BIG DATA PLATFORM
COVID TWEET ANALYSIS

Bhadri Vaidhyanathan

SUMMARY

- □ 100Mil Tweets on Covid between Oct2021-Jan2022 reviewed and analyzed in the GCP using PySpark in JupyterLab
- □ Tech Stack:
  - □ GCP
  - □ Apache Spark – PySpark
  - □ JupyterLab
- □ Analysis themes:
  - □ Top Tweeters
  - □ Top Tweeter Personas
  - □ Location Analysis
  - □ Timeline Analysis

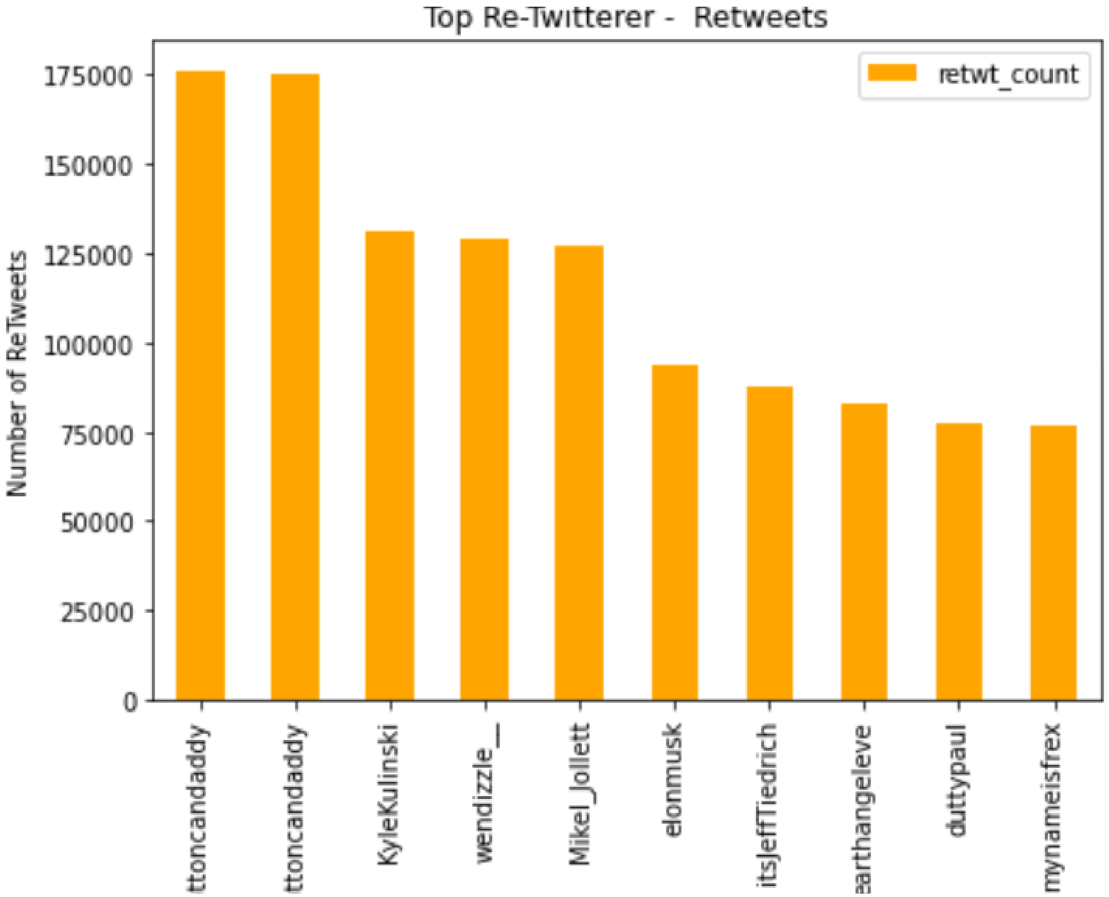
▪ Parquet format selected for maximum efficiency in size and analysis process time

