DATA ENGINEERING

Real Time YouTube Comment Sentiment Analysis

Presented by -

Rohan Mahaveer
Manoj Saligrama Harisha
Namratha Prakash
Valeska Joshna Dsouza

INTRODUCTION

- Domain: real-time analysis of YouTube comments.

- **Goal:** Build an end-to-end pipeline from ingestion → batch processing → streaming → ML predictions → storage and monitoring

This Pipeline Delivers:

- Automated, serverless GCP architecture that transforms chaotic comment streams into actionable sentiment metrics

- From Raw Comments → Strategic Insights .

PROBLEM STATEMENT

- YouTube comments contain rich signals about viewer satisfaction, content quality, and emerging issues.
- Manual review and purely batch analytics do not scale for high-volume channels or real-time moderation.
- We need a cloud-based system that can analyse comment sentiment both historically (batch) and in near real-time (streaming).
- Our project focuses on real-time YouTube comment sentiment analysis using a two-phase big data architecture

TWO PHASE ARCHITECTURE

Phase 1: Training

Batch pipeline in Airflow + BigQuery for cleaning, aggregation, feature engineering, and BigQuery ML model training.

Phase 2: Inference

Streaming via Pub/Sub and Dataflow, applying the trained BigQuery ML model to new comments hourly and storing predictions in BigQuery.

OBJECTIVES

- Design a simple, scalable cloud architecture for real-time YouTube comment sentiment analysis with 4–5 core GCP components
- Implement a batch pipeline (Airflow → BigQuery) to ingest, clean, aggregate comments and train a sentiment model with BigQuery ML
- Implement a streaming pipeline (Pub/Sub → Dataflow → BigQuery) that scores new comments in micro-batches every 30–60 seconds and runs hourly sentiment analytics
- Apply Little's Law and basic queuing theory to estimate throughput, latency, and worker/consumer sizing for both batch and streaming phases
- Set up basic monitoring and logging for data volume, processing time, and error rates, and provide clear documentation, diagrams, and demo instructions

