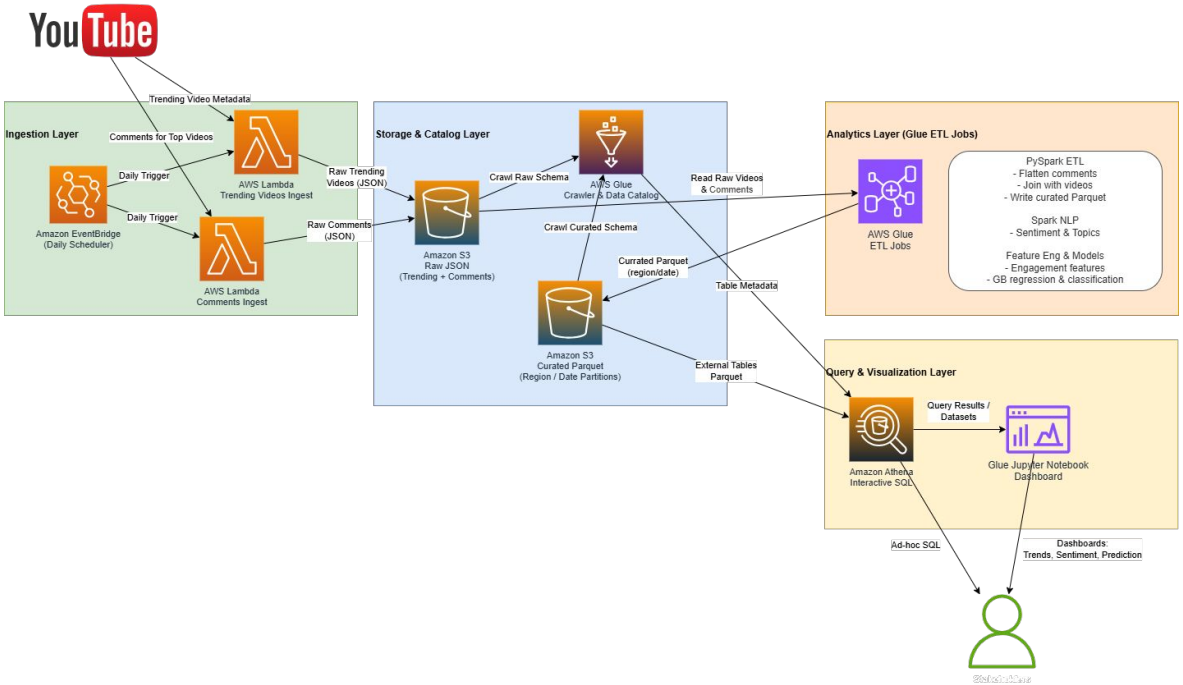
YouTube Trending Pipeline

CS 6705 Applied Cloud Computing - Kevin Bell, Jacob Child

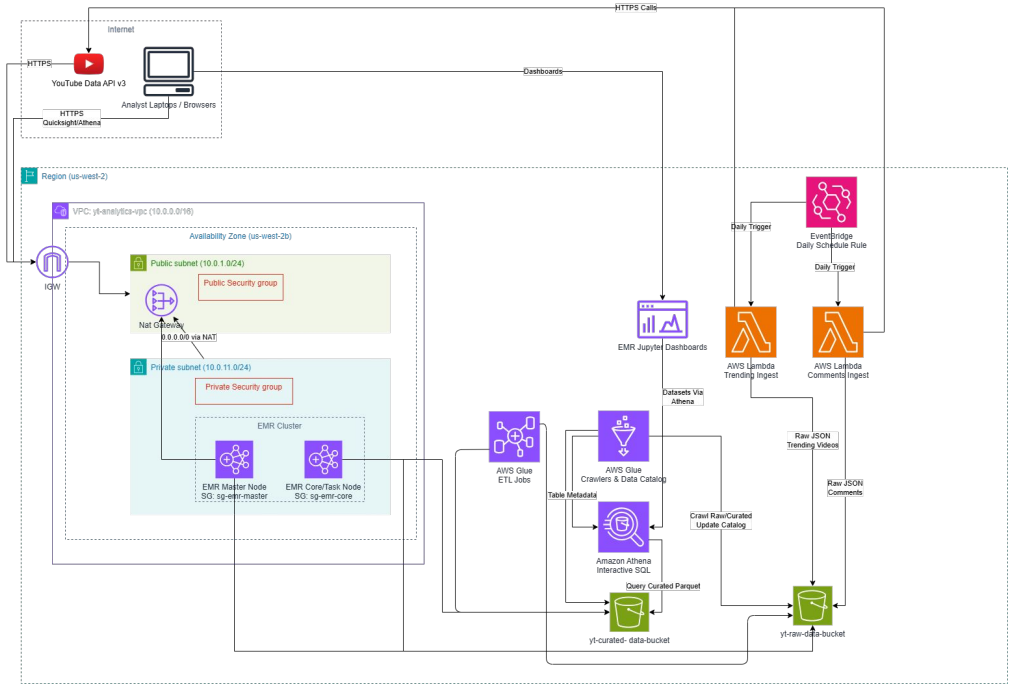
Project Goal

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



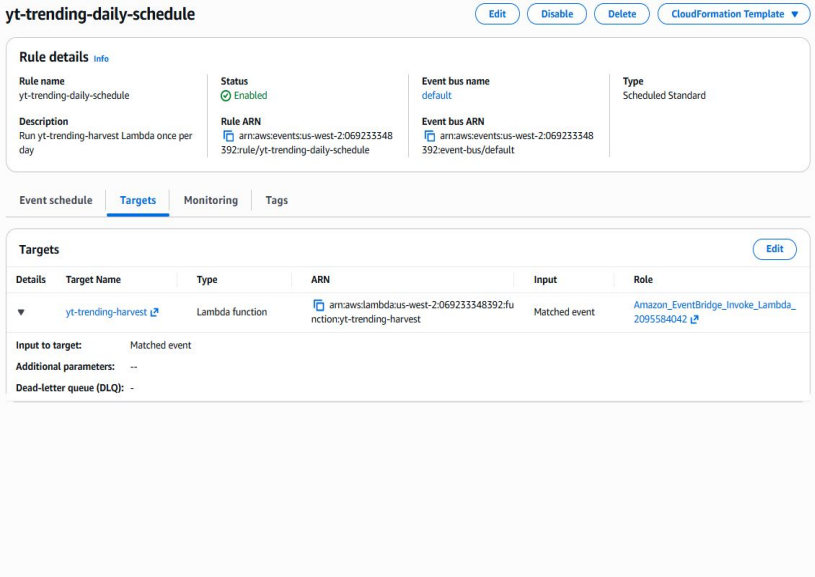
AWS Services Used

- 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)

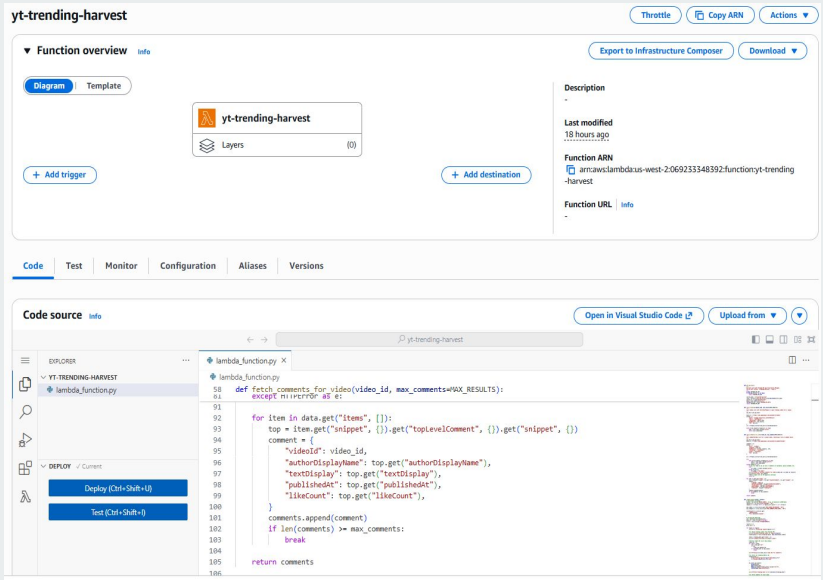
- Set up a VPC to have a place to run this pipeline and manage access to the data
- Created a public and private subnet as well as different security groups to facilitate access
- Created an Internet Gateway (IGW) to allow parts of our cloud to access the internet
- Created a Network Attached Translation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

