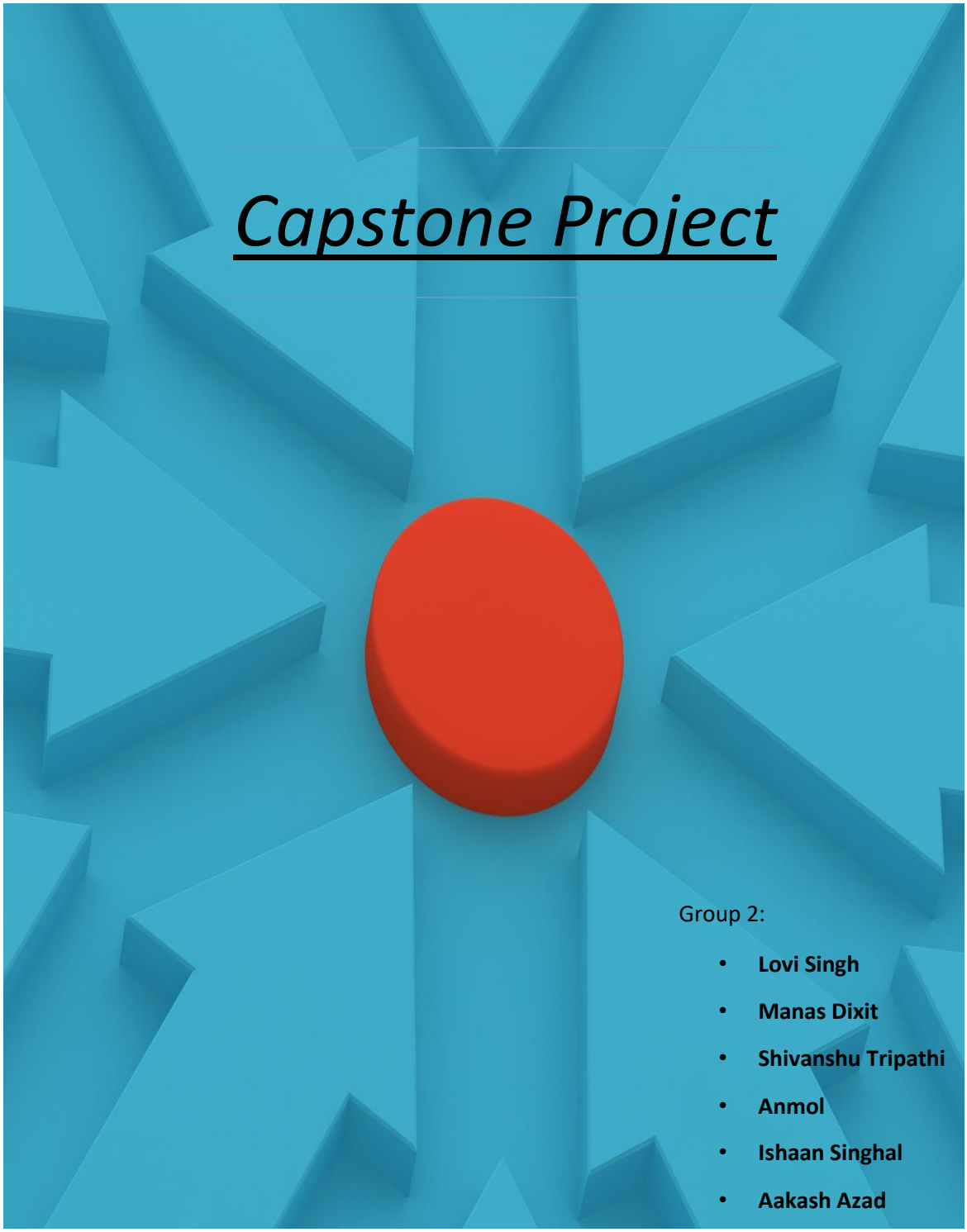
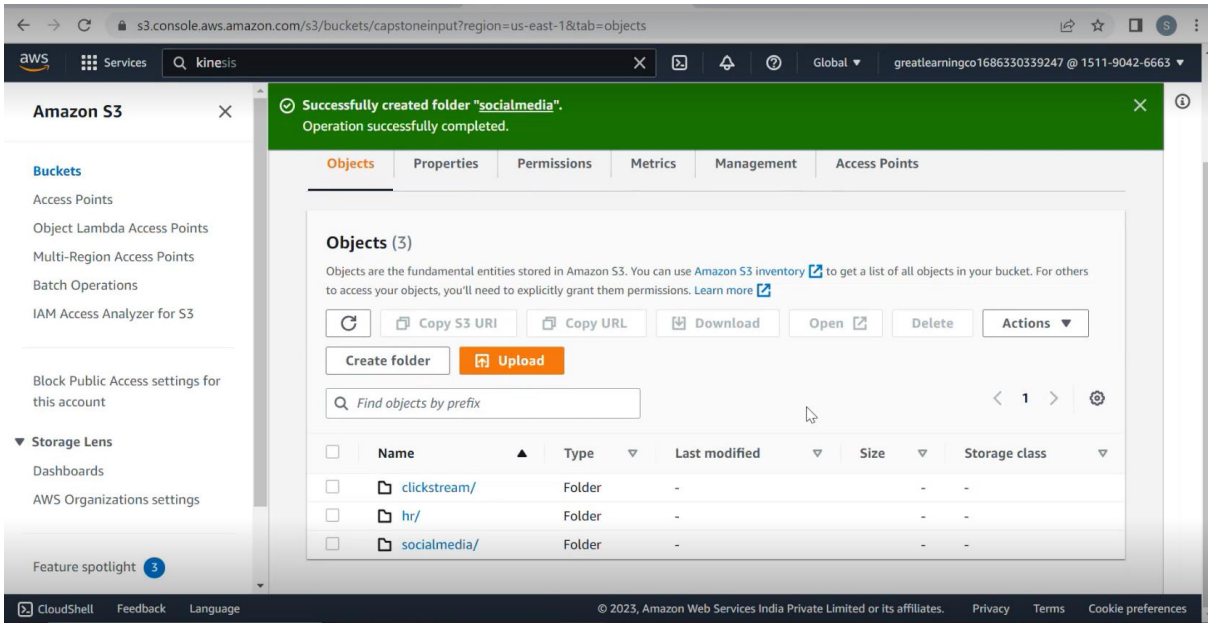


This report contains all the required screenshots as required.
The detailed explanation of each step is given in the attached PowerPoint.

