



# Database Management Systems

## Twitter Search Application

Final Project

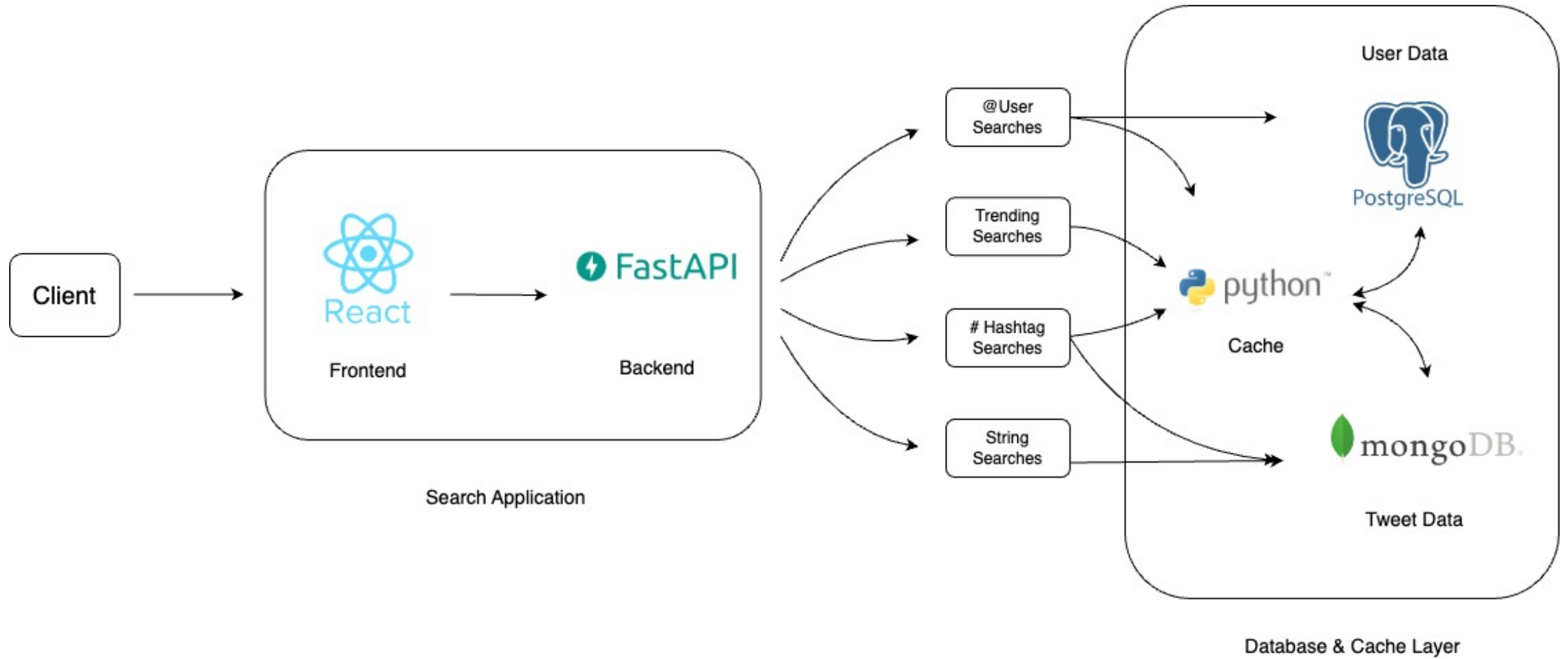
Team 31  
- Sasank Chindirala

# Data Curation

- The data used for our application was sourced from the corona-out-2 and corona-out-3 datasets provided.
- Each file was read line-by-line, parsed and processed before storing the needed user and tweet object models in the database.
- The individual line was itself a nested json object, that had important information inside the nested key's like “retweeted\_status” and “quoted\_status”




```
{
  "created_at": "Sun Apr 12 18:27:25 +0000 2020",
  "id": 1249403767180668930,
  "id_str": "1249403767180668930",
  "text": "RT @nuffsaidny: wishing death on people is weirdo behavior.",
  "source": "\u003ca href=\"http://twitter.com/download/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c/a\u003e",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "retweeted_status": {
      "quoted_status_id": 1249315454797168641,
      "quoted_status_id_str": "1249315454797168641",
      "quoted_status": {
        "quoted_status_permalink": {
          "is_quote_status": true,
          "quote_count": 0,
          "reply_count": 0,
          "retweet_count": 0,
          "favorite_count": 0,
          "entities": {
            "favorited": false,
            "retweeted": false,
            "filter_level": "low",
            "lang": "en",
            "timestamp_ms": "1586716045552"
          }
        }
      }
    }
  }
}
```

# System Architecture



# User Data

- An index on screen\_name was created for efficient searches
- Another multi-column index was created for faster sorting of the search results.
- Username based searches were ranked on the basis of the number of followers and the number of tweets posted.

total_rows 	table_size 	column_count 
bigint	text	bigint
108043	35 MB	12

Column	Type	Collation	Nullable	Default	Storage
id	character varying(255)		not null		extended
name	character varying(255)				extended
screen_name	character varying(255)				extended
verified	boolean				plain
location	character varying(255)				extended
description	character varying(255)				extended
followers_count	bigint				plain
friends_count	bigint				plain
favourites_count	bigint				plain
statuses_count	bigint				plain
tweets_count	bigint				plain
created_at	timestamp without time zone				plain

# Tweet Data

- An index on the text field was created for efficient searches.
- A field of tweet\_score was created for ranking the search results that assigns a weighted score to each tweet based on the number of likes and retweets that a particular tweet had.

