

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 characteristics 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 weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and serves millions of users. How should you design the frontend to respond to a database failure?

- A. Issue a command to restart the database servers.
- B. Retry the query with exponential backoff, up to a cap of 15 minutes.
- C. Retry the query every second until it comes back online to minimize staleness of data.
- D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:

<https://cloud.google.com/sql/docs/mysql/manage-connections#backoff>

NEW QUESTION 3

- (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 4

- (Exam Topic 1)

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

- A. Use a row key of the form <timestamp>.
- B. Use a row key of the form <sensorid>.
- C. Use a row key of the form <timestamp>#<sensorid>.
- D. Use a row key of the form >#<sensorid>#<timestamp>.

Answer: A

NEW QUESTION 5

- (Exam Topic 1)

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- A. Use Google Stackdriver Audit Logs to review data access.
- B. Get the identity and access management (IAM) policy of each table.
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Answer: A

NEW QUESTION 6

- (Exam Topic 1)

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.

Answer: B

NEW QUESTION 7

- (Exam Topic 1)

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

- A. Update the current pipeline and use the drain flag.
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Answer: D

NEW QUESTION 8

- (Exam Topic 1)

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.

Which Google database service should you use?

- A. Cloud SQL
- B. BigQuery
- C. Cloud Bigtable
- D. Cloud Datastore

Answer: A

NEW QUESTION 9

- (Exam Topic 1)

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

- A. Grant the consultant the Viewer role on the project.
- B. Grant the consultant the Cloud Dataflow Developer role on the project.
- C. Create a service account and allow the consultant to log on with it.
- D. Create an anonymized sample of the data for the consultant to work with in a different project.

Answer: C

NEW QUESTION 10

- (Exam Topic 1)

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

- A. Add a node to the MySQL cluster and build an OLAP cube there.
- B. Use an ETL tool to load the data from MySQL into Google BigQuery.
- C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
- D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Answer: C

NEW QUESTION 10

- (Exam Topic 1)

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

- A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type
- B. Reload the data.
- C. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row
- D. Reference the column TS instead of the column DT from now on.
- E. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP value
- F. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
- G. Add two columns to the table CLICK_STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type

- H. Reload all data in append mod
- I. For each appended row, set the value of IS_NEW to tru
- J. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.
- K. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP value
- L. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP typ
- M. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now o
- N. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

NEW QUESTION 12

- (Exam Topic 1)

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

- A. The message body for the sensor event is too large.
- B. Your custom endpoint has an out-of-date SSL certificate.
- C. The Cloud Pub/Sub topic has too many messages published to it.
- D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

Answer: B

NEW QUESTION 15

- (Exam Topic 1)

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
- C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
- D. Partition the table into smaller tables, with one for each clini
- E. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: C

NEW QUESTION 19

- (Exam Topic 1)

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

- A. Assign global unique identifiers (GUID) to each data entry.
- B. Compute the hash value of each data entry, and compare it with all historical data.
- C. Store each data entry as the primary key in a separate database and apply an index.
- D. Maintain a database table to store the hash value and other metadata for each data entry.

Answer: D

NEW QUESTION 20

- (Exam Topic 1)

Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

- A. Threading
- B. Serialization
- C. Dropout Methods
- D. Dimensionality Reduction

Answer: C

Explanation:

Reference

<https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505>

NEW QUESTION 25

- (Exam Topic 1)

You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time. What should you do?

- A. Send the data to Google Cloud Datastore and then export to BigQuery.
- B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
- C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
- D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Answer: B

NEW QUESTION 27

- (Exam Topic 2)

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

- A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
- B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub.
- C. Use the NOW () function in BigQuery to record the event's time.
- D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Answer: B

NEW QUESTION 28

- (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 30

- (Exam Topic 3)

You need to compose visualizations for operations teams with the following requirements: Which approach meets the requirements?

- A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
- B. 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.
- C. 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.
- D. 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.

Answer: C

NEW QUESTION 32

- (Exam Topic 3)

MJTelco is building a custom interface to share data. They have these requirements:

- They need to do aggregations over their petabyte-scale datasets.
- They need to scan specific time range rows with a very fast response time (milliseconds). Which combination of Google Cloud Platform products should you recommend?