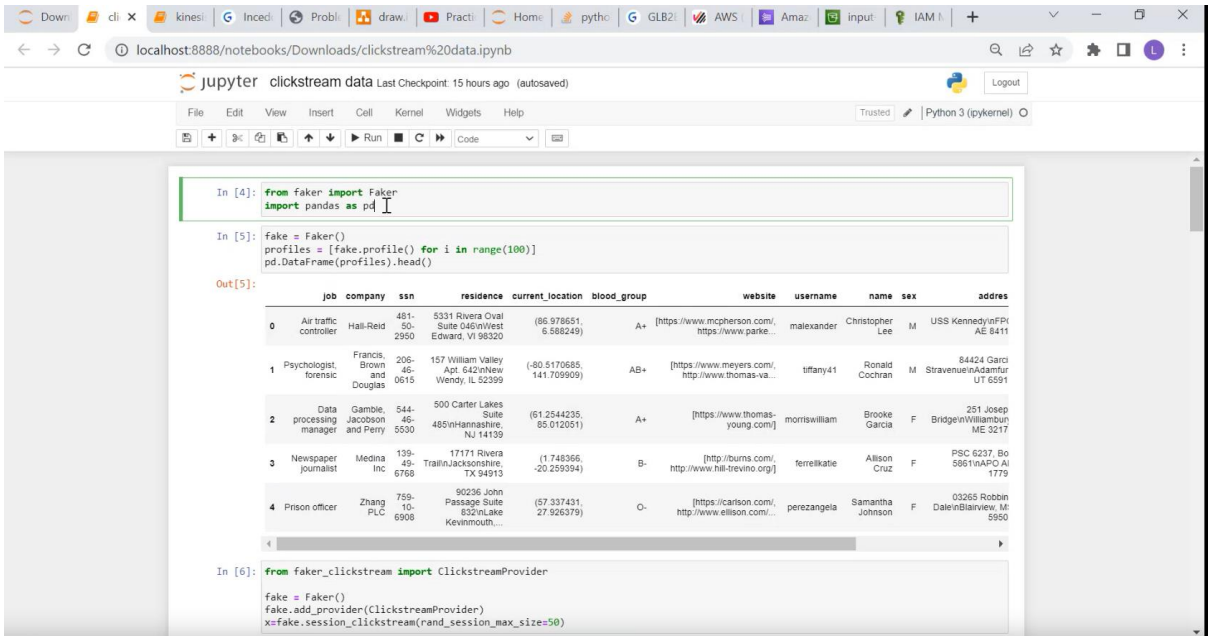
S3 bucket

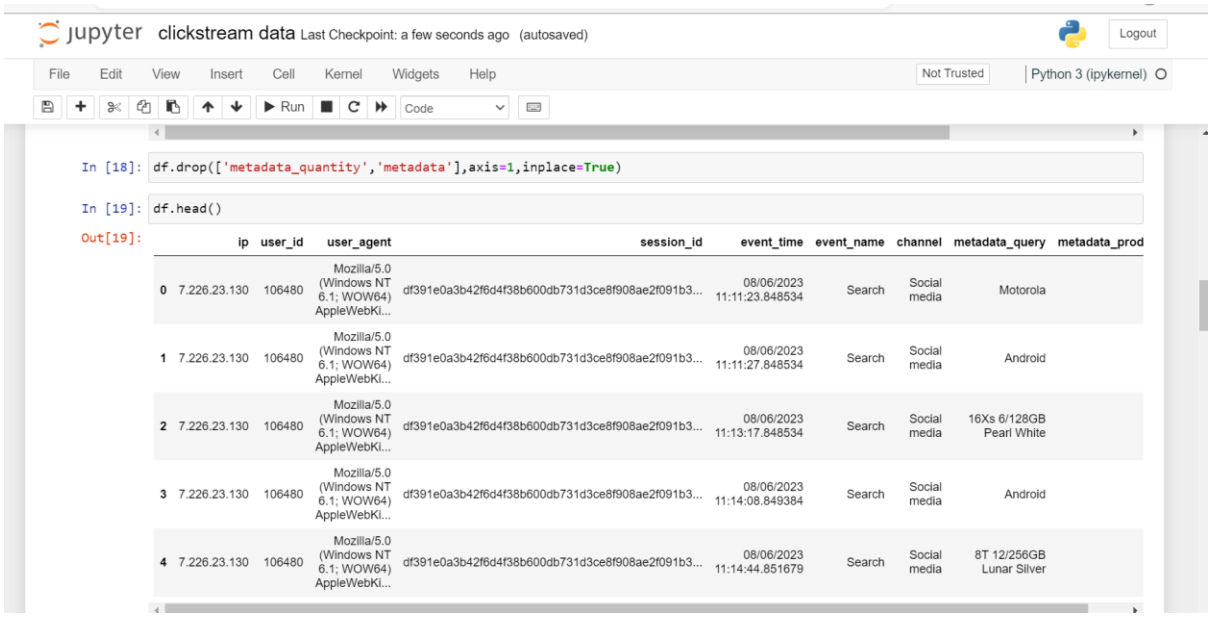
Clickstream and social media is coming from Kinesis and hr data is given



CLICKSTREAM DATA

Data is coming from kinesis



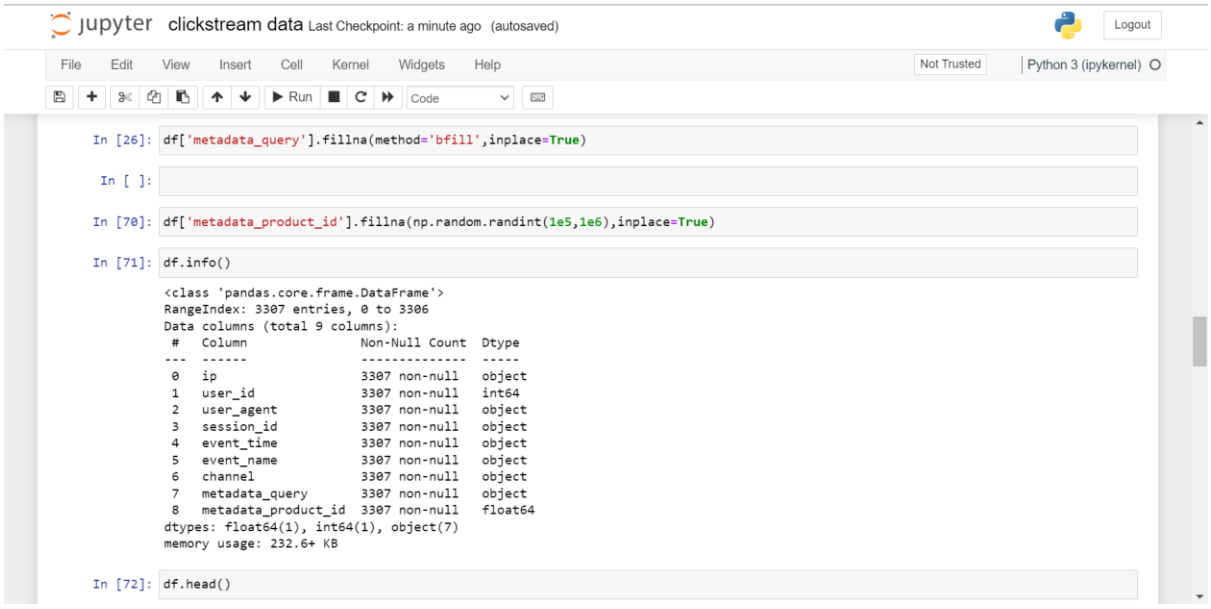


In [18]:
```
df.drop(['metadata_quantity','metadata'],axis=1,inplace=True)
```

In [19]:
```
df.head()
```

Out[19]:

	ip	user_id	user_agent	session_id	event_time	event_name	channel	metadata_query	metadata_prod
0	7.226.23.130	106480	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:11:23.848534	Search	Social media	Motorola	
1	7.226.23.130	106480	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:11:27.848534	Search	Social media	Android	
2	7.226.23.130	106480	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:13:17.848534	Search	Social media	16Xs 6/128GB Pearl White	
3	7.226.23.130	106480	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:14:08.849384	Search	Social media	Android	
4	7.226.23.130	106480	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:14:44.851679	Search	Social media	8T 12/256GB Lunar Silver	



In [26]:
```
df['metadata_query'].fillna(method='bfill',inplace=True)
```

In [ ]:
```
```

In [70]:
```
df['metadata_product_id'].fillna(np.random.randint(1e5,1e6),inplace=True)
```

