

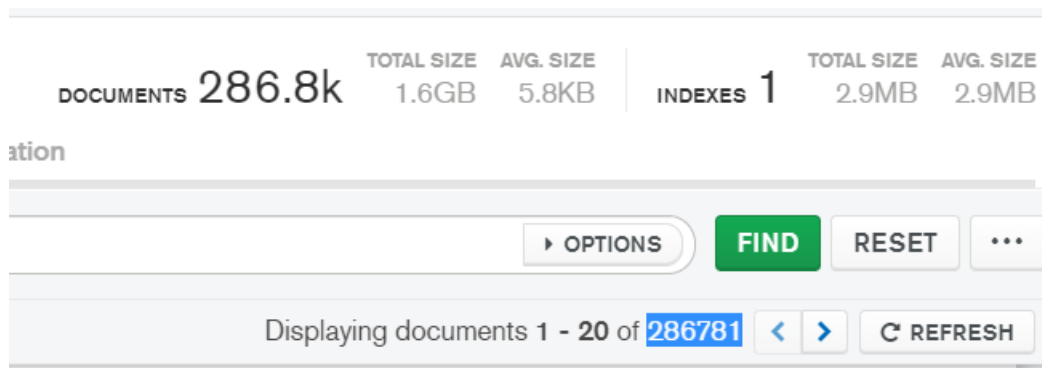
# **Semi-Structure   DataProcessing   With NoSQL Database Server**

## Technology Used –

- Pycharm
- Python 3.0 as an Interpreter
- Microsoft SQL
- MongoDB Compass

## Project Explanation

- 1) Tweets gathered – 286781
- 2) Total File Size – 1.6GB




- 3) Extra Credit Part is done, tweets directly went to MongoDB using Python to make the process automatic for big data processing.
- 4) Topic on which tweets are gathered - Presidential Election with Donald Trump and Joe Biden

## MONGODB COMPASS SET UP -


### MongoDB Community Server

#### Available Downloads


Version


4.4.1 (current) 

Platform

Windows 

Package

msi 

 **Download**

Copy Link

[Current releases & packages](#)  
[Development releases](#)  
[Archived releases](#)  
[Changelog](#)  
[Release Notes](#)

### Step 2-

#### New Connection FAVORITE

[Paste connection string](#)