TOOLS USED

Orchestration Tool

Airflow

Streaming Platform

Pub/Sub

Data Store

BigQuery

Pipeline between Pub/Sub & BigQuery

Dataflow

Model Training

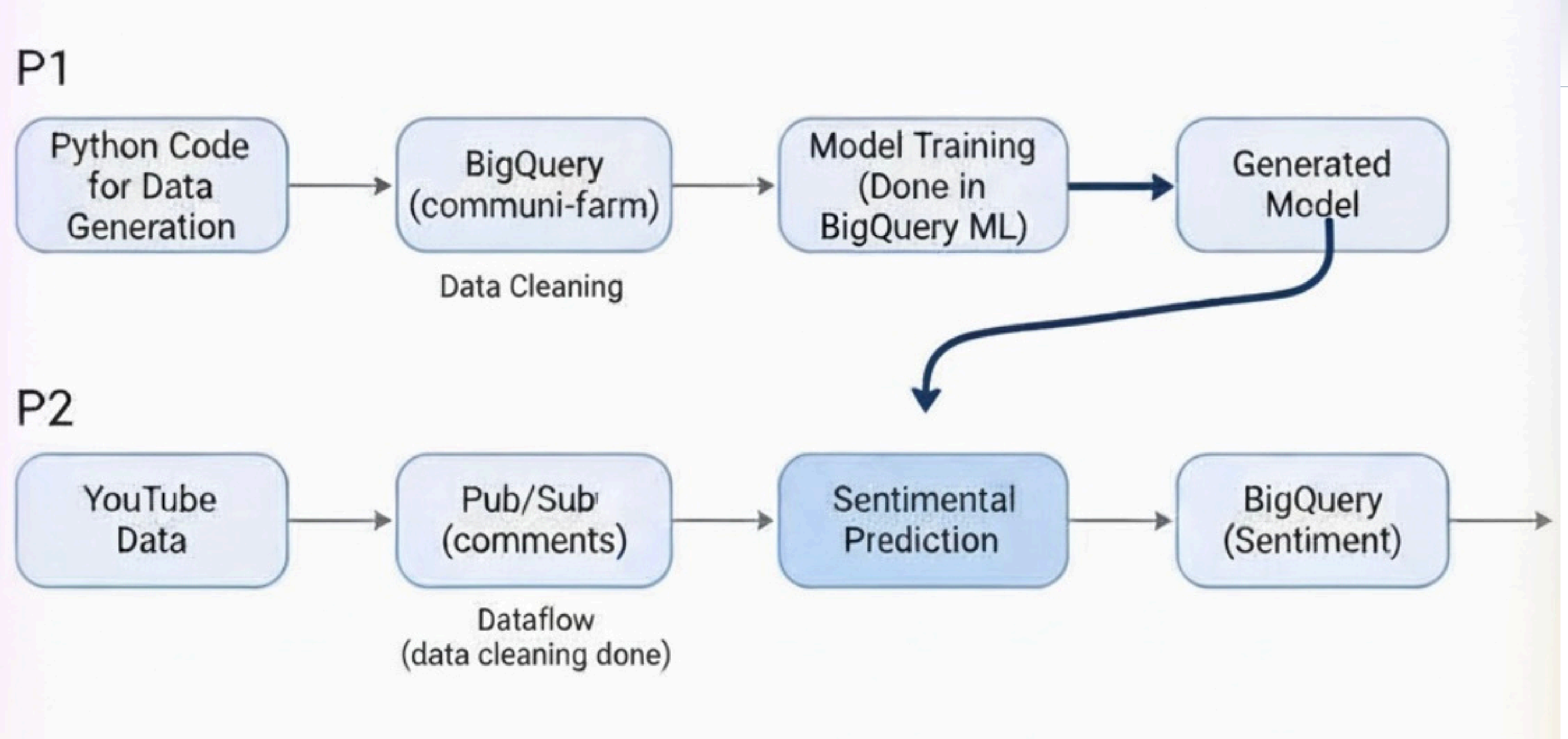
BigQueryML

PROBLEM STATEMENT

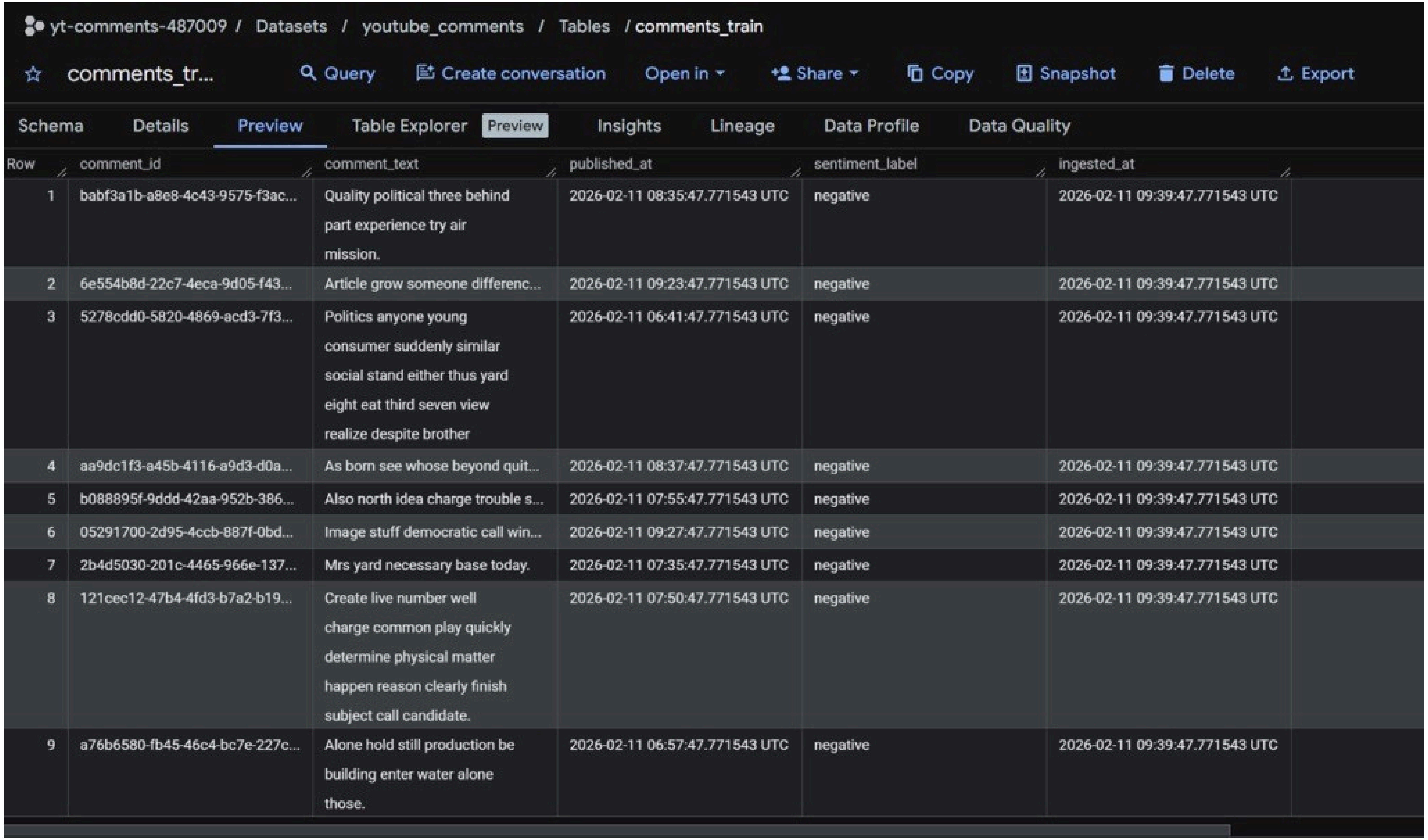
- YouTube comments contain rich signals about viewer satisfaction, content quality, and emerging issues.
- Manual review and purely batch analytics do not scale for high-volume channels or real-time moderation.
- We need a cloud-based system that can analyse comment sentiment both historically (batch) and in near real-time (streaming).
- Our project focuses on real-time YouTube comment sentiment analysis using a two-phase big data architecture

