



Capstone Project

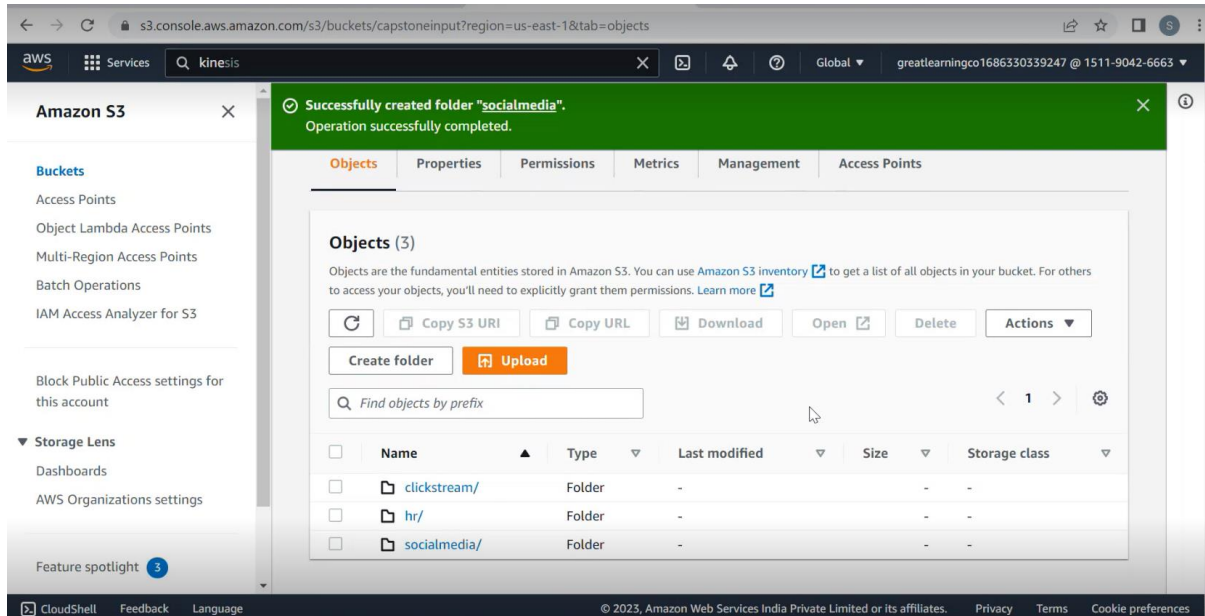
Group 2:

- Lovi Singh
- Manas Dixit
- Shivanshu Tripathi
- Anmol
- Ishaan Singhal
- Aakash Azad

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

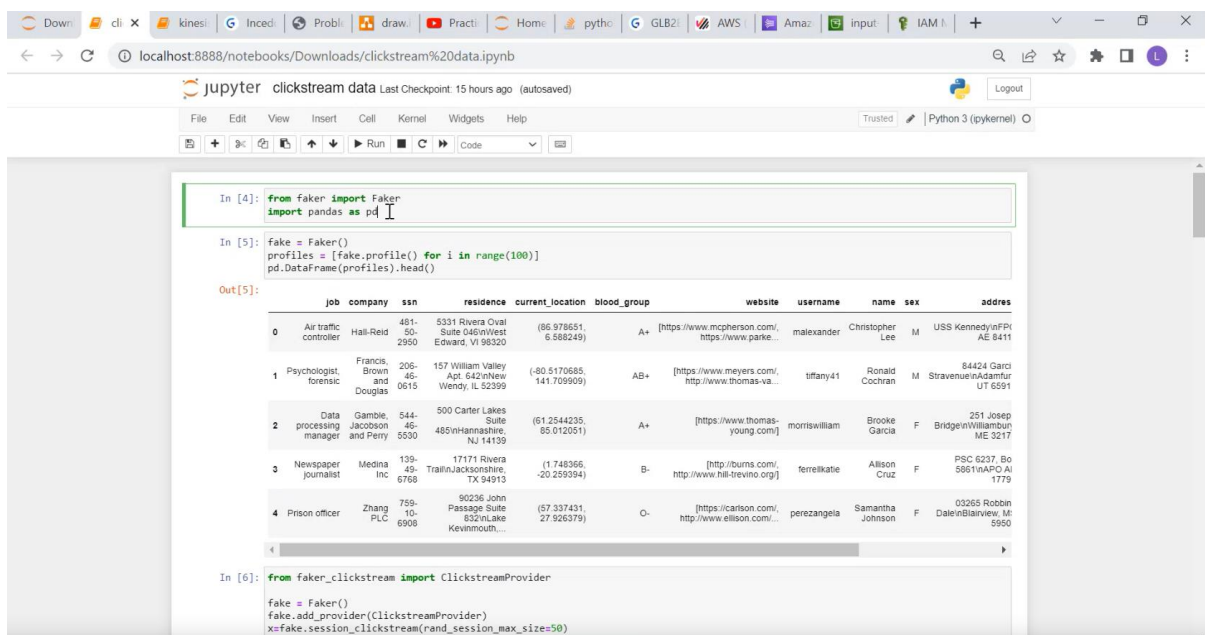
S3 bucket

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



CLICKSTREAM DATA

Data is coming from kinesis



jupyter clickstream data Last Checkpoint: a few seconds ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

```
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) AppleWebKit...	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) AppleWebKit...	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) AppleWebKit...	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) AppleWebKit...	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) AppleWebKit...	df391e0a3b42f6d4f38b600db731d3ce8f908ae2f091b3...	08/06/2023 11:14:44.851679	Search	Social media	8T 12/256GB Lunar Silver	

jupyter clickstream data Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

