PES UNIVERSITY

100 feet Ring Road, BSK 3rd Stage
Bengaluru 560085

Department of Computer Science and Engineering
B. Tech. CSE - 6th Semester
Jan – May 2023

UE20CS343

DATABASE TECHNOLOGIES (DBT)

Project Report

Performing Stream Processing and Batch Processing on tweets

TEAM #: 5

PES1UG20CS286: Pragathya Baskar
PES1UG20CS045: Ananya Prakash
PES1UG21CS810: Divya M
PES1UG20CS265: Nidhi R Jois

Class of Prof. K S Srinivas

Pragathya
Ananya
Divya
Nidhi

Table of Contents

1. Introduction

- The task involves the use of various technologies and frameworks to process streaming data and run batch queries on the same data. The aim is to compare the performance and accuracy of processing the data in both streaming and batch modes.
- Apache Spark Streaming and Spark SQL will be used to execute multiple workloads on the input data. These workloads will include Spark SQL queries to perform actions, transformations, and aggregations on the input data.
- Apache Kafka Streaming will be used to publish and subscribe to the results or produce and consume from three or more topics. The data will be stored in a DBMS of choice such as Postgres or MySQL.
- The input data for the computation examples will be a streaming data source such as Twitter feed (tweets).

2.Installation of Software [include version #s and URLs]

➔Streaming Tools Used

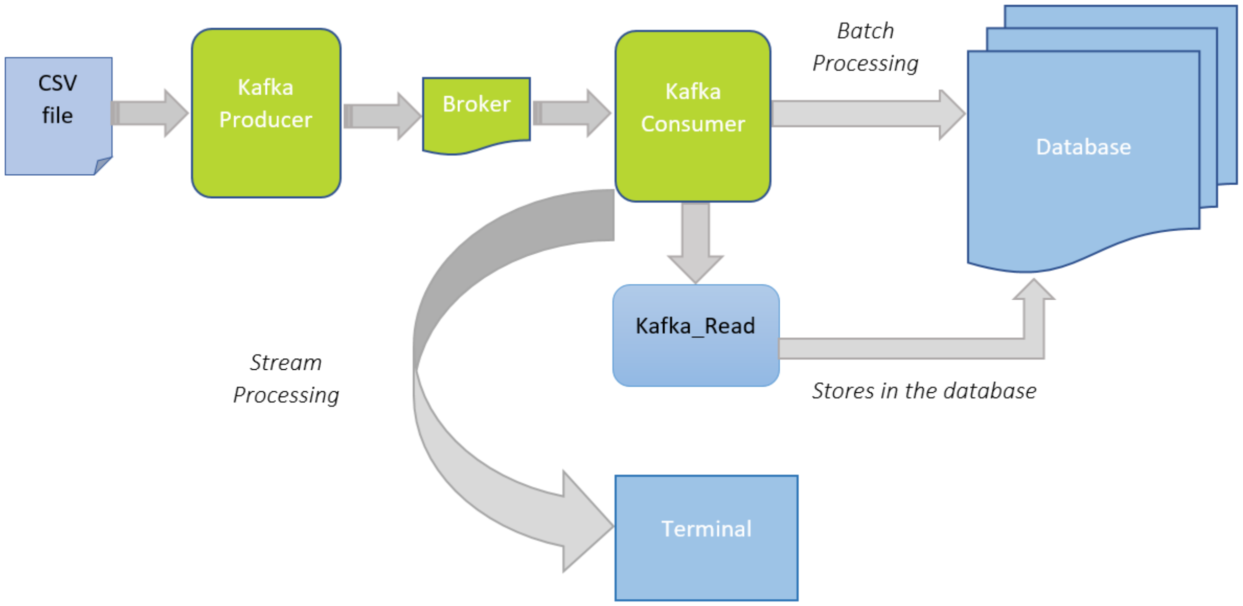
- Apache Spark Streaming:
  - Version # : Spark 3.4.0
  - URL: https://spark.apache.org/downloads.html
- Apache Kafka Streaming:
  - Version #: Kafka 3.4.0
  - URL: https://kafka.apache.org/downloads

➔DBMS Used:
- ◆MySQL database

3.Problem Description

- Apache Spark Streaming and Spark SQL is used to execute multiple workloads on the input data.
- These workloads will include Spark SQL queries to perform actions, transformations, and aggregations on the input data.
- Apache Kafka Streaming will be used to publish and subscribe to the results or produce and consume from three or more topics.
- The data will be stored in the MySQL database.

4.Architecture Diagram



5. Input Data

- Source
  - Kaggle dataset
- Description
  - The dataset containing tweets is taken from a producer which publishes it to kafka as topic and value.
  - Then kafka will store the data in MySQL database after doing some preprocessing.
  - Then the consumer will subscribe to a topic and perform stream and batch processing.

