PES UNIVERSITY

100 feet Ring Road, BSK 3rd Stage
Bengaluru 560085

Department of Computer Science and Engineering
B. Tech. CSE - 6th Semester
Jan – May 2022

UE19CS344
DATABASE TECHNOLOGIES (DBT)

Project Report

Real-Time Twitter IPL Data Streaming and Processing using Spark and Kafka

PES2UG19CS219: Manjunath Y
PES2UG19CS249: Naveen S Nelogal
PES2UG19CS312: Raghav S K
PES2UG19CS312: Vivek S D

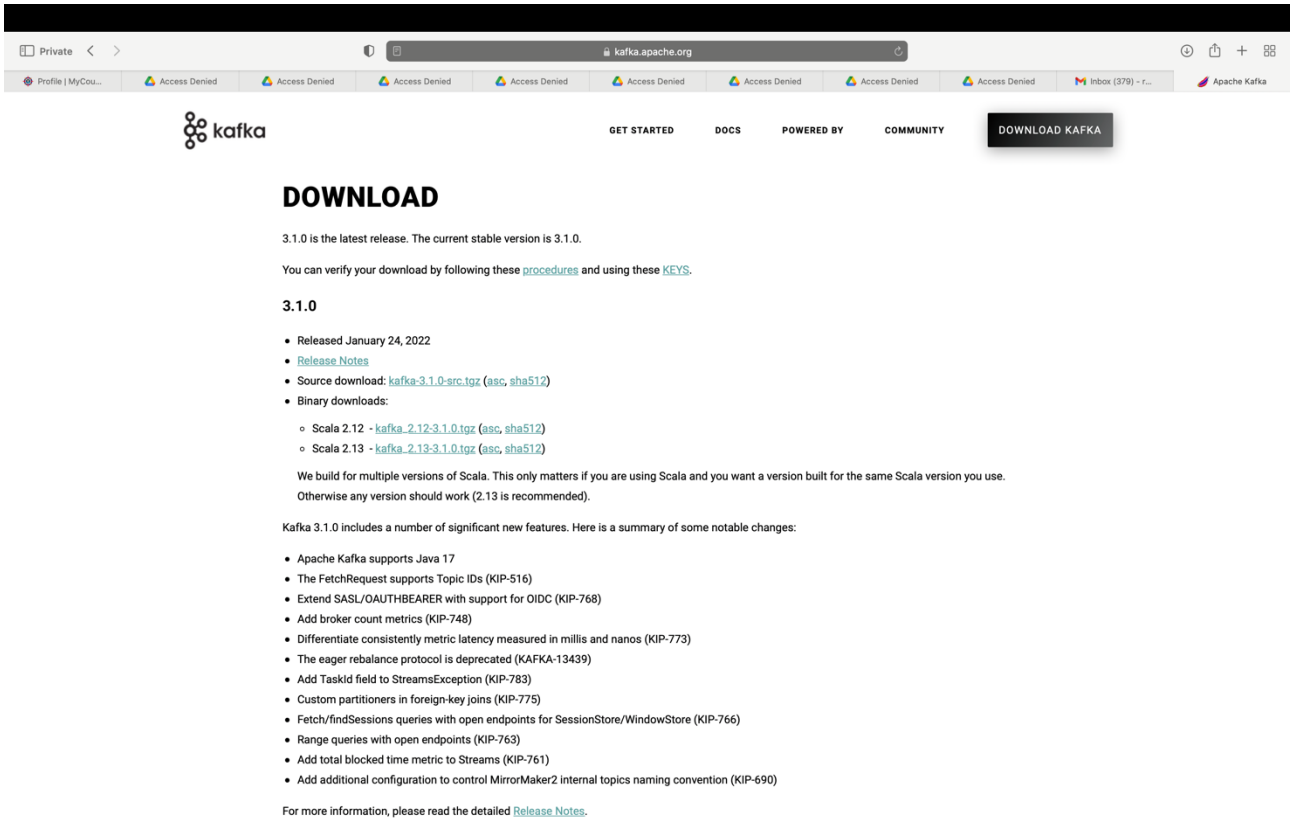
Table of Contents

Introduction

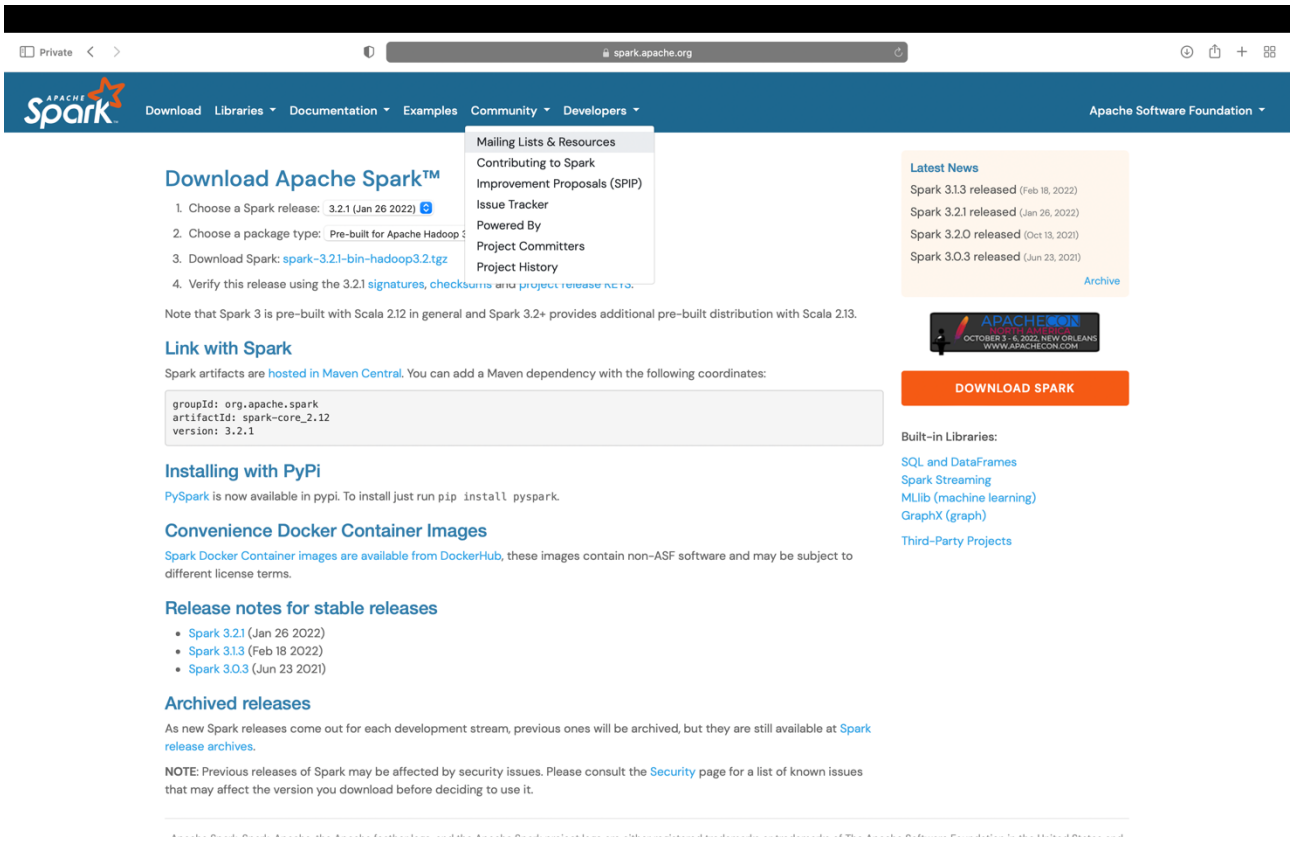
The project is focused on streaming of data i.e. tweets from Twitter using the Twitter API and producing the data to a topic on Kafka. The tweets are filtered using the tag “IPL”. The tweets are then streamed and produced to a topic on Kafka. The consumer later consumes the same tweets by subscribing to the topic on Kafka. Spark Streaming is used for the same and the tweets are streamed and then queries are executed on the same using Spark SQL. The data is then stored to MySQL database and batch processing is done on the stored data and SQL queries are performed on the data.