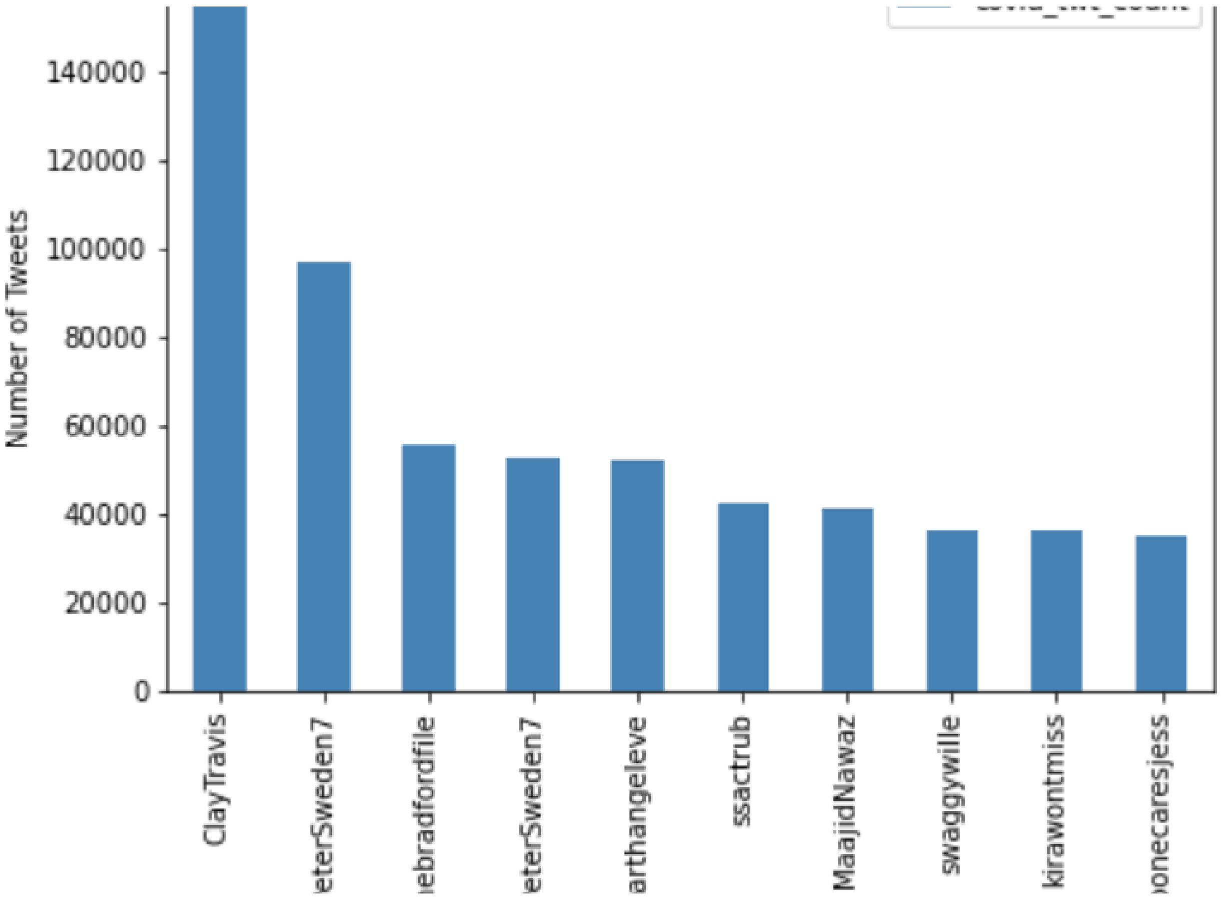
DATA PROFILE

- Size: 617.8 GB
- Data dates: Oct 15,2021 to Jan 25,2022
- Form: Tabular data with 100Mil rows or tweets and related meta data such as userid, timestamp, location and description.
- Preprocessing:
  - Filtering
    - Tweet and Retweets with ‘covid’: 61Mil rows (60% of original)
    - 50% of the extract are retweets (30Mil)
- Data stored as parquet partitioned by year, month and date

