CMPE 266
BIG DATA ENGINEERING AND ANALYTICS

STOCK PRICE PERFORMANCE PERCEPTION USING TWITTER STREAMS

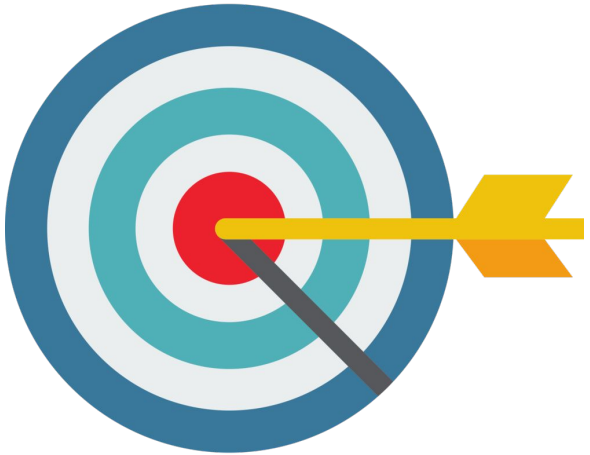
Aakash Alurkar	Aniket Deshpande	Madhu P. Kalluraya	Shivani Mangal	Zainab Khan
013716729	013532519	013708071	012530362	010120812

Introduction

- Only historical analysis of stock data is not sufficient to predict the Stock Market
- Human-emotions have tremendous impact on our decision making as an Investor
- Twitter trends display our collective human emotions real-time
- Hence, we are using Twitter to predict how human-emotions are impacting stock market real-time
- AWS cloud suite will be used to process, analyze and present real-time streaming Twitter/Stock data

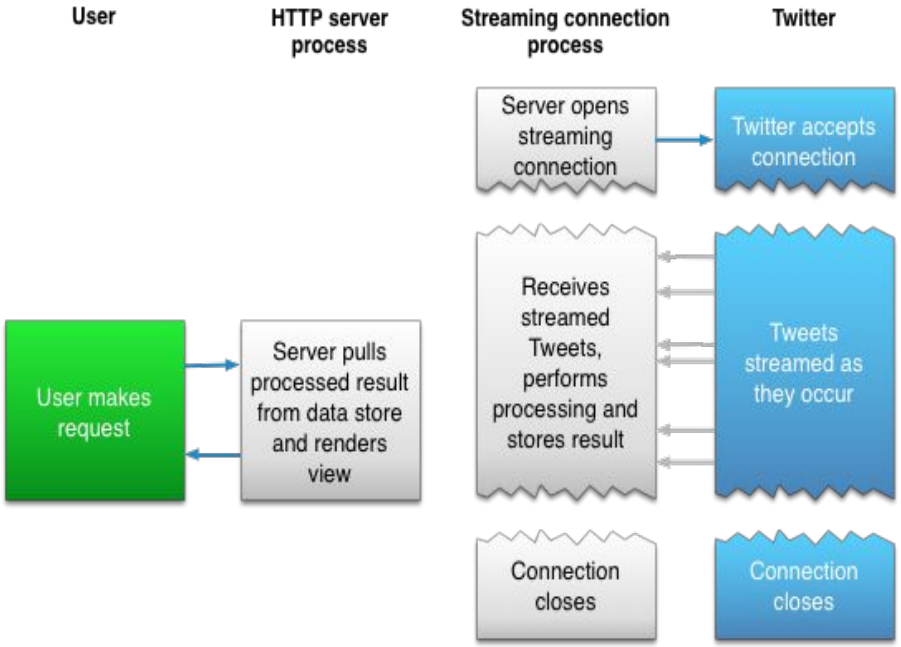
Aim

- Implement a streaming data pipeline and get realtime Twitter Streams for a particular company.
- Measure the tweet sentiments of the real-time streaming data.
- Implement a cloud based solution for the entire infrastructure.
- View the results using interactive dashboards.