DATA GENERATION & CLEANING

- Python script generates synthetic YouTube comments
- **Faker** is used to create synthetic YouTube comments for training
- Script continuously inserts rows into the comments_train table in **BigQuery**
- Cleaned data is stored in BigQuery and later used as the training dataset for the sentiment model



DATA GENERATION & CLEANING



Row	comment_id	comment_text	published_at	sentiment_label	ingested_at
1	babf3a1b-a8e8-4c43-9575-f3ac...	Quality political three behind part experience try air mission.	2026-02-11 08:35:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
2	6e554b8d-22c7-4eca-9d05-f43...	Article grow someone differenc...	2026-02-11 09:23:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
3	5278cdd0-5820-4869-acd3-7f3...	Politics anyone young consumer suddenly similar social stand either thus yard eight eat third seven view realize despite brother	2026-02-11 06:41:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
4	aa9dc1f3-a45b-4116-a9d3-d0a...	As born see whose beyond quit...	2026-02-11 08:37:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
5	b088895f-9ddd-42aa-952b-386...	Also north idea charge trouble s...	2026-02-11 07:55:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
6	05291700-2d95-4ccb-887f-0bd...	Image stuff democratic call win...	2026-02-11 09:27:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
7	2b4d5030-201c-4465-966e-137...	Mrs yard necessary base today.	2026-02-11 07:35:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
8	121cec12-47b4-4fd3-b7a2-b19...	Create live number well charge common play quickly determine physical matter happen reason clearly finish subject call candidate.	2026-02-11 07:50:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC
9	a76b6580-fb45-46c4-bc7e-227c...	Alone hold still production be building enter water alone those.	2026-02-11 06:57:47.771543 UTC	negative	2026-02-11 09:39:47.771543 UTC

