

The background features a complex graphic with teal and orange wavy lines, network diagrams with nodes and edges, and several computer monitors displaying video player interfaces. A prominent bar chart with an upward-pointing arrow is on the right side.

# YouTube Trending Pipeline

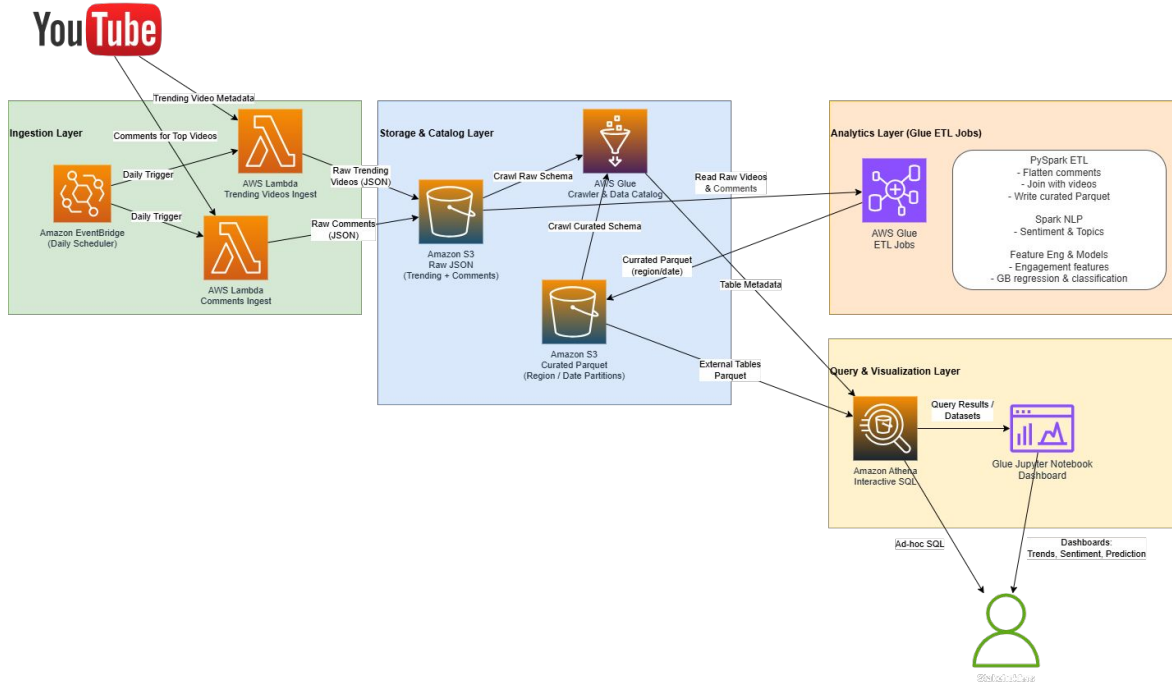
CS 6705 Applied Cloud Computing - Kevin Bell, Jacob Child



