

What's new in BQ AI/ML

Xi Cheng, Jenny Ortiz
on behalf of the BQ AI/ML (BQML) team

Mar 13, 2026

Today's agenda



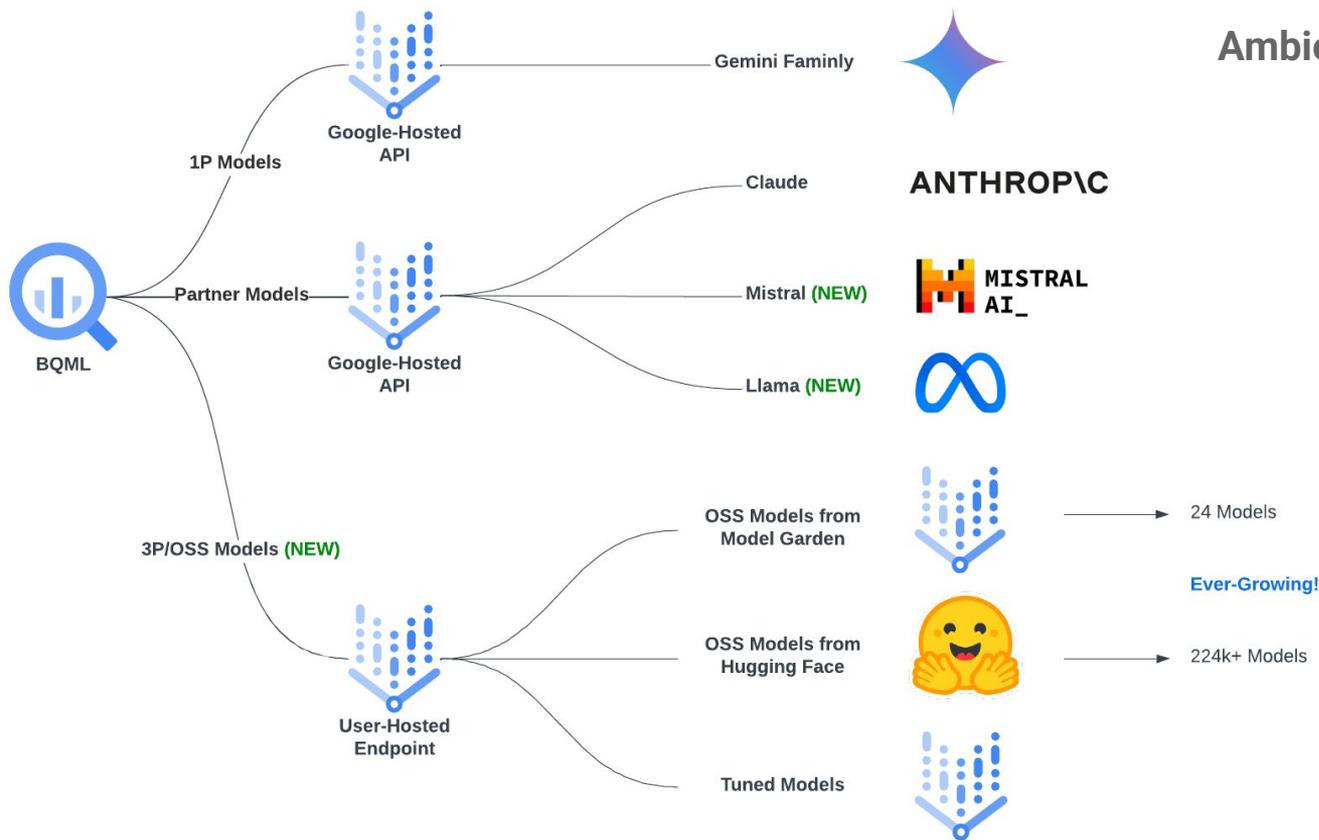
- 1 GenAI - Feature offerings
- 2 GenAI - Reliability, Scalability, Usability enhancement
- 3 GenAI meets predictive AI/ML
- 4 Agents

Today's agenda



- 1 GenAI - Feature offerings
- 2 GenAI - Reliability, Scalability, Usability enhancement
- 3 GenAI meets predictive AI/ML
- 4 Agents

LLM Inference



Ambient LLM choices

Besides the ambient LLM choices...



