

Team 15 

Twitter Search Application

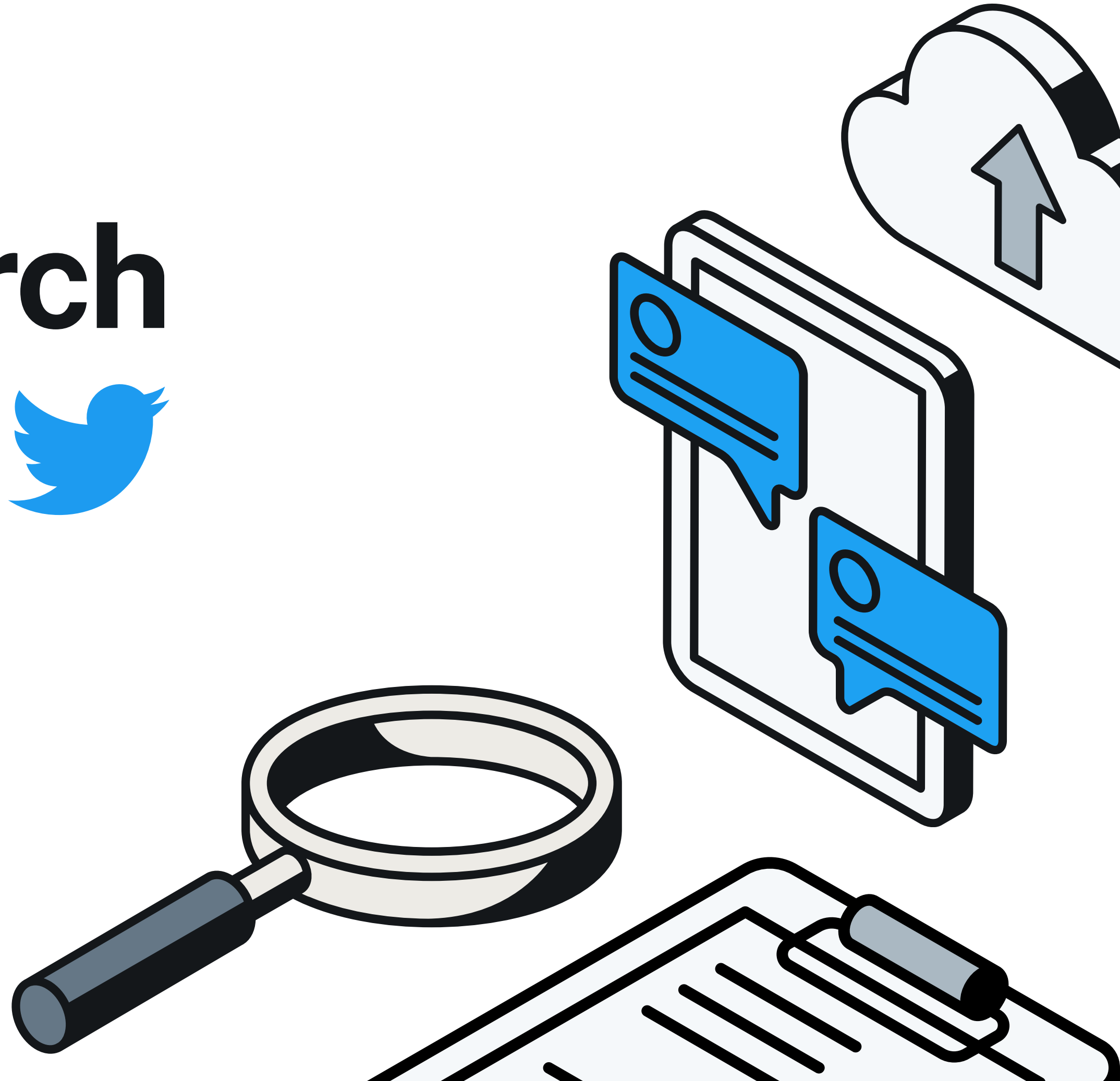
Team Members:

Hsiao-Chun Hung

Yuyue Sun

Yu Wang

Zhuoer Liu



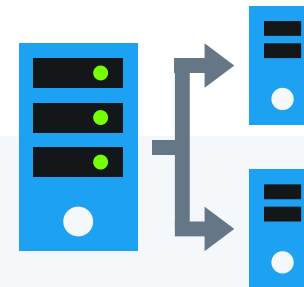
Intro & Methodology

This project aims to create a search application for a Twitter dataset using relational and non-relational databases. Users can search tweets by usernames, strings, hashtags. Additionally, users can access features such as viewing rankings of top users or top retweets.

