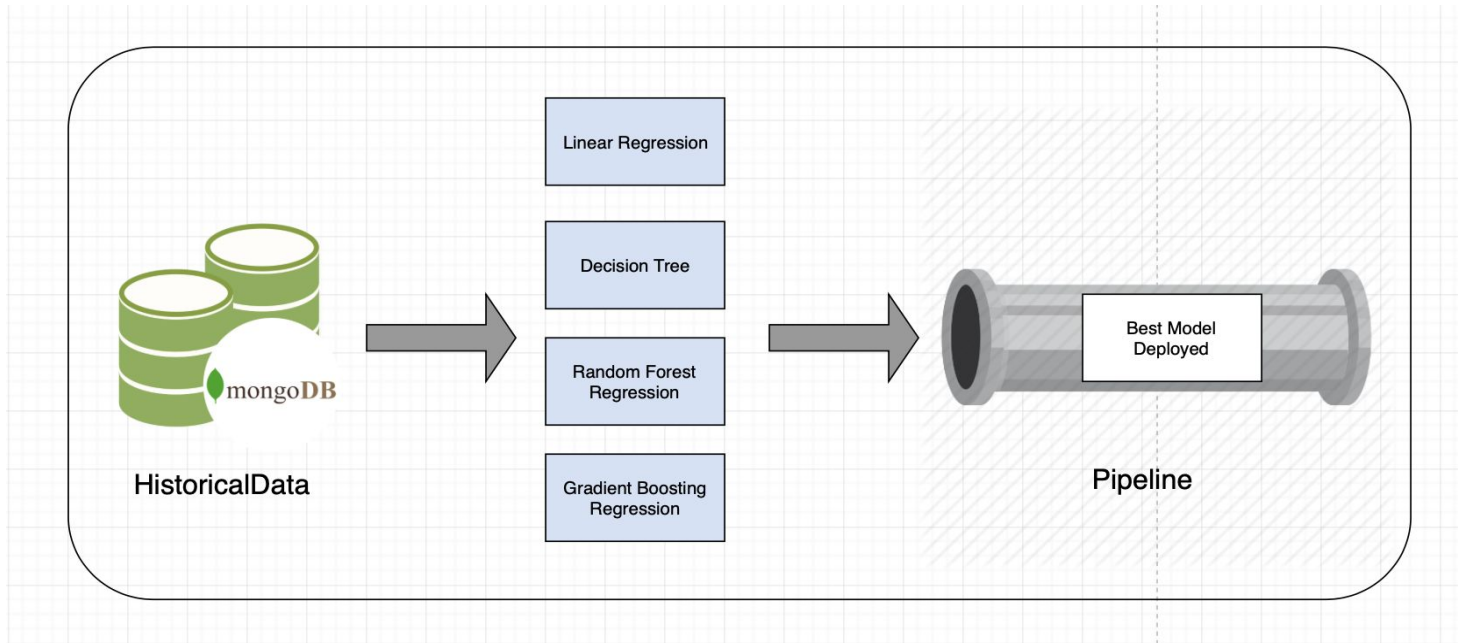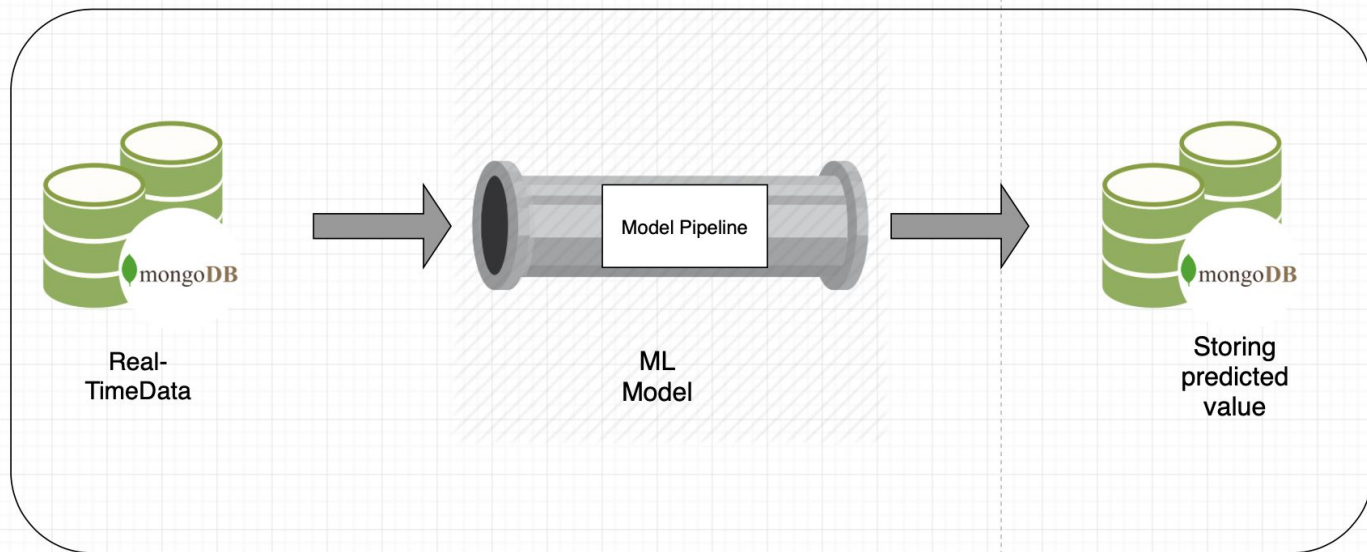- Fiduciary Call (Long Call on the Options Market + Risk Free Bond)