## tweets

**Storage size:** 30.61 MB

**Documents:** 134 K

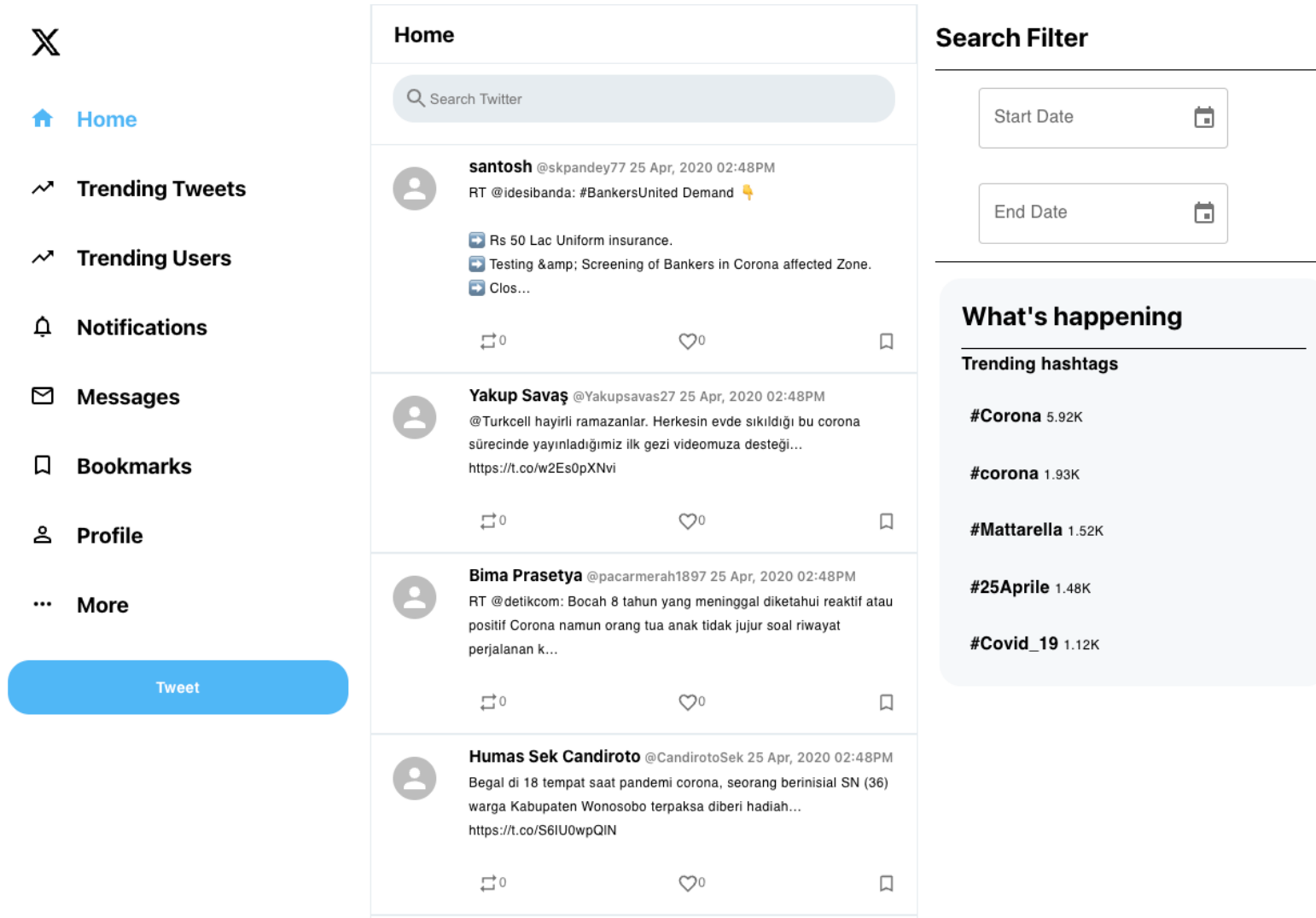
**Avg. document size:** 439.00 B

**Indexes:** 2

**Total index size:** 31.57 MB

```
_id: ObjectId('662964076cc8b70cf9abe575')
tweet_id: "1249402922309423107"
text: "In Turkey, there are 300 thousand prisoners and 150 thousand prison em..."
▶ hashtag: Array (empty)
user_id: "1055885344736993280"
user_name: "no Comment"
user_screen_name: "lastcavalry61"
likes_count: 5
retweet_count: 21
source_tweet_id: 0
tweet_score: 14.6
created_at: "2020-04-12 18:24:04"
```

# Application User Interface



# String Search

X

Home

Trending Tweets

Trending Users

Notifications

Messages

Bookmarks

Profile

More

Tweet

Home

doctors

Pawan Kalyan @PawanKalyan 22 Mar, 2020 11:40AM

We salute to all the Doctors, Nurses, health workers, sanitary workers, media and police for fighting against coron... <https://t.co/gb6eQWglNp>

1

76808

Chiranjeevi Konidela @KChiruTweets 20 Apr, 2020 04:11PM

We can never thank them enough, all our doctors, health workers, police, sanitation workers and media, our frontli... <https://t.co/HR4KAeh76s>

1

11289

Pawan Kalyan @PawanKalyan 21 Apr, 2020 09:54AM

Let's respect and honour our Doctors and medical staff.. Just convey your heartfelt gratitude to all the Doctors, N... <https://t.co/c5e8MECt51>

1

10334

Cyberabad Police @cyberabadpolice 18 Apr, 2020 10:24AM

Appreciations to Doctors, Police, Sanitation workers and Media who are front line fighters against Corona Virus. su... <https://t.co/CNmkw7HcBr>