- A. Cloud Datastore and Cloud Bigtable
- B. Cloud Bigtable and Cloud SQL
- C. BigQuery and Cloud Bigtable
- D. BigQuery and Cloud Storage

Answer: C

NEW QUESTION 36

- (Exam Topic 3)

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

- A. Create a table called tracking_table and include a DATE column.
- B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
- C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
- D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Answer: B

NEW QUESTION 38

- (Exam Topic 4)

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: BDF

NEW QUESTION 43

- (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 44

- (Exam Topic 4)

You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

- > The user profile: What the user likes and doesn't like to eat
- > The user account information: Name, address, preferred meal times
- > The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

- A. BigQuery
- B. Cloud SQL
- C. Cloud Bigtable
- D. Cloud Datastore

Answer: A

NEW QUESTION 47

- (Exam Topic 4)

You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery
- C. Store the data in Google Cloud Datastor
- D. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
- E. Store the data in a file in a regional Google Cloud Storage bucke
- F. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Answer: A

NEW QUESTION 50

- (Exam Topic 5)

The Dataflow SDKs have been recently transitioned into which Apache service?

- A. Apache Spark
- B. Apache Hadoop
- C. Apache Kafka
- D. Apache Beam

Answer: D

Explanation:

Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive Reference: <https://cloud.google.com/dataflow/docs/>

NEW QUESTION 51

- (Exam Topic 5)

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

- A. BETWEEN
- B. WHERE
- C. SELECT
- D. LIMIT

Answer: C

Explanation:

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

NEW QUESTION 54

- (Exam Topic 5)

Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

- A. An hourly watermark
- B. An event time trigger
- C. The with Allowed Lateness method
- D. A processing time trigger

Answer: D

Explanation:

When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.

Processing time triggers. These triggers operate on the processing time – the time when the data element is processed at any given stage in the pipeline.

Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.

Reference: <https://beam.apache.org/documentation/programming-guide/#triggers>

NEW QUESTION 56

- (Exam Topic 5)

You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

- A. PCollection
- B. Transform
- C. Pipeline
- D. Sink API

Answer: B

Explanation:

In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and

other complex programming structure to create a branching pipeline.
Reference: <https://cloud.google.com/dataflow/model/programming-model>

NEW QUESTION 58

- (Exam Topic 5)

What are two of the characteristics of using online prediction rather than batch prediction?

- A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
- B. Predictions are returned in the response message.
- C. Predictions are written to output files in a Cloud Storage location that you specify.
- D. It is optimized to minimize the latency of serving predictions.

Answer: BD

Explanation:

Online prediction

Optimized to minimize the latency of serving predictions.

Predictions returned in the response message. Batch prediction

Optimized to handle a high volume of instances in a job and to run more complex models. Predictions written to output files in a Cloud Storage location that you specify.

Reference:

https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION 62

- (Exam Topic 5)

How can you get a neural network to learn about relationships between categories in a categorical feature?

- A. Create a multi-hot column
- B. Create a one-hot column
- C. Create a hash bucket
- D. Create an embedding column

Answer: D

Explanation:

There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.

Both of these problems can be solved by representing a categorical feature with an embedding

column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).

You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.

Reference:

<https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-dee>

NEW QUESTION 63

- (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 64

- (Exam Topic 5)

The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

- A. Cloud Dataflow connector
- B. DataFlow SDK
- C. BigQuery API
- D. BigQuery Data Transfer Service

Answer: A

Explanation:

The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.

Reference: <https://cloud.google.com/bigtable/docs/dataflow-hbase>

NEW QUESTION 66

- (Exam Topic 5)

You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.

Tom,555 X street Tim,553 Y street Sam, 111 Z street

Which operation is best suited for the above data processing requirement?

- A. ParDo
- B. Sink API
- C. Source API
- D. Data extraction

Answer: A

Explanation:

In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.

Reference: <https://cloud.google.com/dataflow/model/par-do>

NEW QUESTION 71

- (Exam Topic 5)

Which TensorFlow function can you use to configure a categorical column if you don't know all of the possible values for that column?

- A. categorical_column_with_vocabulary_list
- B. categorical_column_with_hash_bucket
- C. categorical_column_with_unknown_values
- D. sparse_column_with_keys

Answer: B

Explanation:

If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an auto-incremental ID starting from 0.

What if we don't know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column occupation will be hashed to an integer ID as we encounter them in training.

Reference: <https://www.tensorflow.org/tutorials/wide>

NEW QUESTION 72

- (Exam Topic 5)

What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?

- A. Sessions
- B. OutputCriteria
- C. Windows
- D. Triggers

Answer: D

Explanation:

Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Tri>

NEW QUESTION 75

- (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 77

- (Exam Topic 5)

To give a user read permission for only the first three columns of a table, which access control method would you use?

- A. Primitive role
- B. Predefined role
- C. Authorized view
- D. It's not possible to give access to only the first three columns of a table.