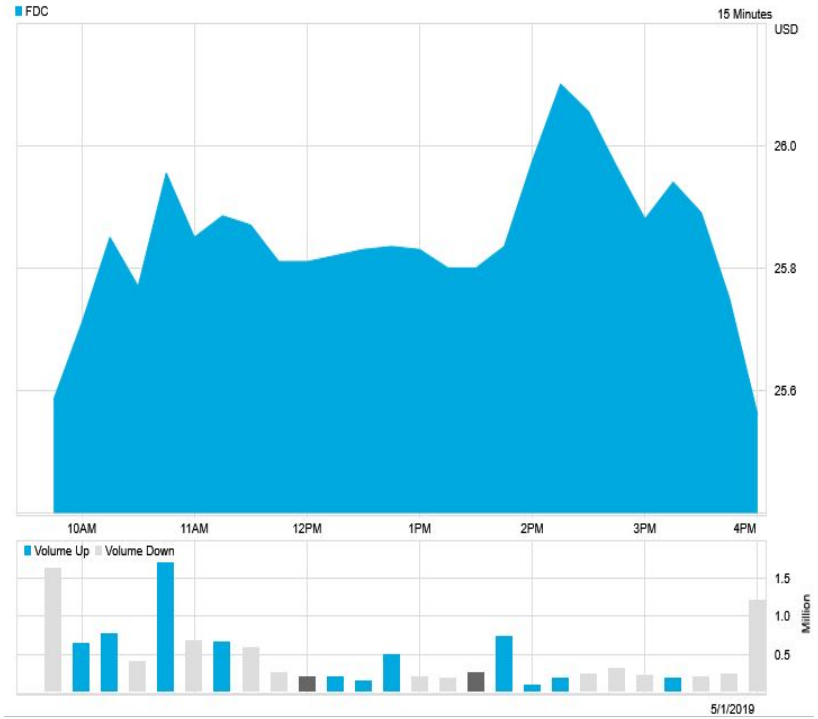


Why interact with a twitter API ?

- Twitter Developers API provides streaming endpoints to access the incoming tweets in real-time.
- The API response consists of the tweet data alongwith the delimited JSON-encoded activities, system messages, and blank lines.
- Realtime streams of data are initiated by sending a HTTP GET command to your custom URL
- Filter realtime streaming tweets with keywords, hashtags and other filters



Why interact with a Stock API ?

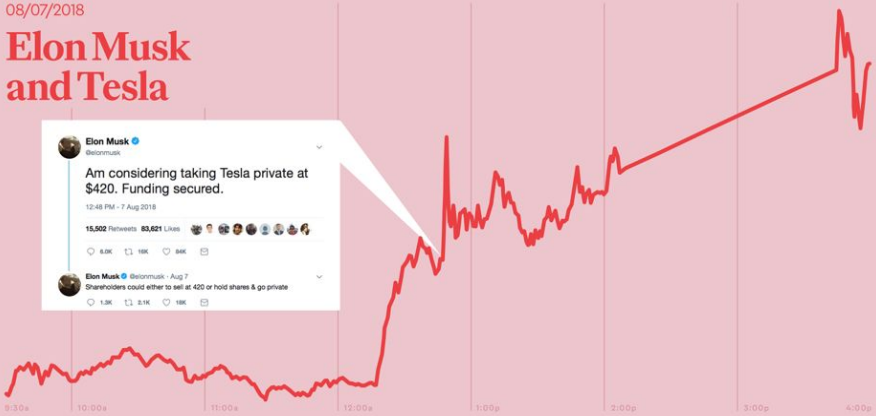


- Dynamic and realtime data which fits perfectly with the streaming data ecosystem use case.
- Ability to view stock prices for a particular ticker (company) and rate limiting it with a frequency filter.
- REST API for stock data with a JSON response.
- Can be correlated with Twitter data as most of the sentiments are expressed on Twitter which correlate to day trading stock prices.
- Historical Intraday trading prices provide us with stock price fluctuation with user reactions to announcements of any company.

How Tweets affect the Stock Prices

08/07/2018

Elon Musk and Tesla



02/21/2018

Kylie Jenner and Snapchat

01/05/2017

Trump and Toyota

08/13/2013

Carl Icahn and Apple

Streaming Data in 2019

- Ability to view data and decrease decision-making time.
- Infrastructure cost reduced due to competing players and technological improvements.
- Analyze streaming data without storing the data
- Many open source streaming options (ActiveMQ, RabbitMQ, Apache Kafka, Apache Spark, Apache Flint)
- AWS Kinesis and AWS Firehose provide streaming infrastructures in the cloud and the ability to integrate it with Data Warehouses(Redshift), Interactive Dashboards (Quicksight), etc.