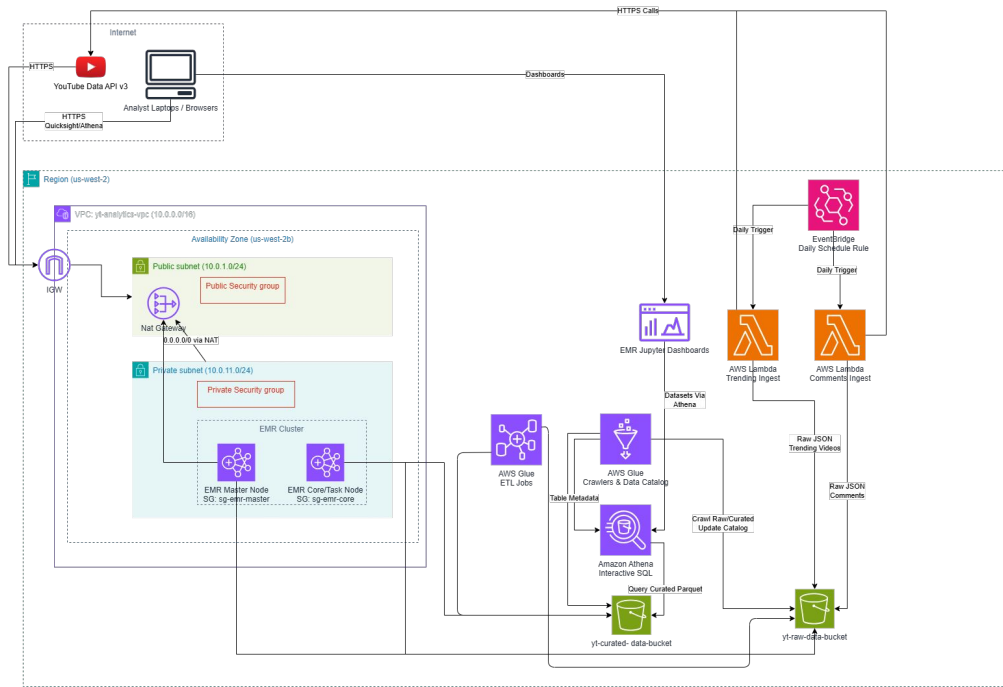
## Project Goal

Our project will build a cloud-based YouTube analytics pipeline that retrieves trending-video metadata and comment text through the YouTube Data API v3, processes it on AWS, and performs descriptive and predictive analytics at scale using Glue ETL jobs.

# Project Flow Diagram



# Project Network Infrastructure





# AWS Services Used

- AWS VPC
- AWS EventBridge
- AWS Secrets
- AWS Lambda
- Amazon S3
- AWS Glue Workflow
- AWS Glue Crawlers
- AWS Glue ETJ Jobs
- AWS Glue ETL Notebook
- Amazon Athena



# AWS Virtual Private Cloud (VPC)

- Set up a VPC to have a place to run this pipeline and manage access to the data
- Created a public and private subnet as well as different security groups to facilitate access
- Created an Internet Gateway (IGW) to allow parts of our cloud to access the internet
- Created a Network Attached Translation (NAT) gateway to connect private components securely to the internet

# Amazon EventBridge

yt-trending-daily-schedule

[Edit](#) [Disable](#) [Delete](#) [CloudFormation Template](#) ▼

## Rule details [info](#)

**Rule name**  
yt-trending-daily-schedule

**Status**  
Enabled

**Event bus name**  
[default](#)

**Type**  
Scheduled Standard

**Description**  
Run yt-trending-harvest Lambda once per day

**Rule ARN**  
[arn:aws:events:us-west-2:069233348392:rule/yt-trending-daily-schedule](#)

**Event bus ARN**  
[arn:aws:events:us-west-2:069233348392:event-bus/default](#)

[Event schedule](#) [Targets](#) [Monitoring](#) [Tags](#)

## Targets [Edit](#)

Details	Target Name	Type	ARN	Input	Role
▼	<a href="#">yt-trending-harvest</a> <a href="#">🔗</a>	Lambda function	<a href="#">arn:aws:lambda:us-west-2:069233348392:function:yt-trending-harvest</a>	Matched event	<a href="#">Amazon_EventBridge_Invoke_Lambda_2095584042</a> <a href="#">🔗</a>
Input to target: Matched event					
Additional parameters: --					
Dead-letter queue (DLQ): -					

- Used to schedule and launch AWS Lambda jobs everyday at 6 am.

# AWS Lambda

The screenshot displays the AWS Lambda console for a function named 'yt-trending-harvest'. The top section, 'Function overview', includes buttons for 'Throttle', 'Copy ARN', and 'Actions'. It also features links to 'Export to Infrastructure Composer' and 'Download'. A 'Diagram' tab is selected, showing a visual representation of the function and its layers. A 'Description' panel on the right provides details: 'Last modified 18 hours ago', 'Function ARN arn:aws:lambda:us-west-2:069233348392:function:yt-trending-harvest', and 'Function URL'. Below the overview, tabs for 'Code', 'Test', 'Monitor', 'Configuration', 'Aliases', and 'Versions' are visible. The 'Code source' tab is active, showing a code editor with a Python script. The script defines a lambda function that fetches comments for a video, processes them, and returns the results. The code is as follows:

```
100 def fetch_comments_for_video(video_id, max_comments=MAX_RESULTS):
101     except msvcrt.getch() as e:
102         pass
103
104     for item in data.get("items", []):
105         top = item.get("snippet", {}).get("topLevelComment", {}).get("snippet", {})
106         comment = {
107             "videoId": video_id,
108             "authorDisplayName": top.get("authorDisplayName"),
109             "textDisplay": top.get("textDisplay"),
110             "publishedAt": top.get("publishedAt"),
111             "likeCount": top.get("likeCount"),
112         }
113         comments.append(comment)
114         if len(comments) >= max_comments:
115             break
116     return comments
```

