

# Database Technologies

## Course-Project

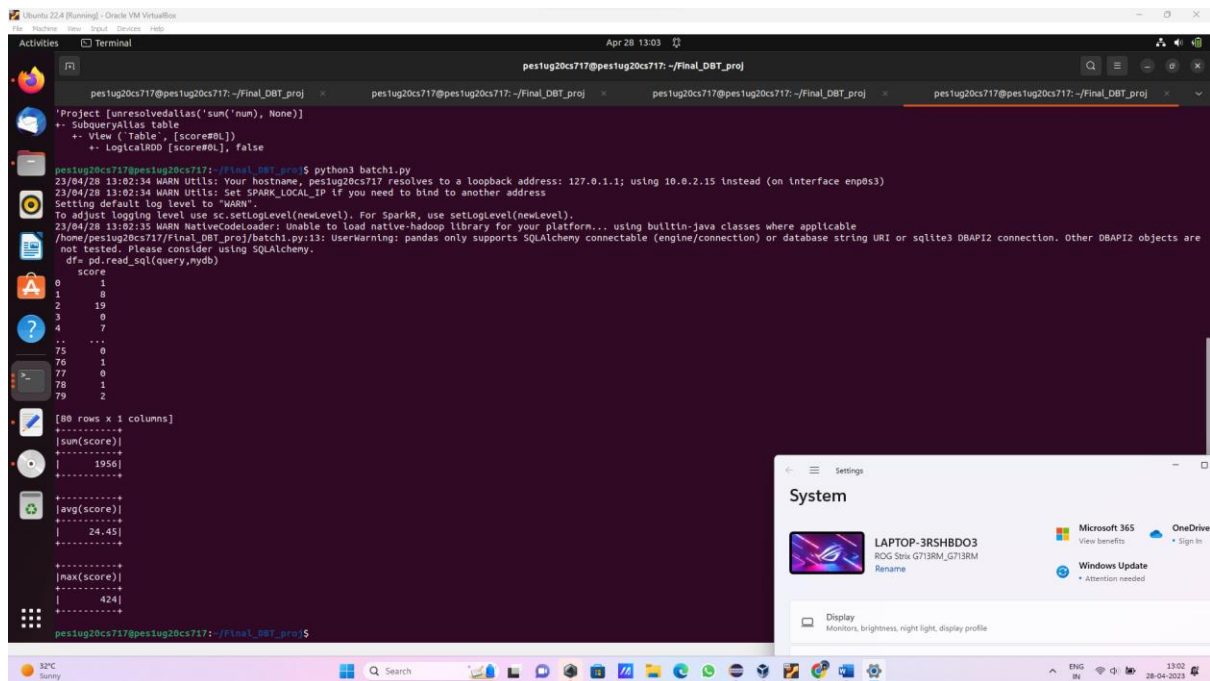
Name: Shal Ritvik Sinha

SRN: PES1UG20CS717

**Problem Statement:** Comparison between spark streaming and spark batch processing.

**Technique applied:** Tumbling Windows

**Batch mode:**

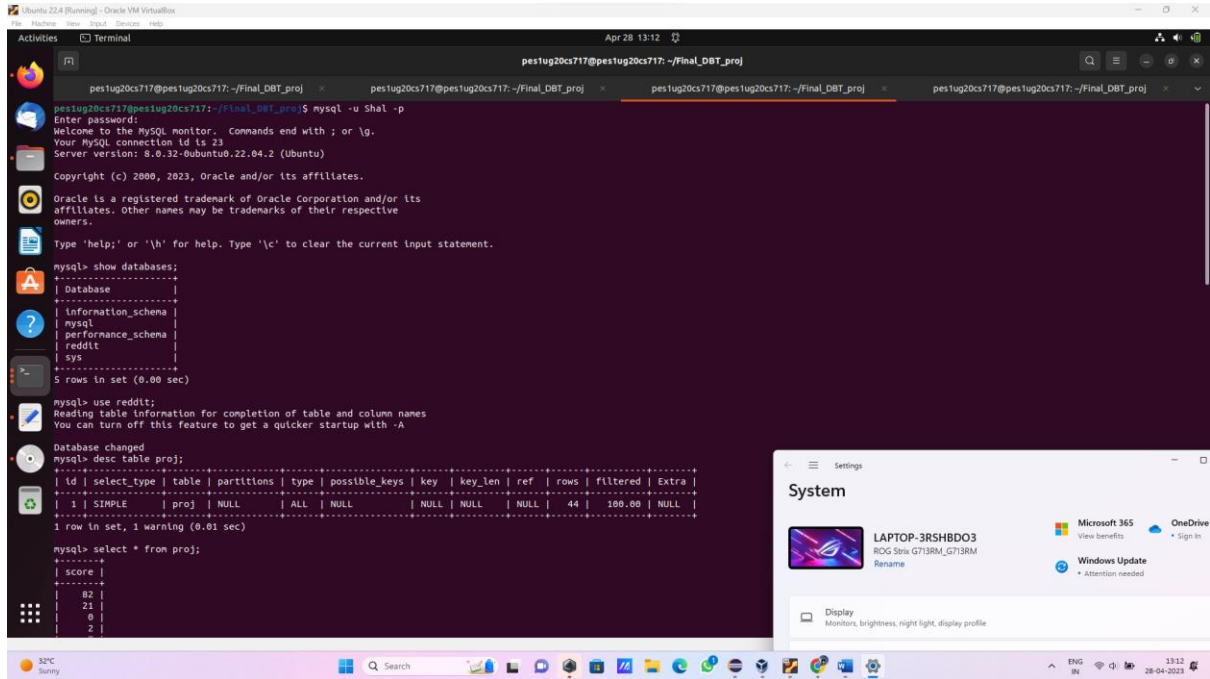


The screenshot shows a terminal window in a virtual machine (Ubuntu 22.4) running a Spark batch job. The terminal output displays the execution of a Python script named `batch1.py`. The script processes a dataset and outputs the following summary statistics:

```
[88 rows x 1 columns]
+-----+
|sum(score)|
+-----+
|      1956|
+-----+
|avg(score)|
+-----+
|      24.45|
+-----+
|max(score)|
+-----+
|       424|
+-----+
```

Below the terminal window, a Windows Settings window is visible, showing the system information for a laptop (LAPTOP-3RSHBDO3) and the Windows Update status.