Architecture

DATA   COLLECT   PROCESS   STORE   ANALYZE

Alpha Vantage

kafka API

API

STREAM

TOPIC

Spark Streaming

mongoDB

Spark SQL

MACHINE LEARNING MODEL

ML Pipeline Construction

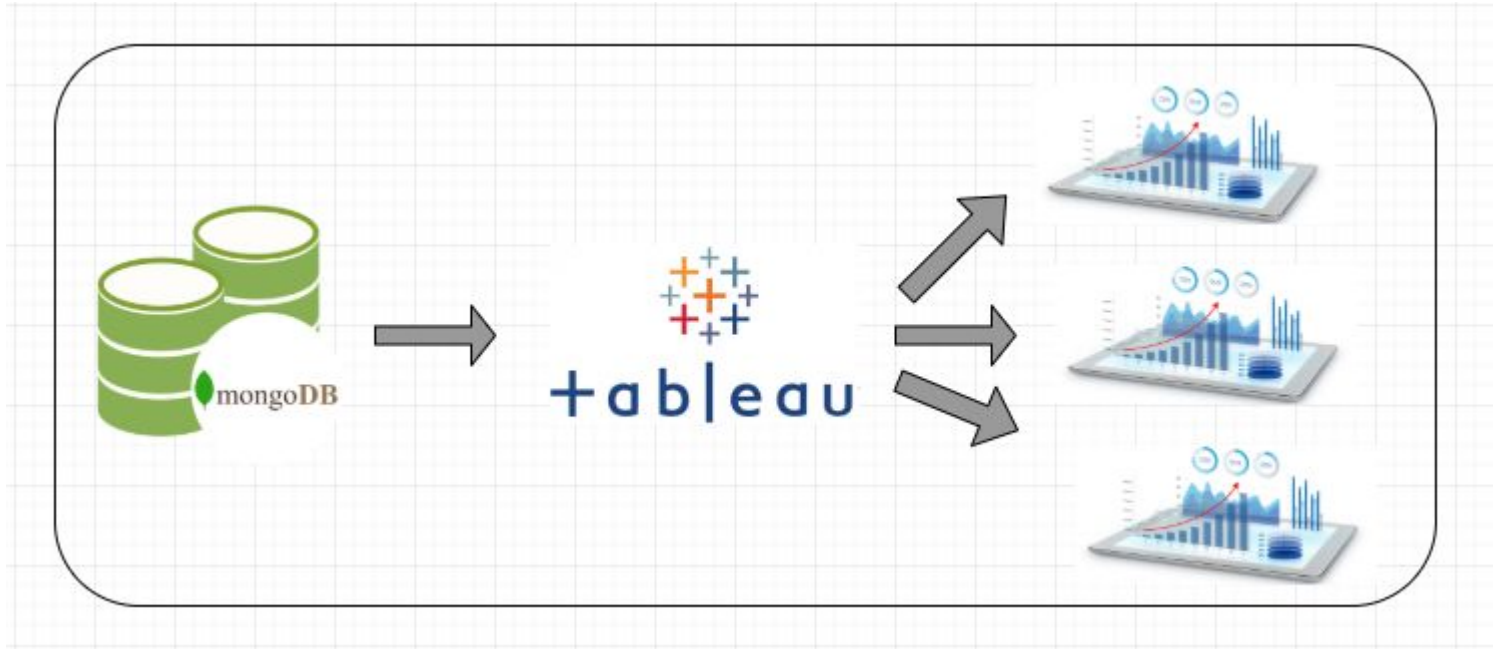# Real-Time Prediction

Trading Strategy

Assumption:

- The market value of a security is corrupted by noise
  - High Frequency Trading
  - Market Sentiment
  - False Information
