#coronavirus

Yuhsiang Hong
Jiazhang Cai
Hang Qi
Rahul Malhotra

Synopsis

- Data Collection
- Data Storage
- Indexing
- Sample Queries
- Caching
- Search Application
- Possible Improvements
- Conclusions

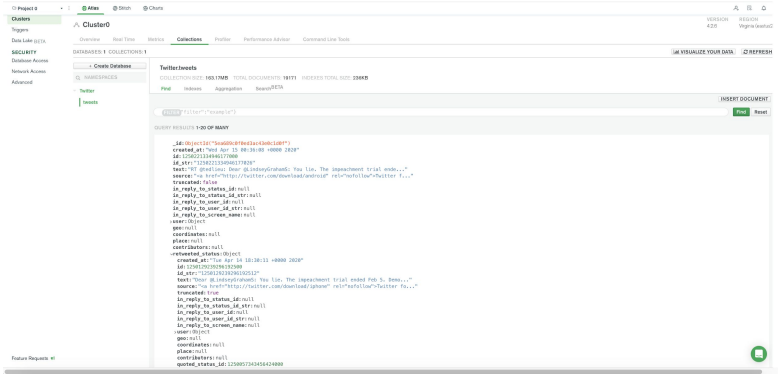
Data Collection

- We collected tweets which contained the hashtag “coronavirus”
  - saved as JSON objects
- We obtained a little over 19,000 tweets (about 180 MB) on April 14th in about 10 minutes
  - Very fast as expected due to the popularity of the topic
- We also expect there to be a lot of sub-queries and “sub-hashtags” since it has had such a profound impact on daily life
  - ex.) quarantine, social distancing, economy

Data Storage

- To store the collected data, we separated it into the “tweets” data and “user” data
- For the “tweets” data, we stored it using MongoDB, a non-relational database, since the data for all tweets is not the same
    - ex.) some tweets quote other tweets
- For the “user” data, we stored it using PostgreSQL, a relational database, since the data is consistent across users

Tweets Data



- We stored the tweets in two different ways:
  - 1) Pymongo
  - 2) Cloud version of MongoDB
- Using Pymongo, we were able to implement MongoDB using Jupyter Notebook and work on the indexing, caching, and queries here
- However, viewing the data in Jupyter Notebook can be messy
  - Atlas offers a nice way to view data, as well as it have it uploaded to the cloud

Users Data

- Each user data contains 38 or 39 keys
  - Some users don't have "profile_banner_url"
- Keep the keys that we will be used for querying or further analysis such as "id", "name", "followers_count", and drop the rest that are not important to us
  - Remain 11 keys in each user data
- Since each user's id is unique, we can set Twitter user's 'id' as our primary key


```
{'id': 531629036,
 'id_str': '531629036',
 'name': 'Creeds Cannon',
 'screen_name': 'ThucydidesTried',
 'location': '⭐TEXAS⭐',
 'url': None,
 'description': 'Free Market, Strong Def, Ltd Gvt. Chronicling the decline of the Republic & the fight to save Her.
🇺🇸#MAGA #Cruz #Trump #Qanon',
 'translator_type': 'none',
 'protected': False,
 'verified': False,
 'followers_count': 11697,
 'friends_count': 12307,
 'listed_count': 72,
 'favourites_count': 73023,
 'statuses_count': 107291,
 'created_at': 'Tue Mar 20 20:53:16 +0000 2012',
 'utc_offset': None,
 'time_zone': None,
 'geo_enabled': True,
 'lang': None,
 'contributors_enabled': False,
 'is_translator': False,
 'profile_background_color': '000000',
 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
 'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png',
 'profile_background_tile': False,
 'profile_link_color': 'ABB8C2',
 'profile_sidebar_border_color': '000000',
 'profile_sidebar_fill_color': '000000',
 'profile_text_color': '000000',
 'profile_use_background_image': False,
 'profile_image_url': 'http://pbs.twimg.com/profile_images/1158158098625957888/LcyR_ws7_normal.jpg',
 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1158158098625957888/LcyR_ws7_normal.jpg',
 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/531629036/1539050323',
 'default_profile': False,
 'default_profile_image': False,
 'following': None,
 'follow_request_sent': None,
 'notifications': None}
```



```
{'id': 531629036,
 'name': 'Creeds Cannon',
 'screen_name': 'ThucydidesTried',
 'protected': False,
 'verified': False,
 'followers_count': 11697,
 'friends_count': 12307,
 'listed_count': 72,
 'favourites_count': 73023,
 'statuses_count': 107291,
 'created_at': 'Tue Mar 20 20:53:16 +0000 2012'}
```

