PES UNIVERSITY
100 feet Ring Road, BSK 3rd Stage
Bengaluru 560085 INDIA

Department of Computer Science and Engineering
B. Tech. CSE – 6th Semester
Jan – May 2024

UE21CS343BB3
DATABASE TECHNOLOGIES (DBT)

PROJECT REPORT
on

Performing Stream Processing and Batch Processing on Instagram comments

Submitted by : Team #: 691_698_850_912

Name :VISHNU	SRN :850	6th L sec	Name 3:SANJANA	SRN:912	6th L sec
Name: RACHANA	SRN :691	6th L sec	Name 4:VARSHA	SRN:698	6th L sec

Class of Prof. Raghu B. A.

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.

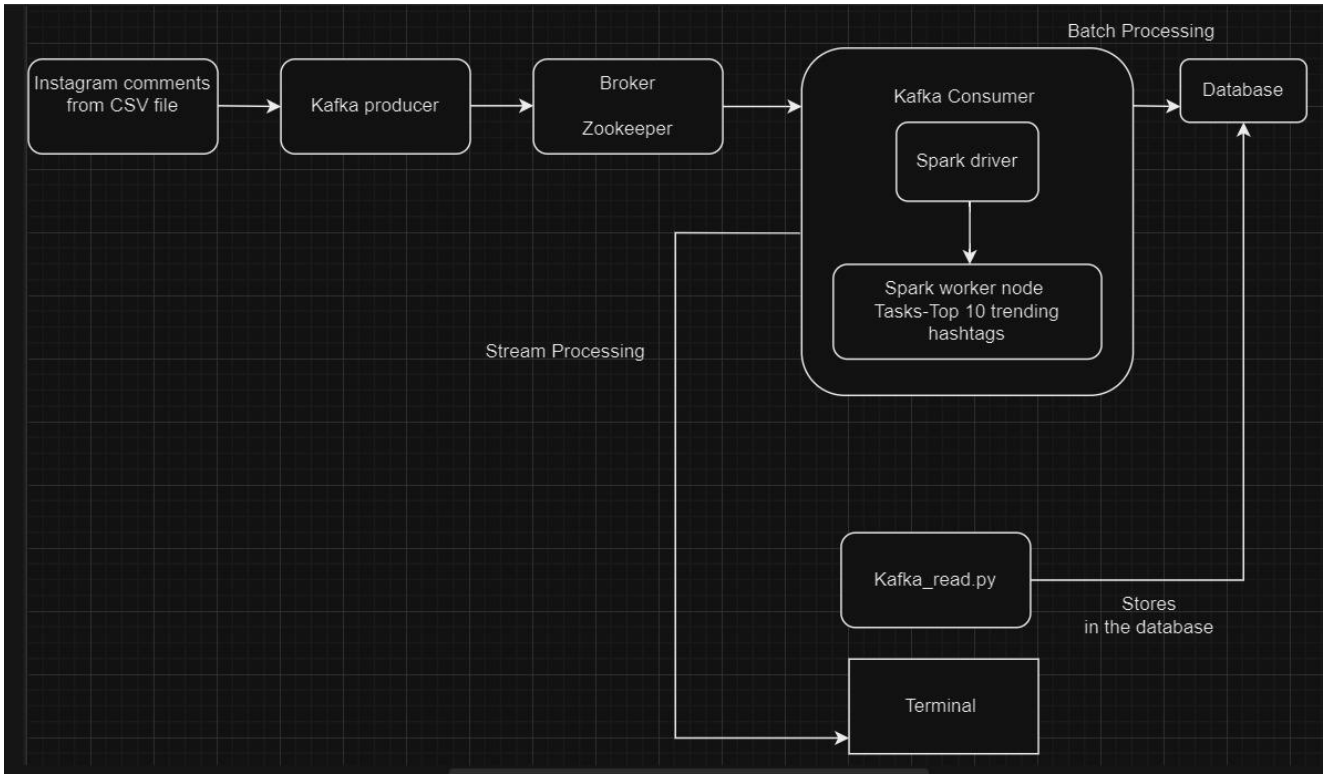
The aim is to compare the performance and accuracy of processing the data in both streaming and batch modes.

- 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.

.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.