Installation of Software

- Installing Kafka



- Installing Spark



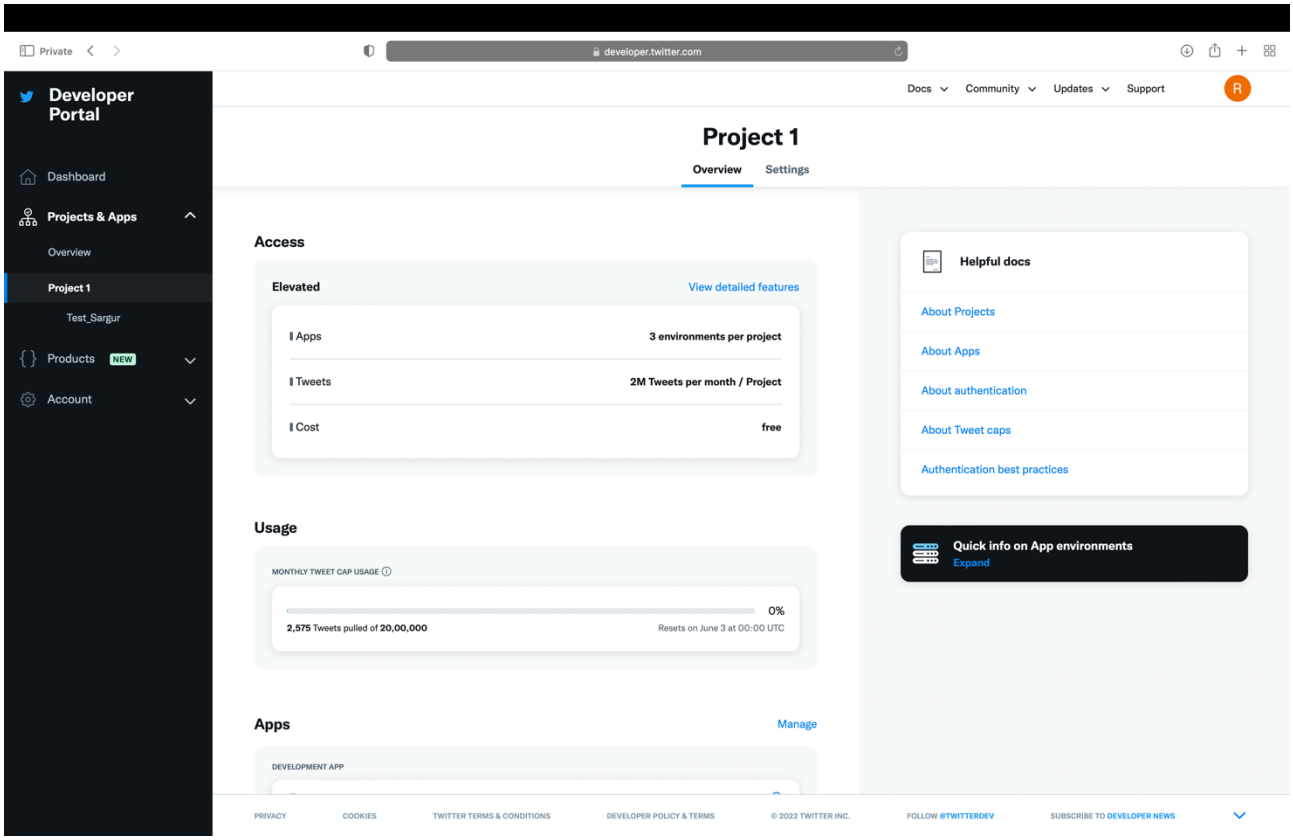
- Installing Python libraries using pip
  - pip install pyspark
  - pip install kafka-python
  - pip install tweepy

Input Data

Source of the input data is Realtime data from Twitter API

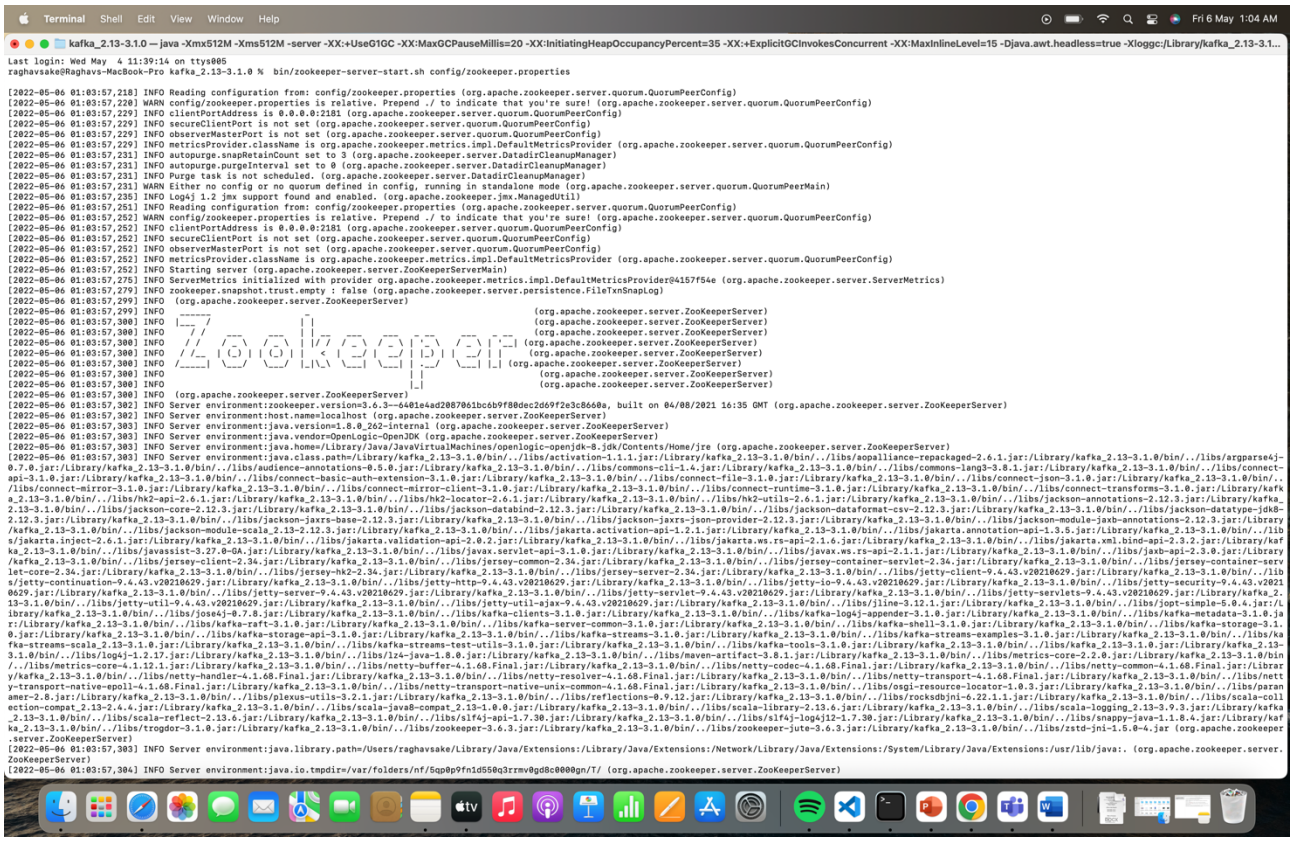
The twitter API streams data from twitter. The tweets are streamed and the data is in json format. The main field is the text field which consists of the main text the tweet contains and the link to pictures uploaded in the tweet and link to the tweet.