Amazon EventBridge



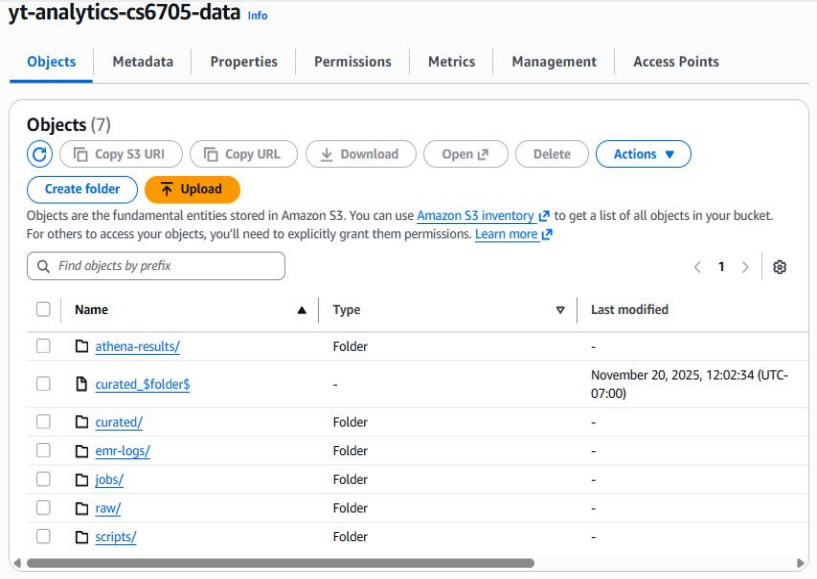
- Used to schedule and launch AWS Lambda jobs everyday at 6 am.

AWS Lambda



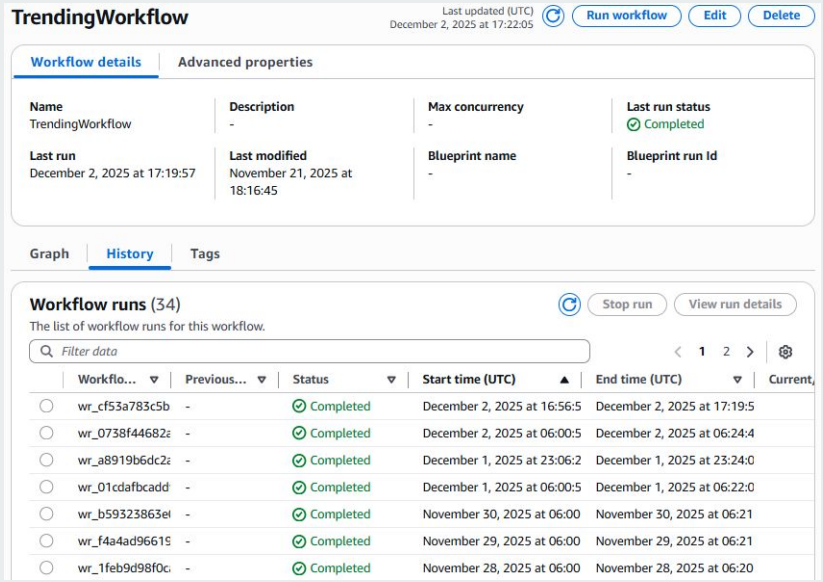
- Wrote a Python script that was run as a Lambda job
- The Python script utilized the YouTube API key stored in AWS Secrets to pull trending video data and the comments associated with those trending videos.
- This Lambda job also launches the AWS Glue Workflow

