PR OFESSIONAL-DATA-ENGINEER$^{Q\&As}$

Professional Data Engineer on Google Cloud Platform

Pass Google PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Following Questions and Answers are all new published by Google Official Exam Center

QUESTION 1

What are two of the benefits of using denormalized data structures in BigQuery?

A. Reduces the amount of data processed, reduces the amount of storage required

B. Increases query speed, makes queries simpler

C. Reduces the amount of storage required, increases query speed

D. Reduces the amount of data processed, increases query speed

Correct Answer: B

Denormalization increases query speed for tables with billions of rows because BigQuery\\'s performance degrades when doing JOINs on large tables, but with a denormalized data structure, you don\\'t have to use JOINs, since all of the data

has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses.

Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.

Reference:

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

QUESTION 2

You work for a large ecommerce company. You store your customers order data in Bigtable. You have a garbage collection policy set to delete the data after 30 days and the number of versions is set to 1. When the data analysts run a query to report total customer spending, the analysts sometimes see customer data that is older than 30 days. You need to ensure that the analysts do not see customer data older than 30 days while minimizing cost and overhead. What should you do?

A. Set the expiring values of the column families to 30 days and set the number of versions to 2.

B. Use a timestamp range filter in the query to fetch the customer\\'s data for a specific range.

C. Set the expiring values of the column families to 29 days and keep the number of versions to 1.

D. Schedule a job daily to scan the data in the table and delete data older than 30 days.

Correct Answer: B

By using a timestamp range filter in the query, you can ensure that the analysts only see the customer data that is within the desired time range, regardless of the garbage collection policy1. This option is the most cost-effective and simple way to avoid fetching data that is marked for deletion by garbage collection, as it does not require changing the existing policy or creating additional jobs. You can use the Bigtable client libraries or the cbt CLI to apply a timestamp range filter to your read requests2. Option A is not effective, as it increases the number of versions to 2, which may cause more data to be retained and increase the storage costs. Option C is not reliable, as it reduces the expiring values to 29 days, which may not match the actual data arrival and usage patterns. Option D is not efficient, as it requires scheduling a job daily to scan and delete the data, which may incur additional overhead and complexity. Moreover, none of these options guarantee that the data older than 30 days will be immediately deleted, as garbage collection is an asynchronous process that can take up to a week to remove the data3. References:

1: Filters | Cloud Bigtable Documentation | Google Cloud

2: Read data | Cloud Bigtable Documentation | Google Cloud

3: Garbage collection overview | Cloud Bigtable Documentation | Google Cloud

QUESTION 3

You are designing a system that requires an ACID-compliant database. You must ensure that the system requires minimal human intervention in case of a failure. What should you do?

A. Configure a Cloud SQL for MySQL instance with point-in-time recovery enabled.

B. Configure a Cloud SQL for PostgreSQL instance with high availability enabled.

C. Configure a Bigtable instance with more than one cluster.

D. Configure a BJgQuery table with a multi-region configuration.

Correct Answer: B

The best option to meet the ACID compliance and minimal human intervention requirements is to configure a Cloud SQL for PostgreSQL instance with high availability enabled. Key reasons: Cloud SQL for PostgreSQL provides full ACID compliance, unlike Bigtable which provides only atomicity and consistency guarantees. Enabling high availability removes the need for manual failover as Cloud SQL will automatically failover to a standby replica if the leader instance goes down. Point-in-time recovery in MySQL requires manual intervention to restore data if needed. BigQuery does not provide transactional guarantees required for an ACID database. Therefore, a Cloud SQL for PostgreSQL instance with high availability meets the ACID and minimal intervention requirements best. The automatic failover will ensure availability and uptime without administrative effort.

QUESTION 4

Which TensorFlow function can you use to configure a categorical column if you don\\'t know all of the possible values for that column?

A. categorical_column_with_vocabulary_list

B. categorical_column_with_hash_bucket

C. categorical_column_with_unknown_values

D. sparse_column_with_keys

Correct Answer: B

If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an auto-incremental ID starting from 0. What if we don\\'t know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column

occupation will be hashed to an integer ID as we encounter them in training. Reference:
https://www.tensorflow.org/tutorials/wide

QUESTION 5

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

1.

You will batch-load the posts once per day and run them through the Cloud Natural Language API.

2.

You will extract topics and sentiment from the posts.

3.

You must store the raw posts for archiving and reprocessing.

4.

You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

A. Store the social media posts and the data extracted from the API in BigQuery.

B. Store the social media posts and the data extracted from the API in Cloud SQL.

C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.

D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Correct Answer: D