Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

NEW QUESTION 1
- (Exam Topic 1)
You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

A. There are very few occurrences of mutations relative to normal samples.
B. There are roughly equal occurrences of both normal and mutated samples in the database.
C. You expect future mutations to have different features from the mutated samples in the database.
D. You expect future mutations to have similar features to the mutated samples in the database.
E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AD

Explanation:
Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. https://en.wikipedia.org/wiki/Anomaly_detection

NEW QUESTION 2
- (Exam Topic 1)
Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:
Syntax error : Expected end of statement but got "-" at [4:11] SELECT age
FROM
bigquery-public-data.noaa_gsod.gsod WHERE
age != 99
AND_TABLE_SUFFIX = '1929' ORDER BY
age DESC
Which table name will make the SQL statement work correctly?

A. 'bigquery-public-data.noaa_gsod.gsod'
B. bigquery-public-data.noaa_gsod.gsod*
C. 'bigquery-public-data.noaa_gsod.gsod'*`
D. 'bigquery-public-data.noaa_gsod.gsod*`

Answer: D

NEW QUESTION 3
- (Exam Topic 1)
You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.
B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer: D

Explanation:
https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

NEW QUESTION 4
- (Exam Topic 1)
You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.
B. Combine highly co-dependent features into one representative feature.
C. Instead of feeding in each feature individually, average their values in batches of 3.
D. Remove the features that have null values for more than 50% of the training records.

Answer: B

NEW QUESTION 5
- (Exam Topic 1)
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

NEW QUESTION 6
- (Exam Topic 1)
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

A. Supervised learning to determine which transactions are most likely to be fraudulent.
B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
C. Clustering to divide the transactions into N categories based on feature similarity.
D. Supervised learning to predict the location of a transaction.
E. Reinforcement learning to predict the location of a transaction.
F. Unsupervised learning to predict the location of a transaction.

Answer: BCD

NEW QUESTION 7
- (Exam Topic 3)
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

Answer: D

NEW QUESTION 8
- (Exam Topic 3)
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

NEW QUESTION 9
- (Exam Topic 3)
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)

A. Ensure all the tables are included in global dataset.
B. Ensure each table is included in a dataset for a region.
C. Adjust the settings for each table to allow a related region-based security group view access.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Adjust the settings for each dataset to allow a related region-based security group view access.

Answer: BD

NEW QUESTION 10
- (Exam Topic 4)
You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

A. Manually configure the index in your index config as follows:
```
Indexes:
-kind: Movie
     Properties:
     -name: actors
     name: date_released
-kind: Movie
     Properties:
     -name: tags
     name: date_released
```

B. Manually configure the index in your index config as follows:
```
Indexes:
     -kind: Movie
          Properties:
          -name: actors
          -name: tags
-name: date_published
```

C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'

D. Set the following in your entity options: exclude_from_indexes = 'date_published'

A. Option A
B. Option B.
C. Option C
D. Option D

Answer: A

NEW QUESTION 10
- (Exam Topic 5)
Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?

A. A sequential numeric ID
B. A timestamp followed by a stock symbol
C. A non-sequential numeric ID
D. A stock symbol followed by a timestamp

Answer: AB

Explanation:
using a timestamp as the first element of a row key can cause a variety of problems.
In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill
that node; and then move onto the next node in the cluster, resulting in hotspotting.
Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table.
However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.
[https://cloud.google.com/bigtable/docs/schema-design]
Reference:
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

NEW QUESTION 11
- (Exam Topic 5)
What are two methods that can be used to denormalize tables in BigQuery?

A. 1) Split table into multiple tables; 2) Use a partitioned table
B. 1) Join tables into one table; 2) Use nested repeated fields
C. 1) Use a partitioned table; 2) Join tables into one table
D. 1) Use nested repeated fields; 2) Use a partitioned table

Answer: B

Explanation:
The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are
dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information.
The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing
records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a

JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.
Reference: https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION 12
- (Exam Topic 5)
Which is not a valid reason for poor Cloud Bigtable performance?

A. The workload isn't appropriate for Cloud Bigtable.
B. The table's schema is not designed correctly.
C. The Cloud Bigtable cluster has too many nodes.
D. There are issues with the network connection.

Answer: C

Explanation:
The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.
Reference: https://cloud.google.com/bigtable/docs/performance

NEW QUESTION 15
- (Exam Topic 5)
Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.
SELECT person FROM `project1.example.table1` WHERE city = "London" How would you correct the error?

A. Add ", UNNEST(person)" before the WHERE clause.
B. Change "person" to "person.city".
C. Change "person" to "city.person".
D. Add ", UNNEST(city)" before the WHERE clause.