Amazon S3



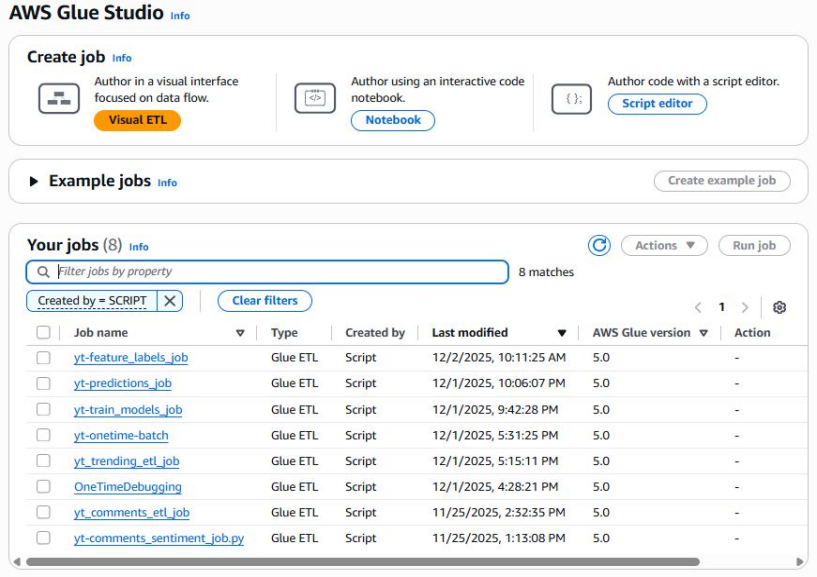
- Created a storage location in S3 to store
  - Raw data pulled from YouTube
  - Curated data flattened by ETL Jobs
  - Predicted data created by ETL jobs
  - Scripts used by Glue

AWS Glue Workflow



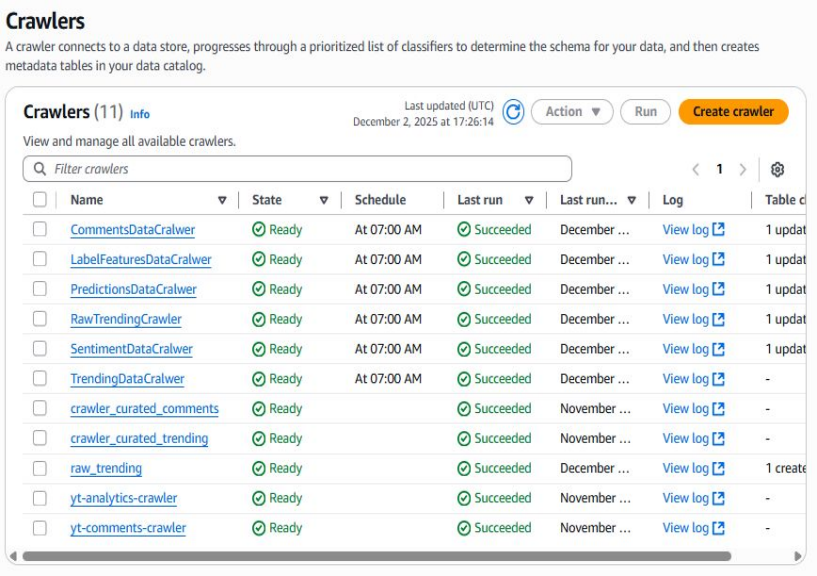
- Created a workflow that connects all of our ETL scripts.
  - Trending ETL Job
  - Comments ETL Job
  - Comments Sentiment ETL Job
  - Feature Label ETL Job
  - Tran Model ETL Job
  - Predictions ETL Job