LLM tuning and evaluation

Supports LoRA fine tuning via CREATE MODEL syntax

Supports LLM evaluation via ML.Evaluate function.



Grounding & Safety attributes

Integrated with Google search grounding to avoid hallucination.

The Gemini model filters the responses it returns based on the safete attributes users specify

New



Table and scalar functions

```
SELECT
  movie_title,
  AI.GENERATE (
    "translate to french",
    movie_title)
FROM
  Movies_table
-----
AI.GENERATE_TEXT is the TVF
```

New



Multimodal

Work with any type and any combination of unstructured data: text, image, video, audio, document.

New



Structured Data output

```
SELECT
  AI.GENERATE("some content",
    STRUCT("age INT64,
  medications ARRAY<STRING>"
  AS output_schema))
```

AI.GENERATE_TABLE is the TVF

Extract graph [demo](#)

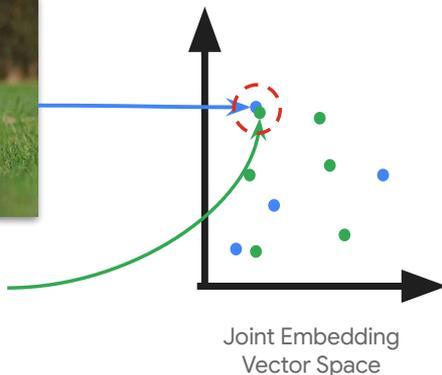
Embeddings

- **AI.embed** (scalar), **AI.Generate_embedding** (TVF)
 - Text embedding generation
 - Multimodal embedding: embed text, image and video in the same semantic space with the same dimensionality.
- **AI.similarity** (scalar)
 - Semantic distance between pair of text, pair of images, or between text and image
 - **Vector indexing and search** for more scalable solutions.

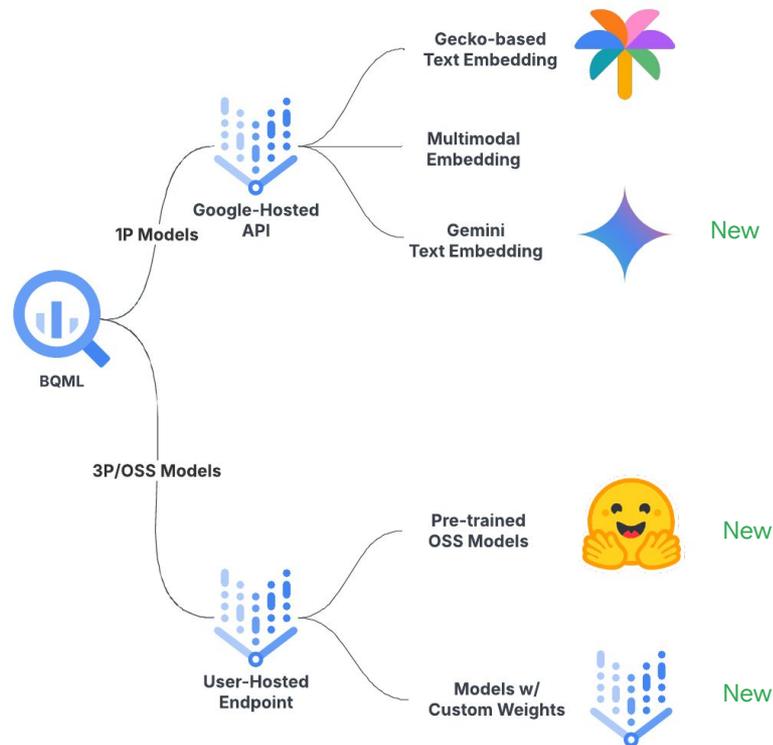


Text:

“a dog in a grass field playing fetch with a red ball”