Hostname

More Options

Hostname

localhost


Port

27017

SRV Record

☐

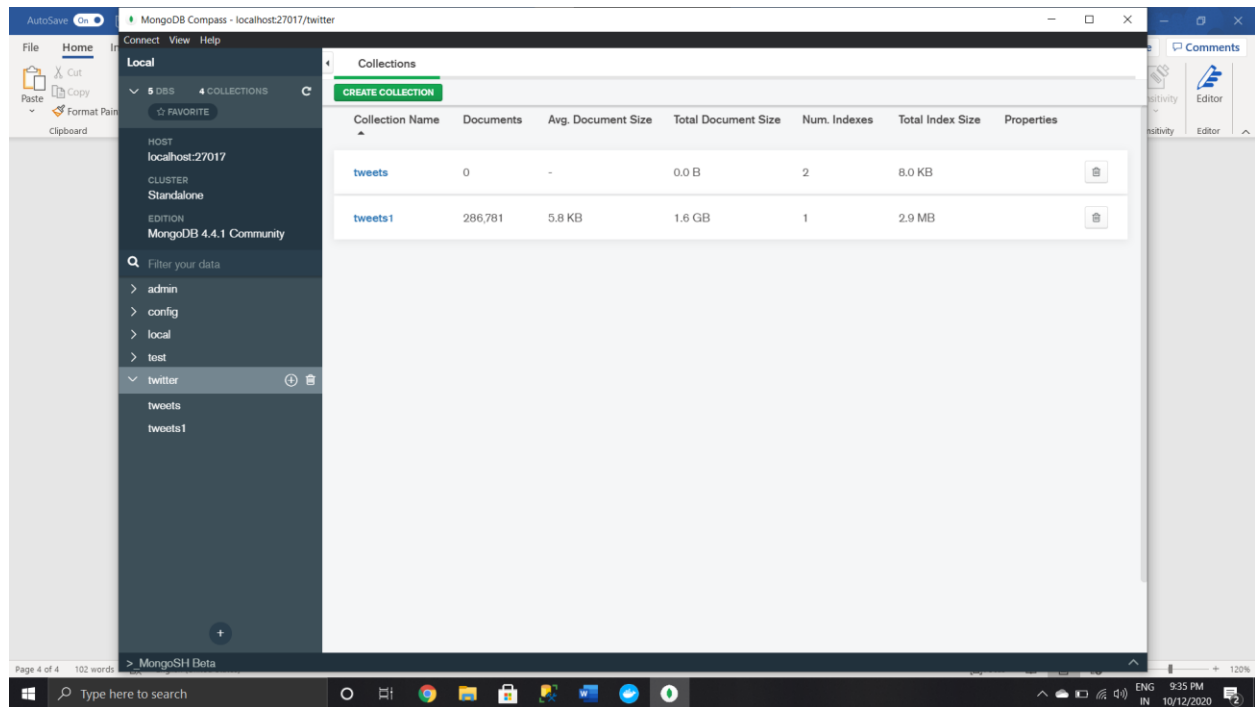
Authentication

None 

CONNECT

### Step 3 –

Created the noSql database which I'm going to use as an automatic big data processing.



## Code Description

```
lab3.py × lab3-1.py × lab3 new.py × cpl.py ×
import tweepy
import json
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import pymongo
from pymongo import MongoClient
```

I have imported this library

- 1) Pymongo – with the help of this library, I connected my code with the mongodb compass
- 2) Tweepy – with the help of this library, I gathered tweets.
- 3) Json – to load and dump all the tweets.

```
access_token = '124237897407010000'
access_token_secret = 'b0fFRRUMDXv'
consumer_token = 'b0fFRRUMDXv'
consumer_token_secret = 'b0fFRRUMDXv'
```

Insert your token here

In this part of the code, I have initialized all of my twitter developer credentials.

```
auth = tweepy.OAuthHandler(consumer_token, consumer_token_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

conn = MongoClient('localhost', 27017)
db = conn.twitter
db.tweets.create_index([("id", pymongo.ASCENDING)], unique=True)
collection = db.tweets
```

In this part of the code, I have established all the authentication and connection with twitter\_api and MongoDB

```

try:
    for tweet in tweepy.Cursor(api.search, q='Presidential Election', lang="en").items(50000):
        # tweet_text = tweet.text
        time = tweet.created_at
        # tweeter = tweet.user.screen_name
        # tweet_dict = {"tweet_text" : tweet_text.strip(), "timestamp" : str(time), "user" : tweeter,
        #               "location" : tweet.user.location, "entities" : tweet.user.entities}
        tweet_json = json.dumps(tweet._json)
        print(tweet_json)
        try:
            with open("Presidential Election.json", "a") as file: #appending the JSON data to the file
                file.write(tweet_json+"\n")
        except:
            pass
    except tweepy.TweepError:
        time.sleep(60)

```

This part of the code, I have used my authentication to find all the tweets regarding presidential election and specified and all of those should be in language English.

And with every parsed tweet I have write one column in my json.

Just in case if any error occurred – I have initialized that program should sleep for 60 seconds

```

class streamdata(StreamListener):
    def on_data(self, data):
        # Load the Tweet into the variable "tweet"
        try:
            tweet = json.loads(data)
            # Pull important data from the tweet to store in the database.
            tweet_id = tweet['id_str']
            tweet_time = tweet['retweeted_status']['created_at']
            username = tweet['user']['screen_name']
            person_name = tweet['user']['name']
            followers_count = tweet['user']['followers_count']
            friends_count = tweet['user']['friends_count']
            user_location = tweet['user']['location']
            text = tweet['text']
            hashtags = tweet['entities']['hashtags']

```

In this part, according to this code all this attributed element should be in json.

```

# Load all of the extracted Tweet data into the variable "tweet" that will be stored into the database
tweet_obj = {
    'id':tweet_id,
    'created':tweet_time,
    'username':username,
    'name':person_name,
    'followers':followers_count,
    'friends': friends_count,
    'tweet_text':text,
    'location':user_location,
    'hashtags':hashtags
}

# Save the refined Tweet data to MongoDB
collection.insert_one(tweet_obj)
return True
except:
    pass

```

This code will insert every tweet one by one in the noSql(mongodb).

```

if __name__ == "__main__":
    authentication = OAuthHandler(consumer_token, consumer_token_secret)
    authentication.set_access_token(access_token, access_token_secret)
    tweeStream = Stream(authentication, streamdata(num_tweets=100000))
    tweeStream.filter(track=["Presidential Election"], languages=['en'])

```

On this part, the code start execution on its own.