- Wrote a Python script that was run as a Lambda job
- The Python script utilized the YouTube API key stored in AWS Secrets to pull trending video data and the comments associated with those trending videos.
- This Lambda job also launches the AWS Glue Workflow



# Amazon S3

yt-analytics-cs6705-data [info](#)

[Objects](#) | [Metadata](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

**Objects (7)**

[Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified
<input type="checkbox"/>	<a href="#">athena-results/</a>	Folder	-
<input type="checkbox"/>	<a href="#">curated_\$folder\$</a>	-	November 20, 2025, 12:02:34 (UTC-07:00)
<input type="checkbox"/>	<a href="#">curated/</a>	Folder	-
<input type="checkbox"/>	<a href="#">emr-logs/</a>	Folder	-
<input type="checkbox"/>	<a href="#">jobs/</a>	Folder	-
<input type="checkbox"/>	<a href="#">raw/</a>	Folder	-
<input type="checkbox"/>	<a href="#">scripts/</a>	Folder	-

- Created a storage location in S3 to store
  - Raw data pulled from YouTube
  - Curated data flattened by ETL Jobs
  - Predicted data created by ETL jobs
  - Scripts used by Glue

# AWS Glue Workflow

**TrendingWorkflow**

Last updated (UTC)  
December 2, 2025 at 17:22:05

Run workflowEditDelete

Workflow details

Advanced properties

**Name**  
TrendingWorkflow

**Description**  
-

**Max concurrency**  
-

**Last run status**  
Completed

**Last run**  
December 2, 2025 at 17:19:57

**Last modified**  
November 21, 2025 at 18:16:45

**Blueprint name**  
-

**Blueprint run id**  
-

Graph

History

Tags

**Workflow runs (34)**

The list of workflow runs for this workflow.

Filter data

< 1 2 > ⚙

	Workflo...	Previous...	Status	Start time (UTC)	End time (UTC)	Current
<input type="radio"/>	wr_cf53a783c5b	-	Completed	December 2, 2025 at 16:56:5	December 2, 2025 at 17:19:5	
<input type="radio"/>	wr_0738f44682e	-	Completed	December 2, 2025 at 06:00:5	December 2, 2025 at 06:24:4	
<input type="radio"/>	wr_a8919b6dc2i	-	Completed	December 1, 2025 at 23:06:2	December 1, 2025 at 23:24:0	
<input type="radio"/>	wr_01cdfbcbadd	-	Completed	December 1, 2025 at 06:00:5	December 1, 2025 at 06:22:0	
<input type="radio"/>	wr_b59323863ei	-	Completed	November 30, 2025 at 06:00	November 30, 2025 at 06:21	
<input type="radio"/>	wr_f4a4ad9661s	-	Completed	November 29, 2025 at 06:00	November 29, 2025 at 06:21	
<input type="radio"/>	wr_1feb9d98f0c	-	Completed	November 28, 2025 at 06:00	November 28, 2025 at 06:20	

- Created a workflow that connects all of our ETL scripts.
  - Trending ETL Job
  - Comments ETL Job
  - Comments Sentiment ETL Job
  - Feature Label ETL Job
  - Tran Model ETL Job
  - Predictions ETL Job

# AWS Glue ETL Jobs

**AWS Glue Studio** [Info](#)

**Create job** [Info](#)

Author in a visual interface focused on data flow. [Visual ETL](#)

Author using an interactive code notebook. [Notebook](#)

Author code with a script editor. [Script editor](#)

► **Example jobs** [Info](#) [Create example job](#)

**Your jobs (8)** [Info](#)

🔍 Filter jobs by property 8 matches

Created by = SCRIPT X [Clear filters](#)