## In mysql:



```
peslug20cs717@peslug20cs717: ~/Final_DBT_proj
peslug20cs717@peslug20cs717:~/Final_DBT_proj$ mysql -u Shal -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 23
Server version: 8.0.32-0ubuntu22.04.2 (Ubuntu)

Copyright (c) 2000, 2023, 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.

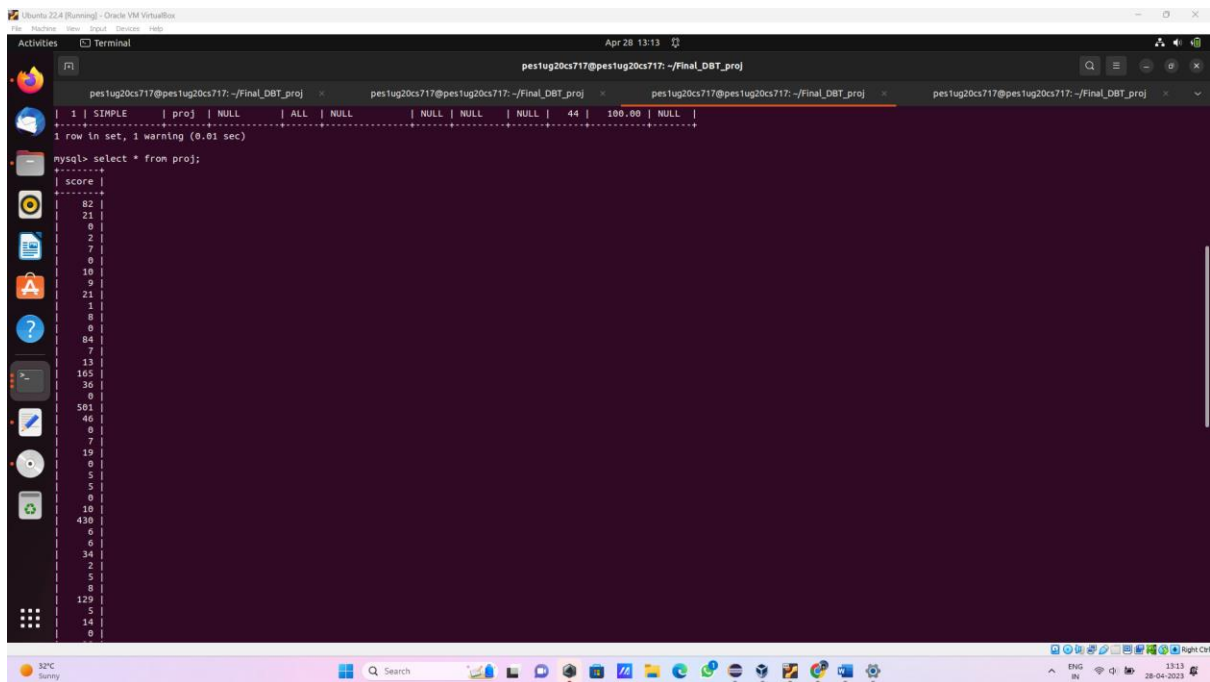
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mysql |
| performance_schema |
| reddit |
| sys |
+-----+
5 rows in set (0.00 sec)

mysql> use reddit;
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> desc table proj;
+-----+
| Id | select_type | table | partitions | type | possible_keys | key | key_len | ref | rows | filtered | Extra |
+-----+
| 1 | SIMPLE | proj | NULL | ALL | NULL | NULL | NULL | NULL | 44 | 100.00 | NULL |
+-----+
1 row in set, 1 warning (0.01 sec)

mysql> select * from proj;
+-----+
| score |
+-----+
| 82 |
| 21 |
| 0 |
| 2 |
| 7 |
| 0 |
| 10 |
| 9 |
| 21 |
| 1 |
| 8 |
| 0 |
| 84 |
| 7 |
| 13 |
| 105 |
| 36 |
| 0 |
| 501 |
| 46 |
| 0 |
| 7 |
| 19 |
| 0 |
| 5 |
| 5 |
| 0 |
| 10 |
| 430 |
| 6 |
| 6 |
| 34 |
| 2 |
| 5 |
| 8 |
| 129 |
| 5 |
| 14 |
| 0 |
+-----+
```



```