## Output Results

All the tweets are stored here.

MongoDB Compass - localhost:27017/twitter.tweets1

Connect View Collection Help

Local

5 DBS

COLLECTIONS

FAVORITE

HOST

localhost:27017

CLUSTER

Standalone

EDITION

MongoDB 4.4.1 Community

Filter your data

admin

config

local

test

twitter

tweets

tweets1

twitter.tweets1 Documents

DOCUMENTS 286.8k

TOTAL SIZE 1.6GB

AVG. SIZE 5.8KB

INDEXES 1

TOTAL SIZE 2.9MB

AVG. SIZE 2.9MB

Documents

Aggregations

Schema

Explain Plan

Indexes

Validation

0 FILTER

ADD DATA

VIEW

Displaying documents 1 - 20 of 286781

REFRESH

# tweets1

	_id ObjectId	created_at String	id Int64	id_str String	text String	source String
1	5f83cc3f96e27ba8ce4f7d3e	"Mon Oct 12 03:23:36 +0000 2020"	1315493292147179520	"1315493292147179520"	"RT @charliekirk11: Joe Biden i"	"ca href="http://t
2	5f83cc3f96e27ba8ce4f7d3f	"Mon Oct 12 03:23:36 +0000 2020"	1315493292168052743	"1315493292168052743"	"RT @ABC: BREAKING: Joe Biden h"	"ca href="http://t
3	5f83cc3f96e27ba8ce4f7d40	"Mon Oct 12 03:23:36 +0000 2020"	1315493294265307137	"1315493294265307137"	"RT @AndrewMcCarthy: Joe Biden"	"ca href="http://t
4	5f83cc3f96e27ba8ce4f7d41	"Mon Oct 12 03:23:36 +0000 2020"	1315493294550339585	"1315493294550339585"	"RT @ajitdata: Biden tried sab"	"ca href="http://t
5	5f83cc3f96e27ba8ce4f7d42	"Mon Oct 12 03:23:37 +0000 2020"	1315493295737368576	"1315493295737368576"	"RT @HelenKennedy: Joe Biden's"	"ca href="http://t
6	5f83cc3f96e27ba8ce4f7d43	"Mon Oct 12 03:23:37 +0000 2020"	1315493296530104320	"1315493296530104320"	"RT @BreitbartNews: Vice Presid"	"ca href="http://t
7	5f83cc3f96e27ba8ce4f7d44	"Mon Oct 12 03:23:37 +0000 2020"	1315493297322950658	"1315493297322950658"	"RT @RimansHisMouth: 50 millio"	"ca href="http://t
8	5f83cc3f96e27ba8ce4f7d45	"Mon Oct 12 03:23:37 +0000 2020"	1315493297427804161	"1315493297427804161"	"RT @realDonaldTrump: Record 96"	"ca href="http://t
9	5f83cc3f96e27ba8ce4f7d46	"Mon Oct 12 03:23:37 +0000 2020"	1315493297520029698	"1315493297520029698"	"RT @MarkMc20: Joe Biden's Cov"	"ca href="https://
10	5f83cc3f96e27ba8ce4f7d47	"Mon Oct 12 03:23:37 +0000 2020"	1315493297599606784	"1315493297599606784"	"RT @realDonaldTrump: Record 96"	"ca href="http://t
11	5f83cc3f96e27ba8ce4f7d48	"Mon Oct 12 03:23:37 +0000 2020"	1315493297583017984	"1315493297583017984"	"RT @allinwithchris: Why won't"	"ca href="http://t
12	5f83cc3f96e27ba8ce4f7d49	"Mon Oct 12 03:23:37 +0000 2020"	1315493297847234568	"1315493297847234568"	"RT @realDonaldTrump: Record 96"	"ca href="http://t
13	5f83cc3f96e27ba8ce4f7d4a	"Mon Oct 12 03:23:37 +0000 2020"	1315493297838845952	"1315493297838845952"	"RT @Herriage: Helmut Norpeth"	"ca href="http://t
14	5f83cc3f96e27ba8ce4f7d4b	"Mon Oct 12 03:23:37 +0000 2020"	1315493298019147777	"1315493298019147777"	"RT @charliekirk11: 175 differe"	"ca href="http://t
15	5f83cc3f96e27ba8ce4f7d4c	"Mon Oct 12 03:23:37 +0000 2020"	1315493298232942595	"1315493298232942595"	"RT @emptywheel: Joe Biden shou"	"ca href="http://t
16	5f83cc3f96e27ba8ce4f7d4d	"Mon Oct 12 03:23:37 +0000 2020"	1315493298367352832	"1315493298367352832"	"RT @KlayTravis: And most of al"	"ca href="http://t