PHASE 1: MODEL TRAINING ➝ Generated Model

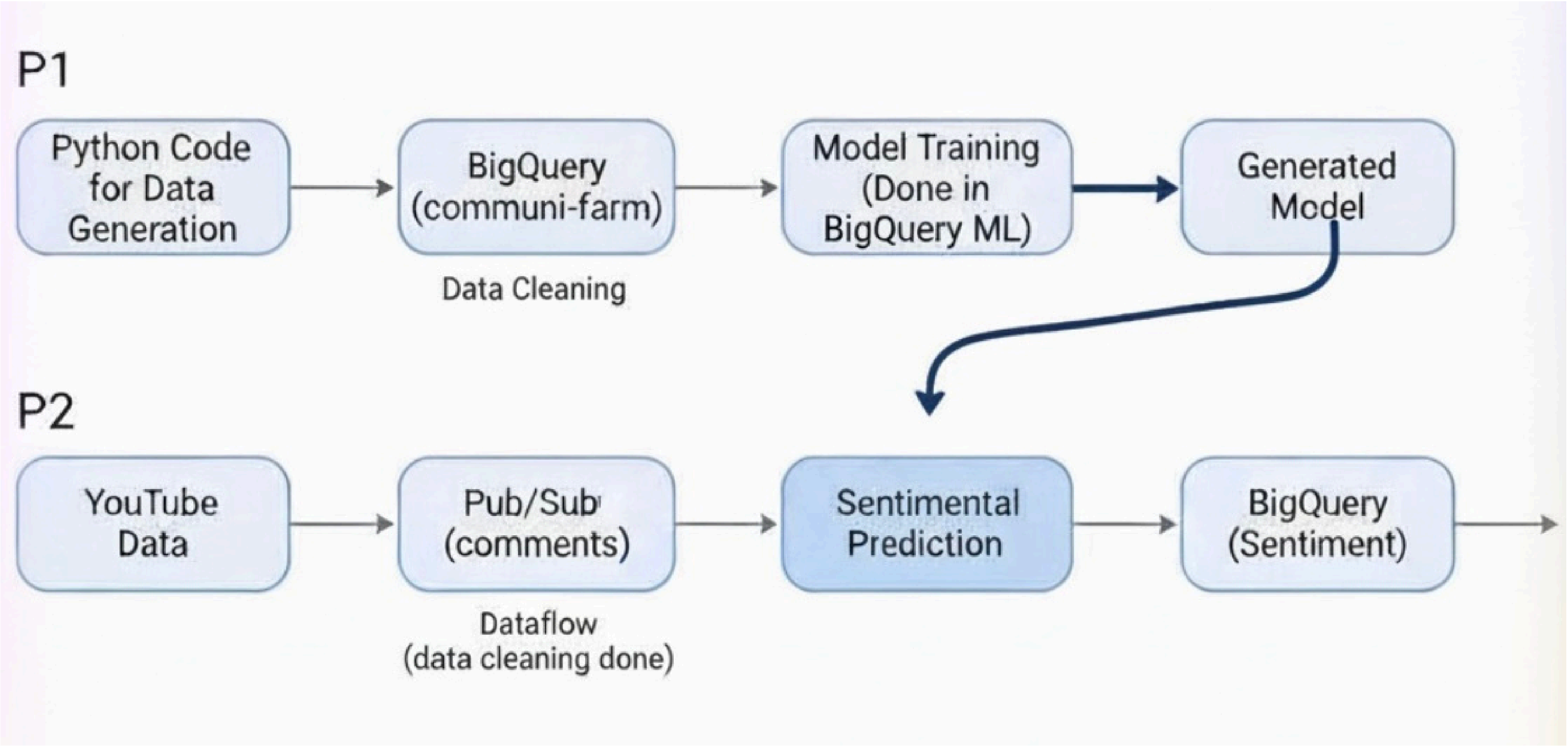
4-STEP PIPELINE ➝ GENERATED MODEL

- PYTHON DATA → communi-farm dataset
  Generate labeled training dataset

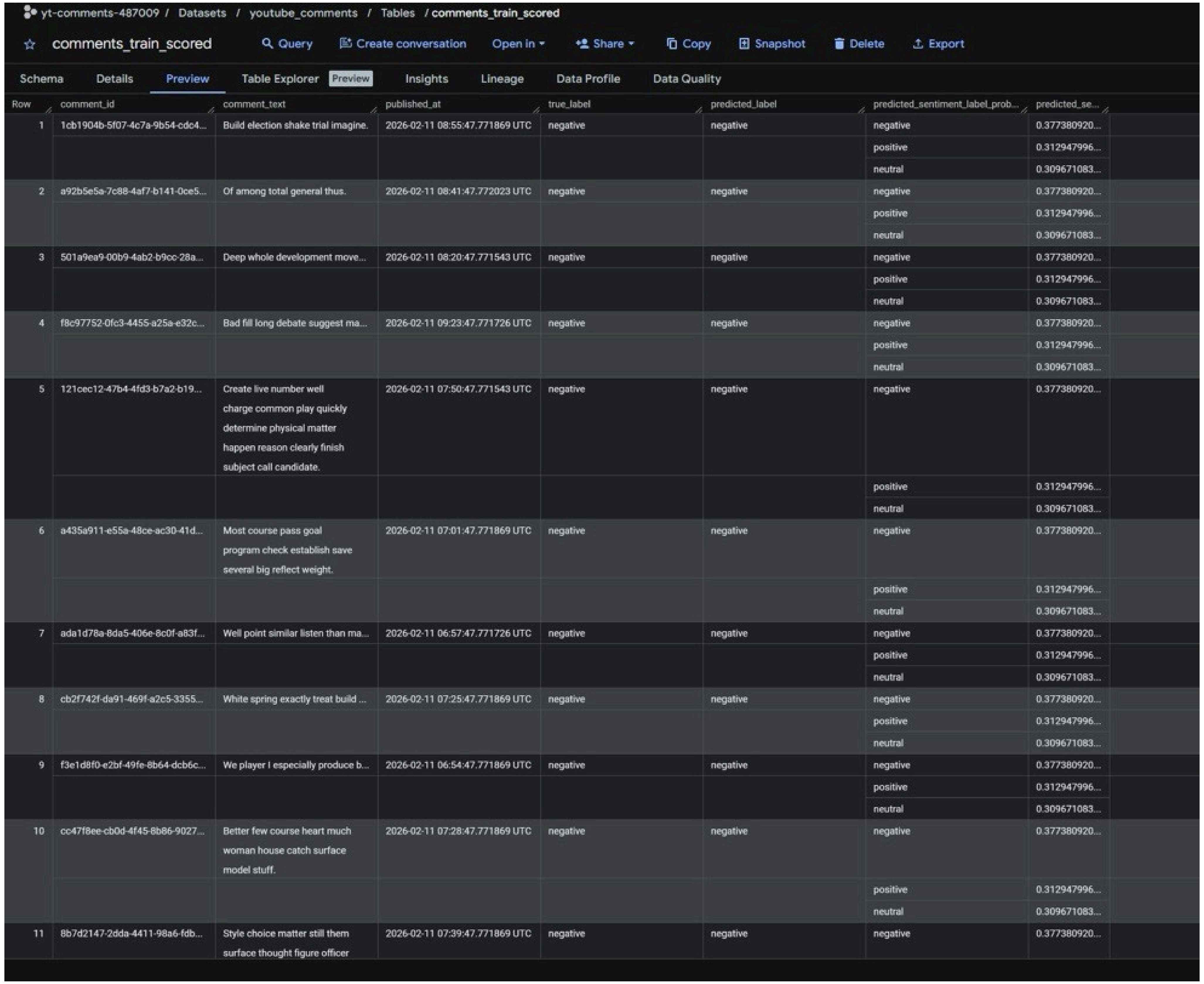
- BIGQUERY CLEANING
  Transform raw text for ML

- BIGQUERY ML TRAINING CREATE MODEL sentiment_model (LOGISTIC_REG)
  Serverless model training.