Answer: C

Explanation:

An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view.

When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.

Reference: <https://cloud.google.com/bigquery/docs/views#authorized-views>

NEW QUESTION 82

- (Exam Topic 5)

Which of the following is not possible using primitive roles?

- A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
- B. Give UserA owner access and UserB editor access for all datasets in a project.
- C. Give a user access to view all datasets in a project, but not run queries on them.
- D. Give GroupA owner access and GroupB editor access for all datasets in a project.

Answer: C

Explanation:

Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.

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

NEW QUESTION 83

- (Exam Topic 5)

When you design a Google Cloud Bigtable schema it is recommended that you .

- A. Avoid schema designs that are based on NoSQL concepts
- B. Create schema designs that are based on a relational database design
- C. Avoid schema designs that require atomicity across rows
- D. Create schema designs that require atomicity across rows

Answer: C

Explanation:

All operations are atomic at the row level. For example, if you update two rows in a table, it's possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.

Reference: <https://cloud.google.com/bigtable/docs/schema-design#row-keys>

NEW QUESTION 87

- (Exam Topic 5)

Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?

- A. Preemptible workers cannot use persistent disk.
- B. Preemptible workers cannot store data.
- C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
- D. A Dataproc cluster cannot have only preemptible workers.

Answer: BD

Explanation:

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster: Processing only—Since preemptibles can be reclaimed at any time, preemptible workers do not store data.

Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.

No preemptible-only clusters—To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters.

Persistent disk size—As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS.

The managed group automatically re-adds workers lost due to reclamation as capacity permits. Reference:

<https://cloud.google.com/dataproc/docs/concepts/preemptible-vms>

NEW QUESTION 89

- (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 90

- (Exam Topic 5)

Which of the following is NOT one of the three main types of triggers that Dataflow supports?

- A. Trigger based on element size in bytes
- B. Trigger that is a combination of other triggers
- C. Trigger based on element count
- D. Trigger based on time

Answer: A

Explanation:

There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way

Reference: <https://cloud.google.com/dataflow/model/triggers>

NEW QUESTION 92

- (Exam Topic 5)

Which Java SDK class can you use to run your Dataflow programs locally?

- A. LocalRunner
- B. DirectPipelineRunner
- C. MachineRunner
- D. LocalPipelineRunner

Answer: B

Explanation:

DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun>

NEW QUESTION 93

- (Exam Topic 5)

Which of these sources can you not load data into BigQuery from?

- A. File upload
- B. Google Drive
- C. Google Cloud Storage
- D. Google Cloud SQL

Answer: D

Explanation:

You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.

Reference: <https://cloud.google.com/bigquery/loading-data>

NEW QUESTION 96

- (Exam Topic 5)

Which of the following are feature engineering techniques? (Select 2 answers)

- A. Hidden feature layers
- B. Feature prioritization
- C. Crossed feature columns
- D. Bucketization of a continuous feature

Answer: CD

Explanation:

Selecting and crafting the right set of feature columns is key to learning an effective model.

Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.

Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.

Reference: https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION 99

- (Exam Topic 5)

To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

- A. gcloud ml-engine local train
- B. gcloud ml-engine jobs submit training
- C. gcloud ml-engine jobs submit training local
- D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .

Answer: A

Explanation:

gcloud ml-engine local train - run a Cloud ML Engine training job locally

This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.

This is especially useful in the case of testing distributed models, as it allows you to validate that you are properly interacting with the Cloud ML Engine cluster configuration.

Reference: <https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train>

NEW QUESTION 104

- (Exam Topic 5)

Does Dataflow process batch data pipelines or streaming data pipelines?

- A. Only Batch Data Pipelines
- B. Both Batch and Streaming Data Pipelines
- C. Only Streaming Data Pipelines
- D. None of the above

Answer: B

Explanation:

Dataflow is a unified processing model, and can execute both streaming and batch data pipelines Reference: <https://cloud.google.com/dataflow/>

NEW QUESTION 105

- (Exam Topic 5)

Which of these are examples of a value in a sparse vector? (Select 2 answers.)

- A. [0, 5, 0, 0, 0, 0]
- B. [0, 0, 0, 1, 0, 0, 1]
- C. [0, 1]
- D. [1, 0, 0, 0, 0, 0, 0]

Answer: CD

Explanation:

Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).

[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1. [0, 5, 0, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s. Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION 107

- (Exam Topic 5)

All Google Cloud Bigtable client requests go through a front-end server they are sent to a Cloud Bigtable node.

- A. before
- B. after
- C. only if
- D. once

Answer: A

Explanation:

In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.