- The model constructed is immune to these noises
- Conclusion:
  - Real Value > Predicted Value ---> Security is overvalued ---> Sell/Short Position
  - Real Value < Predicted Value ---> Security is undervalued ---> Buy/Long Position
  - Real Value == Predicted Value ---> Security is properly valued ---> Hold Position
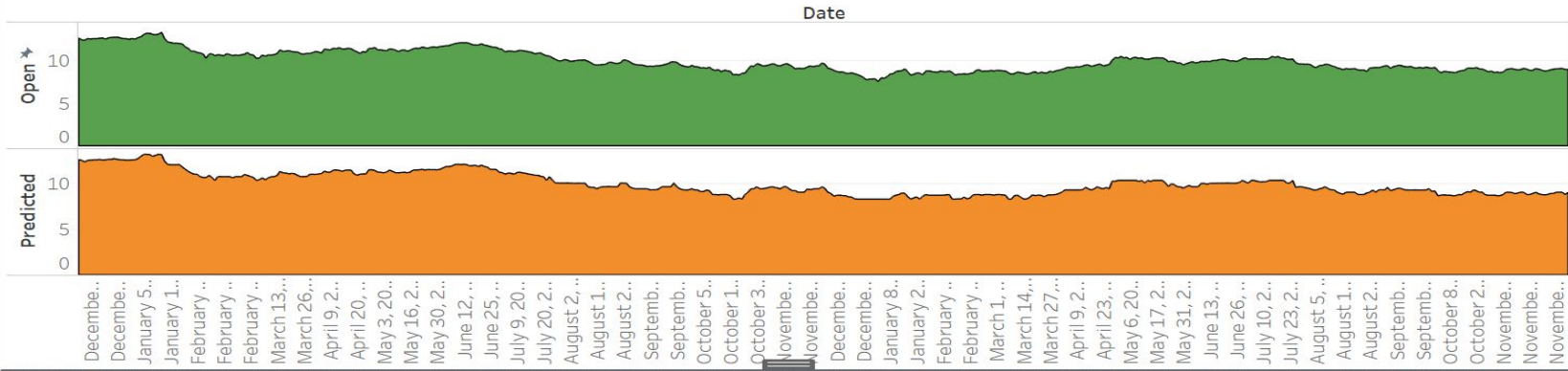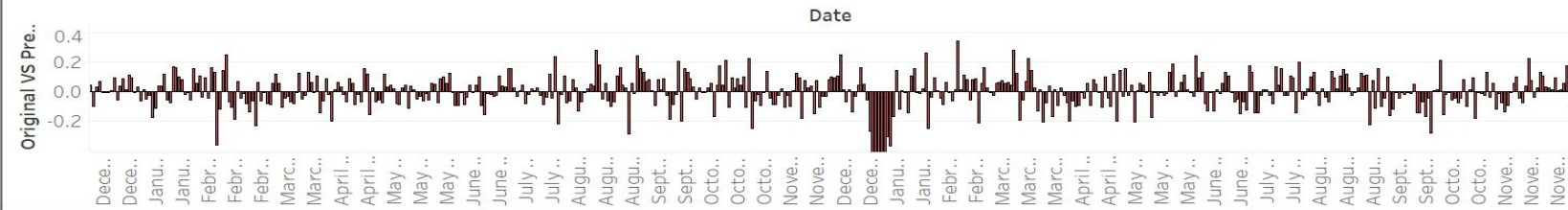
Data
Analysis

