pre-presentation: completed

MSDS 694: SPEC TOP DATA SCI – FINAL-PROJECT

Twitter Dataset Application

Team 6

Jinwoo Lee

Jongsoo Han

Kunal Jangam

Heon Park

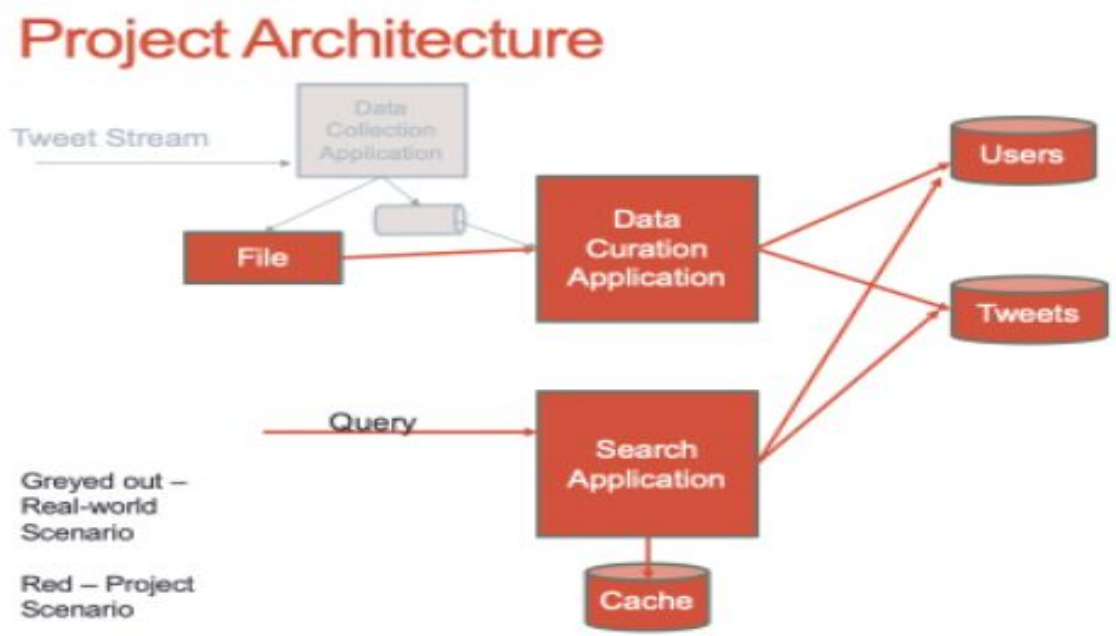
Table of Contents

Dataset

- Twitter Dataset
  - Data in Json Format
  - Over 90,000 tweets
  - More than 20 fields

Persisted Data Model and Datastores

- Relational Database: PostgreSQL
  - User Information
- Non-Relational Database: MongoDB
  - Tweet Information



Design: User Table

- Screen name - primary key
- Ranking by follower count

	screen_name [PK] character varying (100)	user_id bigint	followers_count integer	tweet_ids bigint[]
1	nuffsaidny	16144221	50	{1249378751349231616}
2	umesh_agr	268218622	24	{1249326224964345857}
3	meysimek	1193535233242664960	4	{1249403114614075400}
4	biannagolodryga	14135350	6343	{1249316363681910784}
5	ani_royal007	3917836273	74	{1248907321947783170}

- Store list of tweet_ids user made and user retweeted
  - Compare tweet's user_id with this user's id to determine if it was tweeted or retweeted

Design: Tweet Database

- id - primary key
- text
- created_at time
- hashtags
- user_id
- Ranking by interactions: favorite_count + retweet_count
- Store list of retweet objects: who retweeted when?

Tweet Database Improving Response Times

- **id index**
  - internal search when inserting
- **created_at index**
  - searching by time range
- **user_id index**
  - searching by user
- **Sorting text field**
  - Stored in additional column text_sorted
  - List of lower case words, remove punctuation and duplicates

Using binary search

- **PROS**
  - Faster search: O(nlogn) instead of O(n^2)
- **CONS**
  - Need to store an extra data column
  - Slower insert: text needs to be sorted
  - Might not be optimal

Search Application Design

- Search by
  - String, user, hashtag, time range
  - Order the tweets by interactions
- Drill-down search features
  - Tweet Text
  - Meta Data: (Author, When, How many times retweeted)