Answer: A

Explanation:
To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_resu

NEW QUESTION 19
- (Exam Topic 5)
Which of the following IAM roles does your Compute Engine account require to be able to run pipeline jobs?

A. dataflow.worker
B. dataflow.compute
C. dataflow.developer
D. dataflow.viewer

Answer: A

Explanation:
The dataflow.worker role provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline
Reference: https://cloud.google.com/dataflow/access-control

NEW QUESTION 24
- (Exam Topic 5)
Why do you need to split a machine learning dataset into training data and test data?

A. So you can try two different sets of features
B. To make sure your model is generalized for more than just the training data
C. To allow you to create unit tests in your code
D. So you can use one dataset for a wide model and one for a deep model

Answer: B

Explanation:
The flaw with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting.
Reference: https://machinelearningmastery.com/a-simple-intuition-for-overfitting/

NEW QUESTION 26
- (Exam Topic 5)
Which of these numbers are adjusted by a neural network as it learns from a training dataset (select 2 answers)?

A. Weights
B. Biases
C. Continuous features
D. Input values

Answer: AB

Explanation:
A neural network is a simple mechanism that's implemented with basic math. The only difference between the traditional programming model and a neural network is that you let the computer determine the parameters (weights and bias) by learning from training datasets.
Reference:
https://cloud.google.com/blog/big-data/2016/07/understanding-neural-networks-with-tensorflow-playground

NEW QUESTION 31
- (Exam Topic 5)
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

A. The wide model is used for memorization, while the deep model is used for generalization.
B. A good use for the wide and deep model is a recommender system.
C. The wide model is used for generalization, while the deep model is used for memorization.
D. A good use for the wide and deep model is a small-scale linear regression problem.

Answer: AB

Explanation:
Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.
Reference: https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

NEW QUESTION 33
- (Exam Topic 5)
Which of the following statements is NOT true regarding Bigtable access roles?

A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
C. You can configure access control only at the project level.
D. To give a user access to only one table in a project, you must configure access through your application.

Answer: B

Explanation:
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 38
- (Exam Topic 5)
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

Answer: B

Explanation:
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

NEW QUESTION 43
- (Exam Topic 5)
Which of these statements about exporting data from BigQuery is false?

A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
B. The only supported export destination is Google Cloud Storage.
C. Data can only be exported in JSON or Avro format.
D. The only compression option available is GZIP.

Answer: C

Explanation:
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 44
- (Exam Topic 6)
You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and

need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.
What should you do?

A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Answer: B

NEW QUESTION 49
- (Exam Topic 6)
You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

A. Deploy small Kafka clusters in your data centers to buffer events.
B. Have the data acquisition devices publish data to Cloud Pub/Sub.
C. Establish a Cloud Interconnect between all remote data centers and Google.
D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

Answer: B

NEW QUESTION 52
- (Exam Topic 6)
You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

A. Partition by transaction time; cluster by state first, then city then store ID
B. Partition by transaction tome cluster by store ID first, then city, then stale
C. Top-level cluster by stale first, then city then store
D. Top-level cluster by store ID first, then city then state.

Answer: C

NEW QUESTION 54
- (Exam Topic 6)
The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
D. Import the new records from the CSV file into a new BigQuery tabl
E. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

Answer: D

NEW QUESTION 55
- (Exam Topic 6)
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

Answer: B

NEW QUESTION 56
- (Exam Topic 6)
You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.
B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Answer: B

NEW QUESTION 60
- (Exam Topic 6)
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 dat
E. 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
F. Create a temporary table in Cloud Spanner that will act as a buffer for new dat
G. 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: AB

NEW QUESTION 65
- (Exam Topic 6)
You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.
What should you do?

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
B. Place the MariaDB instances in an Instance Group with a Health Check.
C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
D. Install the StackDriver Agent and configure the MySQL plugin.

Answer: C

NEW QUESTION 67
- (Exam Topic 6)
You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longtitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.
What should you do?

A. Provide latitude and longtitude as input vectors to your neural net.
B. Create a numeric column from a feature cross of latitude and longtitude.
C. Create a feature cross of latitude and longtitude, bucketize at the minute level and use L1 regularizationduring optimization.
D. Create a feature cross of latitude and longtitude, bucketize it at the minute level and use L2 regularization during optimization.

Answer: C

Explanation:
Reference https://cloud.google.com/bigquery/docs/gis-dataa

NEW QUESTION 72
- (Exam Topic 6)
You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

A. Use bq load to load a batch of sensor data every 60 seconds.
B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
C. Use the INSERT statement to insert a batch of data every 60 seconds.
D. Use the MERGE statement to apply updates in batch every 60 seconds.