Data Analytics:

Data Analytics

1. Live data Streaming: Fetching Open stock price and Predicted stock price getting fed in MongoDB on real time basis
2. Connecting MongoDB with Tableau using BI Connector - Simba for Real Time data visualisation
3. In Tableau, presenting dynamic charts (changing with respect to Date-time) as per the data input
4. Showing Error (Original - Predicated stock price) , to guide user, if to sell or purchase the stock
5. Showing how change in Open and Predicted Open stock affects the Volume of the stock

Data Analytics (a)

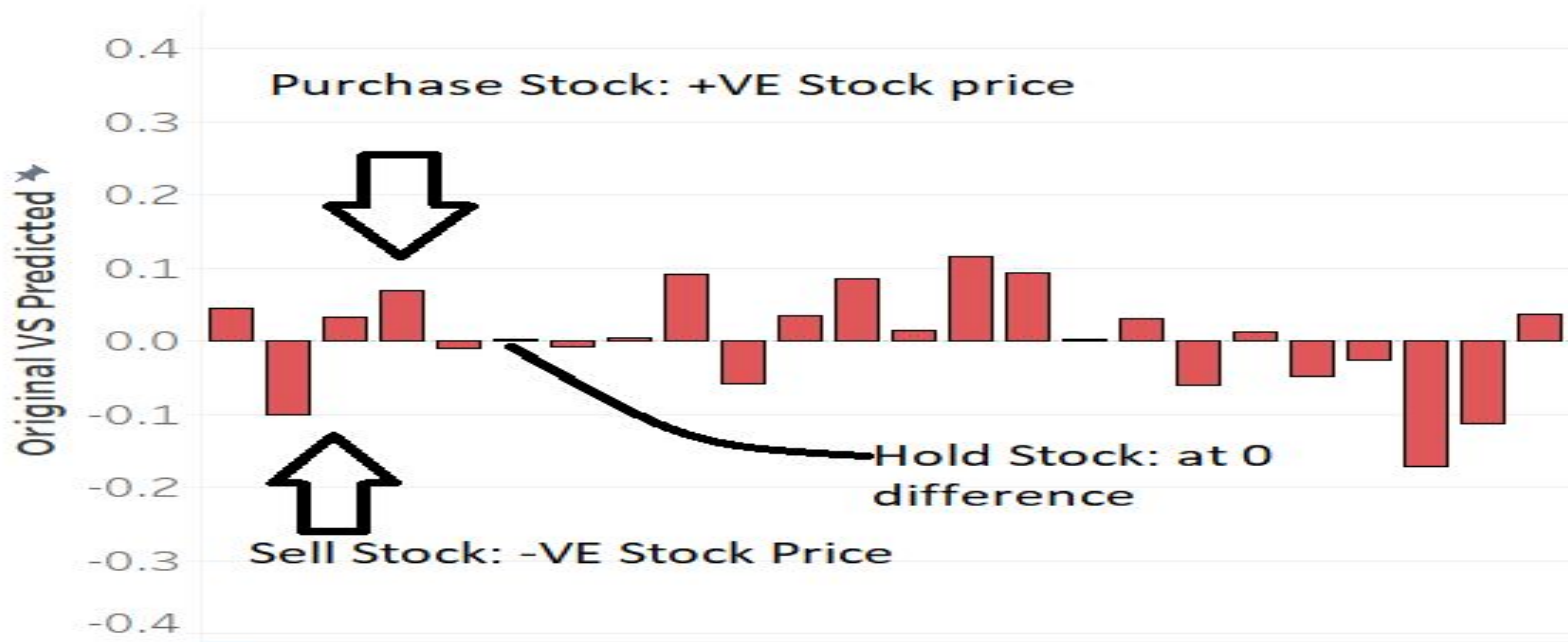## Open and Predicted Open prices of Stock