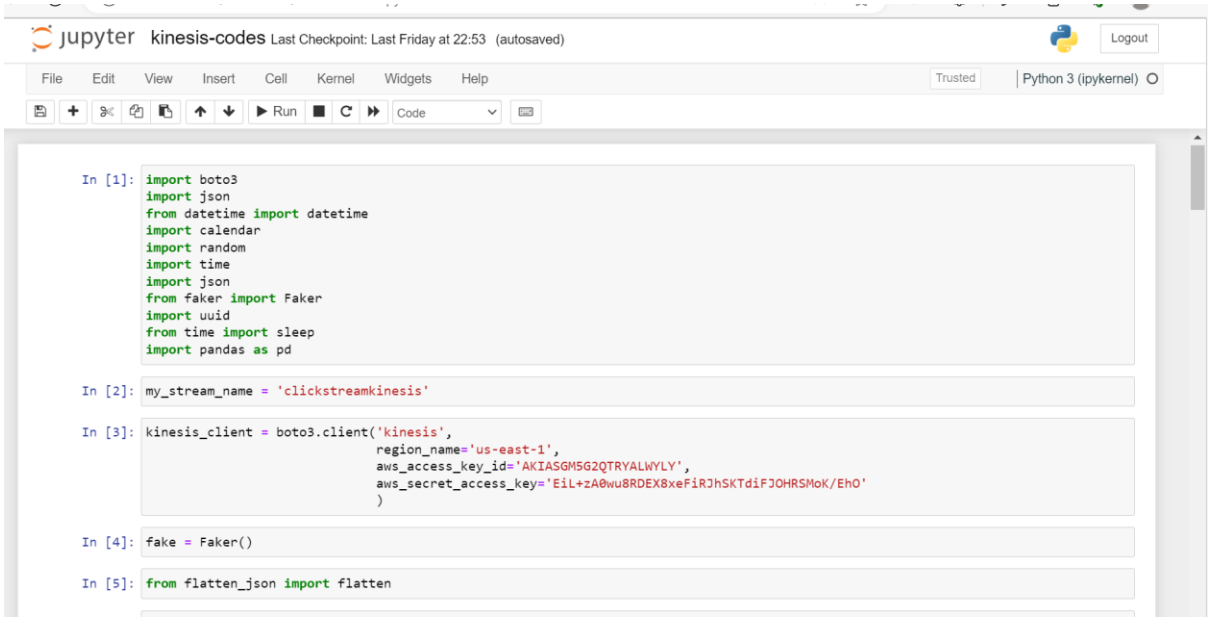
In [71]:
```
df.info()
```

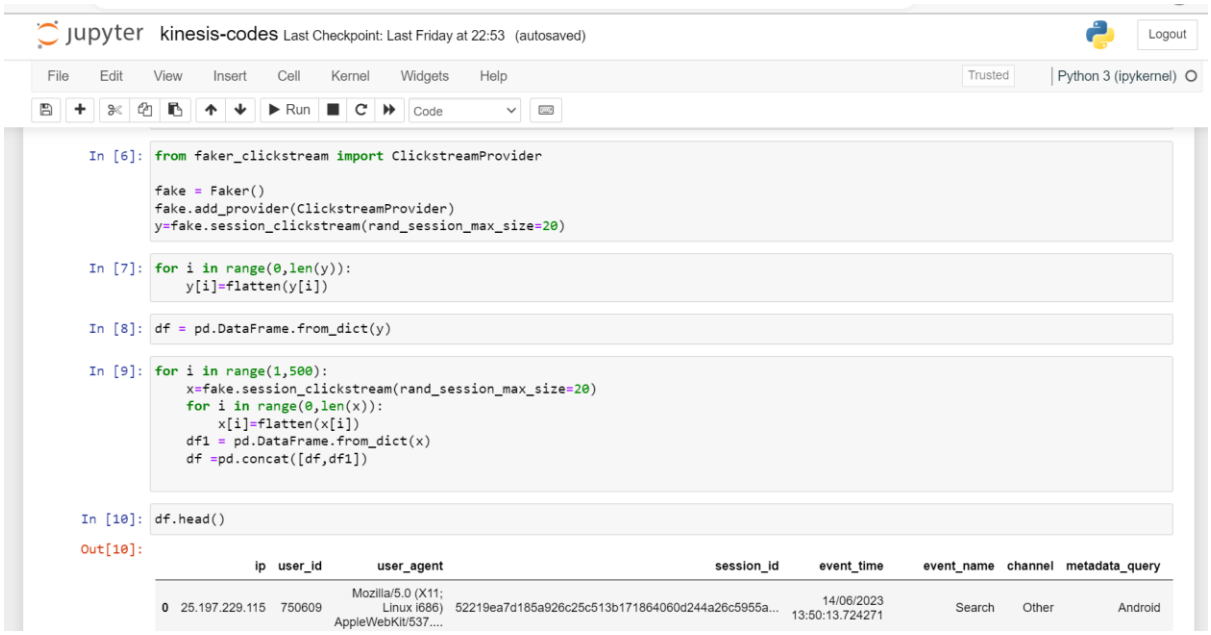
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3307 entries, 0 to 3306
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ip                  3307 non-null   object
 1   user_id             3307 non-null   int64
 2   user_agent          3307 non-null   object
 3   session_id          3307 non-null   object
 4   event_time          3307 non-null   object
 5   event_name          3307 non-null   object
 6   channel             3307 non-null   object
 7   metadata_query      3307 non-null   object
 8   metadata_product_id  3307 non-null   float64
dtypes: float64(1), int64(1), object(7)
memory usage: 232.6+ KB
```

In [72]:
```
df.head()
```

KINESIS CODE

Sending API data to Firehose





	ip	user_id	user_agent	session_id	event_time	event_name	channel	metadata_query
0	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:50:13.724271	Search	Other	Android

jupyter kinesis-codes Last Checkpoint: Last Friday at 22:53 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

```
In [11]: df.drop(['metadata_quantity'],axis=1,inplace=True)

In [12]: df['metadata_query'].fillna(method='bfill',inplace=True)

In [13]: import numpy as np

In [14]: df['metadata_product_id'].fillna(np.random.randint(1e5,1e6),inplace=True)

In [ ]:

In [15]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5139 entries, 0 to 17
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ip                   5139 non-null   object
 1   user_id              5139 non-null   int64
 2   user_agent           5139 non-null   object
 3   session_id           5139 non-null   object
 4   event_time           5139 non-null   object
 5   event_name           5139 non-null   object
 6   channel              5139 non-null   object
 7   metadata_query       5139 non-null   object
```

jupyter kinesis-codes Last Checkpoint: Last Friday at 22:53 (autosaved)

Logout

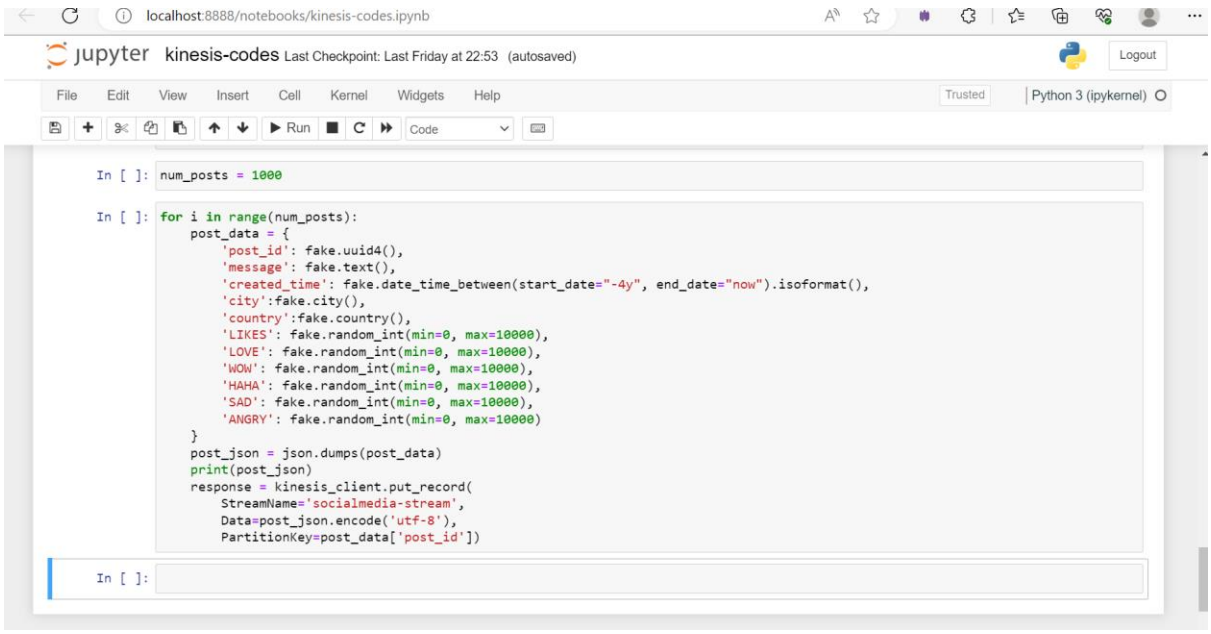
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

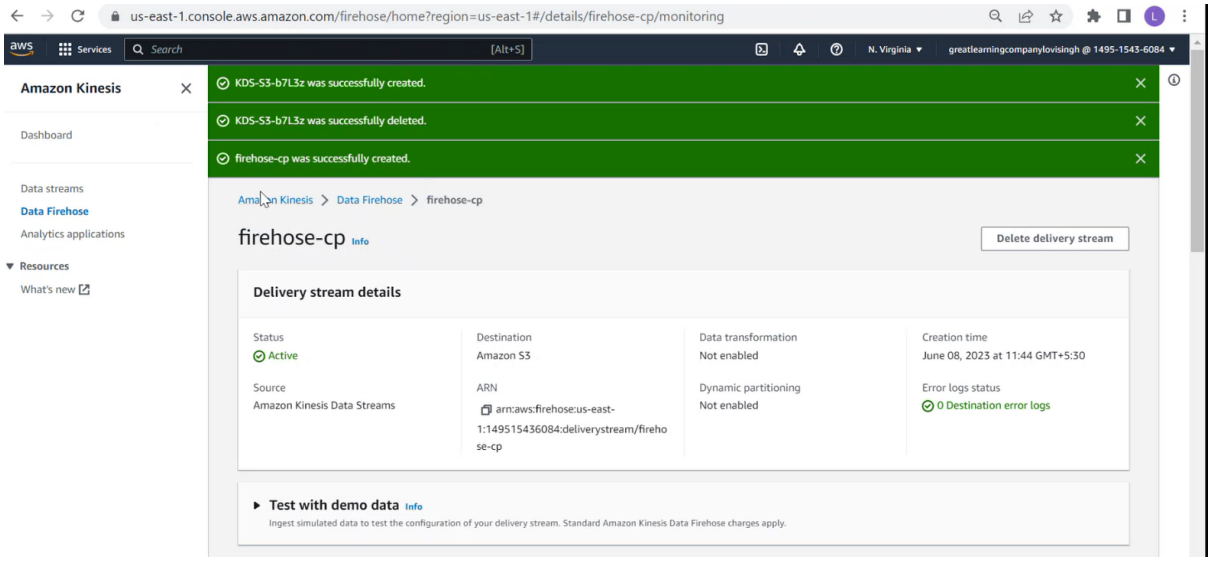
	ip	user_id	user_agent	session_id	event_time	event_name	channel	metadata_query
0	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:50:13.724271	Search	Other	Android
1	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:50:46.724271	IncreaseQuantity	Other	iOS
2	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:51:42.724271	DecreaseQuantity	Other	iOS
3	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:51:54.724271	Search	Other	iOS
4	25.197.229.115	750609	Mozilla/5.0 (X11; Linux i686) AppleWebKit/537....	52219ea7d185a926c25c513b171864060d244a26c5955a...	14/06/2023 13:53:42.724271	DecreaseQuantity	Other	Android

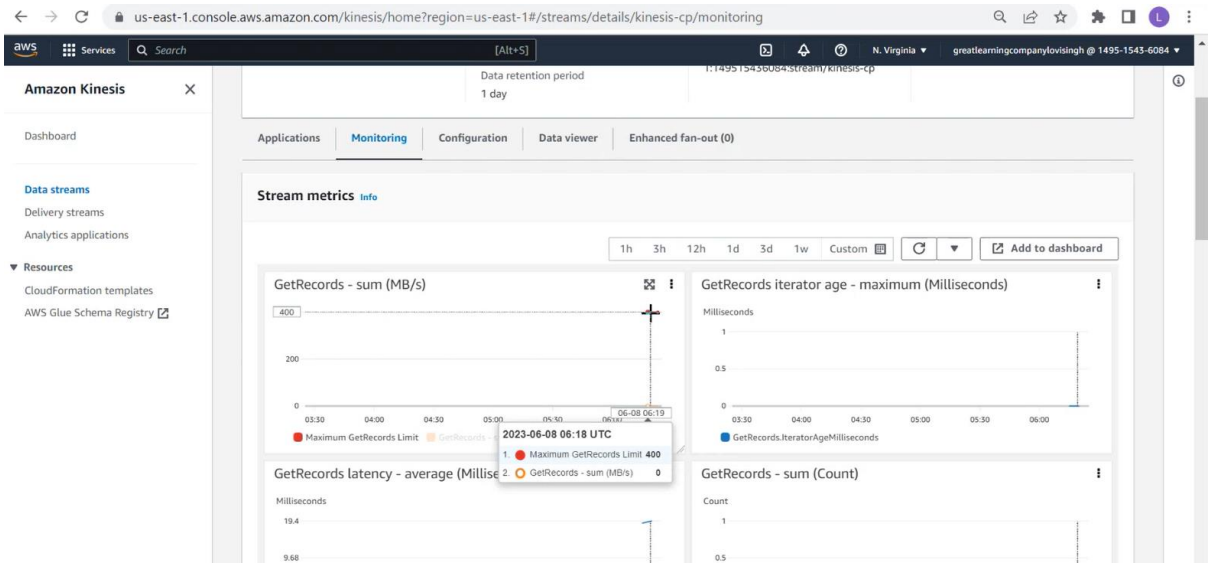
```
In [17]: df['metadata_product_id']=df['metadata_product_id'].astype('int')