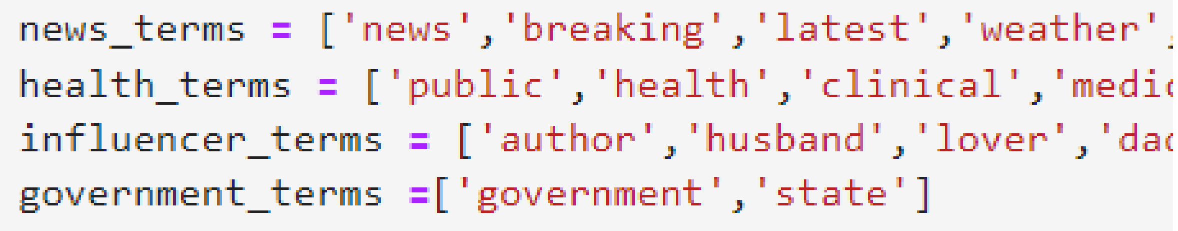
TOP TWEETERS

- ❑Total number of Unique Tweeters is 2Mil
- The top tweeters and re-tweeters as shown below:





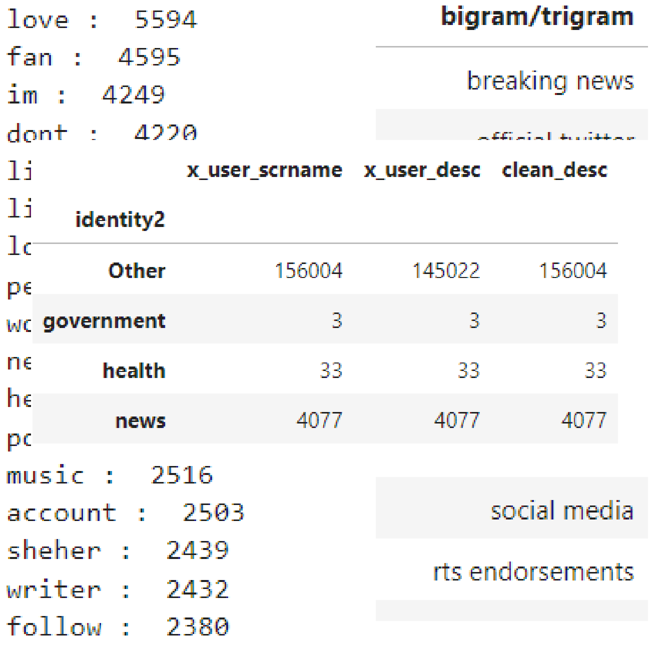
TWEETER PERSONAS



```
news_terms = ['news','breaking','latest','weather'
health_terms = ['public','health','clinical','medi
influencer_terms = ['author','husband','lover','da
government_terms =['government','state']
```

Methodology

- 2Mil users analysed
- User description used to identify personas
  - Descriptions broken down and analyzed and 4 common themes identified based on frequency of key terms
- NLP operations performed:
  - Description cleaned (stop words, punctuations etc) using Gensim
  - Word Count and Bi/Trigram analysis performed to and 4 groups, News, Government, health entity, Influencer and Others, created manually using the top terms.
- Based on the number of words in the description from these lists, the entities were classified