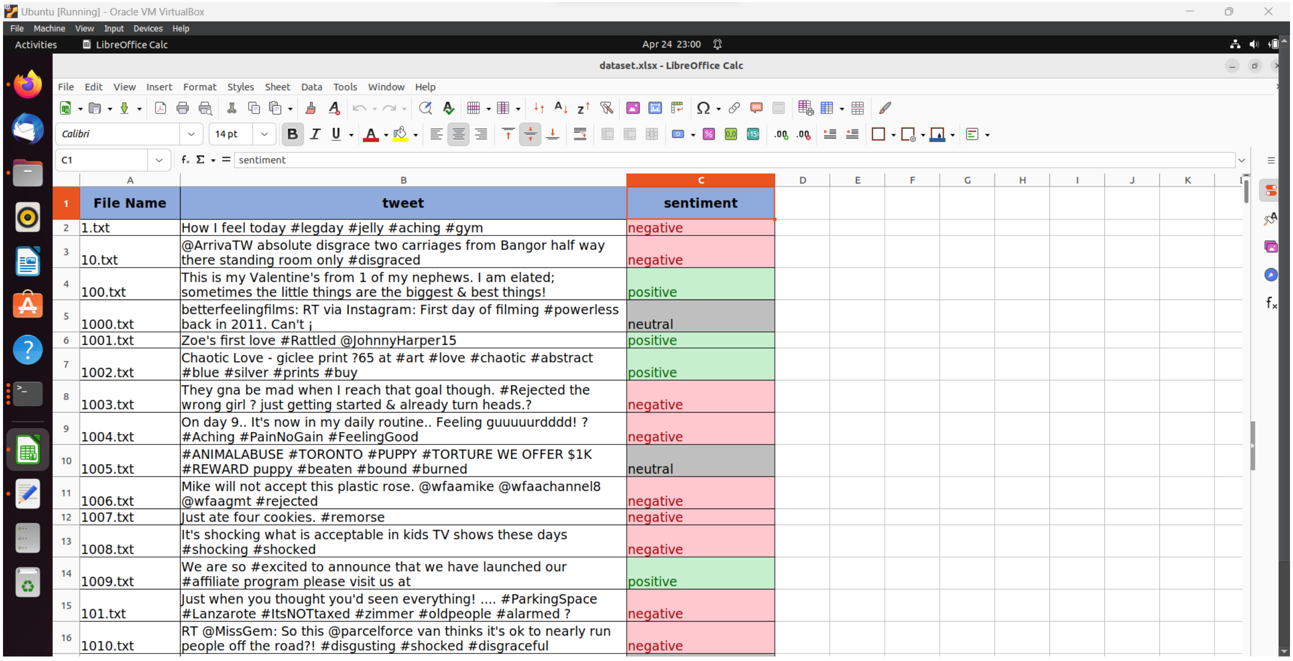
6. Streaming Mode Experiment

- Windows: Ubuntu
- The type of window used here: tumbling window
- Workloads:
  - For Spark, data processing tasks such as batch processing and streaming processing.
- Code like SQL Scripts : Spark SQL and Pyspark
  - Spark SQL code used in consumer_batch.py and consumer_stream.py
  - positive_df = spark.sql("SELECT * FROM tweets WHERE topic = 'positive'")

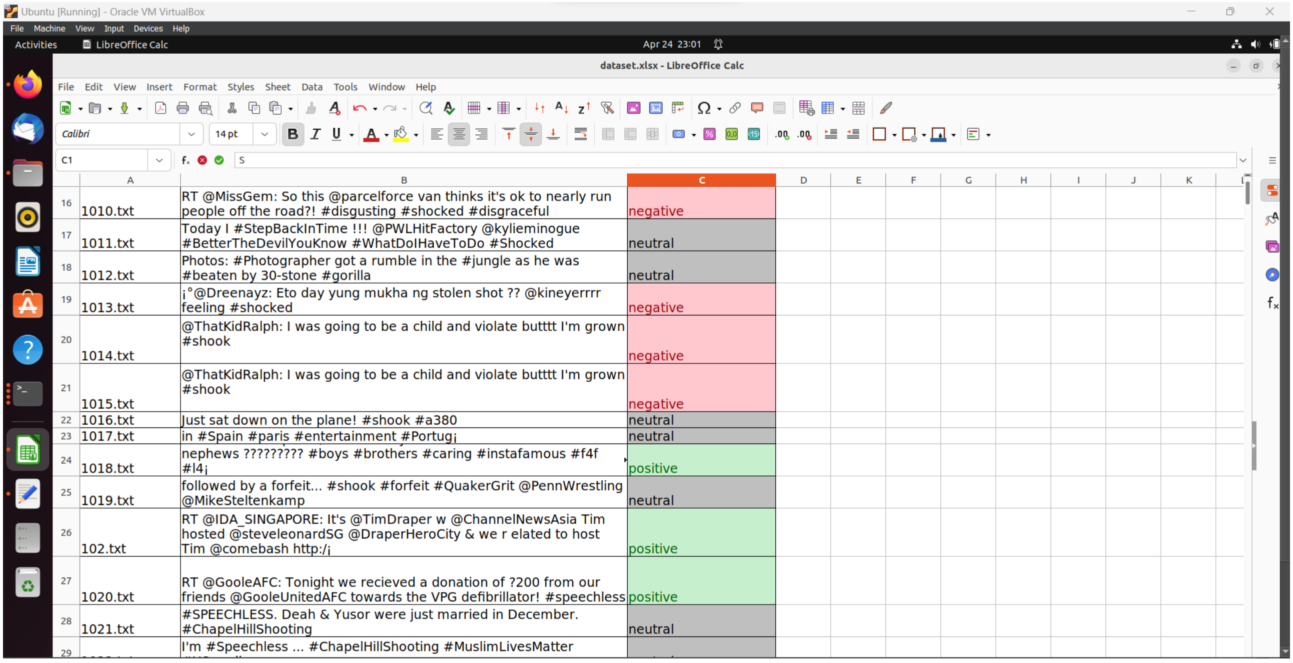
- Inputs and Corresponding Results
  - Input is a twitter dataset from which the producer read and publishes the topics to the kafka broker
  - The result is selecting all tweets of topic 'positive' and then counting the number of hashtags in the tweet

Input screenShot:

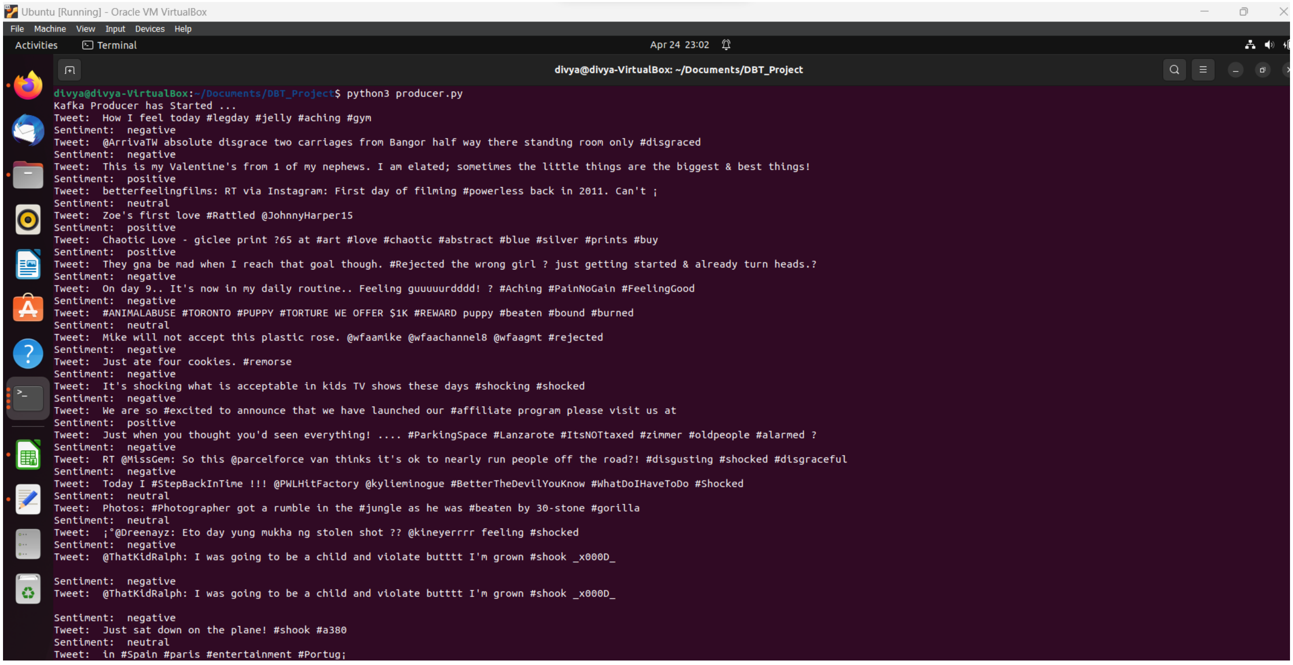
Dataset.xlsx


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
LibreOffice Calc
Apr 24 23:00
dataset.xlsx - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Calibri
14 pt
C1
sentiment
A B C D E F G H I J K