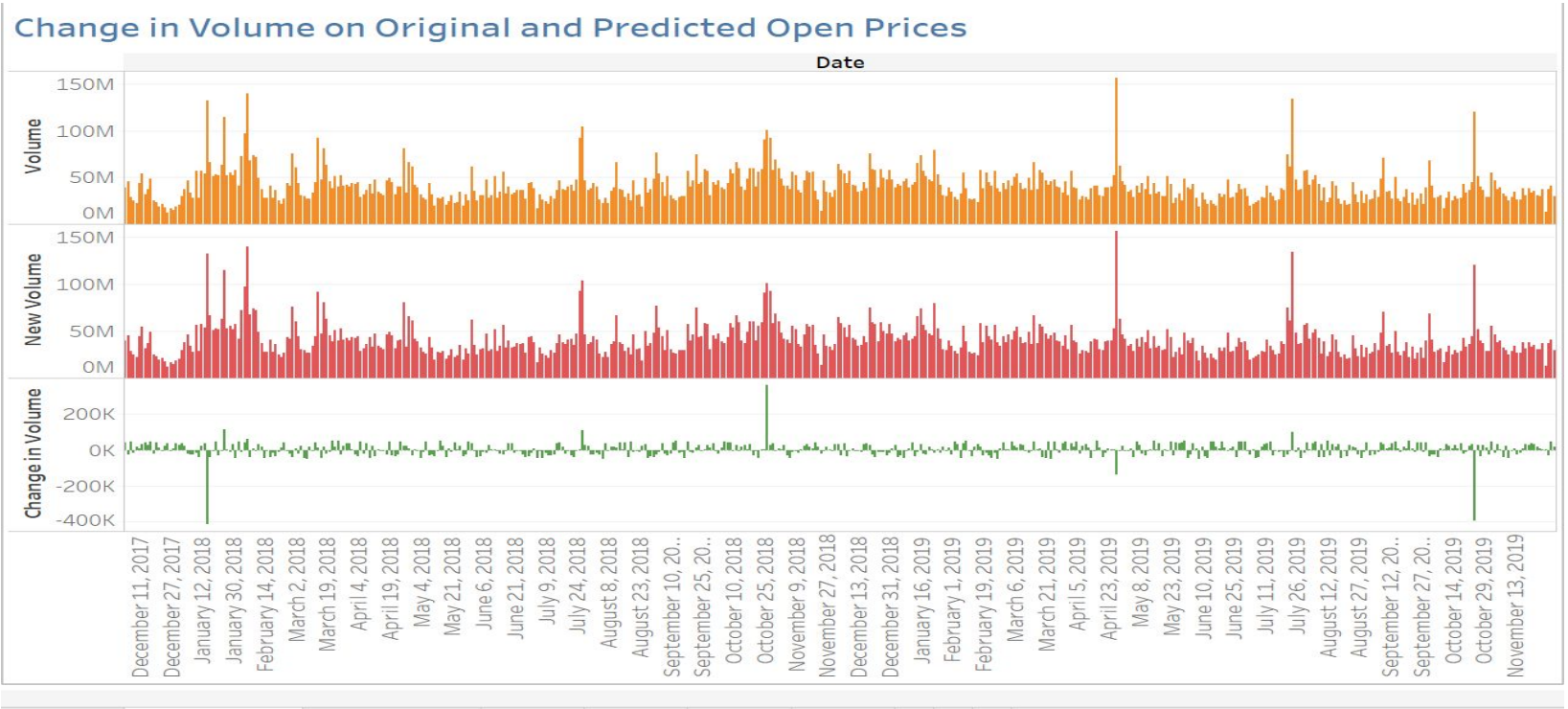## Difference between Open and Predicted Open prices of Stock

