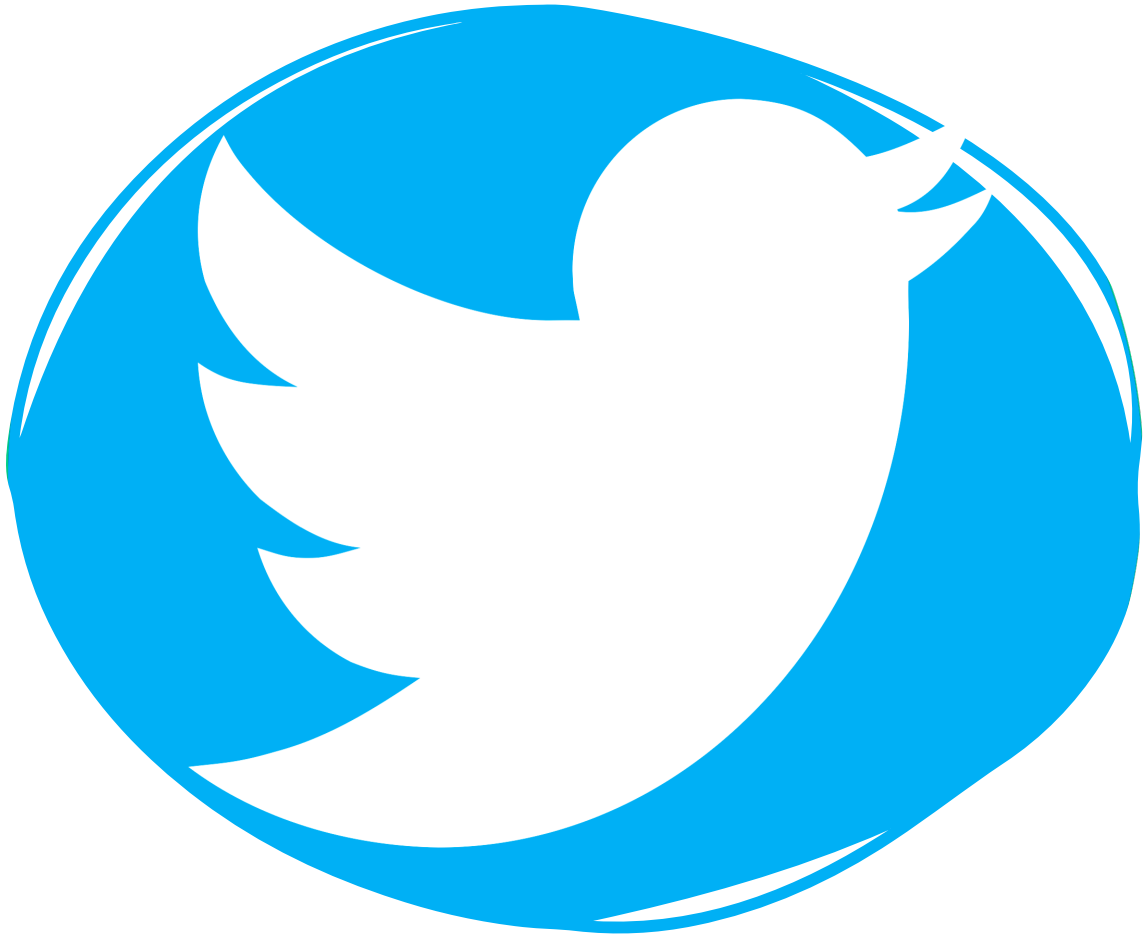


# COVID-19 TWITTER DATA ANALYSIS USING SPARK


- Abhay Kumar Singh
- Mayank Birla
- Prathiksha R Prasad
- Vedangi Bengali
- Vindhya Ningegowda

# AGENDA



Introduction

Cloud Infrastructure

Data Collection

Spark Analytics

Web Application

Challenges

Demo

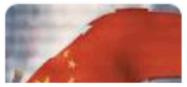
# INTRODUCTION

TNW

Twitter sees 900% increase in hate speech towards China — because coronavirus

US President Donald Trump has often labeled the coronavirus as the "Chinese virus." However, he's not alone in blaming China, the origin of ...

1 month ago



San Francisco Chronicle

Coronavirus Twitter: These Bay Area doctors are educating the public and becoming social media stars

Word that the Bay Area may be flattening the coronavirus curve swept across the nation this week thanks in part to a few UCSF physicians ...

3 weeks ago

CNET

Twitter says it's removed more than 1,100 misleading coronavirus tweets

Misinformation about the novel coronavirus is an ongoing problem networks like Twitter. Image by Pixabay; illustration by CNET.

