

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

1. Introduction
2. Installation of Software
3. Input Data
4. Streaming Mode Experiment
5. Batch mode Experiment
6. Comparison of Streaming vs Batch mode
Results
7. Conclusion

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

3.1.0 is the latest release. The current stable version is 3.1.0.

You can verify your download by following these [procedures](#) and using these [KEYS](#).

3.1.0

- Released January 24, 2022
- [Release Notes](#)
- Source download: [kafka-3.1.0-src.tgz \(asc, sha512\)](#)
- Binary downloads:
 - Scala 2.12 - [kafka_2.12-3.1.0.tgz \(asc, sha512\)](#)
 - Scala 2.13 - [kafka_2.13-3.1.0.tgz \(asc, sha512\)](#)

We build for multiple versions of Scala. This only matters if you are using Scala and you want a version built for the same Scala version you use. Otherwise any version should work (2.13 is recommended).

Kafka 3.1.0 includes a number of significant new features. Here is a summary of some notable changes:

- Apache Kafka supports Java 17
- The FetchRequest supports Topic IDs (KIP-516)
- Extend SASL/OAUTHBEARER with support for OIDC (KIP-768)
- Add broker count metrics (KIP-748)
- Differentiate consistently metric latency measured in millis and nanos (KIP-773)
- The eager rebalance protocol is deprecated (KAFKA-13439)
- Add TaskId field to StreamsException (KIP-783)
- Custom partitioners in foreign-key joins (KIP-775)
- Fetch/findSessions queries with open endpoints for SessionStore/WindowStore (KIP-766)
- Range queries with open endpoints (KIP-763)
- Add total blocked time metric to Streams (KIP-761)
- Add additional configuration to control MirrorMaker2 internal topics naming convention (KIP-690)

For more information, please read the detailed [Release Notes](#).

- Installing Spark

The screenshot shows the Apache Spark website at spark.apache.org. The 'Community' dropdown menu is open, displaying options such as 'Mailing Lists & Resources', 'Contributing to Spark', 'Improvement Proposals (SPIP)', 'Issue Tracker', 'Powered By', 'Project Committers', and 'Project History'. To the right of the main content area, there is a 'Latest News' section with links to recent releases: 'Spark 3.1.3 released (Feb 18, 2022)', 'Spark 3.2.1 released (Jan 26, 2022)', 'Spark 3.2.0 released (Oct 13, 2021)', and 'Spark 3.0.3 released (Jun 23, 2021)'. Below this is an 'Archive' link. On the far right, there is an 'APACHECON OCTOBER 3 - 6, 2022, NEW ORLEANS WWW.APACHECON.COM' logo and a large orange 'DOWNLOAD SPARK' button. The main content area includes sections for 'Download Apache Spark™', 'Link with Spark', 'Installing with PyPi', 'Convenience Docker Container Images', 'Release notes for stable releases' (listing Spark 3.2.1, 3.1.3, and 3.0.3), and 'Archived releases'.

- 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.

The screenshot shows the Twitter Developer Portal interface. The left sidebar is dark with white text, showing 'Developer Portal' at the top, followed by 'Dashboard', 'Projects & Apps' (with 'Overview' selected), 'Project 1' (selected), 'Test_Sargur', 'Products NEW', and 'Account'. The main content area has a light background. At the top, it says 'Project 1' with tabs for 'Overview' (selected) and 'Settings'. Below this, there are three sections: 'Access', 'Usage', and 'Apps'. The 'Access' section shows 'Elevated' status with '3 environments per project', '2M Tweets per month / Project', and 'free' cost. The 'Usage' section shows 'MONTHLY TWEET CAP USAGE' with a progress bar at 0% (2,575 Tweets pulled of 20,00,000) that resets on June 3 at 00:00 UTC. The 'Apps' section shows a 'DEVELOPMENT APP' entry with a 'Manage' button. On the right side, there's a sidebar titled 'Helpful docs' with links to 'About Projects', 'About Apps', 'About authentication', 'About Tweet caps', and 'Authentication best practices'. Below this is a dark callout box with 'Quick info on App environments' and an 'Expand' button. At the bottom of the page, there are links for 'PRIVACY', 'COOKIES', 'TWITTER TERMS & CONDITIONS', 'DEVELOPER POLICY & TERMS', '© 2022 TWITTER INC.', 'FOLLOW @TWITTERDEV', 'SUBSCRIBE TO DEVELOPER NEWS', and a small dropdown menu.

