Youtube Video Trending

APAN 5450 Group 7:

Xiaoting Teng
Emily Xiao
Seline Yin
Keqi Yu
Chang Yuan

TABLE OF CONTENTS

01 Business case

Business Problem
Data Source & Description

02 Cloud System

Design
Cloud Architecture

03 Team Responsibilities

Area of Responsibility

04 Implementation & Results

Database
Compute

Background & Business Plan

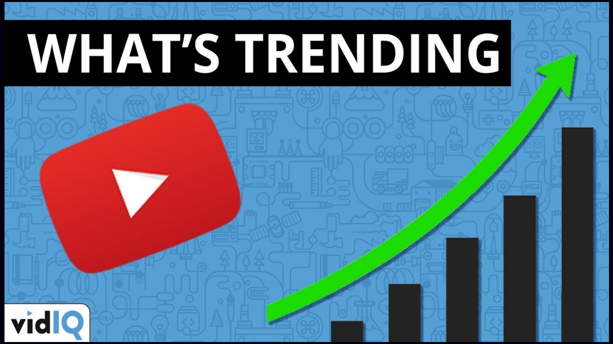
Provide Recommendations for Youtube Channel:

- ❑ Knowing what is the best publishing time for videos could help businesses target advertisements specific to viewer country and taste
- ❑ Explore the key factors that influence viewers' preferences and content based on variables such as
  a. Likes & Dislikes
  b. Publish time
  c. Views count
  d. Comment Counts

Data Source & Description

- The dataset is a version of Youtube Trending Videos Statistics from Kaggle which includings Youtube videos that are most popular on a daily basis. The Size of Data is 2.79 GB in total, Including Json files and csv across 11 regions around the world. And based on the main business plan, our team plans to use 6 regions (1.5 GB) to do our analysis, respectively USA, Canada, France, Russia, South Korea and Japan.

- Data Link: https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset



Team Responsibilities

Xiaoting

VPC Setup and EC2 instance setup with key pairs

Seline

Upload and restore data in S3

Keqi

Clean data in AWS Glue

Chang

Integrate Data from S3 and configuration

Emily

Classification in AWS SageMaker

Could Architecture Description

❑ Amazon EC2 (future use) – The reason why we believe EC2 can be used in future is after we trained our model to solve YouTuber's need and easy to access. Flask of EC2 might be used with EC2.

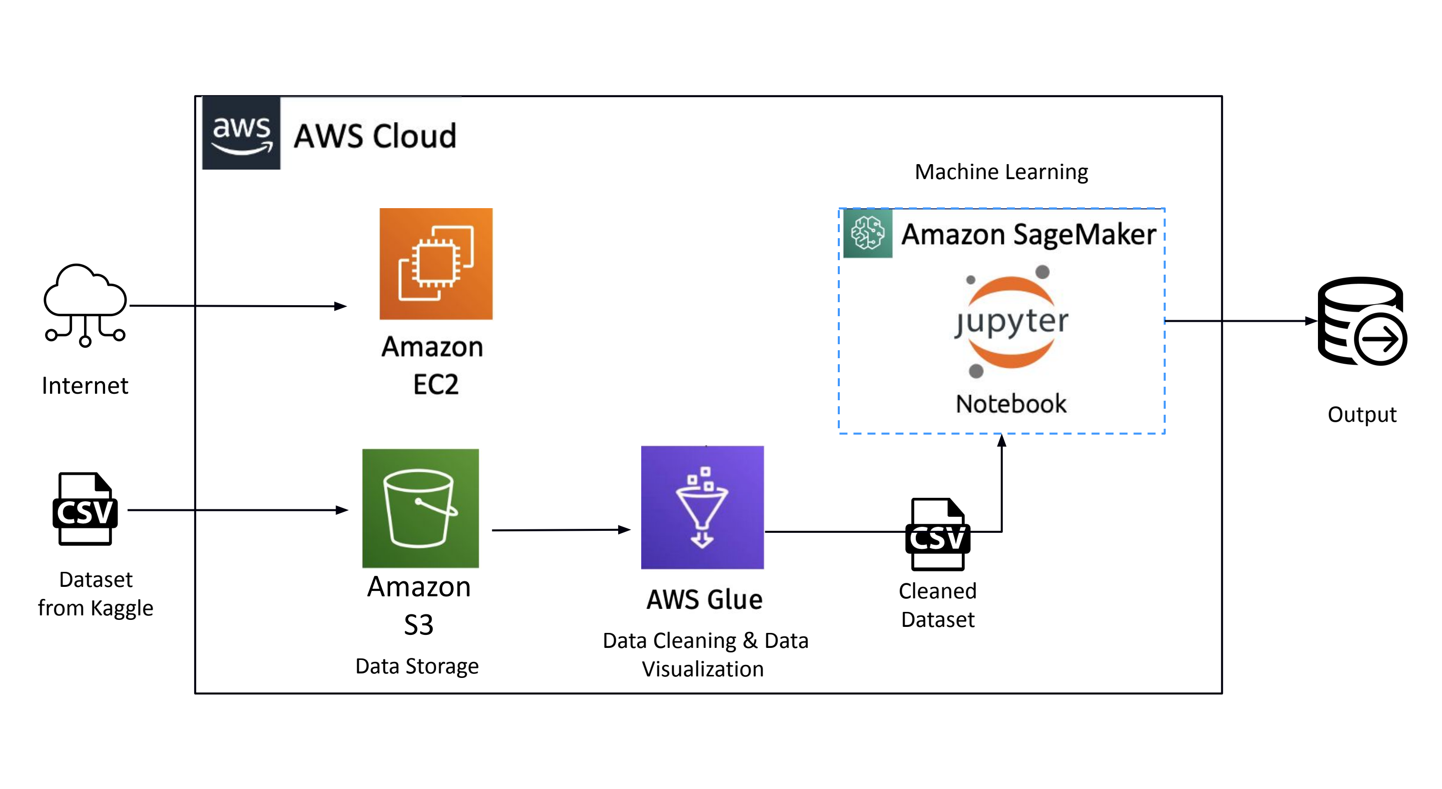
❑ Amazon Virtual Private Cloud (Amazon VPC) – AWS VPC uses security groups as a firewall to control traffic at the instance level, while it also uses network access control lists as a firewall to control traffic at the subnet level. VPC provides much more granular control over security.