AWS Glue ETL Jobs



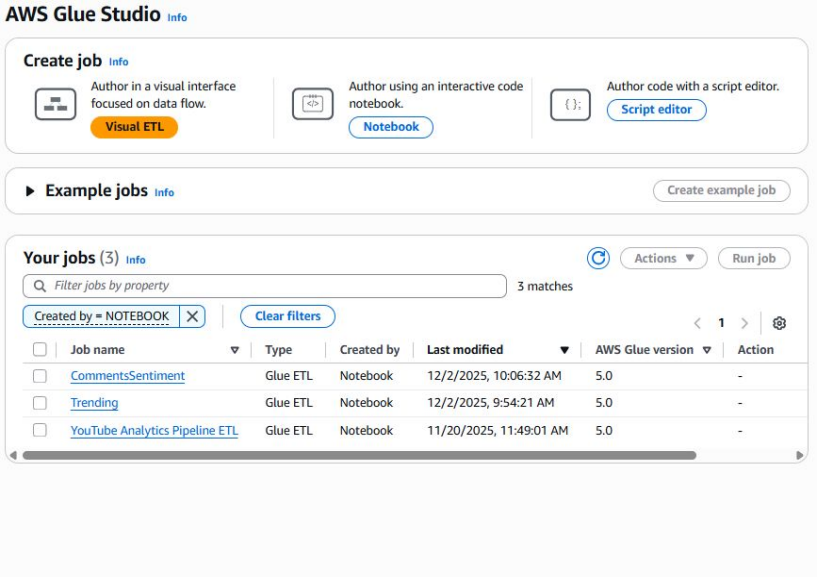
- Trending ETL Job
  - Flattens the raw trending video data
- Comments ETL Job
  - Flattens the raw comments
- Comments Sentiment ETL Job
  - Pulls out the sentiment data from the comments raw data
- Feature Label ETL Job
  - Creates labels for training the model
- Tran Model ETL Job
  - Updates the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

AWS Glue Crawlers



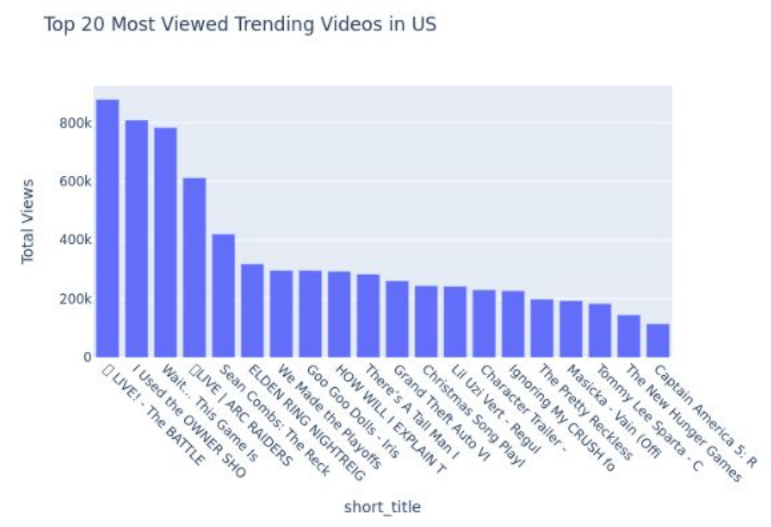
- Utilized crawlers to look over the parquet files in the S3 curated area
- Generated tables to store the schema structure to use Athena to query the data

AWS Glue ETL Notebook



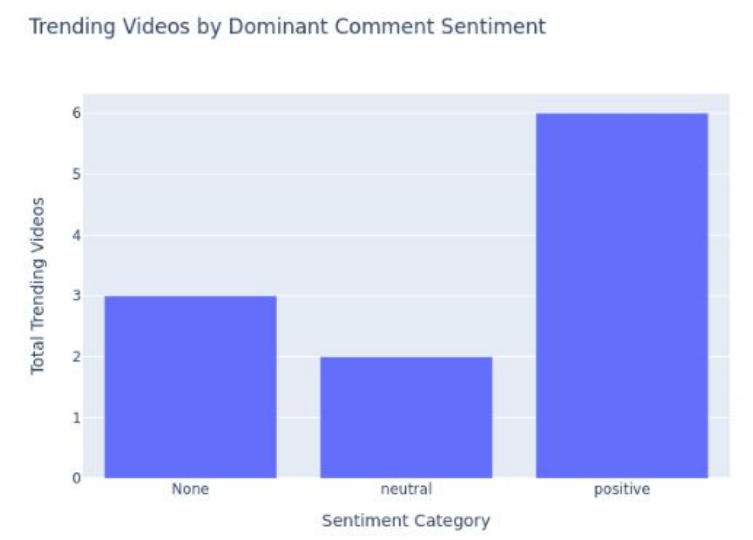
- Created Glue ETL notebooks to read the data and process it using python and panda
- Created plotly graphs to act as a dashboard since Quicksight wasn't available