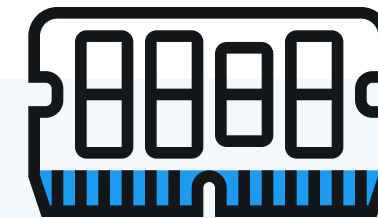


Exploratory Data Analysis

Study data trends and determine the best data loading structure

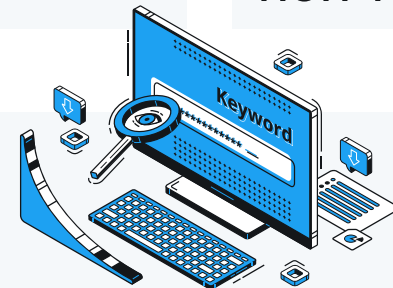


Data Loading



Caching

Cache frequently searched tweets for faster access



Search Application

Smart queries pulling from both cache and disk



Output & Frontend

Generates output on the frontend website

Dataset



corona-out-3

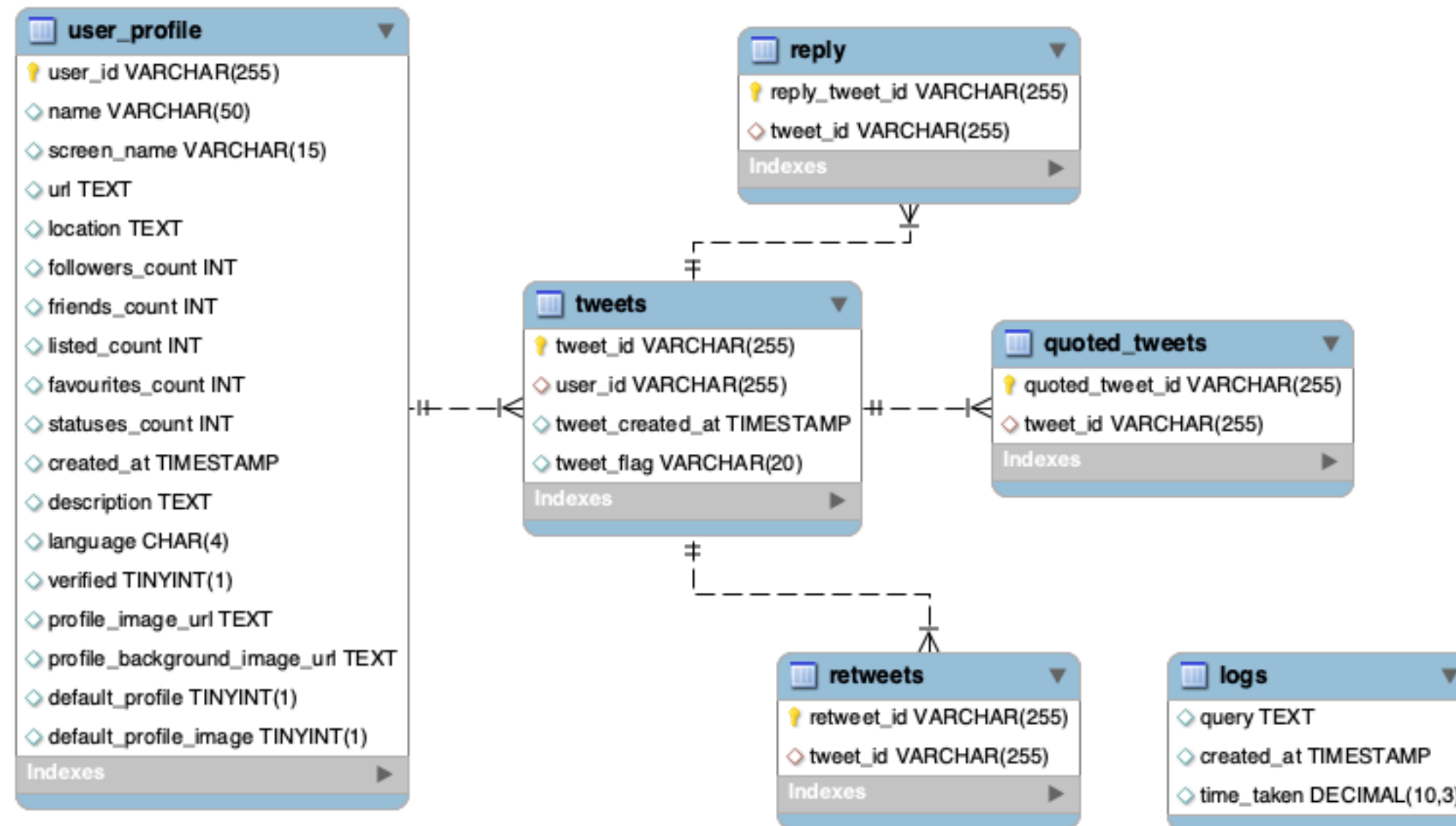
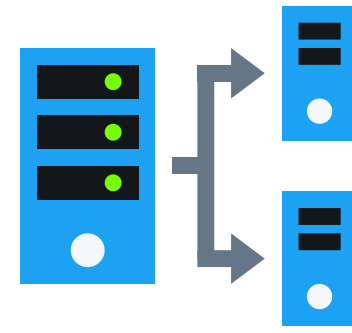
- Number of lines= 101916
- Number of tweets= 40804
- Number of retweets= 61112
- Number of unique tweets/retweets= 101894