PostgreSQL

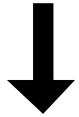
	id [PK] bigint	name text	screen_name text	protected boolean	verified boolean	followers_count bigint	friends_count integer	listed_count integer	favourites_count bigint	statuses_count bigint	created_at text
1	47868745023488	🔥 Sim...	Simargl4	false	false	3435	3540	0	4853	6505	Tue Jan 15 05:...
2	53338411245569	José	rjose316	false	false	14	68	0	2608	568	Fri Nov 08 17:5...
3	156549858	mceb72	MCEB72	false	false	5592	6087	12	19474	39468	Thu Jun 17 06:...
4	55374944	Debbie ...	debbiered15	false	false	3678	4969	4	97725	50267	Thu Jul 09 21:...
5	335037182	La Gue...	MayVenezolana	false	false	47420	2696	129	41370	192387	Thu Jul 14 01:...
6	72009411772416	Shark R...	SharkRadioNet	false	false	1072	2	14	276	57159	Mon Feb 10 20...
7	173211118	Healthy...	HealthAdv0cates	false	false	570	362	0	213	7757	Sat Jul 31 19:0...
8	58597033492484	Julio M...	ManonellasJulio	false	false	225	112	1	27600	9219	Sat Apr 08 03:5...
9	474029401	melelani	melelani22	false	false	185	73	8	75951	10394	Wed Jan 25 15...
10	15831179	Stay H...	YuriArtibise	false	false	9723	8304	676	98632	31811	Wed Aug 13 01...

- Install pgAdmin to show and check the users table in database
- Use Python on Jupyter Notebook to interact with PostgreSQL database by importing psycopg2 package
    - Twitter data is downloaded as a JSON file through Jupyter Notebook
    - Easy to query data from PostgreSQL and MongoDB on Jupyter Notebook
- Create a users table in PostgreSQL
    - The 11 keys that we decided to keep are our columns in the users table
    - Set id as our primary key

```
postgre_select_query = "SELECT * FROM users ORDER BY followers_count DESC LIMIT 3"
```

PostgreSQL

- Each user's data is an object with pairs of keys and values in JSON. Therefore, we need to break the structure down.
  - Use "for" loop to iterate all the keys in each user data
  - Implement "if" function to select the keys we want to keep
  - Create a temporary tuple to store values selected by the keys
  - Use cursor.execute() and connection.commit() functions to import the tuple into users table in PostgreSQL
- Create indexes for certain columns such as id and numbers of followers
- Test the ability to search data on Jupyter Notebook



```
There are 3 querying results

NO.1 user
id : 1115874631
name : CGTN
screen_name : CGTNOfficial
protected : False
verified : True
followers_count : 14024195
friends_count : 56
listed_count : 8412
favourites_count : 68
statuses_count : 117674
created_at : Thu Jan 24 03:18:59 +0000 2013

NO.2 user
id : 37034483
name : NDTV
screen_name : ndtv
protected : False
verified : True
followers_count : 12616711
friends_count : 15
listed_count : 12732
favourites_count : 0
statuses_count : 704225
created_at : Fri May 01 20:34:48 +0000 2009

NO.3 user
id : 16676396
name : El Universal
screen_name : El_Universal_Mx
protected : False
verified : True
followers_count : 5512589
friends_count : 14014
listed_count : 25234
favourites_count : 29497
statuses_count : 864128
created_at : Fri Oct 10 00:09:06 +0000 2008
```

Indexing and Sample Queries

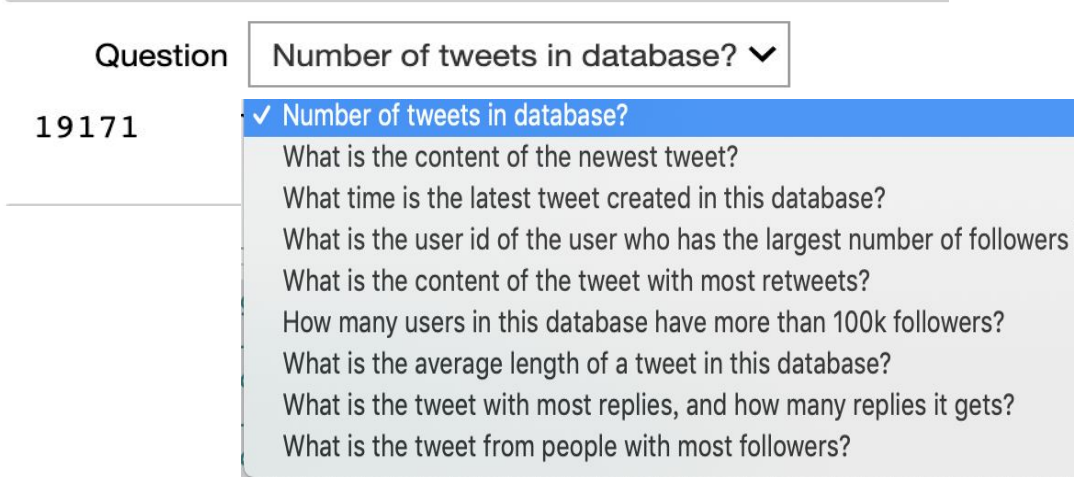
- Create indexes for fast access
  - 1) Number of their followers
  - 2) Created time
  - 3) Number of retweets
  - 4) Number of replies
- Create some sample queries
  - 1) Total number of tweets: 19171
  - 2) Most recent tweet time
  - 3) user id with largest # followers
  - 4) some more...

```
Wed Apr 15 00:56:34 +0000 2020
0.0024022199995670235
```