peslug20cs717@peslug20cs717: ~/Final_DBT_proj
peslug20cs717@peslug20cs717:~/Final_DBT_proj$ mysql -u Shal -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 23
Server version: 8.0.32-0ubuntu22.04.2 (Ubuntu)

Copyright (c) 2000, 2023, 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.

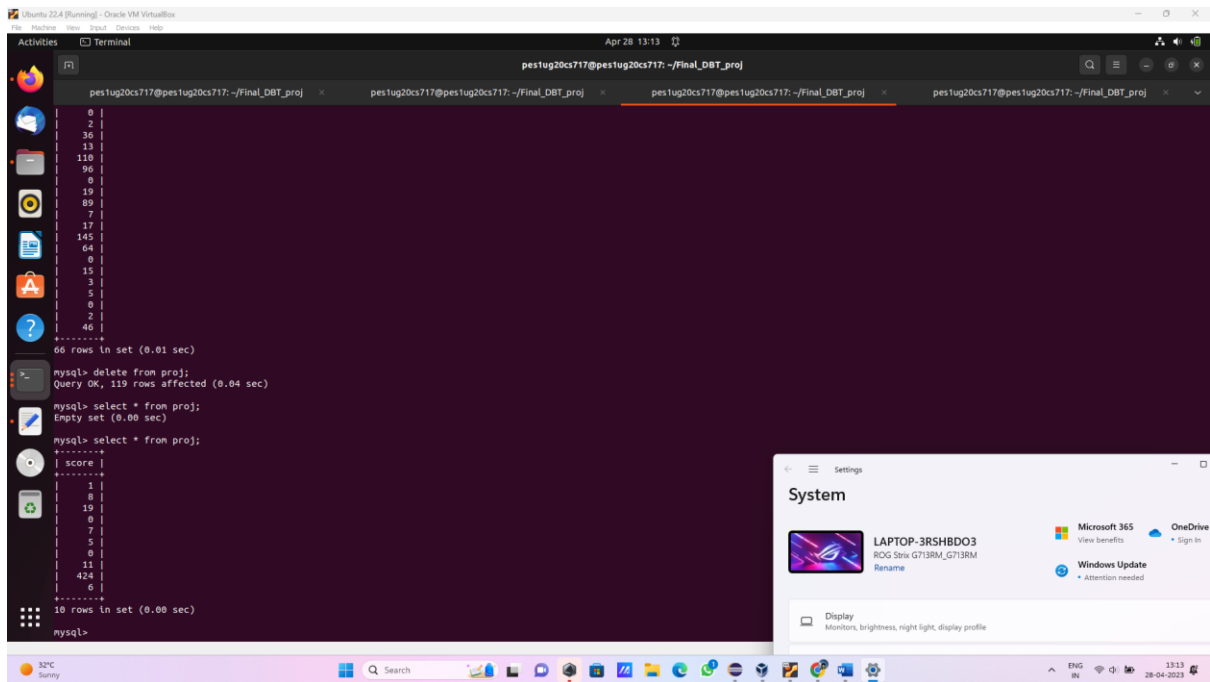
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mysql |
| performance_schema |
| reddit |
| sys |
+-----+
5 rows in set (0.00 sec)

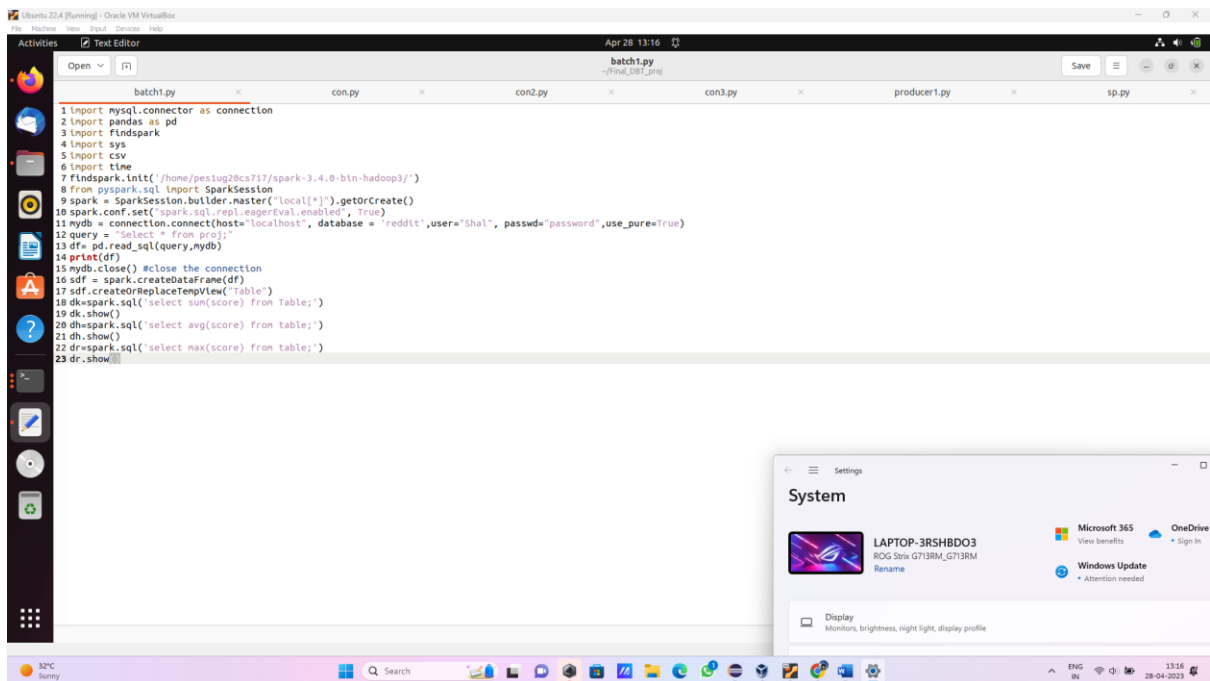
mysql> use reddit;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> desc table proj;
+-----+
| Id | select_type | table | partitions | type | possible_keys | key | key_len | ref | rows | filtered | Extra |
+-----+
| 1 | SIMPLE | proj | NULL | ALL | NULL | NULL | NULL | NULL | 44 | 100.00 | NULL |
+-----+
1 row in set, 1 warning (0.01 sec)

mysql> select * from proj;
+-----+
| score |
+-----+
| 82 |
| 21 |
| 0 |
| 2 |
| 7 |
| 0 |
| 10 |
| 9 |
| 21 |
| 1 |
| 8 |
| 0 |
| 84 |
| 7 |
| 13 |
| 105 |
| 36 |
| 0 |
| 501 |
| 46 |
| 0 |
| 7 |
| 19 |
| 0 |
| 5 |
| 5 |
| 0 |
| 10 |
| 430 |
| 6 |
| 6 |
| 34 |
| 2 |
| 5 |
| 8 |
| 129 |
| 5 |
| 14 |
| 0 |
+-----+
```