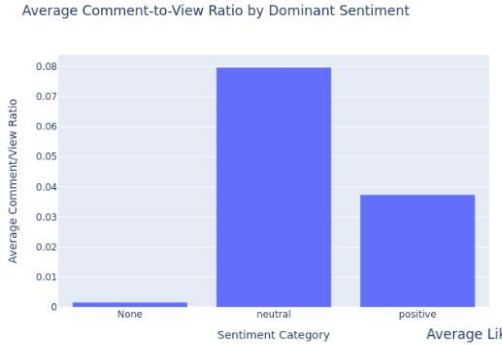
Trending Notebook

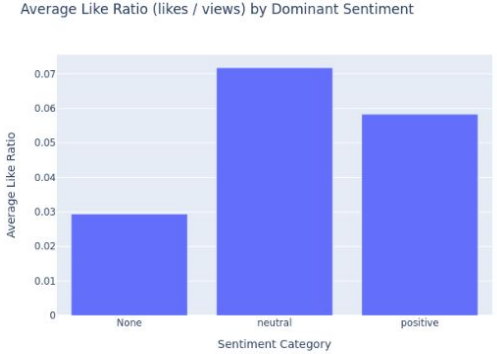


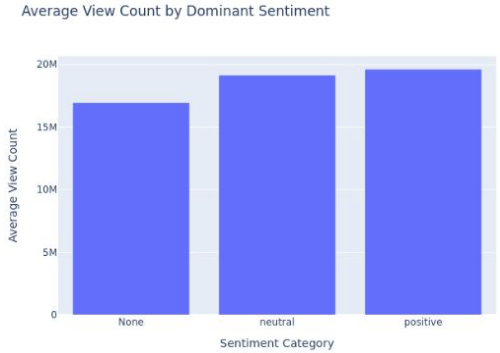
Region	Title	Stay Trending	Probability
GB	[Special Clip] ATEEZ(□□□□) □□ & □□ '□□□□	Yes	75.23%
GB	The Pretty Reckless - Where Are You Chri	Yes	69.82%
CA	The Pretty Reckless - Where Are You Chri	Yes	69.81%
US	DDG - Yea I Kno (Music Video)	Yes	59.03%
GB	JP Fans - Ruzhowa (REMIX)	Yes	57.04%
CA	BAROTA (MUSIC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

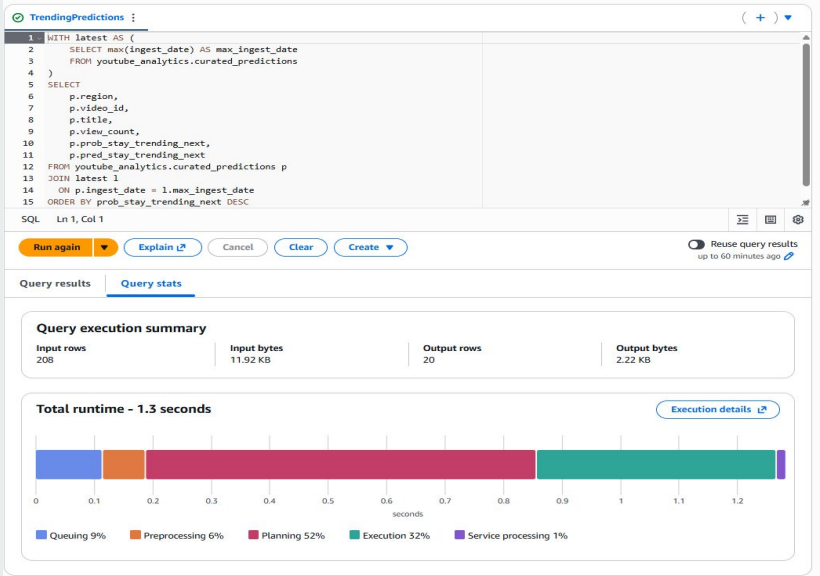








Amazon Athena



- 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