DUMPSARENA

Google Professional Data Engineer Exam

Google Professional-Data-Engineer

Version Demo

Total Demo Questions: 10

Total Premium Questions: 184

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

QUESTION NO: 1

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.

You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis

B. HBase

C. MySQL

D. MongoDB

E. Cassandra

F. HDFS with Hive

ANSWER: B D F

QUESTION NO: 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 PCollection represents a single row in the table.

ANSWER: D

QUESTION NO: 3

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

A. Denormalize the data as must as possible.

B. Preserve the structure of the data as much as possible.

C. Use BigQuery UPDATE to further reduce the size of the dataset.

D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.

E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

ANSWER: D E

QUESTION NO: 4

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? (Choose two.)

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.

B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.

C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.

D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.

E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

ANSWER: C E

QUESTION NO: 5

You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples

B. Reduce the number of training examples

C. Use a smaller set of features

D. Use a larger set of features

E. Increase the regularization parameters

F. Decrease the regularization parameters

ANSWER: A D F

QUESTION NO: 6

MJTelco Case Study

Company Overview

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

▪ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations. ▪ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments – development/test, staging, and production – to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements

- Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. - Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

- Provide reliable and timely access to data for analysis from distributed research workers

- Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements

- Ensure secure and efficient transport and storage of telemetry data

- Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone

B. The number of workers

C. The disk size per worker

D. The maximum number of workers

ANSWER: A

QUESTION NO: 7

You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum

CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

A. Increase the number of max workers

B. Use a larger instance type for your Cloud Dataflow workers

C. Change the zone of your Cloud Dataflow pipeline to run in us-central1

D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery

E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

ANSWER: B E

QUESTION NO: 8

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: A

QUESTION NO: 9

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.

C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

ANSWER: B

QUESTION NO: 10

Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.

B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.

C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.

D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.

ANSWER: D