.Architecture Diagram



2.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.Input Data ○ ● Source

• ○ Kaggle dataset

○ ● Description

- ○ The dataset containing comments 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

4. 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 instagram_comments WHERE topic = 'positive'")

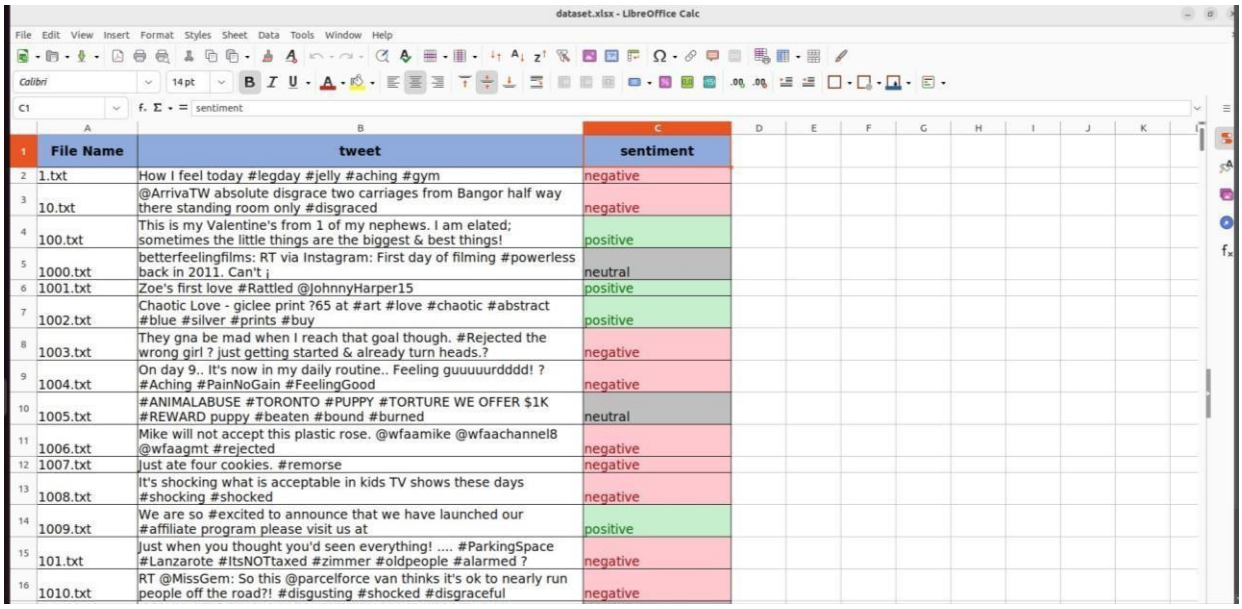
○ ● Inputs and Corresponding Results

- ○ Input is a Instagram dataset from which the producer read and publishes the topics to the kafka broker

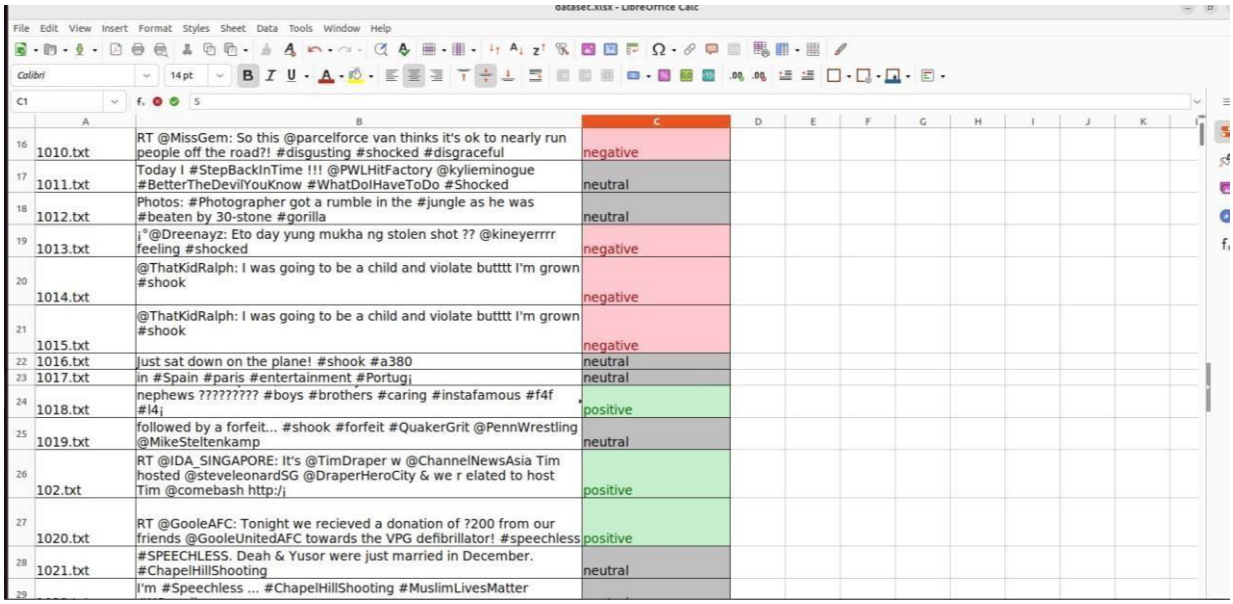
- The result is selecting all instagram_comments of topic ‘positive’ and then counting the number of hashtags in the instagram_comments.

