EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 45 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 45
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are training a TensorFlow model on a structured dataset with 100 billion records stored in several CSV files. You need to improve the input/output execution performance. What should you do?

A. Load the data into BigQuery, and read the data from BigQuery.

B. Load the data into Cloud Bigtable, and read the data from Bigtable.

C. Convert the CSV files into shards of TFRecords, and store the data in Cloud Storage.

D. Convert the CSV files into shards of TFRecords, and store the data in the Hadoop Distributed File System (HDFS).

Show Suggested Answer

by burnout at *July 3, 2021, 2:15 p.m.*

Comments

Type your comment...

Submit

ralf_cc Highly Voted 3 years, 4 months ago
C - not enough info in the question, but C is the "most correct" one

upvoted 25 times

PhilipKoku Most Recent 5 months ago

Selected Answer: C

C) The most suitable option for improving input/output execution performance in this scenario is C. Convert the CSV files into shards of TFRecords and store the data in Cloud Storage. This approach leverages the efficiency of TFRecords and the scalability of Cloud Storage, aligning with TensorFlow best practices.

upvoted 1 times

fragkris 11 months ago

Selected Answer: C

C is the google reccomended approach.

upvoted 1 times

Sum_Sum 11 months, 3 weeks ago

C is the correct one as BQ will not help you with performance

upvoted 1 times

peetTech 1 year, 1 month ago

Selected Answer: C

C https://datascience.stackexchange.com/questions/16318/what-is-the-benefit-of-splitting-tfrecord-file-into-shards#:~:text=Splitting%20TFRecord%20files%20into%20shards,them%20through%20a%20training%20process.

upvoted 2 times

peetTech 1 year, 1 month ago

C https://datascience.stackexchange.com/questions/16318/what-is-the-benefit-of-splitting-tfrecord-file-into-shards#:~:text=Splitting%20TFRecord%20files%20into%20shards,them%20through%20a%20training%20process.

upvoted 1 times

ftl 1 year, 1 month ago

bard: The correct answer is:

C. Convert the CSV files into shards of TFRecords, and store the data in Cloud Storage.
TFRecords is a TensorFlow-specific binary format that is optimized for performance. Converting the CSV files into TFRecords will improve the input/output execution performance. Sharding the TFRecords will allow the data to be read in parallel, which will further improve performance.

The other options are not as likely to improve performance.

Loading the data into BigQuery or Cloud Bigtable will add an additional layer of abstraction, which can slow down performance.
Storing the TFRecords in HDFS is not likely to improve performance, as HDFS is not optimized for TensorFlow.

upvoted 1 times

tavva_prudhvi 1 year, 2 months ago

Using BigQuery or Bigtable may not be the most efficient option for input/output operations with TensorFlow. Storing the data in HDFS may be an option, but Cloud Storage is generally a more scalable and cost-effective solution.

upvoted 1 times

PST21 1 year, 5 months ago

While Bigtable can offer high-performance I/O capabilities, it is important to note that it is primarily designed for structured data storage and real-time access patterns. In this scenario, the focus is on optimizing input/output execution performance, and using TFRecords in Cloud Storage aligns well with that goal.

upvoted 1 times

Voyager2 1 year, 5 months ago

Selected Answer: A

A. Load the data into BigQuery, and read the data from BigQuery.
https://cloud.google.com/blog/products/ai-machine-learning/tensorflow-enterprise-makes-accessing-data-on-google-cloud-faster-and-easier
Precisely on this link provided in other comments it whos that the best shot with tfrecords is: 18752 Records per second. In the same report it shows that bigquery is morethan 40000 recors per second

upvoted 2 times

tavva_prudhvi 1 year, 3 months ago

BigQuery is designed for running large-scale analytical queries, not for serving input pipelines for machine learning models like TensorFlow. BigQuery's strength is in its ability to handle complex queries over vast amounts of data, but it may not provide the optimal performance for the specific task of feeding data into a TensorFlow model.

On the other hand, converting the CSV files into shards of TFRecords and storing them in Cloud Storage (Option C) will provide better performance because TFRecords is a format designed specifically for TensorFlow. It allows for efficient

storage and retrieval of data, making it a more suitable choice for improving the input/output execution performance. Additionally, Cloud Storage provides high throughput and low-latency data access, which is beneficial for training large-scale TensorFlow models.

upvoted 3 times

M25 1 year, 6 months ago

Selected Answer: C

Went with C

upvoted 2 times

shankalman717 1 year, 8 months ago

Selected Answer: C

Cloud Bigtable is typically used to process unstructured data, such as time-series data, logs, or other types of data that do not conform to a fixed schema. However, Cloud Bigtable can also be used to store structured data if necessary, such as in the case of a key-value store or a database that does not require complex relational queries.

upvoted 1 times

shankalman717 1 year, 8 months ago

Selected Answer: C

Option C, converting the CSV files into shards of TFRecords and storing the data in Cloud Storage, is the most appropriate solution for improving input/output execution performance in this scenario

upvoted 1 times

behzadsw 1 year, 10 months ago

Selected Answer: A

https://cloud.google.com/architecture/ml-on-gcp-best-practices#store-tabular-data-in-bigquery
BigQuery for structured data, cloud storage for unstructed data

upvoted 4 times

ShePiDai 1 year, 5 months ago

agree. BigQuery and Cloud Storage have effectively identical storage performance, where BigQuery is optimised for structured dataset and GCS for unstructured.

upvoted 1 times

Mohamed_Mossad 2 years, 4 months ago

Selected Answer: D

"100 billion records stored in several CSV files" that means we deal with distributed big data problem , so HDFS is very suitable , Will choose D

upvoted 1 times

hoai_nam_1512 2 years, 2 months ago

HDFS will require more resources
100 bil record is processed fine with Cloud Storage object

upvoted 2 times

David_ml 2 years, 6 months ago

Answer is C. TFRecords in cloud storage for big data is the recommended practice by Google for training TF models.

upvoted 4 times

giaZ 2 years, 7 months ago

Selected Answer: C

Google best practices: Use Cloud Storage buckets and directories to group the shards of data (either sharded TFRecord files if using Tensorflow, or Avro if using any other framework). Aim for files of at least 100Mb, and 100 - 10000 shards.

upvoted 3 times

Load full discussion...

Start Learning for free

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.