San José State University

CMPE 266 — Big Data Engineering and Analytics

Project Milestone II

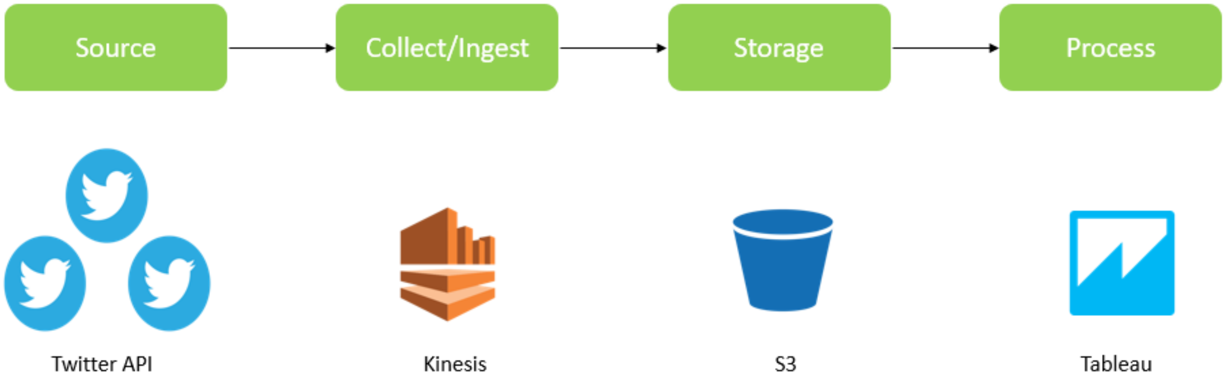
Created By:

Aakash Alurkar	013716729
Aniket Deshpande	013532519
Madhu Prasanna Kalluraya	013708071
Shivani Mangal	012530362
Zainab Khan	010120812

STOCK PRICE PERFORMANCE PERCEPTION USING TWITTER STREAMS

1.0 Project Idea

As per the random walk theory, “changes in stock prices have the same distribution and are independent of each other, therefore, the past movement or trend of a stock price or market cannot be used to predict its future movement”. However, it is also known that human emotions deeply affect decision making and an individual’s behaviour. This forms one of the prime concepts of behavioural economics. Thus, collective mood swings of humans are bound to affect stock market. This is because our emotions govern how we feel; and how we feel, affects how we behave as an investor. Our emotions help in determining if our decisions are beneficial or harmful. The raw indicators of swinging emotions in today’s world are twitter feeds, blog posts and Reddit feed. Through news articles, one can analyze the emotion of the masses. So can twitter feed be used to support in making decisions for the stock market? Through this project we aim to analyze the same.



1.1 Motivation

In one of his speeches, Abraham Lincoln once said,”In this age, in this country, public sentiment is everything. With it, nothing can fail; against it, nothing can succeed. Whoever molds public sentiment goes deeper than he who enacts statutes, or pronounces judicial decisions.” Social Media, in the form of twitter and facebook feeds is holding this true in the 21st Century as well. A skilled NLP analyst can mine patterns out of tweets to make conclusions about how the market in general feels about a stock, a company or a brand. John Bollen, in one of his research papers, claimed that Twitter data could predict the Dow Jones Industrial Average with 87.6% accuracy (Bollen, 2010). As per an article by MIT, “an analysis of almost 10 million tweets from 2008 shows how they can be used to predict stock market movements up to 6 days in advance.”

In recent times we have seen tweets that turned the market completely upside down, thus cementing the above mentioned theories. When Elon Musk tweeted about taking Tesla private and it’s funding being secured, Tesla shares ended at 11%.

1.2 Objective

Our Big Data application will analyze incoming twitter feed to gauge the market sentiment for particular brand. It will also capture the live stock market feed to check the correlation between market prices and sentiment. The user will be provided with both the information in form of a visualization dashboard to make an informed decision when investing in the market.

2.0 Technologies Used

2.1 Data Source

We take inputs from the following sources.

2.1.1 Twitter API

To ingest the real-time twitter data Twitter Developers API provides the following parameters.