Caching and Search Applications

- **Search application UI**

- **Store some answers as cache**



Caching and Search Applications

- Cost of time to query total number of tweets(from cache or not)

```
19171
0.0001838299995142733 7
```

```
19171
3.261247906999415
```

- Cost of time to query the content of newest tweet (from cache or not)

```
RT @TarekFatah: Pakistanis in Karachi defying orders not to congregate i
mosques by creating makeshift mosques on rooftops. Working hard t…
5.2732999392901 55e-05
```

```
RT @TarekFatah: Pakistanis in Karachi defying orders not to congregate in
mosques by creating makeshift mosques on rooftops. Working hard t…
0.011365051000211679
```

Optimize Search Applications

- **New user interface**

- **LRU Cache**

  (red line)


```
No. 1
Wed Apr 15 00:56:34 +0000 2020 User ID: 22091137 User Name: Bas
u Ghosh Das
RT @TarekFatah: Pakistanis in Karachi defying orders not to con
gregate in mosques by creating makeshift mosques on rooftops. W
orking hard t…

{('Find Newest Tweets', 1): <__main__.LinkedNode object at 0x11
a461278>}
No. 1
Wed Apr 15 00:56:34 +0000 2020 User ID: 22091137 User Name: Bas
u Ghosh Das
RT @TarekFatah: Pakistanis in Karachi defying orders not to con
gregate in mosques by creating makeshift mosques on rooftops. W
orking hard t…
```

Optimize Search Applications

- My LRU cache

```
SearchApplicationOne()
```

Question: Find Newest Tweets

Top: 21

```
Wed Apr 15 00:56:32 +0000 2020 User ID: 619554146 User Name: MoeT
RT @TarekFatah: Pakistanis in Karachi defying orders not to congregate in mosques by creating makeshift mosques on
rooftops. Working hard t…

{('Find Newest Tweets', 2): <__main__.LinkedNode object at 0x11a461278>, ('Find Newest Tweets', 3): <__main__.Linke
dNode object at 0x10dcd9048>, ('Find Newest Tweets', 4): <__main__.LinkedNode object at 0x10e341e10>, ('Find Newest
Tweets', 5): <__main__.LinkedNode object at 0x10d54cda0>, ('Find Newest Tweets', 6): <__main__.LinkedNode object at
0x10e2b8518>, ('Find Newest Tweets', 7): <__main__.LinkedNode object at 0x10e2b8160>, ('Find Newest Tweets', 8): <_
_main__.LinkedNode object at 0x10e27a0f0>, ('Find Newest Tweets', 9): <__main__.LinkedNode object at 0x11a23c128>,
('Find Newest Tweets', 10): <__main__.LinkedNode object at 0x10e27c438>, ('Find Newest Tweets', 11): <__main__.Link
edNode object at 0x10e274f28>, ('Find Newest Tweets', 12): <__main__.LinkedNode object at 0x10e266fd0>, ('Find Newe
st Tweets', 13): <__main__.LinkedNode object at 0x10e27bfd0>, ('Find Newest Tweets', 14): <__main__.LinkedNode obje
ct at 0x10e277fd0>, ('Find Newest Tweets', 15): <__main__.LinkedNode object at 0x10e291fd0>, ('Find Newest Tweets',
16): <__main__.LinkedNode object at 0x10e36afd0>, ('Find Newest Tweets', 17): <__main__.LinkedNode object at 0x10e3
33fd0>, ('Find Newest Tweets', 18): <__main__.LinkedNode object at 0x10e362fd0>, ('Find Newest Tweets', 19): <__mai
n__.LinkedNode object at 0x10e34dfd0>, ('Find Newest Tweets', 20): <__main__.LinkedNode object at 0x10e347fd0>, ('F
ind Newest Tweets', 21): <__main__.LinkedNode object at 0x10e334dd8>}
```

Possible Improvements for the Future

- Making our dataset dynamic would allow us to continuously collect new tweets
  - Since our topic is so prevalent in the news, this would allow our application keep up to date with what is going on
  - However, this would require us to consider how to scale the database
    - Luckily, MongoDB supports scaling through its use of shards (could cost money for servers)
- Modifying cache to also consider popularity of tweets
  - keep newer AND more prevalent tweets readily available
  - likely to be accessed more
- Language filter
  - detect different languages and mark the tweets as such

Conclusions

- Some things we learned:
  - How to collect and analyze real data from a source that most of us use everyday
  - Even when dealing with one type of data, tweets, you can usual multiple, different kinds of databases to store it
  - The importance of indexing and caching in order to speed up queries, which is even more important for larger databases
  - Some domain knowledge when it comes to working with Twitter and designing our own search application
    - Got to see how tweets are stored and the various attributes that come with each tweet