```
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

jupyterkinesis-codesLast Checkpoint: Last Friday at 22:53 (autosaved)

FileEditViewInsertCellKernelWidgetsHelpTrustedPython 3 (ipykernel)

In [1]:

import boto3
import json
from datetime import datetime
import calendar
import random
import time
import json
from faker import Faker
import uuid
from time import sleep
import pandas as pd

In [2]:

my_stream_name = 'clickstreamkinesis'

In [3]:

kinesis_client = boto3.client('kinesis',
 region_name='us-east-1',
 aws_access_key_id='AKIASGM5G2QTRYALWYLY',
 aws_secret_access_key='EiL+zA0wu8RDEX8xeFiRjhSKTdifJOHRSMoK/EhO'
)

In [4]:

fake = Faker()

In [5]:

from flatten_json import flatten

In [6]:

from faker_clickstream import ClickstreamProvider

fake = Faker()
fake.add_provider(ClickstreamProvider)
y=fake.session_clickstream(rand_session_max_size=20)

In [7]:

for i in range(0,len(y)):
 y[i]=flatten(y[i])

In [8]:

df = pd.DataFrame.from_dict(y)

In [9]:

for i in range(1,500):
 x=fake.session_clickstream(rand_session_max_size=20)
 for i in range(0,len(x)):
 x[i]=flatten(x[i])
 df1 = pd.DataFrame.from_dict(x)
 df =pd.concat([df,df1])

In [10]:

df.head()

Out[10]:

	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....	52219ea7d185a926c25c513b7171864060d244a26c5955a...	14/06/2023 13:50:13.724271	Search	Other	Android

```
jupyter kinesis-codes Last Checkpoint: Last Friday at 22:53 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
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
8 metadata_product_id 5139 non-null object
```

```
jupyter kinesis-codes Last Checkpoint: Last Friday at 22:53 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
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)
```

```
localhost:8888/notebooks/kinesis-codes.ipynb
jupyter kinesis-codes Last Checkpoint: Last Friday at 22:53 (autosaved)
Python 3 (ipykernel)

In [ ]: num_posts = 1000

In [ ]: for i in range(num_posts):
    post_data = {
        'post_id': fake.uuid4(),
        'message': fake.text(),
        'created_time': fake.date_time_between(start_date="-4y", end_date="now").isoformat(),
        'city': fake.city(),
        'country': fake.country(),
        'LIKES': fake.random_int(min=0, max=10000),
        'LOVE': fake.random_int(min=0, max=10000),
        'WOW': fake.random_int(min=0, max=10000),
        'HAHA': fake.random_int(min=0, max=10000),
        'SAD': fake.random_int(min=0, max=10000),
        'ANGRY': fake.random_int(min=0, max=10000)
    }
    post_json = json.dumps(post_data)
    print(post_json)
    response = kinesis_client.put_record(
        StreamName='socialmedia-stream',
        Data=post_json.encode('utf-8'),
        PartitionKey=post_data['post_id'])

In [ ]:
```