> MongoDB Beta

Type here to search

1003 PM

10/12/2020

Query to find all the Donald Trump and Joe Biden tweets, for this I have used regex

```
{ text: { $regex: /^@realDonaldTrump/i } }
```

```
{ text: { $regex: /^@JoeBiden/i } }
```



MongoDB Compass - localhost:27017/twitter.tweets1

Connect View Collection Help

Local

5 DBS 4 COLLECTIONS

☆ FAVORITE

HOST localhost:27017

CLUSTER Standalone

EDITION MongoDB 4.4.1 Community

Q Filter your data

> admin

> config

> local

> test

> twitter

tweets

tweets1 ...

> MongoDB Beta

twitter.tweets1 Documents

DOCUMENTS 286.8k TOTAL SIZE 1.6GB AVG. SIZE 5.8KB INDEXES 1 TOTAL SIZE 2.9MB AVG. SIZE 2.9MB

Documents Aggregations Schema Explain Plan Indexes Validation

0 FILTER { text: { \$regex: /@joeliden/ } } OPTIONS FIND RESET ...

+ ADD DATA + VIEW

Displaying documents 1 - 20 of 675 C REFRESH

```
{
  "_id": ObjectId("5f83cc5f96e27ba8ce47fe0b"),
  "created_at": "Mon Oct 12 03:24:00 +0000 2020",
  "id": "1315493430009769089",
  "id_str": "1315493430009769089",
  "text": "@joeliden people are not stupid enough to vote for you and then find o...",
  "source": "a href='http://twitter.com/download/iphone' rel='nofollow'>twitter fo...",
  "truncated": true,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": 939091,
  "in_reply_to_user_id_str": "939091",
  "in_reply_to_screen_name": "joeliden",
  "user": {
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "quoted_status_id": 1315341500725600371,
    "quoted_status_id_str": "1315341500725600371",
    "quoted_status": {
      "is_quote_status": true,
      "extended_tweet": {
        "quote_count": 0,
        "reply_count": 0
      }
    }
  }
}
```

SHOW 8 MORE FIELDS

```
{
  "_id": ObjectId("5f83cc7596e27ba8ce47fef"),
  "created_at": "Mon Oct 12 03:24:31 +0000 2020",
  "id": "1315493525958475776",
  "id_str": "1315493525958475776",
  "text": "a href='http://twitter.com/download/iphone' rel='nofollow'>twitter fo..."
}
```

MongoDB Compass - localhost:27017/twitter.tweets1

Connect View Collection Help

Local

5 DBS 4 COLLECTIONS

☆ FAVORITE

HOST localhost:27017

CLUSTER Standalone

EDITION MongoDB 4.4.1 Community

Q Filter your data

> admin

> config

> local

> test

> twitter

tweets

tweets1 ...