3 weeks ago

CNN International

Doctors turn to Twitter and TikTok to share coronavirus news

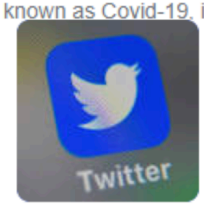
... has been using Twitter to share information about coronavirus, also known as Covid-19, including personal protective equipment for medical ...

Washington Post

On Twitter, almost 60 percent of false claims about coronavirus remain online — without a warning label

More than half of the misinformation about the coronavirus pandemic that has been debunked by fact checkers remains on Twitter without any ...

2 weeks ago



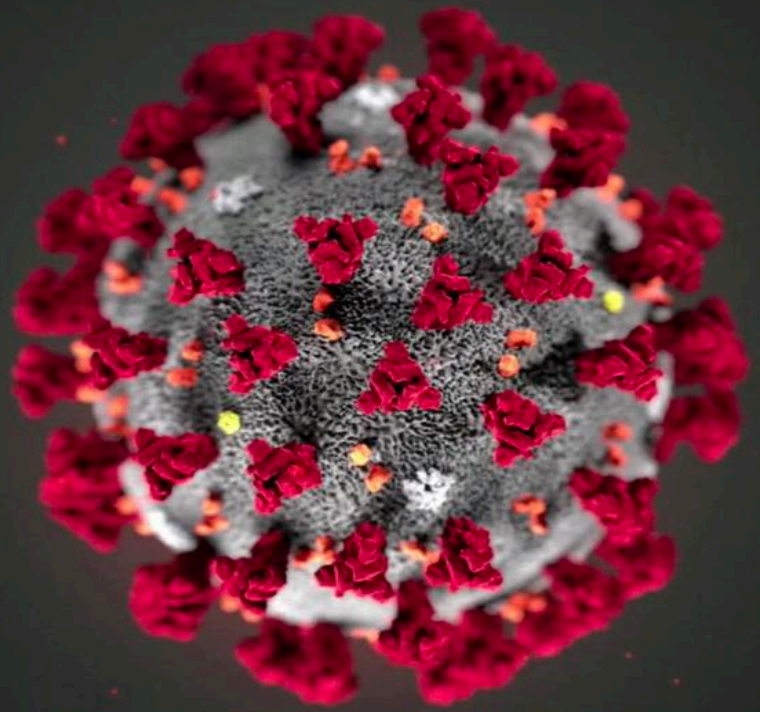

Vanity Fair

Conservative Voices Are Pumping Out Coronavirus Misinformation on Twitter

Twitter has frequently been criticized for acting slowly, if at all, to curb the spread of misinformation. But with an unprecedented pandemic ...

3 weeks ago

# PROBLEM STATEMENT



More than 560 million tweets about Covid19 since the 1st of January of 2020.

Twitter has large amount of structured updated data, making it suitable for employing ML models.

Apache Spark – distributed general-purpose cluster-computing framework to process big data from Twitter stream

Web Application – to stream processed twitter data.