Embedding Generation



Today's agenda



- 1 GenAI - Feature offerings
- 2 GenAI - Reliability, Scalability, Usability enhancement
- 3 GenAI meets predictive AI/ML
- 4 Agents

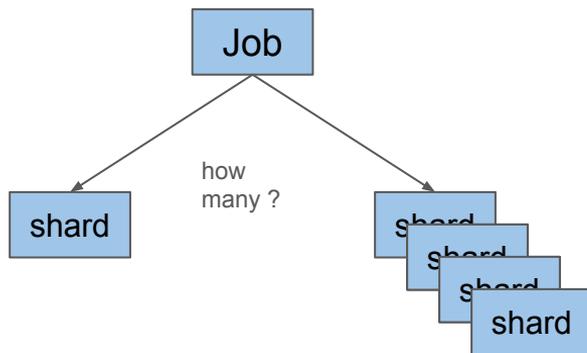
Challenge 1: resource

BQ doesn't own inference resource. We use Vertex LLM online endpoint for batch inference.

- **LLM quota capacity is limited and non-guaranteed.**
- **Quota competition:** Batch inference has lower priority than online inference

Challenge 2 - infra

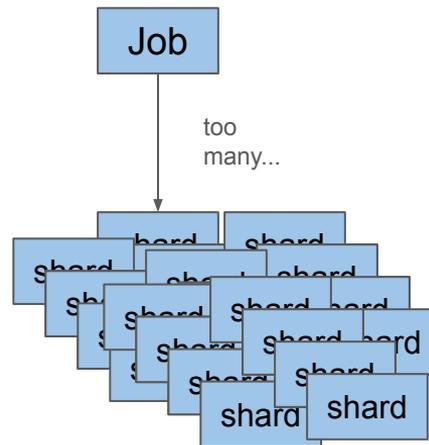
1. Non-deterministic number of active shards



Rate Control Challenge: The runtime `#shard` workers is not deterministic.

- planned `#shard` varies on different input
- runtime `#shards` might change due to slot contention, etc.

2. Too many shards hurts the coordination of total concurrency

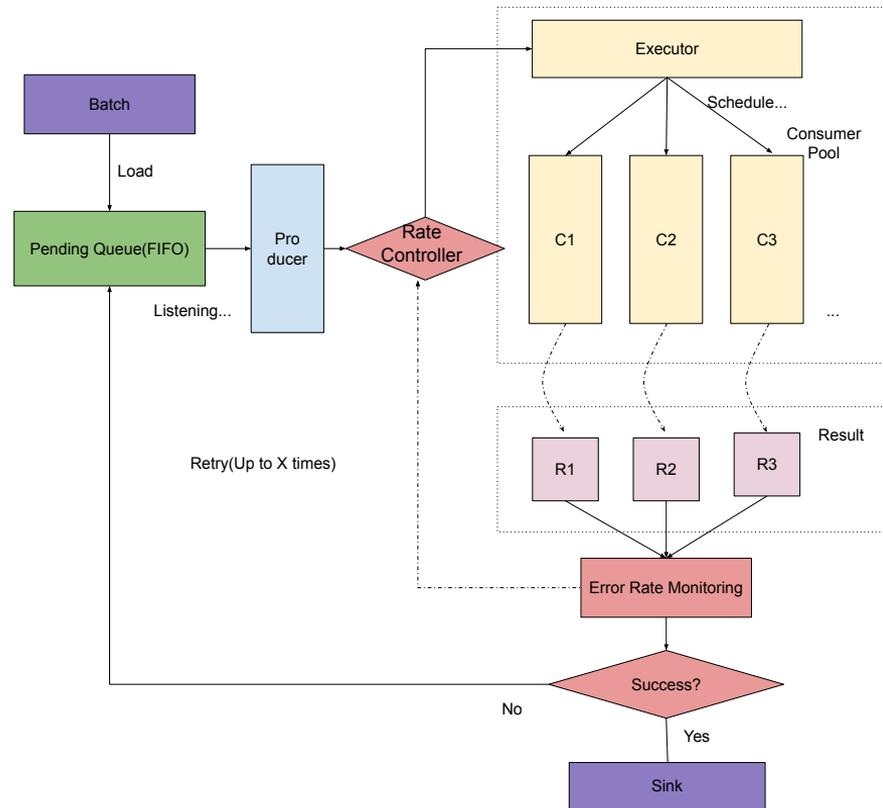


For remote inference, having too many shards (>100) doesn't help to speed up the inference, but brings challenges in coordinating concurrency among the different shards.