FireHose

Sending data to s3 bucket via firehose

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

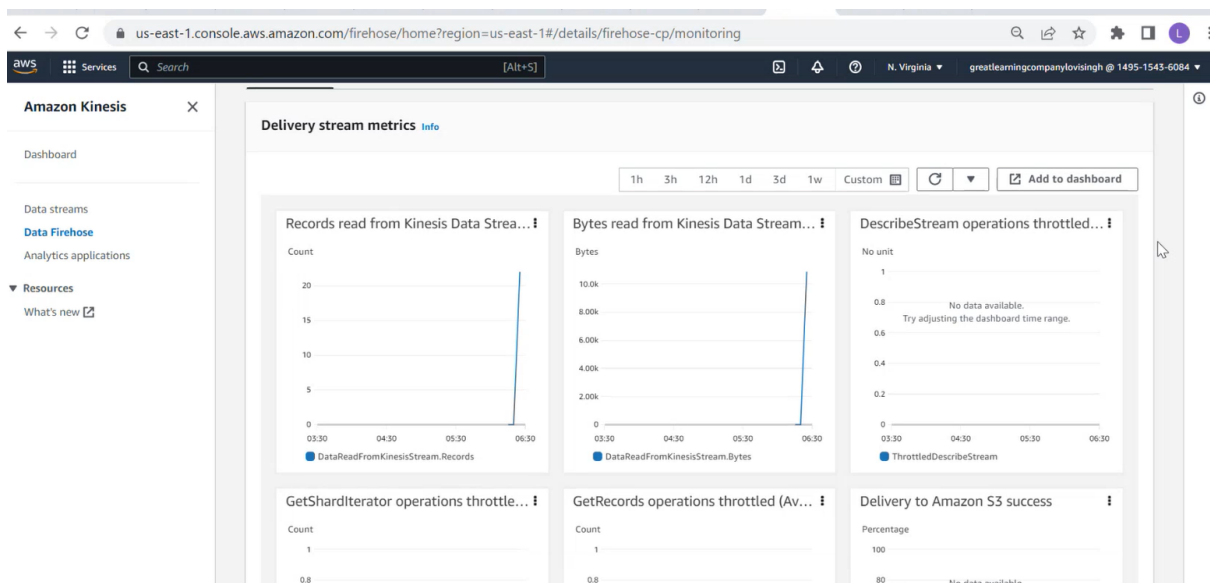
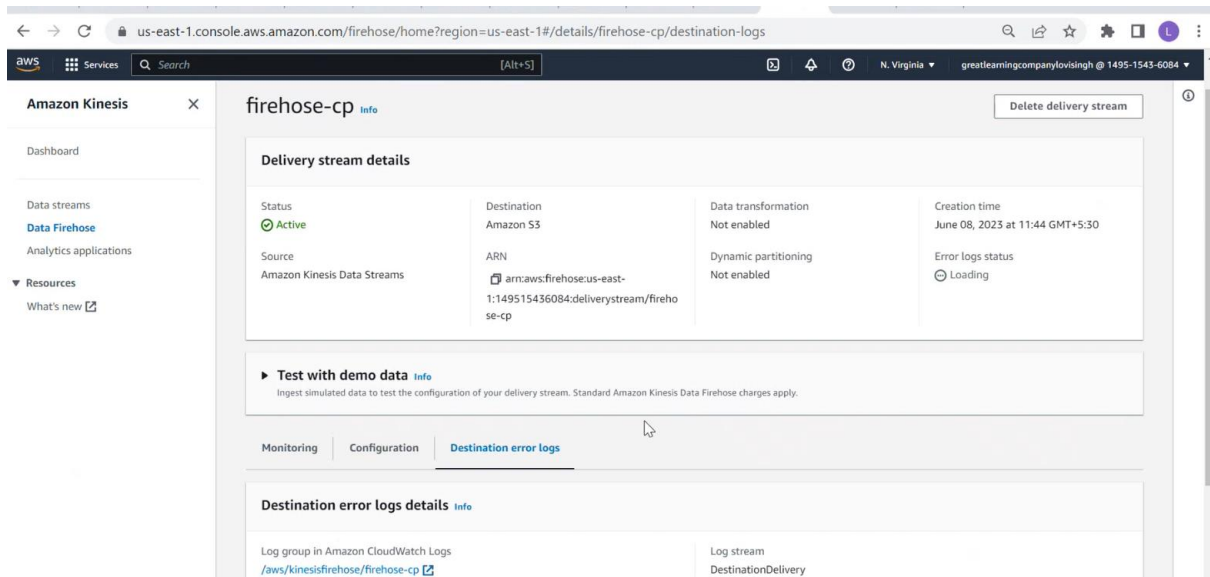
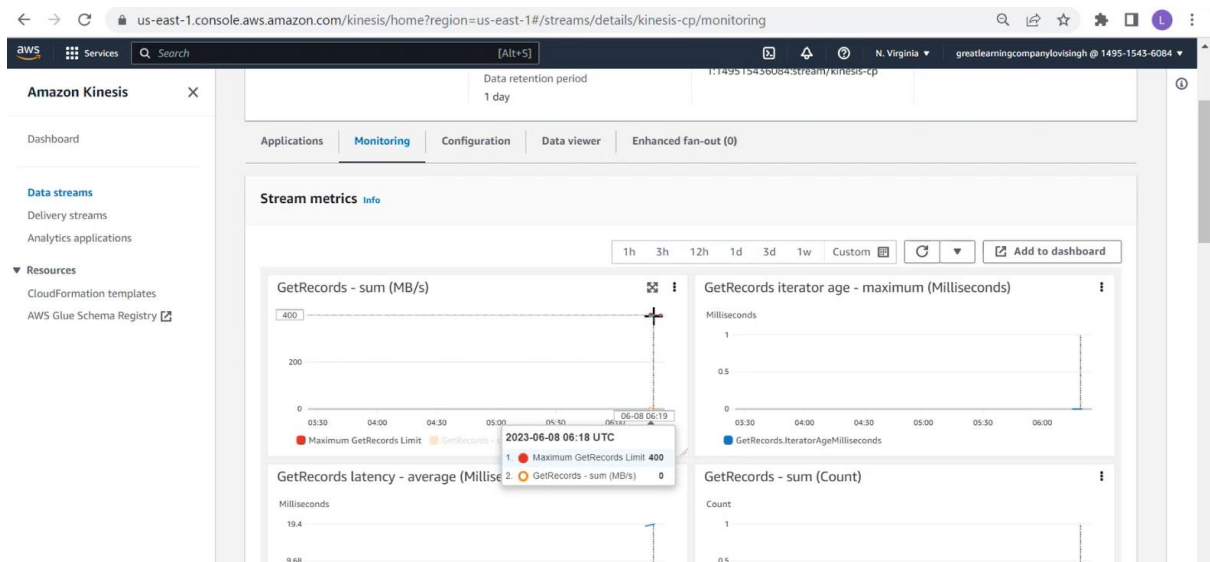
firehose-cp