<input type="checkbox"/>	Job name	Type	Created by	Last modified	AWS Glue version	Action
<input type="checkbox"/>	<a href="#">yt-feature_labels_job</a>	Glue ETL	Script	12/2/2025, 10:11:25 AM	5.0	-
<input type="checkbox"/>	<a href="#">yt-predictions_job</a>	Glue ETL	Script	12/1/2025, 10:06:07 PM	5.0	-
<input type="checkbox"/>	<a href="#">yt-train_models_job</a>	Glue ETL	Script	12/1/2025, 9:42:28 PM	5.0	-
<input type="checkbox"/>	<a href="#">yt-onetime-batch</a>	Glue ETL	Script	12/1/2025, 5:31:25 PM	5.0	-
<input type="checkbox"/>	<a href="#">yt_trending_etl_job</a>	Glue ETL	Script	12/1/2025, 5:15:11 PM	5.0	-
<input type="checkbox"/>	<a href="#">OneTimeDebugging</a>	Glue ETL	Script	12/1/2025, 4:28:21 PM	5.0	-
<input type="checkbox"/>	<a href="#">yt_comments_etl_job</a>	Glue ETL	Script	11/25/2025, 2:32:35 PM	5.0	-
<input type="checkbox"/>	<a href="#">yt-comments_sentiment_job.py</a>	Glue ETL	Script	11/25/2025, 1:13:08 PM	5.0	-

- Trending ETL Job
  - Flattens the raw trending video data
- Comments ETL Job
  - Flattens the raw comments
- Comments Sentiment ETL Job
  - Pulls out the sentiment data from the comments raw data
- Feature Label ETL Job
  - Creates labels for training the model
- Tran Model ETL Job
  - Updates the model with data from the feature label job
- Predictions ETL Job
  - Generates prediction data

# 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.

Crawlers (11) [Info](#)

Last updated (UTC)  
December 2, 2025 at 17:26:14

Action ▼ Run Create crawler

View and manage all available crawlers.

< 1 > ⚙

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run...	Log	Table d
<input type="checkbox"/>	<a href="#">CommentsDataCralwer</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	1 updat
<input type="checkbox"/>	<a href="#">LabelFeaturesDataCralwer</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	1 updat
<input type="checkbox"/>	<a href="#">PredictionsDataCralwer</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	1 updat
<input type="checkbox"/>	<a href="#">RawTrendingCrawler</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	1 updat
<input type="checkbox"/>	<a href="#">SentimentDataCralwer</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	1 updat
<input type="checkbox"/>	<a href="#">TrendingDataCralwer</a>	✔ Ready	At 07:00 AM	✔ Succeeded	December ...	<a href="#">View log</a>	-
<input type="checkbox"/>	<a href="#">crawler_curated_comments</a>	✔ Ready		✔ Succeeded	November ...	<a href="#">View log</a>	-
<input type="checkbox"/>	<a href="#">crawler_curated_trending</a>	✔ Ready		✔ Succeeded	November ...	<a href="#">View log</a>	-
<input type="checkbox"/>	<a href="#">raw_trending</a>	✔ Ready		✔ Succeeded	December ...	<a href="#">View log</a>	1 create
<input type="checkbox"/>	<a href="#">yt-analytics-crawler</a>	✔ Ready		✔ Succeeded	November ...	<a href="#">View log</a>	-
<input type="checkbox"/>	<a href="#">yt-comments-crawler</a>	✔ Ready		✔ Succeeded	November ...	<a href="#">View log</a>	-

- Utilized crawlers to look over the parquet files in the S3 curated area
- Generated tables to store the schema structure to use Athena to query the data

# AWS Glue ETL Notebook

AWS Glue Studio [Info](#)

Create job [Info](#)



Author in a visual interface focused on data flow.

Visual ETL



Author using an interactive code notebook.

Notebook



Author code with a script editor.