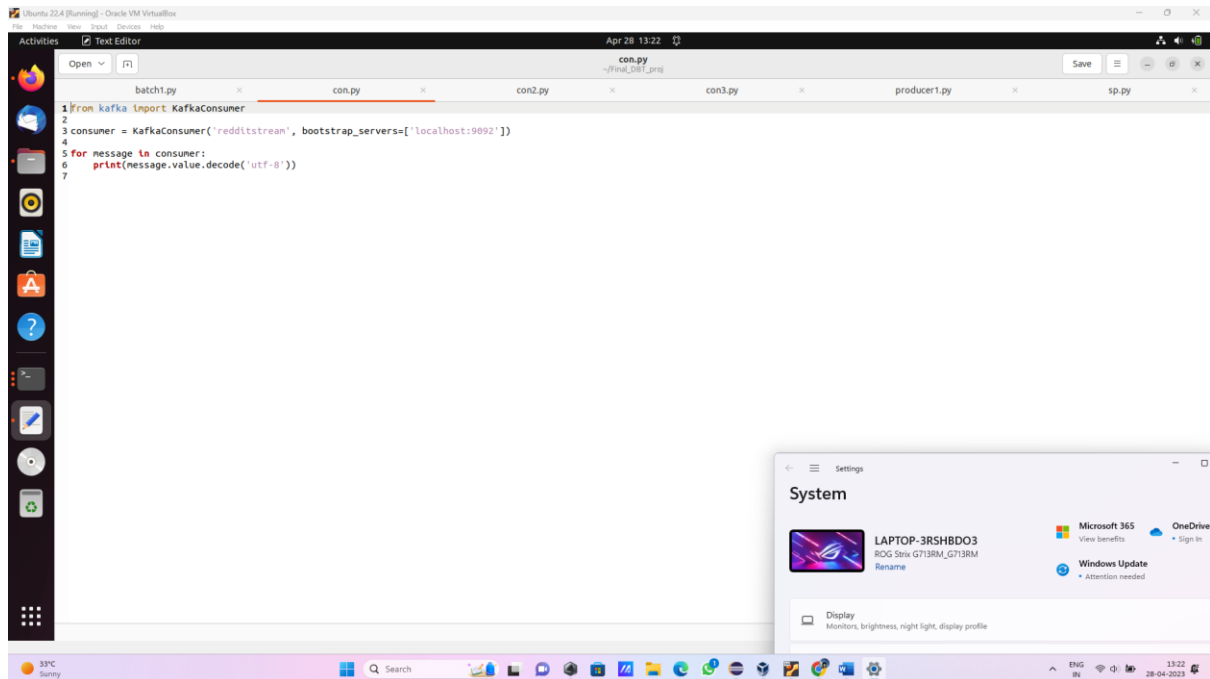


## Batch file

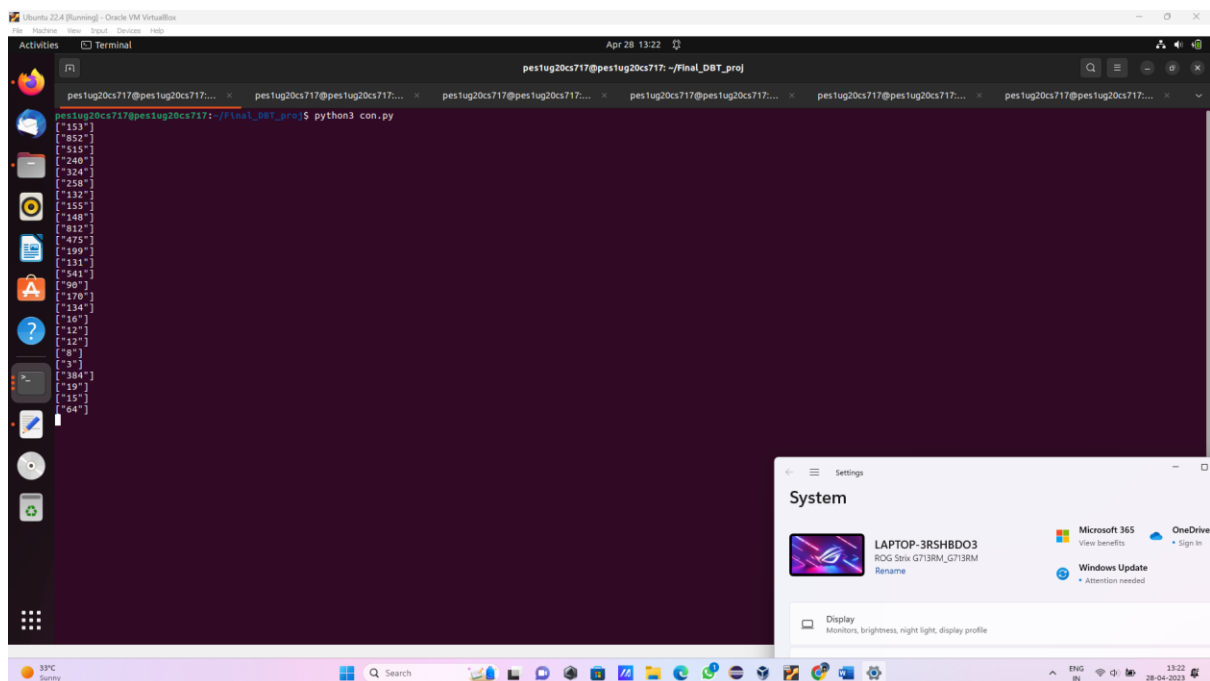