Standard streaming API request parameters

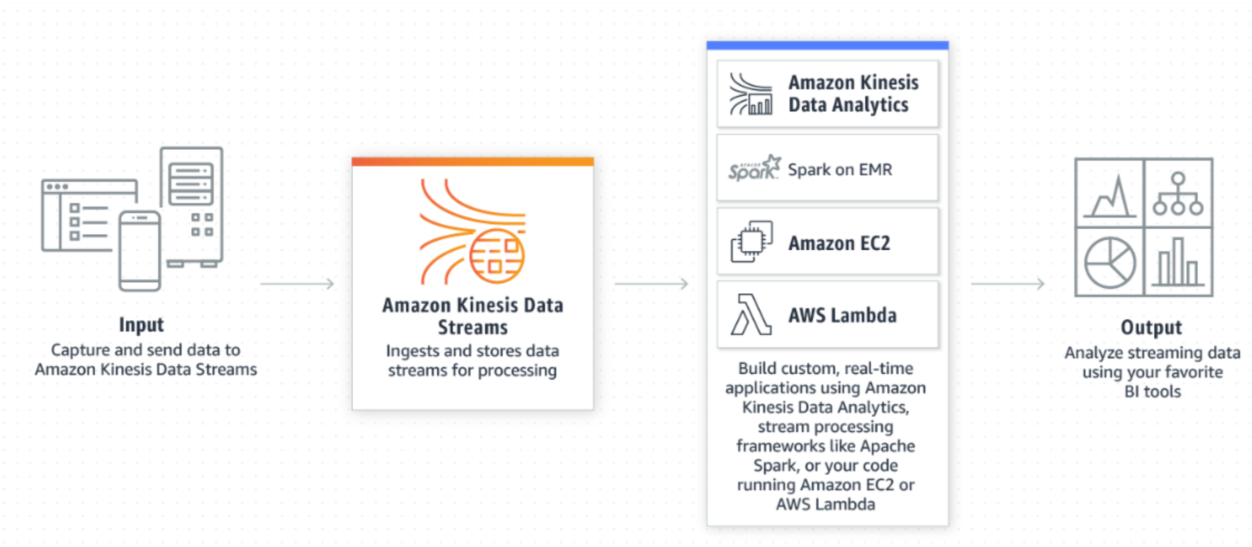
delimited
stall_warnings
filter_level
language
follow
track
locations
count

Any or all of these parameters can be used to ingest the real-time Twitter Streaming Data. After connecting to the API using an HTTP response we can parse the real-time incoming data.

To ensure the real-time performance of the tweet-stream, we will implement ingesting the feed using AWS KDS(Kinesis Data Streams) which provides the infrastructure to handle such a huge volume of big data in the cloud at near real-time speeds which is cost effective and durable.

2.1.2 Stock API

We have used Alpha Vantage as our Stock data source. Alpha Vantage Inc. is a leading provider of free APIs for realtime and historical data on stocks, forex (FX), and digital/crypto currencies.

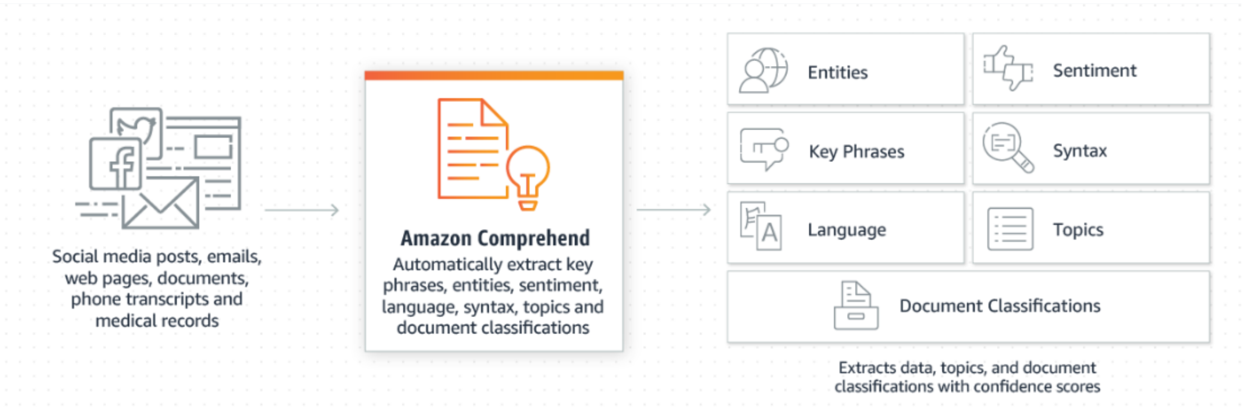

Input
Capture and send data to Amazon Kinesis Data Streams
Amazon Kinesis Data Streams
Ingests and stores data streams for processing
Amazon Kinesis Data Analytics
Spark on EMR
Amazon EC2
AWS Lambda
Build custom, real-time applications using Amazon Kinesis Data Analytics, stream processing frameworks like Apache Spark, or your code running Amazon EC2 or AWS Lambda
Output
Analyze streaming data using your favorite BI tools

2.2 Data Storage - Amazon S3

For storing the data we will use Amazon S3 which will work with Kinesis Data Streams in tandem and provide the appropriate solution to store and query data which is volatile in nature (hot data).

2.3 Natural Language Processing - Amazon Comprehend

For the Sentiment Analysis module **Amazon Translate** and **Amazon Comprehend** are implemented. They will perform comprehension and NLP on the real time Twitter data.