The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.

When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.

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

NEW QUESTION 109

- (Exam Topic 5)

You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output.

Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

- A. Cancel
- B. Drain
- C. Stop
- D. Finish

Answer: B

Explanation:

Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.

Reference: <https://cloud.google.com/dataflow/pipelines/stopping-a-pipeline>

NEW QUESTION 110

- (Exam Topic 5)

Which of the following is NOT true about Dataflow pipelines?

- A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner
- B. Dataflow pipelines can consume data from other Google Cloud services
- C. Dataflow pipelines can be programmed in Java
- D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources

Answer: A

Explanation:

Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs

Reference: <https://cloud.google.com/dataflow/>

NEW QUESTION 113

- (Exam Topic 5)

When a Cloud Bigtable node fails, is lost.

- A. all data
- B. no data
- C. the last transaction
- D. the time dimension

Answer: B

Explanation:

A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.

Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:

Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud

Bigtable simply updates the pointers for each node.

Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.

When a Cloud Bigtable node fails, no data is lost Reference: <https://cloud.google.com/bigtable/docs/overview>

NEW QUESTION 114

- (Exam Topic 5)

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

- A. Storage, queries, and streaming inserts
- B. Storage, queries, and loading data from a file
- C. Storage, queries, and exporting data
- D. Queries and streaming inserts

Answer: A

Explanation:

Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.

Reference: <https://cloud.google.com/bigquery/pricing>

NEW QUESTION 116

- (Exam Topic 5)

Which of the following is not true about Dataflow pipelines?

- A. Pipelines are a set of operations
- B. Pipelines represent a data processing job
- C. Pipelines represent a directed graph of steps
- D. Pipelines can share data between instances

Answer: D

Explanation:

The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms

Reference: <https://cloud.google.com/dataflow/model/pipelines>

NEW QUESTION 118

- (Exam Topic 5)

Cloud Dataproc is a managed Apache Hadoop and Apache service.

- A. Blaze
- B. Spark
- C. Fire

D. Ignite

Answer: B

Explanation:

Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.

Reference: <https://cloud.google.com/dataproc/docs/>

NEW QUESTION 122

- (Exam Topic 5)

If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

- A. Do not use a production instance.
- B. Run your test for at least 10 minutes.
- C. Before you test, run a heavy pre-test for several minutes.
- D. Use at least 300 GB of data.

Answer: A

Explanation:

If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:

Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.

Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.

Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.

Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

Reference: <https://cloud.google.com/bigtable/docs/performance>

NEW QUESTION 125

- (Exam Topic 5)

When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

- A. Your gcloud does not have access to the BigQuery resources
- B. BigQuery cannot be accessed from local machines
- C. You are missing gcloud on your machine
- D. Pipelines cannot be run locally

Answer: A

Explanation:

When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink

Reference:

<https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun>

NEW QUESTION 126

- (Exam Topic 5)

Cloud Bigtable is a recommended option for storing very large amounts of ____?

- A. multi-keyed data with very high latency
- B. multi-keyed data with very low latency
- C. single-keyed data with very low latency
- D. single-keyed data with very high latency

Answer: C

Explanation:

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. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

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

NEW QUESTION 128

- (Exam Topic 5)

Which Google Cloud Platform service is an alternative to Hadoop with Hive?

- A. Cloud Dataflow
- B. Cloud Bigtable
- C. BigQuery
- D. Cloud Datastore

Answer: C

Explanation:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis. Google BigQuery is an enterprise data warehouse. Reference: https://en.wikipedia.org/wiki/Apache_Hive

NEW QUESTION 133

- (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 138

- (Exam Topic 5)

Which methods can be used to reduce the number of rows processed by BigQuery?

- A. Splitting tables into multiple tables; putting data in partitions
- B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
- C. Putting data in partitions; using the LIMIT clause
- D. Splitting tables into multiple tables; using the LIMIT clause

Answer: A

Explanation:

If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.

If you use the LIMIT clause, BigQuery will still process the entire table. Reference: <https://cloud.google.com/bigquery/docs/partitioned-tables>

NEW QUESTION 142

- (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 144

- (Exam Topic 6)

An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application.

They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?

- A. BigQuery
- B. Cloud SQL
- C. Cloud BigTable
- D. Cloud Datastore

Answer: C

Explanation:

Reference: <https://cloud.google.com/solutions/business-intelligence/>

NEW QUESTION 146

- (Exam Topic 6)

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

- A. Publisher throughput quota is too small.
- B. Total outstanding messages exceed the 10-MB maximum.
- C. Error handling in the subscriber code is not handling run-time errors properly.
- D. The subscriber code cannot keep up with the messages.
- E. The subscriber code does not acknowledge the messages that it pulls.

