



VCE & PDF

GeekCert.com

<https://www.geekcert.com/professional-data-engineer.html>
2024 Latest geekcert PROFESSIONAL-DATA-ENGINEER PDF and VCE
dumps Download

PROFESSIONAL-DATA-ENGINEER^{Q&As}

Professional Data Engineer on Google Cloud Platform

**Pass Google PROFESSIONAL-DATA-ENGINEER
Exam with 100% Guarantee**

Free Download Real Questions & Answers **PDF** and **VCE** file from:

<https://www.geekcert.com/professional-data-engineer.html>

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Google
Official Exam Center



VCE & PDF

GeekCert.com

<https://www.geekcert.com/professional-data-engineer.html>

2024 Latest geekcert PROFESSIONAL-DATA-ENGINEER PDF and VCE
dumps Download

- ⚙️ **Instant Download** After Purchase
- ⚙️ **100% Money Back** Guarantee
- ⚙️ **365 Days** Free Update
- ⚙️ **800,000+** Satisfied Customers





QUESTION 1

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

- A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.
- B. Create a new view with BigQuery that does not include a column which city information.
- C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2 3, 4, or 5, and then use that number to represent the city in the model.
- D. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file and upload that as part of your model to BigQuery ML.

Correct Answer: C

QUESTION 2

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

- A. PigLatin using Pig
- B. HiveQL using Hive
- C. Java using MapReduce
- D. Python using MapReduce

Correct Answer: D

QUESTION 3

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user,

and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first.

What should you do?

- A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
- B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud



Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.

C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

Correct Answer: C

QUESTION 4

You have data located in BigQuery that is used to generate reports for your company. You have noticed some weekly executive report fields do not correspond to format according to company standards for example, report errors include different telephone formats and different country code identifiers. This is a frequent issue, so you need to create a recurring job to normalize the data. You want a quick solution that requires no coding What should you do?

A. Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.

B. Use BigQuery and GoogleSQL to normalize the data, and schedule recurring queries in BigQuery.

C. Create a Spark job and submit it to Dataproc Serverless.

D. Use Dataflow SQL to create a job that normalizes the data, and that after the first run of the job, schedule the pipeline to execute recurrently.

Correct Answer: A

Cloud Data Fusion is a fully managed, cloud-native data integration service that allows you to build and manage data pipelines with a graphical interface. Wrangler is a feature of Cloud Data Fusion that enables you to interactively explore, clean, and transform data using a spreadsheet-like UI. You can use Wrangler to normalize the data in BigQuery by applying various directives, such as parsing, formatting, replacing, and validating data. You can also preview the results and export the wrangled data to BigQuery or other destinations. You can then set up a recurring job in Cloud Data Fusion to run the Wrangler pipeline on a schedule, such as weekly or daily. This way, you can create a quick and code-free solution to normalize the data for your reports. References: Cloud Data Fusion overview Wrangler overview Wrangle data from BigQuery [Scheduling pipelines]

QUESTION 5

What are all of the BigQuery operations that Google charges for?

A. Storage, queries, and streaming inserts

B. Storage, queries, and loading data from a file

C. Storage, queries, and exporting data

D. Queries and streaming inserts

Correct Answer: A



VCE & PDF

GeekCert.com

<https://www.geekcert.com/professional-data-engineer.html>

2024 Latest geekcert PROFESSIONAL-DATA-ENGINEER PDF and VCE
dumps Download

Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations. Reference: <https://cloud.google.com/bigquery/pricing>

[Latest PROFESSIONAL-
DATA-ENGINEER Dumps](#)

[PROFESSIONAL-DATA-
ENGINEER PDF Dumps](#)

[PROFESSIONAL-DATA-
ENGINEER Braindumps](#)