Google Discussions



## EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 171 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 171
Topic #: 1
[All Professional Data Engineer Questions]

You work for a large real estate firm and are preparing 6 TB of home sales data to be used for machine learning. You will use SQL to transform the data and use
BigQuery ML to create a machine learning model. You plan to use the model for predictions against a raw dataset that has not been transformed. How should you set up your workflow in order to prevent skew at prediction time?

A. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

B. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. Before requesting predictions, use a saved query to transform your raw input data, and then use ML.EVALUATE.

C. Use a BigQuery view to define your preprocessing logic. When creating your model, use the view as your model training data. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

D. Preprocess all data using Dataflow. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any further transformations on the input data.

Show Suggested Answer

by AWSandeep at *Sept. 2, 2022, 7:44 p.m.*

### Comments

Type your comment...

Submit

**AWSandeep** Highly Voted 2 years, 8 months ago

Selected Answer: A

A. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

Using the TRANSFORM clause, you can specify all preprocessing during model creation. The preprocessing is automatically applied during the prediction and evaluation phases of machine learning.

Reference: https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform

upvoted 15 times

**zellck** Highly Voted 2 years, 5 months ago

Selected Answer: A

A is the answer.

https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform
Using the TRANSFORM clause, you can specify all preprocessing during model creation. The preprocessing is automatically applied during the prediction and evaluation phases of machine learning

upvoted 6 times

**f74ca0c** Most Recent 4 months ago

Selected Answer: C

C. Use a BigQuery view to define your preprocessing logic. When creating your model, use the view as your model training data. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

Explanation:
Preventing Data Skew:

Training-serving skew occurs when the transformations applied to training data are not identically applied to prediction data. Using a BigQuery view ensures consistent preprocessing for both training and prediction.
Advantages of BigQuery Views:

Views encapsulate preprocessing logic, ensuring that the same transformations are applied whenever the view is queried. By referencing the view during both training and prediction, you eliminate the need for manual transformations and the risk of discrepancies.

upvoted 1 times

**SamuelTsch** 6 months, 1 week ago

Selected Answer: A

A

upvoted 1 times

**Lenifia** 10 months ago

Selected Answer: B

The key to preventing skew in machine learning models is to ensure that the same data preprocessing steps are applied consistently to both the training data and the prediction data. In option B, the TRANSFORM clause in BigQuery ML is used to define preprocessing steps during model creation, and a saved query is used to apply the same transformations to the raw input data before making predictions. This ensures consistency and prevents skew. The ML.EVALUATE function is then used to evaluate the model's performance on the transformed prediction data. This is the recommended workflow

upvoted 2 times

**Matt_108** 1 year, 3 months ago

Selected Answer: A

Option A

upvoted 1 times

**Prudvi3266** 2 years ago

Selected Answer: A

A is correct answer if we use TRANSFORM clause in BigQuery no need to use any transform while evaluating and predicting https://cloud.google.com/bigquery/docs/bigqueryml-transform

upvoted 3 times

**Kvk117** 2 years, 3 months ago

Selected Answer: A

A is the correct answer

A is the correct answer

upvoted 2 times

**jkhong** 2 years, 4 months ago

Selected Answer: A

Problem: Skew

One thing that I overlooked when answering previously is that B, C does not address skew. When we preprocess our training data, we need to save our scaled factors somewhere, and when performing predictions on our test data, we need to use the scaling factors of our training data to predict the results.

ML.EVALUATE already incorporates preprocessing steps for our test data using the saved scaled factors.

upvoted 3 times

**GCPSharon** 2 years, 6 months ago

Selected Answer: C

Stew prediction time by remove the preprocessing!

upvoted 1 times

**TNT87** 2 years, 7 months ago

Selected Answer: A

https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform
Ans A

upvoted 4 times

**ducc** 2 years, 8 months ago

Selected Answer: A

This query's nested SELECT statement and FROM clause are the same as those in the CREATE MODEL query. Because the TRANSFORM clause is used in training, you don't need to specify the specific columns and transformations. They are automatically restored.

Reference: https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform

upvoted 2 times