In [18]: for i in range(0,df.shape[0]):
             pk=df.columns[4]
             json_data=df.iloc[i].to_dict()
             final_data=json.dumps(json_data)+'\n'
             print(json_data)
             kinesis_client.put_record(StreamName=my_stream_name,Data=final_data,PartitionKey=pk)
             sleep(3)
```

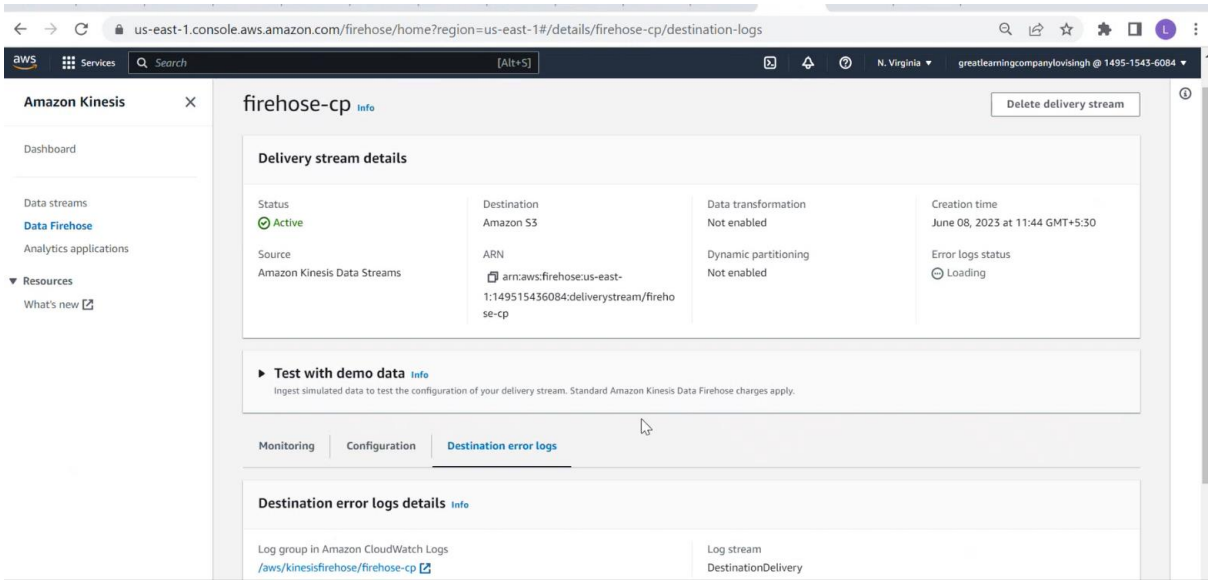


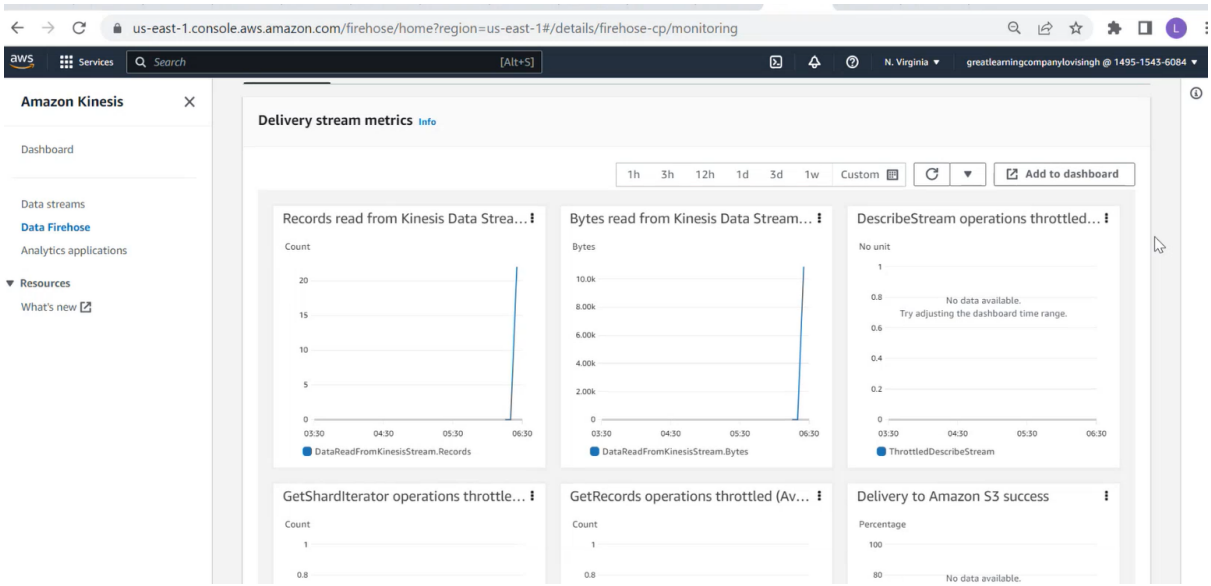
FireHose

Sending data to s3 bucket via firehose



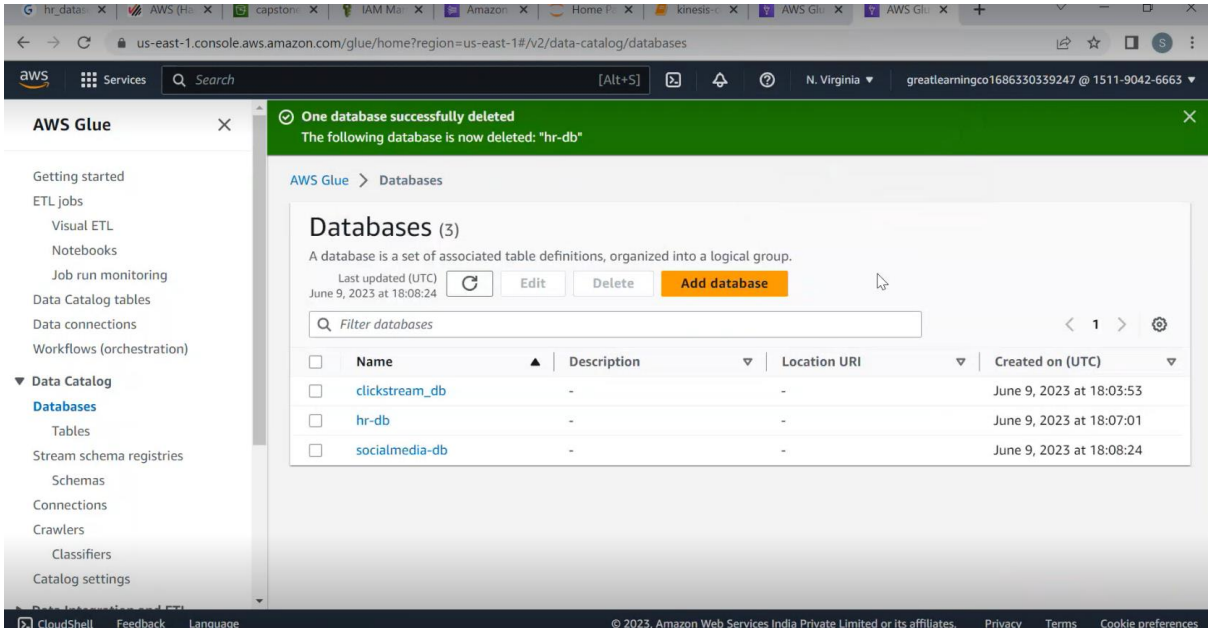

us-east-1.console.aws.amazon.com/kinesis/home?region=us-east-1#/streams/details/kinesis-cp/monitoring
Amazon Kinesis
Dashboard
Data streams
Delivery streams
Analytics applications
Resources
CloudFormation templates
AWS Glue Schema Registry
Data retention period
1 day
Applications
Monitoring
Configuration
Data viewer
Enhanced fan-out (0)
Stream metrics Info
1h 3h 12h 1d 3d 1w Custom
Add to dashboard
GetRecords - sum (MB/s)
GetRecords iterator age - maximum (Milliseconds)
Milliseconds
Maximum GetRecords Limit
GetRecords.IteratorAgeMilliseconds
2023-06-08 06:18 UTC
Maximum GetRecords Limit 400
GetRecords - sum (MB/s) 0
GetRecords latency - average (Millise
GetRecords - sum (Count)
Count


us-east-1.console.aws.amazon.com/firehose/home?region=us-east-1#/details/firehose-cp/destination-logs
Amazon Kinesis
Dashboard
Data streams
Data Firehose
Analytics applications
Resources
What's new
firehose-cp Info
Delete delivery stream
Delivery stream details
Status
Active
Source
Amazon Kinesis Data Streams
Destination
Amazon S3
ARN