❑ Amazon Simple Storage Service (S3) – Our group use S3 because it can assist the team store data at the lowest cost, backup, and restore data, as well as providing great monitoring to ensure data security, S3 is essential in this research. Also sine S3 offers rich security controls, which means it benefits from a data center and network architecture built to meet the requirements of most security-sensitive organizations. There is also need to set up the own securities plans such as take it own control and permission for the system.

❑ Amazon Glue – Since AWS Glue is a serverless data integration service that makes data preparation simpler, faster, and cheaper. Thus, our group use it to help data cleaning part and visualization.

❑ Amazon Sagemaker – Prepare, build, train, and deploy high-quality machine learning models quickly by bringing together a broad set of capabilities purpose-built for machine learning. So our group use Sagemarker for machine learning to predict audience behaviors and count views.

ETL Process


AWS Cloud
Machine Learning
Amazon SageMaker
jupyter
Notebook
Internet
Amazon EC2
Output
CSV
Dataset from Kaggle
Amazon S3
Data Storage
AWS Glue
Data Cleaning & Data Visualization
CSV
Cleaned Dataset

Security Plan

IAM

Identity and Access Management since we cannot create the role right now, we use the default LabRole

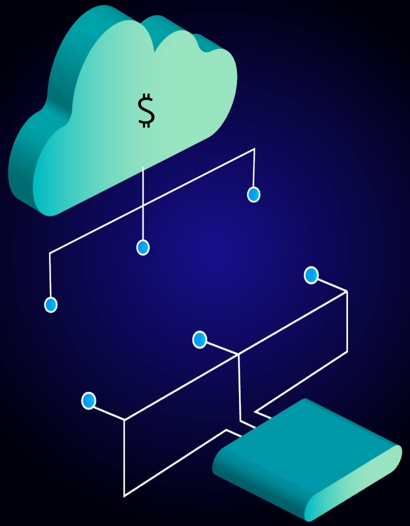
VPC

Make public subnets so the group can use various platforms and devices to access, analyze, and execute analysis on the data.

NAT Gateway

Permit proxy access to application servers for a group of servers in a public subnet using the internet

Cost Analysis



Total Cost:

★ Upfront: 29.78 USD

★ Monthly: 55.10 USD

AWS VPC

★ Total NAT Gateway usage and data processing cost (Monthly): 32.94 USD

★ Total PrivateLink endpoints and data processing cost (Monthly): 7.32 USD

AWS EC2 Instance

★ Amazon EC2 Instance Savings Plans instances (Monthly): 2.48 USD

★ Amazon EC2 instances (Upfront): 29.78 USD

★ Total Monthly cost: 3.28 USD

AWS S3

★ Total Monthly cost: 0.5 USD

★ AWS Glue Monthly 8.86USD

★ Sagemaker Monthly ml.t3.medium 3.00USD

Success Criteria

Quantitative

Cost-effectiveness

The cost of implementing technologies for applicable data and storage is saved for further analysis, which the analysts can directly analyze by using this prepared AWS technologies

Data Size

The data is securely and successfully imported into the Cloud ready to analyze and predict variables of interest

Qualitative

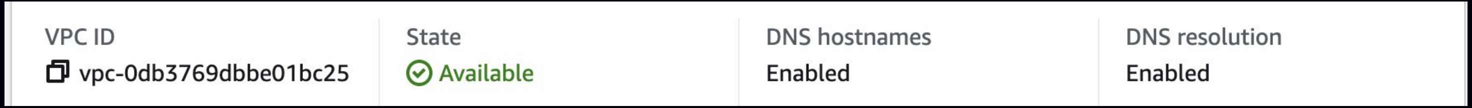
Data Quality

Assume raw data from data resources are up-to-date and well-collected, the processed data should be reliable, and the search result from processed data should be accurate.

