Welcome to ExamTopics

EXAMTOPICS

- Expert Verified, Online, **Free**.

MENU

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 33 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 33
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have a demand forecasting pipeline in production that uses Dataflow to preprocess raw data prior to model training and prediction. During preprocessing, you employ Z-score normalization on data stored in BigQuery and write it back to BigQuery. New training data is added every week. You want to make the process more efficient by minimizing computation time and manual intervention. What should you do?

A. Normalize the data using Google Kubernetes Engine.

B. Translate the normalization algorithm into SQL for use with BigQuery.

C. Use the normalizer_fn argument in TensorFlow's Feature Column API.

D. Normalize the data with Apache Spark using the Dataproc connector for BigQuery.

Show Suggested Answer

by maartenalexander at *June 22, 2021, 12:27 p.m.*

Comments

Type your comment...

Submit

maartenalexander Highly Voted 3 years, 4 months ago

B. I think. BiqQuery definitely minimizes computational time for normalization. I think it would also minimize manual intervention. For data normalization in dataflow you'd have to pass in values of mean and standard deviation as a side-input. That seems more work than a simple SQL query

upvoted 21 times

93alejandrosanchez 3 years ago

I agree that B would definitely get the job done. But wouldn't D work as well and keep all the data pre-processing in Dataflow?

upvoted 2 times

kaike_reis 2 years, 11 months ago

Dataflow uses Beam, different from dataproc that uses Spark.

I think that D would be wrong because we would add one more service into the pipeline for a simple transformation (minus the mean and divide by std).

upvoted 4 times

PhilipKoku Most Recent 5 months ago

Selected Answer: B

B) Using BigQuery

upvoted 1 times

Sum_Sum 11 months, 3 weeks ago

Selected Answer: B

z-scores is very easy to do in BQ - no need for more complex solutions

upvoted 1 times

elenamatay 1 year, 1 month ago

B. All that maartenalexander said, + BigQuery already has a function for that: https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-standard-scaler , we could even schedule the query for calculating this automatically :)

upvoted 2 times

aaggii 1 year, 3 months ago

Selected Answer: C

Every week when new data is loaded mean and standard deviation is calculated for it and passed as parameter to calculate z score at serving
https://towardsdatascience.com/how-to-normalize-features-in-tensorflow-5b7b0e3a4177

upvoted 1 times

tavva_prudhvi 1 year, 3 months ago

owever, in the given scenario, you are using Dataflow for preprocessing and BigQuery for storing data.

To make the process more efficient by minimizing computation time and manual intervention, you should still opt for option B: Translate the normalization algorithm into SQL for use with BigQuery. This way, you can perform the normalization directly in BigQuery, which will save time and resources compared to using an external tool.

upvoted 1 times

SamuelTsch 1 year, 4 months ago

Selected Answer: B

A, D usually need additional configuration, which could cost much more time.

upvoted 1 times

M25 1 year, 6 months ago

Selected Answer: B

Went with B

upvoted 2 times

SergioRubiano 1 year, 7 months ago

Selected Answer: B

Best way is B

upvoted 2 times

Fatiy 1 year, 8 months ago

Selected Answer: D

Option D is the best solution because Apache Spark provides a distributed computing platform that can handle large-scale data processing with ease. By using the Dataproc connector for BigQuery, Spark can read data directly from BigQuery and perform the normalization process in a distributed manner. This can significantly reduce computation time and manual intervention. Option A is not a good solution because Kubernetes is a container orchestration platform that does not directly provide data normalization capabilities. Option B is not a good solution because Z-score normalization is a data transformation technique that cannot be easily translated into SQL. Option C is not a good solution because the normalizer_fn argument in TensorFlow's Feature Column API is only applicable for feature normalization during model training, not for data preprocessing.

upvoted 2 times

ares81 1 year, 9 months ago

Selected Answer: B

Best way to proceed is B.

upvoted 2 times

> **Fatiy** 1 year, 8 months ago
>
> SQL is not as flexible as other programming languages like Python, which can limit the ability to customize the normalization process or incorporate new features in the future.
>
> upvoted 1 times

Mohamed_Mossad 2 years, 4 months ago

Selected Answer: B

B is the most efficient as you will not load --> process --> save , no you will only write some sql in bigquery and voila :D

upvoted 4 times

baimus 2 years, 7 months ago

It's B, bigquery can do this internally, no need for dataflow

upvoted 2 times

> **Fatiy** 1 year, 8 months ago
>
> SQL is not as flexible as other programming languages like Python, which can limit the ability to customize the normalization process or incorporate new features in the future.
>
> upvoted 1 times

xiaoF 2 years, 9 months ago

Selected Answer: B

I agree with B.

upvoted 2 times

alashin 3 years, 4 months ago

B. I agree with B as well.

upvoted 3 times

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.