

YouTube Trending Pipeline

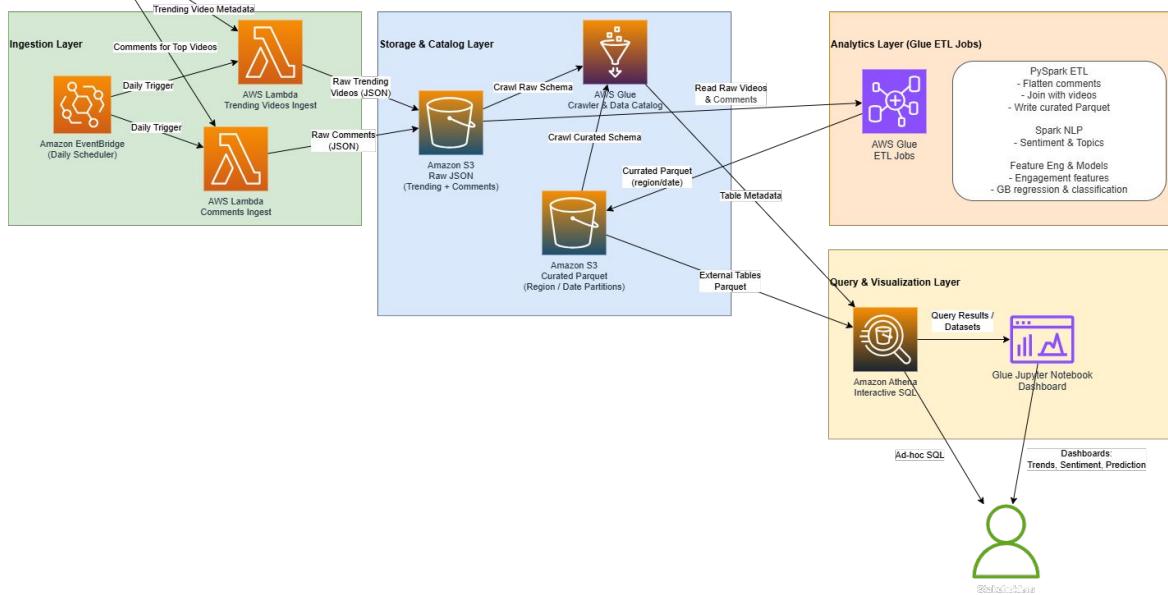
CS 6705 Applied Cloud Computing - Kevin Bell, Jacob Child



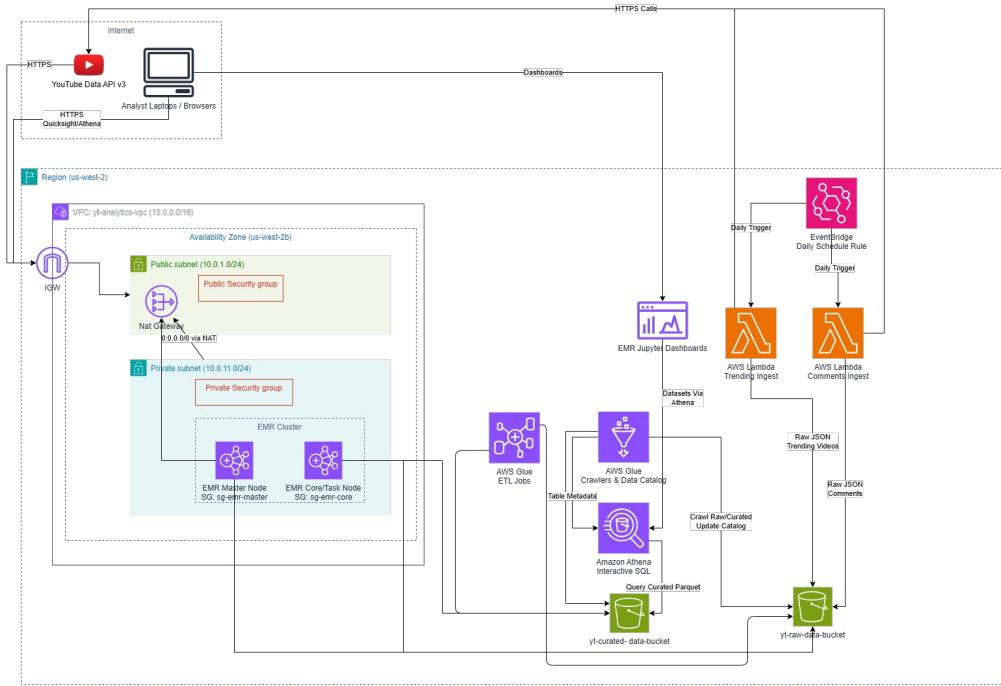
Project Goal

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-2:069233548392:rule/yt-trending-daily-schedule	Event bus ARN arn:aws:events:us-west-2:069233548392:event-bus/default	