	File Name	tweet	sentiment
2	1.txt	How I feel today #legday #jelly #aching #gym	negative
3	10.txt	@ArrivaTW absolute disgrace two carriages from Bangor half way there standing room only #disgraced	negative
4	100.txt	This is my Valentine's from 1 of my nephews. I am elated; sometimes the little things are the biggest & best things!	positive
5	1000.txt	betterfeelingfilms: RT via Instagram: First day of filming #powerless back in 2011. Can't ¡	neutral
6	1001.txt	Zoe's first love #Rattled @JohnnyHarper15	positive
7	1002.txt	Chaotic Love - giclee print ?65 at #art #love #chaotic #abstract #blue #silver #prints #buy	positive
8	1003.txt	They gna be mad when I reach that goal though. #Rejected the wrong girl ? just getting started & already turn heads.?	negative
9	1004.txt	On day 9.. It's now in my daily routine.. Feeling guuuuurdddd! ? #Aching #PainNoGain #FeelingGood	negative
10	1005.txt	#ANIMALABUSE #TORONTO #PUPPY #TORTURE WE OFFER $1K #REWARD puppy #beaten #bound #burned	neutral
11	1006.txt	Mike will not accept this plastic rose. @wfaamike @wfaachannel8 @wfaagmt #rejected	negative
12	1007.txt	Just ate four cookies. #remorse	negative
13	1008.txt	It's shocking what is acceptable in kids TV shows these days #shocking #shocked	negative
14	1009.txt	We are so #excited to announce that we have launched our #affiliate program please visit us at	positive
15	101.txt	Just when you thought you'd seen everything! .... #ParkingSpace #Lanzarote #ItsNOTtaxed #zimmer #oldpeople #alarmed ?	negative
16	1010.txt	RT @MissGem: So this @parcelforce van thinks it's ok to nearly run people off the road?! #disgusting #shocked #disgraceful	negative


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
LibreOffice Calc
Apr 24 23:01
dataset.xlsx - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Calibri
14 pt
C1
S
A B C D E F G H I J K

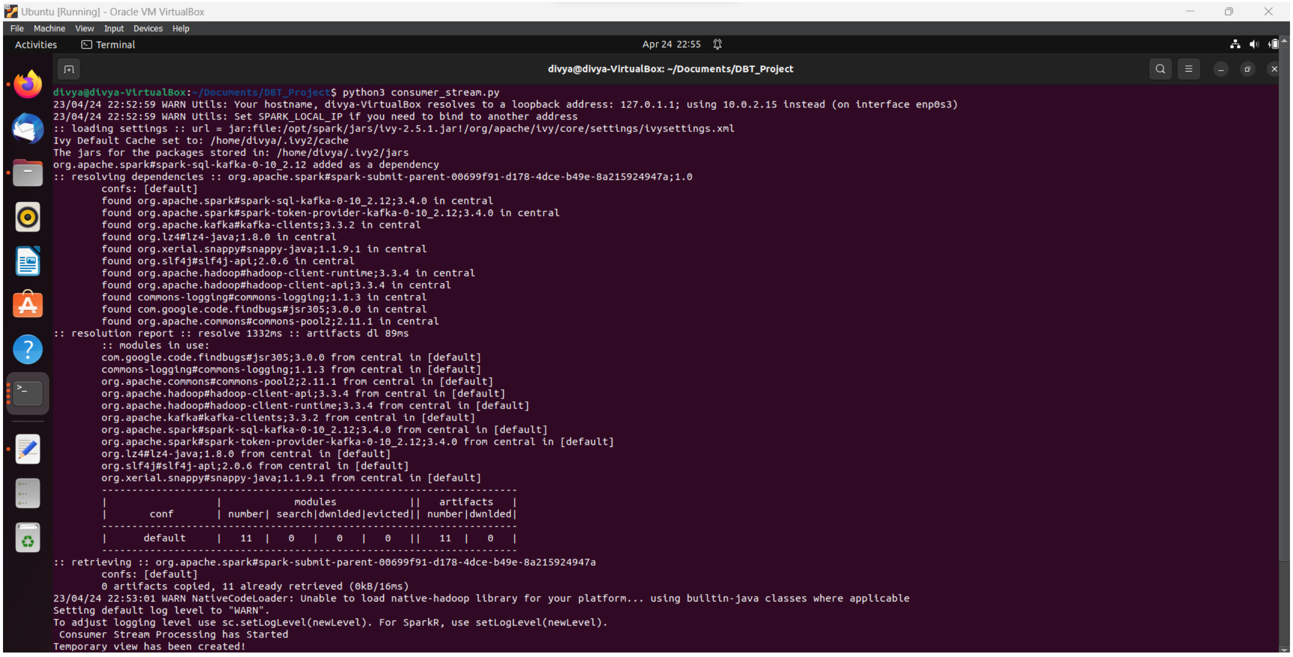
	A	B	C
16	1010.txt	RT @MissGem: So this @parcelforce van thinks it's ok to nearly run people off the road?! #disgusting #shocked #disgraceful	negative
17	1011.txt	Today I #StepBackInTime !!! @PWLHitFactory @kylieminogue #BetterTheDevilYouKnow #WhatDoIHaveToDo #Shocked	neutral
18	1012.txt	Photos: #Photographer got a rumble in the #jungle as he was #beaten by 30-stone #gorilla	neutral
19	1013.txt	¡°@Dreenayz: Eto day yung mukha ng stolen shot ?? @kineyerrrr feeling #shocked	negative
20	1014.txt	@ThatKidRalph: I was going to be a child and violate butttt I'm grown #shook	negative
21	1015.txt	@ThatKidRalph: I was going to be a child and violate butttt I'm grown #shook	negative
22	1016.txt	Just sat down on the plane! #shook #a380	neutral
23	1017.txt	in #Spain #paris #entertainment #Portug¡	neutral
24	1018.txt	nephews ????????? #boys #brothers #caring #instafamous #f4f #l4¡	positive
25	1019.txt	followed by a forfeit... #shook #forfeit #QuakerGrit @PennWrestling @MikeSteltenkamp	neutral
26	102.txt	RT @IDA_SINGAPORE: It's @TimDraper w @ChannelNewsAsia Tim hosted @steveleonardSG @DraperHeroCity & we r elated to host Tim @comebash http:/¡	positive
27	1020.txt	RT @GooleAFC: Tonight we recieved a donation of ?200 from our friends @GooleUnitedAFC towards the VPG defibrillator! #speechless	positive
28	1021.txt	#SPEECHLESS. Deah & Yusor were just married in December. #ChapelHillShooting	neutral
29		I'm #Speechless ... #ChapelHillShooting #MuslimLivesMatter	