Amazon Kinesis Firehose

Flink

APACHE kafka®

Spark Streaming

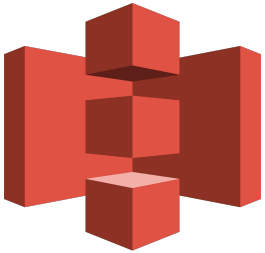
Stack used for Cloud

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud.

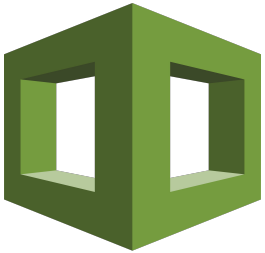
Amazon Athena is a serverless, interactive query service that makes it easy to analyze big data in S3 using standard SQL.

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text.

Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data stores and analytics tools.

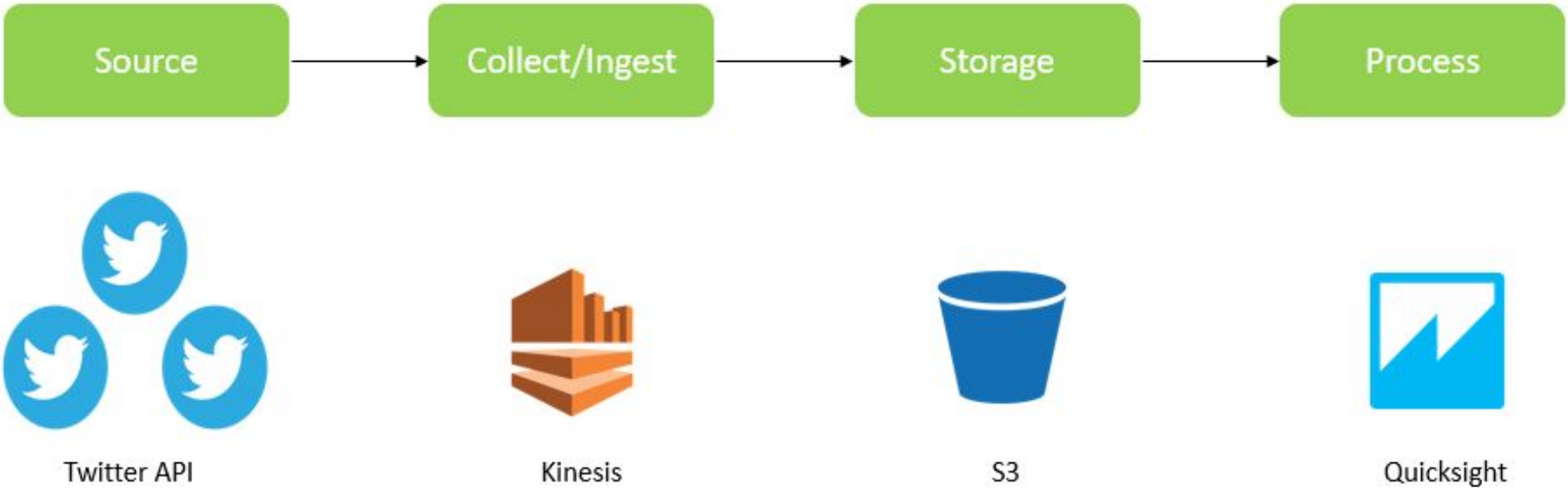


Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance.



AWS CloudFormation provides a common language for you to describe and provision all the infrastructure resources in your cloud environment.