Meaningful

The analysis will provide meaning to social reality by understanding how an individual subjectively perceived.

Implementation

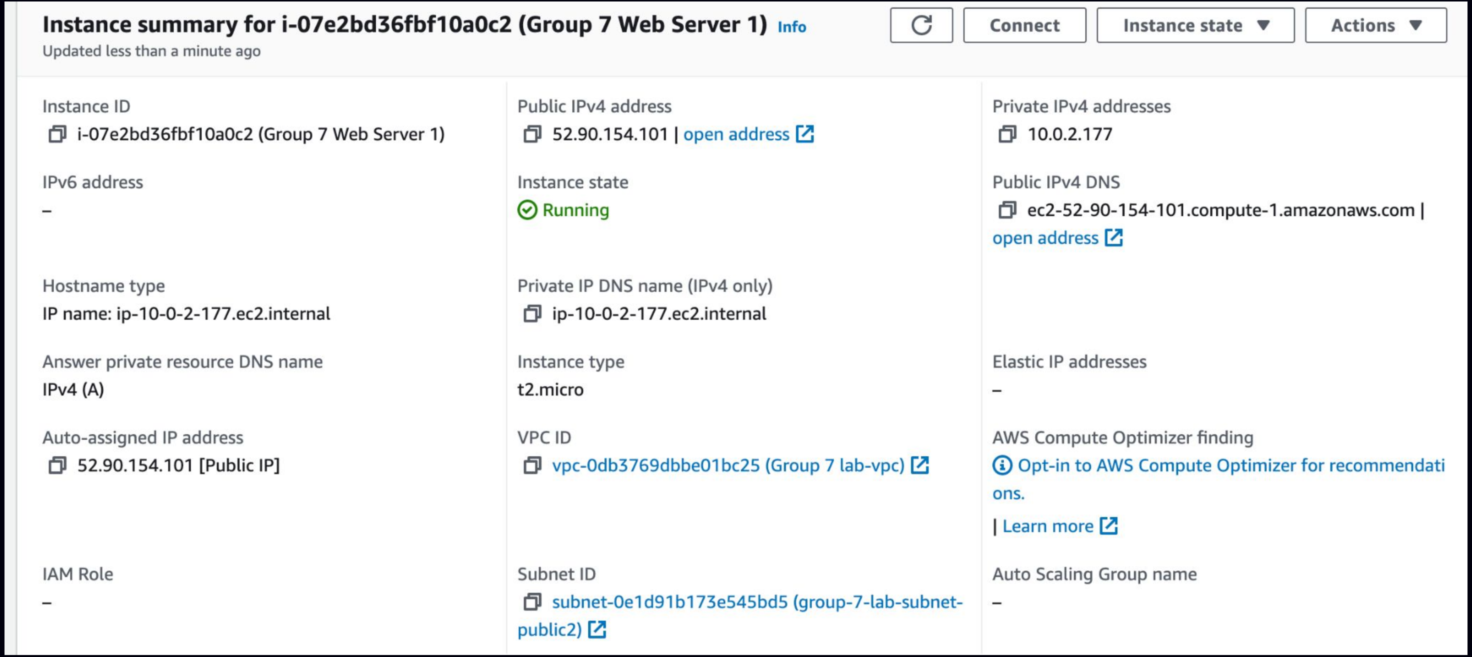


VPC ID
vpc-0db3769dbbe01bc25

State
Available

DNS hostnames
Enabled

DNS resolution
Enabled



Instance summary for i-07e2bd36fbf10a0c2 (Group 7 Web Server 1)

Updated less than a minute ago

Instance ID
i-07e2bd36fbf10a0c2 (Group 7 Web Server 1)

IPv6 address
–

Hostname type
IP name: ip-10-0-2-177.ec2.internal

Answer private resource DNS name
IPv4 (A)

Auto-assigned IP address
52.90.154.101 [Public IP]

IAM Role
–

Public IPv4 address
52.90.154.101 | open address

Instance state
Running

Private IP DNS name (IPv4 only)
ip-10-0-2-177.ec2.internal

Instance type
t2.micro

VPC ID
vpc-0db3769dbbe01bc25 (Group 7 lab-vpc)

Subnet ID
subnet-0e1d91b173e545bd5 (group-7-lab-subnet-public2)