Twitter data is considered big data due to the sheer volume, velocity and the variety of the data.



Streaming Mode Experiment

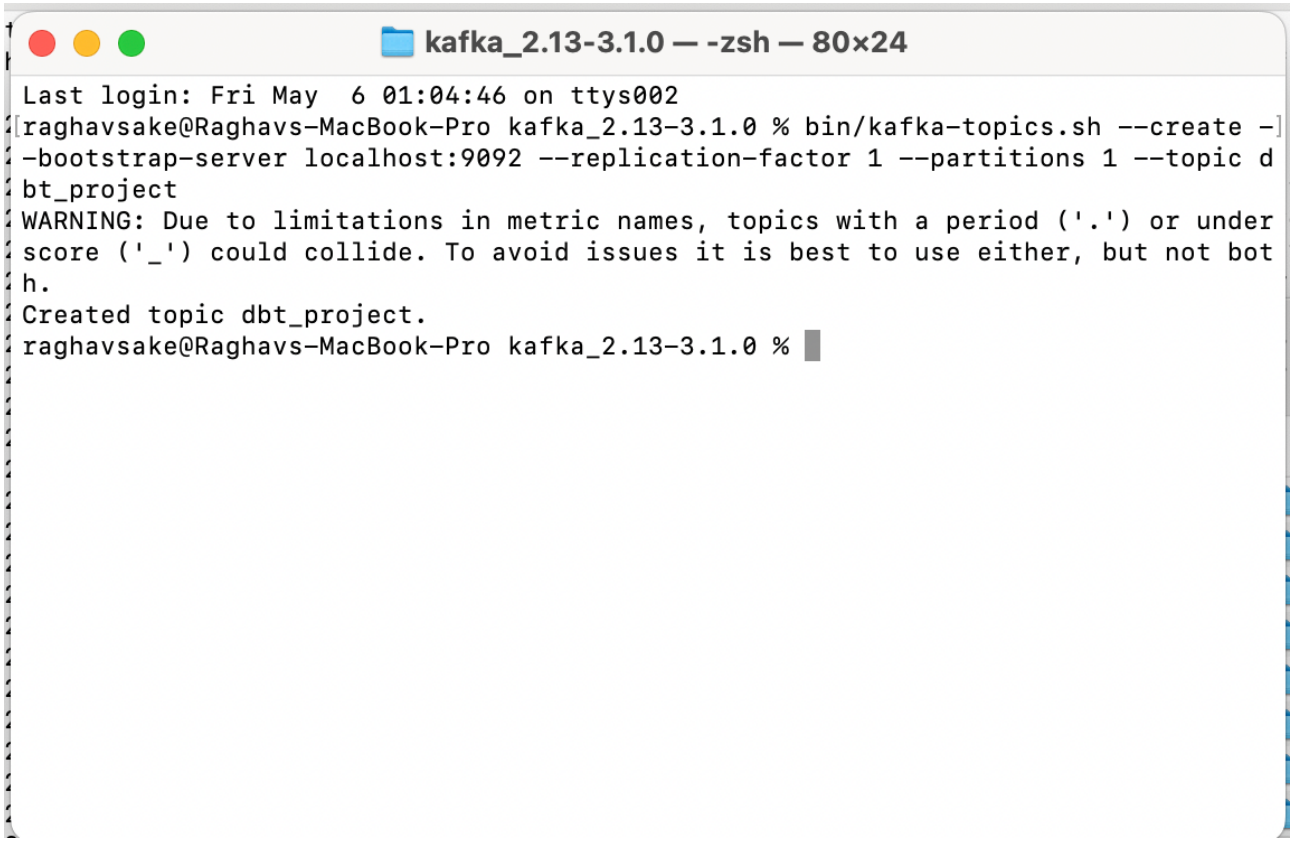
Starting ZooKeeper



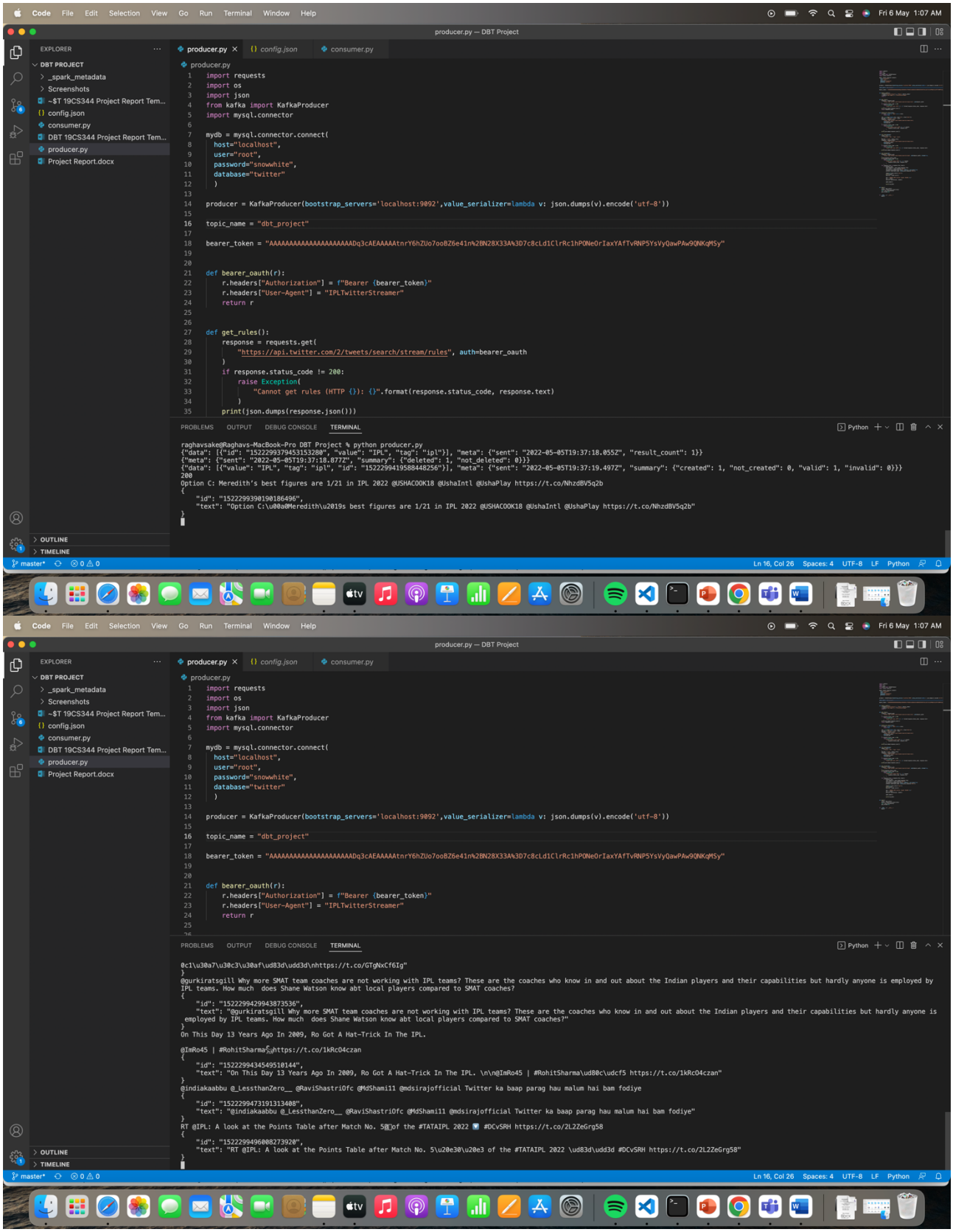
Starting Kafka Server

```
Last login: Fri May  6 01:03:46 on ttys000
raghavsake@Raghavs-MacBook-Pro kafka_2.13-3.1.0 % bin/kafka-server-start.sh config/server.properties
[2022-05-06 01:04:26,201] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2022-05-06 01:04:26,518] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2022-05-06 01:04:26,646] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2022-05-06 01:04:26,683] INFO starting (kafka.server.KafkaServer)
[2022-05-06 01:04:26,683] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2022-05-06 01:04:26,700] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2022-05-06 01:04:26,707] INFO Client environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:host.name=localhost (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.version=1.8.0_262-internal (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.vendor=OpenLogic-OpenJDK (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.home=/Library/Java/JavaVirtualMachines/openlogic-openjdk-8.jdk/Contents/Home/jre (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.class.path=/Library/kafka_2.13-3.1.0/bin/../libs/activation-1.1.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/aopalliance-repackaged-2.6.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/argparse4j-
0.7.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/audience-annotations-0.5.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/commons-cli-1.4.jar:/Library/kafka_2.13-3.1.0/bin/../libs/commons-lang3-3.8.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-
api-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-basic-auth-extension-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-file-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-json-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/..
/libs/connect-mirror-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-mirror-client-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-runtime-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/connect-transforms-3.1.0.jar:/Library/kafk
a_2.13-3.1.0/bin/../libs/hk2-api-2.6.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/hk2-locator-2.6.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/hk2-utils-2.6.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-annotations-2.12.3.jar:/Library/kafka_
2.13-3.1.0/bin/../libs/jackson-core-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-databind-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-dataformat-csv-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-datatype-jdk8-
2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-jaxrs-base-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-jaxrs-json-provider-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jackson-module-jaxb-annotations-2.12.3.jar:/Library
/kafka_2.13-3.1.0/bin/../libs/jackson-module-scala_2.13-2.12.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jakarta.activation-api-1.2.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jakarta.annotation-api-1.3.5.jar:/Library/kafka_2.13-3.1.0/bin/../lib
s/jakarta.inject-2.6.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jakarta.validation-api-2.0.2.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jakarta.ws.rs-api-2.1.6.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jakarta.xml.bind-api-2.3.2.jar:/Library/kaf
ka_2.13-3.1.0/bin/../libs/javassist-3.27.0-GA.jar:/Library/kafka_2.13-3.1.0/bin/../libs/javax.servlet-api-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/javax.ws.rs-api-2.1.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jaxb-api-2.3.0.jar:/Library
/kafka_2.13-3.1.0/bin/../libs/jersey-client-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jersey-common-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jersey-container-servlet-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jersey-container-serv
let-core-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jersey-hk2-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jersey-server-2.34.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-client-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../lib
s/jetty-continuation-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-http-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-io-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-security-9.4.43.v2021
0629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-server-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-servlet-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-servlets-9.4.43.v20210629.jar:/Library/kafka_2.
13-3.1.0/bin/../libs/jetty-util-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jetty-util-ajax-9.4.43.v20210629.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jline-3.12.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/jopt-simple-5.0.4.jar:/L
ibrary/kafka_2.13-3.1.0/bin/../libs/jose4j-0.7.8.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-clients-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-log4j-appender-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-metadata-3.1.0.ja
r:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-raft-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-server-common-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-shell-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-storage-3.1.
0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-storage-api-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-streams-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-streams-examples-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/ka