Category	Fields
Basic Tweet Info	Created At, ID, ID Str, Text, Source, Language
User Information	User ID, ID Str, Screen Name, Followers Count, Friends Count
Engagement Metrics	Quote Count, Reply Count, Retweet Count, Favorite Count
Additional Metadata	Geo, Coordinates, Place, Entities in Retweet
Retweeted Status	Original Text, Truncated, Extended Tweet, Original User, URLs



Data Load

Relational



Relational database: tables

user_profile

tweets

quoted_tweets

retweets

reply

Indexing:

on 'tweet_id' field of 'tweets' table

special characters:

charset = utf8mb4

collation = utf8mb4_unicode_ci

Data Load

Non-Relational



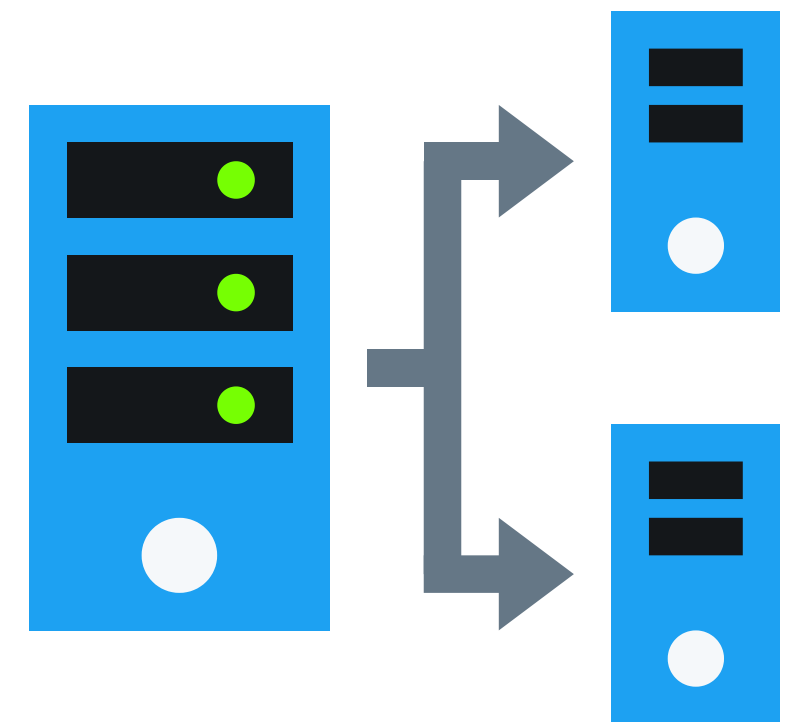
Data Processing

Extracting necessary fields from a JSON line:

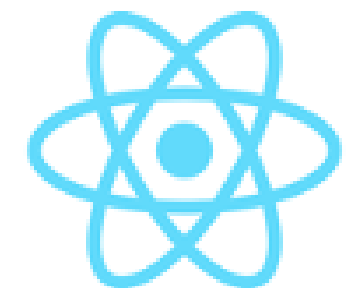
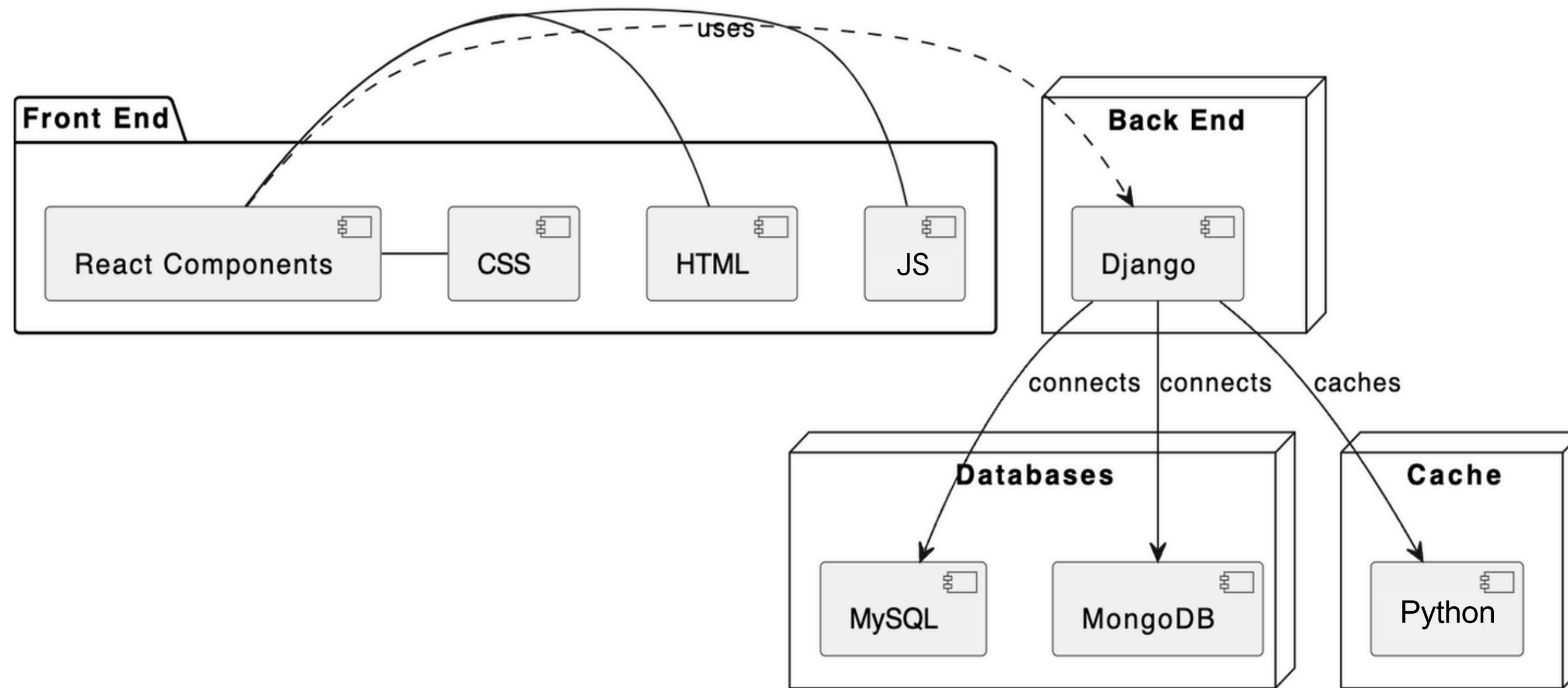
- Basic Fields: Directly transfer fields such as **text**, **source**, and **created_at**.
- Nested Fields: Extract nested values like **user_id** from `doc['user']['id_str']`.
- Complex Fields: Serialize dictionaries and nested data (e.g. **entities**, **retweeted_status**) into JSON strings to preserve structure and simplify storage.