Answer: C

NEW QUESTION 77
- (Exam Topic 6)
A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?

A. Change the VM type to n2-highmem-32
B. Change the VM type to e2 standard-32
C. Train the model using a VM with a GPU hardware accelerator
D. Train the model using a VM with a TPU hardware accelerator

Answer: C

NEW QUESTION 81
- (Exam Topic 6)
You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

A. Convert all daily log tables into date-partitioned tables

B. Convert the sharded tables into a single partitioned table
C. Enable query caching so you can cache data from previous months
D. Create separate views to cover each month, and query from these views

Answer: A

NEW QUESTION 86
- (Exam Topic 6)
An organization maintains a Google BigQuery dataset that contains tables with user-level datA. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

A. Create and share an authorized view that provides the aggregate results.
B. Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Answer: D

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

NEW QUESTION 89
- (Exam Topic 6)
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

A. Cloud Scheduler
B. Cloud Dataflow
C. Cloud Functions
D. Cloud Composer

Answer: A

NEW QUESTION 94
- (Exam Topic 6)
You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

A. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows
B. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations
C. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations
D. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

Answer: A

NEW QUESTION 96
- (Exam Topic 6)
Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of datA. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

A. Encrypted on Cloud Storage with user-supplied encryption key
B. A separate decryption key will be given to each authorized user.
C. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
D. In Cloud SQL, with separate database user names to each use
E. The Cloud SQL Admin activity logs will be used to provide the auditability.
F. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Answer: B

NEW QUESTION 98
- (Exam Topic 6)
Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

A. Set a single global window to capture all the data.
B. Set sliding windows to capture all the lagged data.
C. Use watermarks and timestamps to capture the lagged data.
D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Answer: B

NEW QUESTION 102
- (Exam Topic 6)

You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

A. Store and process the entire dataset in BigQuery.
B. Store and process the entire dataset in Cloud Bigtable.
C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
D. Store the warm data as files in Cloud Storage, and store the active data in BigQuer
E. Keep this ratio as 80% warm and 20% active.

Answer: C

NEW QUESTION 104
- (Exam Topic 6)
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 fil
B. Process the file with Apache Hadoop to identify which user bid first.
C. Have each application server write the bid events to Cloud Pub/Sub as they occu
D. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
E. Set up a MySQL database for each application server to write bid events int
F. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
G. Have each application server write the bid events to Google Cloud Pub/Sub as they occu
H. Use a pull subscription to pull the bid events using Google Cloud Dataflo
I. Give the bid for each item to the userin the bid event that is processed first.

Answer: C

NEW QUESTION 108
- (Exam Topic 6)
You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.
What should you do?

A. Increase the size of the dataset by collecting additional data.
B. Train a linear regression to predict a credit default risk score.
C. Remove the bias from the data and collect applications that have been declined loans.
D. Match loan applicants with their social profiles to enable feature engineering.

Answer: B

NEW QUESTION 110
- (Exam Topic 6)
You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage.Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.

Answer: A

NEW QUESTION 114
- (Exam Topic 6)
You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Answer: C

NEW QUESTION 115
- (Exam Topic 6)
You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

A. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery
B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.
C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table

D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery

Answer: A

NEW QUESTION 120
- (Exam Topic 6)
You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storag
B. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
C. Use Cloud Bigtable for storag
D. Link as permanent tables in BigQuery for query.
E. Use Cloud Storage for storag
F. Link as permanent tables in BigQuery for query.
G. Use Cloud Storage for storag
H. Link as temporary tables in BigQuery for query.

Answer: A

NEW QUESTION 125
- (Exam Topic 6)
You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service About 70% of customer requests are simple requests that are solved within 10 intents The remaining 30% of inquiries require much longer, more complicated requests Which intents should you automate first?

A. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests
B. Automate the more complicated requests first because those require more of the agents' time
C. Automate a blend of the shortest and longest intents to be representative of all intents
D. Automate intents in places where common words such as "payment" appear only once so the software isn't confused

Answer: A

NEW QUESTION 129
- (Exam Topic 6)
You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?

A. BigQuery
B. Cloud Bigtable
C. Cloud Datastore
D. Cloud SQL for PostgreSQL

Answer: A

NEW QUESTION 134
- (Exam Topic 6)
You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

A. Use Cloud TPUs without any additional adjustment to your code.
B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
D. Stay on CPUs, and increase the size of the cluster you're training your model on.