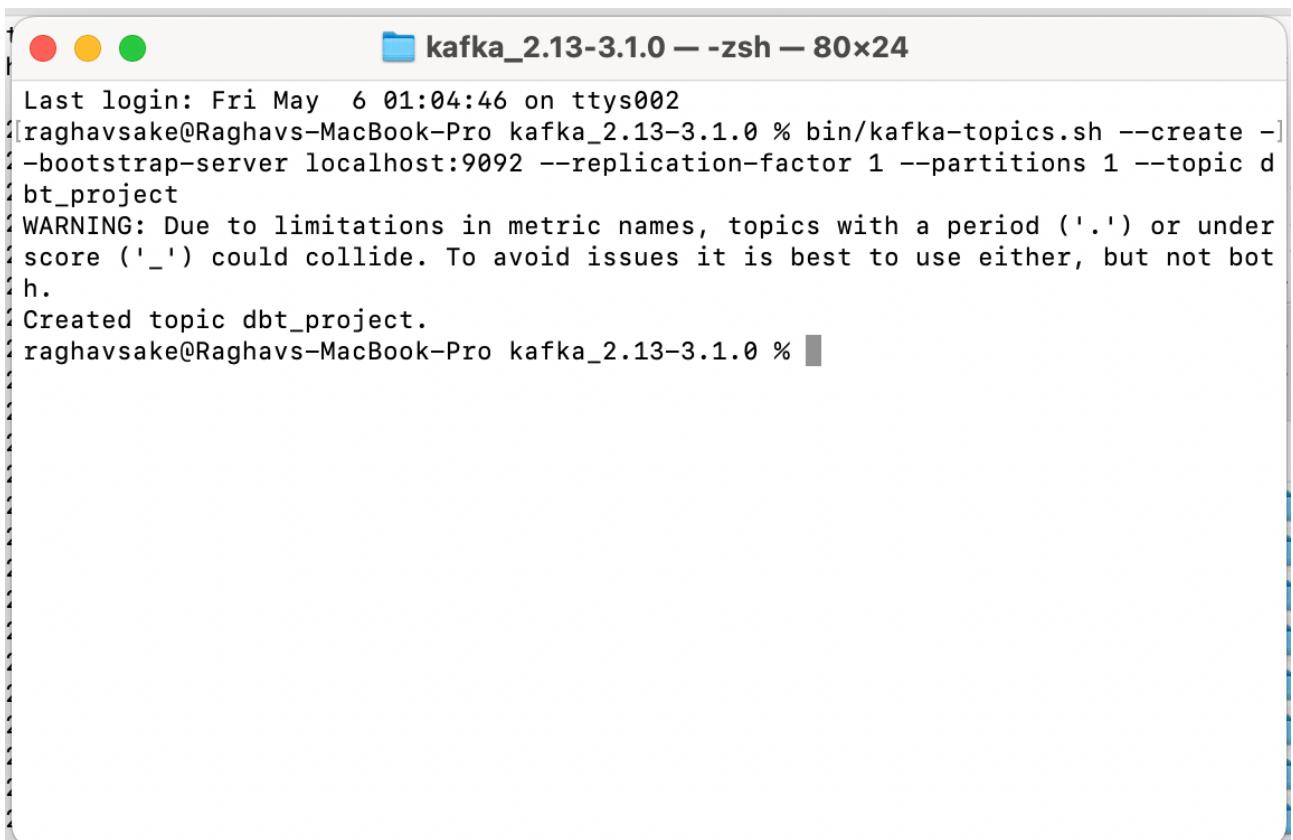
Streaming Mode Experiment

Starting ZooKeeper

Starting Kafka Server

```
Last login: Fri May 6 04:58:46 on ttv-0888
aghavashankar@MacBook-Pro:kafka-2.13-3.1.0$ bin/kafka-server-start.sh config/server.properties
[2022-05-06 01:04:26,293] INFO Registered kafka:type=Controller,MBean [kafka.utils.LogControllerRegistration$]
[2022-05-06 01:04:26,538] INFO Setting -D jdk.iiis.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2022-05-06 01:04:26,646] INFO Using default signal handler for: TEMP, INT, HUP (org.apache.kafka.common.LoggingSignalHandler)
[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 01:04:26,767] INFO Client environment:zookeeper.version=2.13.1.0-alpha-14619703-2022-05-06-01:04:26-2d9f72e8c668, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.vendor=OpenJDK OpenJDK JRE (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.vendor.version=11.0.11+11-b2043.9-11247-20220506-01:04:26-2d9f72e8c668 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.version=1.8.0_261-b14 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.home=/Library/Java/JavaVirtualMachines/openjdk-11.0.11-fcs/lib/jre (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.library.path=/Library/Java/JavaVirtualMachines/openjdk-11.0.11-fcs/jdk/lib/amd64 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.io.tmpdir=/Library/Java/JavaVirtualMachines/openjdk-11.0.11-fcs/jdk/tmp (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:java.compiler=NONE (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,767] INFO Client environment:os.name=Mac OS X (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:os.name=Mac OS X (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:os.arch=x86_64 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:os.version=10.16 (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:user.name=aghavashankar (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:user.dir=/Users/aghavashankar (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:memory_free=281M (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:memory_total=1024M (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Client environment:memroy_total=1024M (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Initiating initial connection, connecting to localhost:2181 sessionTimeout=18000 watcher=kafka.zookeeper.ZooKeeperClient$ZooKeeperClientWatcher@91153bc (org.apache.zookeeper.ZooKeeper)