Social media posts, emails, web pages, documents, phone transcripts and medical records
Amazon Comprehend
Automatically extract key phrases, entities, sentiment, language, syntax, topics and document classifications
Entities
Sentiment
Key Phrases
Syntax
Language
Topics
Document Classifications
Extracts data, topics, and document classifications with confidence scores

2.4 Serverless Compute - Amazon Lambda

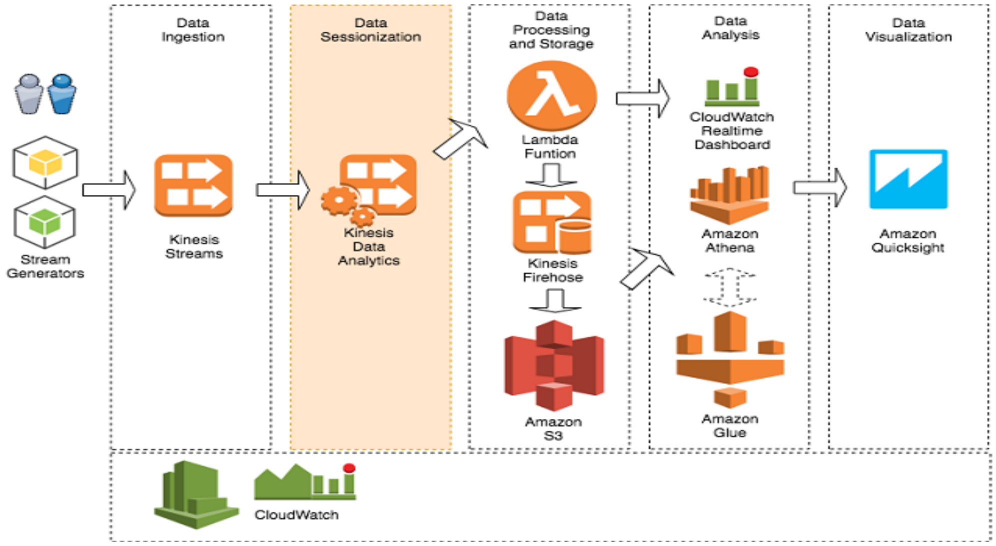
Amazon Lambda is used as a trigger for object creation in S3 buckets. This is detected by Amazon Comprehend and Amazon Athena.

2.5 Serverless Querying - Amazon Athena

Amazon Athena will quickly enable us to leverage and query the volume of data quickly stored in S3 with the help of AWS Glue crawler for optimization.

2.6 Data Analysis - AWS Quicksight

AWS Quicksight will enable rapid prototyping of dashboards for dynamic visualizations and insights in the real time data.



3.0 Features List

- **Ingestion of live streaming data**

  The streaming data is ingested continuously from Twitter. This allows for near-real-time analytics being displayed to the user. The user can then act based on this data such as selling or buying stock.

- **Storage of data**

  The incoming data from the data source will be stored. This is an intermediate dataset because the data needs to be processed. Data transformations and cleaning be applied to this raw intermediately dataset.

- **Machine learning**

  Sentiment analysis on the processed data will be done using NLP (Natural Language Processing).

