CS-GY 6513 - Big Data Final Project

Under the Supervision of Prof. Juan Rodriguez

Project Title: Job Market Analysis System (Insightsise)

By

Name	Net ID
Apoorva Ganapati	ag8159
Manan Chawda	mrc9419
Rohan Sardana	rs7445
Srijan Malhotra	sm9439

NEW YORK UNIVERSITY

Fall 2022

Table of Contents

Introduction

Motivation

There are several reasons why it might be useful to create a job market analysis system:

1. To help job seekers identify in-demand fields and find job opportunities that match their skills and interests.
2. To help employers identify potential candidates and understand the local job market.
3. To help policymakers and government agencies understand the state of the job market and identify areas for economic development.
4. To help job seekers and employers negotiate salaries and benefits, as understanding the job market can give both parties a better sense of what is fair and competitive.
5. To help job seekers and employers make informed decisions about career development and workforce planning.

Overall, a job market analysis system can provide valuable insights and support decision-making in the job market.

Problem Statement

We aim to solve the problem of job market analysis by developing a system that will leverage real-time data from multiple sources for streaming, modeling, and visualization that will be used by candidates to make decisions on the job application or switch.

Market Analysis - Big Data Problem!

Any company's performance over recent years can be assessed to gauge decisions regarding future investments one could make in a company and the company's employee retention and attrition rates.

We have large-scale data available for the companies:

1. **Financial Data:** By analyzing large volumes of financial data, such as stock price, revenue, expenses, and profits, organizations can get a sense of how well a company is performing financially. This can be used to identify trends and patterns in financial performance and to inform business decisions.
2. **Social Media Data:** A positive sentiment on social media can be correlated with the good performance of any company. Especially, in the influencer age, where few tweets by profiles with credentials could influence sales of any product, companies thrive on social

media index and Net Promoter Rate to increase revenue.

3. **Employee Reviews Data:** Based on the data of employee reviews on portals like Glassdoor, Blind, etc. We could derive meaningful insights regarding the workplace environment, internal product successes and failures, and company-wide operational sentiments to assess if a company could be joined at the given moment.

As we collect the 3 types of data from multiple sources, we could leverage big data technologies to stream, store and analyze market sentiments and financials about a company. We talk about the data sources in the following section.

Datasource for Job Market Analysis

Twitter

A Twitter dataset is a collection of data from the social media platform Twitter. Some types of content you might find in a Twitter dataset include:

1. Tweets: A tweet is a message or update that can be up to 280 characters long, and is typically posted by an individual or organization on Twitter. A Twitter dataset may include the text, images, and other media included in tweets, as well as metadata such as the time and date the tweet was posted, the number of likes and replies it received, and any hashtags or links included in the tweet.
2. User profiles: A Twitter dataset may include information about individual users, such as their username, profile picture, location, and description.
3. Hashtags and mentions: Twitter users often use hashtags to categorize and make their tweets more discoverable, and may also mention other users in their tweets. A Twitter dataset may include data on the hashtags and mentions used in tweets.
4. Retweets and replies: A Twitter dataset may include information on the tweets that users have shared or replied to.

Using **Tweepy** for Streaming data, we would ingest real-time tweets for the companies, that would be stored for building insights and visualizations.

Yahoo Finance

Yahoo Finance is a financial news and data website that provides a wide range of information and tools related to the stock market and personal finance. Some of the content that we found relevant:

1. Stock quotes and market data: Yahoo Finance provides real-time stock quotes, as

well as historical price data, market summaries, and other financial metrics.

2. News articles and analysis: Yahoo Finance publishes news articles and analyses on a wide range of financial topics, including market trends, company news, and economic indicators.
3. Tools and resources: Yahoo Finance offers a variety of tools and resources for investors and personal finance enthusiasts, including portfolio tracking, stock screener, and retirement planning tools.
4. Market commentary and opinion: Yahoo Finance features commentary and analysis from financial experts and industry insiders on current events and trends in the financial markets.

We are interested in using real-time and historical stock price data that can be time-stamped in seconds, minutes, weeks, months, and years. Our real-time data would be fetched in intervals of 2s and historical data per minute can be retrieved for the last 60 days.

Blind Reviews

Blind is a professional networking app that allows users to connect with each other and share anonymous feedback and reviews about work experience and the companies they work for. The app as the name suggests collected 'Blind' data which means the identities of the users who post data, is masked.