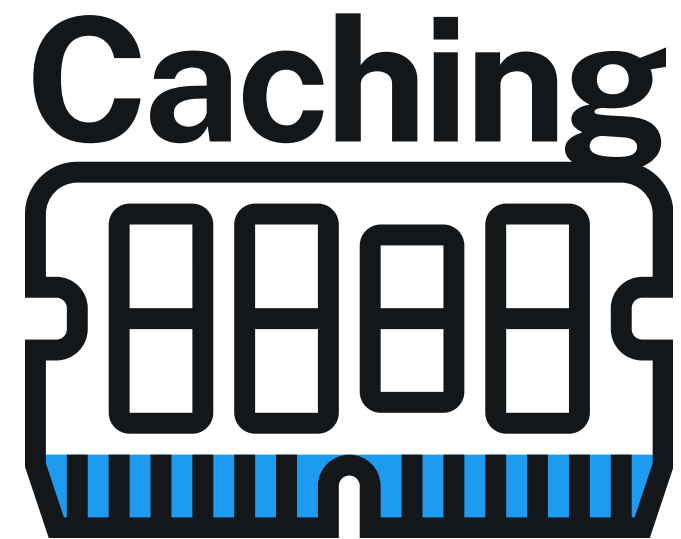
Indexing Strategy

- Unique Index on **id_str**: Establish a unique index on `id_str` to prevent duplicates and enhance query performance, as `id_str` is the unique identifier for each tweet.
- Upsert Operations: Use `update_one` with **upsert=True** to either update existing documents or insert new ones, optimizing data integrity and performance without manual checks.



Search Application





System Design:

Uses CacheManager with LRU caching (1024 results).
Preloads data from diskCache.json on start-up.



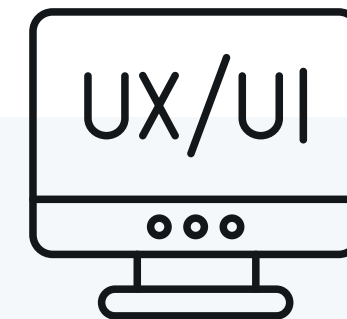
Query Processing:

Queries checked against cache for speed, prompting database retrieval on misses; cache managed with putQuery



Maintenance and Efficiency:

Entries removed via delQuery; regular cache snapshots ensure data integrity, and asynchronous saving maintains performance.



Enhanced User Experience:

Efficient cache management showcases quick responses and optimized resource utilization in web applications.

Caching

The CacheManager implemented in the Twitter search app substantially reduces latency by avoiding redundant database queries, thus providing rapid data access.

Results

	Username	Tweet String	Hashtags	Time (seconds)
Without Cache	Joseph Andrew	virus	corona	2.90
With Cache	Joseph Andrew	virus	corona	0.001