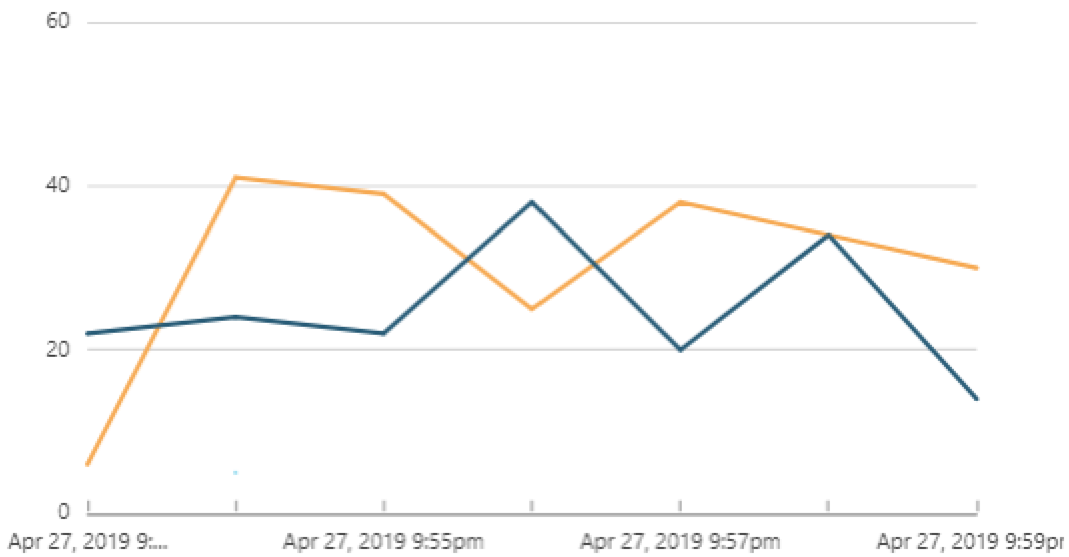
- **Near-Real-Time Analytics**

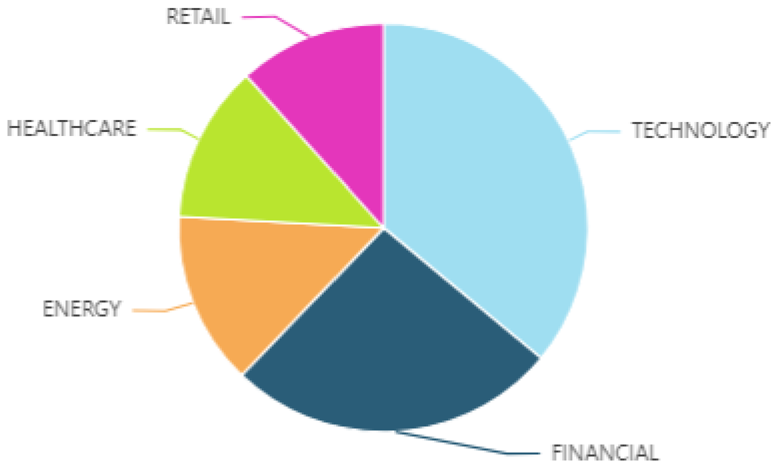
  Since the data being used is live streaming data, it will give the user near-real-time analytics.

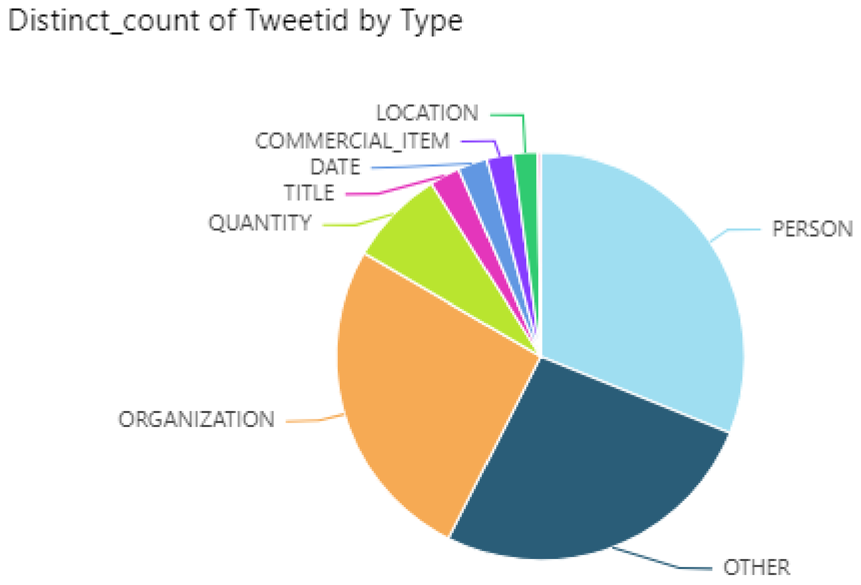
- **Dashboard for visualization**

  There will be a dashboard provided for visualization. Users can also run queries via RedShift to extract additional insights.