Private IPv4 addresses
10.0.2.177

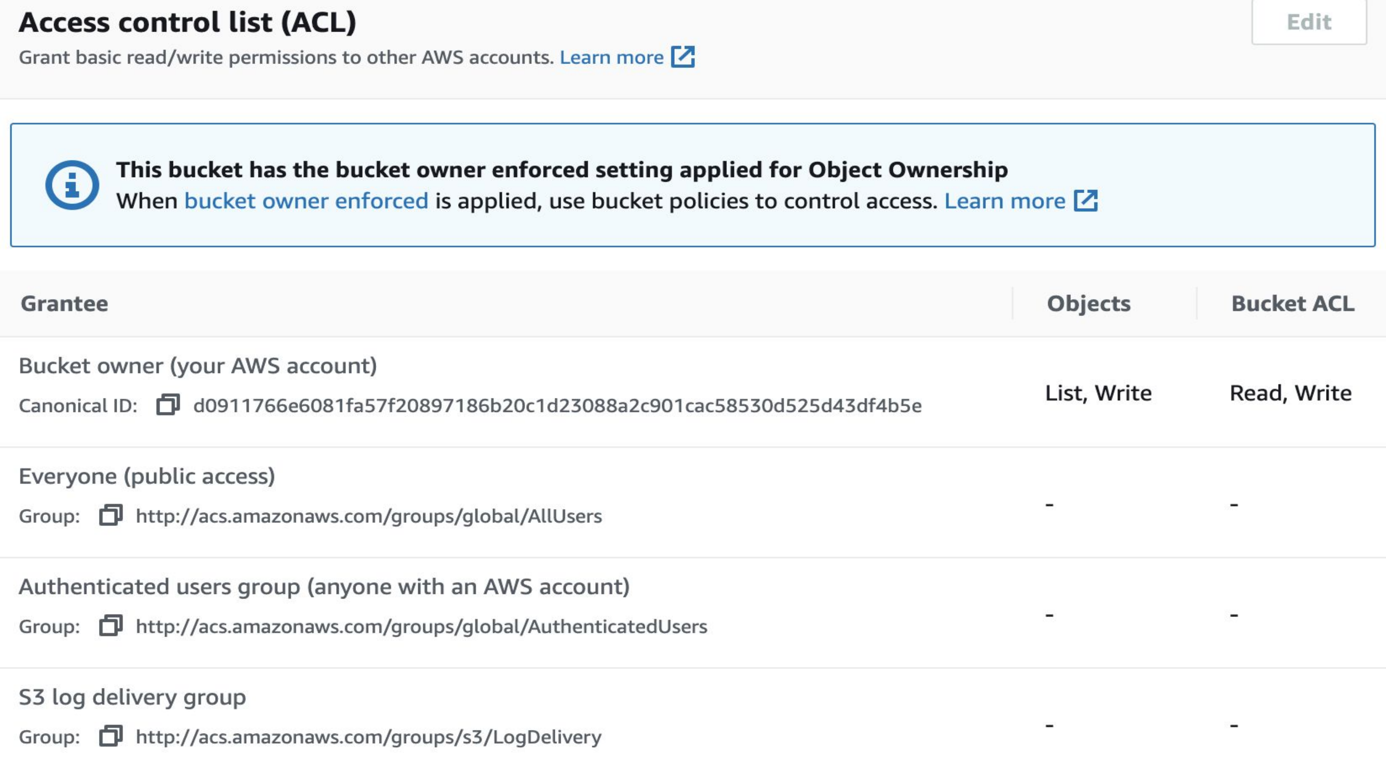
Public IPv4 DNS
ec2-52-90-154-101.compute-1.amazonaws.com | open address

Elastic IP addresses
–

AWS Compute Optimizer finding
Opt-in to AWS Compute Optimizer for recommendations.
| Learn more

Auto Scaling Group name
–

Setting up the VPC
Setting up key paris and Launch EC2 Instance



Access control list (ACL)

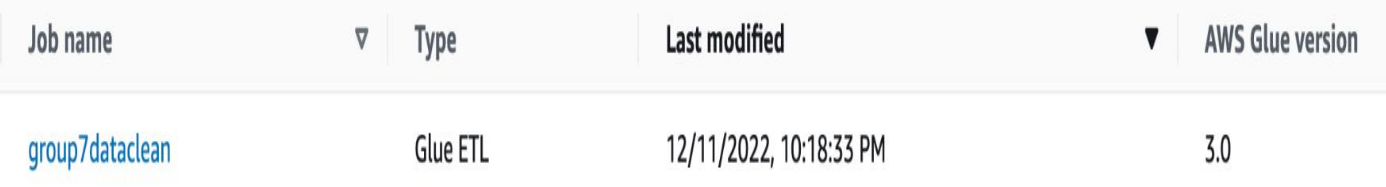
Grant basic read/write permissions to other AWS accounts. Learn more

This bucket has the bucket owner enforced setting applied for Object Ownership
When bucket owner enforced is applied, use bucket policies to control access. Learn more

Grantee	Objects	Bucket ACL
Bucket owner (your AWS account) Canonical ID: d0911766e6081fa57f20897186b20c1d23088a2c901cac58530d525d43df4b5e	List, Write	Read, Write
Everyone (public access) Group: http://acs.amazonaws.com/groups/global/AllUsers	-	-
Authenticated users group (anyone with an AWS account) Group: http://acs.amazonaws.com/groups/global/AuthenticatedUsers	-	-
S3 log delivery group Group: http://acs.amazonaws.com/groups/s3/LogDelivery	-	-

Created bucket and upload dataset in S3 with subsequent access.