Search Filter

Start Date

End Date

What's happening

Trending hashtags

#Corona 5.92K

#corona 1.93K

#Mattarella 1.52K



#25Aprile 1.48K

#Covid\_19 1.12K

# @Username Search


Home

@purohit

PINKY RAJPUROHIT (ABP NEWS)   @MADRASSAN\_PINKY

DR.SANJEEV RAJPUROHIT@DRSANJEEVRAJP4

NALINI\_PUROHIT@NALINI51PUROHIT

DR DHIMANT PUROHIT  @DHIMANTPUROHIT

APARNAA PUROHIT@APARNAAPUROHIT

Search

Star

End

What

Trendin

#Coroi

#coror



# @Username Search – Tweet Drilldown

X

Home

Trending Tweets

Trending Users

Notifications

Messages

Bookmarks

Home

@purohit

Pinky Rajpurohit (ABP News) 🇮🇳 @Madrassan\_Pinky

24 Apr, 2020 10:27AM

Tiruppur Police catches #COVIDIOTS who don't follow lockdown strictly and lock them up with a Covid19 patient in an...  
<https://t.co/8UGvBf4FJE>

1

86

Search Filter

Start Date


End Date


What's happening


Trending hashtags


#Corona 5.92K


# #Hashtag Search





 Home


 Trending Tweets


 Trending Users

 Notifications

 Messages


 Bookmarks


 Profile


 More


Tweet


Home


 #corona


**Frank Figliuzzi** @FrankFigliuzzi13 Apr, 2020 02:10PM  
Mob boss asks his crew if that guy who likes science is a snitch.  
#fauci #corona


 1


 9217





**Chad Ellsworth** @chad\_ellsworth19 Apr, 2020 06:53PM  
Happening Now in San Diego in response to closing walking trails and beaches. #COVID19 #corona #encinitas #sandiego  
<https://t.co/ukmyMbx2yq>


 1


 8097




**Abdullah T.R #coronavirus** @TraderAT1222 Feb, 2020 04:09PM  
#coronavirus #corona  
#koronavirus #korona  
♀ GENLER ve IRK/VİRÜS♂  
♂ VIRUS ve ÇALANLAR♂  
👉 HIV/CCRS 'suz bebek👉  
Da... <https://t.co/pQqvJmaAyx>


 1

 6824



Search Filter

Start Date

End Date

What's happening

Trending hashtags

#Corona5.92K

#corona1.93K


#Mattarella1.52K

#25Aprile1.48K

#Covid\_191.12K

# Trending Search


- Trending Tweets




- Home
- Trending Tweets**
- Trending Users
- Notifications
- Messages
- Bookmarks
- Profile
- More

Tweet


Home

**Pre K** @stayfree\_ 04 Mar, 2020 05:31PM  
ALERT!!!!!!  
The corona virus can be spread through money. If you have any money at home, put on some gloves, put al... <https://t.co/juJjDpFN3l>


1128502

**gilbert** @HtownBabyG 13 Mar, 2020 12:43AM  
"corona virus enters my body"  
The 4 Flintstone gummies I ate in 2005: <https://t.co/3STfdlQtaT>

811062

**Dill** @Khydill 18 Mar, 2020 05:51PM  
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

764405

**Mr Dre** @MrDre\_ 13 Mar, 2020 11:47PM  
If I gave you 100 skittles and told you 3 of them could kill you.... I'm sure you would avoid the fucking skittles

598511

Search Filter

Start Date

End Date

What's happening

Trending hashtags

#Corona 5.92K


#corona 1.93K

#Mattarella 1.52K

#25Aprile 1.48K

#Covid\_19 1.12K

- Trending Users



- Home
- Trending Tweets
- Trending Users**
- Notifications
- Messages
- Bookmarks
- Profile
- More