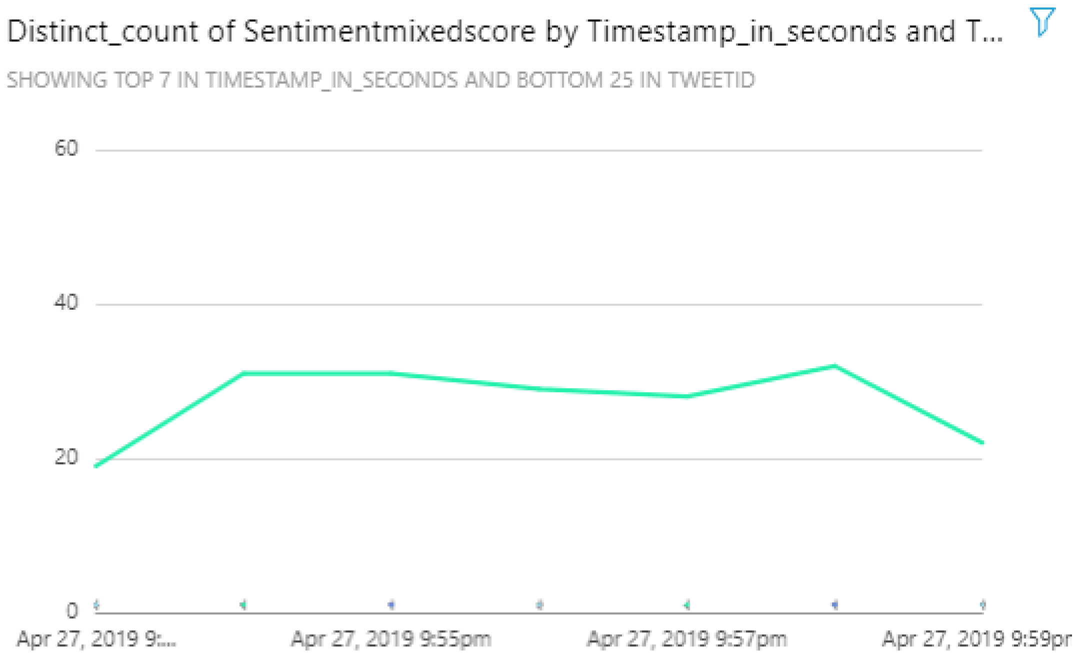
4.0 Sample Demo Screenshots


Count of Records by Sentiment and Timestamp_in_seconds


Sum of Price by Sector

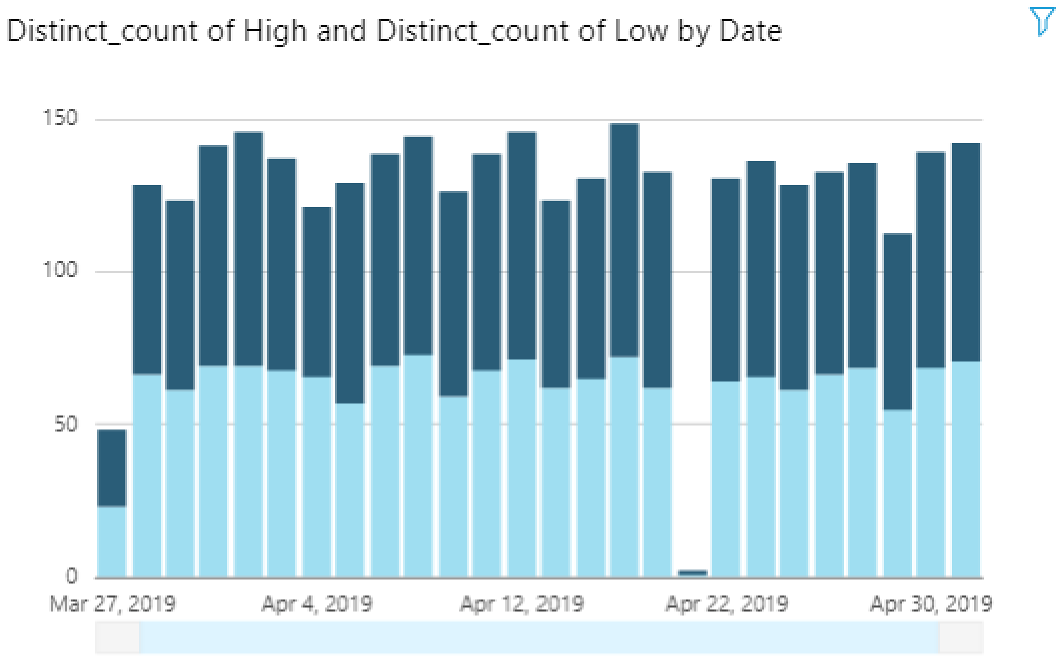


Distinct_count of Tweetid by Type



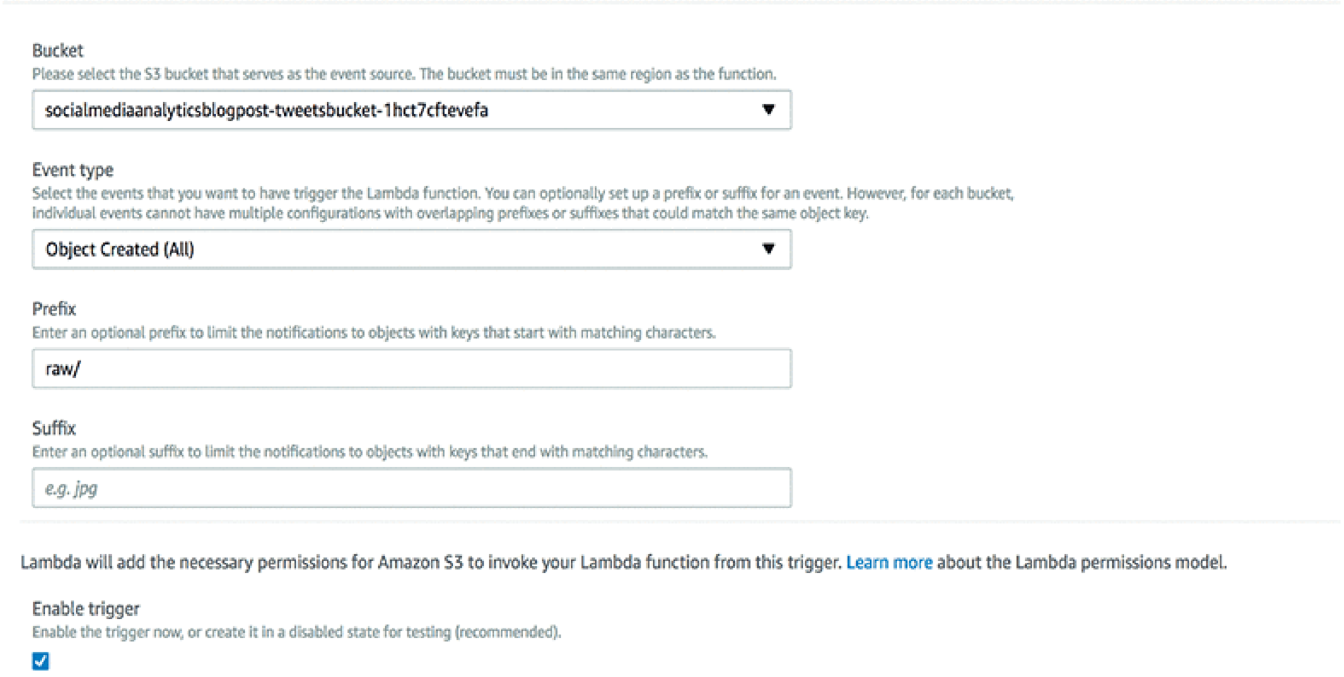
Distinct_count of Sentimentmixedscore by Timestamp_in_seconds and T...