Within-Shard Concurrency Control

A universal architecture optimized for client-side batch processing with appropriate error handling.

- **Scalability**
 - Multithreading
 - Precise & deterministic rate control
- **Reliability**
 - Retry failed tasks at end of the queue
 - Distributing the risk evenly among all tasks
- **Adaptive Rate Control**
 - Feedback the historical success rate to the rate controller

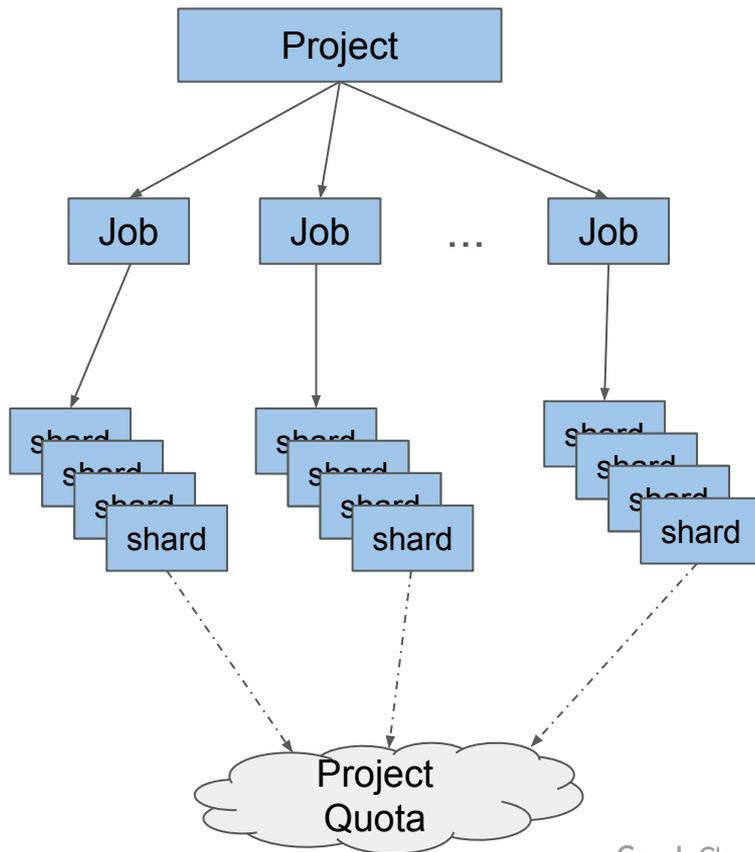


Enforce User Quota

Each shard worker is unaware of the other shards.