> MongoDB Beta

twitter.tweets1 Documents

DOCUMENTS 286.8k TOTAL SIZE 1.6GB AVG. SIZE 5.8KB INDEXES 1 TOTAL SIZE 2.9MB AVG. SIZE 2.9MB

Documents Aggregations Schema Explain Plan Indexes Validation

0 FILTER { text: { \$regex: /@realDonaldTrump/ } } OPTIONS FIND RESET ...

+ ADD DATA + VIEW

Displaying documents 1 - 20 of 1379 C REFRESH

```
{
  "_id": ObjectId("5f83cc69a20408859e4b4f5"),
  "created_at": "Mon Oct 12 03:34:38 +0000 2020",
  "id": "1315490608935753734",
  "id_str": "1315490608935753734",
  "text": "@realDonaldTrump I'm old enough to remember when Trump colluded with R...",
  "display_text_range": Array,
  "source": "a href='https://mobile.twitter.com' rel='nofollow'>Twitter Web App/a...",
  "truncated": true,
  "in_reply_to_status_id": 1315480195990781952,
  "in_reply_to_status_id_str": "1315480195990781952",
  "in_reply_to_user_id": 25073877,
  "in_reply_to_user_id_str": "25073877",
  "in_reply_to_screen_name": "realDonaldTrump",
  "user": {
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "is_quote_status": false,
    "extended_tweet": {
      "quote_count": 0,
      "reply_count": 0,
      "retweet_count": 0,
      "favorite_count": 0
    },
    "entities": {
      "hashtags": null,
      "mentions": null,
      "urls": null,
      "media": null
    }
  }
}
```

SHOW 8 MORE FIELDS

```
{
  "_id": ObjectId("5f83cc61a20408859e4c4f2"),
  "created_at": "Mon Oct 12 03:34:51 +0000 2020",
  "id": "1315496125881671680",
  "id_str": "1315496125881671680",
  "text": "@realDonaldTrump Next year will be the BEST EVER because Joe Biden will..."
}
```

[illegible][illegible]