SHOWING TOP 7 IN TIMESTAMP_IN_SECONDS AND BOTTOM 25 IN TWEETID


Distinct_count of High and Distinct_count of Low by Date
150
100
50
0
Mar 27, 2019
Apr 4, 2019
Apr 12, 2019
Apr 22, 2019
Apr 30, 2019

5.0 AWS Configuration Screenshots

Setting up S3 Buckets -



Bucket
Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.
socialmediaanalyticsblogpost-tweetsbucket-1hct7cftevefa

Event type
Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.
Object Created (All)

Prefix
Enter an optional prefix to limit the notifications to objects with keys that start with matching characters.
raw/

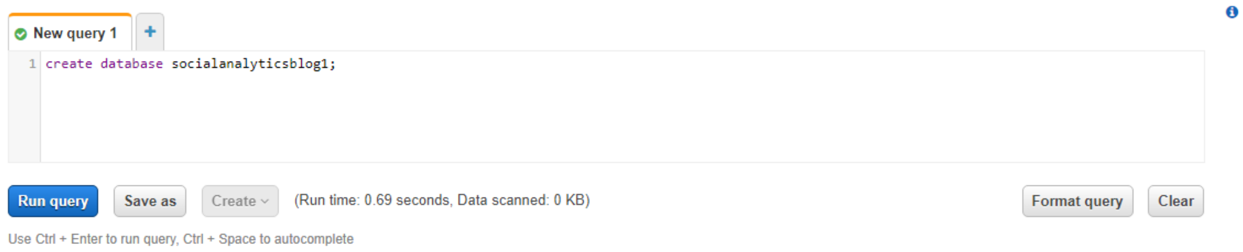
Suffix
Enter an optional suffix to limit the notifications to objects with keys that end with matching characters.
e.g. jpg

Lambda will add the necessary permissions for Amazon S3 to invoke your Lambda function from this trigger. Learn more about the Lambda permissions model.

Enable trigger
Enable the trigger now, or create it in a disabled state for testing (recommended).
☑

We have deployed 4 S3 Buckets here. They hold datasets for Twitter and Stock data.

Athena -
To create a new database in Athena -



```
create database socialanalyticsblog1;
```

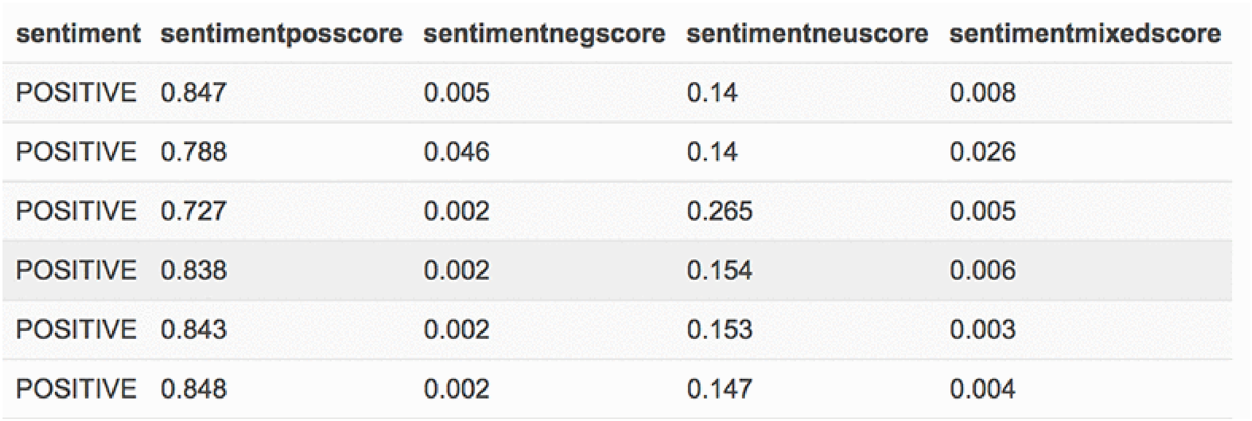
(Run time: 0.69 seconds, Data scanned: 0 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Create a Tweets table, followed by the entities and sentiment table. We can now run queries to investigate the data that has been collected.
If we want to analyze the positive tweets and see their corresponding score in sentiment analysis, we can implement the following query,

```
select * from socialanalyticsblog1.tweet_sentiments where sentiment =
'POSITIVE' limit 20;
```