# CLOUD INFRASTRUCTURE

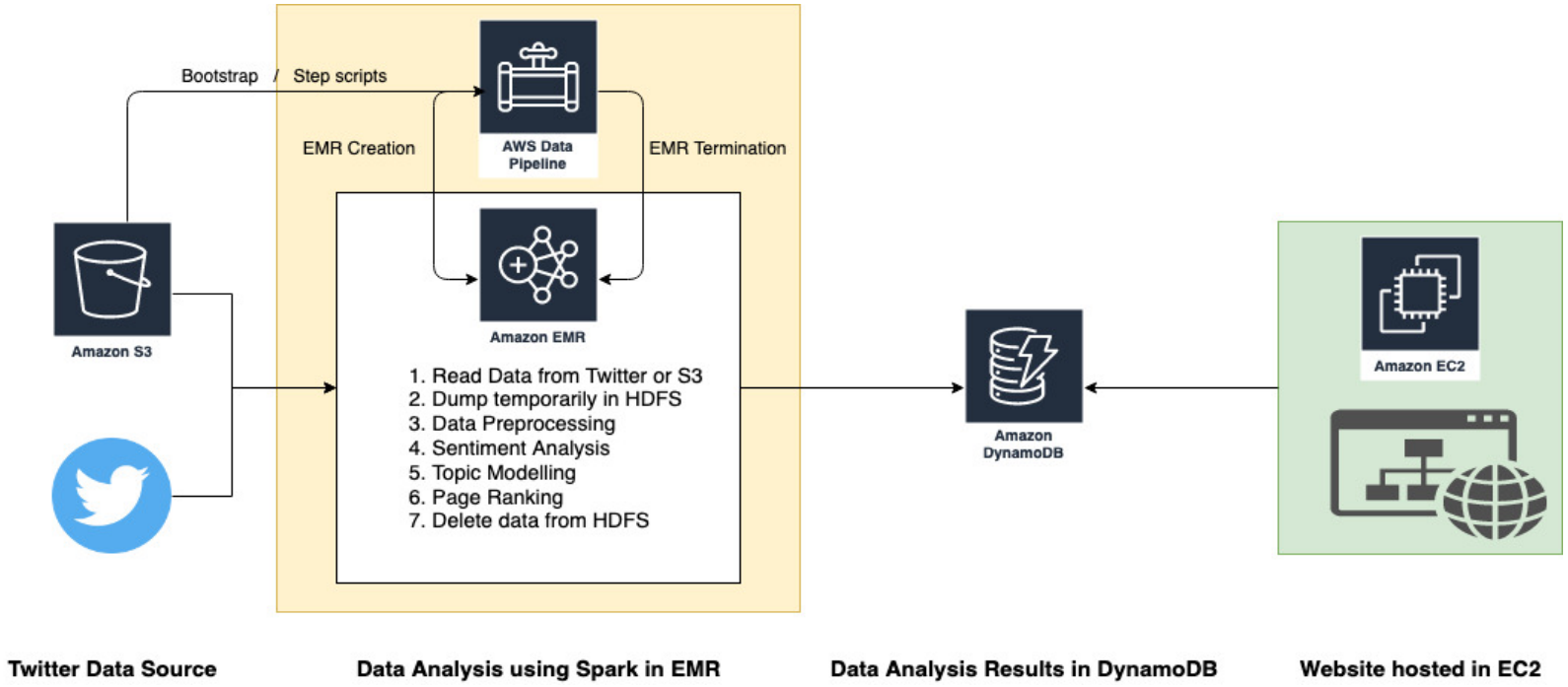
Built an end-to-end twitter data analysis pipeline on AWS.

Used several AWS cloud services:

- Elastic Mapreduce Cluster
- HDFS
- DynamoDb Storage
- S3
- EC2
- Data Pipeline

# AWS PIPELINE

- For historic data, tweets were pulled from S3.
- Daily Tweets are pulled from Twitter APIs.
- AWS Data-pipeline schedules the creation of daily cluster and script execution.
- Data Preprocessing and analysis is done in EMR using Spark and results are stored in DynamoDB
- Website is hosted on an EC2 instance.



# DATA COLLECTION

## Phase 1

- Worked on Covid-19 data available on Kaggle.
- Static data. Consists of tweets for the month of March.

## Phase 2

- Used twitter APIs through *tweepy* to collect daily data and capture the latest trends.
- Analytics modules are run daily to reflect new data.
- Filtered tweets for "Covid-19" related terms.
- Stored the tweets on HDFS for analysis.

# SPARK ANALYTICS

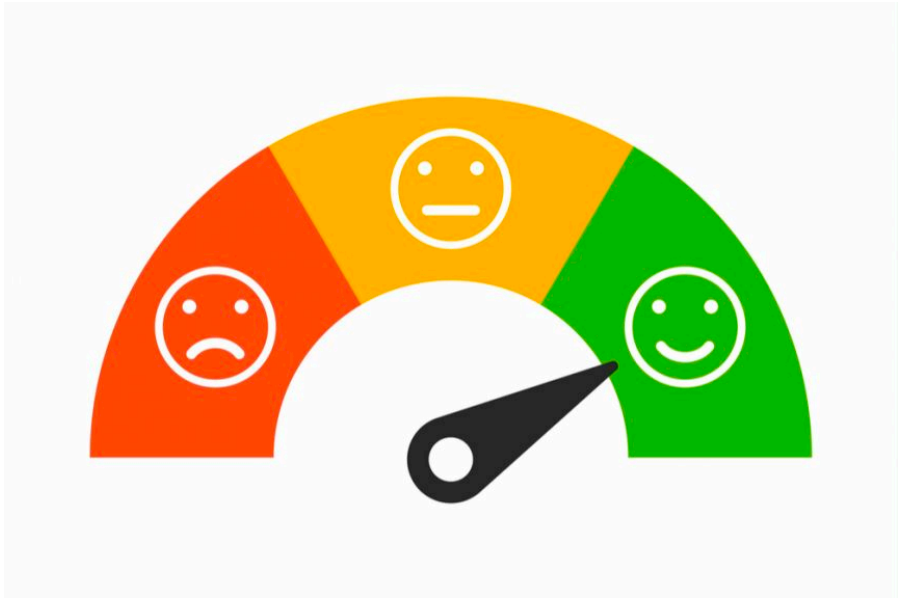
Performed core data processing and analysis using Spark on EMR cluster.