Why use Blind data?

Blind data does not contain spam - Posts are created anonymously by users who are verified with their company email addresses.

Exploratory Data Analysis and Transformation

Review Sentiment	Number of Reviews
Positive	32966
Neutral	3412
Negative	4032

Fig 1. Sentiment Count of Blind reviews

Fig 2. WordCloud for all the reviews from Blind

Open	High	Low	Close	Volume	Dividends	Stock Splits

Fig 3. Real-time stock price from Tweepy

The following transformations were performed on the datasets:

Twitter Data:

1. Timestamps from Tweepy follow different DateTime formats and timezones. We had to convert them to follow the EST format to maintain consistency
2. Stock prices from Tweepy had *"Open, High, Low, Close, Volume, Dividends, Stock Splits"*. We only used the closing prices and dropped all other prices as they were not relevant to us

Yahoo Finance Data:

1. Timestamps from yahoo finance API follow different DateTime formats and timezones
2. We had to convert them to follow the EST format to maintain consistency

Blind Data:

1. Concatenated columns *"Description, Pros, Cons"* into "Review"
2. Dropped reviews with null value so that it is not accounted for during sentiment analysis and WordCloud generation
3. Lowercasing, removal of stopwords in the review text for sentiment analysis and WordCloud generation

Sentiment Analysis:

For sentiment analysis, we have used nltk's sentiment analysis. Each review word is represented by a tuple (sentence). The sentence is tokenized, so it is represented by a list of strings. Then the polarity score of the sentence is calculated using SentimentIntensityAnalyzer and the compound score is used. The Compound Score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive). We then used the compound score to categorize the review as positive or negative or neutral as follows:

Negative	if	the compound score is less than -0.25
Neutral	if	the compound score is between -0.25 and 0.25
Positive	if	the compound score is greater than 0.25

Architecture

Big Data Technologies

Kafka:

Apache Kafka is an open-source distributed event streaming platform used for building real-time data pipelines and streaming applications. It is designed to handle high volumes of data with low latency and provides a publish-subscribe model for messaging.

Kibana:

Kibana is an open-source data visualization and exploration platform that is used to analyze and visualize data stored in Elasticsearch, a search and analytics engine. Kibana is part of the Elastic Stack, a set of tools for data ingestion, enrichment, storage, and analysis.

PySpark:

Apache PySpark is an open-source Python library for big data processing and analysis, built on top of the Apache Spark distributed computing platform. PySpark is designed to make it easy to work with large datasets and perform complex data operations using a simple API.

ElasticSearch:

Elasticsearch is an open-source search and analytics engine based on the Lucene library. It is designed to provide fast and flexible search and analysis capabilities for a wide range

of applications, including full-text search, structured search, and geospatial search.

