PR OFESSIONAL-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.pass2lead.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

⚙ **Instant Download** After Purchase

⚙ **100% Money Back** Guarantee

⚙ **365 Days** Free Update

⚙ **800,000+** Satisfied Customers

QUESTION 1

Which SQL keyword can be used to reduce the number of columns processed by BigQuery?

A. BETWEEN

B. WHERE

C. SELECT

D. LIMIT

Correct Answer: C

SELECT allows you to query specific columns rather than the whole table.

LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.

Reference:

https://cloud.google.com/bigquery/launch-checklist#architecture_design_and_development_checklist

QUESTION 2

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.

B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.

C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.

D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

Correct Answer: B

QUESTION 3

You need ads data to serve AI models and historical data tor analytics longtail and outlier data points need to be identified

You want to cleanse the data n near-reel time before running it through AI models

What should you do?

A. Use BigQuery to ingest prepare and then analyze the data and then run queries to create views

B. Use Cloud Storage as a data warehouse shell scripts tor processing, and BigQuery to create views tor desired datasets

C. Use Dataflow to identity longtail and outber data points programmatically with BigQuery as a sink

D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery

Correct Answer: A

QUESTION 4

Flowlogistic\\'s management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably.

Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage

B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD

C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage

D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Correct Answer: A

QUESTION 5

MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day.

Which schema should you use?

A. Rowkey: date#device_idColumn data: data_point

B. Rowkey: dateColumn data: device_id, data_point

C. Rowkey: device_idColumn data: date, data_point

D. Rowkey: data_pointColumn data: device_id, date

E. Rowkey: date#data_pointColumn data: device_id

Correct Answer: A

![Pass2Lead](https://Pass2Lead.com)
QUESTION 6

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

A. Create a new view over events using standard SQL

B. Create a new partitioned table using a standard SQL query

C. Create a new view over events_partitioned using standard SQL

D. Create a service account for the ODBC connection to use for authentication

E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

Correct Answer: CD

A standard SQL query cannot reference a view defined using legacy SQL syntax. In order to connect through ODBC connection, we need to use standard SQL. So, we need to create a new view over events_partitioned table using standard SQL which is C. Need service account to connect through ODBC which is option D. Check the links below. I am not sure about A whether we can create a view over another view which was built using legacy SQL https://cloud.google.com/bigquery/docs/views https://cloud.google.com/community/tutorials/bigquery-from-excel https://www.simba.com/products/BigQuery/doc/ODBC_InstallGuide/mac/content/odbc/bq/configuring/authenticating/serviceaccount.htm

QUESTION 7

Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

A. Field promotion

B. Randomization

C. Salting

D. Hashing

Correct Answer: A

By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.

Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

QUESTION 8

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.

Correct Answer: D

QUESTION 9

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes

B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project

C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

Correct Answer: B

QUESTION 10

You work for an advertising company, and you\\'ve developed a Spark ML model to predict click-through rates at advertisement blocks. You\\'ve been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Cloud ML Engine

C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Correct Answer: C

QUESTION 11

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don\'t get slots to execute their query and you need to correct this. You\'d like to avoid introducing new projects to your account.

What should you do?

A. Convert your batch BQ queries into interactive BQ queries.

B. Create an additional project to overcome the 2K on-demand per-project quota.

C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.

D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Correct Answer: C

Reference https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

QUESTION 12

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

A. Batch job, PubSubIO, side-inputs

B. Streaming job, PubSubIO, JdbcIO, side-outputs

C. Streaming job, PubSubIO, BigQueryIO, side-inputs

D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Correct Answer: C

QUESTION 13

Cloud Dataproc charges you only for what you really use with _____ billing.

A. month-by-month

B. minute-by-minute

C. week-by-week

D. hour-by-hour

![Pass2Lead](https://Pass2Lead.com)
Correct Answer: B

One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period. Reference: https://cloud.google.com/dataproc/docs/concepts/overview

---

QUESTION 14

Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

A. Hive

B. Pig

C. YARN

D. Spark

Correct Answer: ABD

Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs. Reference: https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

---

QUESTION 15

Google Cloud Bigtable indexes a single value in each row. This value is called the _____.

A. primary key

B. unique key

C. row key

D. master key

Correct Answer: C

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key.

Reference: https://cloud.google.com/bigtable/docs/overview

Latest PROFESSIONAL-DATA-ENGINEER Dumps   **PROFESSIONAL-DATA-ENGINEER Practice Test**   **PROFESSIONAL-DATA-ENGINEER Braindumps**