Search By String ¶

```
%%time
search_by_text()
```

```
Enter a word you want to search:
covid
Searching for covid ...
CPU times: user 750 ms, sys: 76.4 ms, total: 827 ms
Wall time: 1.54 s


[{'Tweet Text': 'even though kitorg positive covid  , kitorg
tion ki… https://t.co/vSoHG6DwQi',
  'Author of Tweet': 'errikaaaaaaaa',
  'When Created:': '2020-03-27T13:45:29',
  'Retweet Count': 30880},
 {'Tweet Text': 'I want our youth to play their role in helpi
ce whi… https://t.co/9dzONjCGan',
  'Author of Tweet': 'ImranKhanPTI',
  'When Created:': '2020-04-01T10:31:47',
  'Retweet Count': 5940},
```

Search By User ¶

```
%%time
search_by_user("B_King69")
```

```
CPU times: user 1.87 ms, sys: 1.36 ms, total: 3.
Wall time: 14.5 ms


[{'Tweet Text': 'É isto, ou vou morrer sem ar ou
  'Author of Tweet': 'B_King69',
  'When Created:': '2020-04-25T12:21:41',
  'Retweet Count': 0}]
```

Search By HashTag

```
%%time
search_by_ht()
```

```
Enter a word you want to search:
covid
Searching for covid ...
CPU times: user 713 ms, sys: 18.1 ms, total: 731 ms
Wall time: 1.26 s


[{'Tweet Text': 'खुद्दार मेरे शहर का, आज भूख से मर गया.\n\nराशन
co/8SLuO76Vv9',
  'Author of Tweet': 'upcoprahul',
  'When Created:': '2020-04-11T18:38:01',
  'Retweet Count': 331},
 {'Tweet Text': 'Popular Front of India Corona Relief
D… https://t.co/hlHzYz5esx',
  'Author of Tweet': 'AnisPFI',
  'When Created:': '2020-04-24T14:25:26',
  'Retweet Count': 106},
```

Search Application Design

Search By Time Range

```
t0 = time.time()
print(time_range_search())
t1 = time.time()
total = t1-t0
print(f"Time spent :{total}")
```

```
Enter a start date of your search(in the format of YEAR-MONTH-DAY ex)2020-04-12): You can press enter to not set the start dat
e.
2020-04-01
Enter a start date time of your search (in the format of HOUR:MINUTE:SECOND ex)18:26:00):
00:00:00
Enter a end date of your search (in the format of YEAR-MONTH-DAY ex)2020-04-12): You can press enter to not set the start dat
e.
2020-04-05
Enter a end date time of of your search (in the format of HOUR:MINUTE:SECOND ex)18:26:00):
00:00:00
Searching from 2020-04-01T00:00:00 to 2020-04-05T00:00:00

{'Tweet Text': 'My daddy beat the corona virus! Pour one for him tonight.', 'Author of Tweet': 'HeartThrobNever', 'When Create
d:': '2020-04-03T23:46:44', 'Retweet Count': 10883}

{'Tweet Text': 'When Corona Virus  is over, let's spend our holidays in India, eat in local restaurants, buy local meats and v
eggie… https://t.co/6GOxNRAFuy', 'Author of Tweet': 'SirPareshRawal', 'When Created:': '2020-04-03T09:09:11', 'Retweet Count': 2112
5}

{'Tweet Text': 'Listen to how she says corona virus Lmfaooo kids so slow 😴😴😴 https://t.co/ip0SsKPCJm', 'Author of Tweet': 'WoahDe
reQuise', 'When Created:': '2020-04-02T02:55:47', 'Retweet Count': 22335}
```

Search Application Design

- Drill-down search features
  - Show who retweeted the current tweet
  - Show other tweets by the author