Project Flow



API - Real-time Tweet Streaming

- The Twitter API accepts the connection
- The GET /{stream-type}/:stream API establishes a persistent connection to the Firehose stream, through which the realtime data will be delivered. API connects and retrieves the real-time twitter stream.
- This returns a JSON format data
- Particularly in Amazon Kinesis, the streaming has a limit of
  5 minutes of streaming data or
  5 MB of streaming data
- The twitter API closes the connection after the set constraints

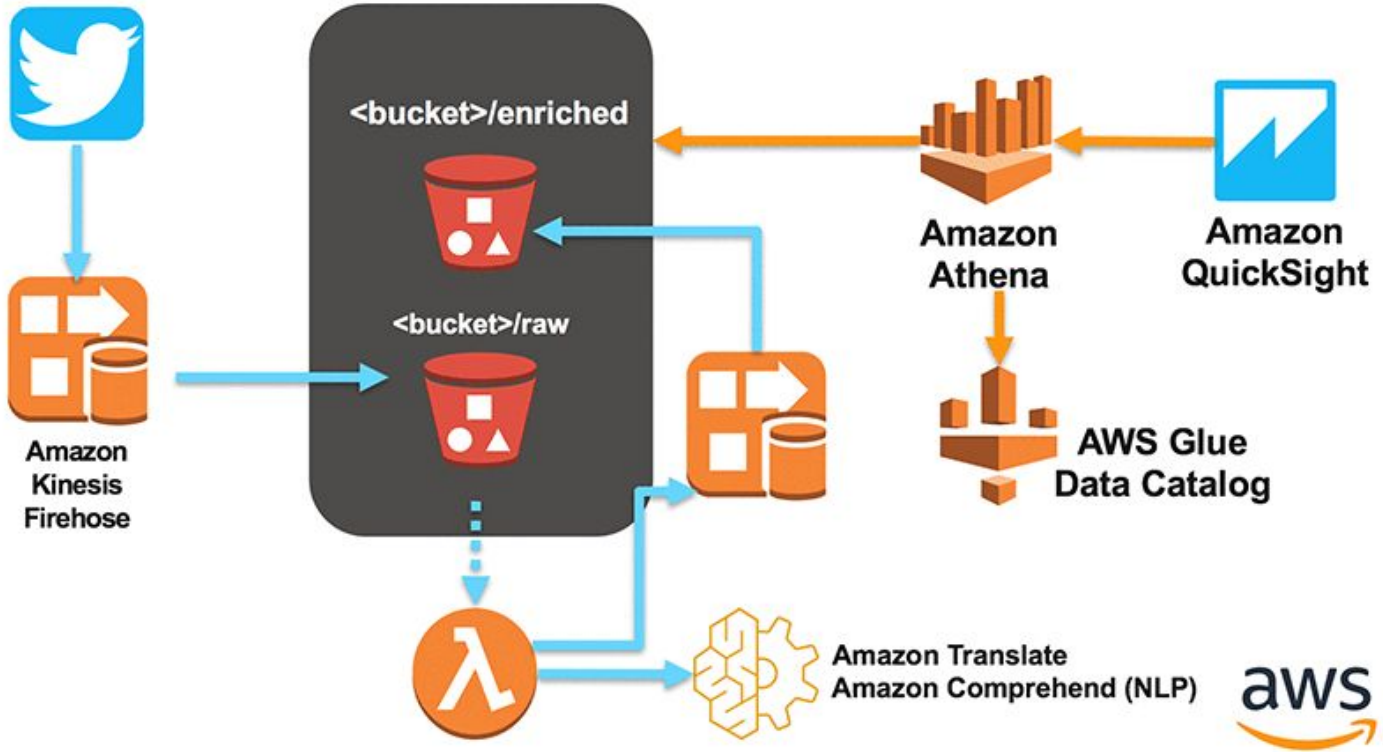

```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future of the
Twitter API
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news,
  "place": {
  },
  "entities": {
    "hashtags": [
    ],
    "urls": [
      {
        "url": "https:\/\/t.co\/XweGngmxlP",
        "unwound": {
          "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
          "title": "Building the Future of the Twitter API Platform"
        }
      }
    ],
    "user_mentions": [
    ]
  }
}
```