Explored various areas.

- Sentiment Analysis
- Topic Modelling
- Graph Analytics

# SPARK ANALYTICS - SENTIMENT ANALYSIS

- Intent.
  - To understand what the reaction of the general population is Positive or Negative to what extent.
- Technical Details.
  - Used polarity to classify each tweet as positive or negative.
  - Summarized the number of positive and negative tweets by date.
  - Stored the counts on DynamoDB.
- Tools Used.
  - NLTK Vader tools with Spark User Defined Functions (UDF).

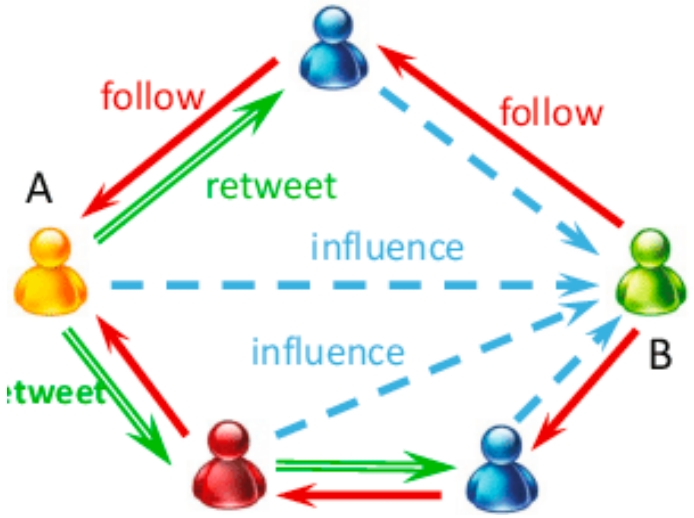


# SPARK ANALYTICS - TOPIC MODELLING

- Intent - To understand theme of Covid-19 related tweets
- Technical Details.
  - Pre-processed tweets to exclude stopwords.
  - Trained Latent Direchlet Allocation model to cluster related terms and learn 10 inherent topics.
  - Obtained top 10 words along with their weights from each topic.
  - Generated word cloud for words used in tweets
- Tools Used.
  - Spark ML Library

