Google Discussions

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 83 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 83

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to design an architecture that serves asynchronous predictions to determine whether a particular mission-critical machine part will fail. Your system collects data from multiple sensors from the machine. You want to build a model that will predict a failure in the next N minutes, given the average of each sensor's data from the past 12 hours. How should you design the architecture?

A. 1. HTTP requests are sent by the sensors to your ML model, which is deployed as a microservice and exposes a REST API for prediction
2. Your application queries a Vertex AI endpoint where you deployed your model.
3. Responses are received by the caller application as soon as the model produces the prediction.

B. 1. Events are sent by the sensors to Pub/Sub, consumed in real time, and processed by a Dataflow stream processing pipeline.
2. The pipeline invokes the model for prediction and sends the predictions to another Pub/Sub topic.
3. Pub/Sub messages containing predictions are then consumed by a downstream system for monitoring.

C. 1. Export your data to Cloud Storage using Dataflow.
2. Submit a Vertex AI batch prediction job that uses your trained model in Cloud Storage to perform scoring on the preprocessed data.
3. Export the batch prediction job outputs from Cloud Storage and import them into Cloud SQL.

D. 1. Export the data to Cloud Storage using the BigQuery command-line tool
2. Submit a Vertex AI batch prediction job that uses your trained model in Cloud Storage to perform scoring on the preprocessed data.
3. Export the batch prediction job outputs from Cloud Storage and import them into BigQuery

3. Export the batch prediction job outputs from Cloud Storage and import them into BigQuery.

Show Suggested Answer

by LearnSodas at *Dec. 11, 2022, 6:57 p.m.*

Comments

Type your comment...

Submit

PhilipKoku 4 months, 2 weeks ago

Selected Answer: B

B) Pub/Sub & DataFlow

upvoted 1 times

inc_dev_ml_001 6 months ago

Selected Answer: C

The simplest solution that can support an eventual batch prediction (triggered by pub/sub) even the semi-real time prediction.

upvoted 1 times

Werner123 7 months, 3 weeks ago

Selected Answer: B

Needs to be real time not batch. The data needs to be processed as a stream since multiple sensors are used. pawan94 is right. https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction

upvoted 1 times

pawan94 9 months, 2 weeks ago

Here you go to the answer provided by google itself. I don't understand why would people use batch prediction when they its sensor data and online prediction is as well asynchronous.
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#offline_batch_prediction:~:text=Predictive%20maintenance%3A%20asynchronously%20predicting%20whether%20a%20particular%20machine%20part%20will%20fail%20in%20the%20next%20N%20minutes%2C%20given%20the%20averages%20of%20the%20sensor%27s%20data%20in%20the%20past%2030%20minutes.

upvoted 2 times

vale_76_na_xxx 10 months, 1 week ago

it refers to asincronou prediction I' go with C

upvoted 1 times

rosenr0 1 year, 4 months ago

Selected Answer: D

D.
I think we have to query data from the past 12 hours for the prediction, and that's the reason for exporting the data to Cloud Storage.
Also, the predictions don't have to be real time.

upvoted 2 times

M25 1 year, 5 months ago

Selected Answer: B

Went with B

upvoted 1 times

JamesDoe 1 year, 6 months ago

Selected Answer: B

B.
Online prediction, and need decoupling with Pub/Sub to make it asynchronous. Option A is synchronous.

upvoted 2 times

tavva_prudhvi 1 year, 7 months ago

Option C may not be the best choice for this use case because it involves using a batch prediction job in Vertex AI to perform scoring on preprocessed data. Batch prediction jobs are more suitable for scenarios where data is processed in batches, and results can be generated over a longer period, such as daily or weekly.

results can be generated over a longer period, such as daily or weekly.

In this use case, the requirement is to predict whether a machine part will fail in the next N minutes, given the average of each sensor's data from the past 12 hours. Therefore, real-time processing and prediction are necessary. Batch prediction jobs are not designed for real-time processing, and there may be a delay in receiving the predictions.

Option B, on the other hand, is designed for real-time processing and prediction. The Pub/Sub and Dataflow components allow for real-time processing of incoming sensor data, and the trained ML model can be invoked for prediction in real-time. This makes it ideal for mission-critical applications where timely predictions are essential.

upvoted 2 times

tavva_prudhvi 1 year, 7 months ago

Its B, This architecture leverages the strengths of Pub/Sub, Dataflow, and Vertex AI. The system collects data from multiple sensors, which sends events to Pub/Sub. Pub/Sub can handle the high volume of incoming data and can buffer messages to prevent data loss. A Dataflow stream processing pipeline can consume the events in real-time and perform feature engineering and data preprocessing before invoking the trained ML model for prediction. The predictions are then sent to another Pub/Sub topic, where they can be consumed by a downstream system for monitoring.

This architecture is highly scalable, resilient, and efficient, as it can handle large volumes of data and perform real-time processing and prediction. It also separates concerns by using a separate pipeline for data processing and another for prediction, making it easier to maintain and modify the system.

upvoted 1 times

enghabeth 1 year, 8 months ago

Selected Answer: B

if you have sensors inyour architecture.. you need pub/sub...

upvoted 1 times

John_Pongthorn 1 year, 8 months ago

Selected Answer: B

B is most likely . if you search asynchronous on this page. it appears in
the question wants to focus on online prediction with asynchronous mode.
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#online_real-time_prediction
and the question is the same as what has been explained in this section obviously. it is as below.
Predictive maintenance: asynchronously predicting whether a particular machine part will fail in the next N minutes, given the averages of the sensor's data in the past 30 minutes.

afte that, you can take a closer look at figure3 and read what it try to describe

C and D it is the offline solution but you opt to use different tools.
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#offline_batch_prediction

upvoted 2 times

John_Pongthorn 1 year, 8 months ago

Asycnchromoue preciction = Batch prediction
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#offline_batch_prediction

upvoted 1 times

John_Pongthorn 1 year, 8 months ago

Asynchronous prediction = Batch prediction, It is incorrect because I am reckless to read this article, Admin can delete my shitty comment above. I was mistaken

upvoted 1 times

hiromi 1 year, 10 months ago

Selected Answer: B

B
"Predictive maintenance: asynchronously predicting whether a particular machine part will fail in the next N minutes, given the averages of the sensor's data in the past 30 minutes."
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#offline_batch_prediction

upvoted 3 times

hiromi 1 year, 10 months ago

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

upvoted 1 times

mil_spyro 1 year, 10 months ago

Selected Answer: B

Answer is B.
https://cloud.google.com/architecture/minimizing-predictive-serving-latency-in-machine-learning#handling_dynamic_real-

time_features

upvoted 1 times

ares81 1 year, 10 months ago

Selected Answer: C

C, for me.

upvoted 1 times

seifou 1 year, 10 months ago

Selected Answer: C

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

upvoted 1 times

Load full discussion...

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.