Twitter Search

User's Name

User's ScreenName

User Verification

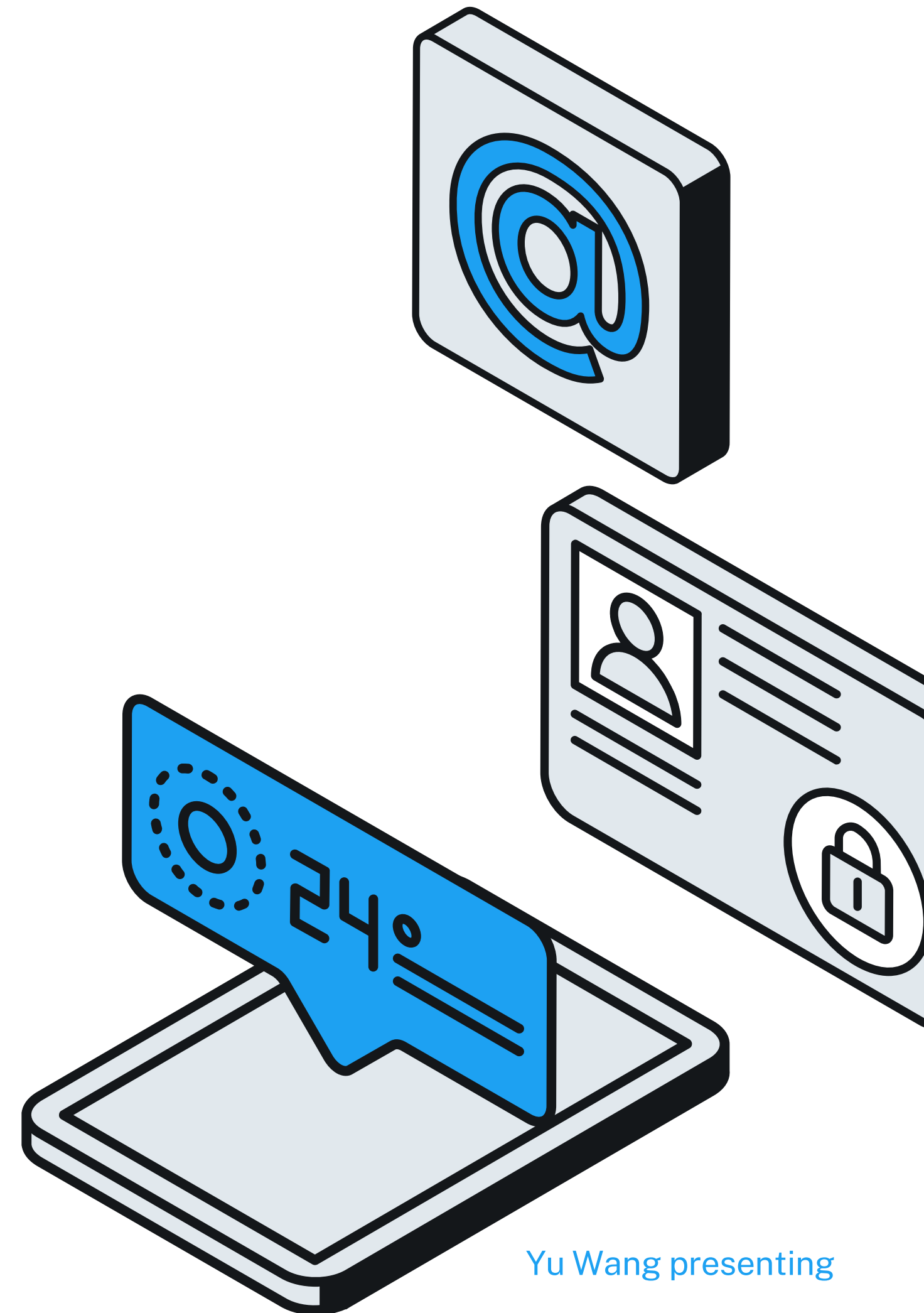
Tweet String

Hashtags

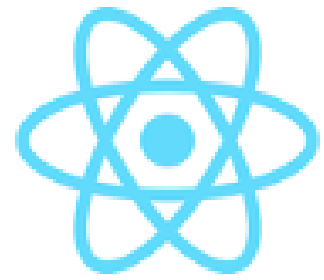
Tweet Sensitivity

Tweet Content Type

Date Time Range



UI



Log in



 Username

 Password



Login

Register



 Username

 Password

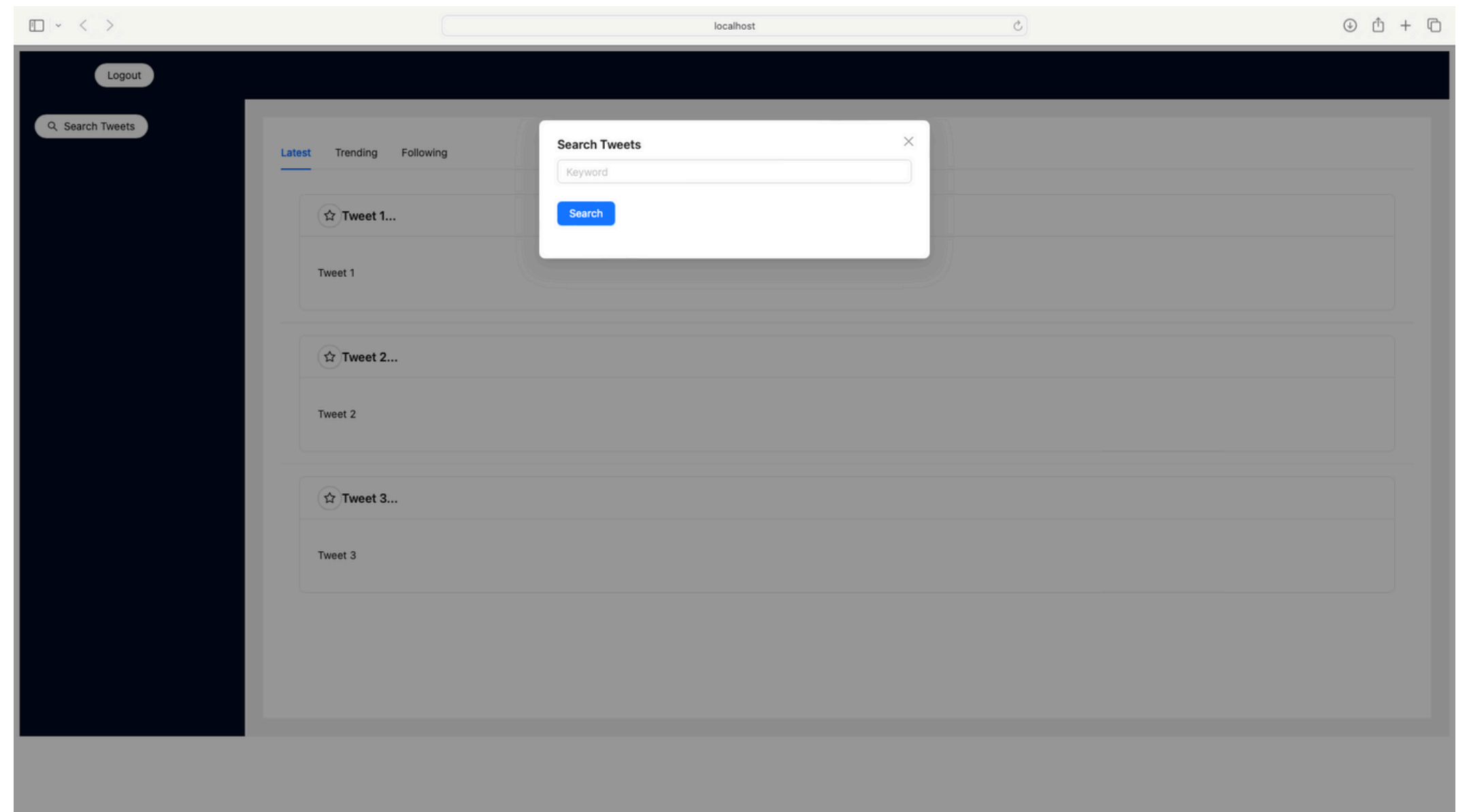
firstname

lastname

Register

Login

Register



OUTPUT

Search by names containing “ky”

X

Twitter Search Application

Search

Top-Level Metrics

Total Time taken for search to run= 16.475786924362183 seconds.

Show

10

entries

Search:

tweet_created_at	tweet_flag	name	screen_name	verified	retweet_count_y	quoted_count	reply_count_y
2020-04-25 13:55:23	original	Mike Lisansky	brynmont	0	64	0	0
2020-04-25 13:22:47	original	SKY QUIZON	skyquizon	0	23	0	0
2020-04-25 13:39:07	original	Kyle Whitmire	WarOnDumb	1	23	1	0
2020-04-25 13:11:31	original	pinky singh	pinkysi40475597	0	5	1	0
2020-04-25 13:05:50	original	Mr Malky	MrMalky	0	4	1	0
2020-04-25 13:11:34	original	Lou Lewinsky 2	lewinskylou2	0	4	1	0
2020-04-25 12:24:47	original	Nicky Monreau	NickyMonreau	0	2	0	1
2020-04-25 12:31:21	original	Helena Malikyar	HelenaMalikyar	0	2	0	0
2020-04-25 12:46:12	original	Nicky Bouwers	NickyBouwers	0	2	0	0
2020-04-25 12:32:54	original	Askyourbudget.com	askyourbudget	0	1	0	0

Showing 1 to 10 of 547 entries

Previous

1

2

3

4

5


...

55

Next

OUTPUT

Search by screen_name “sivaetb”

 Twitter Search Application

SearchTop-Level Metrics

Total Time taken for search to run= 1.539538860321045 seconds.