GENERATED MODEL READY P2 Predictions



SENTIMENT RESULTS (P1 PIPELINE)



Row	comment_id	comment_text	published_at	true_label	predicted_label	predicted_sentiment_label_prob...	predicted_se...
1	1cb1904b-5f07-4c7a-9b54-cdc4...	Build election shake trial imagine.	2026-02-11 08:55:47.771869 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
2	a92b5e5a-7c88-4af7-b141-0ce5...	Of among total general thus.	2026-02-11 08:41:47.772023 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
3	501a9ea9-00b9-4ab2-b9cc-28a...	Deep whole development move...	2026-02-11 08:20:47.771543 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
4	f8c97752-0fc3-4455-a25a-e32c...	Bad fill long debate suggest ma...	2026-02-11 09:23:47.771726 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
5	121cec12-47b4-4fd3-b7a2-b19...	Create live number well charge common play quickly determine physical matter happen reason clearly finish subject call candidate.	2026-02-11 07:50:47.771543 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
6	a435a911-e55a-48ce-ac30-41d...	Most course pass goal program check establish save several big reflect weight.	2026-02-11 07:01:47.771869 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
7	ada1d78a-8da5-406e-8c0f-a83f...	Well point similar listen than ma...	2026-02-11 06:57:47.771726 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
8	cb2f742f-da91-469f-a2c5-3355...	White spring exactly treat build ...	2026-02-11 07:25:47.771869 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
9	f3e1d8f0-e2bf-49fe-8b64-dcb6c...	We player I especially produce b...	2026-02-11 06:54:47.771869 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
10	cc47f8ee-cb0d-4f45-8b86-9027...	Better few course heart much woman house catch surface model stuff.	2026-02-11 07:28:47.771869 UTC	negative	negative	negative positive neutral	0.377380920... 0.312947996... 0.309671083...
11	8b7d2147-2dda-4411-98a6-fdb...	Style choice matter still them surface thought figure officer	2026-02-11 07:39:47.771869 UTC	negative	negative	negative	0.377380920...

PROCESS 2

YOUTUBE COMMENT DATA AND PUB/SUB

PIPELINE ARCHITECTURE

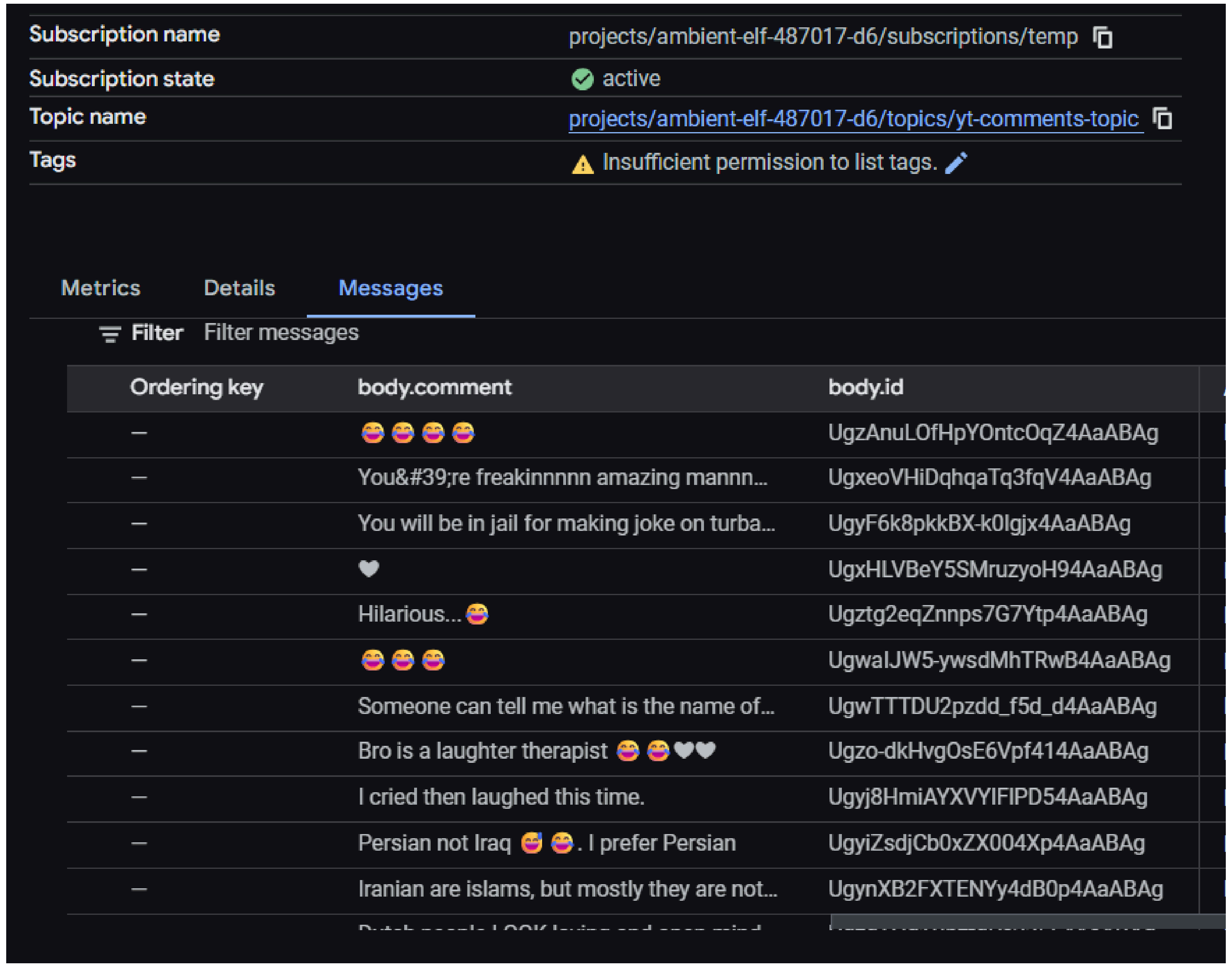
1.**YouTube Data API v3** - Source of video data
2.**Cloud Pub/Sub** - Message broker for streaming
3.**Cloud Run** - Serverless compute for processing
4.**Push Subscription** - Event-driven integration