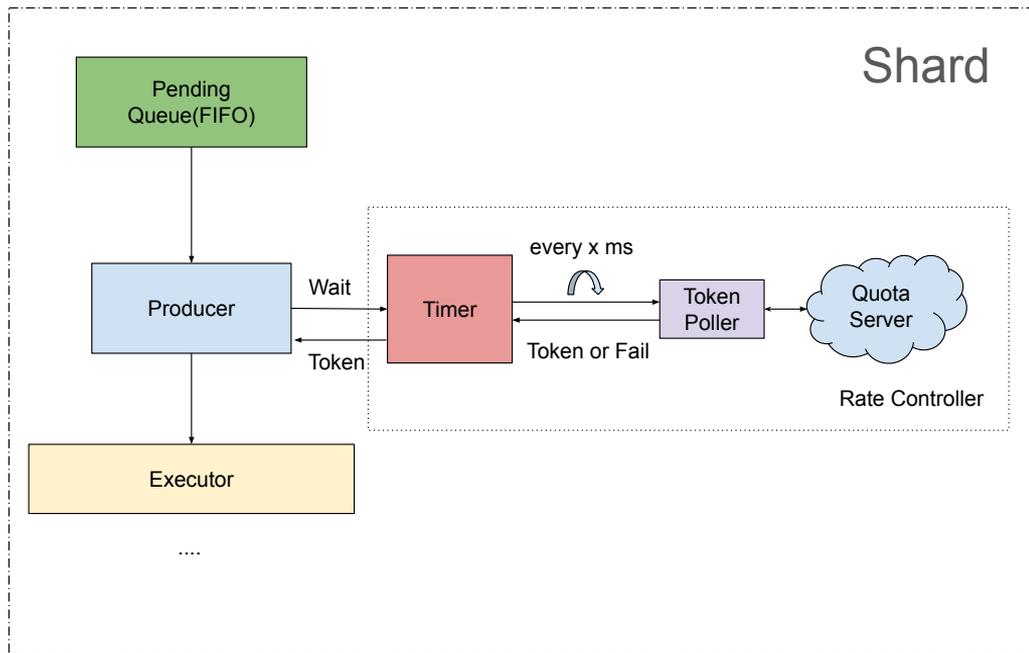
...

We need a way to enforce project-wide user quota across all shard workers.



Enforce User Quota

- Centralized quota server manages the project quota.
 - All shards & all query jobs within the same project use the same quota key.
- Token poller at each shard worker that polls the quota periodically.
 - Analogy: Ticket bots refresh & collect the next available tickets.



Reliability enhancement

- 99.99% generative AI queries can finish without a single row failure.
- A row success rate of over 99.99% across all query jobs.
- [SQL script](#) or the [Dataform package](#), allowing you to iteratively retry failed rows to get to a 100% row success rate.

Scalability enhancement

- 1P Gemini Models
 - Dynamic shared quota (DSQ) for PayGo Users: highly non-guaranteed quota.
 - [100x+ scalability gain] Gemini-2.5-flash for 500 input and 50 output tokens -> 9.3 million rows/(6h-) job
- 1P Embedding Models
 - Token-based batching to pack as many rows as possible into a single request, subject to the token limit.
 - 30x+ scalability gain
 - 1500 QPM (default quota) and 50 tokens/row -> 80 million rows / (6h-) job
 - 10000 QPM (request without manually approval) and 50 tokens/row -> 500 million rows/(6h-) job

Usability enhancement

- Permission setup: End user credential (EUC) support, instead of using more complicated BigQuery connections.
- Gemini Global endpoint support for higher availability across the world.
- Automatic quota sync from Vertex AI for endpoints with fixed per-project quota.
- Managed 3P/OSS GenAI Inference - users no longer need to deploy the LLM themselves

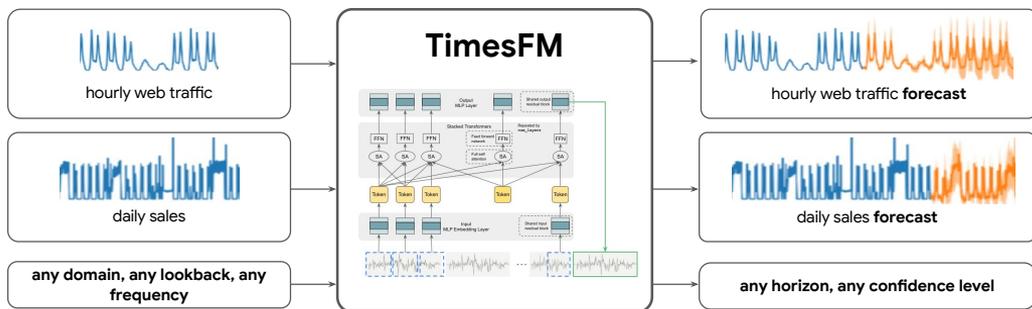
Today's agenda



- 1 GenAI - feature offerings
- 2 GenAI - Reliability, Scalability, Usability enhancement
- 3 GenAI meets predictive AI/ML
- 4 Agents