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

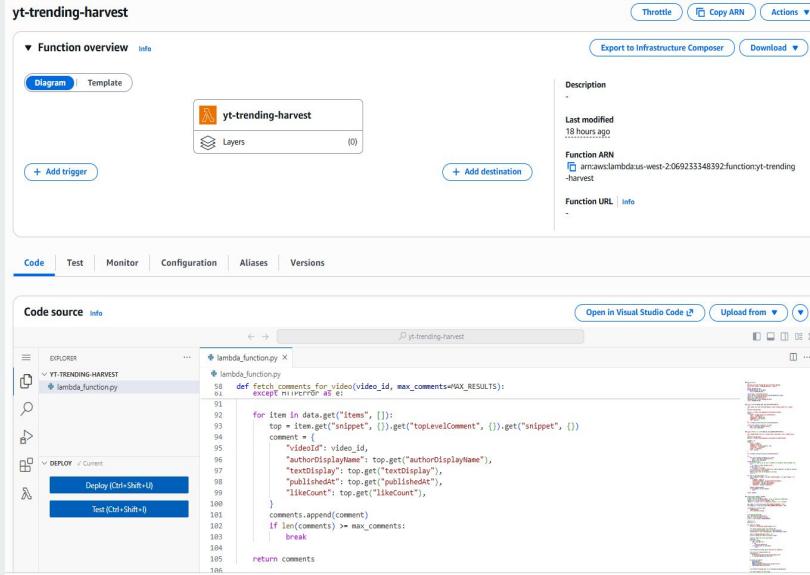
Targets

Details	Target Name	Type	ARN	Input	Role
▼	yt-trending-harvest	Lambda function	arn:aws:lambda:us-west-2:06923348392:function:yt-trending-harvest	Matched event	Amazon_EventBridge_Invoke_Lambda_2095584042

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



- 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)

[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"/>	athena-results/	Folder	-
<input type="checkbox"/>	curated_\$folder\$	-	November 20, 2025, 12:02:34 (UTC-07:00)
<input type="checkbox"/>	curated/	Folder	-
<input type="checkbox"/>	emr-logs/	Folder	-
<input type="checkbox"/>	jobs/	Folder	-
<input type="checkbox"/>	raw/	Folder	-
<input type="checkbox"/>	scripts/	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 workflow](#) [Edit](#) [Delete](#)

[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.

[View run details](#)

Workflow...	Status	Start time (UTC)	End time (UTC)	Current
wr_cf53a783c5b	Completed	December 2, 2025 at 16:56:5	December 2, 2025 at 17:19:5	
wr_0738f44682e	Completed	December 2, 2025 at 06:00:5	December 2, 2025 at 06:24:4	
wr_a8919b6dc2e	Completed	December 1, 2025 at 23:06:2	December 1, 2025 at 23:24:0	
wr_01cdafbcadd	Completed	December 1, 2025 at 06:00:5	December 1, 2025 at 06:22:0	
wr_b59323863el	Completed	November 30, 2025 at 06:00	November 30, 2025 at 06:21	
wr_f4a4ad96619	Completed	November 29, 2025 at 06:00	November 29, 2025 at 06:21	
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

The screenshot shows the AWS Glue Studio interface. At the top, there are three options for creating a job: "Visual ETL" (selected), "Notebook", and "Script editor". Below this is a section titled "Example jobs" with a "Create example job" button. The main area is titled "Your jobs (8)" and shows a table of existing jobs. The table has columns for Job name, Type, Created by, Last modified, AWS Glue version, and Action. The jobs listed are:

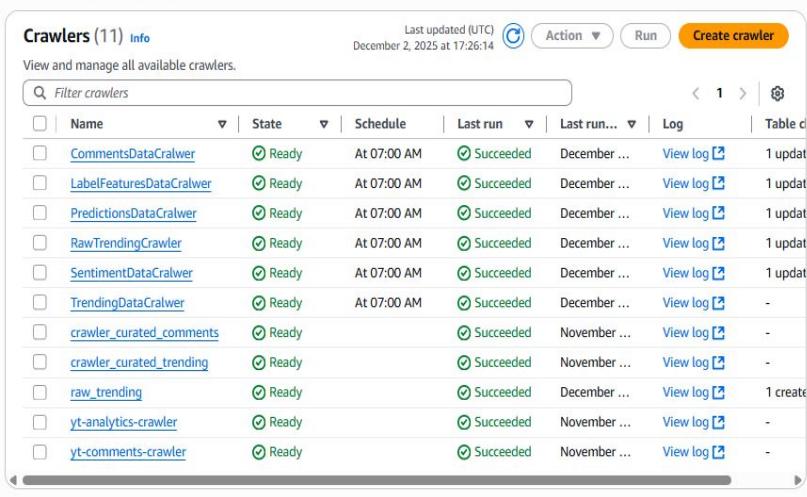
Action	Job name	Type	Created by	Last modified	AWS Glue version	Action
	yt-feature_labels_job	Glue ETL	Script	12/2/2025, 10:11:25 AM	5.0	-
	yt-predictions_job	Glue ETL	Script	12/1/2025, 10:06:07 PM	5.0	-
	yt-train_models_job	Glue ETL	Script	12/1/2025, 9:42:28 PM	5.0	-
	yt-onetime-batch	Glue ETL	Script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



The screenshot shows the AWS Glue Crawler list interface. At the top, it says "Crawlers (11) Info" and "Last updated (UTC) December 2, 2025 at 17:26:14". There are buttons for "Action", "Run", and "Create crawler". Below is a search bar labeled "Filter crawlers". A table lists 11 crawlers:

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

- 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](#)

Filter jobs by property 3 matches

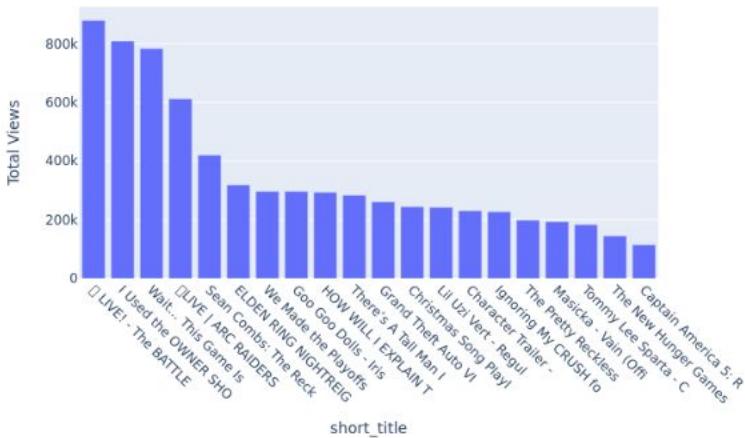
Created by = NOTEBOOK

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

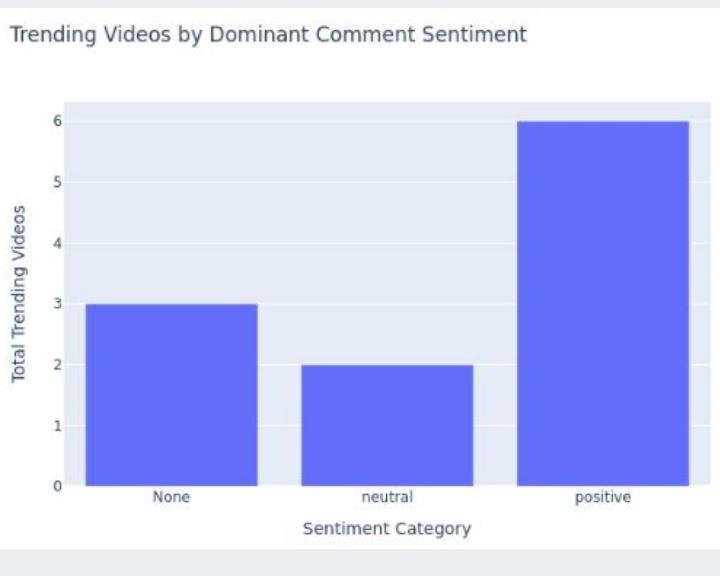
Top 20 Most Viewed Trending Videos in US