Answer: CD

NEW QUESTION 151

- (Exam Topic 6)

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

- A. Increase the CPU size on your server.
- B. Increase the size of the Google Persistent Disk on your server.
- C. Increase your network bandwidth from your datacenter to GCP.
- D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Answer: C

NEW QUESTION 156

- (Exam Topic 6)

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

- The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured
- Support for publish/subscribe semantics on hundreds of topics
- Retain per-key ordering

Which system should you choose?

- A. Apache Kafka
- B. Cloud Storage
- C. Cloud Pub/Sub
- D. Firebase Cloud Messaging

Answer: A

NEW QUESTION 158

- (Exam Topic 6)

You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

- Decoupling producer from consumer
- Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
- Near real-time SQL query
- Maintain at least 2 years of historical data, which will be queried with SQ Which pipeline should you use to meet these requirements?

- A. Create an application that provides an AP
- B. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
- C. Create an application that writes to a Cloud SQL database to store the dat
- D. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
- E. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
- F. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

Answer: A

NEW QUESTION 162

- (Exam Topic 6)

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

- A. Export the records from the database as an Avro fil
- B. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- C. Export the records from the database as an Avro fil
- D. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- E. Export the records from the database into a CSV fil
- F. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storag
- G. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
- H. Export the records from the database as an Avro fil
- I. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storag
- J. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Answer: A

NEW QUESTION 164

- (Exam Topic 6)

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

- A. Cloud Spanner
- B. Cloud Bigtable
- C. Cloud Firestore
- D. Cloud SQL

Answer: D

NEW QUESTION 166

- (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 table.
- 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 167

- (Exam Topic 6)

You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increase the size of vocabularies or n-grams used.

Answer: D

NEW QUESTION 168

- (Exam Topic 6)

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

- A. Perform hyperparameter tuning
- B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
- C. Deploy the model and measure the real-world AUC; it's always higher because of generalization
- D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

Answer: A

Explanation:

<https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568>

NEW QUESTION 173

- (Exam Topic 6)

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

- A. Modify the transform MapReduce jobs to apply sensor calibration before they do anything else.
- B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
- C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
- D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Answer: A

NEW QUESTION 178

- (Exam Topic 6)

Government regulations in the banking industry mandate the protection of client's personally identifiable information (PII). Your company requires PII to be access controlled, encrypted, and compliant with major data protection standards. In addition to using Cloud Data Loss Prevention (Cloud DLP), you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?

- A. Assign the required identity and Access Management (IAM) roles to every employee, and create a single service account to access protected resources
- B. Use one service account to access a Cloud SQL database and use separate service accounts for each human user
- C. Use Cloud Storage to comply with major data protection standard
- D. Use one service account shared by all users
- E. Use Cloud Storage to comply with major data protection standard
- F. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group

Answer: D

NEW QUESTION 180

- (Exam Topic 6)

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

Answer: C

Explanation:

<https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml>

NEW QUESTION 182

- (Exam Topic 6)

Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS) in a colocation facility. The cluster is designed to support peak usage on the system, however, many jobs are batch in nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more server offerings in order to take advantage of the cloud. Because of the timing of their contract renewal with the colocation facility they have only 2 months for their initial migration. How should you recommend they approach their upcoming migration strategy so they can maximize their cost savings in the cloud while still executing the migration in time?

- A. Migrate the workloads to Dataproc plus HOPS, modernize later
- B. Migrate the workloads to Dataproc plus Cloud Storage modernize later
- C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery
- D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery

Answer: D

NEW QUESTION 184

- (Exam Topic 6)

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

- A. Create an API using App Engine to receive and send messages to the applications
- B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them
- C. Create a table on Cloud SQL, and insert and delete rows with the job information
- D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Answer: A

NEW QUESTION 186

- (Exam Topic 6)

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

- A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
- B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
- C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
- D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API
- E. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Answer: D

NEW QUESTION 187

- (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 data
- 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 data
- 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 189

- (Exam Topic 6)

You need to choose a database for a new project that has the following requirements:

- > Fully managed
- > Able to automatically scale up
- > Transactionally consistent
- > Able to scale up to 6 TB
- > Able to be queried using SQL Which database do you choose?

- A. Cloud SQL
- B. Cloud Bigtable
- C. Cloud Spanner
- D. Cloud Datastore

Answer: C

NEW QUESTION 191

- (Exam Topic 6)

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products or features of the platform. What should you do?