TimesFM in BigQuery

State of the art forecasting and anomaly detection at your analysts fingertips

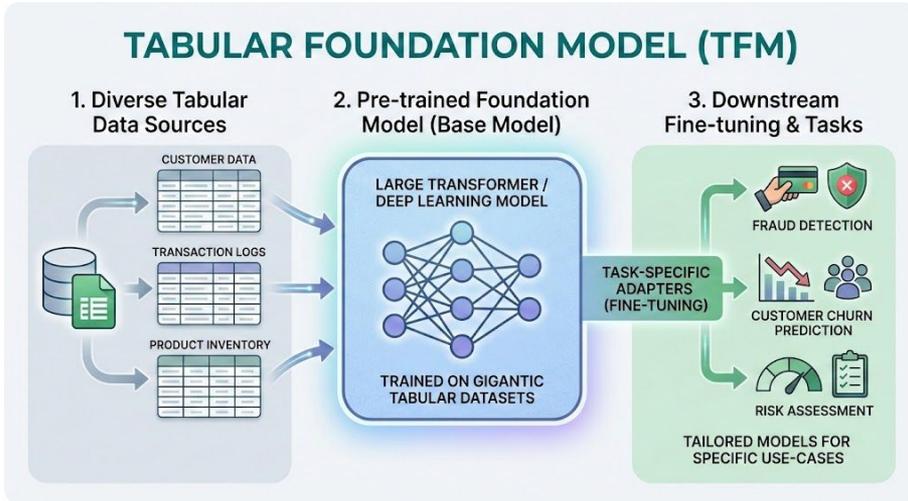


- **Easy to use:** Zero shot forecasting / anomaly detection inference; just provide an input time series and forecast
- **Integrated into BigQuery:** No model to train, no infra to manage
- **Accurate:** Gains on standard time series benchmarks vs statistical, deep learning and other foundational models.
- **Fast and scalable:** Forecast and detect anomalies in millions of time series in minutes

```
AI.FORECAST(  
  (SELECT id1, id2, date, data  
FROM table),  
  horizon => 30,  
  id_cols => ["id1", "id2"],  
  data_col => "data",  
  timestamp_col => "date")
```

```
AI.Detect_anomalies(  
  (SELECT id1, id2, date, data FROM  
table),  
  cutoff_timestamp => '2026-01-01',  
  id_cols => ["id1", "id2"],  
  data_col => "data",  
  timestamp_col => "date")
```

TabularFM in BigQuery



```
SELECT *
FROM AI.PREDICT (
  training_data => (TABLE my_labeled_transactions),
  prediction_data => (TABLE my_new_transactions),
  model_name => 'tabularFM',
  label_col => 'is_fraud'
)
```

Today's agenda



- 1 GenAI - feature offerings
- 2 GenAI - Reliability, Scalability, Usability enhancement
- 3 GenAI meets predictive AI/ML
- 4 Agents

BQ AI/ML in Conversational Analytics

- Conversational analytics in **BigQuery** lets you chat with agents about your data using natural language.
 - Create agents over the data
 - Ask questions about specific relations in BQ
- BQ AI/ML functions supported
 - AI.FORECAST
 - AI.DETECT_ANOMALIES
 - AI.GENERATE
 - ...
 - More to come

What is the schema for this dataset?

Retrieved context for 1 table.

Analyzing Forecasting Needs

I've determined the user wants a forecast for bike trips in April 2026. My next step is to prepare for the BigQuery AI Expert Agent. Before that, I've checked the schema, noting `start_time` as the time series, and the need to count trips. Now I'm ready to proceed with `QUERY_GENERATION` and anticipate instructions for `AI.FORECAST`.

Assessing Data Range and Limits

I've shifted my focus to the data itself. My current plan is to determine the date range of the `bikeshare_trips` data and count the number of data points. This is crucial for verifying that the total historical data plus the forecast horizon does not exceed 1000 rows. This step will also help me understand the recency of the data, which may affect the forecast horizon I use.