[2022-05-06 01:04:26,768] INFO Jute buffer value is 1494304 Bytes (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,770] INFO Jute buffer value is 1494304 Bytes (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,770] INFO [ZooKeeperClient Kafka server] Waiting until connected. (kafka.zookeeper.ZooKeeperClient)
[2022-05-06 01:04:26,771] INFO Open socket connection to server localhost/127.0.0.1:2181. (org.apache.zookeeper.ClientCnxn)
[2022-05-06 01:04:26,771] INFO Upgrading connection from existing connection to latest FinalizedFeaturesAndEpoch(features=Features(), epoch=0). (kafka.server.FinalizedFeatureCache)
[2022-05-06 01:04:27,112] INFO Cluster ID = 7YeFuFretA@jmh3kEdg (kafka.server.KafkaServer)
[2022-05-06 01:04:27,181] INFO KafkaConfig values:
advertised.host.name null
advertised.port null
advertised.protocol.name = null
alter.log.dir.replication.quota.window.size.seconds = 11
alter.log.dir.replication.quota.window.size.seconds = 1
advertised.compression.type = none
auto.create.topics.enable = true
auto.leader.rebalance.enable = true
background.threads = 10
...
```

Creating a Kafka Topic



A screenshot of a macOS terminal window titled "kafka_2.13-3.1.0 -- zsh -- 80x24". The window shows the command "bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic dbt_project" being run. It includes a warning about metric name collisions and confirms the creation of the topic "dbt_project".

```
Last login: Fri May  6 01:04:46 on ttys002
[raghavsake@Raghavs-MacBook-Pro kafka_2.13-3.1.0 % bin/kafka-topics.sh --create --
--bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic dbt_
project
WARNING: Due to limitations in metric names, topics with a period ('.') or under
score ('_') could collide. To avoid issues it is best to use either, but not bot
h.
Created topic dbt_project.
raghavsake@Raghavs-MacBook-Pro kafka_2.13-3.1.0 %
```