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Apr 24 23:02
divya@divya-VirtualBox: ~/Documents/DBT_Project

```
divya@divya-VirtualBox:~/Documents/DBT_Project$ python3 producer.py
Kafka Producer has Started ...
Tweet:  How I feel today #legday #jelly #aching #gym
Sentiment:  negative
Tweet:  @ArrivaTW absolute disgrace two carriages from Bangor half way there standing room only #disgraced
Sentiment:  negative
Tweet:  This is my Valentine's from 1 of my nephews. I am elated; sometimes the little things are the biggest & best things!
Sentiment:  positive
Tweet:  betterfeelingfilms: RT via Instagram: First day of filming #powerless back in 2011. Can't ¡
Sentiment:  neutral
Tweet:  Zoe's first love #Rattled @JohnnyHarper15
Sentiment:  positive
Tweet:  Chaotic Love - giclee print ?65 at #art #love #chaotic #abstract #blue #silver #prints #buy
Sentiment:  positive
Tweet:  They gna be mad when I reach that goal though. #Rejected the wrong girl ? just getting started & already turn heads.?
Sentiment:  negative
Tweet:  On day 9.. It's now in my daily routine.. Feeling guuuuurdddd! ? #Aching #PainNoGain #FeelingGood
Sentiment:  negative
Tweet:  #ANIMALABUSE #TORONTO #PUPPY #TORTURE WE OFFER $1K #REWARD puppy #beaten #bound #burned
Sentiment:  neutral
Tweet:  Mike will not accept this plastic rose. @wfaamike @wfaachannel8 @wfaagmt #rejected
Sentiment:  negative
Tweet:  Just ate four cookies. #remorse
Sentiment:  negative
Tweet:  It's shocking what is acceptable in kids TV shows these days #shocking #shocked
Sentiment:  negative
Tweet:  We are so #excited to announce that we have launched our #affiliate program please visit us at
Sentiment:  positive
Tweet:  Just when you thought you'd seen everything! .... #ParkingSpace #Lanzarote #ItsNOTtaxed #zimmer #oldpeople #alarmed ?
Sentiment:  negative
Tweet:  RT @MissGem: So this @parcelforce van thinks it's ok to nearly run people off the road?! #disgusting #shocked #disgraceful
Sentiment:  negative
Tweet:  Today I #StepBackInTime !!! @PWLHitFactory @kylieminogue #BetterTheDevilYouKnow #WhatDoIHaveToDo #Shocked
Sentiment:  neutral
Tweet:  Photos: #Photographer got a rumble in the #jungle as he was #beaten by 30-stone #gorilla
Sentiment:  neutral
Tweet:  ¡°@Dreenayz: Eto day yung mukha ng stolen shot ?? @kineyerrrr feeling #shocked
Sentiment:  negative
Tweet:  @ThatKidRalph: I was going to be a child and violate butttt I'm grown #shook _x000D_

Sentiment:  negative
Tweet:  @ThatKidRalph: I was going to be a child and violate butttt I'm grown #shook _x000D_

Sentiment:  negative
Tweet:  Just sat down on the plane! #shook #a380
Sentiment:  neutral
Tweet:  in #Spain #paris #entertainment #Portug¡
```

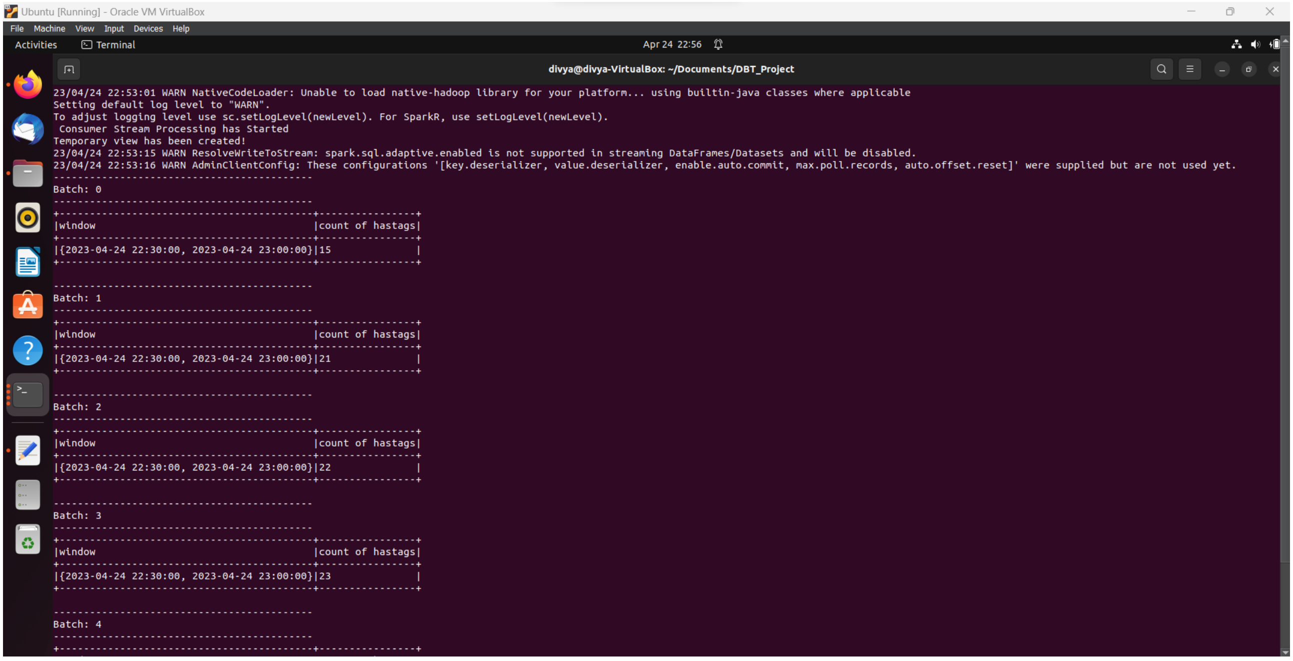
Output Screenshot:

Pragathya
Ananya
Divya
Nidhi



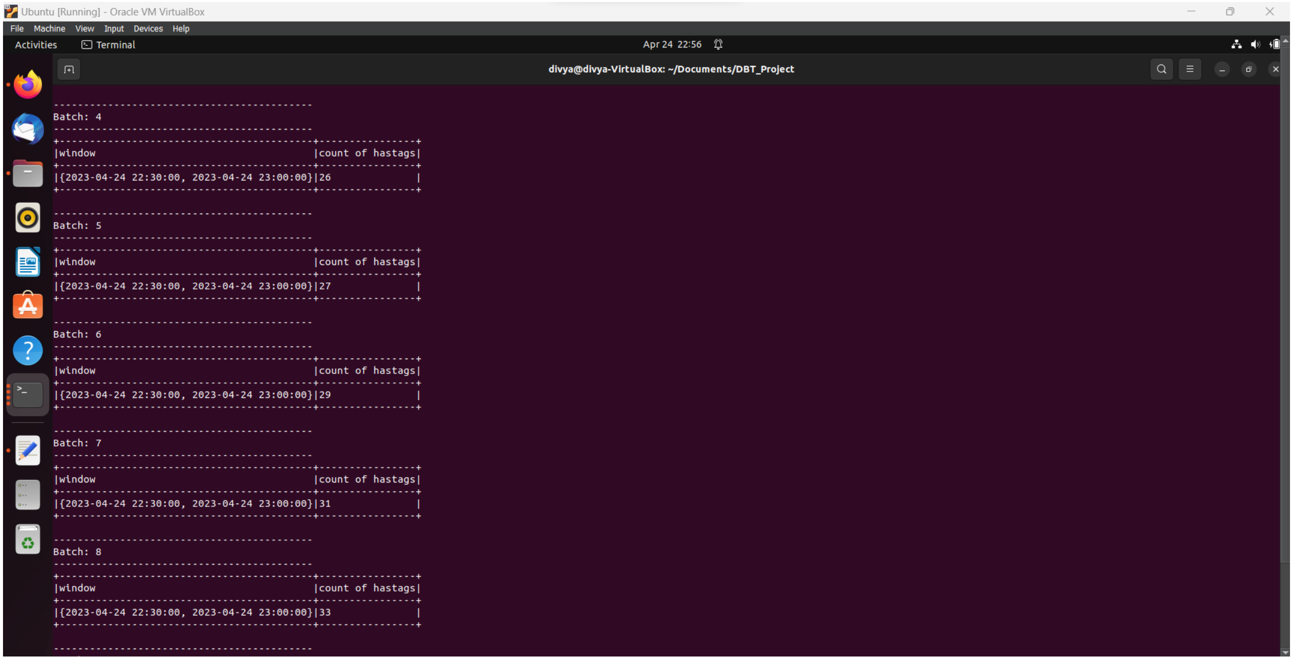
Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Apr 24 22:55
divya@divya-VirtualBox: ~/Documents/DBT_Project

```
divya@divya-VirtualBox:~/Documents/DBT_Project$ python3 consumer_stream.py
23/04/24 22:52:59 WARN Utils: Your hostname, divya-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
23/04/24 22:52:59 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
:: loading settings :: url = jar:file:/opt/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/divya/.ivy2/cache
The jars for the packages stored in: /home/divya/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-00699f91-d178-4dce-b49e-8a215924947a;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.4.0 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.4.0 in central
        found org.apache.kafka#kafka-clients;3.3.2 in central
        found org.lz4#lz4-java;1.8.0 in central
        found org.xerial.snappy#snappy-java;1.1.9.1 in central
        found org.slf4j#slf4j-api;2.0.6 in central
        found org.apache.hadoop#hadoop-client-runtime;3.3.4 in central
        found org.apache.hadoop#hadoop-client-api;3.3.4 in central
        found commons-logging#commons-logging;1.1.3 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found org.apache.commons#commons-pool2;2.11.1 in central
:: resolution report :: resolve 1332ms :: artifacts dl 89ms
        :: modules in use:
        com.google.code.findbugs#jsr305;3.0.0 from central in [default]
        commons-logging#commons-logging;1.1.3 from central in [default]
        org.apache.commons#commons-pool2;2.11.1 from central in [default]
        org.apache.hadoop#hadoop-client-api;3.3.4 from central in [default]
        org.apache.hadoop#hadoop-client-runtime;3.3.4 from central in [default]
        org.apache.kafka#kafka-clients;3.3.2 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.4.0 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.4.0 from central in [default]
        org.lz4#lz4-java;1.8.0 from central in [default]
        org.slf4j#slf4j-api;2.0.6 from central in [default]
        org.xerial.snappy#snappy-java;1.1.9.1 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   11  |   0   |   0   |   0   ||   11  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-00699f91-d178-4dce-b49e-8a215924947a
        confs: [default]
        0 artifacts copied, 11 already retrieved (0kB/16ms)
23/04/24 22:53:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
 Consumer Stream Processing has Started
Temporary view has been created!
```



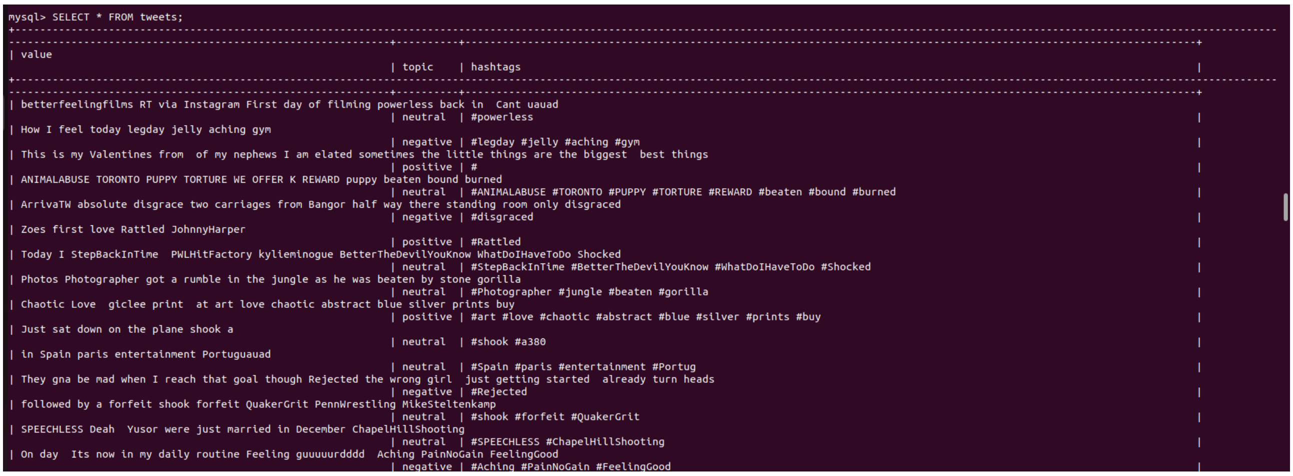
Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Apr 24 22:56
divya@divya-VirtualBox: ~/Documents/DBT_Project

```
23/04/24 22:53:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
 Consumer Stream Processing has Started
Temporary view has been created!
23/04/24 22:53:15 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
23/04/24 22:53:16 WARN AdminClientConfig: These configurations '[key.deserializer, value.deserializer, enable.auto.commit, max.poll.records, auto.offset.reset]' were supplied but are not used yet.
-------------------------------------------
Batch: 0
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|15              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|21              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|22              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|23              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 4
-------------------------------------------
+------------------------------------------+----------------+
```


