



Vendor: Google

Exam Code: Professional-Data-Engineer

Exam Name: Professional Data Engineer

Version: DEMO

QUESTION 1

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources.

How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
- C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
- D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: C

Explanation:

It provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.

QUESTION 2

You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

- No interaction by the user on the site for 1 hour
- Has added more than \$30 worth of products to the basket
- Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- A. Use a fixed-time window with a duration of 60 minutes.
- B. Use a sliding time window with a duration of 60 minutes.
- C. Use a session window with a gap time duration of 60 minutes.
- D. Use a global window with a time based trigger with a delay of 60 minutes.

Answer: C

Explanation:

It will send a message per user after that user is inactive for 60 minutes. Session window works well for capturing a session per user basis.

QUESTION 3

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristics support this method? (Choose two.)

- A. There are very few occurrences of mutations relative to normal samples.
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.

- D. You expect future mutations to have similar features to the mutated samples in the database.
- E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AD

Explanation:

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

https://en.wikipedia.org/wiki/Anomaly_detection

QUESTION 4

You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

- A. Re-write the application to load accumulated data every 2 minutes.
- B. Convert the streaming insert code to batch load for individual messages.
- C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
- D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Answer: D

Explanation:

The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written to storage.

QUESTION 5

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- A. Use Google Stackdriver Audit Logs to review data access.
- B. Get the identity and access management (IAM) policy of each table
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Answer: A

Explanation:

First we need to know who is accessing what then we can create suitable policies. Stackdriver is used to track access logs for Bigquery.

QUESTION 6

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What

should you do?

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Answer: D

Explanation:

Dataproc is used to migrate Hadoop and Spark jobs on GCP. Dataproc with GCS connected through Google Cloud Storage connector helps store data after the life of the cluster. When the job is high I/O intensive, then we need to create a small persistent disk.

QUESTION 7

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

- A. Supervised learning to determine which transactions are most likely to be fraudulent.
- B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
- C. Clustering to divide the transactions into N categories based on feature similarity.
- D. Supervised learning to predict the location of a transaction.
- E. Reinforcement learning to predict the location of a transaction.
- F. Unsupervised learning to predict the location of a transaction.

Answer: BCD

Explanation:

Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features.

QUESTION 8

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

- A. The message body for the sensor event is too large.
- B. Your custom endpoint has an out-of-date SSL certificate.
- C. The Cloud Pub/Sub topic has too many messages published to it.
- D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

Answer: D

Explanation:

Until or unless the message is not acknowledged within defined ack window period for every message, we will get duplicate (number of retries to send message can be defined).

<https://cloud.google.com/pubsub/docs/troubleshooting#dupes>

QUESTION 9

Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

```
# Syntax error : Expected end of statement but got "-" at [4:11] SELECT
age
FROM
bigquery-public-data.noaa_gsod.gsod
WHERE
age != 99
AND_TABLE_SUFFIX = `1929`
ORDER BY
age DESC
```

Which table name will make the SQL statement work correctly?

- A. `bigquery-public-data.noaa_gsod.gsod`
- B. bigquery-public-data.noaa_gsod.gsod*
- C. `bigquery-public-data.noaa_gsod.gsod`*
- D. `bigquery-public-data.noaa_gsod.gsod`*

Answer: D

Explanation:

It follows the correct wildcard syntax of enclosing the table name in backticks and including the * wildcard character.

QUESTION 10

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiently?

- A. Assign global unique identifiers (GUID) to each data entry.
- B. Compute the hash value of each data entry, and compare it with all historical data.
- C. Store each data entry as the primary key in a separate database and apply an index.
- D. Maintain a database table to store the hash value and other metadata for each data entry.

Answer: D

Explanation:

Using Hash values we can remove duplicate values from a database. Hashvalues will be same for duplicate data and thus can be easily rejected.

QUESTION 11

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

- A. Grant the consultant the Viewer role on the project.
- B. Grant the consultant the Cloud Dataflow Developer role on the project.
- C. Create a service account and allow the consultant to log on with it.