Streaming data into kafka topic

```
producer.py — DBT Project
```

```
producer.py
1 import requests
2 import os
3 import json
4 from kafka import KafkaProducer
5 import mysql.connector
6
7 mydb = mysql.connector.connect(
8     host="localhost",
9     user="root",
10    password="snowwhite",
11    database="twitter"
12 )
13
14 producer = KafkaProducer(bootstrap_servers='localhost:9092', value_serializer=lambda v: json.dumps(v).encode('utf-8'))
15
16 topic_name = "dbt_project"
17
18 bearer_token = "AAAAAAAAAAAAAAAAdq3cAEAAAAtnrY6hZUo7ooB26e41n%2BN2X33A%3D7c8cLd1ClRc1hP0Ne0rIaxYAfTvRNPsYsVyQawPaw9QNkqfSy"
19
20
21 def bearer_oauth(r):
22     r.headers["Authorization"] = f"Bearer {bearer_token}"
23     r.headers["User-Agent"] = "IPLTwitterStreamer"
24     return r
25
26
27 def get_rules():
28     response = requests.get(
29         "https://api.twitter.com/2/tweets/search/stream/rules", auth=bearer_oauth
30     )
31     if response.status_code != 200:
32         raise Exception(
33             "Cannot get rules (HTTP {}): {}".format(response.status_code, response.text)
34         )
35     print(json.dumps(response.json()))

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
python producer.py
("data": {"id": "1522299379451532898", "value": "IPL", "tag": "IPL"}, "meta": {"sent": "2022-05-05T19:37:18.055Z", "result_count": 1})
("meta": {"sent": "2022-05-05T19:37:18.077Z", "summary": {"deleted": 1, "not_deleted": 0}})
("data": {"value": "IPL", "tag": "IPL", "id": "152229941958848256"}, "meta": {"sent": "2022-05-05T19:37:19.497Z", "summary": {"created": 1, "not_created": 0, "valid": 1, "invalid": 0}})
200
Option C: Meredith's best figures are 1/21 in IPL 2022 @USHACOOK18 @UshaIntl @UshaPlay https://t.co/NhzdBV5q2b
{
    "id": "1522299398190186496",
    "text": "Option C:@Usha0@Meredith @Usha2019s best figures are 1/21 in IPL 2022 @USHACOOK18 @UshaIntl @UshaPlay https://t.co/NhzdBV5q2b"
}
```

Ln 16, Col 26 Spaces: 4 UTF-8 LF Python

master ⌂ 0 0 0

Fri 6 May 1:07 AM

```
producer.py — DBT Project
```

```
producer.py
1 import requests
2 import os
3 import json
4 from kafka import KafkaProducer
5 import mysql.connector
6
7 mydb = mysql.connector.connect(
8     host="localhost",
9     user="root",
10    password="snowwhite",
11    database="twitter"
12 )
13
14 producer = KafkaProducer(bootstrap_servers='localhost:9092', value_serializer=lambda v: json.dumps(v).encode('utf-8'))
15
16 topic_name = "dbt_project"
17
18 bearer_token = "AAAAAAAAAAAAAAAAdq3cAEAAAAtnrY6hZUo7ooB26e41n%2BN2X33A%3D7c8cLd1ClRc1hP0Ne0rIaxYAfTvRNPsYsVyQawPaw9QNkqfSy"
19
20
21 def bearer_oauth(r):
22     r.headers["Authorization"] = f"Bearer {bearer_token}"
23     r.headers["User-Agent"] = "IPLTwitterStreamer"
24     return r
25
26
27 def get_rules():
28     response = requests.get(
29         "https://api.twitter.com/2/tweets/search/stream/rules", auth=bearer_oauth
30     )
31     if response.status_code != 200:
32         raise Exception(
33             "Cannot get rules (HTTP {}): {}".format(response.status_code, response.text)
34         )
35     print(json.dumps(response.json()))

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