Answer: B

NEW QUESTION 137
- (Exam Topic 6)
You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data tittered by location_id and device_version with the following query:

```
SELECT
    MAX(temperature)
FROM
    acme_iot_data.sensors
WHERE
    create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
    AND location_id = "SW1W9TQ"
    AND device_version = "202007r3"
```

You want to optimize your queries for cost and performance. How should you structure your data?

A. Partition table data by create_date, location_id and device_version
B. Partition table data by create_date cluster table data by tocation_id and device_version
C. Cluster table data by create_date location_id and device_version
D. Cluster table data by create_date, partition by location and device_version

Answer: C

NEW QUESTION 139
- (Exam Topic 6)
You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API Following Google's best practices, you have both a staging and a production table for the data How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging
B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging
C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours
D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes

Answer: D

NEW QUESTION 141
- (Exam Topic 6)
You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?

A. Use Cloud Vision AutoML with the existing dataset.
B. Use Cloud Vision AutoML, but reduce your dataset twice.
C. Use Cloud Vision API by providing custom labels as recognition hints.
D. Train your own image recognition model leveraging transfer learning techniques.

Answer: A

NEW QUESTION 146
- (Exam Topic 6)
After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.
What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.
B. Select random samples from the tables using the HASH() function and compare the samples.
C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sortin
D. Compare the hashes of each table.
E. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer: B

NEW QUESTION 151
- (Exam Topic 6)
Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

❯ Single global endpoint

❯ ANSI SQL support

❯ Consistent access to the most up-to-date data
What should you do?

A. Implement BigQuery with no region selected for storage or processing.
B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.
C. Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.
D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

Answer: B

NEW QUESTION 152
- (Exam Topic 6)
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?

A. They have not assigned the timestamp, which causes the job to fail
B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Answer: C

NEW QUESTION 157
- (Exam Topic 6)
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

Answer: C

Explanation:
A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data. For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).
https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design

NEW QUESTION 158
- (Exam Topic 6)
You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.
What should you do?

A. Deploy a Kafka cluster on GCE VM Instance
B. Configure your on-prem cluster to mirror your topics to the cluster running in GC
C. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
D. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connecto
E. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
F. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connecto
G. Use a Dataflow job to read fron PubSub and write to GCS.
H. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connecto
I. Use a Dataflow job to read fron PubSub and write to GCS.

Answer: A

NEW QUESTION 163
- (Exam Topic 6)
You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

A. An alert based on a decrease of subscription/num_undelivered_messages for the source and a rate of change increase of instance/storage/used_bytes for the destination
B. An alert based on an increase of subscription/num_undelivered_messages for the source and a rate of change decrease of instance/storage/used_bytes for the destination
C. An alert based on a decrease of instance/storage/used_bytes for the source and a rate of change increase of subscription/num_undelivered_messages for the destination
D. An alert based on an increase of instance/storage/used_bytes for the source and a rate of change decrease of subscription/num_undelivered_messages for the destination

Answer: B

NEW QUESTION 164
- (Exam Topic 6)
You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on nonkey columns. What should you do?

A. Use Cloud SQL for storag

B. Add secondary indexes to support query patterns.
C. Use Cloud SQL for storag
D. Use Cloud Dataflow to transform data to support query patterns.
E. Use Cloud Spanner for storag
F. Add secondary indexes to support query patterns.
G. Use Cloud Spanner for storag
H. Use Cloud Dataflow to transform data to support query patterns.

Answer: D

Explanation:

Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

NEW QUESTION 167
- (Exam Topic 6)
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

A. Use gcloud kms keys create to create a symmetric ke
B. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
C. Use gcloud kms keys create to create a symmetric ke
D. Then use gcloud kms encrypt to encrypt each archival file with the ke
E. Use gsutil cp to upload each encrypted file to the Cloud Storage bucke
F. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.
G. Specify customer-supplied encryption key (CSEK) in the .boto configuration fil
H. Use gsutil cp to upload each archival file to the Cloud Storage bucke
I. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
J. Specify customer-supplied encryption key (CSEK) in the .boto configuration fil
K. Use gsutil cp to upload each archival file to the Cloud Storage bucke
L. Save the CSEK in a different project that only the security team can access.

Answer: B

NEW QUESTION 169
......

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

Professional-Data-Engineer Practice Exam Features:

* Professional-Data-Engineer Questions and Answers Updated Frequently

* Professional-Data-Engineer Practice Questions Verified by Expert Senior Certified Staff

* Professional-Data-Engineer Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* Professional-Data-Engineer Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
Order The Professional-Data-Engineer Practice Test Here