# Match Query

Filter

{ "user.favourites\_count": { \$lte: NumberInt(7000) } }

Options

MongoDB Compass - localhost:27017/tweets.tweets1

Connect View Collection Help

Local

5 DBS 4 COLLECTIONS

HOST localhost:27017

CLUSTER Standalone

EDITION MongoDB 4.4.1 Community

Filter your data

admin

config

local

test

twitter

tweets

tweets1

twitter.tweets1 Documents

Documents Aggregations Schema Explain Plan Indexes Validation

DOCUMENTS 286.8k TOTAL SIZE 1.6GB AVG. SIZE 5.8KB INDEXES 1 TOTAL SIZE 2.9MB AVG. SIZE 2.9MB

Filter { "user.favourites\_count": { \$lte: NumberInt(7000) } }

Options FIND RESET

Displaying documents 1 - 20 of 76831

Refresh

#	tweets1	_id ObjectId	created_at String	id Int64	id_str String	text String	source String
1		5f83cc3f96e27ba8ce4f7542	"Mon Oct 12 03:23:37 +0000 2020"	1315493295737368576	"1315493295737368576"	"RT @HelenKennedy: Joe Biden's"	"ca href="http://t"
2		5f83cc3f96e27ba8ce4f754a	"Mon Oct 12 03:23:37 +0000 2020"	1315493297838845952	"1315493297838845952"	"RT @Kierriage: Helmut Norpeth"	"ca href="http://t"
3		5f83cc3f96e27ba8ce4f754f	"Mon Oct 12 03:23:37 +0000 2020"	1315493298476326912	"1315493298476326912"	"Just one more example that the"	"ca href="http://t"
4		5f83cc4096e27ba8ce4f7553	"Mon Oct 12 03:23:38 +0000 2020"	1315493300242206720	"1315493300242206720"	"RT @thebradfordfile: LOL. Zero"	"ca href="http://t"
5		5f83cc4096e27ba8ce4f7557	"Mon Oct 12 03:23:38 +0000 2020"	1315493302616031232	"1315493302616031232"	"RT @mepadellan: I'm all in for"	"ca href="https://"
6		5f83cc4196e27ba8ce4f755b	"Mon Oct 12 03:23:39 +0000 2020"	1315493304189277187	"1315493304189277187"	"RT @DineshMSouza: Tell us Joe"	"ca href="http://t"
7		5f83cc4196e27ba8ce4f755c	"Mon Oct 12 03:23:39 +0000 2020"	1315493304222488706	"1315493304222488706"	"RT @JuddLegum: Joe Biden won't"	"ca href="http://t"
8		5f83cc4196e27ba8ce4f7560	"Mon Oct 12 03:23:39 +0000 2020"	1315493304952188928	"1315493304952188928"	"RT @charliekirk11: Joe Biden i"	"ca href="http://t"
9		5f83cc4196e27ba8ce4f7561	"Mon Oct 12 03:23:39 +0000 2020"	1315493304914644992	"1315493304914644992"	"RT @if_you_see_kaye: If Mitch"	"ca href="https://"
10		5f83cc4196e27ba8ce4f7567	"Mon Oct 12 03:23:39 +0000 2020"	1315493307355598848	"1315493307355598848"	"RT @jjdpower1978: @JoeBiden RE"	"ca href="http://t"
11		5f83cc4296e27ba8ce4f7569	"Mon Oct 12 03:23:40 +0000 2020"	1315493308391710720	"1315493308391710720"	"@NBCNewsTHINK @gerardush Rene"	"ca href="http://t"
12		5f83cc4296e27ba8ce4f756b	"Mon Oct 12 03:23:40 +0000 2020"	1315493308794171394	"1315493308794171394"	"RT @jff510: I just saw an ad f"	"ca href="http://t"
13		5f83cc4296e27ba8ce4f7571	"Mon Oct 12 03:23:40 +0000 2020"	1315493310920038077	"1315493310920038077"	"RT @emptywheel: Joe Biden shou"	"ca href="http://t"
14		5f83cc4296e27ba8ce4f7574	"Mon Oct 12 03:23:40 +0000 2020"	1315493311038267392	"1315493311038267392"	"RT @louobobs: Maybe the Pollst"	"ca href="http://t"
15		5f83cc4296e27ba8ce4f7575	"Mon Oct 12 03:23:40 +0000 2020"	1315493311474528257	"1315493311474528257"	"RT @ilucyhamah: Let that sink"	"ca href="http://t"
16		5f83cc4296e27ba8ce4f757a	"Mon Oct 12 03:23:40 +0000 2020"	1315493312296558592	"1315493312296558592"	"RT @jff510: I just saw an ad f"	"ca href="http://t"

MongoSH Beta

Type here to search

ENG 10:30 PM 10/12/2020

# Group Query

The screenshot displays the MongoDB Compass interface for a local instance. The left sidebar shows the database structure with 'tweets' and 'tweets1' collections. The main panel shows the 'tweets' collection with a filter applied: `{ "user.followers_count": { "$lte": NumberInt(225) }, "user.friends_count": { "$gt": NumberInt(27) } }`. The results show documents 1 to 20 of 94455. The first document is expanded, showing fields like `_id`, `created_at`, `id`, `id_str`, `text`, `source`, `truncated`, `in_reply_to_status_id`, `in_reply_to_status_id_str`, `in_reply_to_user_id`, `in_reply_to_user_id_str`, `in_reply_to_screen_name`, `user`, `geo`, `coordinates`, `place`, `contributors`, `retweeted_status`, `is_quote_status`, `quote_count`, `reply_count`, `retweet_count`, `favorite_count`, `entities`, and `favorited`.

MongoDB Compass - localhost:27017/tweets.tweets1

Connect View Collection Help

Local

5 DBS 4 COLLECTIONS

localhost:27017

CLUSTER Standalone

EDITION MongoDB 4.4.1 Community

Filter your data

admin

config

local

test

twitter

tweets

tweets1

twitter.tweets1

Documents

DOCUMENTS 286.8k TOTAL SIZE 1.6GB AVG. SIZE 5.8KB INDEXES 1 TOTAL SIZE 2.9MB AVG. SIZE 2.9MB

Documents Aggregations Schema Explain Plan Indexes Validation

FILTER { "user.followers\_count": { "\$lte": NumberInt(225) }, "user.friends\_count": { "\$gt": NumberInt(27) } }