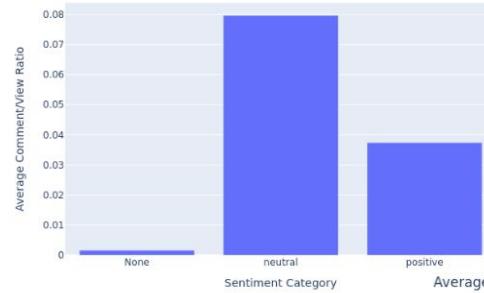
Region	Title	Stay Trending	Probability
GB	[Special Clip] ATEEZ(에이티즈) 『』 & 『』 '『』『』'	Yes	75.23%
GB	The Pretty Reckless - Where Are You Chri	Yes	69.82%
CA	The Pretty Reckless - Where Are You Chri	Yes	69.81%
US	DDG - Yea I Kno (Music Video)	Yes	59.03%
GB	JP Fans - Ruzhowa (REMIX)	Yes	57.04%
CA	BAROTA (MUSIC 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 『』 '『』 (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

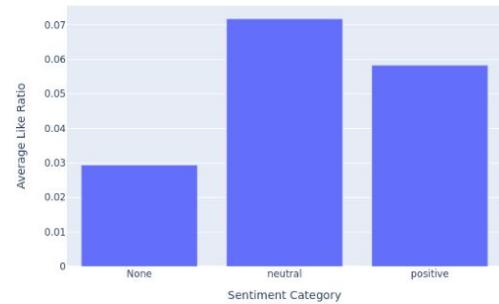
Trending Videos by Dominant Comment Sentiment



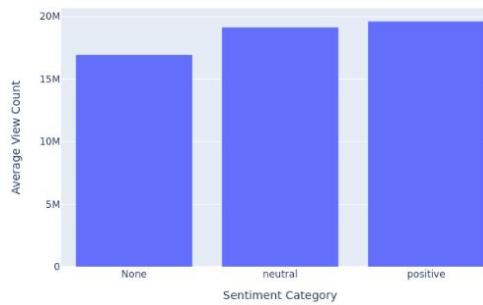
Average Comment-to-View Ratio by Dominant Sentiment



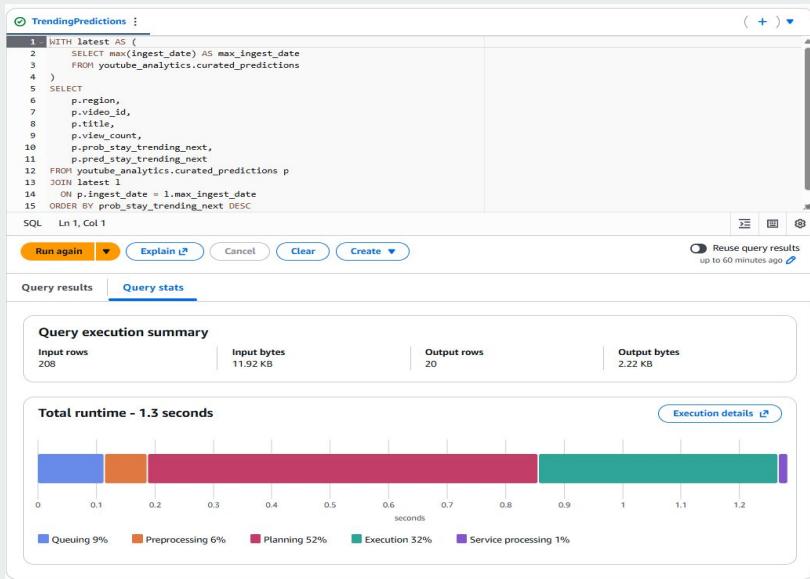
Average Like Ratio (likes / views) by Dominant Sentiment



Average View Count by Dominant Sentiment



Amazon Athena



The screenshot shows the Amazon Athena console interface. At the top, there's a code editor window titled "TrendingPredictions" containing the following SQL query:

```
1 WITH latest AS (
2     SELECT max(ingest_date) AS max_ingest_date
3         FROM youtube_analytics.curated_predictions
4 )
5   SELECT
6     p.region,
7     p.video_id,
8     p.title,
9     p.view_count,
10    p.prob_stay_trending_next,
11    p.pred_stay_trending_next
12  FROM youtube_analytics.curated_predictions p
13  JOIN latest l
14  ON p.ingest_date = l.max_ingest_date
15 ORDER BY prob_stay_trending_next DESC
```

Below the code editor are several buttons: "Run again", "Explain", "Cancel", "Clear", and "Create". To the right of these buttons is a "Reuse query results" toggle switch. The "Query stats" tab is selected, showing the "Query execution summary" and "Total runtime - 1.3 seconds". The "Execution details" chart shows the breakdown of the total runtime by task type:

Task Type	Percentage
Queuing	9%
Preprocessing	6%
Planning	52%
Execution	32%
Service processing	1%

- 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