Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 163 DISCUSSI…

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 163
Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a retail company. You have built a model that predicts a coupon to offer an ecommerce customer at checkout based on the items in their cart. When a customer goes to checkout, your serving pipeline, which is hosted on Google Cloud, joins the customer's existing cart with a row in a BigQuery table that contains the customers' historic purchase behavior and uses that as the model's input. The web team is reporting that your model is returning predictions too slowly to load the coupon offer with the rest of the web page. How should you speed up your model's predictions?

A. Attach an NVIDIA P100 GPU to your deployed model's instance.

B. Use a low latency database for the customers' historic purchase behavior.

C. Deploy your model to more instances behind a load balancer to distribute traffic.

D. Create a materialized view in BigQuery with the necessary data for predictions.

Show Suggested Answer

by kalle_balle at *Jan. 7, 2024, 3:50 a.m.*

Comments

Type your comment...

Submit

inc_dev_ml_001 2 months, 2 weeks ago

Selected Answer: B

It says that you have to join the cart data, so you can't use the materialized view because it means that you should materialize the view every time a new cart shows up. So use a low latency DB it's the only way

upvoted 1 times

inc_dev_ml_001 4 months ago

Selected Answer: B

In my opinion the materialized view could be the best way but it says that the cart data have to join with historic behaviour so it's impossibile to have all the needed data for the prediction in the materialized view because cart data are not in the database.

upvoted 1 times

SausageMuffins 5 months, 1 week ago

Selected Answer: D

Both B and D in theory does reduce latency but B implies that we might need to migrate the database to another low latency database. This migration and setup might incur additional costs and effort.

In contrast, creating a materialized view seems much more straight forward since there is already a preexisting big query table mentioned in the question.

upvoted 1 times

Ria_1989 5 months, 1 week ago

Coupon to offer an ecommerce customer at checkout based on the items in their cart not the customer historic behaviour. That's creating confusion while choosing B.

upvoted 1 times

fitri001 6 months ago

Selected Answer: D

Reduced Join Cost: Joining the customer's cart with their purchase history in BigQuery during each prediction can be slow. A materialized view pre-computes and stores the join results, eliminating the need for repetitive joins and significantly reducing latency.
Targeted Data Access: Materialized views allow you to specify the exact columns needed for prediction, minimizing data transferred between BigQuery and your serving pipeline.

upvoted 2 times

> **pinimichele01** 6 months ago
>
> https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction
>
> i'm not sure that bq is the best option, what do you think?
>
> upvoted 1 times

gscharly 6 months, 1 week ago

Selected Answer: B

https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction

"Analytical data stores such as BigQuery are not engineered for low-latency singleton read operations, where the result is a single row with many columns."

upvoted 3 times

guilhermebutzke 8 months, 1 week ago

Selected Answer: B

I changed my mind.

B: Im read a lot this page

https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction

If the web team is reporting that the model is returning predictions too slowly to load the coupon offer with the rest of the web page, it suggests that the bottleneck might indeed be in the inference process rather than in data retrieval or processing. Given that the model is deployed on Google Cloud, choosing a low-latency database makes it suitable for scenarios where quick access to data is crucial, such as real-time predictions for web applications.

Option D: While pre-aggregating data in BigQuery can improve query speed, it might not be as efficient as a low-latency database for frequently accessed data like customer purchase history.

upvoted 3 times

guilhermebutzke 8 months, 2 weeks ago

Selected Answer: D

Firstly, I believe the correct choice should be B. This is supported by a comprehensive Google page discussing methods to minimize real-time prediction latency. In this resource, they don't mention using a BigQuery view but instead suggest precomputing and lookup approaches to minimize prediction time.

https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction

However, I will stick with option D because it's not clear whether option B suggests changing the entire database or just utilizing it as a preliminary step for online prediction.

upvoted 1 times

guilhermebutzke 8 months, 1 week ago

I change for B

upvoted 1 times

sonicclasps 8 months, 3 weeks ago

Selected Answer: D

Queries that use materialized views are generally faster and consume fewer resources than queries that retrieve the same data only from the base tables. Materialized views can significantly improve the performance of workloads that have the characteristic of common and repeated queries.

upvoted 2 times

ddogg 8 months, 3 weeks ago

Selected Answer: D

D. Create a materialized view in BigQuery with the necessary data for predictions.

Here's why:

Current bottleneck: Joining the cart data with the BigQuery table containing historic purchases likely creates the latency bottleneck. Fetching data from BigQuery on every prediction request can be slow.
Materialized view: A materialized view pre-computes and stores the join between the cart data and the relevant historic purchase information in BigQuery. This eliminates the need for real-time joins during prediction, significantly reducing latency.
Faster access: The pre-computed data in the materialized view is readily available within BigQuery, ensuring faster access for your serving pipeline when predicting the coupon offer.
Lower cost: Compared to additional instances or GPU resources, a materialized view can be a more cost-effective solution, especially if prediction requests are frequent.

upvoted 3 times

kalle_balle 9 months, 2 weeks ago

Selected Answer: B

Option B seems most sensible.

upvoted 1 times

Social Media

Facebook , Twitter
YouTube , Reddit
Pinterest

We are the biggest and most updated IT certification exam material website.

Using our own resources, we strive to strengthen the IT professionals community for free.

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.