.

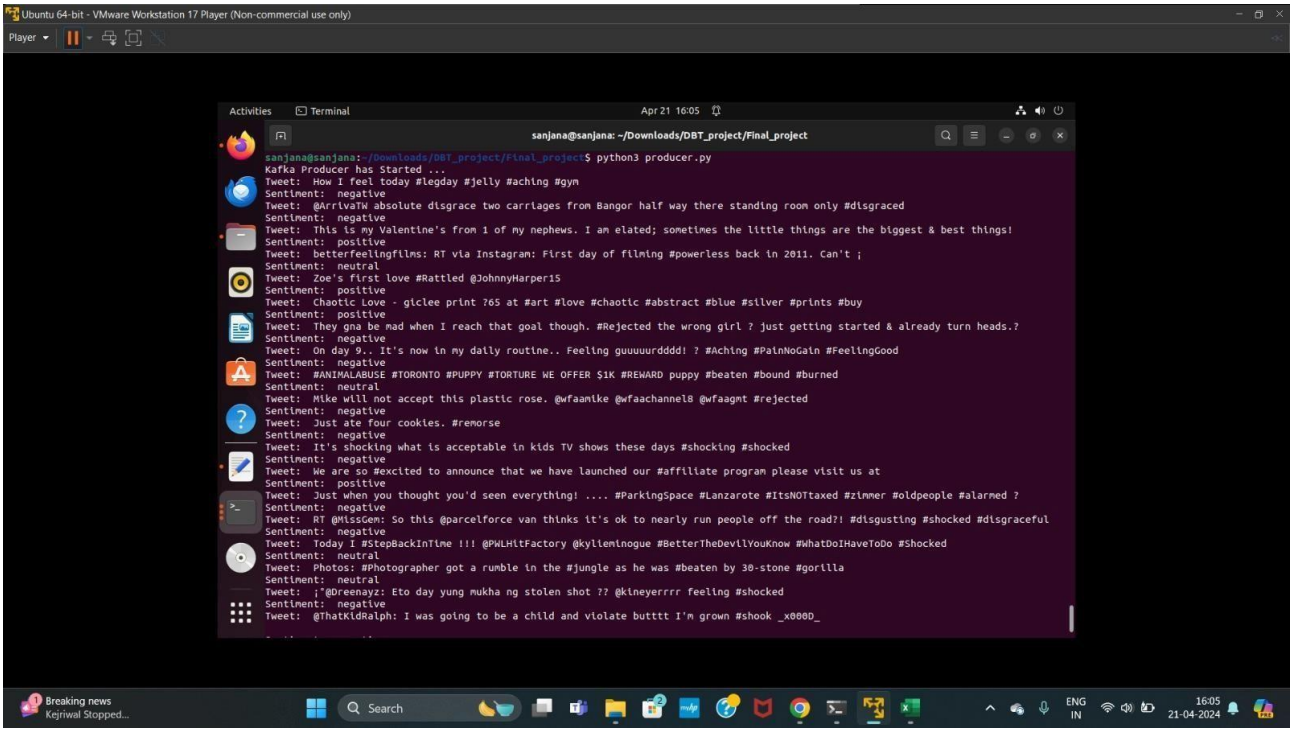
Input screenshot:



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



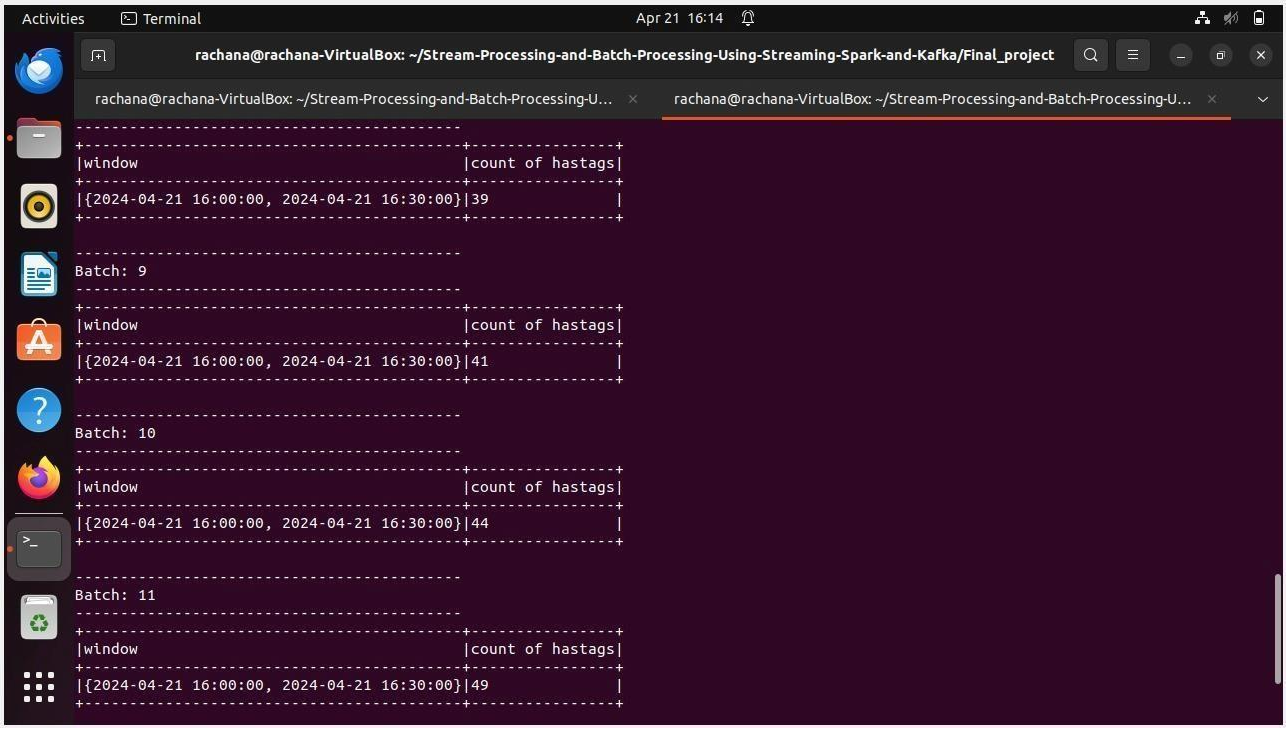
1010.txt	RT @MissGem: So this @parcelforce van thinks it's ok to nearly run people off the road?! #disgusting #shocked #disgraceful	negative
1011.txt	Today I #StepBackInTime !!! @PWLHitFactory @kylieminogue #BetterTheDevilYouKnow #WhatDoIHaveToDo #Shocked	neutral
1012.txt	Photos: #Photographer got a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



