Start your Professional-Data-Engineer Exam Questions Preparation with Updated 333 Questions [Q159-Q173

Start your Professional-Data-Engineer Exam Questions Preparation with Updated 333 Questions A Fully Updated 2024 Professional-Data-Engineer Exam Dumps - PDF Questions and Testing Engine

Google Professional-Data-Engineer certification exam comprises of multiple-choice and multiple-select questions that require a thorough understanding of Google Cloud Platform services such as BigQuery, Google Cloud Storage, and Google Cloud Dataflow. Professional-Data-Engineer exam also tests an individual's knowledge of data processing patterns and best practices, understanding of machine learning models and algorithms, and proficiency in designing and deploying solutions that meet business requirements.

NEW QUESTION 159

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?

- Run a local version of Jupiter on the laptop.
- Grant the user access to Google Cloud Shell.
- Host a visualization tool on a VM on Google Compute Engine.
- Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Explanation/Reference:

NEW QUESTION 160

You are building a model to make clothing recommendations. You know a user&#8217;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?
* Continuously retrain the model on just the new data.
* Continuously retrain the model on a combination of existing data and the new data.
* Train on the existing data while using the new data as your test set.
* Train on the new data while using the existing data as your test set.

NEW QUESTION 161

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?
* Use Cloud Dataprep to find null values in sample source data. Convert all nulls to `none&#8217; using a Cloud Dataproc job.
* Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.
* Use Cloud Dataflow to find null values in sample source data. Convert all nulls to `none&#8217; using a Cloud Dataprep job.
* Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.

NEW QUESTION 162

Which of these statements about exporting data from BigQuery is false?
* To export more than 1 GB of data, you need to put a wildcard in the destination filename.
* The only supported export destination is Google Cloud Storage.
* Data can only be exported in JSON or Avro format.
* The only compression option available is GZIP.
Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.

Reference: https://cloud.google.com/bigquery/docs/exporting-data

NEW QUESTION 163

Which of the following statements is NOT true regarding Bigtable access roles?
* Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
* To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
* You can configure access control only at the project level.
* To give a user access to only one table in a project, you must configure access through your application.
For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:

Read from, but not write to, any table within the project.

Read from and write to any table within the project, but not manage instances.

Read from and write to any table within the project, and manage instances.

Reference: https://cloud.google.com/bigtable/docs/access-control

NEW QUESTION 164

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in How should you design your row key and tables to ensure that you can access the data with the most simple query?

* Create one unique table for all of the indices, and then use the index and timestamp as the row key design
* Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.
* For each index, have a separate table and use a timestamp as the row key design
* For each index, have a separate table and use a reverse timestamp as the row key design

NEW QUESTION 165

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?

* Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional

storage bucket as a final destination.

* Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as

a final destination.

* Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-

Regional storage bucket as a final destination.

* Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional

storage bucket as a final destination.

NEW QUESTION 166

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?

* 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.
* 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.
* 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.
* 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.

NEW QUESTION 167

Flowlogistic&#8217;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?

* Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
* Cloud Pub/Sub, Cloud Dataflow, and Local SSD
* Cloud Pub/Sub, Cloud SQL, and Cloud Storage
* Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

NEW QUESTION 168

Case Study: 2 &#8211; MJTelco

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&#8217;s machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualizations for operations teams with the following requirements:

Which approach meets the requirements?
* Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
* Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
* Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
* Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

NEW QUESTION 169

Your company&#8217;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?

* Put the data into Google Cloud Storage.
* Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
* Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
* Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.
Explanation/Reference:

Reference: https://cloud.google.com/dataproc/

NEW QUESTION 170

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 &#8211; development/test, staging, and production

&#8211; 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&#8217;s machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualizations for operations teams with the following requirements:

The report must include telemetry data from all 50,000 installations for the most resent 6 weeks

-

(sampling once every minute).

The report must not be more than 3 hours delayed from live data.

-

The actionable report should only show suboptimal links.

-

Most suboptimal links should be sorted to the top.

-

Suboptimal links can be grouped and filtered by regional geography.

-

User response time to load the report must be <5 seconds.

-

Which approach meets the requirements?
* Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show

only suboptimal links in a table.
* Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates

the metric, and shows only suboptimal rows in a table in Google Sheets.
* Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries

all rows, applies a function to derive the metric, and then renders results in a table using the Google

charts and visualization API.
* Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to

your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a

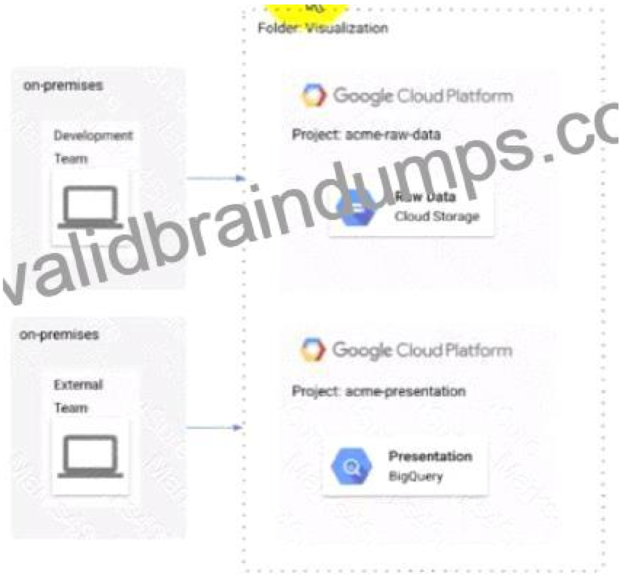
table.

NEW QUESTION 171

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?
* cron
* Cloud Composer
* Cloud Scheduler
* Workflow Templates on Cloud Dataproc

NEW QUESTION 172

The Development and External teams nave the project viewer Identity and Access Management (1AM) role m a folder named Visualization. You want the Development Team to be able to read data from both Cloud Storage and BigQuery, but the External Team should only be able to read data from BigQuery. What should you do?



* Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project
* Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data protect that deny all Ingress traffic from the External Team CIDR range
* Create a VPC Service Controls perimeter containing both protects and BigQuery as a restricted API Add the External Team users to the perimeter s Access Level
* Create a VPC Service Controls perimeter containing both protects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter&#8217;s Access Level

NEW QUESTION 173

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub

streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?
* They have not assigned the timestamp, which causes the job to fail
* They have not set the triggers to accommodate the data coming in late, which causes the job to fail
* They have not applied a global windowing function, which causes the job to fail when the pipeline is

created
* They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Google Professional-Data-Engineer: Google Certified Professional Data Engineer Exam is an essential certification exam for professionals looking to advance their careers in the field of data engineering. Passing Professional-Data-Engineer exam validates a candidate's expertise in designing, building, and managing data processing systems. It also demonstrates their ability to analyze and interpret data, make informed business decisions, and leverage cloud-based data processing systems to achieve business objectives.

Easy Success Google Professional-Data-Engineer Exam in First Try:
https://www.validbraindumps.com/Professional-Data-Engineer-exam-prep.html]