DUMPSQUEEN

Google Professional Data Engineer Exam

Google Professional-Data-Engineer

Version Demo

Total Demo Questions: 10

Total Premium Questions: 184

Buy Premium PDF

QUESTION NO: 1

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of your Cloud Bigtable cluster. Which two actions can you take to accomplish this? (Choose two.)

A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.

B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.

C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.

D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.

E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

ANSWER: A C

QUESTION NO: 2

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.

B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.

C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.

D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

ANSWER: B

QUESTION NO: 3

You are updating the code for a subscriber to a Pub/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. Your subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?

A. Set up the Pub/Sub emulator on your local machine. Validate the behavior of your new subscriber logic before deploying it to production.

B. Create a Pub/Sub snapshot before deploying new subscriber code. Use a Seek operation to re-deliver messages that became available after the snapshot was created.

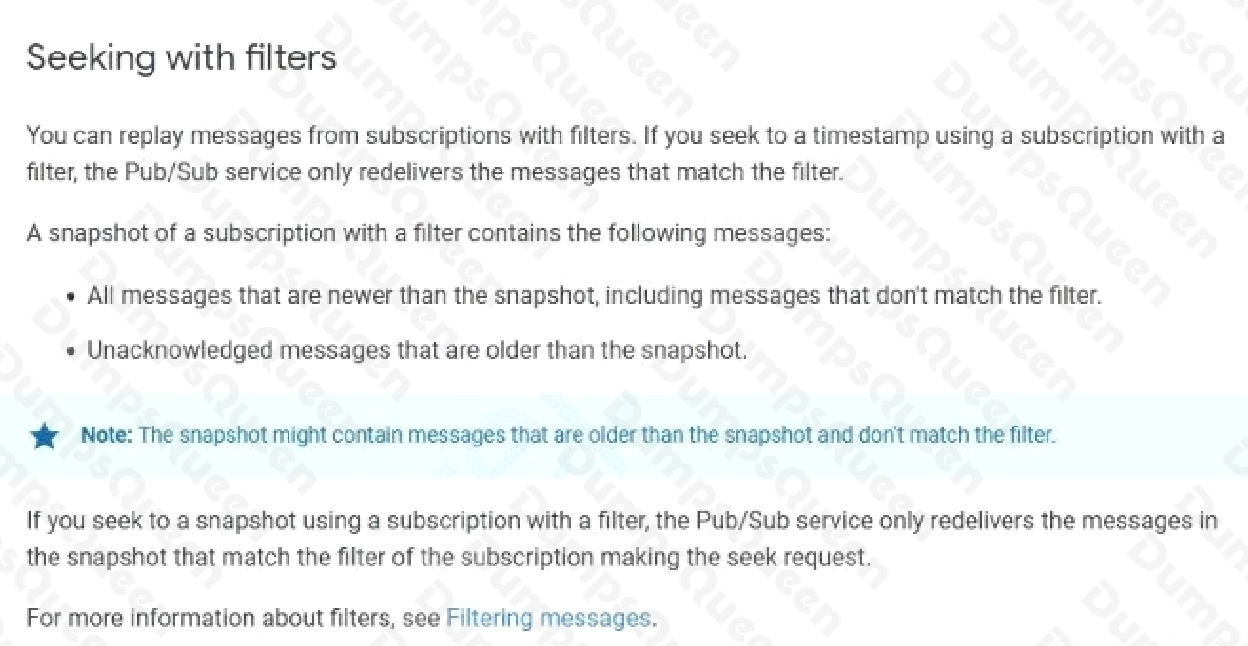
C. Use Cloud Build for your deployment. If an error occurs after deployment, use a Seek operation to locate a timestamp logged by Cloud Build at the start of the deployment.

D. Enable dead-lettering on the Pub/Sub topic to capture messages that aren’t successfully acknowledged. If an error occurs after deployment, re-deliver any messages captured by the dead-letter queue.

ANSWER: C

Explanation:

Reference: https://cloud.google.com/pubsub/docs/replay-overview



Seeking with filters

You can replay messages from subscriptions with filters. If you seek to a timestamp using a subscription with a filter, the Pub/Sub service only redelivers the messages that match the filter.

A snapshot of a subscription with a filter contains the following messages:

- All messages that are newer than the snapshot, including messages that don't match the filter.
- Unacknowledged messages that are older than the snapshot.

★ **Note:** The snapshot might contain messages that are older than the snapshot and don't match the filter.

If you seek to a snapshot using a subscription with a filter, the Pub/Sub service only redelivers the messages in the snapshot that match the filter of the subscription making the seek request.

For more information about filters, see Filtering messages.

QUESTION NO: 4

You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples

B. Reduce the number of training examples

C. Use a smaller set of features

D. Use a larger set of features

E. Increase the regularization parameters

F. Decrease the regularization parameters

ANSWER: A D F

QUESTION NO: 5

You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:

▪ Real-time event stream

▪ ANSI SQL access to real-time stream and historical data ▪ Batch historical exports

Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner

B. Cloud Pub/Sub, Cloud Storage, BigQuery

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

ANSWER: A

QUESTION NO: 6

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data. You noticed that the pipeline is taking longer to complete than expected; what should you do to expedite the Dataflow job?

A. Switch to compressed Avro files.

B. Reduce the batch size.

C. Retry records that throw an error.

D. Use CoGroupByKey instead of the SideInput.

ANSWER: C

QUESTION NO: 7

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

A. Load data into different partitions.

B. Load data into a different dataset for each client.

C. Put each client's BigQuery dataset into a different table.

D. Restrict a client's dataset to approved users.

E. Only allow a service account to access the datasets.

F. Use the appropriate identity and access management (IAM) roles for each client's users.

ANSWER: B D F

QUESTION NO: 8

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

A. Denormalize the data as must as possible.

B. Preserve the structure of the data as much as possible.

C. Use BigQuery UPDATE to further reduce the size of the dataset.

D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.

E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

ANSWER: D E

QUESTION NO: 9

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in. How should you design your row key and tables to ensure that you can access the data with the simplest query?

A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design.

B. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.

C. For each index, have a separate table and use a timestamp as the row key design.

D. For each index, have a separate table and use a reverse timestamp as the row key design.

ANSWER: D

QUESTION NO: 10

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.

You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis

B. HBase

C. MySQL

D. MongoDB

E. Cassandra

F. HDFS with Hive

ANSWER: B D F