Output Screenshot:

```
rachana@rachana-VirtualBox:~/Stream-Processing-and-Batch-Processing-Using-Streaming-Spark-and-Kafka/Final_project$ python3 consumer_
stream_topic1.py
24/04/21 16:10:32 WARN Utils: Your hostname, rachana-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (
on interface enp0s3)
24/04/21 16:10:32 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/rachana/.ivy2/cache
The jars for the packages stored in: /home/rachana/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-23f92490-2bf0-4098-907f-47726713c216;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 1248ms :: artifacts dl 127ms
        :: 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]
```

```
Activities   Terminal                                   Apr 21 16:12
rachana@rachana-VirtualBox: ~/Stream-Processing-and-Batch-Processing-Using-Streaming-Spark-and-Kafka/Final_project
rachana@rachana-VirtualBox: ~/Stream-Processing-and-Batch-Processing-U...   rachana@rachana-VirtualBox: ~/Stream-Processing-and-Batch-Processing-U...
-------------------------------------------
Batch: 0
-------------------------------------------
+----------------------------------------+----------------+
|window                                  |count of hastags|
+----------------------------------------+----------------+
|{2024-04-21 16:00:00, 2024-04-21 16:30:00}|6               |
+----------------------------------------+----------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+----------------------------------------+----------------+
|window                                  |count of hastags|
+----------------------------------------+----------------+
|{2024-04-21 16:00:00, 2024-04-21 16:30:00}|13              |
+----------------------------------------+----------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+----------------------------------------+----------------+
|window                                  |count of hastags|
+----------------------------------------+----------------+
|{2024-04-21 16:00:00, 2024-04-21 16:30:00}|19              |
+----------------------------------------+----------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+----------------------------------------+----------------+
|window                                  |count of hastags|
+----------------------------------------+----------------+
```



The output of kafka_read.py

```
mysql> select * from instagram_comments limit 5;
+-----------------------------------------------------------------------------------------------------------------------------------+----------+------
-----------------------------------------------+---------------------+
| value                                                                                                                             | topic    | hasht
ags                                            | timestamp           |
+-----------------------------------------------------------------------------------------------------------------------------------+----------+------
-----------------------------------------------+---------------------+
| This is my Valentines from  of my nephews I am elated sometimes the little things are the biggest  best things                   | positive |
                                               | 2024-04-21 16:22:42 |
| How I feel today legday jelly aching gym                                                                                          | negative | legda
y #jelly #aching #gym                          | 2024-04-21 16:22:42 |
| Zoes first love Rattled JohnnyHarper                                                                                              | positive | Rattl
ed                                             | 2024-04-21 16:22:42 |
| betterfeelingfilms RT via Instagram First day of filming powerless back in  Cant uauad                                            | neutral  | power
less                                           | 2024-04-21 16:22:42 |
| Chaotic Love  giclee print  at art love chaotic abstract blue silver prints buy                                                   | positive | art #
love #chaotic #abstract #blue #silver #prints #buy | 2024-04-21 16:22:42 |
+-----------------------------------------------------------------------------------------------------------------------------------+----------+------
-----------------------------------------------+---------------------+
5 rows in set (0.00 sec)
```