API - Real-time Stock Prices

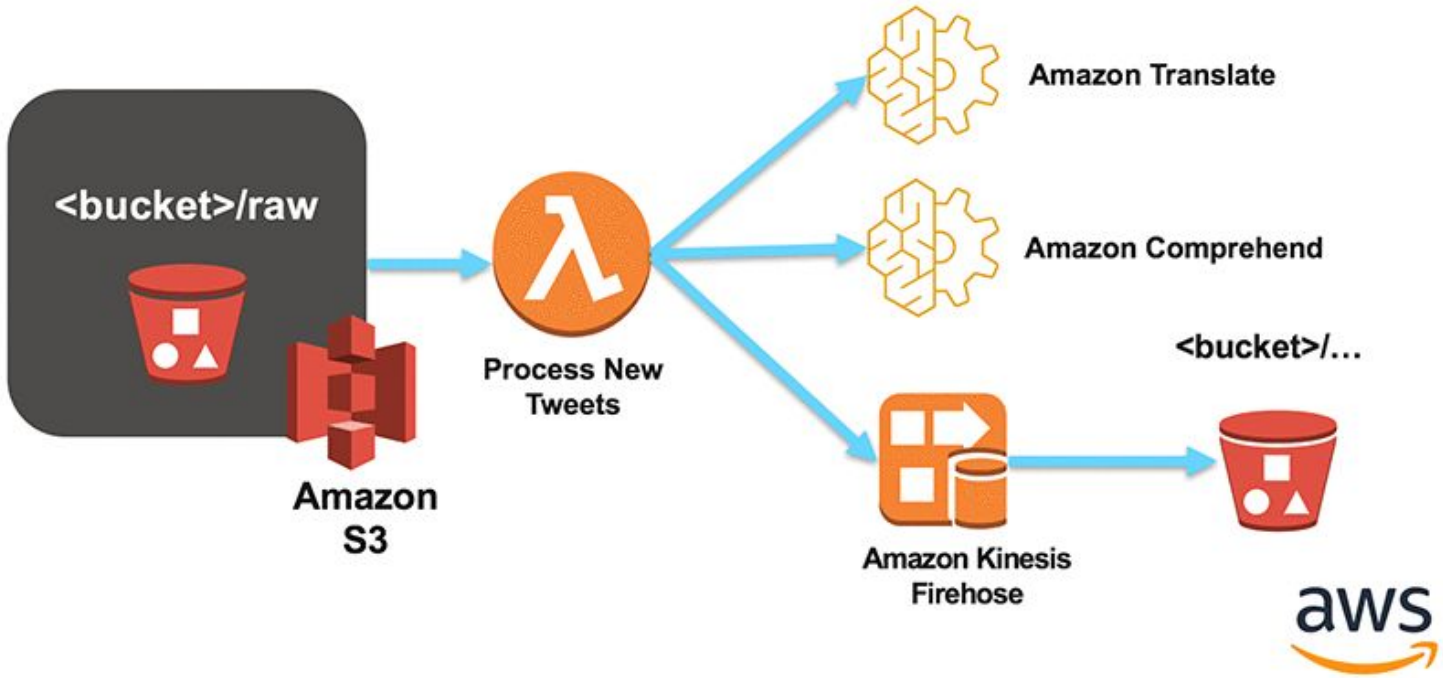
- The Tiingo API provides REST API endpoints to access the stock data.
- Historical data upto 30+ years and real-time intraday trading prices can also be accessed with an API token.
- The API enables 150,000 requests per day.
- Stock data is accessed with a ticker symbol.


```
[
    {
        "ticker": "AMZN",
        "adjClose": 1911.52,
        "adjHigh": 1943.64,
        "adjLow": 1910.55,
        "adjOpen": 1933.09,
        "adjVolume": 3116964,
        "close": 1911.52,
        "date": "2019-05-01T00:00:00+00:00",
        "divCash": 0.0,
        "high": 1943.64,
        "low": 1910.55,
        "open": 1933.09,
        "splitFactor": 1.0,
        "volume": 3116964
    }
]
```