fka-streams-scala_2.13-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-streams-test-utils-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka-tools-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/kafka_2.13-3.1.0.jar:/Library/kafka_2.13-
3.1.0/bin/../libs/log4j-1.2.17.jar:/Library/kafka_2.13-3.1.0/bin/../libs/lz4-java-1.8.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/maven-artifact-3.8.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/metrics-core-2.2.0.jar:/Library/kafka_2.13-3.1.0/bin
/../libs/metrics-core-4.1.12.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-buffer-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-codec-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-common-4.1.68.Final.jar:/Librar
y/kafka_2.13-3.1.0/bin/../libs/netty-handler-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-resolver-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-transport-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/nett
y-transport-native-epoll-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/netty-transport-native-unix-common-4.1.68.Final.jar:/Library/kafka_2.13-3.1.0/bin/../libs/osgi-resource-locator-1.0.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/paran
amer-2.8.jar:/Library/kafka_2.13-3.1.0/bin/../libs/plexus-utils-3.2.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/reflections-0.9.12.jar:/Library/kafka_2.13-3.1.0/bin/../libs/rocksdbjni-6.22.1.1.jar:/Library/kafka_2.13-3.1.0/bin/../libs/scala-coll
ection-compat_2.13-2.4.4.jar:/Library/kafka_2.13-3.1.0/bin/../libs/scala-java8-compat_2.13-1.0.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/scala-library-2.13.6.jar:/Library/kafka_2.13-3.1.0/bin/../libs/scala-logging_2.13-3.9.3.jar:/Library/kafka
_2.13-3.1.0/bin/../libs/scala-reflect-2.13.6.jar:/Library/kafka_2.13-3.1.0/bin/../libs/slf4j-api-1.7.30.jar:/Library/kafka_2.13-3.1.0/bin/../libs/slf4j-log4j12-1.7.30.jar:/Library/kafka_2.13-3.1.0/bin/../libs/snappy-java-1.1.8.4.jar:/Library/kaf
ka_2.13-3.1.0/bin/../libs/trogdor-3.1.0.jar:/Library/kafka_2.13-3.1.0/bin/../libs/zookeeper-3.6.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/zookeeper-jute-3.6.3.jar:/Library/kafka_2.13-3.1.0/bin/../libs/zstd-jni-1.5.0-4.jar (org.apache.zookeeper
.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.library.path=/Users/raghavsake/Library/Java/Extensions:/Library/Java/Extensions:/Network/Library/Java/Extensions:/System/Library/Java/Extensions:/usr/lib/java:. (org.apache.zookeeper.ZooKeep
er)
[2022-05-06 01:04:26,707] INFO Client environment:java.io.tmpdir=/var/folders/nf/5qp0p9fn1d558q3rrmv0gd8c0000gn/T/ (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,707] INFO Client environment:java.compiler=<NA> (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.name=Mac OS X (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.arch=x86_64 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.version=10.16 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:user.name=raghavsake (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:user.home=/Users/raghavsake (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:user.dir=/Library/kafka_2.13-3.1.0 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.memory.free=1011MB (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.memory.max=1024MB (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,708] INFO Client environment:os.memory.total=1024MB (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,710] INFO Initiating client connection, connectString=localhost:2181 sessionTimeout=18000 watcher=kafka.zookeeper.ZooKeeperClient$ZooKeeperClientWatcher$@a1153bc (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,713] INFO jute.maxbuffer value is 4194304 Bytes (org.apache.zookeeper.ClientCnxnSocket)
[2022-05-06 01:04:26,717] INFO zookeeper.request.timeout value is 0. feature enabled=false (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,720] INFO [ZooKeeperClient Kafka server] Waiting until connected. (kafka.zookeeper.ZooKeeperClient)
[2022-05-06 01:04:26,721] INFO Opening socket connection to server localhost/127.0.0.1:2181. (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,721] INFO SASL config status: Will not attempt to authenticate using SASL (unknown error) (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,730] INFO Socket connection established, initiating session, client: /127.0.0.1:51188, server: localhost/127.0.0.1:2181 (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,752] INFO Session establishment complete on server localhost/127.0.0.1:2181, session id = 0x10000dc47380000, negotiated timeout = 18000 (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,758] INFO [ZooKeeperClient Kafka server] Connected. (kafka.zookeeper.ZooKeeperClient)
[2022-05-06 01:04:26,844] INFO [feature-zk-node-event-process-thread]: Starting (kafka.server.FinalizedFeatureChangeListener$ChangeNotificationProcessorThread)
[2022-05-06 01:04:27,104] INFO Updated cache from existing <empty> to latest FinalizedFeaturesAndEpoch(features=Features{}, epoch=0). (kafka.server.FinalizedFeatureCache)
[2022-05-06 01:04:27,112] INFO Cluster ID = 7YEuozFrTeaAfjnNaJk2Eg (kafka.server.KafkaServer)
[2022-05-06 01:04:27,181] INFO KafkaConfig values:
        advertised.listeners = null
        alter.config.policy.class.name = null
        alter.log.dirs.replication.quota.window.num = 11
        alter.log.dirs.replication.quota.window.size.seconds = 1
        authorizer.class.name =
        auto.create.topics.enable = true
        auto.leader.rebalance.enable = true
        background.threads = 10
        broker.heartbeat.interval.ms = 2000
```