- A. Export the information to Cloud Stackdriver, and set up an Alerting policy
- B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
- C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs
- D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Answer: B

NEW QUESTION 193

- (Exam Topic 6)

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.

Answer: C

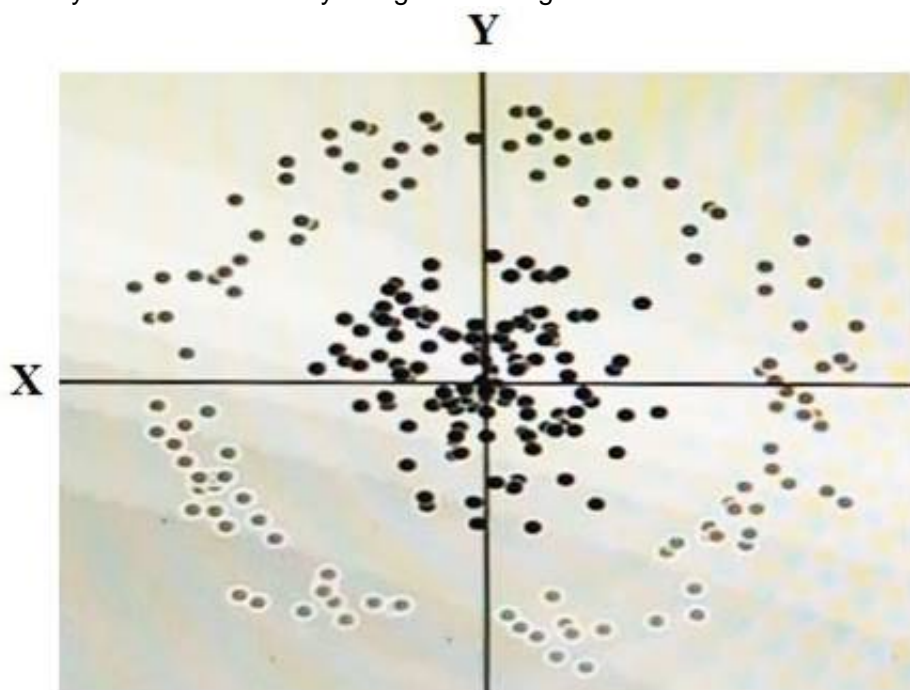
Explanation:

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

NEW QUESTION 196

- (Exam Topic 6)

You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm.



To do this you need to add a synthetic feature. What should the value of that feature be?

- A. $X^2 + Y^2$
- B. X^2
- C. Y^2
- D. $\cos(X)$

Answer: D

NEW QUESTION 198

- (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 202

- (Exam Topic 6)

Your company is migrating its on-premises data warehousing solution to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply daily updates from transactional database sources. Your company wants to use BigQuery to improve its handling of CDC and to optimize the performance of the data warehouse. Source system changes must be available for query in near-real time using trigger-based CDC streams. You need to ensure that changes in the BigQuery reporting table are available with minimal latency and reduced overhead. What should you do? Choose 2 answers.

- A. Perform a DML INSERT UPDATE, or DELETE to replicate each CDC record in the reporting table in real time.
- B. Periodically DELETE outdated records from the reporting table. Periodically use a DML MERGE to simultaneously perform DML INSERT, UPDATE, and DELETE operations in the reporting table.
- C. UPDATE, and DELETE operations in the reporting table.
- D. Insert each new CDC record and corresponding operation type into a staging table in real time.
- E. Insert each new CDC record and corresponding operation type into the reporting table in real time and use a materialized view to expose only the current version of each unique record.

Answer: BD

NEW QUESTION 206

- (Exam Topic 6)

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

- A. Use Cloud Dataproc to run your transformation.
- B. Monitor CPU utilization for the cluster.
- C. Resize the number of worker nodes in your cluster via the command line.
- D. Use Cloud Dataproc to run your transformation.
- E. Use the `diagnose` command to generate an operational output archive.
- F. Locate the bottleneck and adjust cluster resources.
- G. Use Cloud Dataflow to run your transformation.
- H. Monitor the job system lag with Stackdriver.
- I. Use the default autoscaling setting for worker instances.
- J. Use Cloud Dataflow to run your transformation.
- K. Monitor the total execution time for a sampling of jobs.
- L. Configure the job to use non-default Compute Engine machine types when needed.

Answer: B

NEW QUESTION 210

- (Exam Topic 6)

You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

- A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master.
- B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet.
- C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter.
- D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role.