Delivery stream details

Status	Destination	Data transformation	Creation time
Active	Amazon S3	Not enabled	June 08, 2023 at 11:44 GMT+5:30
Source	ARN	Dynamic partitioning	Error logs status
Amazon Kinesis Data Streams	arn:aws:firehose:us-east-1:149515436084:deliverystream/firehose-cp	Not enabled	0 Destination error logs

Test with demo data

Ingest simulated data to test the configuration of your delivery stream. Standard Amazon Kinesis Data Firehose charges apply.



DATABASES

Created database to store table generated by crawler

One database successfully deleted
The following database is now deleted: "hr-db"

AWS Glue > Databases

Databases (3)
A database is a set of associated table definitions, organized into a logical group.
Last updated (UTC) June 9, 2023 at 18:08:24 [Refresh] [Edit] [Delete] [Add database]

Filter databases

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	clickstream_db	-	-	June 9, 2023 at 18:03:53
<input type="checkbox"/>	hr-db	-	-	June 9, 2023 at 18:07:01
<input type="checkbox"/>	socialmedia-db	-	-	June 9, 2023 at 18:08:24

Crawlers

used it to create table

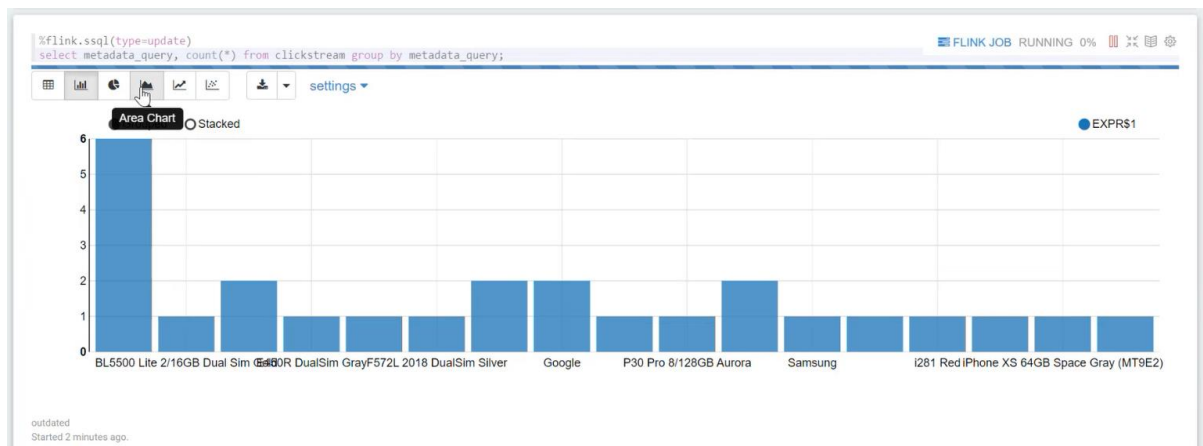
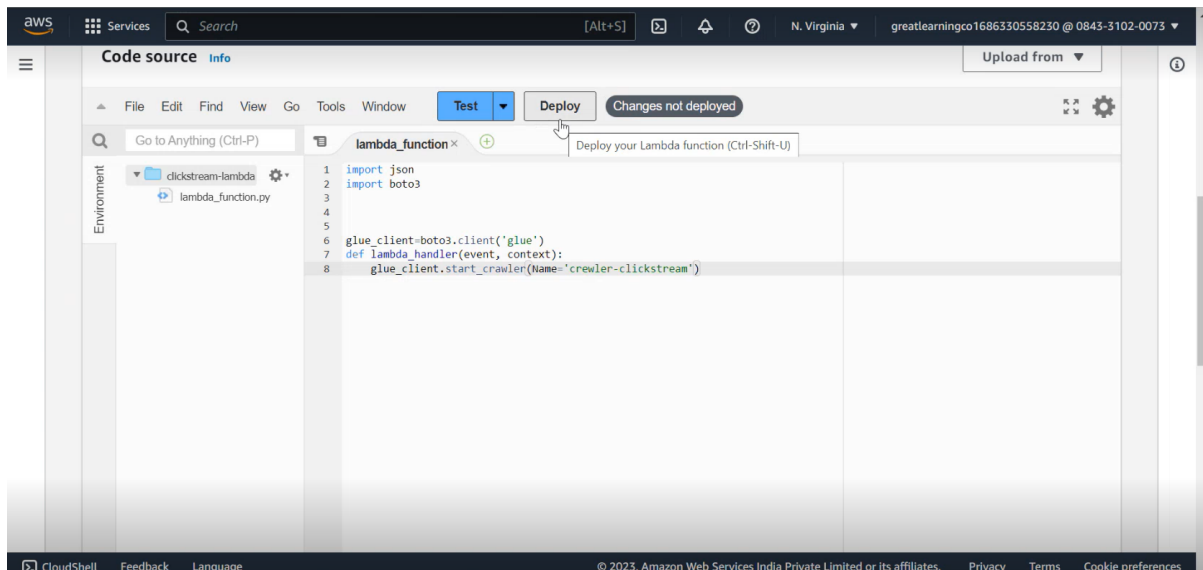
AWS Glue > Crawlers

Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.
Last updated (UTC) June 9, 2023 at 18:34:20 [Refresh] [Action] [Run] [Create crawler]

View and manage all available crawlers.

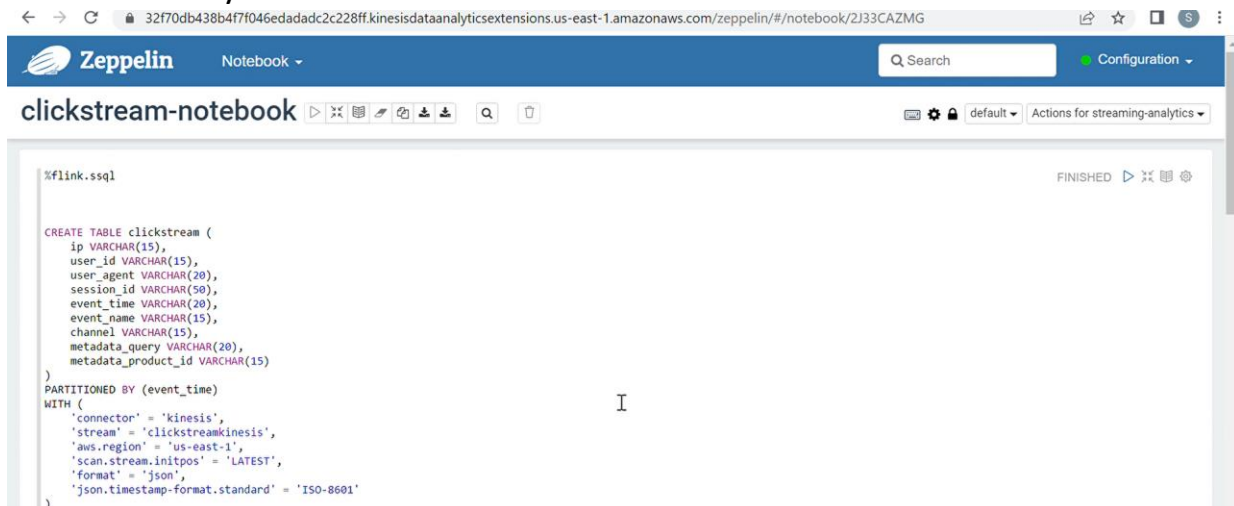
Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run ...	Log
<input type="checkbox"/>	clickstream_crawler	Ready	-	-	-	-
<input type="checkbox"/>	hr-crawler	Ready	-	-	-	-
<input type="checkbox"/>	socialmedia-crawler	Ready	-	-	-	-



Zeppelin

used it to analyse socialmedia data



```
%flink.sql(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 M113/32 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.sql
Drop table clickstream;
```

ERROR

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

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

FINISHED

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

+ Add Paragraph

```
%flink.sql(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 ...

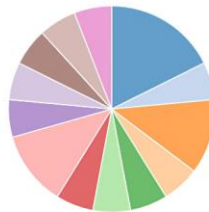
Took 44 sec. Last updated by anonymous at June 10 2023, 11:41:02 AM.

```
%flink.sql(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

The screenshot shows the AWS Athena console interface. At the top, there's a navigation bar with the account ID 'greatlearningco1686582168753' and region 'N. Virginia'. Below this, a tab bar shows 'Query 3', 'Query 4', 'Query 5', 'Query 6', and 'Query 7' (selected). The SQL editor displays a query: `select department,round(avg(empsatisfaction),2) as department_satisfaction,sum(specialprojectscount) as special_projects from "AwsDataCatalog"."hr-database"."hrdata" group by department order by department_satisfaction desc`. Below the editor are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A status bar indicates 'Completed' with 'Time in queue: 184 ms', 'Run time: 590 ms', and 'Data scanned: 70.02 KB'. The 'Query results' tab is active, showing a table with 5 rows and 4 columns: '#', 'department', 'department_satisfaction', and 'special_projects'. The footer contains copyright information for Amazon Web Services India Private Limited.