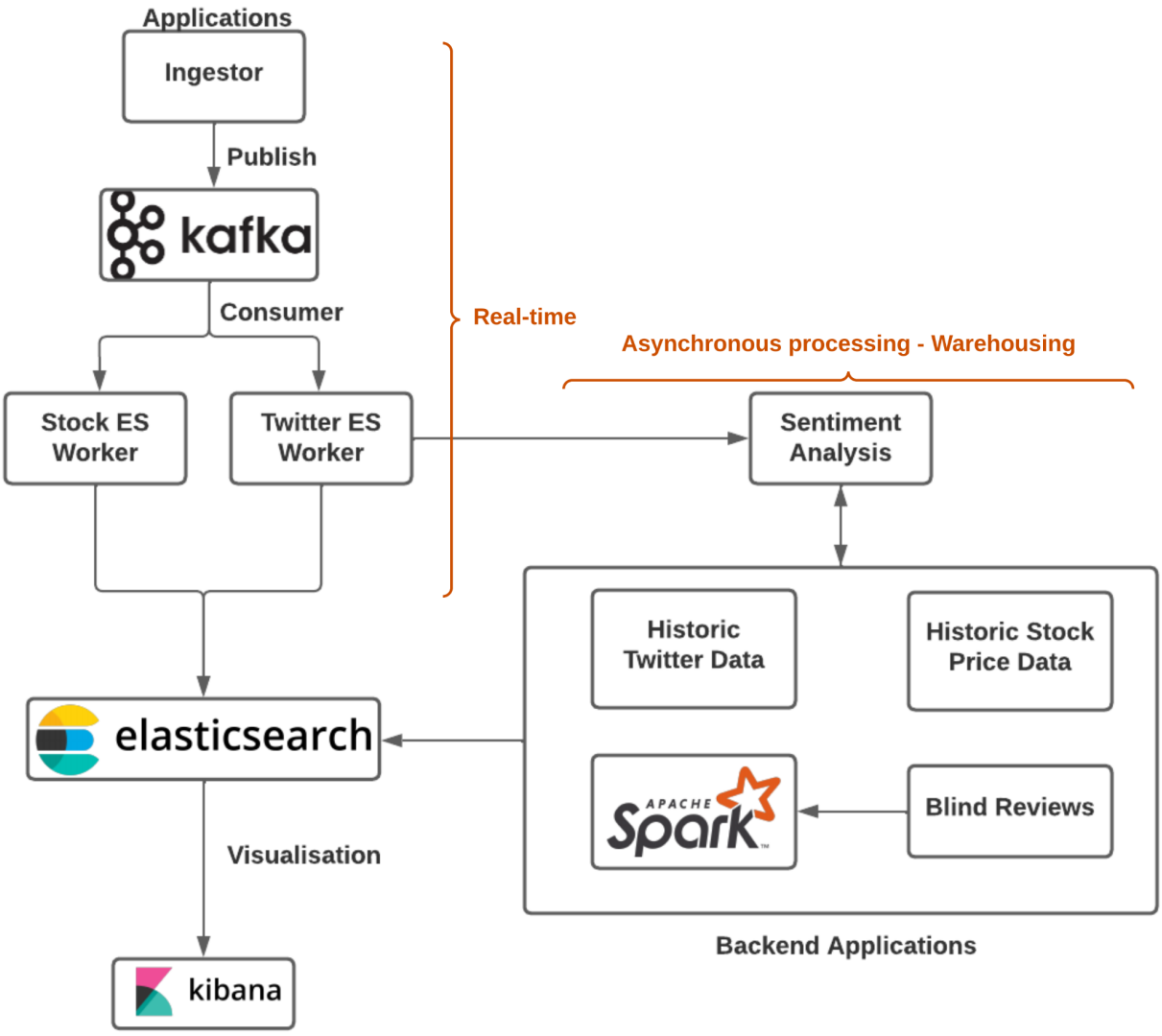

Fig 4. Architecture Diagram

Real Time Streaming:

1. **Ingestor**
   a. We have built two ingestors in a multithreaded application(ingestor.py) that pulls up real-time tweets for 20 companies using Tweepy Streaming and real-time stock price data in seconds using Yfinance.
   b. These ingestors would send data into respective Kafka topics
2. **Kafka**
   a. The kafka would be loaded with two data streams, available for consumers of the

two topics.

3. **Kafka consumer**
    a. We have used two consumers(twitter and stock price) that would do the required cleaning and preprocessing to store the data in ElasticSearch for respective indices.

Asynchronous Processing:

1. **Historic Tweets**
    a. Using a one-time load for historic tweets data, we would populate the ElasticSearch for 7-10 days of tweets and their respective sentiments.
2. **Historic Stock Price**
    a. The stock price data in a time-series manner, based on the timestamp would be loaded for analysis.
3. **Blind Review**
    a. Company specific reviews, with the sentiments of the reviews would be loaded into ElasticSearch for retrieval using the company name.

Idempotency Checks: Ensuring Consistency across our Real-time and Asynchronous system

Maintaining consistency across a real-time system and a warehousing system can be a challenging task, as the two systems may have different requirements and constraints in terms of data processing, storage, and access.

We have followed a **Common Data Model** approach:

In the context of a big data system that uses real-time and asynchronous processing, we have used a common data model at **ElasticSearch** because two systems(real-time vs asynchronous) may have different requirements and constraints in terms of data processing, storage, and access.

To avoid collisions and introduce **idempotency check** in systems, we have used **timestamps** of the **tweets** and **stock** data to ensure that the data is organized and consistent in eventuality.

This ensures that the Elastic Search queries and Visualizations on Kibana are correct over the course of asynchronous and real-time updates. When building a time-series analytics system, the timestamps play a vital role in providing graphical representation and visualization. Moreover, timestamp can be used as a key to filter data for users to identify patterns according to period of time, ex: negative tweets in last X days, stock price over last Y months. Etc.