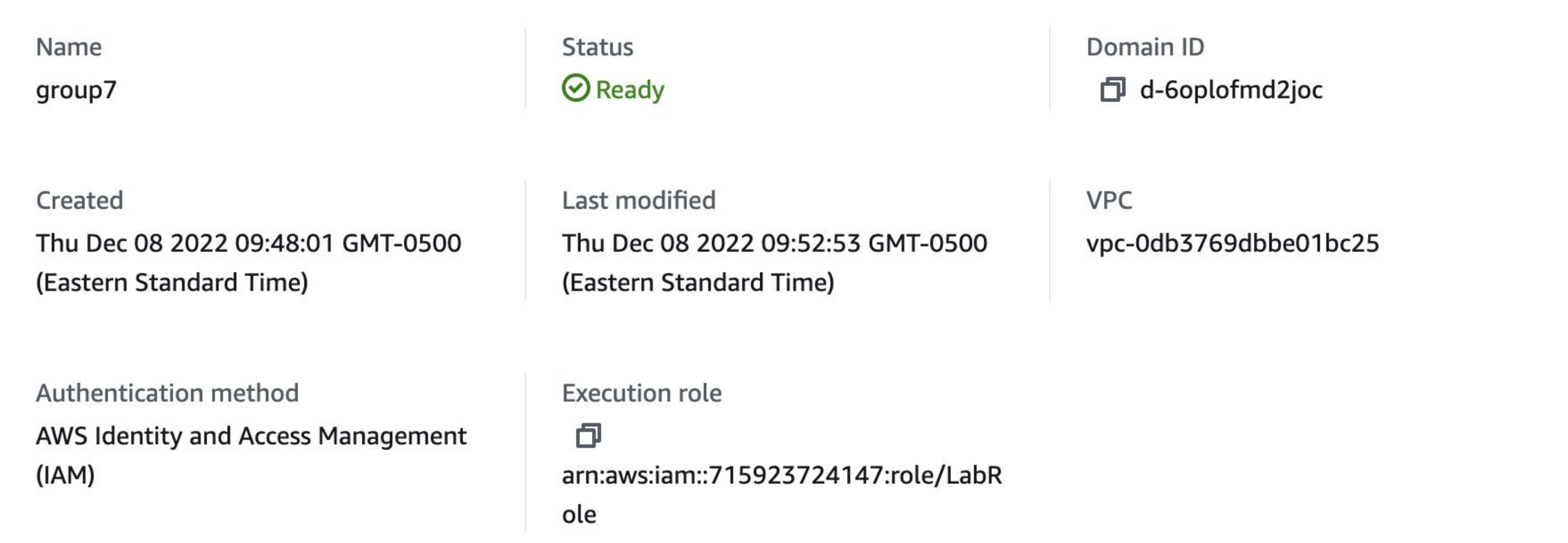
Also for AWS Glue and AWS Sagemaker



Job name	Type	Last modified	AWS Glue version
group7dataclean	Glue ETL	12/11/2022, 10:18:33 PM	3.0

Created a job for cleaning data in AWS Glue.

Created domain of AWS Sagemaker with the VPC setted up for connecting notebook and using Machine Learning in Jupyter Notebook.



Name
group7

Status
✓ Ready

Domain ID
d-6oplofmd2joc

Created
Thu Dec 08 2022 09:48:01 GMT-0500 (Eastern Standard Time)

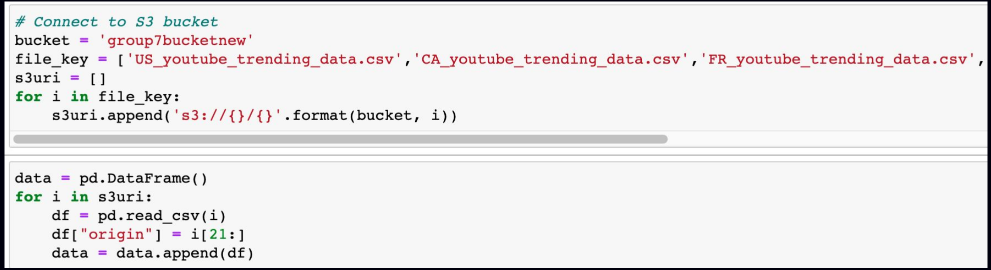
Last modified
Thu Dec 08 2022 09:52:53 GMT-0500 (Eastern Standard Time)

VPC
vpc-0db3769dbbe01bc25

Authentication method
AWS Identity and Access Management (IAM)

Execution role
arn:aws:iam::715923724147:role/LabR
ole

Data Cleaning in AWS Glue



```
# Connect to S3 bucket
bucket = 'group7bucketnew'
file_key = ['US_youtube_trending_data.csv','CA_youtube_trending_data.csv','FR_youtube_trending_data.csv',
s3uri = []
for i in file_key:
    s3uri.append('s3://{}/{}'.format(bucket, i))
```