Tweet

Home

**BARACK OBAMA** @BARACKOBAMA

**DONALD J. TRUMP** @REALDONALDTRUMP

**CNN BREAKING NEWS** @CNNBRK

**NARENDRA MODI** @NARENDRAMODI

**SHAKIRA** @SHAKIRA

Search

What's trending

#Co

#cor

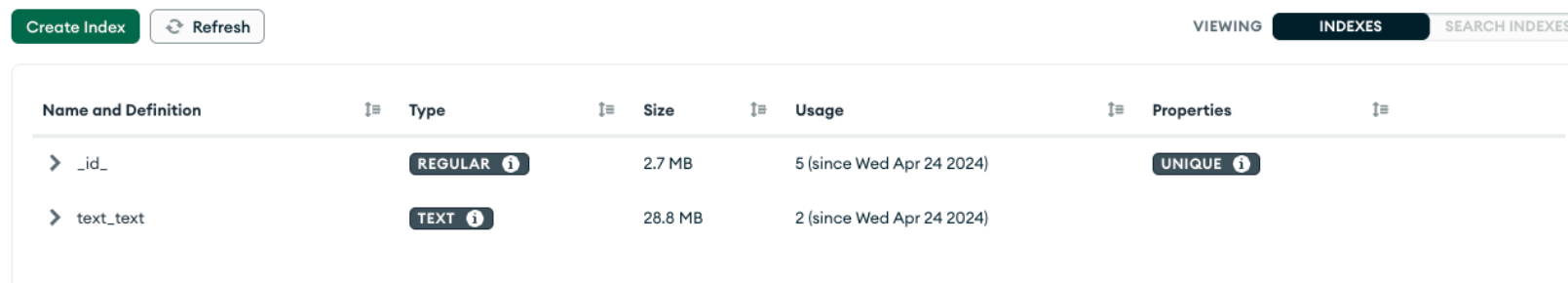
#Ma

#25/

#Co

# Performance Optimizations

- Ranking or search relevance metrics were directly stored in the database schema to avoid on the fly computations
- Indexing
  - MongoDB indexing was done on the text field.

A screenshot of the MongoDB Indexes management interface. At the top, there are buttons for 'Create Index' (green), 'Refresh' (light blue), and tabs for 'VIEWING', 'INDEXES' (selected), and 'SEARCH INDEXES'. Below is a table with columns: Name and Definition, Type, Size, Usage, and Properties. Two indexes are listed: '\_id\_' with type 'REGULAR' and size '2.7 MB', and 'text\_text' with type 'TEXT' and size '28.8 MB'.

Name and Definition	Type	Size	Usage	Properties
> _id_	REGULAR ⓘ	2.7 MB	5 (since Wed Apr 24 2024)	UNIQUE ⓘ
> text_text	TEXT ⓘ	28.8 MB	2 (since Wed Apr 24 2024)	

- PostgreSQL can't use a regular B-tree index on a pattern that starts with a wildcard (screen\_name like %s%). So used pg\_trgm extension to create a GIN (generalized inverted index) that can improve performance for **like** queries that use wildcard characters.

```
Indexes:
  "users_new_pkey" PRIMARY KEY, btree (id)
  "idx_users_ranking" btree (followers_count DESC, tweets_count DESC, verified DESC)
  "idx_users_screen_name" gin (screen_name gin_trgm_ops)
```

# Performance Optimizations (Contd.)

- Cache

- The cache class uses a **least- accessed eviction** strategy to ensure that frequently accessed keys are retained in the cache. This helps optimize the cache's memory and ensures that it can store a large amount of data within the defined limit of **10 MB**.
- The cache class includes a **checkpoint** interval of 5 minutes, which allows the cache to regularly save its contents to disk. This ensures that in case of any failures, the cache can recover its contents and continue serving data to users without disruptions.
- The cache class provides a **TTL** (time-to-live) feature that allows the cache to automatically remove stale data from the cache. With a TTL of one hour, any data that hasn't been accessed within a week will be automatically removed from the cache, ensuring that only the latest and relevant data is stored.

@Username Search

Without Cache

0.04187798500061035 seconds

With Cache

2.7179718017578125e-05 seconds

#Hashtag Search

Without Cache

0.07761096954345703 seconds

With Cache

2.7894973754882812e-05 seconds

Thank You!