## Consumer file

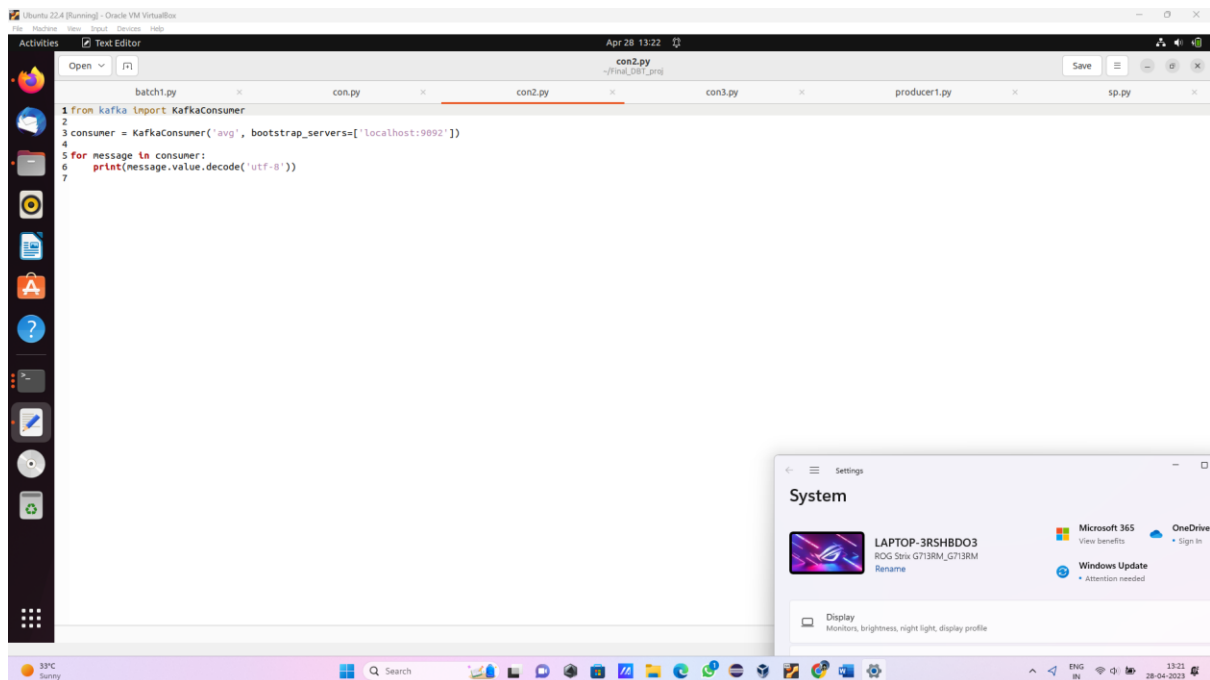
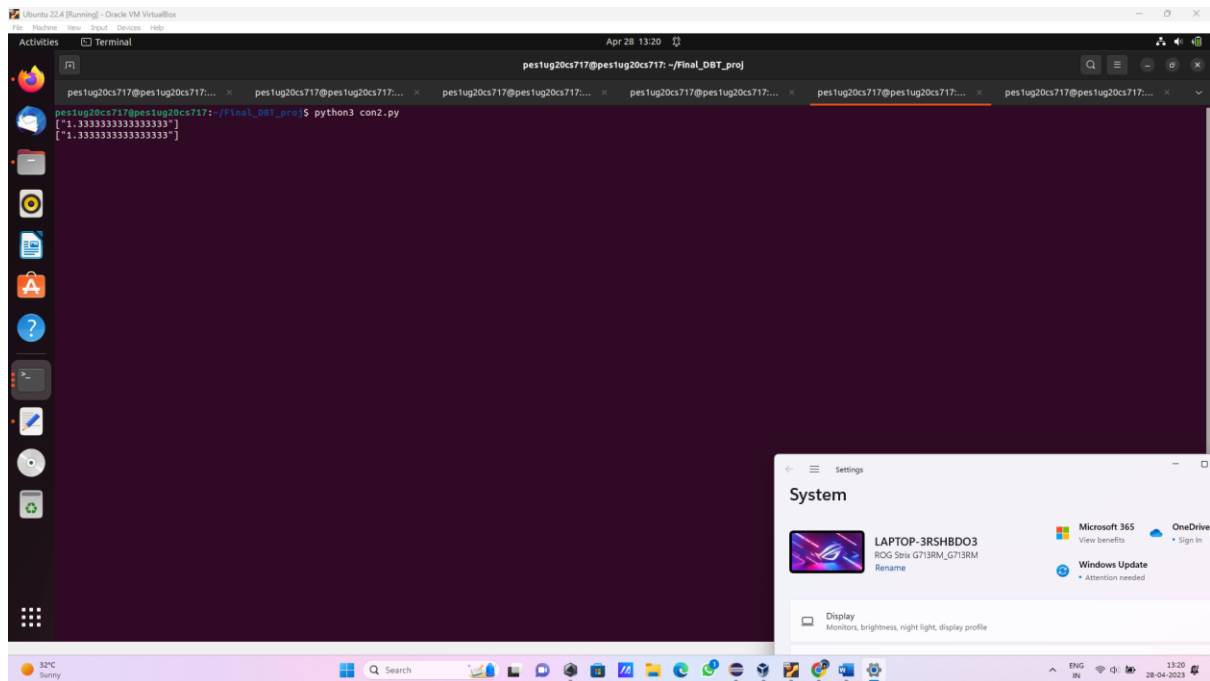
### Consumer1