```
python producer.py
@kurikratsgill Why more SMAT team coaches are not working with IPL teams? These are the coaches who know in and out about the Indian players and their capabilities but hardly anyone is employed by IPL teams. How much does Shane Watson know abt local players compared to SMAT coaches?
{
    "id": "1522299429943873536",
    "text": "@kurikratsgill Why more SMAT team coaches are not working with IPL teams? These are the coaches who know in and out about the Indian players and their capabilities but hardly anyone is employed by ITC teams. How much does Shane Watson know abt local players compared to SMAT coaches?"
}
On This Day 13 Years Ago In 2009, Ro Got A Hat-Trick In The IPL.
@RmRo45 | #RohitSharma https://t.co/1kRc04czan
{
    "id": "1522299434549510144",
    "text": "On This Day 13 Years Ago In 2009, Ro Got A Hat-Trick In The IPL. \n\n@RmRo45 | #RohitSharma@ud80c\udcf5 https://t.co/1kRc04czan"
}
@indiakabbu_@LessthanZero_@RaviShastriOfc@MdShami11@mdsirajofficial Twitter ka baap parag hau malum hai bam fodiy
{
    "id": "1522299473191313408",
    "text": "@indiakabbu_@LessthanZero_@RaviShastriOfc@MdShami11@mdsirajofficial Twitter ka baap parag hau malum hai bam fodiy"
}
RT @IPL: A look at the Points Table after Match No. 58 of the #TATAIPL 2022 #DCvSRH https://t.co/2L2ZeGrg58
{
    "id": "1522299496008273926",
    "text": "RT @IPL: A look at the Points Table after Match No. 58 of the #TATAIPL 2022 \ud83d\udcf3 #DCvSRH https://t.co/2L2ZeGrg58"
}
```

Ln 16, Col 26 Spaces: 4 UTF-8 LF Python

master ⌂ 0 0 0

Fri 6 May 1:07 AM

PES2UG19CS219 – Manjunath Y

PES2UG19CS249 - Naveen S Nelagal

PES2UG19CS312 – Raghav S K

PES2UG19CS461 – Vivek S D

Real-Time Twitter IPL Data Streaming

Page 10 / 20

Subscribing to Kafka Topic and Streaming

```
Last login: Fri May  6 01:05:16 on ttys002
[raghavsake@Raghavs-MacBook-Pro spark-3.2.1-bin-hadoop3.2 % bin/spark-submit --ma]
ster local --class "org.myspark.KafkaStream" --packages org.apache.spark:spark-s
ql-kafka-0-10_2.12:3.0.0 '/Users/raghavsake/Desktop/DBT Project/consumer.py'
22/05/06 01:08:54 WARN Utils: Your hostname, Raghavs-MacBook-Pro.local resolves
to a loopback address: 127.0.0.1; using 192.168.0.124 instead (on interface en0)
22/05/06 01:08:54 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
:: loading settings :: url = jar:file:/Users/raghavsake/Downloads/spark-3.2.1-bi
n-hadoop3.2/jars/ivy-2.5.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /Users/raghavsake/.ivy2/cache
The jars for the packages stored in: /Users/raghavsake/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-4b809adc-6a44-
4195-9266-cad5814dcfdf;1.0
  confs: [default]
    found org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.0 in central
    found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.0 in cen
tral
    found org.apache.kafka#kafka-clients;2.4.1 in central
    found com.github.luben#zstd-jni;1.4.4-3 in central
    found org.lz4#lz4-java;1.7.1 in central
    found org.xerial.snappy#snappy-java;1.1.7.5 in central
    found org.slf4j#slf4j-api;1.7.30 in local-m2-cache
```

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 — 80x24
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 @ufaddal_vohra: IPL chairman said, "there is a slight dip in viewership, but I don't see them having any impact on the media rights sale.."
| 3 | IPL : अचे लोटे 2 अद्य तुम गोपनीय https://t.co/jtyXpBVbhU
| 4 | काम, जरक के रिकॉर्ड पाटीज़ 2 लाख लोटे और दोनों जो पर बहा आय,
IPL में लोटी जो परेंगे वे लोटी जो नहीं नहीं नहीं क्यों?
इसले लोटी जो अद्य लोटी जो जीत हाल है क्या 😊
मार्टिन के CBI जाच होनी याइकी
| 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/8fEy6UVcrt
| 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 #Gtv...
| 12 | RT @FarziCricketter: 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? 🍀