Visualisation

Following is the screenshot of kibana dashboard after querying for **Airbnb and Apple:**

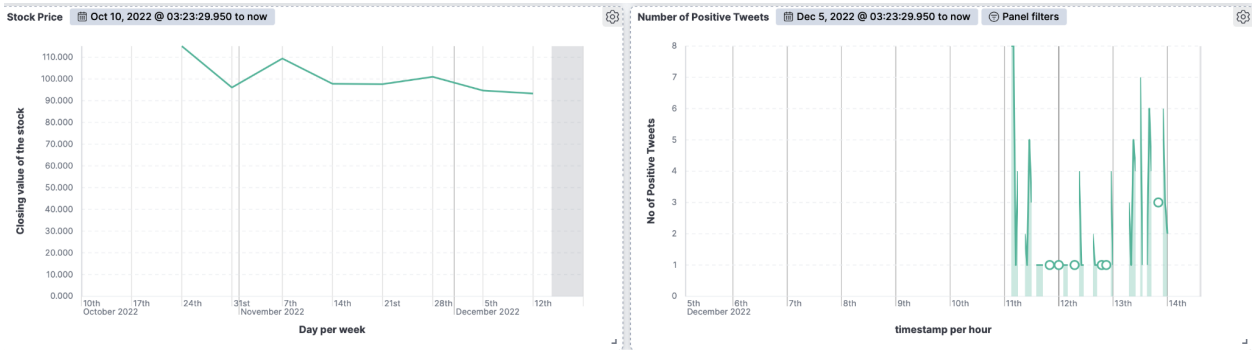

Fig 5. Kibana Dashboard for Airbnb

Fig 6. Blind Reviews Wordcloud for Airbnb

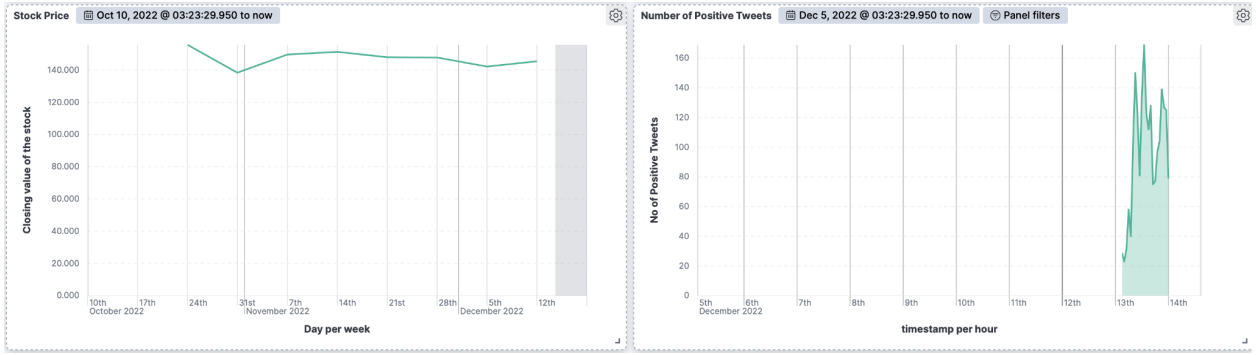

Fig 7. Kibana Dashboard for Apple

Fig 8. Blind Reviews Wordcloud for Apple

Future Scope

1. Dashboards as a Service: Using the big data system, we can provide dashboards as a chrome extension on job portals like Workday, Indeed, Handshake, so that the users can assess if the company would be a right choice for them to join.
2. Using Premium Open API's: Glassdoor, LinkedIn and Twitter provide large-scale and quicker access to their data, that can be ingested to generate more employee reviews and market insights
3. Using Ratings like S&P - Credit ratings for debts and soundness of a company. We can analyze companies' performance data using diverse and verified financial indicators.
4. Private/non-listed companies data ingestion using news and funding information for performance tracking.

References

1. https://www.tweepy.org
2. https://pypi.org/project/yfinance/
3. https://www.elastic.co, https://www.elastic.co/kibana/
4. https://www.nltk.org/_modules/nltk/sentiment/vader.html
5. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.feature.StopWordsRemover.html
6. https://www.confluent.io/kafka-summit-sf18/real-time-market-data-analytics-using-kafka-streams/
7. https://github.com/HarshCasper/Blind-App-Reviews