Sentiment scores are generated as follows:

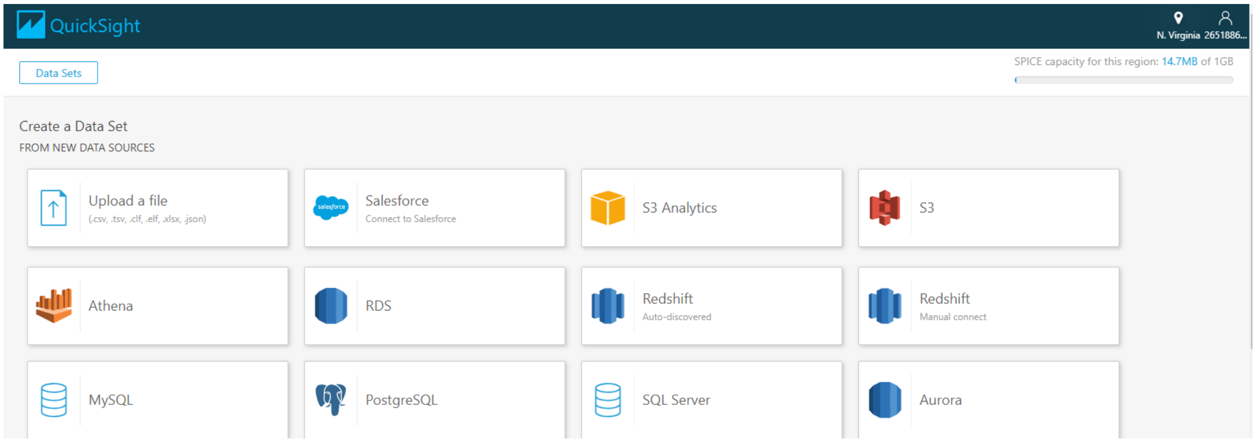


sentiment	sentimentposscore	sentimentnegscore	sentimentneuscore	sentimentmixedscore
POSITIVE	0.847	0.005	0.14	0.008
POSITIVE	0.788	0.046	0.14	0.026
POSITIVE	0.727	0.002	0.265	0.005
POSITIVE	0.838	0.002	0.154	0.006
POSITIVE	0.843	0.002	0.153	0.003
POSITIVE	0.848	0.002	0.147	0.004

QuickSight -
We use AWS QuickSight to provide a visual representation of the sentiment analysis. Here, we use the previously generated Athena tables and tweet_sentiments table to perform our analysis.

```
SELECT  s.*,
        e.entity,
        e.type,
        e.score,
         t.lang as language,
         coordinates.coordinates[1] AS lon,
         coordinates.coordinates[2] AS lat ,
         place.name,
         place.country,
         t.timestamp_ms / 1000 AS timestamp_in_seconds,
         regexp_replace(source,
         '\<.+?\>', '') AS src
FROM socialanalyticsblog.tweets t
JOIN socialanalyticsblog.tweet_sentiments s
    ON (s.tweetid = t.id)
JOIN socialanalyticsblog.tweet_entities e
    ON (e.tweetid = t.id)
```

Click on Save and Visualize. We can start building dashboards as per our analysis requirements.



The dataset to be analyzed is currently in the S3 bucket. We select the buckets that were deployed in the previous steps and begin with the visualization steps.

6.0 Suggested Solution

In this section we describe our suggested solution for each requirement present in the problem statement
Gauging the market with live Twitter analysis:

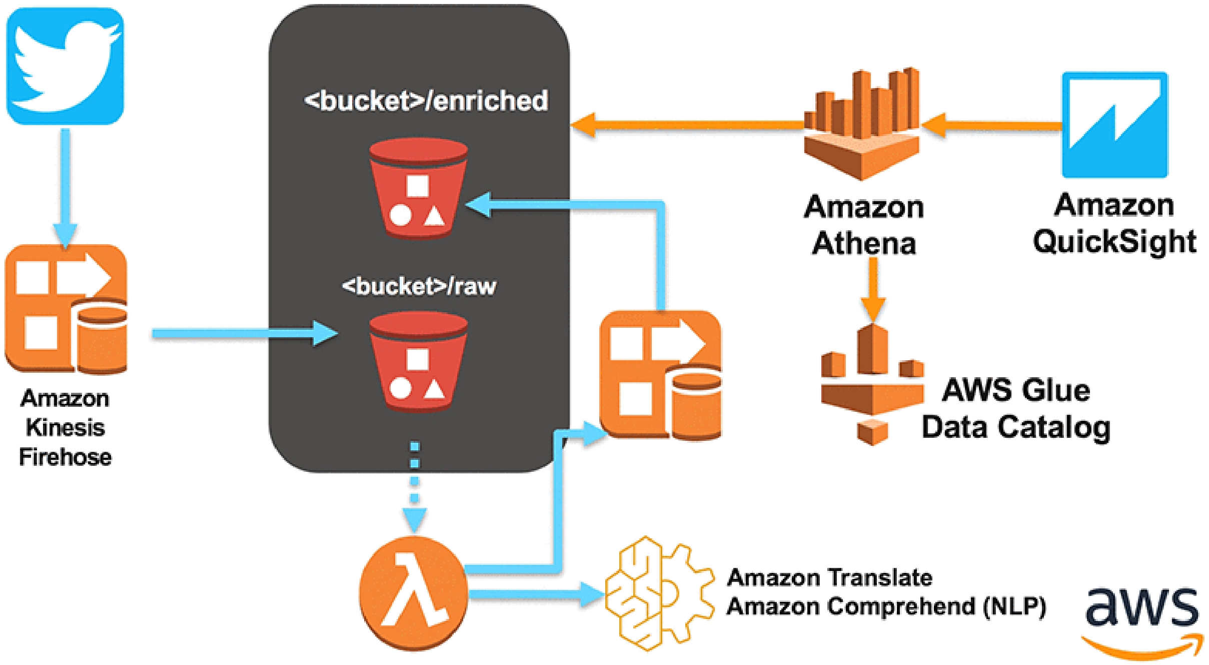
The live Twitter stream will be ingested with Amazon Kinesis for stream processing.
Further processing on streaming data will be done with Amazon Lambda and Amazon Translate and Amazon Comprehend which will extract valuable insight from the Twitter stream.

Gauging the market with live stock market analysis:

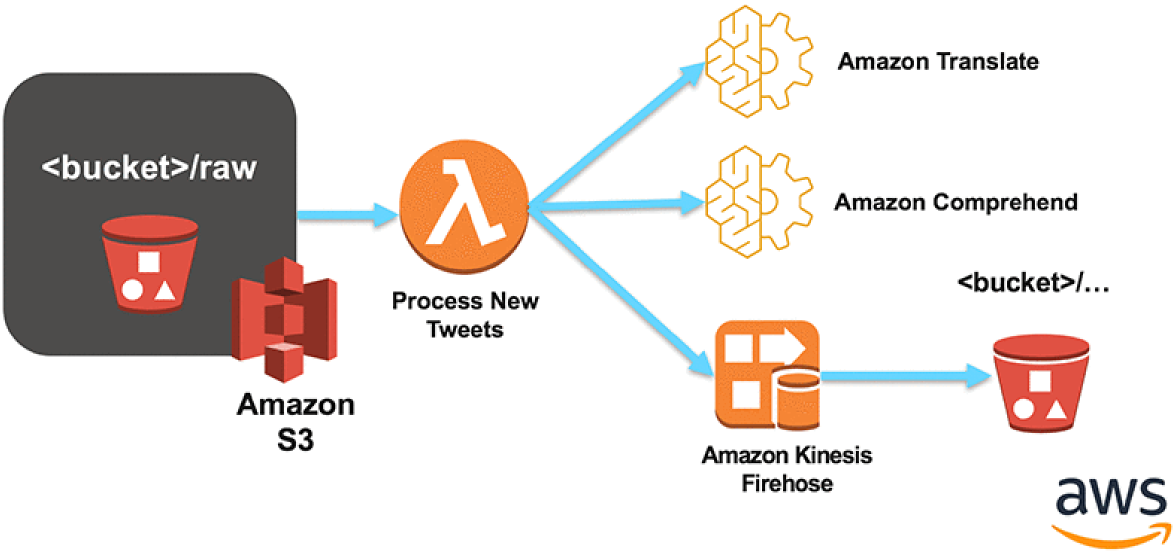
Near real-time analytics will be performed on the live streaming stock market data with the help of Amazon Kinesis stream ingestion and Amazon NLP features such as Amazon Comprehend and Amazon Translate.
Combined market insight from both Twitter and the stock market

Amazon QuickSight's business intelligence features will be used to derive and visualize meaning from the collective data of Twitter and the stock market, which will help the users to make real-time and well-informed decisions.

7.0 Solution Architecture Diagram



Using AWS for storing and analysing Twitter streaming data



AWS Components utilization for process flow

8.0 URL To Github Code Repo

https://github.com/aakashalurkar/cmpe266-project