```

PES2UG19CS219 – Manjunath Y

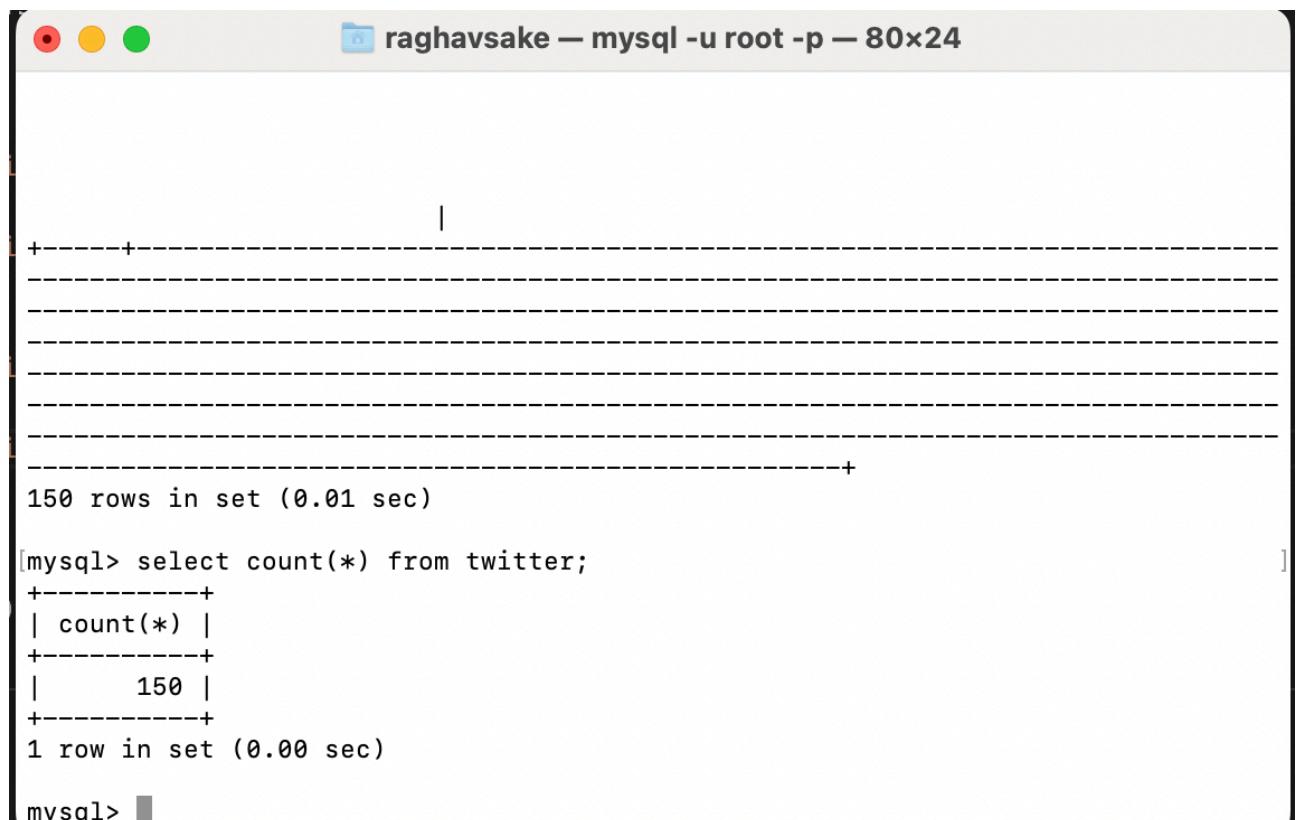
PES2UG19CS249 - Naveen S Nelagal

PES2UG19CS312 – Raghav S K

PES2UG19CS461 – Vivek S D

Real-Time Twitter IPL Data Streaming

Page 16 / 20



```
+-----+  
|  
+-----+  
|  
+-----+  
|  
+-----+  
|  
+-----+  
|  
+-----+  
| 150 |  
+-----+  
1 row in set (0.00 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