Script editor

► Example jobs [Info](#)

Create example job

Your jobs (3) [Info](#)



Actions ▼

Run job

Q Filter jobs by property

3 matches

Created by = NOTEBOOK X

Clear filters

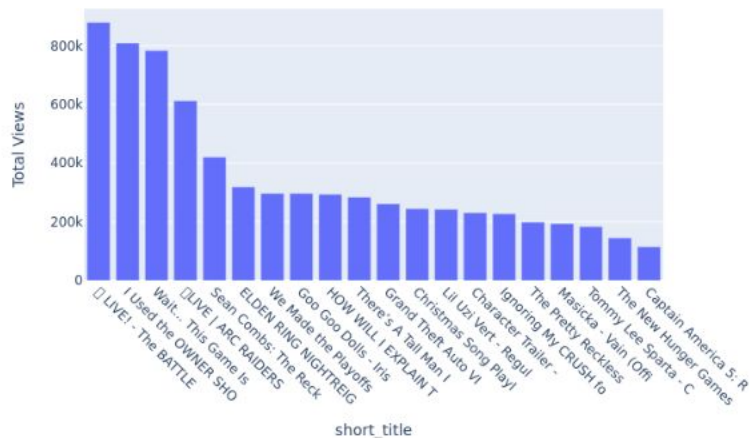
< 1 > ⚙

<input type="checkbox"/>	Job name ▼	Type	Created by	Last modified ▼	AWS Glue version ▼	Action
<input type="checkbox"/>	<a href="#">CommentsSentiment</a>	Glue ETL	Notebook	12/2/2025, 10:06:32 AM	5.0	-
<input type="checkbox"/>	<a href="#">Trending</a>	Glue ETL	Notebook	12/2/2025, 9:54:21 AM	5.0	-
<input type="checkbox"/>	<a href="#">YouTube Analytics Pipeline ETL</a>	Glue ETL	Notebook	11/20/2025, 11:49:01 AM	5.0	-

- Created Glue ETL notebooks to read the data and process it using python and panda
- Created plotly graphs to act as a dashboard since Quicksight wasn't available

# Trending Notebook

Top 20 Most Viewed Trending Videos in US

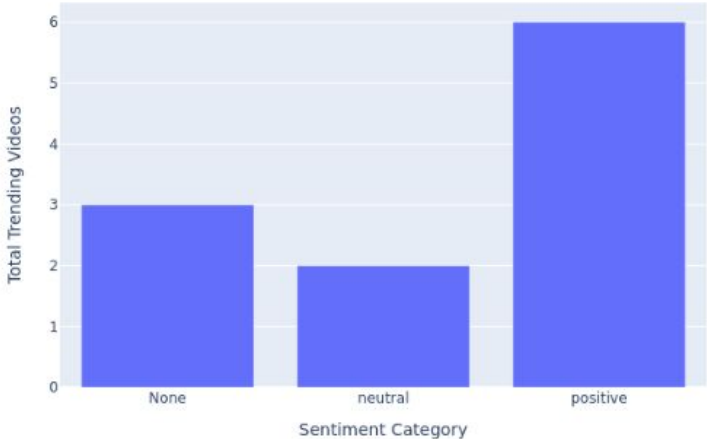


Region	Title	Stay Trending	Probability
GB	[Special Clip] ATEEZ(에이티즈) 00 & 00 '0000	Yes	75.23%
GB	The Pretty Reckless - Where Are You Chri	Yes	69.82%
CA	The Pretty Reckless - Where Are You Chri	Yes	69.81%
US	DDG - Yea I Kno (Music Video)	Yes	59.03%
GB	JP Fans - Ruzhowa (REMIX)	Yes	57.04%
CA	BAROTA (MUSIC VIDEO) SIDHU MOOSE WALA	Yes	56.68%
US	Treaty Oak Revival - Blue Star (Official	Yes	55.42%
GB	BAROTA (MUSIC VIDEO) SIDHU MOOSE WALA	Yes	54.48%
GB	Flavour - The Eagle Has Landed (Official	Yes	52.93%
CA	Character Trailer - "Jahoda: No Hunt in	Yes	51.32%
GB	MAVO - Shakabulizzy Remix (Feat. Davido)	Yes	50.64%
CA	TAEYEON 00 '00 (Panorama)' MV	Yes	50.15%
CA	Sheesha - Surjit Bhullar   Sargi Maan	No	46.18%
US	MICHAEL FLORES X DANI BARRANCO X ALOFOKE	No	46.17%
GB	Kalamkaval Pre release Teaser   Mammoott	No	45.22%
CA	Kalamkaval Pre release Teaser   Mammoott	No	45.22%
CA	Barota	No	44.86%
CA	Nakeyjakey is Outdated	No	44.65%
US	Nakeyjakey is Outdated	No	44.65%
GB	STEAL A BRAINROT GIVEAWAY LIVE   STEAL A	No	44.39%

