EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 231 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 231
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are using Keras and TensorFlow to develop a fraud detection model. Records of customer transactions are stored in a large table in BigQuery. You need to preprocess these records in a cost-effective and efficient way before you use them to train the model. The trained model will be used to perform batch inference in BigQuery. How should you implement the preprocessing workflow?

A. Implement a preprocessing pipeline by using Apache Spark, and run the pipeline on Dataproc. Save the preprocessed data as CSV files in a Cloud Storage bucket.

B. Load the data into a pandas DataFrame. Implement the preprocessing steps using pandas transformations, and train the model directly on the DataFrame.

C. Perform preprocessing in BigQuery by using SQL. Use the BigQueryClient in TensorFlow to read the data directly from BigQuery.

D. Implement a preprocessing pipeline by using Apache Beam, and run the pipeline on Dataflow. Save the preprocessed data as CSV files in a Cloud Storage bucket.

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 7:57 a.m.*

Comments

Type your comment...

Submit

b1a8fae Highly Voted 9 months, 3 weeks ago

Selected Answer: C

Easiest to preprocess the data on BigQuery.

upvoted 5 times

pinimichele01 Most Recent 7 months ago

Selected Answer: C

went with C

upvoted 2 times

pinimichele01 6 months, 1 week ago

Easiest to preprocess the data on BigQuery.

upvoted 2 times

pikachu007 9 months, 3 weeks ago

Selected Answer: C

A. Spark on Dataproc: While powerful, it incurs additional cluster setup and management costs, potentially less cost-effective for this specific use case.
B. pandas DataFrame: Loading large datasets into memory might lead to resource constraints and performance issues, especially for large-scale preprocessing.
D. Apache Beam on Dataflow: While scalable, it introduces extra complexity for managing a separate pipeline and storage for preprocessed data.

upvoted 2 times

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.