Architecture Diagram

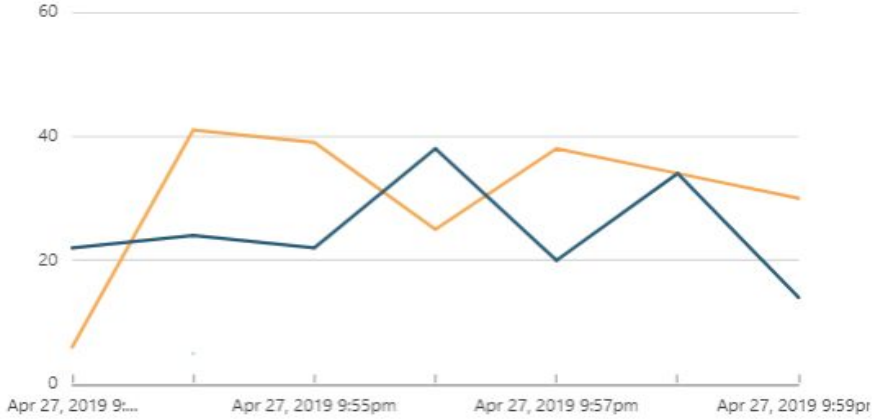


Architecture Diagram



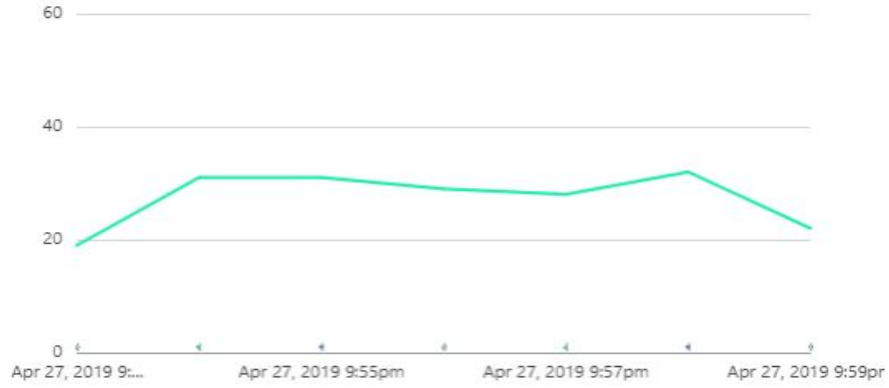
Results

Count of Records by Sentiment and Timestamp_in_seconds

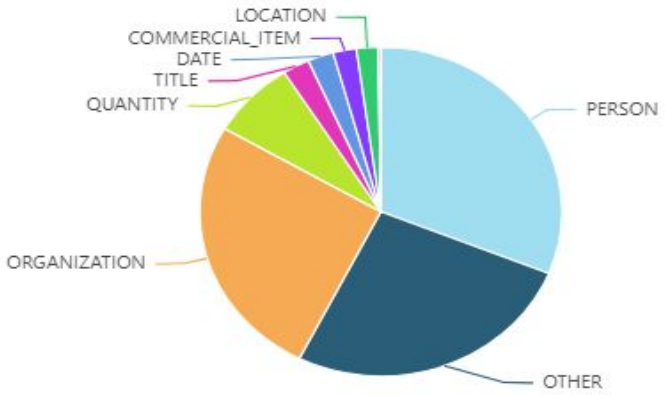


Distinct_count of Sentimentmixedscore by Timestamp_in_seconds and T...