DATA FLOW

1. **Trigger**(Youtube webhook or API poll
2. **Publish** (Message to Pub/Sub topis)
3. **Process** (Cloud Run Handels the Event along with cleaning the data aswell)

ARCHITECTURE BENEFITS

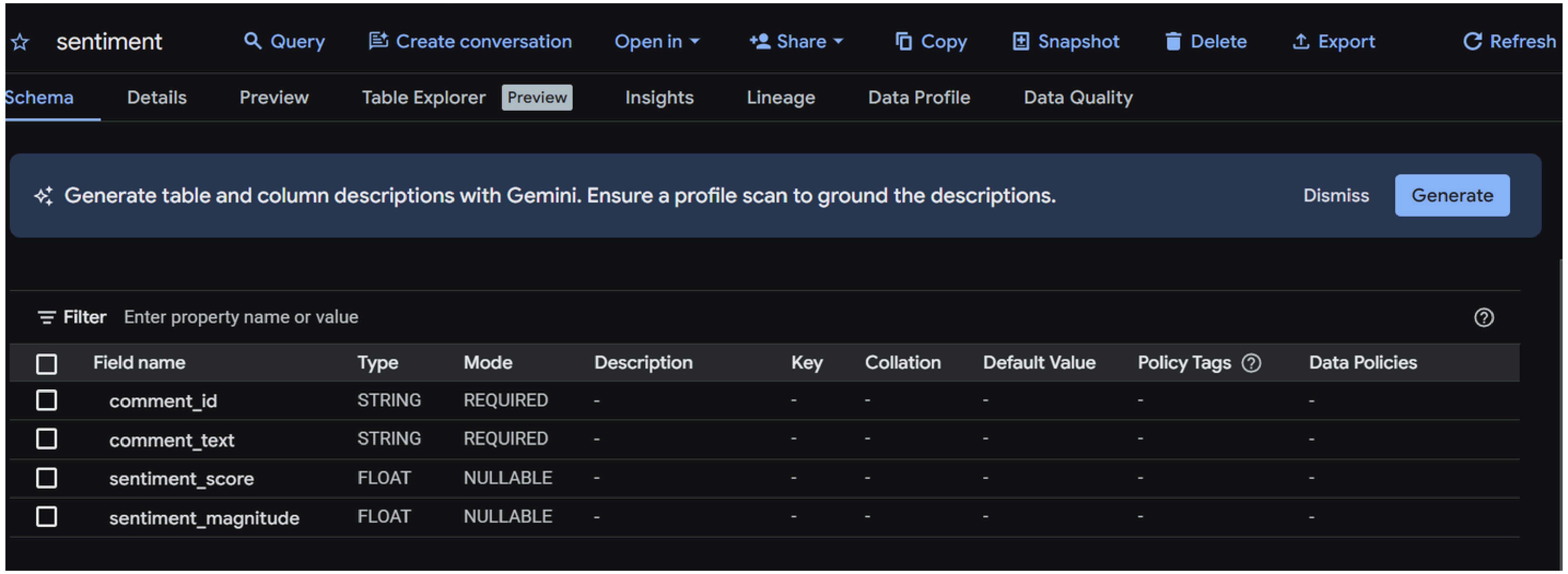
Real-time - Event driven processing with minimal latancy.



Subscription name	projects/ambient-elf-487017-d6/subscriptions/temp
Subscription state	✓ active
Topic name	projects/ambient-elf-487017-d6/topics/yt-comments-topic
Tags	⚠ Insufficient permission to list tags.