Search By Who Retweeted

```
%%time
retweet_info(1238612571193827335)
```

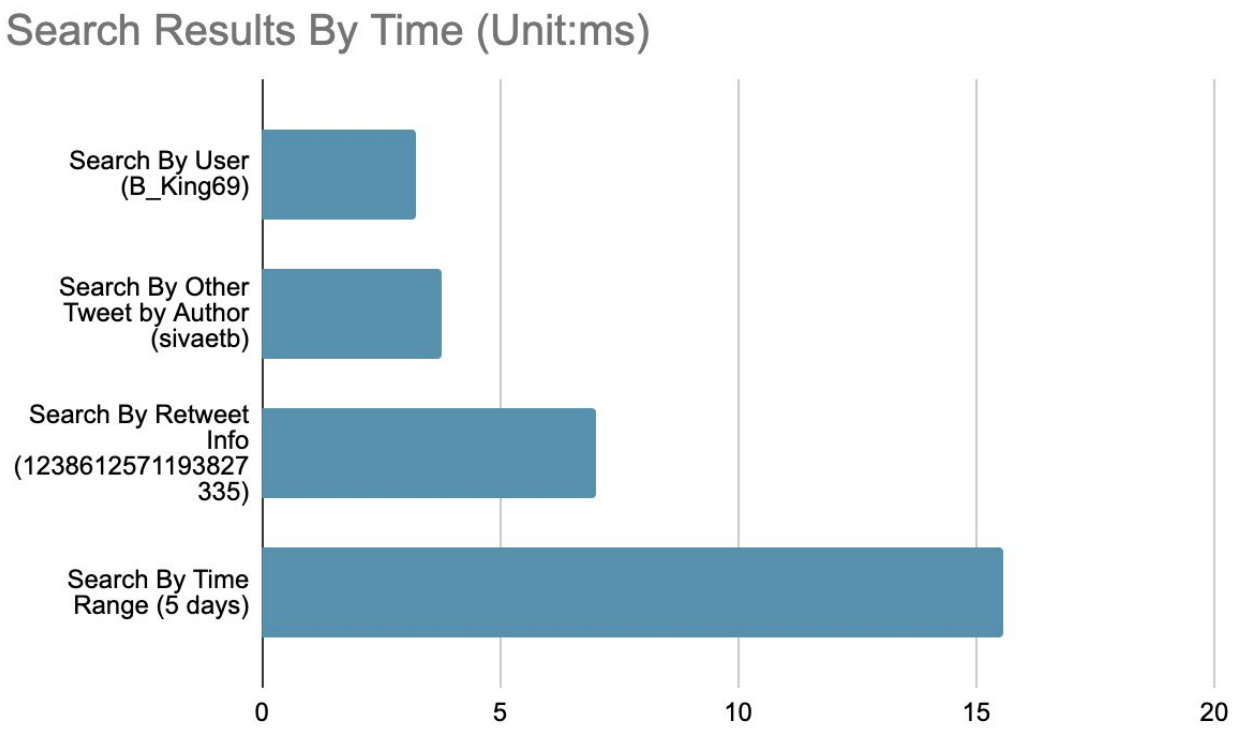
```
Original Tweet  If I gave you 100 skittles and told yo
les
The number of retweet : 3
The Retweet info about this tweet is following
CPU times: user 1.88 ms, sys: 2.57 ms, total: 4.45 ms
Wall time: 4.03 ms


[{'User who retweeted this tweet': 'Dalton642',
  'retweeted_at': '2020-04-12T18:43:16'},
 {'User who retweeted this tweet': 'CartoonGirl35',
  'retweeted_at': '2020-04-12T18:44:19'},
 {'User who retweeted this tweet': 'spooney35',
  'retweeted_at': '2020-04-12T18:45:47'}]
```

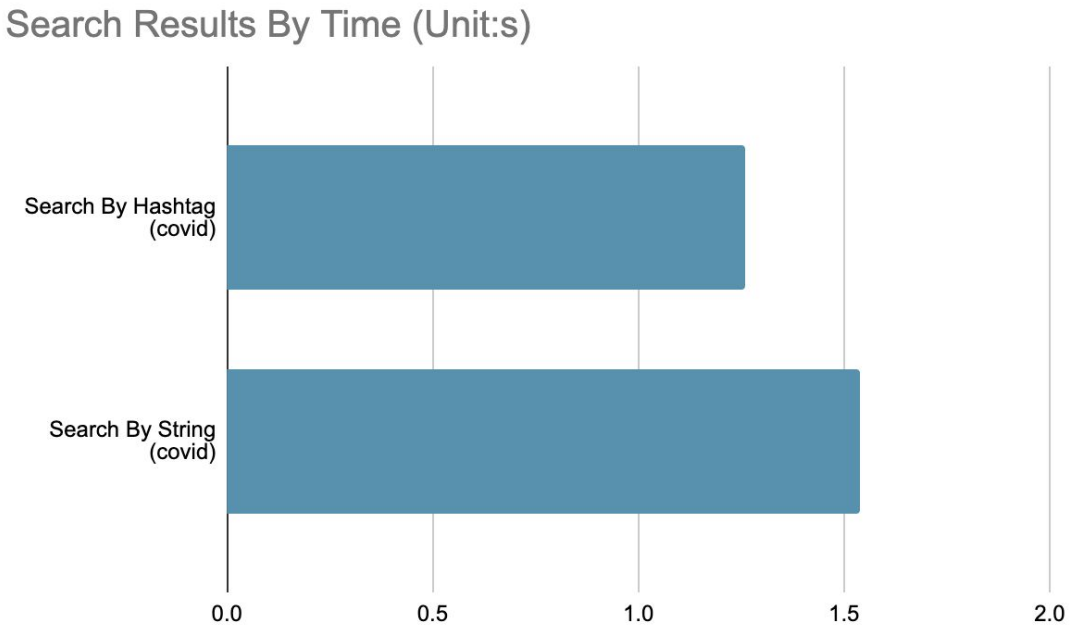
Other Tweet by Author

```
%%time
# input : author, current tweet_id
other_tweet_by_author("sivaetb",1254022804346777601)
```