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Apr 24 22:56
divya@divya-VirtualBox: ~/Documents/DBT_Project

```
-------------------------------------------
Batch: 4
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|26              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 5
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|27              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 6
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|29              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 7
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|31              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 8
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|33              |
+------------------------------------------+----------------+

-------------------------------------------
```

The output of kafka_read.py



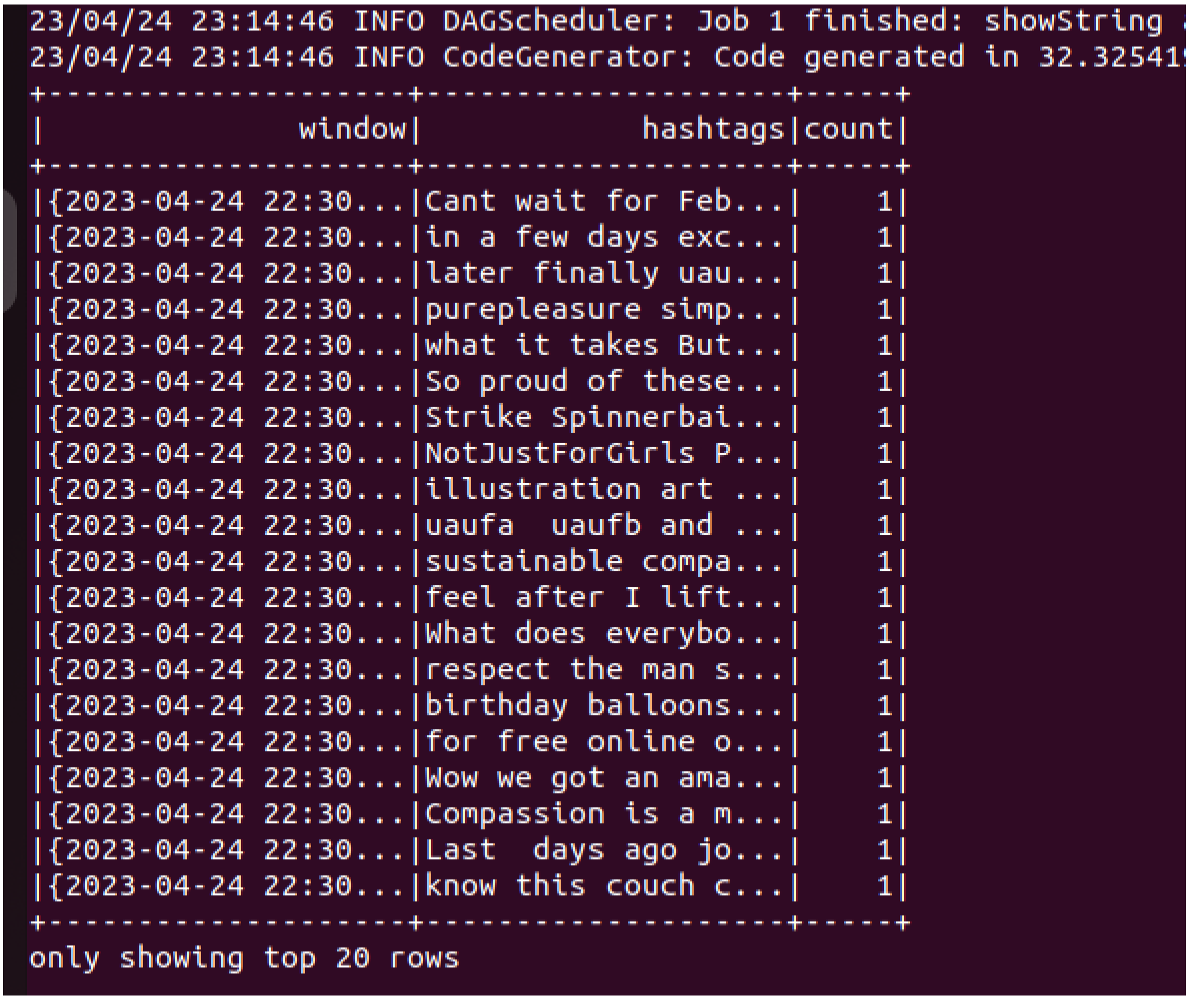
```
mysql> SELECT * FROM tweets;
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------+----------+----------------------------------------------------------------------------------------------------------------------+
| value
                                                             | topic    | hashtags                                                                                                             |
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------+----------+----------------------------------------------------------------------------------------------------------------------+
| betterfeelingfilms RT via Instagram First day of filming powerless back in  Cant uauad
                                                             | neutral  | #powerless                                                                                                           |
| How I feel today legday jelly aching gym
                                                             | negative | #legday #jelly #aching #gym                                                                                          |
| This is my Valentines from  of my nephews I am elated sometimes the little things are the biggest  best things
                                                             | positive | #                                                                                                                    |
| ANIMALABUSE TORONTO PUPPY TORTURE WE OFFER K REWARD puppy beaten bound burned
                                                             | neutral  | #ANIMALABUSE #TORONTO #PUPPY #TORTURE #REWARD #beaten #bound #burned                                                 |
| ArrivaTW absolute disgrace two carriages from Bangor half way there standing room only disgraced
                                                             | negative | #disgraced                                                                                                           |
| Zoes first love Rattled JohnnyHarper
                                                             | positive | #Rattled                                                                                                             |
| Today I StepBackInTime  PWLHitFactory kylieminogue BetterTheDevilYouKnow WhatDoIHaveToDo Shocked
                                                             | neutral  | #StepBackInTime #BetterTheDevilYouKnow #WhatDoIHaveToDo #Shocked                                                     |
| Photos Photographer got a rumble in the jungle as he was beaten by stone gorilla
                                                             | neutral  | #Photographer #jungle #beaten #gorilla                                                                               |
| Chaotic Love  giclee print  at art love chaotic abstract blue silver prints buy
                                                             | positive | #art #love #chaotic #abstract #blue #silver #prints #buy                                                             |
| Just sat down on the plane shook a
                                                             | neutral  | #shook #a380                                                                                                         |
| in Spain paris entertainment Portuguauad
                                                             | neutral  | #Spain #paris #entertainment #Portug                                                                                 |
| They gna be mad when I reach that goal though Rejected the wrong girl  just getting started  already turn heads
                                                             | negative | #Rejected                                                                                                            |
| followed by a forfeit shook forfeit QuakerGrit PennWrestling MikeSteltenkamp
                                                             | neutral  | #shook #forfeit #QuakerGrit                                                                                          |
| SPEECHLESS Deah  Yusor were just married in December ChapelHillShooting
                                                             | neutral  | #SPEECHLESS #ChapelHillShooting                                                                                      |
| On day  Its now in my daily routine Feeling guuuuurdddd  Aching PainNoGain FeelingGood
                                                             | negative | #Aching #PainNoGain #FeelingGood                                                                                     |
```

 it stores the dataset in the database which will be further used by batch processing