Creating a Kafka Topic



Streaming data into kafka topic



Subscribing to Kafka Topic and Streaming

The text in tweets streamed and the count of number of tweets

```
+--------+
|count(1)|
+--------+
|      10|
+--------+

+--------------------------------+
|                            text|
+--------------------------------+
|            Option C: Meredit...|
|【3000円クーポン】脱毛器【2年...|
|            @gurkiratsgill Wh...|
|            On This Day 13 Ye...|
|            @indiakaabbu @_Le...|
|            RT @IPL: A look a...|
|            RT @IPL: In conve...|
|            RT @IPL: In conve...|
|            RT @DelhiCapitals...|
|            @RajivJh65236048 ...|
+--------------------------------+
```

The words split and stored from all tweets and the count

```
+--------+
|count(1)|
+--------+
|     241|
+--------+

+--------------------------------+
|                           words|
+--------------------------------+
|                          Option|
|                  C: Meredith's|
|                            best|
|                         figures|
|                             are|
|                            1/21|
|                              in|
|                             IPL|
|                            2022|
|                     @USHACOOK18|
|                       @UshaIntl|
|                       @UshaPlay|
|              https://t.co/Nhzd...|
|【3000円クーポン】脱毛器【2年...|
|          サロン級IPLパルス技術|
|                      フラッシュ|
|                          脱毛機|
|                             vio|
|                            脱毛|
|                              光|
+--------------------------------+
only showing top 20 rows
```