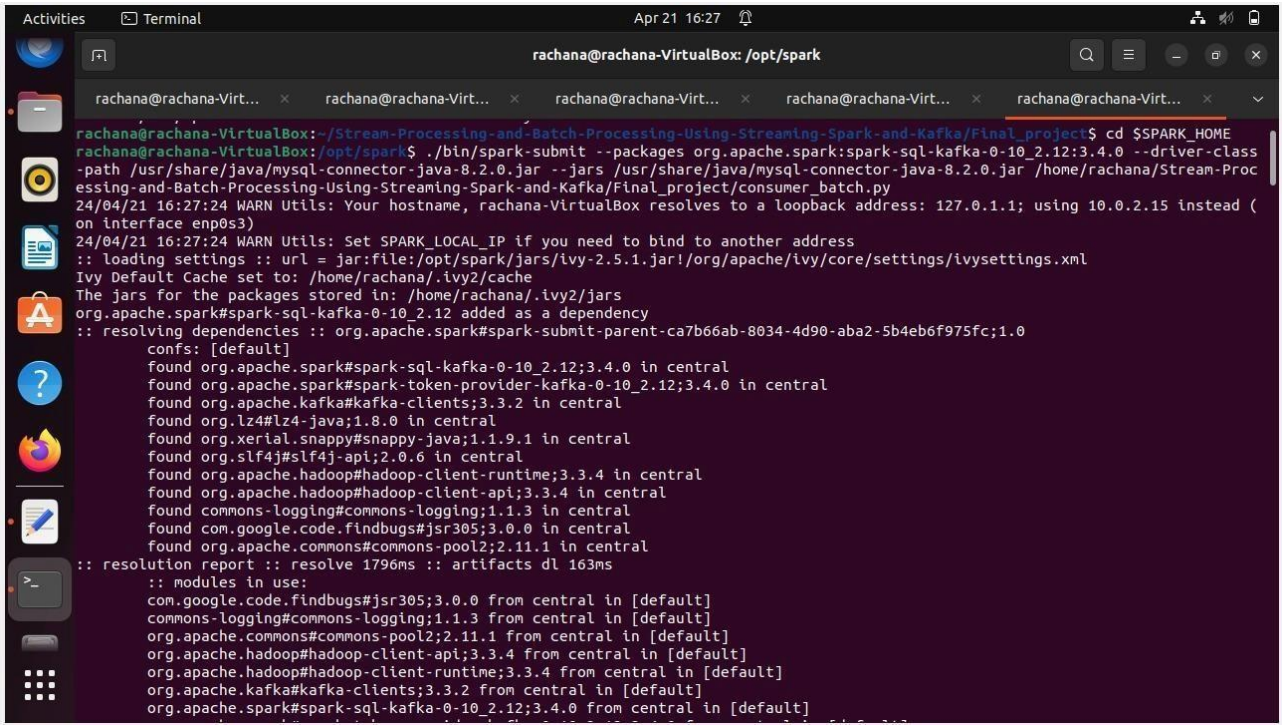
5. Batch Mode Experiment

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

of 14

Output Screenshot

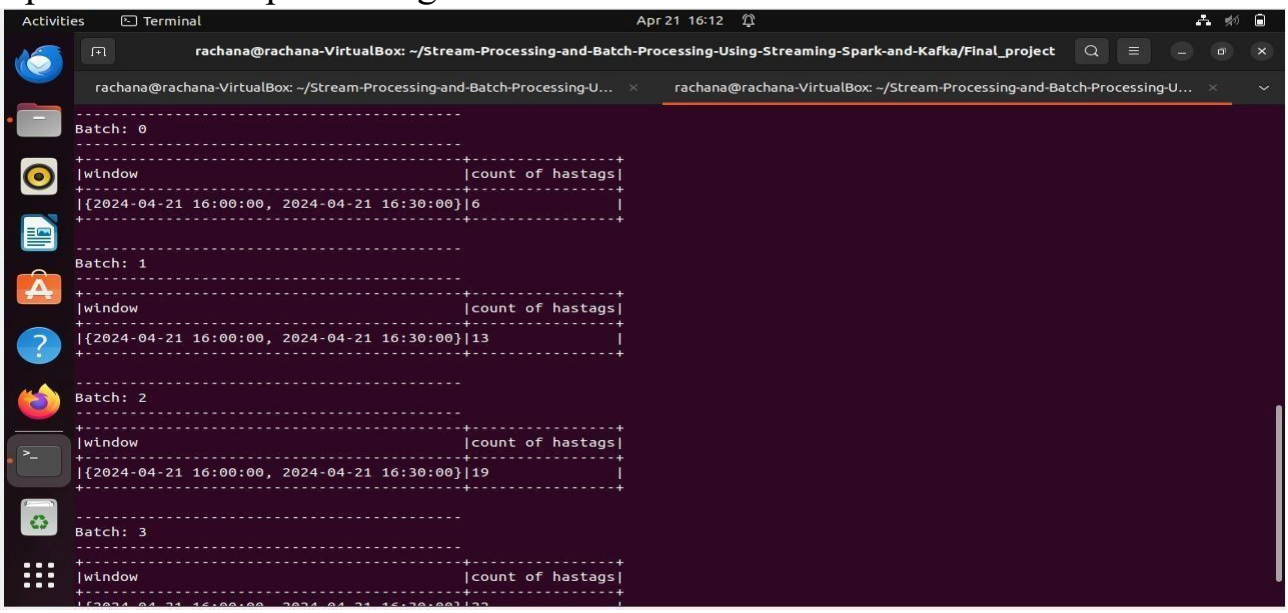
```
rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...

24/04/21 16:28:06 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/04/21 16:28:06 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.699546 s
24/04/21 16:28:07 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 10.0.2.15:34945 in memory (size: 22.5 KiB, free: 434.4 MiB)
24/04/21 16:28:08 INFO CodeGenerator: Code generated in 20.602036 ms
+--------------------+--------------------+-----+
|              window|            hashtags|count|
+--------------------+--------------------+-----+
|{2024-04-21 16:00...|tbt miamipro Apri...|    2|
|{2024-04-21 16:00...|the work of famou...|    2|
|{2024-04-21 16:00...|cool photo not mi...|    2|
|{2024-04-21 16:00...|Kh fitness model ...|    2|
|{2024-04-21 16:00...|illustration art ...|    2|
|{2024-04-21 16:00...|Signs Youre A Mat...|    2|
|{2024-04-21 16:00...|taken aback when ...|    2|
|{2024-04-21 16:00...|Festival  th Febr...|    2|
|{2024-04-21 16:00...|Animals wildlife ...|    1|
|{2024-04-21 16:00...|NEW FancyYhancy W...|    2|
|{2024-04-21 16:00...|hair I done  hair...|    2|
|{2024-04-21 16:00...|bleak midwinter e...|    4|
|{2024-04-21 16:00...|memory and focus ...|    2|
|{2024-04-21 16:00...|Yay They arrived ...|    2|
|{2024-04-21 16:00...|At a forum to dis...|    2|
|{2024-04-21 16:00...|get my order at D...|    2|
|{2024-04-21 16:00...|sunglasses gafas ...|    2|
|{2024-04-21 16:00...|of Bromeliads for...|    2|
|{2024-04-21 16:00...|so excited to ann...|    3|
|{2024-04-21 16:00...|strong awesome RT...|    2|
+--------------------+--------------------+-----+
only showing top 20 rows

24/04/21 16:28:09 INFO SparkContext: Invoking stop() from shutdown hook
24/04/21 16:28:09 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/04/21 16:28:09 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4041
```



