Product Questions: 384

Version: 21.0

Topic 1, Main Questions Set A

Question: 1

Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training dat

a. 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: C

Explanation:

Reference: https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

Question: 2

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

A. Continuously retrain the model on just the new data.

B. Continuously retrain the model on a combination of existing data and the new data.

C. Train on the existing data while using the new data as your test set.

D. Train on the new data while using the existing data as your test set.

Answer: C

Explanation:

https://cloud.google.com/automl-tables/docs/prepare

Question: 3

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:

Question: 4

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: A

Explanation:

Reference: https://support.google.com/datastudio/answer/7020039?hl=en

Question: 5

An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

A. Use federated data sources, and check data in the SQL query.

B. Enable BigQuery monitoring in Google Stackdriver and create an alert.

C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Answer: D

Explanation:

Question: 6

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

A. Issue a command to restart the database servers.

B. Retry the query with exponential backoff, up to a cap of 15 minutes.

C. Retry the query every second until it comes back online to minimize staleness of data.

D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:

https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

Question: 7

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: A

Explanation:

Question: 8

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 dat

a. 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: D

Explanation:

https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

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:

Question: 10

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: B,D,F

Explanation:

Question: 11

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:

Question: 12

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 dat

a. 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

Explanation:

Question: 13

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: A

Explanation:

Question: 14

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 characteristic 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: 15

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 dat

a. 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 tio storage

Question: 16

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 IIAM) 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:

Question: 17

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:

Question: 18

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 dat

a. 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: B,C,D

Explanation:

Question: 19

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: B

Explanation:

Reference:

Question: 20

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: B

Explanation:

Question: 21

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 dat

a. 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: D

Explanation:

Question: 22

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: B

Explanation:

Question: 23

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: B

Explanation:

Question: 24

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: D

Explanation:

Question: 25

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.

B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.

D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: B

Explanation:

Question: 26

You are working on a sensitive project involving private user dat

a. 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: C

Explanation:

Question: 27

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: B

Explanation:

Question: 28

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: D

Explanation:

Question: 29

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

A. Use a row key of the form <timestamp>.

B. Use a row key of the form <sensorid>.

C. Use a row key of the form <timestamp>#<sensorid>.

D. Use a row key of the form >#<sensorid>#<timestamp>.