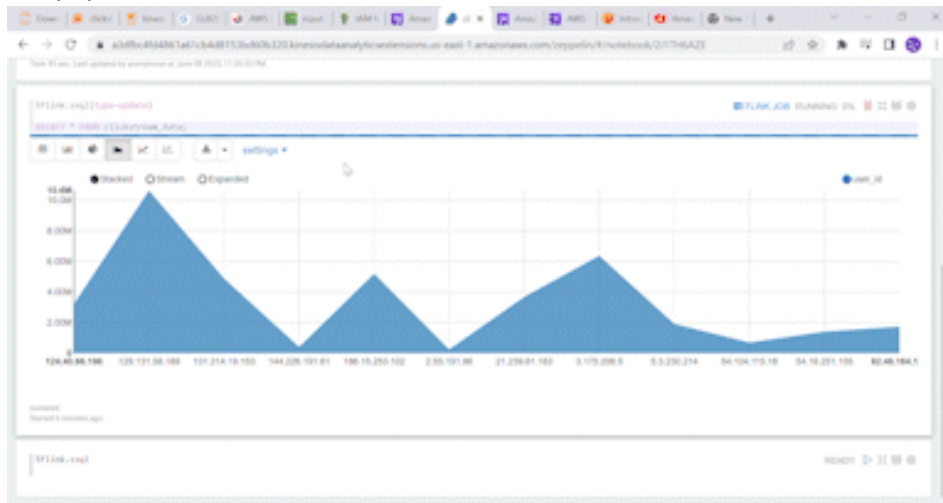
#	department	department_satisfaction	special_projects
1	Software Engineering	4.09	46
2	Sales	4.03	0
3	IT/IS	3.96	296
4	Production	3.86	4
5	Admin Offices	3.56	33

The screenshot shows the AWS Athena console interface for Query 6. The SQL editor displays a query: `SELECT recruitmentsource,count(*) as source_count FROM "AwsDataCatalog"."hr-database"."hrdata" group by recruitmentsource order by source_count desc;`. Below the editor are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A status bar indicates 'Completed' with 'Time in queue: 160 ms', 'Run time: 628 ms', and 'Data scanned: 70.02 KB'. The 'Query results' tab is active, showing a table with 4 rows and 2 columns: '#', 'recruitmentsource', and 'source_count'. The footer contains copyright information for Amazon Web Services India Private Limited.

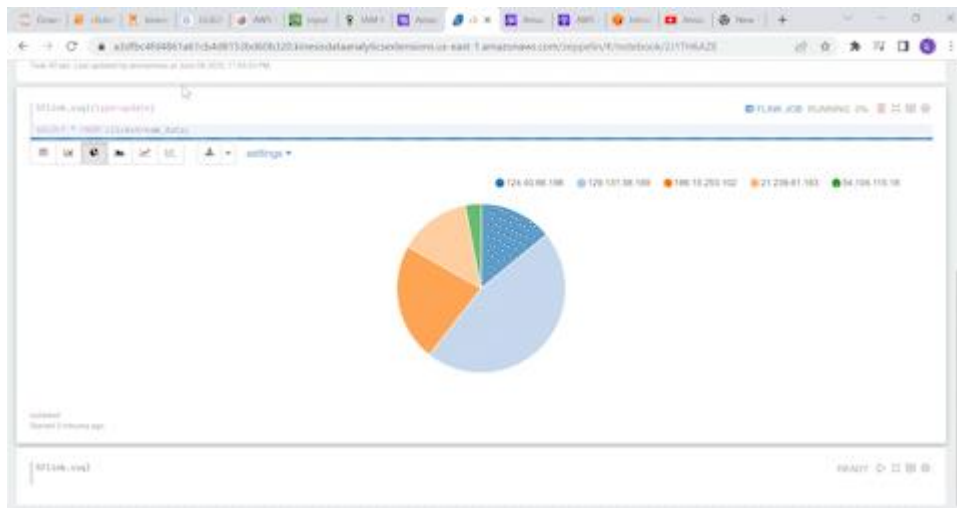
#	recruitmentsource	source_count
1	Indeed	87
2	LinkedIn	76
3	Google Search	49
4	Employee Referral	31