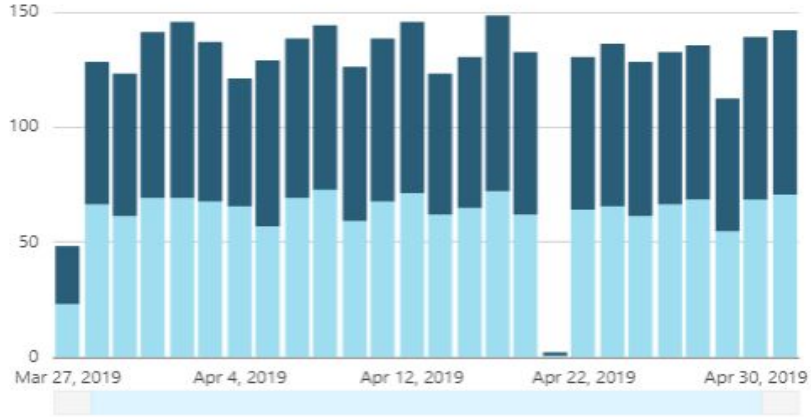
SHOWING TOP 7 IN TIMESTAMP_IN_SECONDS AND BOTTOM 25 IN TWEETID



Distinct_count of Tweetid by Type



Distinct_count of High and Distinct_count of Low by Date



Cost

Your Estimate				
Service Type	Components	Region	Component Price	Service Price
Amazon EC2 Service (US East (N. Virginia))				$81.21
	Compute:	US East (N. Virginia)	$81.21	
Amazon CloudFront Service				$0.10
	Data Transfer Out:	Global	$0.10	
Amazon CloudWatch Service (US East (N. Virginia))				$6
	Dashboard:	US East (N. Virginia)	$6	
Amazon Kinesis Service (US East (N. Virginia))				$12.46
	PUT Payload Unit cost	US East (N. Virginia)	$1.48	
	Shard Hour cost	US East (N. Virginia)	$10.98	
Amazon Comprehend	Amount of Text	US East (N. Virginia)		$16
Amazon Athena	Per Query / Data Dcanned	US East (N. Virginia)		$44.53
	Extended Retention cost	US East (N. Virginia)	$0	
AWS Support (Basic)				$0
	Support for all AWS services:		$0	
		Free Tier Discount:		($0.10)
		Total Monthly Payment:		$160.20

Future Scope for Scalability

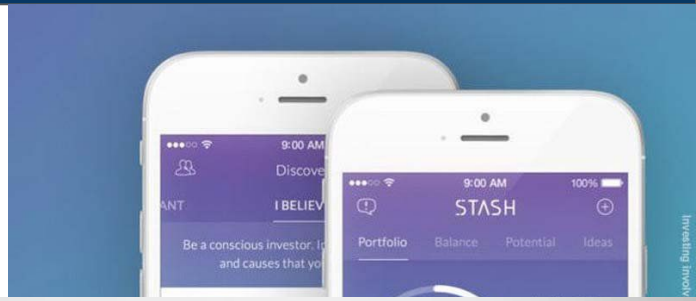
- Adding multiple kinesis streams and group stock by category.
- Implement a distributed architecture over the AWS suite multi-reigon instances enabled.
- Implement manual sharding in a kinesis stream to meet increasing demand.
- Implement a CloudWatch architecture to monitor and scale the platform.
- Integrate Amazon Application Scaling to seamlessly facilitate growth of application

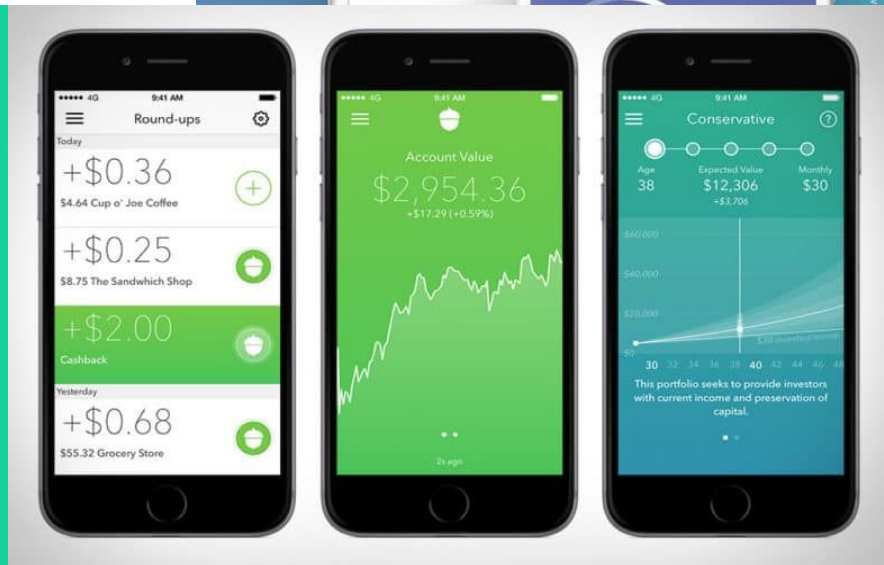
Conclusion

- Amazon Kinesis makes it easy to access and act upon real-time streaming data from an API
- Amazon QuickSight provides dashboard analysis prototyping for quick and actionable insights.
- CloudFront assists the user in building a Big Data and Streaming pipeline to handle the streaming, accessing the stored data and visualizing the data behaviour.
- The cost of implementing an entire pipeline with a powerful infrastructure is minimal for developing a quick PoC.

Use Cases

Applications such as Robinhood, Stash and Acorn which have democratized investing for the general public can use an insight from the stock price-twitter to decide their investments





Thank You