Ordering key	body.comment	body.id
–	😂😂😂😂	UgzAnuLOfHpYOntcOqZ4AaABAg
–	You&#39;re freakinnnnn amazing mannn...	UgxeoVHiDqhqaTq3fqV4AaABAg
–	You will be in jail for making joke on turba...	UgyF6k8pkkBX-k0lgjx4AaABAg
–	🤍	UgxHLVBeY5SMruzyoH94AaABAg
–	Hilarious...😂	Ugztg2eqZnnps7G7Ytp4AaABAg
–	😂😂😂	UgwaIJW5-ywsdMhTRwB4AaABAg
–	Someone can tell me what is the name of...	UgwTTTDU2pzdd_f5d_d4AaABAg
–	Bro is a laughter therapist 😂😂🤍🤍	Ugzo-dkHvgOsE6Vpf414AaABAg
–	I cried then laughed this time.	Ugyj8HmiAYXVYIFIPD54AaABAg
–	Persian not Iraq 😅😂. I prefer Persian	UgyiZsdjCb0xZX004Xp4AaABAg
–	Iranian are islams, but mostly they are not...	UgynXB2FXTENYy4dB0p4AaABAg

SENTIMENT PREDICTICTION

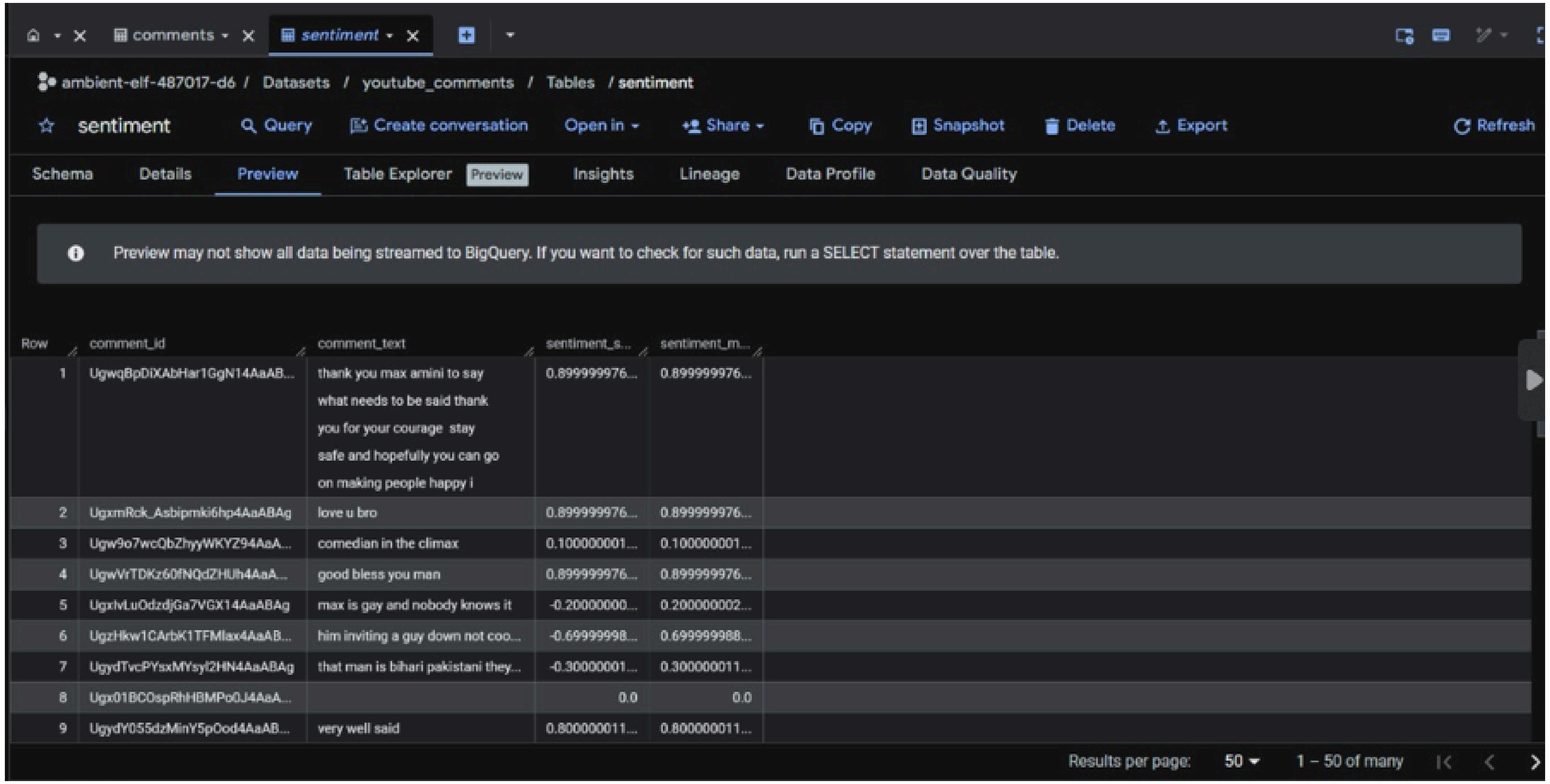


Generate table and column descriptions with Gemini. Ensure a profile scan to ground the descriptions.