6. Comparison of Streaming & Batch Modes

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

Output for Stream processing



Output for batch processing

```
rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...   rachana@rachana-Virt...

24/04/21 16:28:06 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/04/21 16:28:06 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.699546 s
24/04/21 16:28:07 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 10.0.2.15:34945 in memory (size: 22.5 KiB, free: 434.4 MiB)
24/04/21 16:28:08 INFO CodeGenerator: Code generated in 20.602036 ms
+--------------------+--------------------+-----+
|              window|            hashtags|count|
+--------------------+--------------------+-----+
|{2024-04-21 16:00...|tbt miamipro Apri...|    2|
|{2024-04-21 16:00...|the work of famou...|    2|
|{2024-04-21 16:00...|cool photo not mi...|    2|
|{2024-04-21 16:00...|Kh fitness model ...|    2|
|{2024-04-21 16:00...|illustration art ...|    2|
|{2024-04-21 16:00...|Signs Youre A Mat...|    2|
|{2024-04-21 16:00...|taken aback when ...|    2|
|{2024-04-21 16:00...|Festival  th Febr...|    2|
|{2024-04-21 16:00...|Animals wildlife ...|    1|
|{2024-04-21 16:00...|NEW FancyYhancy W...|    2|
|{2024-04-21 16:00...|hair I done  hair...|    2|
|{2024-04-21 16:00...|bleak midwinter e...|    4|
|{2024-04-21 16:00...|memory and focus ...|    2|
|{2024-04-21 16:00...|Yay They arrived ...|    2|
|{2024-04-21 16:00...|At a forum to dis...|    2|
|{2024-04-21 16:00...|get my order at D...|    2|
|{2024-04-21 16:00...|sunglasses gafas ...|    2|
|{2024-04-21 16:00...|of Bromeliads for...|    2|
|{2024-04-21 16:00...|so excited to ann...|    3|
|{2024-04-21 16:00...|strong awesome RT...|    2|
+--------------------+--------------------+-----+
only showing top 20 rows

24/04/21 16:28:09 INFO SparkContext: Invoking stop() from shutdown hook
24/04/21 16:28:09 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/04/21 16:28:09 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4041
```

7. Conclusion

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

8. References

- Streaming Spark Programming guide
- Kafka Streaming overview