D. Create an anonymized sample of the data for the consultant to work with in a different project.

Answer: B

Explanation:

A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs.

<https://cloud.google.com/iam/docs/understanding-service-accounts>

QUESTION 12

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour. The data scientists have written the following code to read the data for a new key features in the logs.

```
BigQueryIO.Read  
.named("ReadLogData")  
.from("clouddataflow-readonly:samples.log_data")
```

You want to improve the performance of this data read. What should you do?

- A. Specify the Tableobject in the code.
- B. Use .fromQuery operation to read specific fields from the table.
- C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
- D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table.

Answer: B

Explanation:

BigQueryIO.read.from() directly reads the whole table from BigQuery. This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from. This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from BigQuery).

BigQueryIO.read.fromQuery() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).

QUESTION 13

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

- A. Update the current pipeline and use the drain flag.
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Answer: B

Explanation:

If any transform names in your pipeline have changed, you must supply a transform mapping and pass it using the --transformNameMapping option.

https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#preventing_compatibility_breaks

QUESTION 14

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

- A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
- B. The performance issue should be resolved over time as the size of the BigData cluster is increased.
- C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
- D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Answer: A

Explanation:

<https://cloud.google.com/bigtable/docs/performance#troubleshooting>

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

QUESTION 15

Which of these statements about BigQuery caching is true?

- A. By default, a query's results are not cached.
- B. BigQuery caches query results for 48 hours.
- C. Query results are cached even if you specify a destination table.
- D. There is no charge for a query that retrieves its results from cache.

Answer: D

Explanation:

When query results are retrieved from a cached results table, you are not charged for the query. BigQuery caches query results for 24 hours, not 48 hours. Query results are not cached if you specify a destination table. A query's results are always cached except under certain conditions, such as if you specify a destination table.

Reference: <https://cloud.google.com/bigquery/querying-data#query-caching>

QUESTION 16

You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.

What should you do?

- A. Provide latitude and longitude as input vectors to your neural net.
- B. Create a numeric column from a feature cross of latitude and longitude.
- C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.
- D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

Answer: C

Explanation:

Use L1 regularization when you need to assign greater importance to more influential features. It shrinks less important feature to 0.

L2 regularization performs better when all input features influence the output & all with the weights are of equal size.

QUESTION 17

You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.

What should you do?

- A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
- B. Place the MariaDB instances in an Instance Group with a Health Check.
- C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
- D. Install the StackDriver Agent and configure the MySQL plugin.

Answer: C

Explanation:

The GitHub repository named google-fluentd-catch-all-config which includes the configuration files for the Logging agent for ingesting the logs from various third-party software packages.

QUESTION 18

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

- A. Export the records from the database as an Avro file.
Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- B. Export the records from the database as an Avro file.
Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- C. Export the records from the database into a CSV file.
Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage.
Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
- D. Export the records from the database as an Avro file.
Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage.

Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Answer: A

Explanation:

Google recommends that enterprises use Transfer Appliance in cases where it would take them over a week to upload data to the cloud via the internet, or when an enterprise needs to migrate over 60 TB of data.

QUESTION 19

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: A

Explanation:

By creating an authorized view one assures that the data is current and avoids taking more storage space (and cost) in order to share a dataset. B and D are not cost optimal and C does not guarantee that the data is kept updated.

QUESTION 20

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 much 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: AD

Explanation:

Denormalization will help in performance by reducing query time, update are not good with bigquery.

Thank You for Trying Our Product

Lead2pass Certification Exam Features:

- ★ More than **99,900** Satisfied Customers Worldwide.
- ★ Average **99.9%** Success Rate.
- ★ **Free Update** to match latest and real exam scenarios.
- ★ **Instant Download** Access! No Setup required.
- ★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.
- ★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.
- ★ **100%** Guaranteed Success or **100%** Money Back Guarantee.
- ★ **Fast**, helpful support **24x7**.



View list of all certification exams: <http://www.lead2pass.com/all-products.html>



Microsoft



ORACLE



CITRIX



JUNIPER
NETWORKS



EMC²
where information lives

10% Discount Coupon Code: ASTR14