☐	Field name	Type	Mode	Description	Key	Collation	Default Value	Policy Tags	Data Policies
☐	comment_id	STRING	REQUIRED	-	-	-	-	-	-
☐	comment_text	STRING	REQUIRED	-	-	-	-	-	-
☐	sentiment_score	FLOAT	NULLABLE	-	-	-	-	-	-
☐	sentiment_magnitude	FLOAT	NULLABLE	-	-	-	-	-	-

- Stores processed sentiment results
- Enables analytical queries
- Can feed downstream ML systems

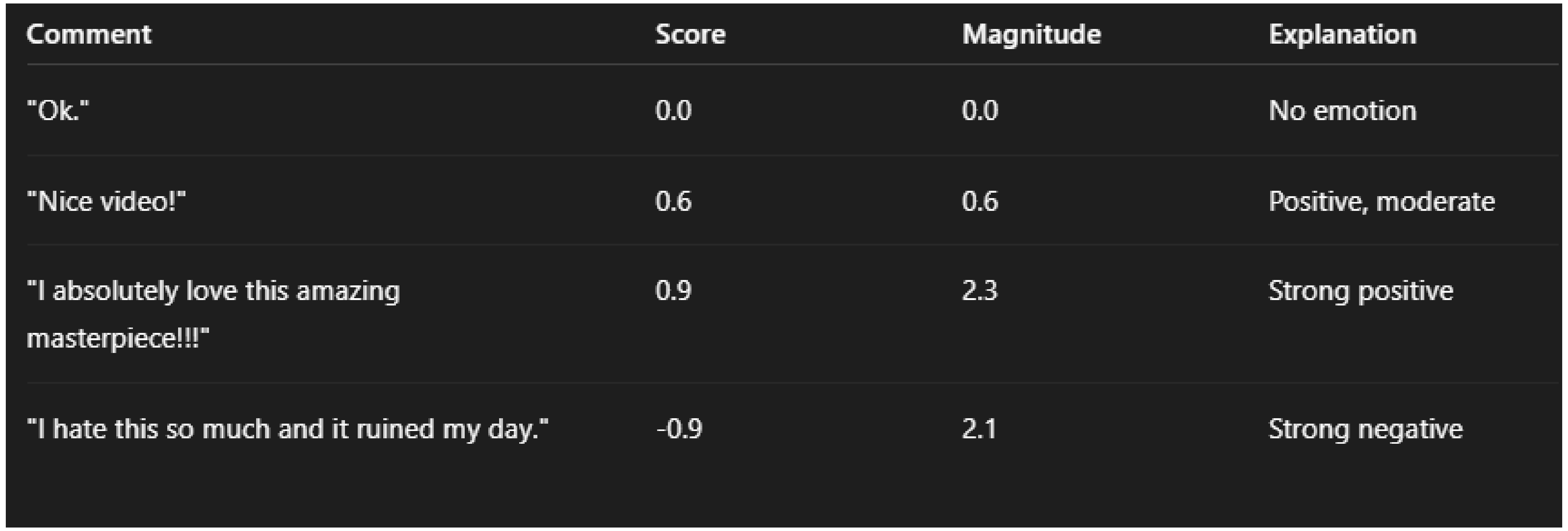
DATA IN BIGQUERY



Preview may not show all data being streamed to BigQuery. If you want to check for such data, run a SELECT statement over the table.

Row	comment_id	comment_text	sentiment_s...	sentiment_m...
1	UgwqBpDiXAbHar1GgN14AaAB...	thank you max amini to say what needs to be said thank you for your courage stay safe and hopefully you can go on making people happy i	0.899999976...	0.899999976...
2	UgxmRck_Asbipmki6hp4AaABAg	love u bro	0.899999976...	0.899999976...
3	Ugw9o7wcQbZhyyWKYZ94AaA...	comedian in the climax	0.100000001...	0.100000001...
4	UgwVrTDKz60fNQdZHUh4AaA...	good bless you man	0.899999976...	0.899999976...
5	UgxlvLuOdzdjGa7VGX14AaABAg	max is gay and nobody knows it	-0.20000000...	0.200000002...
6	UgzHkw1CArbK1TFMlax4AaAB...	him inviting a guy down not coo...	-0.69999998...	0.699999988...
7	UgydTvcPYsxMYsyl2HN4AaABAg	that man is bihari pakistani they...	-0.30000001...	0.300000011...
8	Ugx01BCOspRhHBMPo0J4AaA...		0.0	0.0
9	UgydY055dzMinY5pOod4AaAB...	very well said	0.800000011...	0.800000011...

OUTPUT EXPLANATION



Comment	Score	Magnitude	Explanation
"Ok."	0.0	0.0	No emotion
"Nice video!"	0.6	0.6	Positive, moderate
"I absolutely love this amazing masterpiece!!!"	0.9	2.3	Strong positive
"I hate this so much and it ruined my day."	-0.9	2.1	Strong negative

◆ Sentiment_score

- Range: -1 to +1
- Measures polarity
- +1 → Very positive
- 0 → Neutral
- -1 → Very negative

◆ Sentiment_magnitude

- Range: 0 to ∞
- Measures emotional intensity
- Higher value = stronger emotion

FUTURE SCOPE

- Implement an Airflow (Cloud Composer) DAG to **automate**:
  * Real-time data ingestion
  * Model training
  * Model evaluation
- Apply **performance-based model** selection:
  * Compare evaluation metrics (Accuracy, F1 score).
  * Automatically promote only the best-performing model to production.
- Replace synthetic data generation (Faker) with **real-time YouTube comment** data using YouTube Data API for realistic, production-grade inputs.

Demo.

Thank you