Google Cloud Certified Professional Data Engineer

1. 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 efficiency?

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(s): D

---

2. Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being 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 TableReference object 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 PCollexction represents a single row in the table.

Answer(s): D

---

3. You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine.
Which learning algorithm should you use?

A. Linear regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Answer(s): A

---

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(s): D

---

5. 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(s): B

---

6. 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(s): C

---

7. Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration.
What should you do?

A. Put the data into Google Cloud Storage.

B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.

C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Answer(s): B

---

8. 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(s): D

---

9. Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery.
Which three approaches can you take? (Choose three.)

- [ ] A. Disable writes to certain tables.
- [ ] B. Restrict access to tables by role.
- [ ] C. Ensure that the data is encrypted at all times.
- [ ] D. Restrict BigQuery API access to approved users.
- [ ] E. Segregate data across multiple tables or databases.
- [ ] F. Use Google Stackdriver Audit Logging to determine policy violations.

Answer(s): B D F

---

10. You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive.
What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.

B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.

C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.

D. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to

true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.

E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer(s): D

---

11. Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations.
What should you do?

A. Add a node to the MySQL cluster and build an OLAP cube there.

B. Use an ETL tool to load the data from MySQL into Google BigQuery.

C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.

D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Answer(s): C

---

12. Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks.
What should you do?

A. Run a local version of Jupiter on the laptop.

B. Grant the user access to Google Cloud Shell.

C. Host a visualization tool on a VM on Google Compute Engine.

D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Answer(s): B

---

13. You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old.
What should you do?

A. Disable caching by editing the report settings.

B. Disable caching in BigQuery by editing table details.

C. Refresh your browser tab showing the visualizations.

D. Clear your browser history for the past hour then reload the tab showing the virtualizations.

Answer(s): A

---

14. Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly.
What method can you employ to address this?

A. Threading

B. Serialization

C. Dropout Methods

D. Dimensionality Reduction

Answer(s): C

---

15. You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time.
What should you do?

A. Send the data to Google Cloud Datastore and then export to BigQuery.

B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.

C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.

D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Answer(s): B

---

16. You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy.
What can you do?

A. Eliminate features that are highly correlated to the output labels.

B. Combine highly co-dependent features into one representative feature.

C. Instead of feeding in each feature individually, average their values in batches of 3.

D. Remove the features that have null values for more than 50% of the training records.

Answer(s): B

---

17. You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage

infrastructure scaling.
Which Google database service should you use?

A. Cloud SQL

B. BigQuery

C. Cloud Bigtable

D. Cloud Datastore

Answer(s): A

---

18. Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully.
What should you do next?

A. Check the dashboard application to see if it is not displaying correctly.

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.

D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Answer(s): B

---

19. 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(s): B D F

---

20. You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data.
Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.

B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.

C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.

D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer(s): D