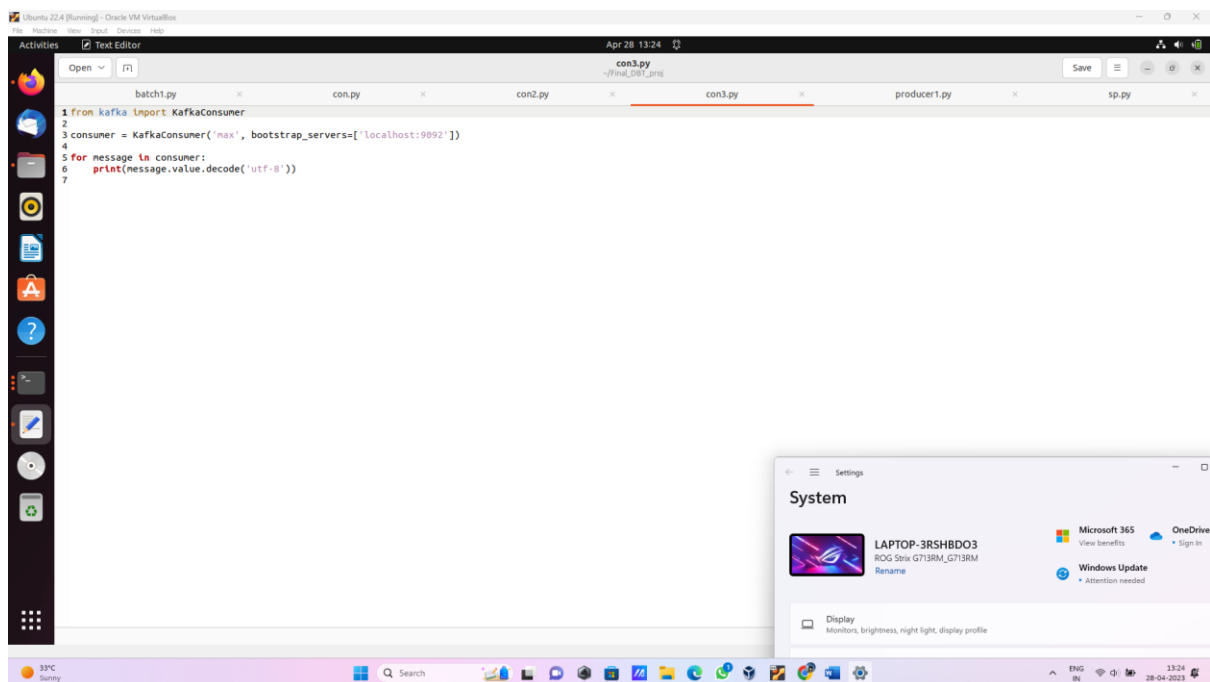
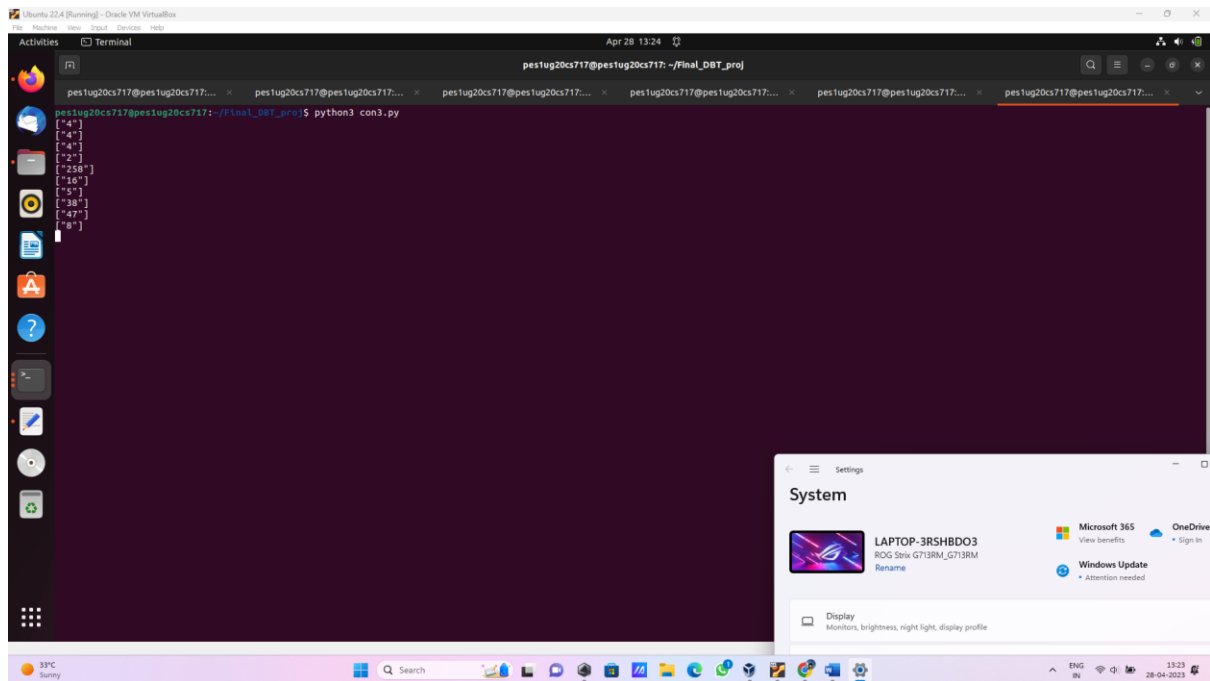
### Running Consumer1