Pragathya
Ananya
Divya
Nidhi

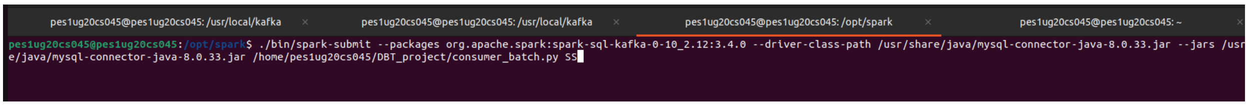
7. Batch Mode Experiment

- Description:
  - Input to the consumer is through the database which has stored the processed tweets which were received from the producer.
- Data Size
  - 375 KB
- Results
  - The result is selecting all tweets of topic ‘positive’ and then counting the number of hashtags in the tweet
- Output ScreenShot



```
23/04/24 23:14:46 INFO DAGScheduler: Job 1 finished: showString
23/04/24 23:14:46 INFO CodeGenerator: Code generated in 32.32541
+--------------------+--------------------+-----+
|              window|            hashtags|count|
+--------------------+--------------------+-----+
|{2023-04-24 22:30...|Cant wait for Feb...|    1|
|{2023-04-24 22:30...|in a few days exc...|    1|
|{2023-04-24 22:30...|later finally uau...|    1|
|{2023-04-24 22:30...|purepleasure simp...|    1|
|{2023-04-24 22:30...|what it takes But...|    1|
|{2023-04-24 22:30...|So proud of these...|    1|
|{2023-04-24 22:30...|Strike Spinnerbai...|    1|
|{2023-04-24 22:30...|NotJustForGirls P...|    1|
|{2023-04-24 22:30...|illustration art ...|    1|
|{2023-04-24 22:30...|uaufa  uaufb and ...|    1|
|{2023-04-24 22:30...|sustainable compa...|    1|
|{2023-04-24 22:30...|feel after I lift...|    1|
|{2023-04-24 22:30...|What does everybo...|    1|
|{2023-04-24 22:30...|respect the man s...|    1|
|{2023-04-24 22:30...|birthday balloons...|    1|
|{2023-04-24 22:30...|for free online o...|    1|
|{2023-04-24 22:30...|Wow we got an ama...|    1|
|{2023-04-24 22:30...|Compassion is a m...|    1|
|{2023-04-24 22:30...|Last  days ago jo...|    1|
|{2023-04-24 22:30...|know this couch c...|    1|
+--------------------+--------------------+-----+
only showing top 20 rows
```

- the command to run consumer_batch.py



pes1ug20cs045@pes1ug20cs045: /usr/local/kafka
pes1ug20cs045@pes1ug20cs045: /usr/local/kafka
pes1ug20cs045@pes1ug20cs045: /opt/spark
pes1ug20cs045@pes1ug20cs045: ~

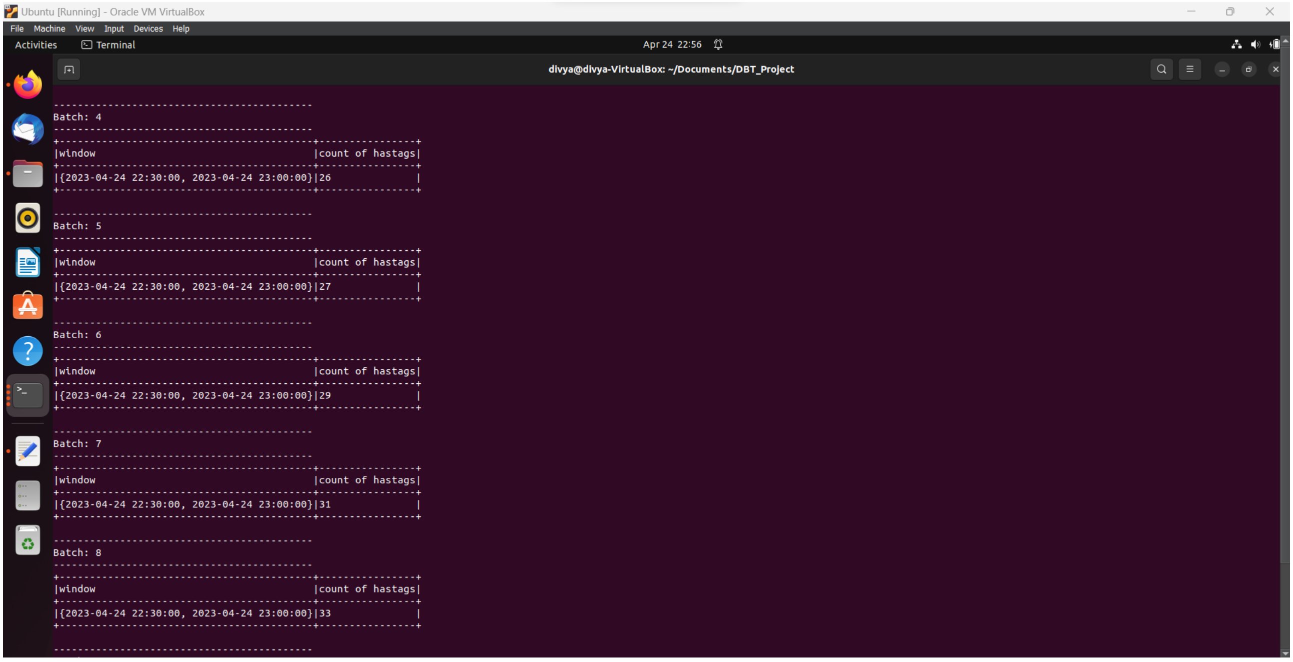
```
pes1ug20cs045@pes1ug20cs045:/opt/spark$ ./bin/spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.4.0 --driver-class-path /usr/share/java/mysql-connector-java-8.0.33.jar --jars /usr
e/java/mysql-connector-java-8.0.33.jar /home/pes1ug20cs045/DBT_project/consumer_batch.py SS
```