```
| 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> 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  
Rohit Sharma - 169  
Jas But...  
| 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 @waadeplaya: 117m SIX by Liam Livingstone. This was quite a hit.  
Liam1689 #IPL2022 #IPL  
https://t.co/f6f5UyZrt  
| 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 @FarziCricketeer: Hardik Pandya has done it. He is no more worried after KKR dropped Venkatesh Iyer. #IPL  
| 13 | 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 @GauravMeena: IPL # ERCP के लिए दर्शक ने जारी कर दी थी वीडियो का अनुवाद फिल्मी। #ERCP_National...  
| 17 | RT @PureWinIndia: This Week's Winners of our Daily #contest are! 🎉  
Sonali Bera  
Sanjay Jain  
Anuradha Karan  
Jhum Jhum Menna  
Neha JaySahni12  
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? 🍞-paneer  
#Punjab ke Chole Bhature? 🍞-paneer  
#Punjab ke Chole Bhature? 🍞-paneer  
#IPL #IPL2022 #usosdayive...  
| 22 | RT @abnijitmejunder: Riots all over Rajasthan to mark the joyous occasion of #EidUlfitr. Jaipur, Jodhpur, Karauli, Nagaur, Alwar...  
Maybe t...  
| 23 | RT @abnijitmejunder: Riots all over Rajasthan to mark the joyous occasion of #EidUlfitr. Jaipur, Jodhpur, Karauli, Nagaur, Alwar...  
More...  
| 25 | RT @susai_yadav2005: काठो के गाली : सुप्रे गाल बेड़ नहीं !
```

PES2UG19CS219 – Manjunath Y

PES2UG19CS249 - Naveen S Nellogal

PES2UG19CS312 – Raghav S K

PES2UG19CS542 – Raghu S R
PES2UG19CS461 – Vivek S D

Real-Time Twitter IPL Data Streaming

Page 18 / 20

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

```
IPL : 3, 85*, 56
One of the grea...
| 146 | @KanesWillow1 @plutolastcomet @SunRisers Agree with you 100
He might not be the best t20 opener but he is the best captain imo
+-----+
-----+
-----+
-----+
5 rows in set (0.00 sec)

mysql> select count(*) from twitter where tweet like '%T20%';
+-----+
| count(*) |
+-----+
|      5   |
+-----+
1 row in set (0.00 sec)

mysql>
mysql> select * from twitter where tweet like '%T20%';
+-----+
| id | tweet           |
+-----+
| 46 | RT @RandomCricketPl: One of the classiest Test openers for India made T20 batting look beautiful when on song. This century was such a deli...
| 55 | RT @RageFanSocial: It's now time for match number 48 in the Indian T20 cricket tournament. In this game #PBKS will be locking horns against...
| 138 | @NMAH@17234875 @Sanju1_ @mufaddal_vohra It should be talked more first of all williamson apart from 2018 ipl hasnt done anything great t20 cricket not in ipl or even he is shit in intl.. and he is captaining a side in ipl.. an indian fra...
nchise captained by underperforming overseas captain |
| 145 | RT @thericabas: Devon Conway's first 3 scores in
Test : 200, 23, 88
ODI : 27, 72, 126
T20I : 41, 65*, 5
IPL : 3, 85*, 56
One of the grea...
| 146 | @KanesWillow1 @plutolastcomet @SunRisers Agree with you 99
He might not be the best t20 opener but he is the best captain imo
+-----+
-----+
-----+
-----+
5 rows in set (0.00 sec)
mysql>
```

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.