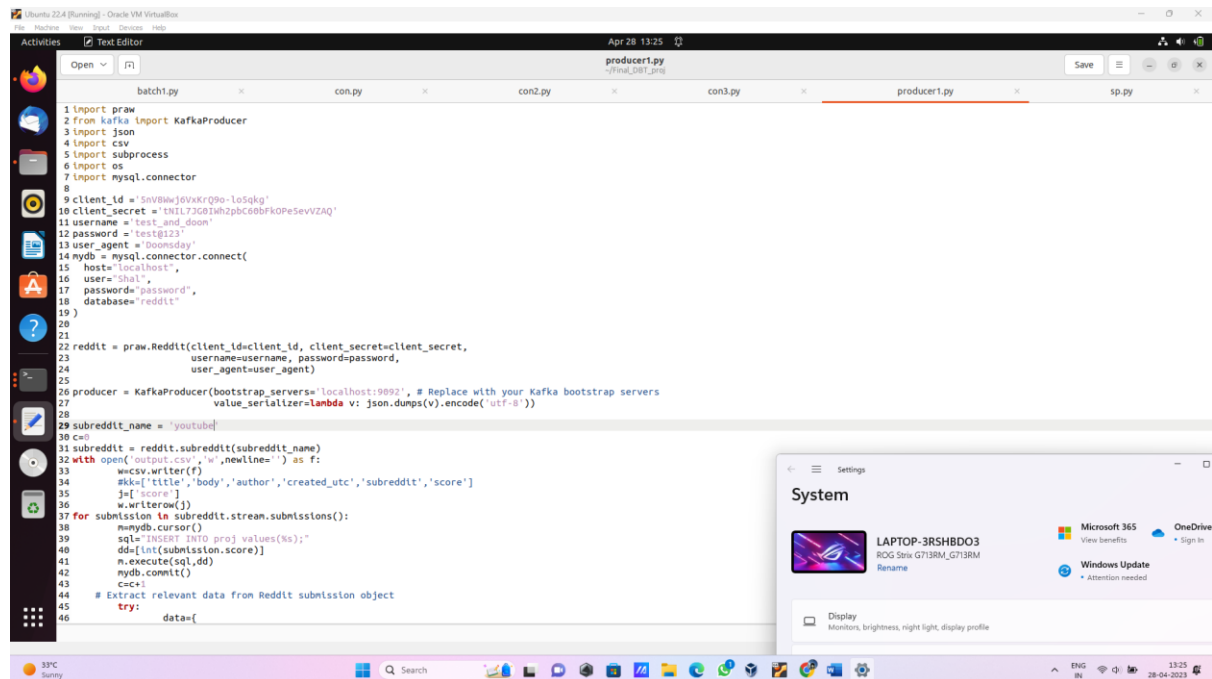
## Running Consumer2



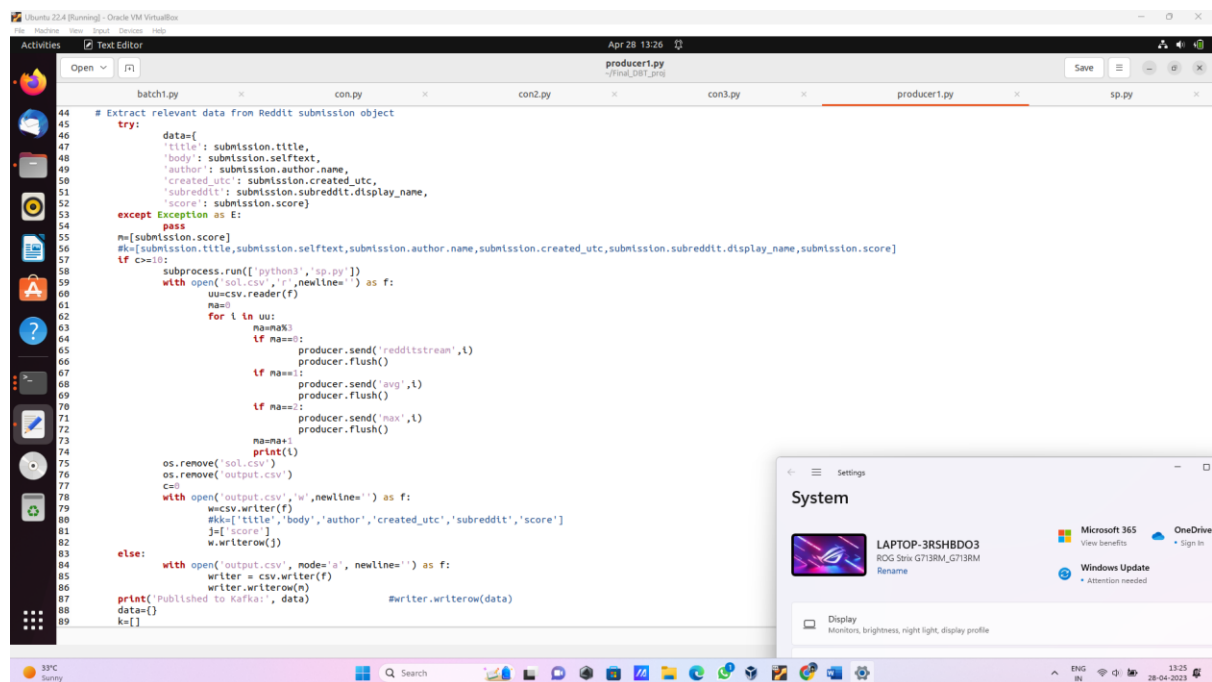
## Running Consumer3



## Producer file



```
1 import praw
2 from kafka import KafkaProducer
3 import json
4 import csv
5 import subprocess
6 import os
7 import mysql.connector
8
9 client_id = '5nV8Wj0vXkrQ9o-1o5qkg'
10 client_secret = 'tNtIL73G01Mh2pbc60BFk0PeSevVZAQ'
11 username = 'test_and_doom'
12 password = 'test@123'
13 user_agent = 'redditbot'
14 mydb = mysql.connector.connect(
15     host='localhost',
16     user='bhal',
17     password='password',
18     database='reddit'
19 )
20
21 reddit = praw.Reddit(client_id=client_id, client_secret=client_secret,
22                     username=username, password=password,
23                     user_agent=user_agent)
24
25 producer = KafkaProducer(bootstrap_servers='localhost:9092', # Replace with your Kafka bootstrap servers
26                          value_serializer=lambda v: json.dumps(v).encode('utf-8'))
27
28 subreddit_name = 'youtube'
29
30 c=1
31 subreddit = reddit.subreddit(subreddit_name)
32 with open('output.csv', 'w', newline='') as f:
33     w = csv.writer(f)
34     #k=[['title', 'body', 'author', 'created_utc', 'subreddit', 'score']]
35     j=['score']
36     w.writerow(j)
37     for submission in subreddit.stream.submissions():
38         n=mydb.cursor()
39         sql="INSERT INTO proj values(%s):"
40         dd=[int(submission.score)]
41         n.execute(sql,dd)
42         mydb.commit()
43         # Extract relevant data from Reddit submission object
44         try:
45             data={
46
```



```
44 # Extract relevant data from Reddit submission object
45 try:
46     data={
47         'title': submission.title,
48         'body': submission.selftext,
49         'author': submission.author.name,
50         'created_utc': submission.created_utc,
51         'subreddit': submission.subreddit.display_name,
52         'score': submission.score
53     }
54 except Exception as e:
55     pass
56 n=[submission.title, submission.selftext, submission.author.name, submission.created_utc, submission.subreddit.display_name, submission.score]
57 #k=[submission.title, submission.selftext, submission.author.name, submission.created_utc, submission.subreddit.display_name, submission.score]
58 if c>=10:
59     subprocess.run(['python3', 'sp.py'])
60     with open('sol.csv', 'r', newline='') as f:
61         uu=csv.reader(f)
62         na=0
63         for l in uu:
64             na+=1
65             if na==0:
66                 producer.send('redditstream', l)
67                 producer.flush()
68             if na==1:
69                 producer.send('avg', l)
70                 producer.flush()
71             if na==2:
72                 producer.send('max', l)
73                 producer.flush()
74             na+=1
75             print(l)
76 os.remove('sol.csv')
77 os.remove('output.csv')
78 c=0
79 with open('output.csv', 'w', newline='') as f:
80     w=csv.writer(f)
81     #k=[['title', 'body', 'author', 'created_utc', 'subreddit', 'score']]
82     j=['score']
83     w.writerow(j)
84 else:
85     with open('output.csv', mode='a', newline='') as f:
86         writer = csv.writer(f)
87         writer.writerow(n)
88         print('Published to Kafka', data) #writer.writerow(data)
89     data={}
90     k=[]
```