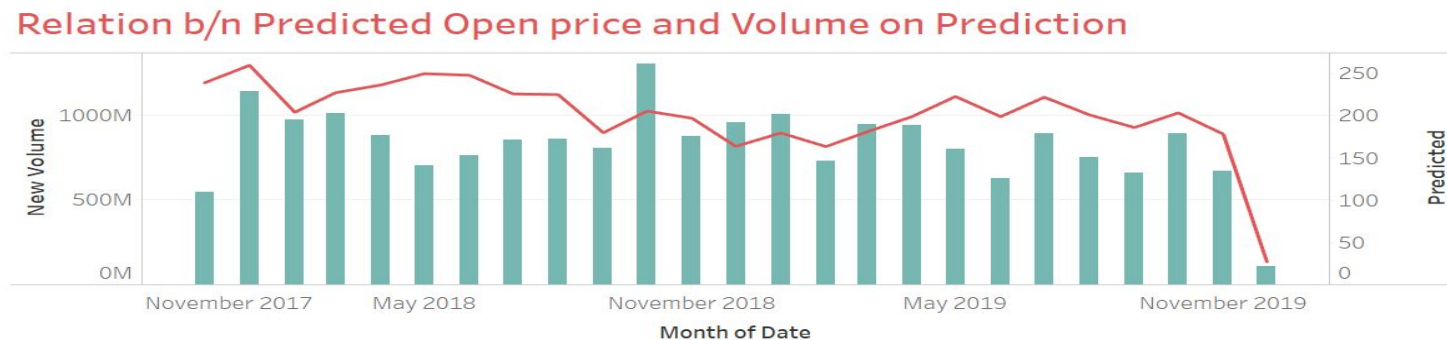# Data Analytics: When to sell, Purchase or Hold a stock

Data Analytics (b)

Change in Volume on Original and Predicted Open Prices
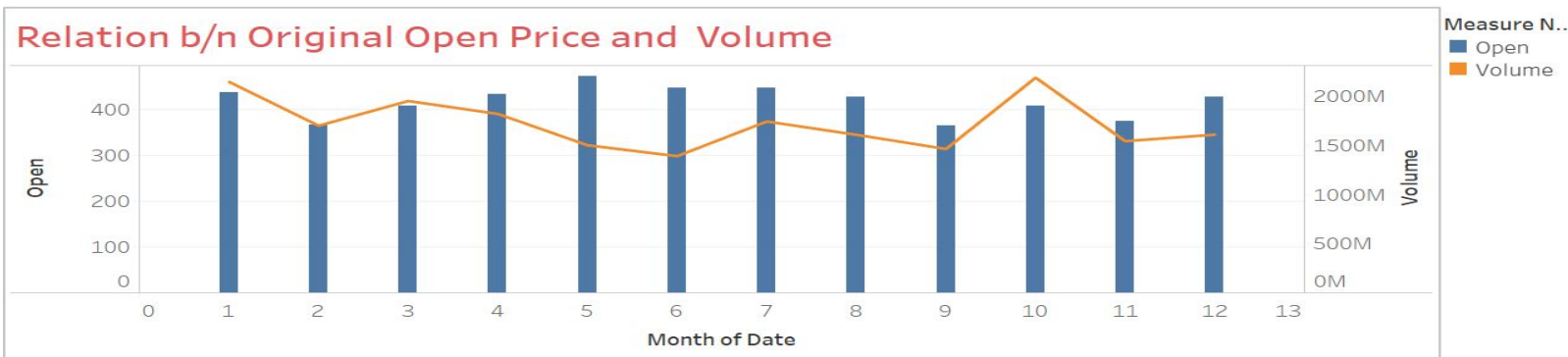
Data Analytics (c)

Acceptance criteria

- Data from API should be fetched in every **5 min** and published with in **5 Seconds** on Kafka Topics ✅

- Consume the data from Kafka topics with in **5 seconds** as it arrived, validate each data and load correct data into database ✅

- Selecting best model depending upon RMSE value (**RMSE < 0.7**) ✅

- Update data in real time interactive dashboards with maximum lag of **2 mins**

Thank You