arn:aws:firehose:us-east-1:149515436084:deliverystream/firehose-cp
Data transformation
Not enabled
Dynamic partitioning
Not enabled
Creation time
June 08, 2023 at 11:44 GMT+5:30
Error logs status
Loading
Test with demo data Info
Ingest simulated data to test the configuration of your delivery stream. Standard Amazon Kinesis Data Firehose charges apply.
Monitoring
Configuration
Destination error logs
Destination error logs details Info
Log group in Amazon CloudWatch Logs
/aws/kinesisfirehose/firehose-cp
Log stream
DestinationDelivery


us-east-1.console.aws.amazon.com/firehose/home?region=us-east-1#/details/firehose-cp/monitoring
Amazon Kinesis
Dashboard
Data streams
Data Firehose
Analytics applications
Resources
What's new
Delivery stream metrics Info
1h 3h 12h 1d 3d 1w Custom
Add to dashboard
Records read from Kinesis Data Strea...
Count
DataReadFromKinesisStream.Records
Bytes read from Kinesis Data Stream...
Bytes
DataReadFromKinesisStream.Bytes
DescribeStream operations throttled...
No unit
No data available.
Try adjusting the dashboard time range.
ThrottledDescribeStream
GetShardIterator operations throttle...
GetRecords operations throttled (Av...
Delivery to Amazon S3 success
Percentage
No data available.

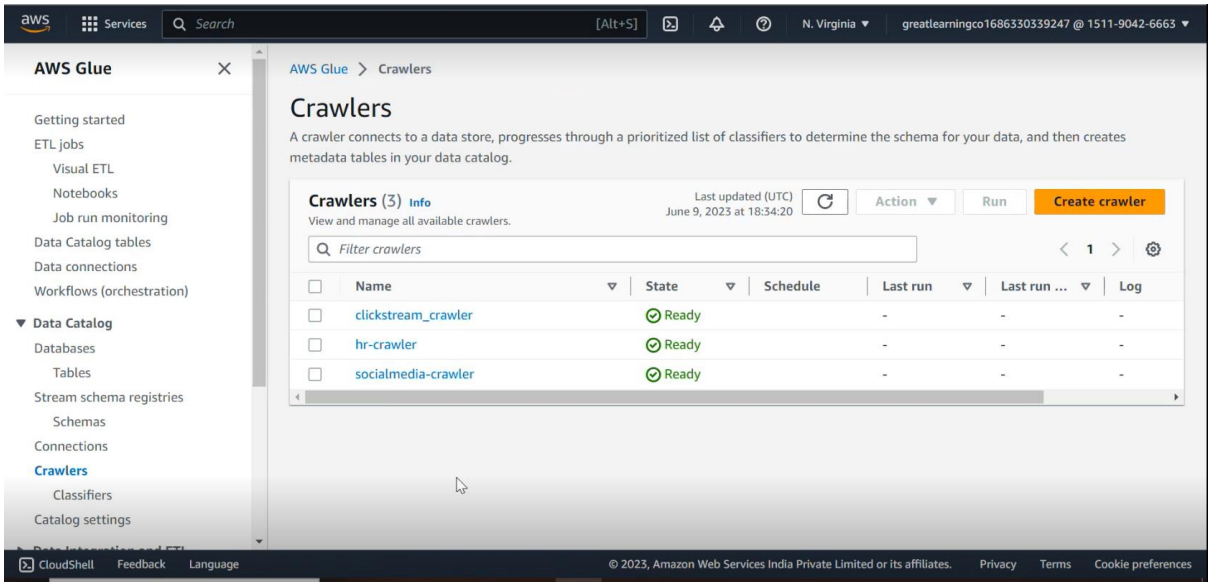
DATABASES

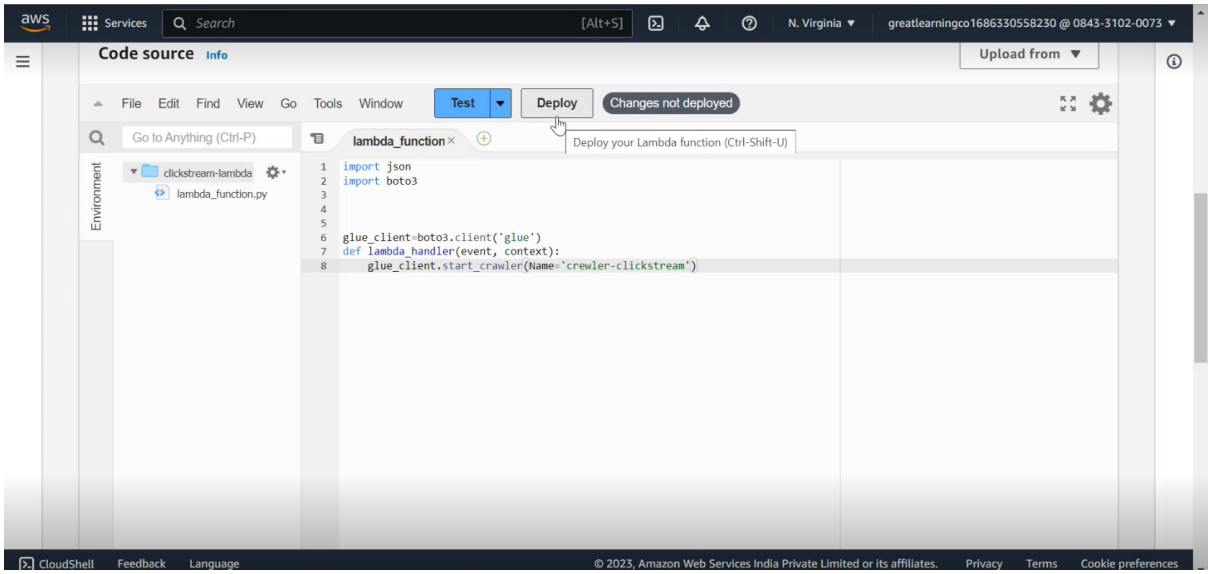
Created database to store table generated by crwaler

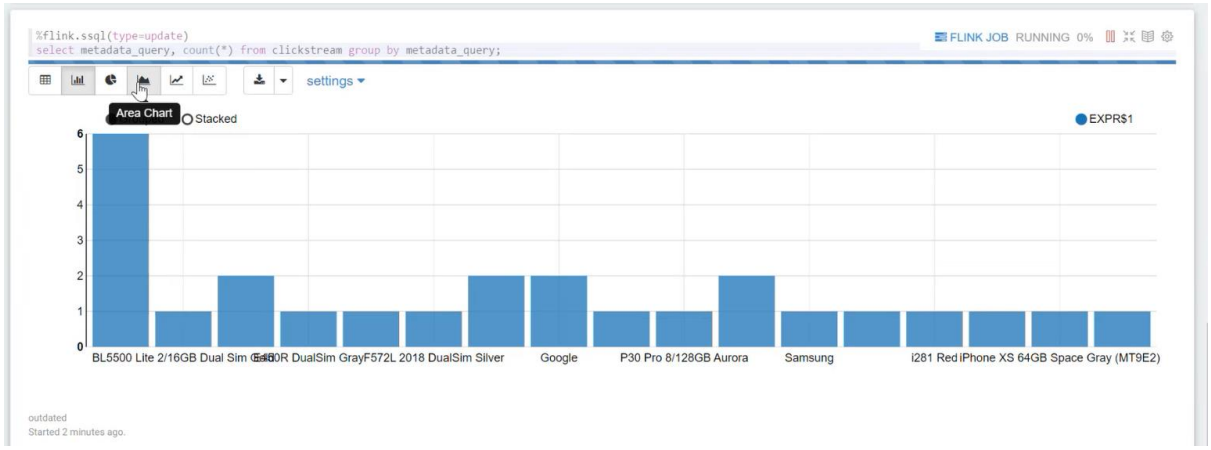


Crawlers

used it to create table

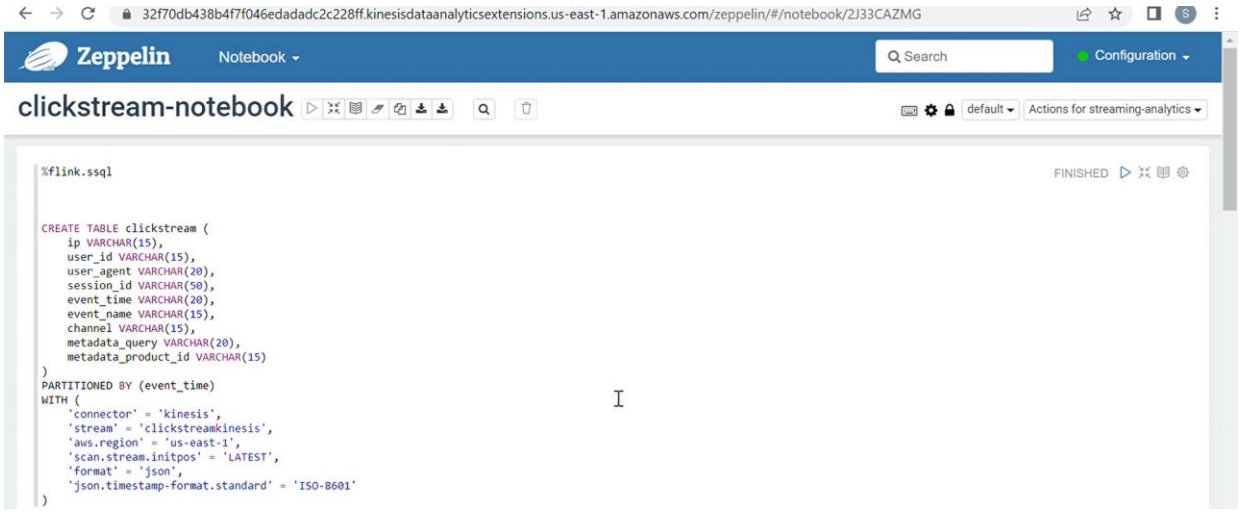






Zeppelin

used it to analyse socialmedia data