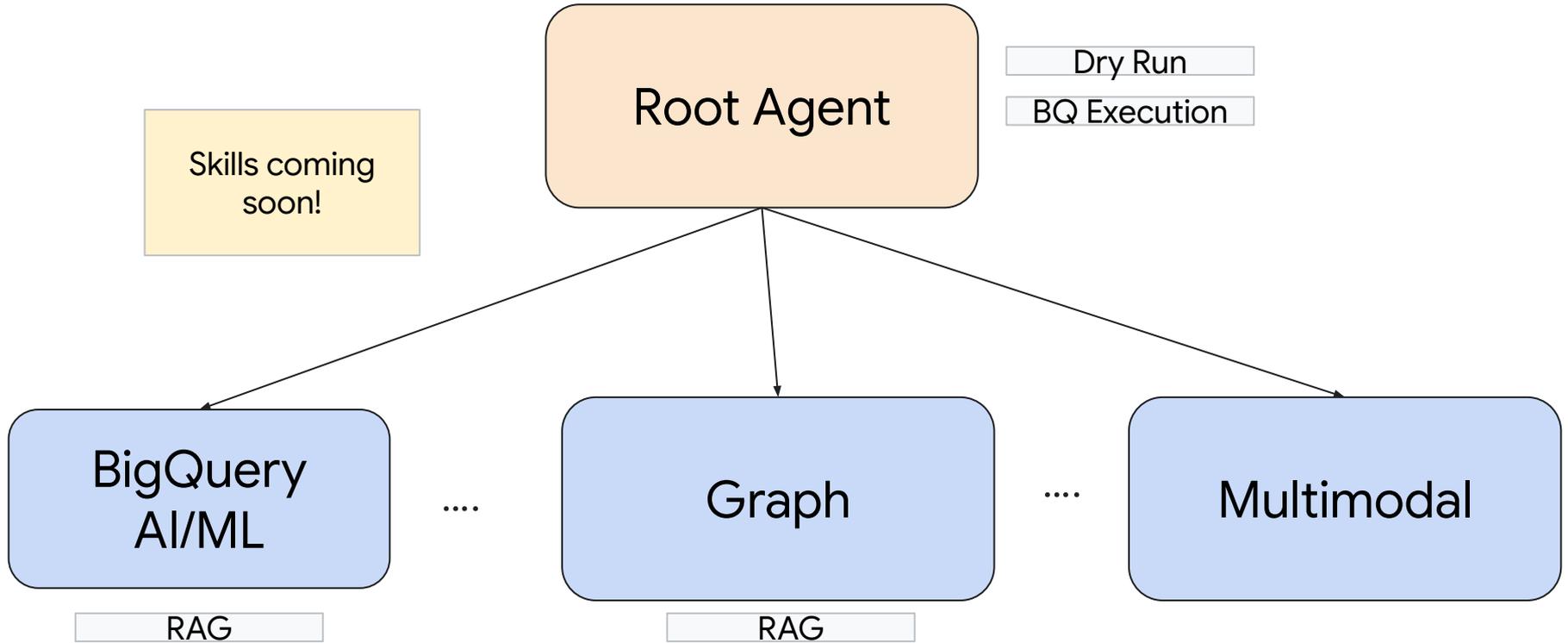
Analyzing Forecasting Approach

I've determined the optimal approach involves forecasting daily trips for the next 30 days, starting from July 1, 2024. To avoid plotting issues, I'll restrict historical data to the last 900 days. Now, I'm generating the SQL query using `AI.FORECAST`, constructing a CTE to aggregate daily counts, filtering to the relevant historical timeframe, and applying the forecast function. I've confirmed that `AI.FORECAST` is the correct function and am now reviewing the SQL structure for completeness.

Analyzing Forecasting Approach ...

Ask a question

BQ AI/ML in Conversational Analytics



Thank you!

chengx@google.com
jeortiz@google.com