8. Comparison of Streaming & Batch Modes

Stream processing is much faster than batch processing as the number of tweets is counted for every 30 min where batch processing it is not real time and it reads data from the dataset

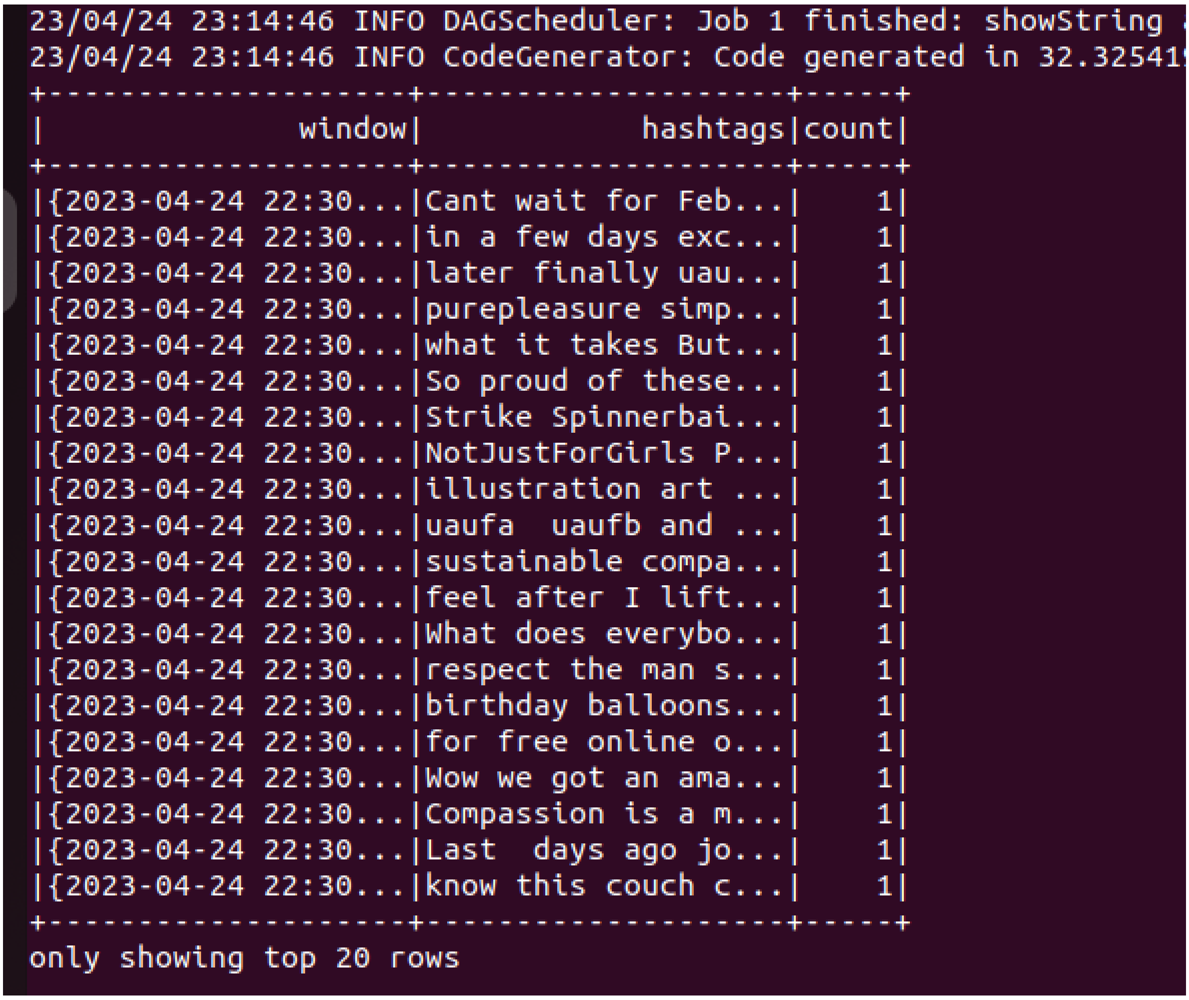
9. Results and Discussion

Output for Stream processing


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Apr 24 22:56
divya@divya-VirtualBox: ~/Documents/DBT_Project

```
-------------------------------------------
Batch: 4
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|26              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 5
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|27              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 6
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|29              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 7
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|31              |
+------------------------------------------+----------------+

-------------------------------------------
Batch: 8
-------------------------------------------
+------------------------------------------+----------------+
|window                                    |count of hastags|
+------------------------------------------+----------------+
|{2023-04-24 22:30:00, 2023-04-24 23:00:00}|33              |
+------------------------------------------+----------------+

-------------------------------------------
```

Output for batch processing



```
23/04/24 23:14:46 INFO DAGScheduler: Job 1 finished: showString
23/04/24 23:14:46 INFO CodeGenerator: Code generated in 32.32541
+--------------------+--------------------+-----+
|              window|            hashtags|count|
+--------------------+--------------------+-----+
|{2023-04-24 22:30...|Cant wait for Feb...|    1|
|{2023-04-24 22:30...|in a few days exc...|    1|
|{2023-04-24 22:30...|later finally uau...|    1|
|{2023-04-24 22:30...|purepleasure simp...|    1|
|{2023-04-24 22:30...|what it takes But...|    1|
|{2023-04-24 22:30...|So proud of these...|    1|
|{2023-04-24 22:30...|Strike Spinnerbai...|    1|
|{2023-04-24 22:30...|NotJustForGirls P...|    1|
|{2023-04-24 22:30...|illustration art ...|    1|
|{2023-04-24 22:30...|uaufa  uaufb and ...|    1|
|{2023-04-24 22:30...|sustainable compa...|    1|
|{2023-04-24 22:30...|feel after I lift...|    1|
|{2023-04-24 22:30...|What does everybo...|    1|
|{2023-04-24 22:30...|respect the man s...|    1|
|{2023-04-24 22:30...|birthday balloons...|    1|
|{2023-04-24 22:30...|for free online o...|    1|
|{2023-04-24 22:30...|Wow we got an ama...|    1|
|{2023-04-24 22:30...|Compassion is a m...|    1|
|{2023-04-24 22:30...|Last  days ago jo...|    1|
|{2023-04-24 22:30...|know this couch c...|    1|
+--------------------+--------------------+-----+
only showing top 20 rows
```

10. Conclusion

Stream processing is much faster than batch processing in certain conditions but here stream processing is faster.

11. References

- Streaming Spark Programming guide
- Kafka Streaming overview

Pragathya
Ananya
Divya
Nidhi