```
%flink.ssql(type=update)
SELECT * FROM clickstream;
```

FLINK JOB ABORT

settings

ip	user_id	user_agent	session_id	event_time	event_name	channel	metadata_query
1.186.17.84	839897	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.130 Safari/537.36	d23dd6d9d4c884ebca8b654f1f211a03ae852dab0ae059fde793dc5d4217ae60	10/06/2023 12:06:57.603498	Search	Social media	Galaxy M115 M11 3/32 Black (SM-M115FZKN)
129.217.167.211	688151	Mozilla/5.0 (Linux; Android 5.0.1; SAMSUNG-SGH-I337 Build/LRX22C) AppleWebKit/537.36 (KHTML, like Gecko)	e065378dfa94c280cbaa3dc7842b338e2e3266a3af7ca44f2120664cd54140ba	10/06/2023 11:21:29.607507	AddToCart	Direct	Motorola

Took 4 min 12 sec. Last updated by anonymous at June 10 2023, 11:49:40 AM. (outdated)

```
%flink.ssql
Drop table clickstream;
```

ERROR

Took 1 sec. Last updated by anonymous at June 10 2023, 11:29:06 AM.

```
%flink.ssql
DROP TABLE clickstream;
```

FINISHED

Took 44 sec. Last updated by anonymous at June 10 2023, 11:41:02 AM.

+ Add Paragraph

```
%flink.ssql(type=update)
select metadata_query, count(*) from clickstream group by metadata_query;
```

FLINK JOB RUNNING 0%

settings

Android BL5500 Lite 2/16GB D... Blade A7 2019 2/32GB... E450R DualSim Gray F284 Balance Dual Si... F572L 2018 DualSim S... Mi 10T Lite 6/128GB ...

Took 44 sec. Last updated by anonymous at June 10 2023, 11:41:02 AM.