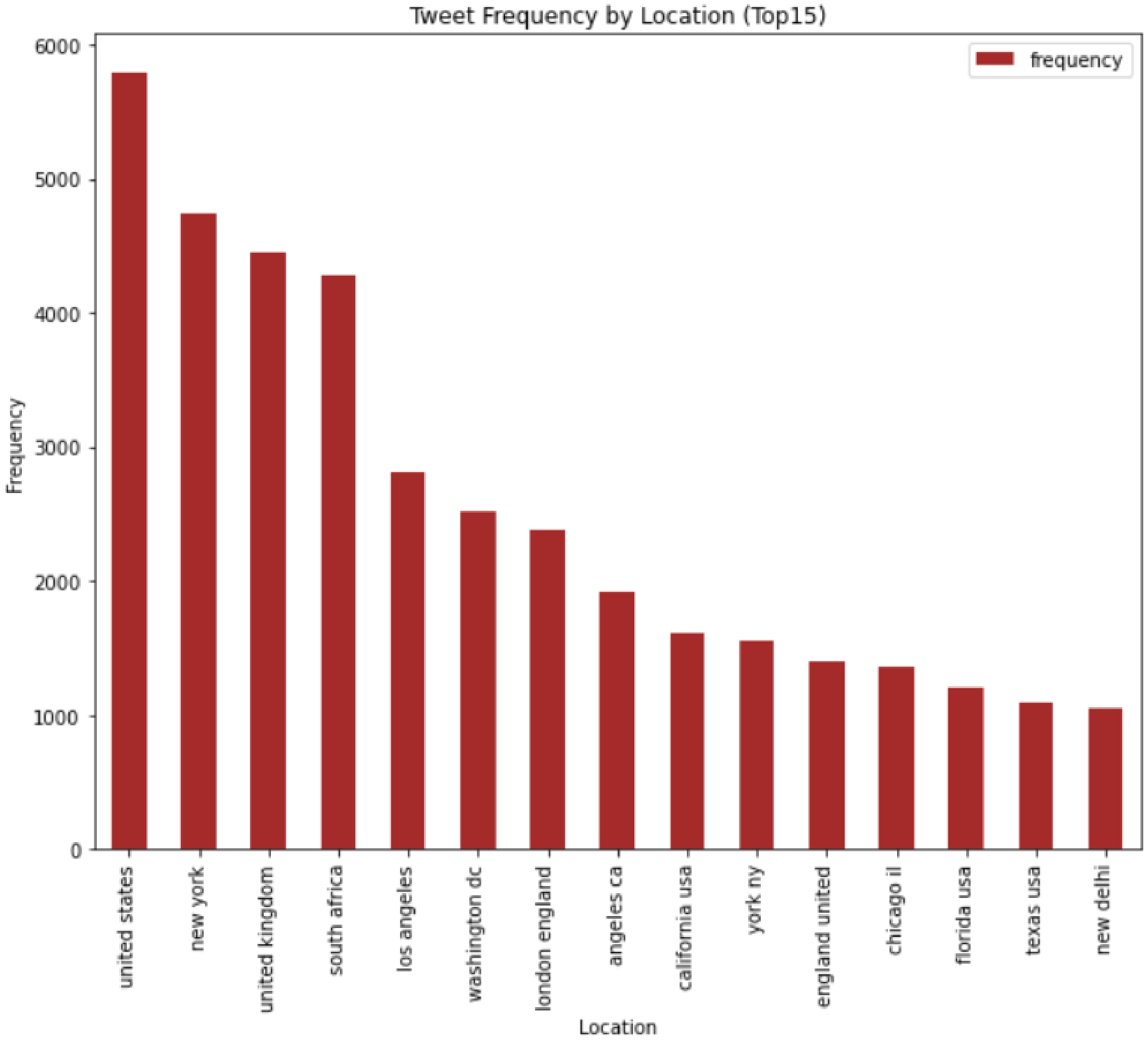
```
love :  5594
fan :  4595
im :  4249
dont :  4220
li
li
lo
pe
wo
ne
he
po
music :  2516
account :  2503
sheher :  2439
writer :  2432
follow :  2380
```

	x_user_scrname	x_user_desc	clean_desc
identity2			
Other	156004	145022	156004
government	3	3	3
health	33	33	33
news	4077	4077	4077

