IT认证电子书

质　量　更　高　　服　务　更　好

半年免费升级服务

http://www.itrenzheng.com

Exam : **Professional Data Engineer**

Title : Google Certified Professional – Data Engineer

Version : DEMO

1. Topic 1, Main Questions Set A

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

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

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

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

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