Show10entries

Search:

uncated	user_id_x	withheld_in_countries	tweet_id	user_id_y	tweet_created_at	tweet_flag	name	screen_name	verific
ue	730576596	None	1254022804346777601	730576596	2020-04-25 12:21:49	original	Sivapriyan E.T.B	sivaetb	1
ue	730576596	None	1254027211109101568	730576596	2020-04-25 12:39:20	original	Sivapriyan E.T.B	sivaetb	1
ue	730576596	None	1254027915869564928	730576596	2020-04-25 12:42:08	original	Sivapriyan E.T.B	sivaetb	1
ue	730576596	None	1254028781418803200	730576596	2020-04-25 12:45:34	original	Sivapriyan E.T.B	sivaetb	1
ue	730576596	None	1254029865579237376	730576596	2020-04-25 12:49:53	original	Sivapriyan E.T.B	sivaetb	1
lse	730576596	None	1254042958904492033	730576596	2020-04-25 13:41:55	retweeted	Sivapriyan E.T.B	sivaetb	1
lse	730576596	None	1254058330609250304	730576596	2020-04-25 14:42:59	retweeted	Sivapriyan E.T.B	sivaetb	1

Showing 1 to 7 of 7 entries

Previous1Next

OUTPUT

Top 10 tweets based on their retweet count

Select Metric

Top 10 Tweets by Retweet Count

Total Time taken for search to run= {1.7775928974151611} seconds.

Show

10

entries

Search:

tweet_count	_id	id_str	possibly_sensitive	text
86	662aea67676c16a8c2b2087c	1254030161403674624	False	Milwaukee's health commissioner has now tied 40 coronavirus infections to the
32	662ae9fe676c16a8c2b1f48a	1254028244166356998	False	Gözün çıksın corona😞 Ülkece asabii olduk 🤪muhtemel psikoloji ektedir 😂😂😂👉
00	662af1bc98079b56c6c980ba	1254045905252167681	False	A MUST READ...Coronavirus Restrictions: Government Bears the Burden of Proof
22	662ae9c7676c16a8c2b1e9dd	1254027175122153479	None	In Illinois, liberal politicians cut sweetheart deals with corrupt union bosses to ke
22	662aeedd98079b56c6c8f315	1254032746361417729	False	Thalapathy fans from Sivakasi Helped the Poor Family who are affected by this c
22	662af03398079b56c6c9343d	1254038838403493889	True	If that's how they are following lockdown/social distancing then Mumbai is sitting
85	662aea55676c16a8c2b20530	1254029837460799490	False	Süleyman Özışık durumu çok iyi açıklamış. \nNedir sizin bu bitmek bilmeyen düşr
03	662aefd598079b56c6c92231	1254037192952995840	False	A maioria dos cientistas acreditam que o coronavírus não veio da China e sim do
78	662ae927676c16a8c2b1cb83	1254024050889949186	None	Glückwunsch: In Sachsen-Anhalt wurden die Daten aller Corona-Infizierten oder
77	662aef5198079b56c6c908ce	1254034810231717889	False	Please, if you read one article today, read this. \nhttps://t.co/pZ7OSbu0V1

Showing 1 to 10 of 10 entries

Previous

1

Next

OUTPUT

Top 10 users based on their follower count

Select Metric

Top 10 Users by Follower Count

Total Time taken for search to run= {0.9565298557281494} seconds.

Show

10

entries

Search:

user_id	name	screen_name	url	location	followers_count	friends_count	listed_count	favourites_count
69183155	detikcom	detikcom	http://www.detik.com	Jakarta, Indonesia	15927642	28	13298	313
62513246	J.K. Rowling	jk_rowling	http://www.jkrowling.com	Scotland	14608046	721	37917	27353
42606652	AajTak	aajtak	http://www.aajtak.in	India	9706667	416	3517	16782
39240673	ABP News	ABPNews	http://abplive.com	India	9562582	248	3941	99
240649814	TIMES NOW	TimesNow	http://www.timesnownews.com	India	9499328	391	5109	4
56304605	Rajdeep Sardesai	sardesairajdeep	http://rajdeepsardesai.net/	New Delhi	8947342	568	8993	7598
24744541	Le Monde	lemondefr	https://www.lemonde.fr	Paris	8808784	628	36505	1609
55507370	tvOneNews	tvOneNews	http://tvonenews.com	Indonesia	8787649	50	8737	4347
23343960	Kompas.com	kompascom	http://www.kompas.com	Indonesia	7678373	23	10617	114
15016209	NTV	ntv	https://www.ntv.com.tr/	Turkey	7429223	29	5436	3

Showing 1 to 10 of 10 entries

Previous

1

Next



Team 15



@Team15DBMS

Thank you!