bigram/trigram
breaking news
official twitter
social media
rts endorsements

LOCATION ANALYSIS

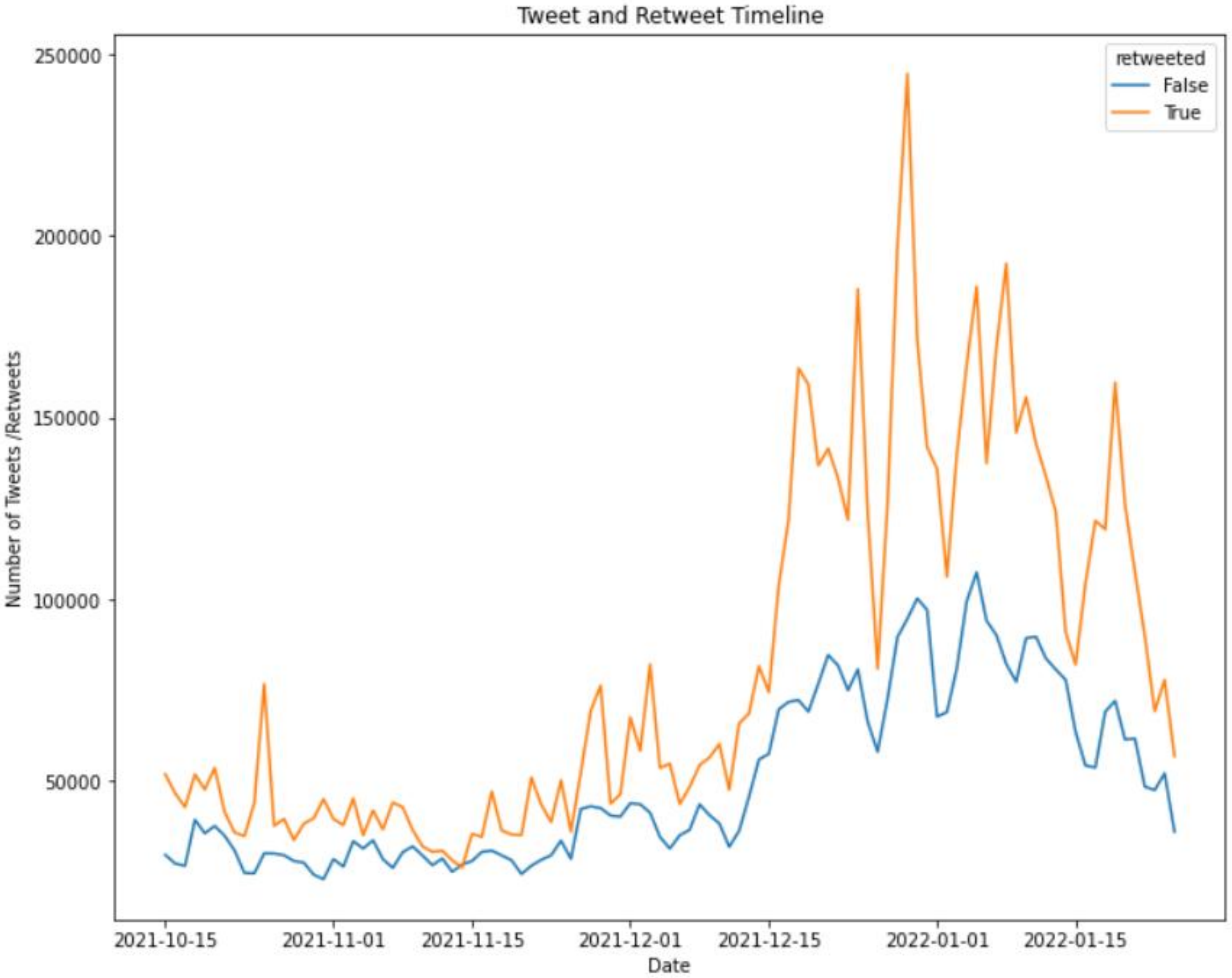
- United States and cities in US have the highest volume of tweet volume between Oct2021 - Jan 2022
- United Kingdom and South Africa also feature in the top list during this period. The discovery of the highly contagious Omicron variant in South Africa and its spread in UK during this time period explains their presence.



https://www.cnn.com/2021/12/02/world/south-africa-omicron-origins-covid-cmd-intl/index.html
https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern

TIME ANALYSIS

- Peak ReTweets in 29th Dec 2021
- Valley on 26th Dec in ReTweets&Twewets due to Christmas and on 1st/2nd 2022 due to New Year's holidays
- Missing Data:
  The tweet and retweet ratio seem to be similar in the Oct/Nov 2021 but the differences peak in Dec2021 abd Jan 2022 indicating a possible gap in data collection



CONCLUSION

- The peak in tweets does match correlate to the surge of the Omicron virus indicating the potential of using the Twitter platform as early warning for spread of contagious disease or occurrence of major disasters
- Identifying Personas of the Users (Individual or Public Institution) enables analysis of identifying key influencers and mediums for passing communication
  - Considering the impact that Twitterer seem to have verification could provide the platform more credibiilty