Query1 - The number of RTs which is basically the number of RTs

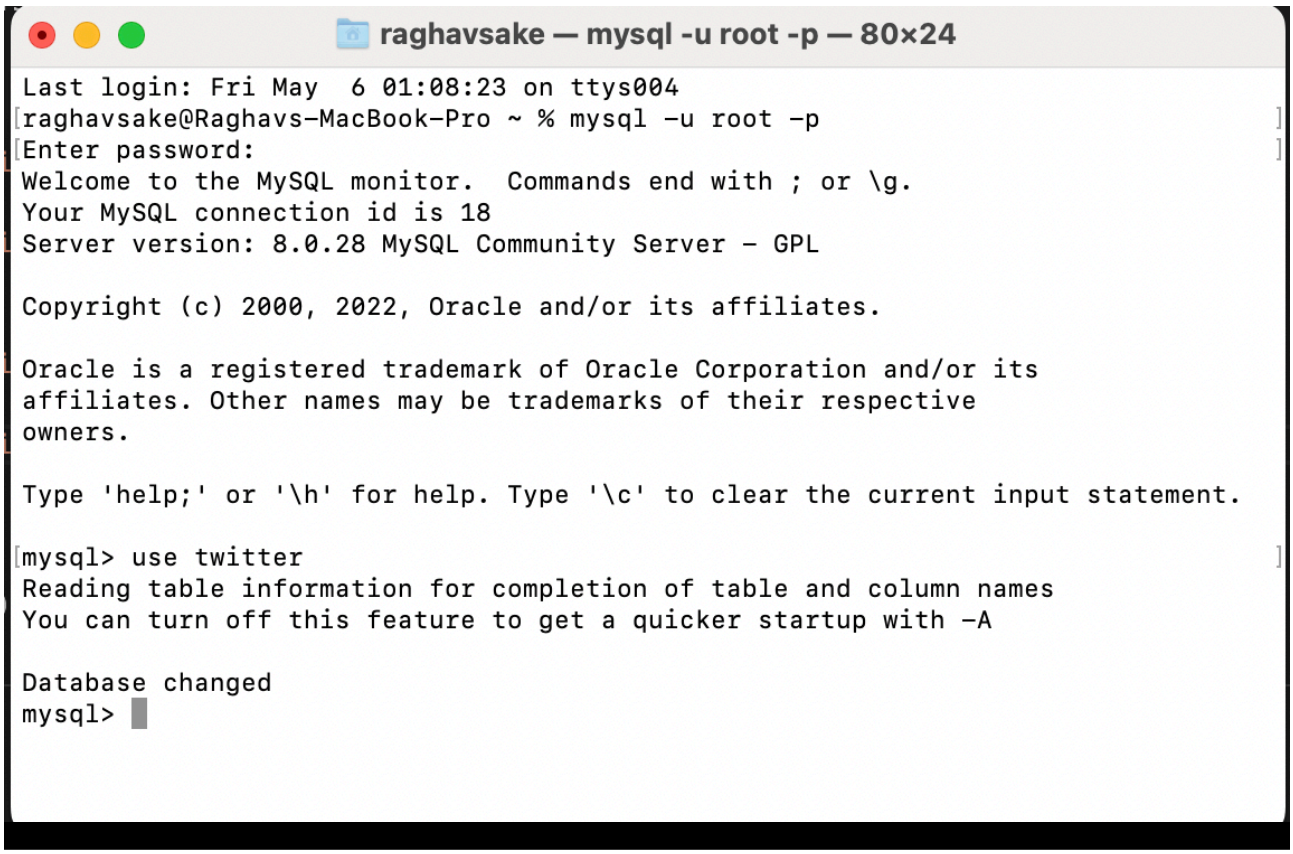
```
+--------+
|count(1)|
+--------+
|       4|
+--------+
```

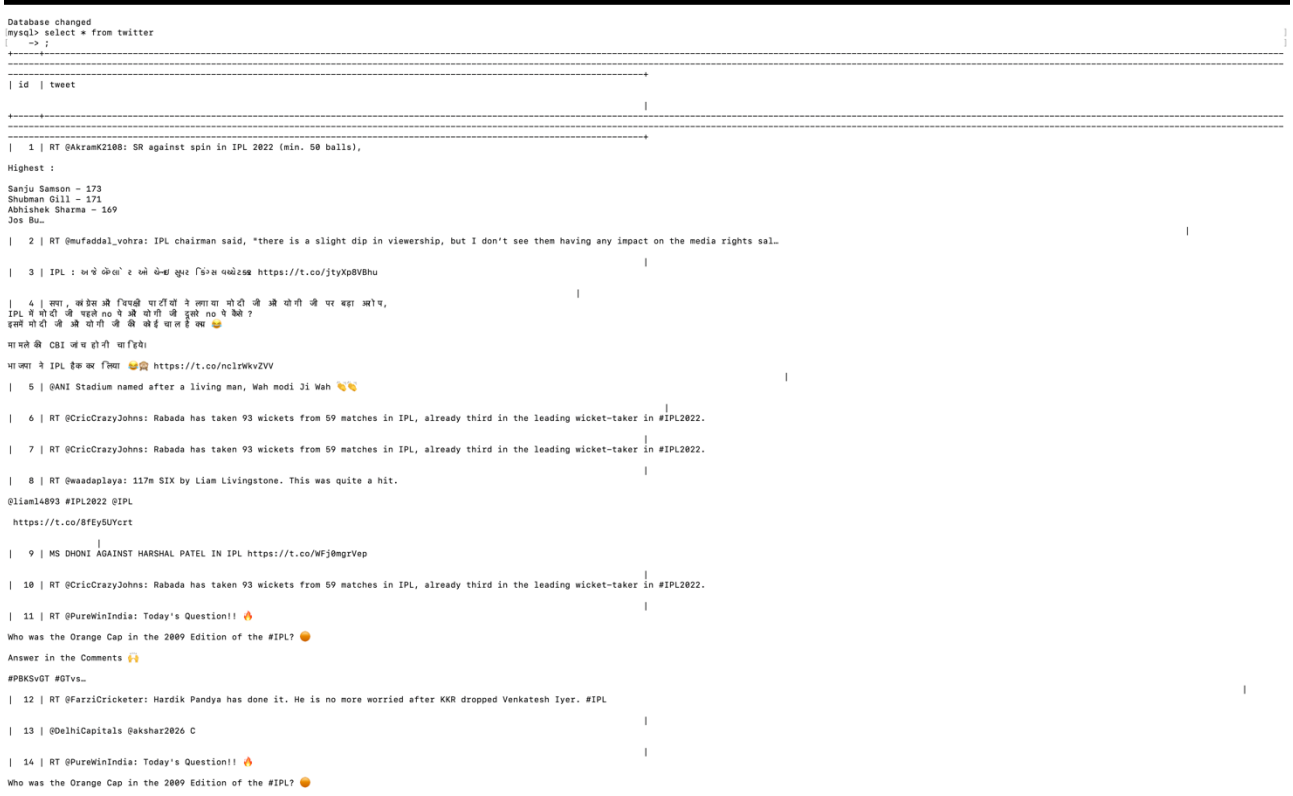
Query2 – The number of tweets containing the word “T20”