Kinesis Real Time Analysis

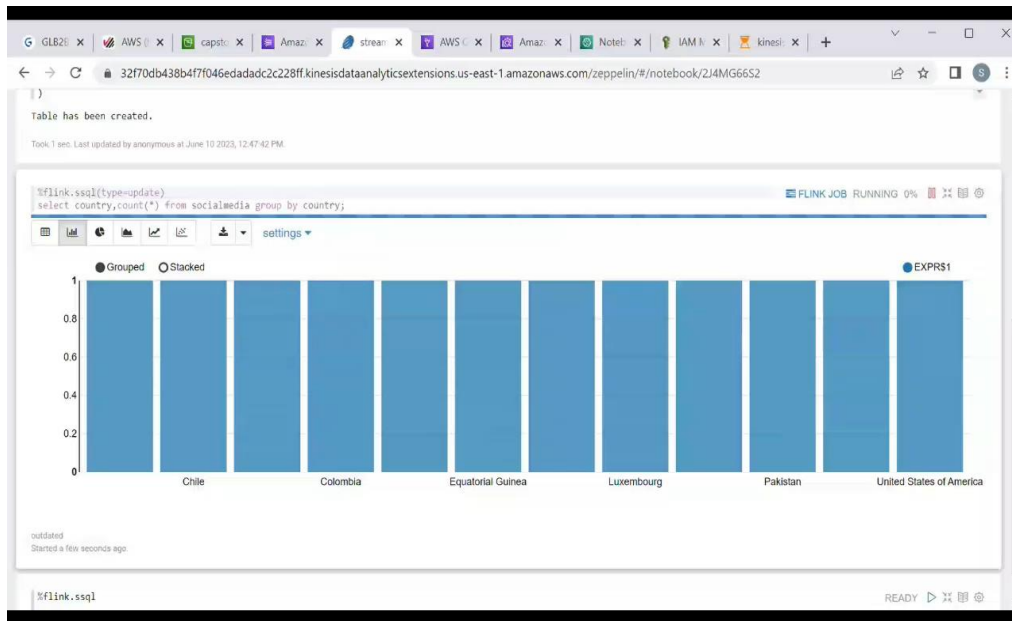
live population of IP address in database



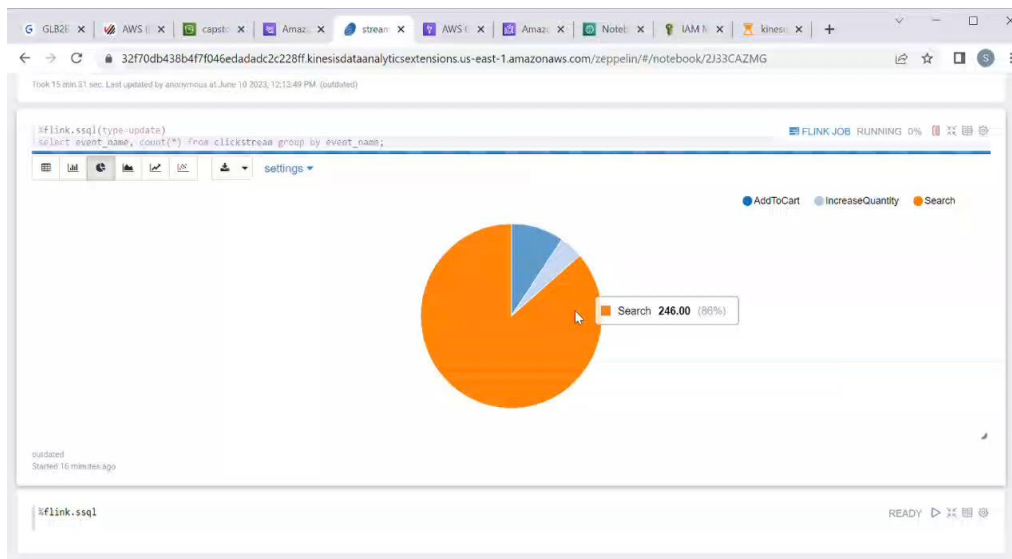
Visualization of IP address populating in Pie Chart



Live traffic on ABC corp website in different countries.