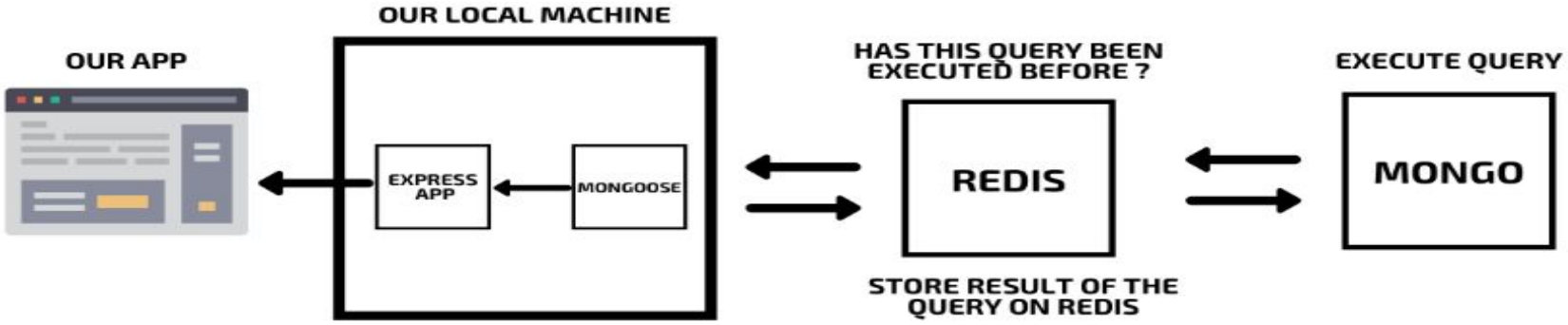
```
CPU times: user 1.92 ms, sys: 1.34 ms, total: 3.26 ms
Wall time: 3.77 ms


[{'Tweet Text': 'Health Minister @Vijayabaskarofl admits containing
and… https://t.co/gSCNnGhniD',
  'Author of Tweet': 'sivaetb',
  'When Created:': '2020-04-25T12:49:53',
  'Retweet Count': 0},
 {'Tweet Text': 'The number of #Covid_19 testing centres in the stat
l… https://t.co/VMOi122Q5u',
  'Author of Tweet': 'sivaetb',
  'When Created:': '2020-04-25T12:45:34',
  'Retweet Count': 0},
 {'Tweet Text': 'A total of 7,707 samples have been tested on Saturd
ients… https://t.co/NZRjVVTxLm',
  'Author of Tweet': 'sivaetb',
  'When Created:': '2020-04-25T12:42:08',
  'Retweet Count': 0},
```