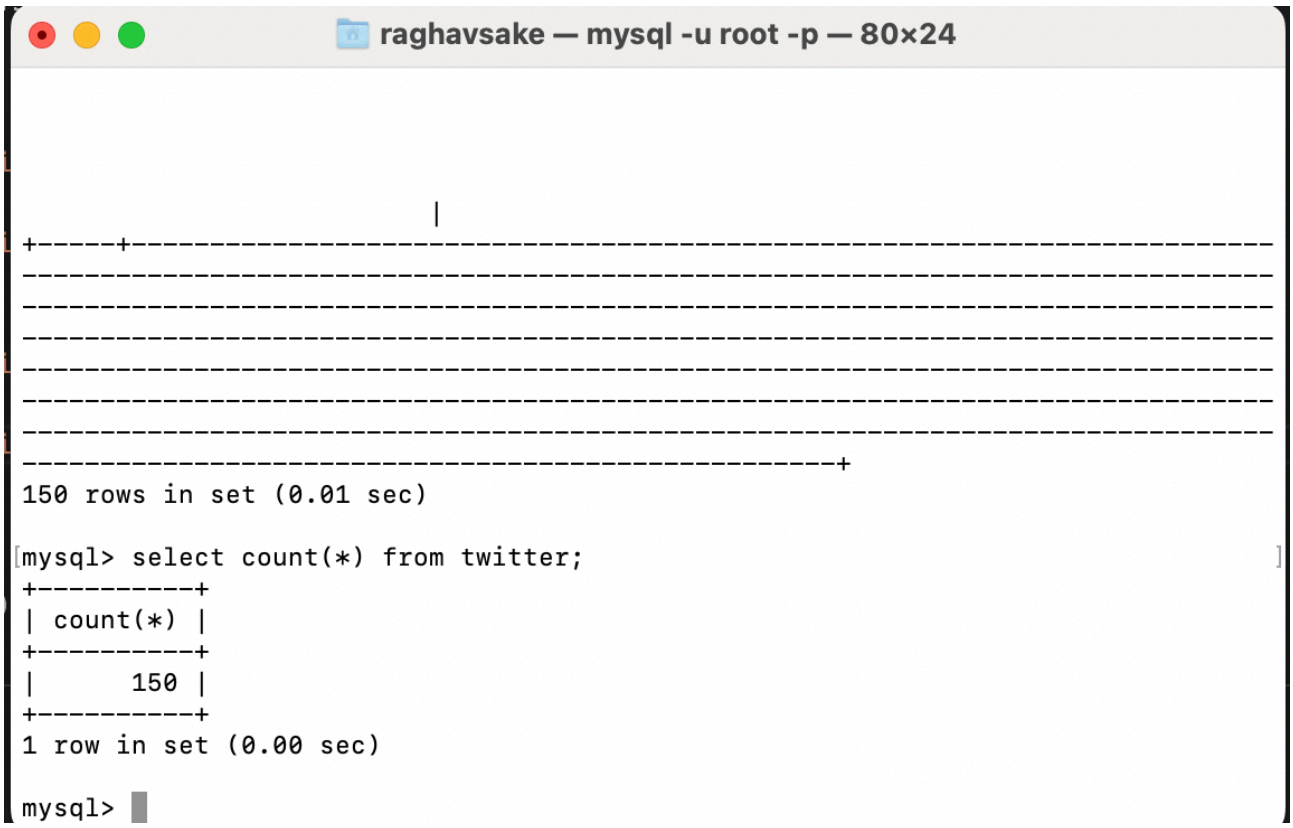
```
+--------+
|count(1)|
+--------+
|       3|
+--------+
```

Batch Mode Experiment

The tweets streamed are also stored in the MySQL database and we try to query the database for the same


raghavsake — mysql -u root -p — 80×24
Last login: Fri May 6 01:08:23 on ttys004
raghavsake@Raghavs-MacBook-Pro ~ % mysql -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 18
Server version: 8.0.28 MySQL Community Server - GPL
Copyright (c) 2000, 2022, Oracle and/or its affiliates.
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
mysql> use twitter
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
Database changed
mysql>


Database changed
mysql> select * from twitter
-> ;
| id | tweet
| 1 | RT @AkramK2108: SR against spin in IPL 2022 (min. 50 balls),
Highest :
Sanju Samson - 173
Shubman Gill - 171
Abhishek Sharma - 169
Jos Bu…
| 2 | RT @mufaddal_vohra: IPL chairman said, "there is a slight dip in viewership, but I don't see them having any impact on the media rights sal…
| 3 | IPL : અને બેસ્ટ ઓલ રાઉન્ડર ઓફ ધ મેચ https://t.co/jtyXpBVBhu
| 4 | क्या , बंदेस अौ विपक्षी पार्टीयों ने लगाया मोदी जी अौ योगी जी पर बड़ा अारोप,
IPL में मोदी जी पहले no पे अौ योगी जी दूसरे no पे खेले ?
इसमें मोदी जी अौ योगी जी की कोई चाल है क्या
मामले की CBI जांच होनी चाहिये।
भाजपा ने IPL हैक कर लिया https://t.co/nclrWkvZVV
| 5 | @ANI Stadium named after a living man, Wah modi Ji Wah
| 6 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
| 7 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
| 8 | RT @waadaplaya: 117m SIX by Liam Livingstone. This was quite a hit.
@liam14893 #IPL2022 @IPL
https://t.co/8fEy5UYcrt
| 9 | MS DHONI AGAINST HARSHAL PATEL IN IPL https://t.co/WFj0mgrVep
| 10 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
| 11 | RT @PureWinIndia: Today's Question!! 🔥
Who was the Orange Cap in the 2009 Edition of the #IPL? 🟠
Answer in the Comments 🙌
#PBKSvGT #GTvs…
| 12 | RT @FarziCricketer: Hardik Pandya has done it. He is no more worried after KKR dropped Venkatesh Iyer. #IPL
| 13 | @DelhiCapitals @akshar2026 C
| 14 | RT @PureWinIndia: Today's Question!! 🔥
Who was the Orange Cap in the 2009 Edition of the #IPL? 🟠