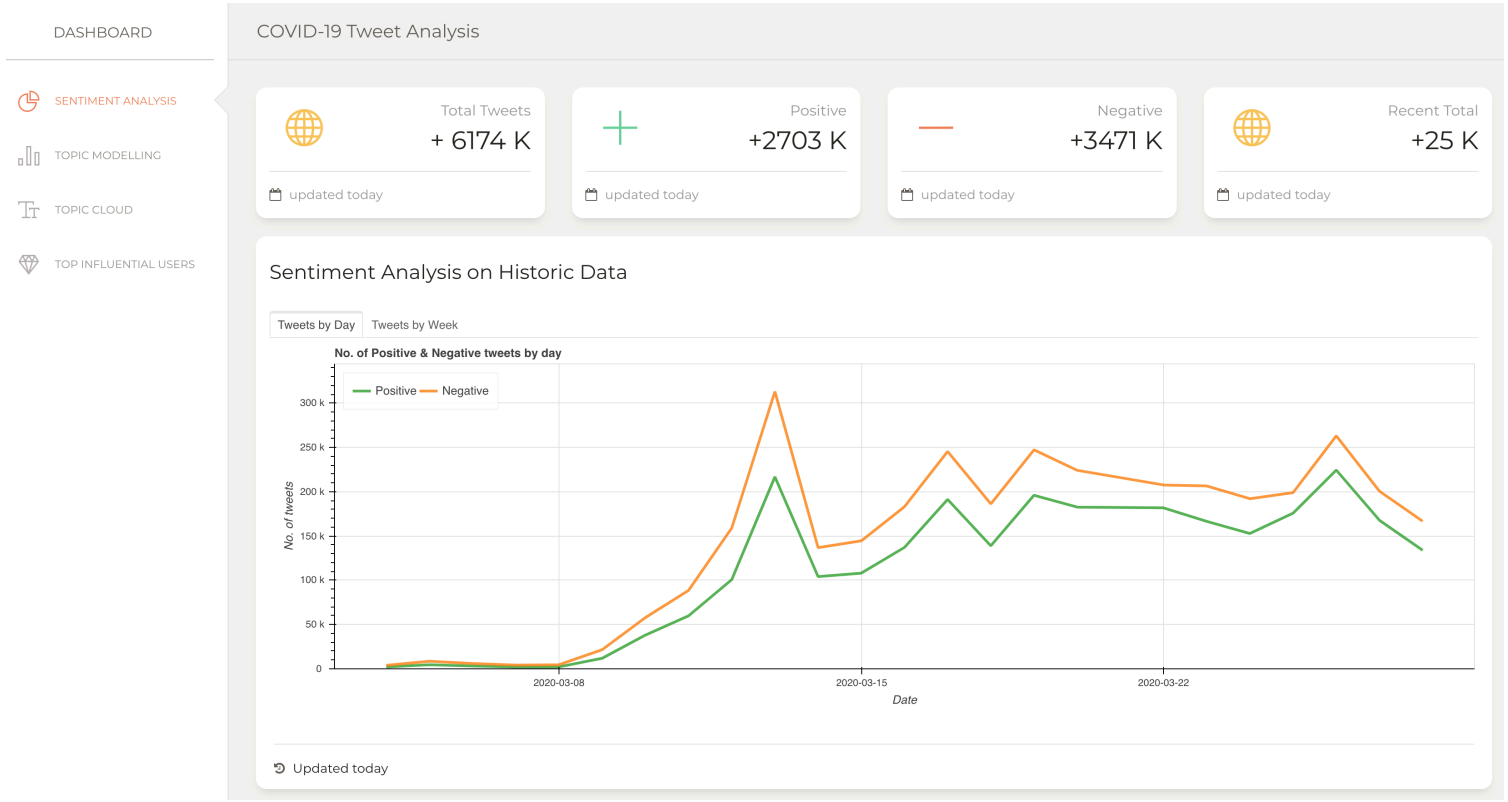
# SPARK ANALYTICS - GRAPH ANALYTICS

- Intent – Determine most mentioned users who are highly influential.
- Technical Details.
  - Constructed a user mentions graph.
  - Applied Pagerank algorithm on the user mentions graph.
  - Extracted the top 10 highly ranked users.
- Tools Used.
  - Spark Graphframes



# WEB APPLICATION

- Intent - A dashboard to visualize Covid-19 tweet analysis results.
- Language and Framework
  - Flask framework with Python backend.
  - Built interactive plots using Bokeh.
  - Server hosted on an EC2 instance.
  - Interactive visualizations.
    - By Sentiment
    - By Topic
    - Top Influential Users



# CHALLENGES

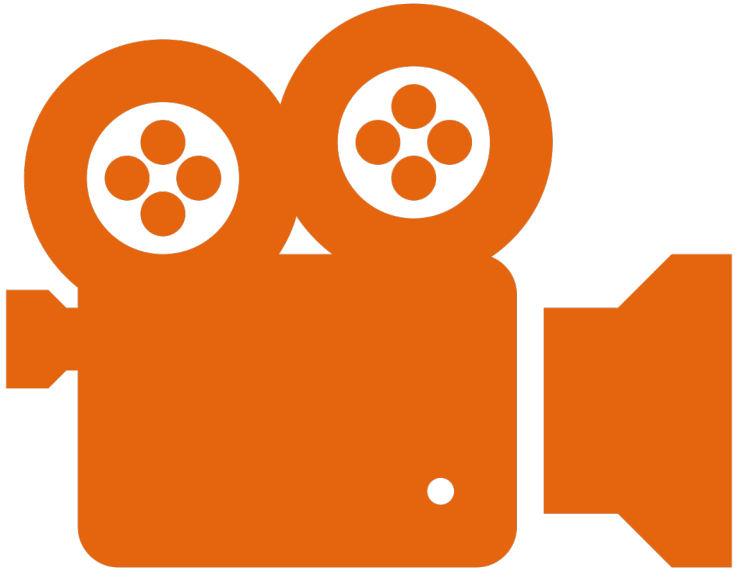
- Data Collection
  - Requested for Twitter Developer account for streaming.
    - Did not go through.
    - Got access 2 days back.
- AWS
  - Deciding the architecture involving the cloud components.
  - EMR Cluster crashed abruptly.

# CHALLENGES

- Data Pipeline
  - EMR Cluster is setup using Bootstrap action.
  - Data streaming and models are built using step function.
  - Error in python PATH variable, blocker for pipeline automation.
- Migrating from traditional ML libraries to Spark based libraries.

# DEMO TIME!

http://ec2-54-219-144-55.us-west-1.compute.amazonaws.com:5000/



# THANK YOU!