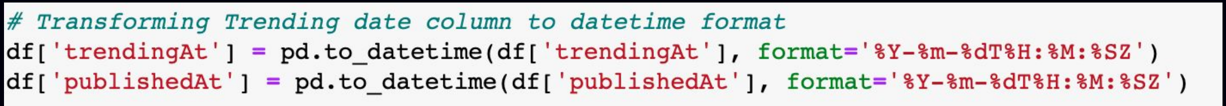
```
data = pd.DataFrame()
for i in s3uri:
    df = pd.read_csv(i)
    df["origin"] = i[21:]
    data = data.append(df)
```

1. **Connect to S3 bucket from AWS Glue and merge all the data files together**

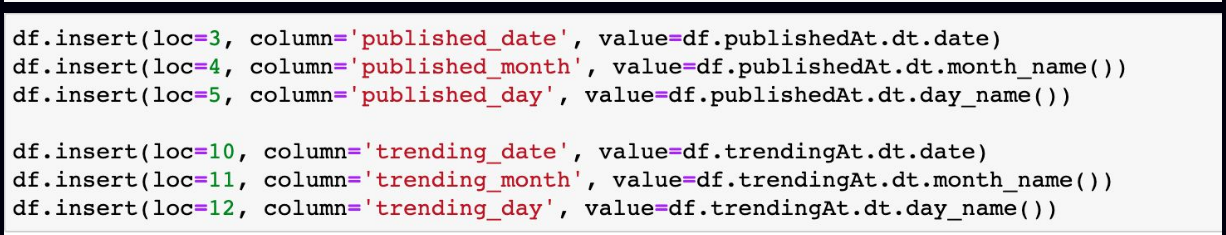
The original dataset has

1012039 rows x 17 columns

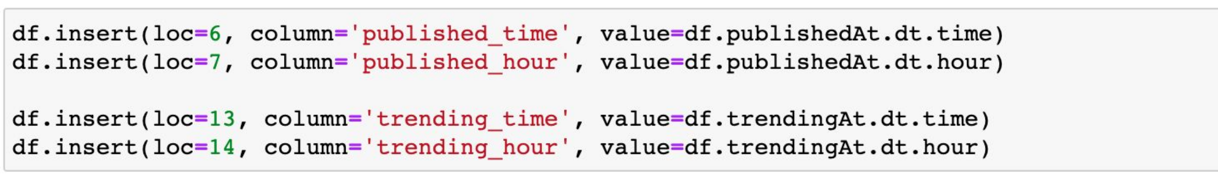
2. Convert Data Type for Date Column to datetime format and convert datetime to date, month, year, time, hour



```
# Transforming Trending date column to datetime format
df['trendingAt'] = pd.to_datetime(df['trendingAt'], format='%Y-%m-%dT%H:%M:%SZ')
df['publishedAt'] = pd.to_datetime(df['publishedAt'], format='%Y-%m-%dT%H:%M:%SZ')
```

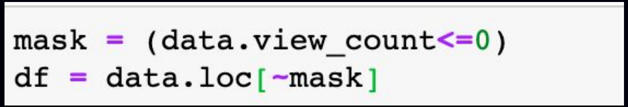


```
df.insert(loc=3, column='published_date', value=df.publishedAt.dt.date)
df.insert(loc=4, column='published_month', value=df.publishedAt.dt.month_name())
df.insert(loc=5, column='published_day', value=df.publishedAt.dt.day_name())

df.insert(loc=10, column='trending_date', value=df.trendingAt.dt.date)
df.insert(loc=11, column='trending_month', value=df.trendingAt.dt.month_name())
df.insert(loc=12, column='trending_day', value=df.trendingAt.dt.day_name())
```



```
df.insert(loc=6, column='published_time', value=df.publishedAt.dt.time)
df.insert(loc=7, column='published_hour', value=df.publishedAt.dt.hour)

df.insert(loc=13, column='trending_time', value=df.trendingAt.dt.time)
df.insert(loc=14, column='trending_hour', value=df.trendingAt.dt.hour)
```