```
%flink.ssql(type=update)
select metadata_query, count(*) from clickstream group by metadata_query;
```

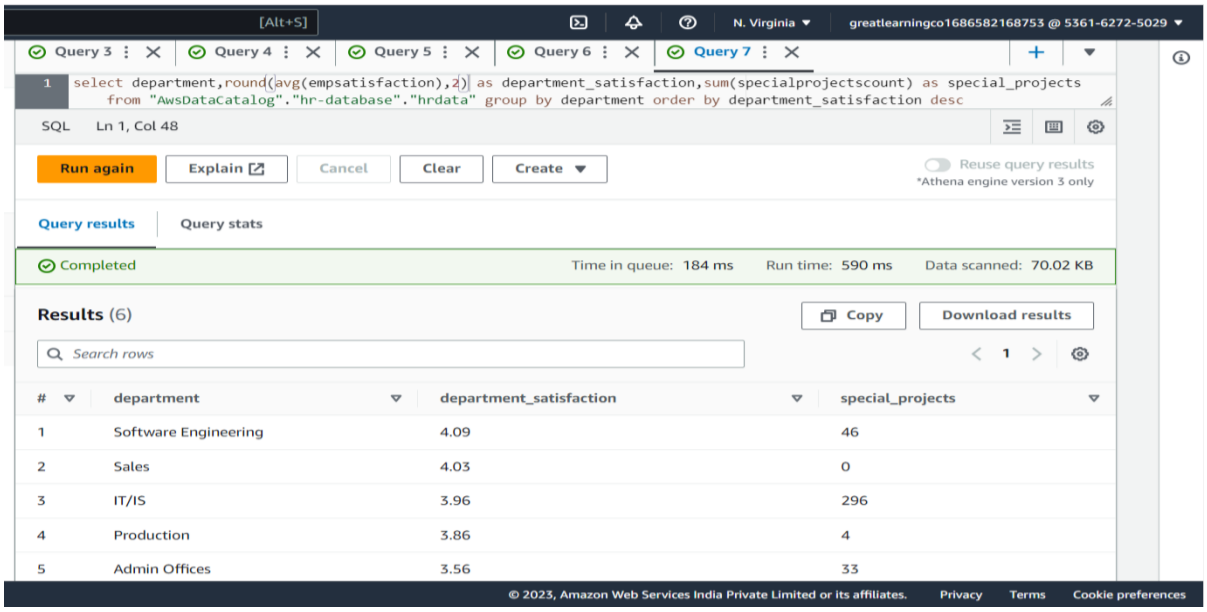
FLINK JOB RUNNING 0%

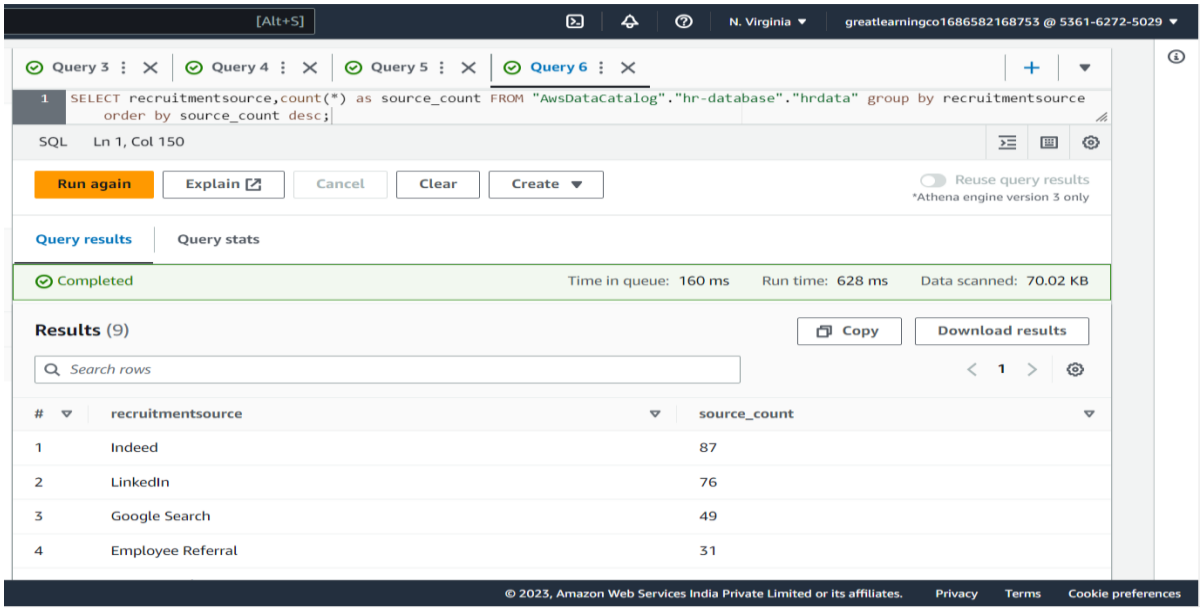
settings

Android BL5500 Lite 2/16GB D... Blade A7 2019 2/32GB... E450R DualSim Gray F284 Balance Dual Si... F572L 2018 DualSim S... Mi 10T Lite 6/128GB ...
Redmi Note 8 Pro 6/6... Samsung Sigma mobile iPhone SE 2020 64GB ... iPhone XS 64GB Space... x-style 35 Screen

Athena

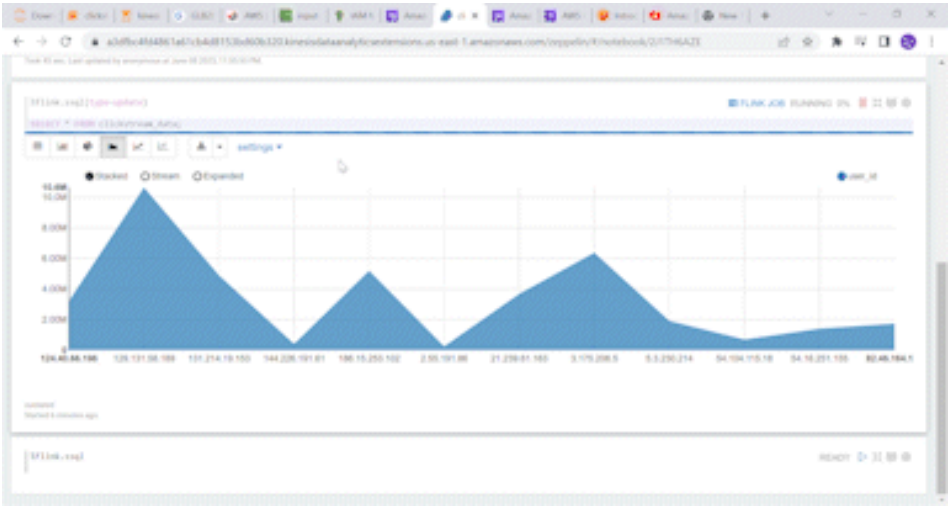
Used it to query streamed data store in s3



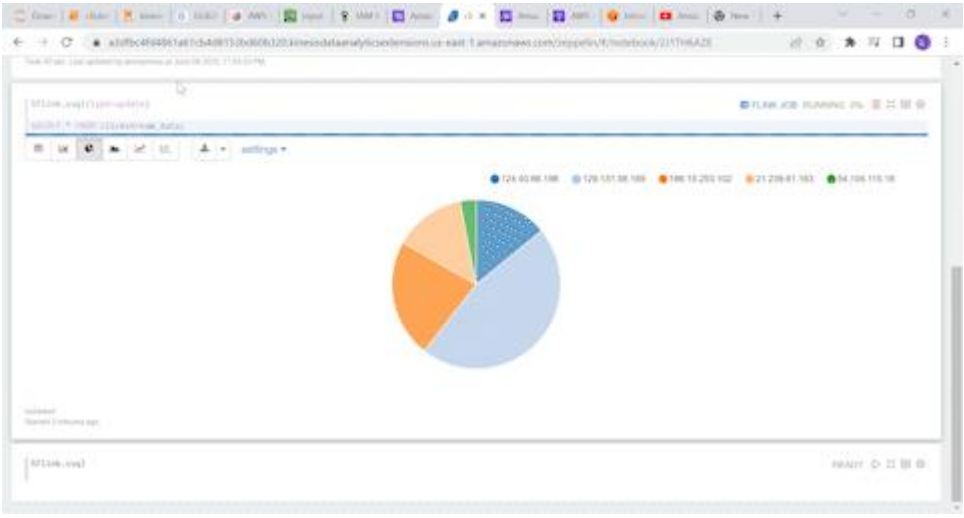


Kinesis Real Time Analysis

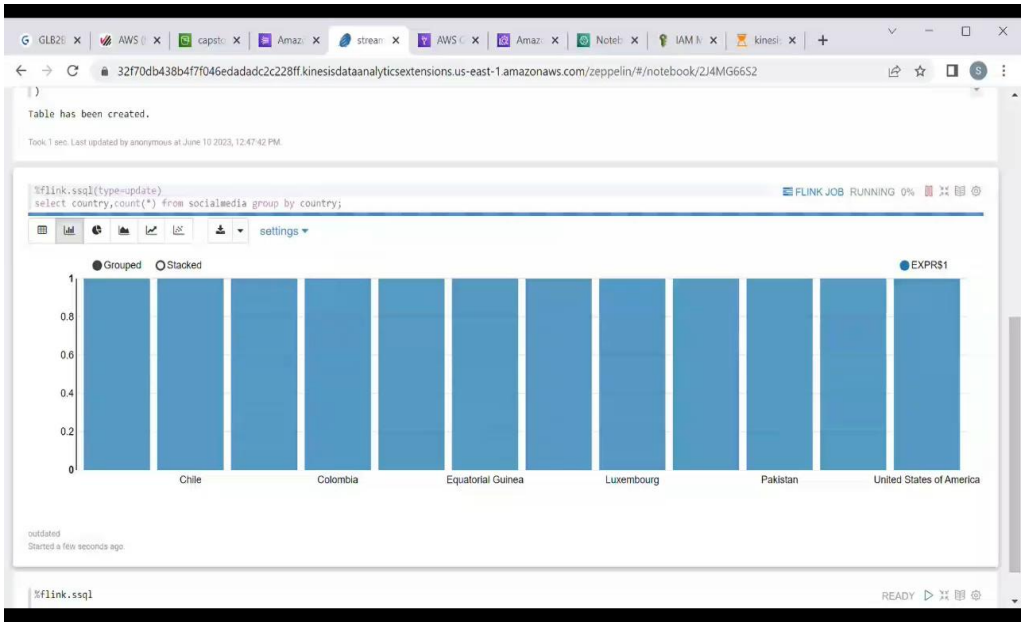
live population of IP address in database