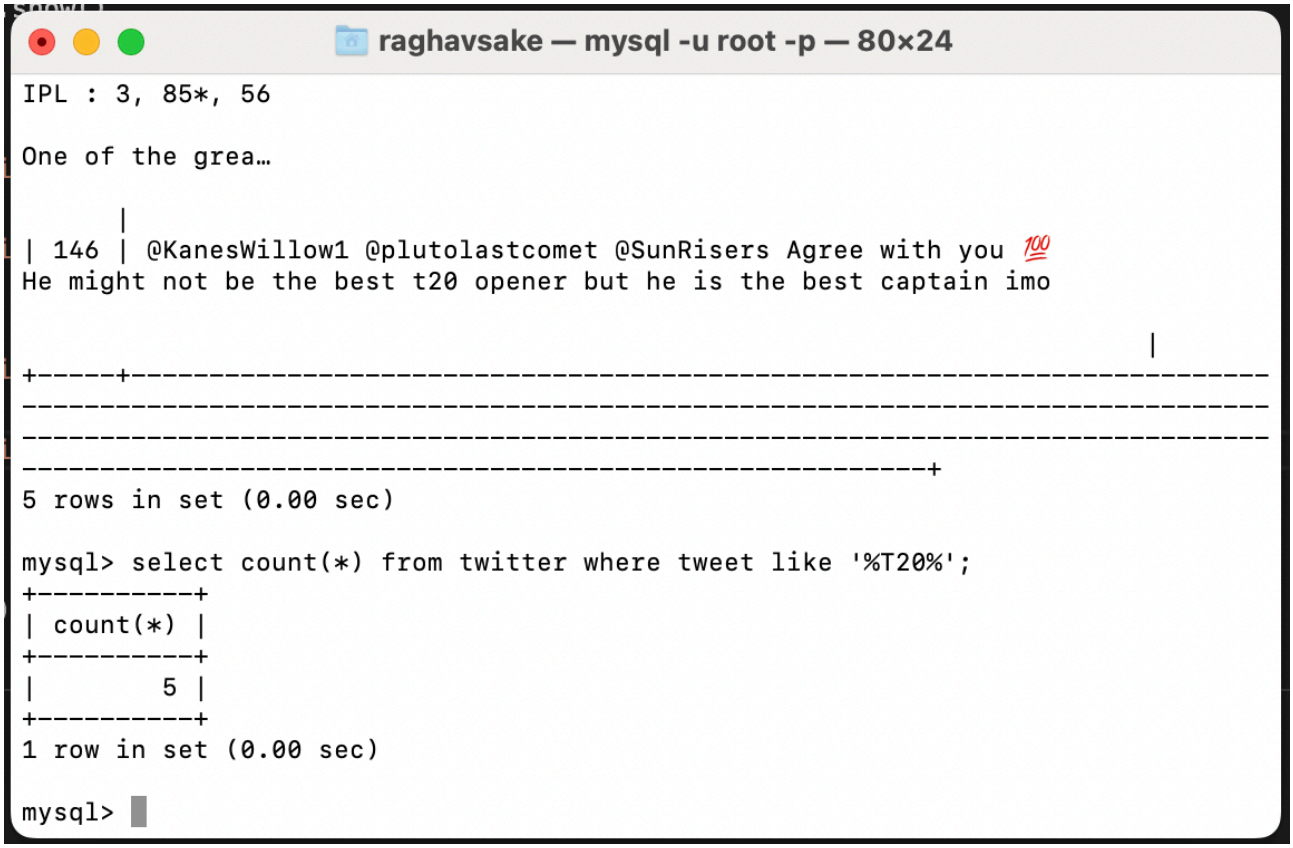

raghavsake — mysql -u root -p — 80×24
150 rows in set (0.01 sec)
mysql> select count(*) from twitter;
| count(*) |
| 150 |
1 row in set (0.00 sec)
mysql>

Query1 - The number of RTs which is basically the number of RTs

```
raghavsake — mysql -u root -p — 80×24
| 149 | RT @mufaddal_vohra: Most Man Of The Match Awards in the IPL:

AB De Villiers - 25.
Chris Gayle - 22.
David Warner - 18*.
Rohit Sharma - 18.…

                                                |
+------+------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-+
98 rows in set (0.01 sec)

mysql> select count(*) from twitter where tweet like '%RT%';
+----------+
| count(*) |
+----------+
|       98 |
+----------+
1 row in set (0.00 sec)

mysql>
```

```
mysql>
mysql> select * from twitter where tweet like '%RT%';
+-----+--------------------------------------------------------------------------
-----------------------------------------------------------------------+
| id  | tweet
                                                                       |
+-----+--------------------------------------------------------------------------
-----------------------------------------------------------------------+
|   1 | RT @AkramK2108: SR against spin in IPL 2022 (min. 50 balls),

Highest :

Sanju Samson - 173
Shubman Gill - 171
Abhishek Sharma - 169
Jos Bu…
|   2 | RT @mufaddal_vohra: IPL chairman said, "there is a slight dip in viewership, but I don't see them having any impact on the media rights sal…
                                                                       |
|   6 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
                                                                       |
|   7 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
                                                                       |
|   8 | RT @waadaplaya: 117m SIX by Liam Livingstone. This was quite a hit.

@liaml4893 #IPL2022 @IPL

 https://t.co/8fEy5UYcrt
|  10 | RT @CricCrazyJohns: Rabada has taken 93 wickets from 59 matches in IPL, already third in the leading wicket-taker in #IPL2022.
                                                                       |
|  11 | RT @PureWinIndia: Today's Question!! 🔥

Who was the Orange Cap in the 2009 Edition of the #IPL? 🟠

Answer in the Comments 🙌

#PBKSvGT #GTvs…
|  12 | RT @FarziCricketer: Hardik Pandya has done it. He is no more worried after KKR dropped Venkatesh Iyer. #IPL
                                                                       |
|  14 | RT @PureWinIndia: Today's Question!! 🔥

Who was the Orange Cap in the 2009 Edition of the #IPL? 🟠

Answer in the Comments 🙌

#PBKSvGT #GTvs…
|  16 | RT @VikramMeena_: IPL में ERCP
 वानखेड़े स्टेडियम में घनश्याम नागल भाई ने ERCP की ओर दर्शकों व मीडिया का ध्यान आकृष्ट किया। #ERCP_Nationa…
                     |
|  17 | RT @PureWinIndia: This Week's Winners of our Daily #contest are🏆

Sonali Bera
Sanjay Jain
Anuradha Karan
Jhum Jhum Manna
@SanjaySahni12
San…
|  18 | RT @CricCrazyJohns: Most 300+ runs season in IPL history:

Dhawan - 13
Rohit - 13
Kohli - 12
Raina - 12
|  20 | RT @PureWinIndia: Aaj kis ki hogi jeet? 🧐

#Gujarat ka Dhokla - Fafda? 😉🔵

 ya

#Punjab ke Chole Bhature? 🏏🔴

#IPL #IPL2022 #tuesdayvibe…
|  22 | RT @abhijitmajumder: Riots all over Rajasthan to mark the joyous occasion of #EidUlFitr. Jaipur, Jodhpur, Karauli, Nagaur, Alwar...
Maybe t…
|  23 | RT @abhijitmajumder: Riots all over Rajasthan to mark the joyous occasion of #EidUlFitr. Jaipur, Jodhpur, Karauli, Nagaur, Alwar...
Maybe t…
|  25 | RT @suraj_yadav2005: करोड़ों की त्रासदी : सुनने वाला कोई नहीं !
```

Query2 – The number of tweets containing the word “T20”



Comparison of Batch Mode Experiment vs Streaming Mode Experiment

The comparison of batch mode vs streaming mode for the following scenarios

- Query has to be done once on the twitter dataset

In this scenario the result of batch mode turns out to be much more efficient and faster. Since data is processed at once we save a lot of RAM memory and also is faster as the entire batch is processed at once. In Stream processing each part of the window is processed and stored and with the velocity of the data produced by the producer it makes it slower.

⇒Time for which data is produced – 15 minutes
⇒Time taken by Stream Processing Queries – 1 minute
⇒Time taken by Batch Processing Queries – 28 seconds

- Repeated Querying on the dataset
  In this scenario the result of the stream mode turns out to be much efficient and faster. The same data has to be queried and thus makes batch processing slower. The intermediate results are stored in stream processing and thus makes it faster

  ⇒ Time for which data is produced – 15 minutes
  ⇒ Time taken by Stream Processing Queries – 1 minute per query
  ⇒ Time taken by Batch Processing Queries – 2 minutes per query + time taken to store

Conclusion

Thus we can conclude that batch processing is useful when a single query has to be executed on entire batch at once and stream processing is better when queries are executed repeatedly on different sizes of data.