Answer: D

NEW QUESTION 211

- (Exam Topic 6)

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

- A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
- B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
- C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Function.
- D. Integrate the package tracking applications with this function.
- E. Use TensorFlow to create a model that is trained on your corpus of images.
- F. Create a Python notebook in Cloud DataLab that uses this model so you can analyze for damaged packages.

Answer: A

NEW QUESTION 213

- (Exam Topic 6)

A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete training. 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 215

- (Exam Topic 6)

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

Answer: C

NEW QUESTION 219

- (Exam Topic 6)

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query --dry-run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID selects a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

- A. Create a separate table for each ID.
- B. Use the LIMIT keyword to reduce the number of rows returned.
- C. Recreate the table with a partitioning column and clustering column.
- D. Use the bq query --maximum_bytes_billed flag to restrict the number of bytes billed.

Answer: C

NEW QUESTION 224

- (Exam Topic 6)

You are migrating an application that tracks library books and information about each book, such as author or year published, from an on-premises data warehouse to BigQuery. In your current relational database, the author information is kept in a separate table and joined to the book information on a common key. Based on Google's recommended practice for schema design, how would you structure the data to ensure optimal speed of queries about the author of each book that has been borrowed?

- A. Keep the schema the same, maintain the different tables for the book and each of the attributes, and query as you are doing today
- B. Create a table that is wide and includes a column for each attribute, including the author's first name, last name, date of birth, etc
- C. Create a table that includes information about the books and authors, but nest the author fields inside the author column
- D. Keep the schema the same, create a view that joins all of the tables, and always query the view

Answer: C

NEW QUESTION 226

- (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 230

- (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 231

- (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 235

- (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 236

- (Exam Topic 6)

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

- A. Cloud Speech-to-Text API
- B. Cloud Natural Language API
- C. Dialogflow Enterprise Edition
- D. Cloud AutoML Natural Language

Answer: C

NEW QUESTION 237

- (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 242

- (Exam Topic 6)

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SidelInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

- A. Switch to compressed Avro files
- B. Reduce the batch size
- C. Retry records that throw an error
- D. Use CoGroupByKey instead of the SidelInput

Answer: B

NEW QUESTION 247

- (Exam Topic 6)

You are using BigQuery and Data Studio to design a customer-facing dashboard that displays large quantities of aggregated data. You expect a high volume of concurrent users. You need to optimize tie dashboard to provide quick visualizations with minimal latency. What should you do?

- A. Use BigQuery BI Engine with materialized views
- B. Use BigQuery BI Engine with streaming data.
- C. Use BigQuery BI Engine with authorized views
- D. Use BigQuery BI Engine with logical reviews

Answer: B

NEW QUESTION 249

- (Exam Topic 6)

Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

- A. Cloud Bigtable
- B. Google BigQuery
- C. Google Cloud Storage
- D. Google Cloud Datastore

Answer: A

Explanation:

Reference: <https://cloud.google.com/bigtable/docs/schema-design-time-series>

NEW QUESTION 251

- (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 255

- (Exam Topic 6)

You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:

- Real-time event stream
- ANSI SQL access to real-time stream and historical data
- Batch historical exports

Which solution should you use?

- A. Cloud Dataflow, Cloud SQL, Cloud Spanner
- B. Cloud Pub/Sub, Cloud Storage, BigQuery
- C. Cloud Dataproc, Cloud Dataflow, BigQuery
- D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

Answer: A

NEW QUESTION 256

- (Exam Topic 6)

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

- A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
- B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
- C. Use the BigQuery streaming the stream changes into a daily inventory movement tabl
- D. Calculate balances in a view that joins it to the historical inventory balance tabl
- E. Update the inventory balance table nightly.
- F. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance tabl
- G. Update the inventory balance table nightly.

Answer: A

NEW QUESTION 258

- (Exam Topic 6)

An aerospace company uses a proprietary data format to store its night data. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?

- A. Use a standard Dataflow pipeline to store the raw data m BigQuery and then transform the format later when the data is used
- B. Write a she script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source
- C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
- D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

Answer: D

NEW QUESTION 261

- (Exam Topic 6)

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

- A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDF
- B. Mount the Hive tables locally.
- C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluste
- D. Mount the Hive tables locally.
- E. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluste
- F. Then run the Hadoop utility to copy them do HDF
- G. Mount the Hive tables from HDFS.
- H. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive table
- I. Replicate external Hive tables to the native ones.
- J. Load the ORC files into BigQuer
- K. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive table
- L. Replicate external Hive tables to the native ones.