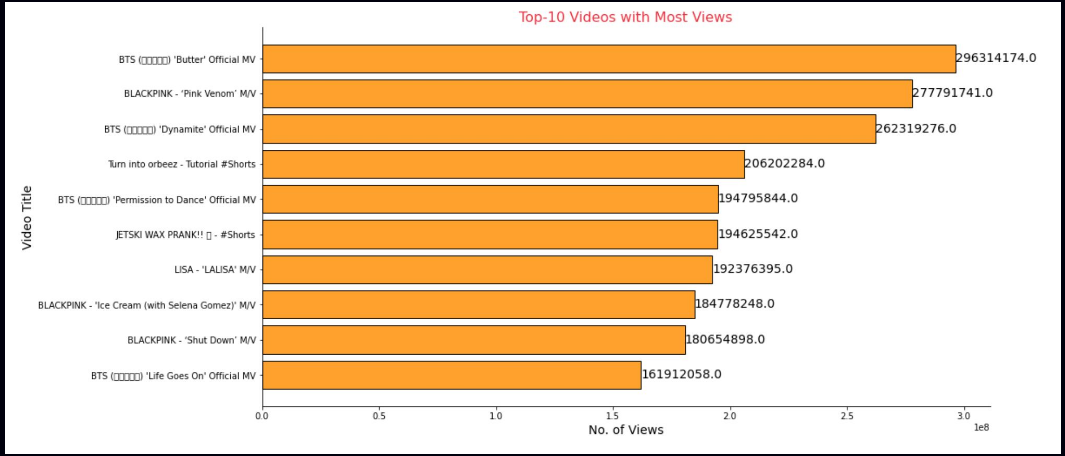
3. Delete abnormal data



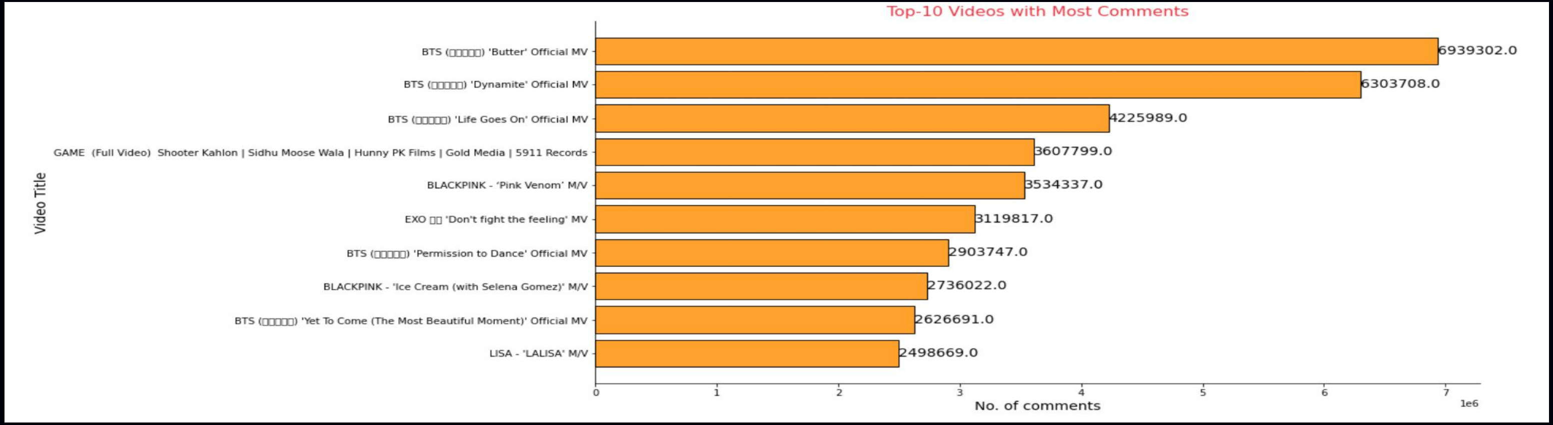
```
mask = (data.view_count<=0)
df = data.loc[~mask]
```

Finally, after the data cleaning, the cleaned dataset has 1011848 rows x 27 columns

Implement Results

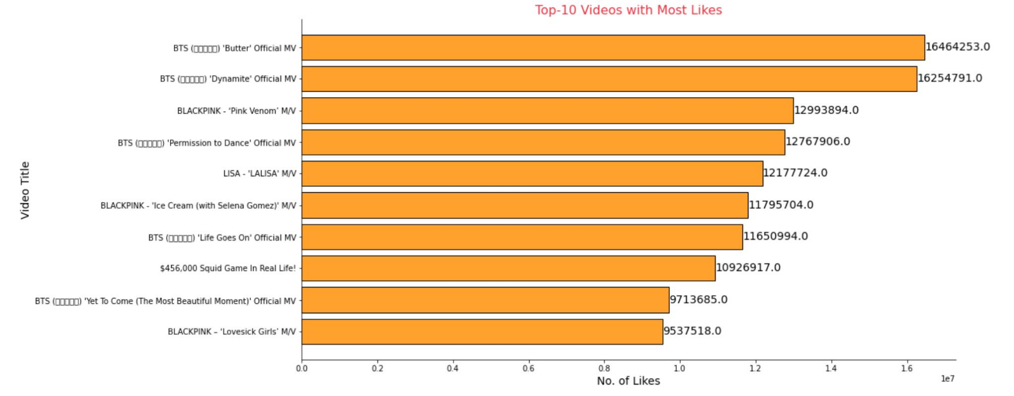


BTS and BlackPink videos make the most of the top 10 videos with most views and most comments