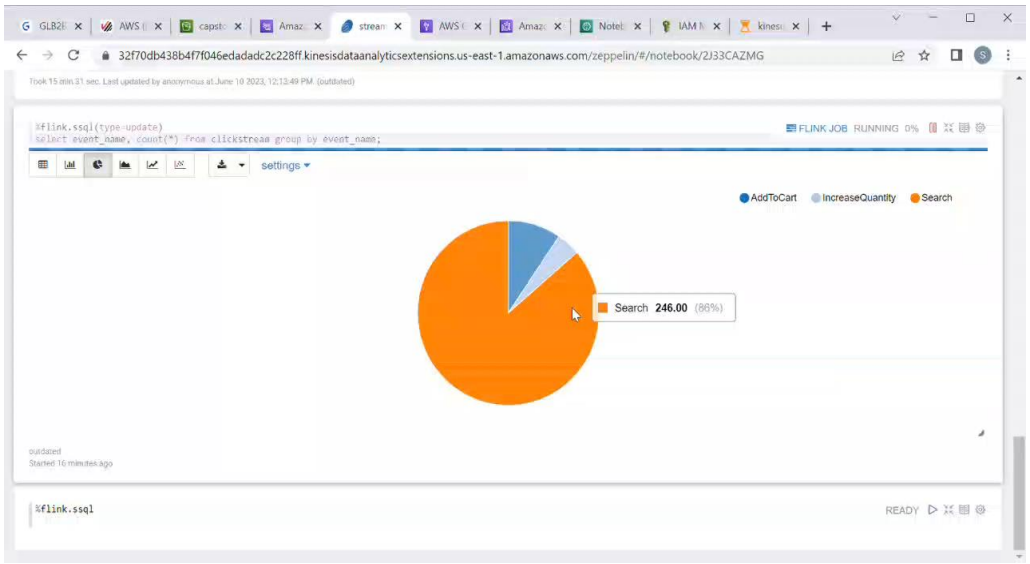
Visualization of IP address populating in Pie Chart



Live traffic on ABC corp website in different countries.

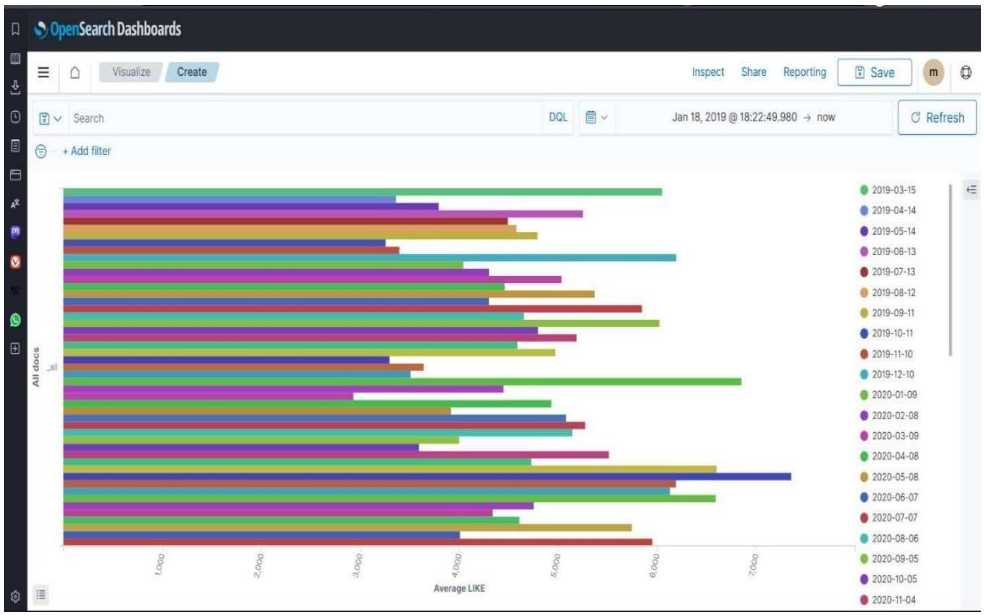


Live User activity on ABC Corporation website

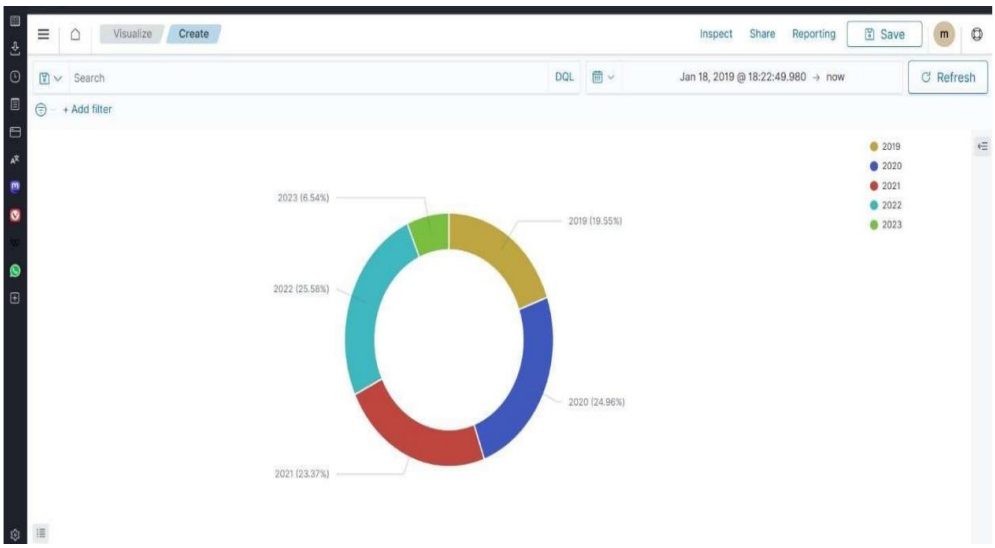


Open Search Dashboard

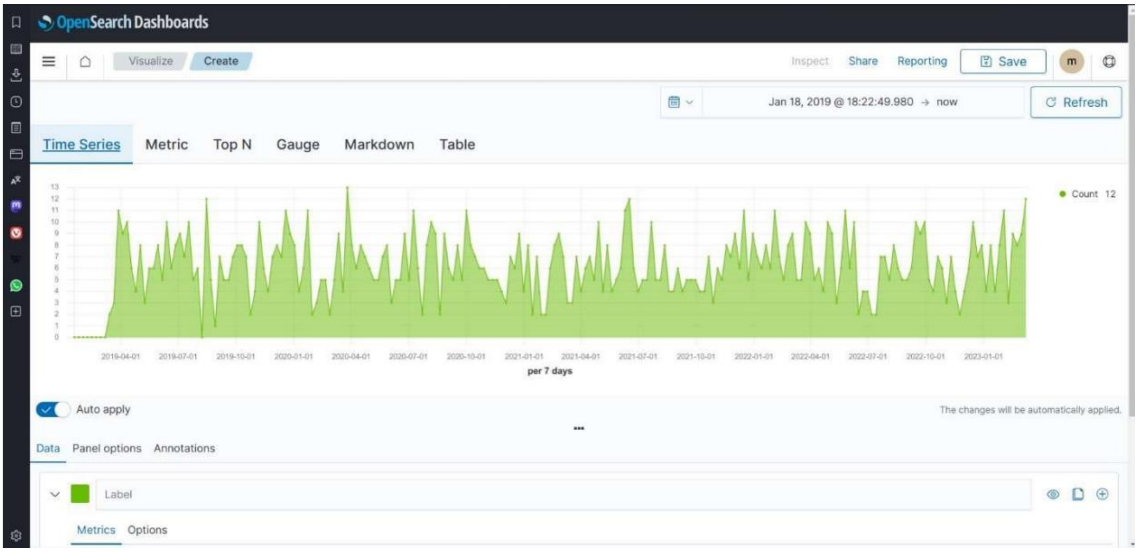
Visualization of Average on monthly basis over the year



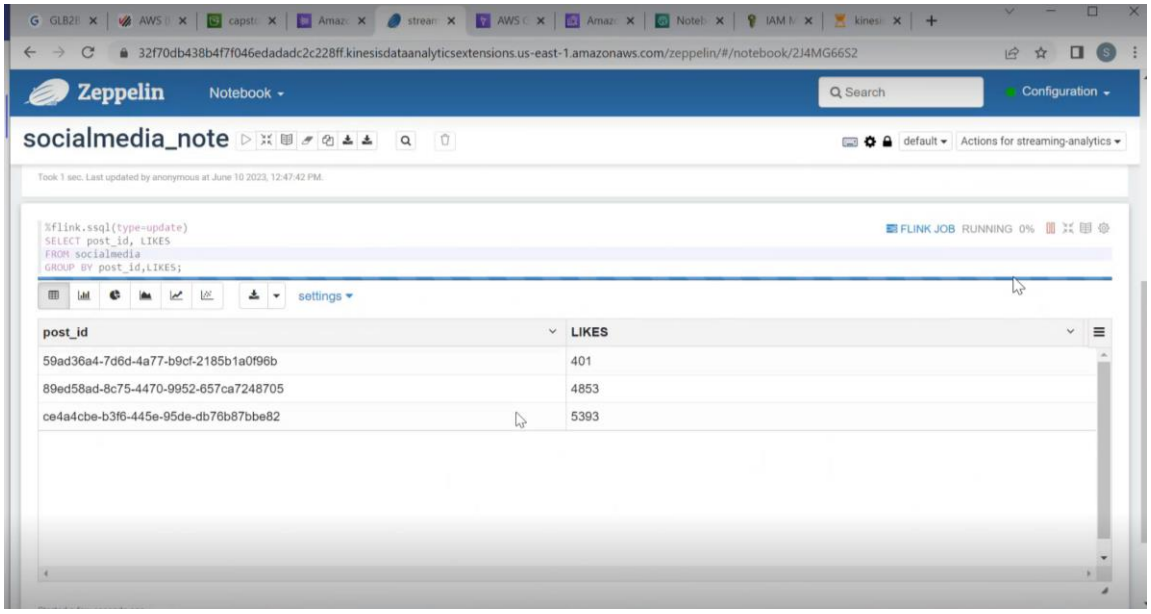
Likes count over the years depicted as percentage



User activity over the years



The data below showing the sum of likes as percentage distribution over the years live.



All the required analysis were performed as per problem statement requirements.

Thank You ☺