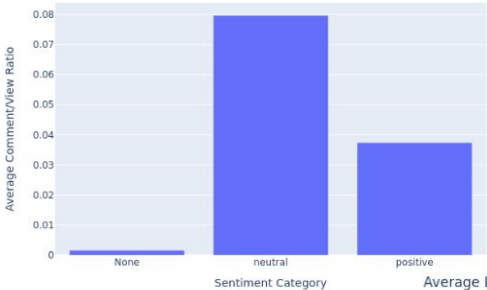


# Comments Notebook

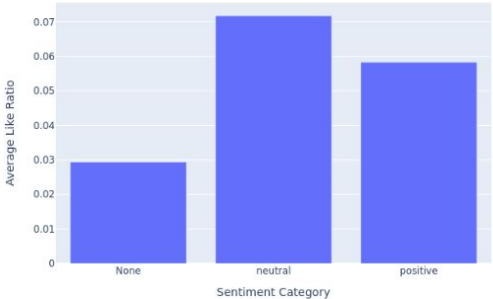
Trending Videos by Dominant Comment Sentiment



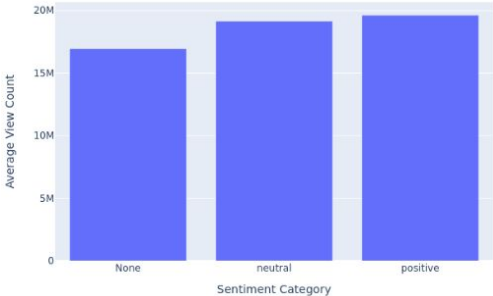
Average Comment-to-View Ratio by Dominant Sentiment



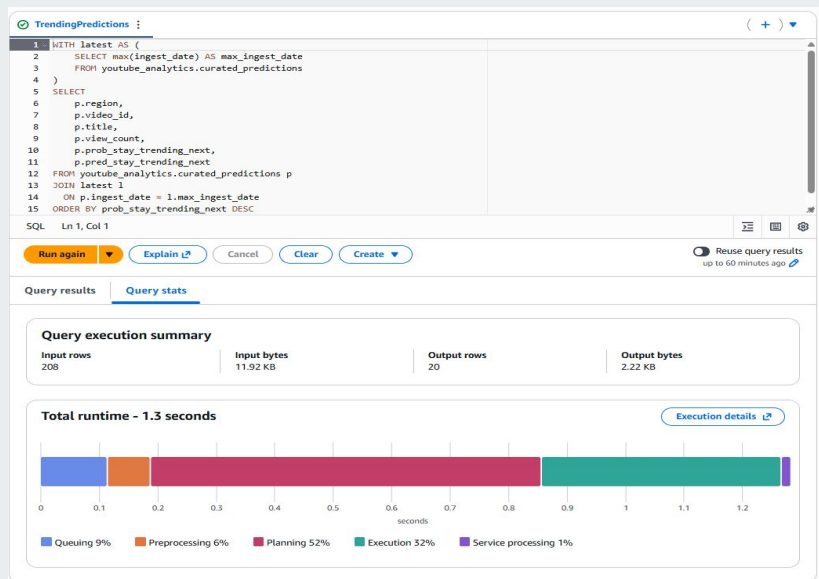
Average Like Ratio (likes / views) by Dominant Sentiment



Average View Count by Dominant Sentiment



# Amazon Athena



- Using the tables created by the glue crawlers, wrote SQL code to verify the data we were generating was correct
- Wrote SQL code to pull tables out with the data we were interested in





# Results

- Working Pipeline
  - Pull data from YouTube
  - Flatten trending and comments data to store in parquet files
  - Extract trending videos sentiment data
  - Populating a model for calculating predictions of videos trending the next day.
  - Run multiple notebooks to present data in a dashboard like environment
- Money Spent
  - ~\$240