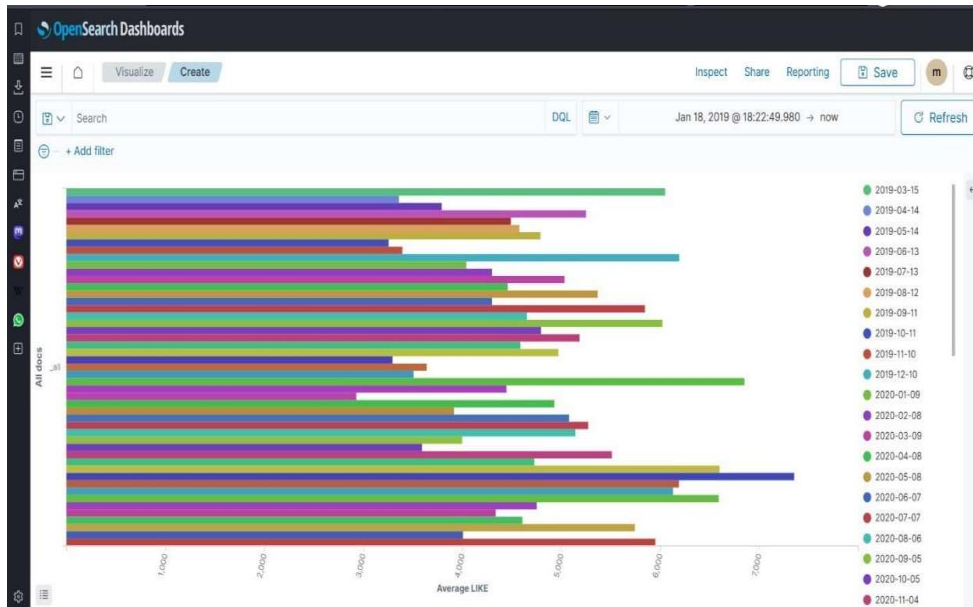


Live User activity on ABC Corporation website

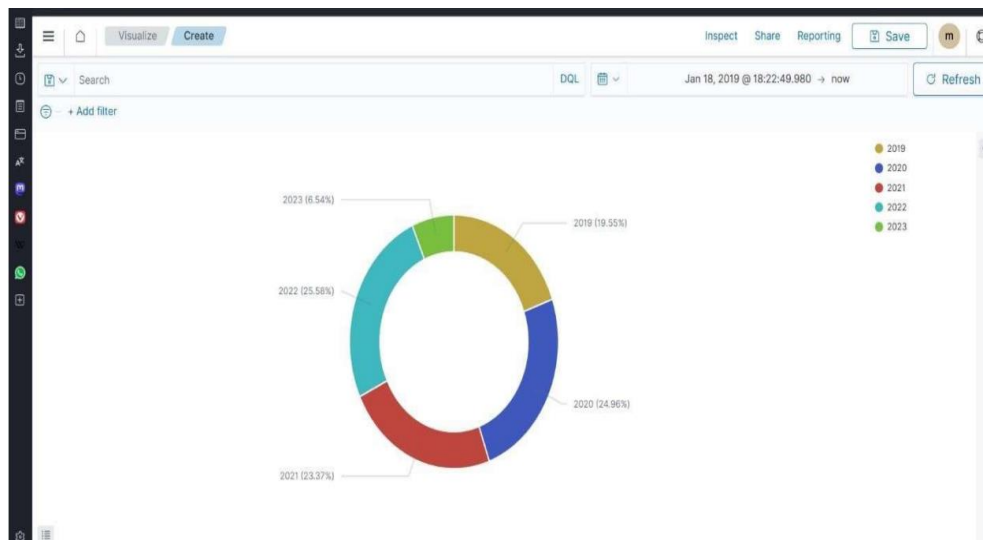


Open Search Dashboard

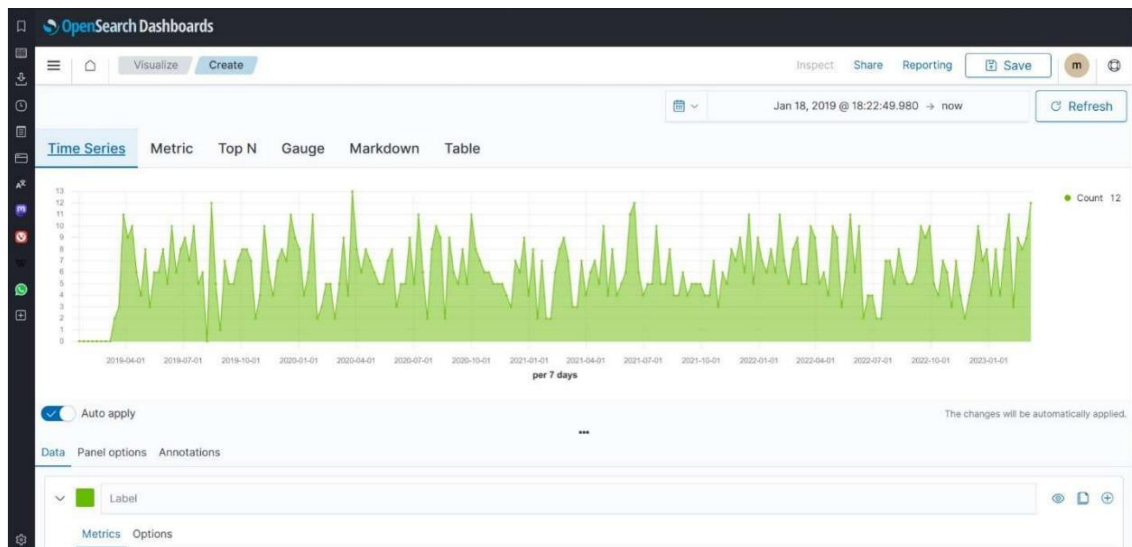
Visualization of Average on monthly basis over the year



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.

The figure is a screenshot of a Zeppelin Notebook interface. It shows a table with two columns: 'post_id' and 'LIKES'. The table contains three rows of data. The first row has a post_id of '59ad36a4-7d6d-4a77-b9cf-2185b1a0f96b' and 401 likes. The second row has a post_id of '89ed58ad-8c75-4470-9952-657ca7248705' and 4853 likes. The third row has a post_id of 'ce4a4cbe-b3f6-445e-95de-db76b87bbe82' and 5393 likes. The notebook interface includes a top navigation bar with 'Zeppelin' and 'Notebook' tabs, and a search bar. Below the table, there are icons for various actions and a 'settings' dropdown menu.

post_id	LIKES
59ad36a4-7d6d-4a77-b9cf-2185b1a0f96b	401
89ed58ad-8c75-4470-9952-657ca7248705	4853
ce4a4cbe-b3f6-445e-95de-db76b87bbe82	5393

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

Thank You 😊