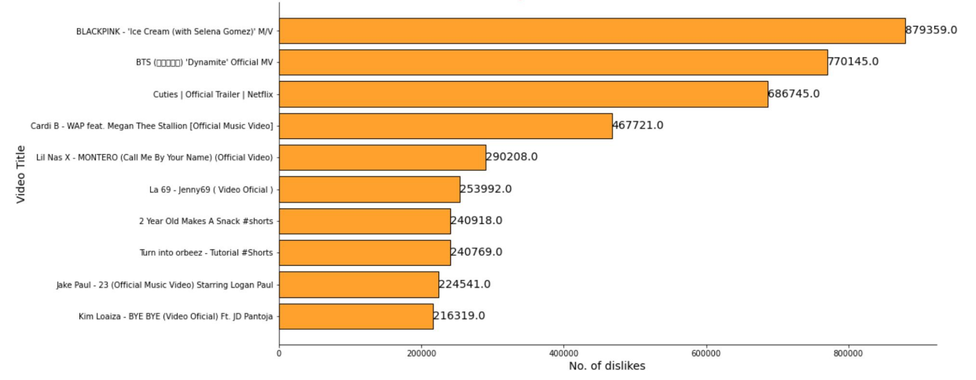
Implement Results

Top-10 Videos with Most Likes



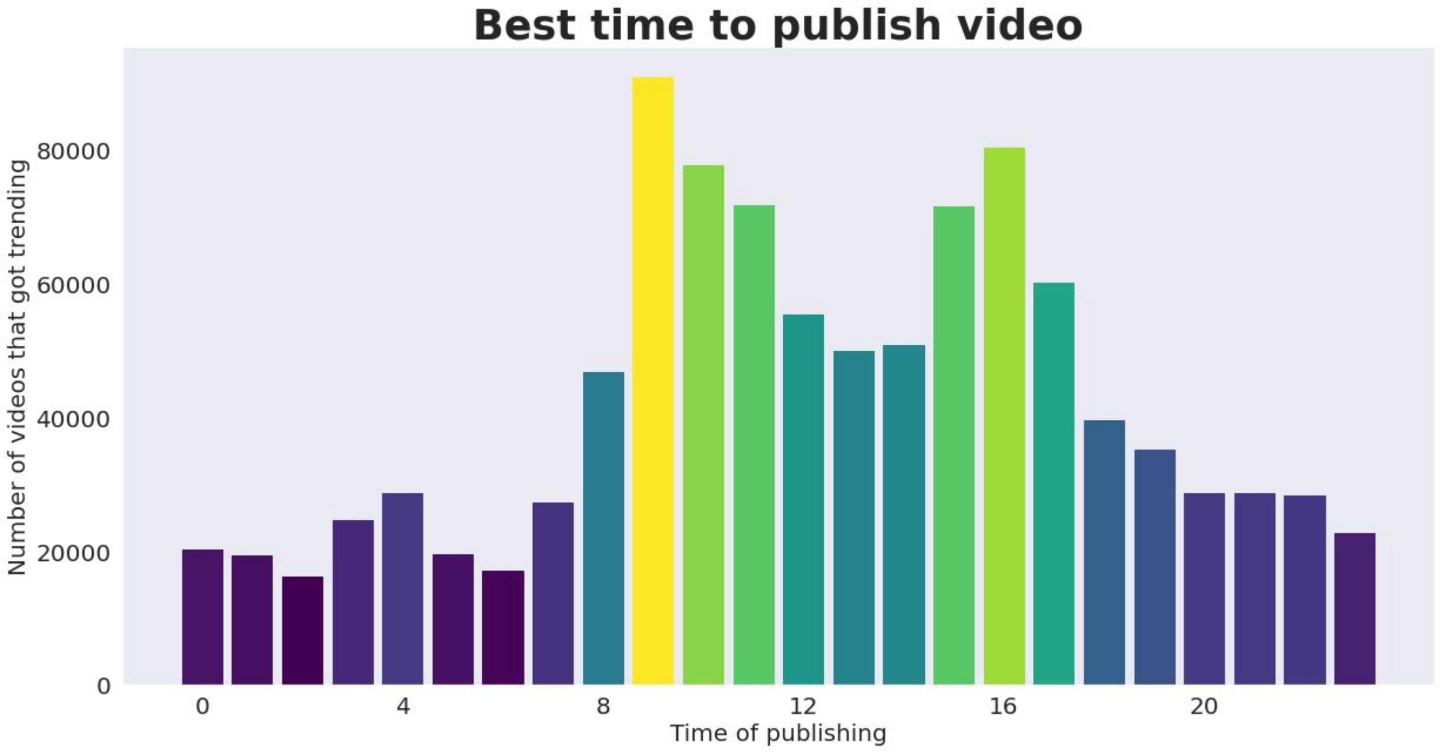
BTS and BlackPink videos make the most of the top 10 videos with most likes, which coincides with the previous slide

The video with most dislikes were also from the music video.



Top-10 Videos with Most Dislikes

Results



This graph can help YouTubers to pick the best time to publish their videos.

Classification Results

Data Split	Training Set: 0.8	Testing Set: 0.2
Model Used	Support Vector Machine Classification	Logistic Regression Classification
RMSE Value	0.1906	0.2527
Usage	Using our model, YouTubers can predict their **View Counts** for their videos and know how many audience would view their videos	

Future Plan

1. Add AWS Key Management Service(KMS) encryption key for additional security
2. Deploy Flask on EC2 for our product development
3. Create IAM roles to achieve specific permissions for future development and maintenance
4. Use AWS RDS database

AWS Key Management Service

Flask

Amazon RDS

Identity and Access Management

THANKS!