Answer: BC

NEW QUESTION 264

- (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 266

- (Exam Topic 6)

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 Dataflo
- B. Use BigQuery for storage and query.
- C. Transform text files to compressed Avro using Cloud Dataflo
- D. Use Cloud Storage and BigQuery permanent linked tables for query.
- E. Compress text files to gzip using the Grid Computing Tool
- F. Use BigQuery for storage and query.
- G. Compress text files to gzip using the Grid Computing Tool
- H. Use Cloud Storage, and then import into Cloud Bigtable for query.

Answer: D

NEW QUESTION 270

- (Exam Topic 6)

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.

What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- C. Create a service account and grant dataset access to that accoun
- D. Use the service account's private key to access the dataset
- E. Create a dummy user and grant dataset access to that use
- F. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

Answer: C

NEW QUESTION 275

- (Exam Topic 6)

You are designing a cloud-native historical data processing system to meet the following conditions:

- The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
- A streaming data pipeline stores new data daily.
- Peformance is not a factor in the solution.

➤ The solution design should maximize availability.
How should you design data storage for this solution?

- A. Create a Cloud Dataproc cluster with high availability
- B. Store the data in HDFS, and perform analysis as needed.
- C. Store the data in BigQuery
- D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- E. Store the data in a regional Cloud Storage bucket
- F. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- G. Store the data in a multi-regional Cloud Storage bucket
- H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: D

NEW QUESTION 278

- (Exam Topic 6)

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

- A. Cloud Dataflow
- B. Cloud Composer
- C. Cloud Dataprep
- D. Cloud Dataproc

Answer: D

NEW QUESTION 283

- (Exam Topic 6)

You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

- A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc
- B. Call the model from your application.
- C. Build and train a classification model with Spark MLlib to generate labels
- D. Build and train a second classification model with Spark MLlib to filter results to match customer preference
- E. Deploy the models using Cloud Dataproc
- F. Call the models from your application.
- G. Build an application that calls the Cloud Video Intelligence API to generate labels
- H. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
- I. Build an application that calls the Cloud Video Intelligence API to generate labels
- J. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Answer: C

NEW QUESTION 285

- (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 custom ops.
- C. Use Cloud GPUs after implementing GPU kernel support for your custom ops.
- D. Stay on CPUs, and increase the size of the cluster you're training your model on.

Answer: B

NEW QUESTION 287

- (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 filtered 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 292

- (Exam Topic 6)

You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

- A. Use the TABLE_DATE_RANGE function
- B. Use the WHERE_PARTITIONTIME pseudo column
- C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
- D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD

Answer: A

Explanation:

Reference:

<https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-ap>

NEW QUESTION 295

- (Exam Topic 6)

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery. How should you securely run this workload?

- A. Restrict the Google Cloud Storage bucket so only you can see the files
- B. Grant the Project Owner role to a service account, and run the job with it
- C. Use a service account with the ability to read the batch files and to write to BigQuery
- D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Answer: B

NEW QUESTION 297

- (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 302

- (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 306

- (Exam Topic 6)

You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20 TB in size. Which database should you choose?

- A. Cloud SQL
- B. Cloud Bigtable
- C. Cloud Spanner
- D. Cloud Datastore

Answer: A

NEW QUESTION 311

- (Exam Topic 6)

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

NEW QUESTION 313

- (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 from 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 from PubSub and write to GCS.

Answer: A

NEW QUESTION 318

- (Exam Topic 6)

You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?

- A. Deploy a Cloud Dataproc cluste
- B. Use a standard persistent disk and 50% preemptible worker
- C. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
- D. Deploy a Cloud Dataproc cluste
- E. Use an SSD persistent disk and 50% preemptible worker
- F. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
- G. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instance
- H. Install the Cloud Storage connector, and store the data in Cloud Storag
- I. Change references in scripts from hdfs:// to gs://
- J. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances.Store data in HDF
- K. Change references in scripts from hdfs:// to gs://

Answer: A

NEW QUESTION 322

- (Exam Topic 6)

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

- A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.
- B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
- C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
- D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Answer: D

NEW QUESTION 326

- (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 330

- (Exam Topic 6)

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

- A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size
- B. Increase the see of your parquet files to ensure them to be 1 GB minimum
- C. Switch to TFRecords format (appr 200 MB per We) instead of parquet files
- D. Switch from HDDs to SSD
- E. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

Answer: A

NEW QUESTION 331

- (Exam Topic 6)

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

- A. Edge TPUs as sensor devices for storing and transmitting the messages.
- B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
- C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
- D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

Answer: C

NEW QUESTION 334

.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questions 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](#)