ADD DATA VIEW

Displaying documents 1 - 20 of 94455

Refresh

Document 1:

```
{
  "_id": ObjectId("5f83cc3e96e27ba8ce4f7d3f"),
  "created_at": "Mon Oct 12 03:23:36 +0000 2020",
  "id": 1315493292168052743,
  "id_str": "1315493292168052743",
  "text": "RT @ABC: BREAKING: Joe Biden holds a 17-point lead over Pres. Trump in ...",
  "source": "a href='\"http://twitter.com/download/iphone\"' rel='nofollow'>twitter fo...",
  "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": {
      "is_quote_status": false,
      "quote_count": 0,
      "reply_count": 0,
      "retweet_count": 0,
      "favorite_count": 0,
      "entities": {
        "favorited": false
      }
    }
  }
}
```

SHOW 4 MORE FIELDS

Document 2:

```
{
  "_id": ObjectId("5f83cc3e96e27ba8ce4f7d41"),
  "created_at": "Mon Oct 12 03:23:36 +0000 2020",
  "id": 1315493294558339585,
  "id_str": "1315493294558339585"
}
```

MongoDB Beta

Type here to search

ENG 7:39 PM 10/12/2020

# Aggregated Pipelining

MongoDB Compass - localhost:27017/tweets.tweets1

Local

5 DBS 4 COLLECTIONS

HOST: localhost:27017

CLUSTER: Standalone

EDITION: MongoDB 4.4.1 Community

Filter your data

- admin
- config
- local
- test
- twitter
- tweets
- tweets1

twitter.tweets1 Aggregations

DOCUMENTS 286.8k TOTAL SIZE 1.6GB AVG. SIZE 5.8KB INDEXES 1 TOTAL SIZE 2.9MB AVG. SIZE 2.9MB

Documents Aggregations Schema Explain Plan Indexes Validation

COLLATION: Untitled Modified SAVE

Output after \$match stage (Sample of 20 documents)

```
1 // **
2 // query: The query in MQL.
3 //
4 //
5 {
6   "user.friends_count": NumberInt(48)
7 }
8
```

Output after \$project stage (Sample of 20 documents)

```
1 // **
2 // specifications: The fields to
3 // include or exclude.
4 //
5 {
6   truncated: false,
7   retweet_count: 0,
8 }
9
```

Sample document 1:

```
{
  "_id": ObjectId("5f83ccc8a5aef43d210831"),
  "created_at": "Mon Oct 12 03:25:54 +0000 2020",
  "id_str": "1315493871170867203",
  "text": "RT @ananavarro: Oh, please . In comparison to Trump, the most chaotic ...",
  "source": "<a href='\"http://twitter.com/download/iphone\"' rel='nofollow'>Twitter fo...",
  "truncated": false
}
```

Sample document 2:

```
{
  "_id": ObjectId("5f83ccf59a20408059ee33ab"),
  "created_at": "Mon Oct 12 03:26:39 +0000 2020",
  "id_str": "1315494060149428224",
  "text": "RT @LabyrinthHeaver: @realDonaldTrump The same poll shows your approva...",
  "source": "<a href='\"http://twitter.com/download/android\"' rel='nofollow'>Twitter f...",
  "truncated": false
}
```

Sample document 3:

```
{
  "_id": ObjectId("5f83ccc8a5aef43d210831"),
  "created_at": "Mon Oct 12 03:25:54 +0000 2020",
  "id_str": "1315493871170867203",
  "text": "RT @ananavarro: Oh, please . In comparison to Trump, the most chaotic ...",
  "source": "<a href='\"http://twitter.com/download/iphone\"' rel='nofollow'>Twitter fo...",
  "truncated": false
}
```

Sample document 4:

```
{
  "_id": ObjectId("5f83ccf59a20408059ee33ab"),
  "created_at": "Mon Oct 12 03:26:39 +0000 2020",
  "id_str": "1315494060149428224",
  "text": "RT @LabyrinthHeaver: @realDonaldTrump The same poll shows your approva...",
  "source": "<a href='\"http://twitter.com/download/android\"' rel='nofollow'>Twitter f...",
  "truncated": false
}
```

MongoSH Beta

Type here to search

ENG 9:12 PM 10/12/2020