Search Application Results





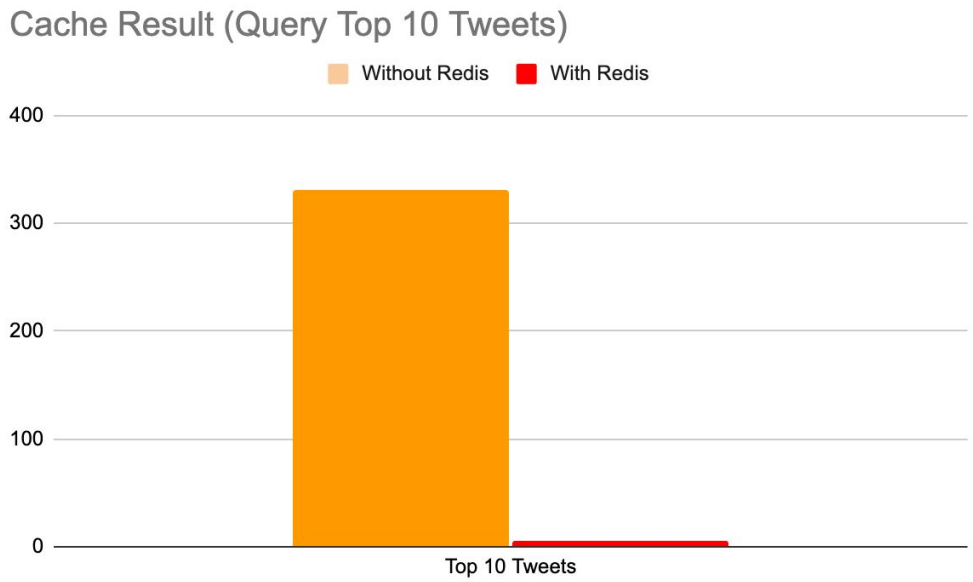
Caching Data with Redis

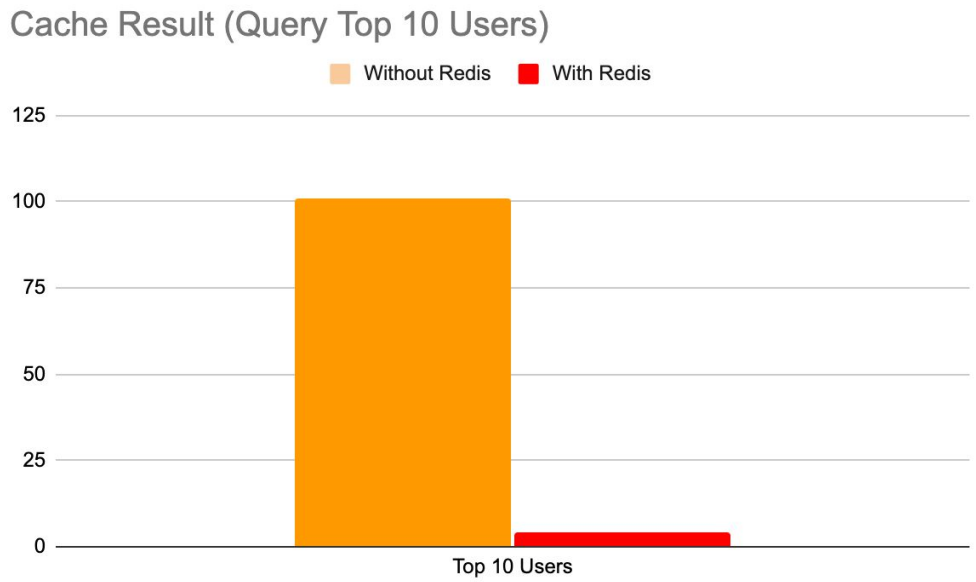


- Top Users
  - Based on their followers count
- Top Tweets
  - Based on the interactions values that our team devised
  - Interactions = (favorite_count + retweet_count)

Cache Result

- Query Top 10 Users and Tweets (Unit : ms)





Cache Result

- Query Top 10 Users and Tweets

```
# get the data from redis
res = redis_get_top10_user()
res
```

```
['detikcom',
 'jk_rowling',
 'aajtak',
 'ABPNews',
 'TimesNow',
 'sardesairajdeep',
 'lemondefr',
 'tvOneNews',
 'kompascom',
 'ntv']
```

```
# get the data from redis
res = redis_get_top10_tweets()
res
```

```
['ALERT!!!!!\nThe corona virus can be spread through money. If you have any money at home, put on some gloves, put al… https://t.co/juJjDpFN3l',
 '*corona virus enters my body*\n\nThe 4 Flintstone gummies I ate in 2005: https://t.co/3STfdlQtaT',
 'When this Corona shit passes we have to promise each other that we're going to tell our kids that we survived a zombie apocalypse in 2020',
 'If I gave you 100 skittles and told you 3 of them could kill you.... I'm sure you would avoid the fucking skittles',
 'THIS MAN IS A GENIUS he figured out the Corona virus problem 😲 https://t.co/EZP7lqTtxV',
 'It wasn't no corona till y'all started balancing brooms in the house, y'all let the devil in',
 'yeah it's a new generation we gon call them quaranteens https://t.co/rJSKmIf7Qp',
 '"corona time "😭😭😭😭 https://t.co/iXBMHVcFoY',
 'Watch this. It shows why we should all do the right thing and stay home to the fullest extent possible. All of us c… https://t.co/GOODRTNI2e',
 'Nobody: \nMe and the homies on our $41 corona trip to the Bahamas: https://t.co/UOoRRJLiGr']
```

THANK YOU