



Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

NEW QUESTION 1

- (Exam Topic 1)

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

- A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
- B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
- C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
- D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: B

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 server 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)

You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

- No interaction by the user on the site for 1 hour
- Has added more than \$30 worth of products to the basket
- Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- A. Use a fixed-time window with a duration of 60 minutes.
- B. Use a sliding time window with a duration of 60 minutes.
- C. Use a session window with a gap time duration of 60 minutes.
- D. Use a global window with a time based trigger with a delay of 60 minutes.

Answer: C

NEW QUESTION 4

- (Exam Topic 1)

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

- A. Continuously retrain the model on just the new data.
- B. Continuously retrain the model on a combination of existing data and the new data.
- C. Train on the existing data while using the new data as your test set.
- D. Train on the new data while using the existing data as your test set.

Answer: C

Explanation:

<https://cloud.google.com/automl-tables/docs/prepare>

NEW QUESTION 5

- (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 6

- (Exam Topic 1)

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

- A. Load data into different partitions.
- B. Load data into a different dataset for each client.
- C. Put each client's BigQuery dataset into a different table.
- D. Restrict a client's dataset to approved users.
- E. Only allow a service account to access the datasets.
- F. Use the appropriate identity and access management (IAM) roles for each client's users.

Answer: BDF

NEW QUESTION 7

- (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 8

- (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 9

- (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 mode
- I. For each appended row, set the value of IS_NEW to true
- 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 type
- M. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on
- N. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

NEW QUESTION 10

- (Exam Topic 1)

An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

- A. Use federated data sources, and check data in the SQL query.
- B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
- C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
- D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Answer: D

NEW QUESTION 15

- (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 19

- (Exam Topic 1)

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.

The data scientists have written the following code to read the data for a new key features in the logs. BigQueryIO.Read

```
.named("ReadLogData")
```

```
.from("clouddataflow-readonly:samples.log_data")
```

You want to improve the performance of this data read. What should you do?

- A. Specify the TableReference object in the code.
- B. Use .fromQuery operation to read specific fields from the table.
- C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
- D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table.

Answer: D

NEW QUESTION 21

- (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 23

- (Exam Topic 4)

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

- A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
- B. Add a new column called FullName to the Users table
- C. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- D. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
- E. Use BigQuery to export the data for the table to a CSV file
- F. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName
- G. Run a BigQuery load job to load the new CSV file into BigQuery.

Answer: C

NEW QUESTION 24

- (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 27

- (Exam Topic 5)

Which of these is NOT a way to customize the software on Dataproc cluster instances?

- A. Set initialization actions
- B. Modify configuration files using cluster properties
- C. Configure the cluster using Cloud Deployment Manager
- D. Log into the master node and make changes from there

Answer: C

Explanation:

You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.

You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster.

When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>]

Reference: <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties>

NEW QUESTION 30

- (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 33

- (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 36

- (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 39

- (Exam Topic 5)

What are the minimum permissions needed for a service account used with Google Dataproc?

- A. Execute to Google Cloud Storage; write to Google Cloud Logging
- B. Write to Google Cloud Storage; read to Google Cloud Logging
- C. Execute to Google Cloud Storage; execute to Google Cloud Logging
- D. Read and write to Google Cloud Storage; write to Google Cloud Logging

Answer: D

Explanation:

Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging.

Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION 41

- (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 46

- (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 48

- (Exam Topic 5)

Why do you need to split a machine learning dataset into training data and test data?

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

Answer: B

Explanation:

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

Reference: <https://machinelearningmastery.com/a-simple-intuition-for-overfitting/>

NEW QUESTION 49

- (Exam Topic 5)

When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and .

- A. zone
- B. node
- C. label
- D. type

Answer: A

Explanation:

At a minimum, you must specify four values when creating a new cluster with the `projects.regions.clusters.create` operation:

The project in which the cluster will be created

The region to use

The name of the cluster

The zone in which the cluster will be created

You can specify many more details beyond these minimum requirements. For example, you can

also specify the number of workers, whether preemptible compute should be used, and the network settings. Reference:

https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

NEW QUESTION 53

- (Exam Topic 5)

When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?

- A. 500 TB
- B. 1 GB
- C. 1 TB
- D. 500 GB

Answer: C

Explanation:

Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

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

NEW QUESTION 56

- (Exam Topic 5)

Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?

- A. Use K-means Clustering to detect faces in the pixels.
- B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
- C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
- D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: C

Explanation:

Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as “deep” learning. So deep is a strictly defined, technical term that means more than one hidden layer.

In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer’s output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.

A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.

Feature engineering is difficult to perform on raw image data.

K- means Clustering is an unsupervised learning method used to categorize unlabeled data. Reference: <https://deeplearning4j.org/neuralnet-overview>

NEW QUESTION 59

- (Exam Topic 5)

What are two of the benefits of using denormalized data structures in BigQuery?

- A. Reduces the amount of data processed, reduces the amount of storage required
- B. Increases query speed, makes queries simpler
- C. Reduces the amount of storage required, increases query speed
- D. Reduces the amount of data processed, increases query speed

Answer: B

Explanation:

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINS on large tables, but with a denormalized data

structure, you don't have to use JOINS, since all of the data has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses.

Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.

Reference:

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION 63

- (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 64

- (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 65

- (Exam Topic 5)

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

- A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
- B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
- C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
- D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

Answer: C

Explanation:

It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,

If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.

It is recommended to avoid creating your Compute Engine instance in a different region from

your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

Reference: <https://cloud.google.com/bigtable/docs/creating-compute-instance>

NEW QUESTION 66

- (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 67

- (Exam Topic 5)

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

- A. month-by-month
- B. minute-by-minute
- C. week-by-week
- D. hour-by-hour

Answer: B

Explanation:

One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.

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

NEW QUESTION 71

- (Exam Topic 5)

When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a proxy.

- A. HTTPS
- B. VPN
- C. SOCKS
- D. HTTP

Answer: C

Explanation:

When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.

Reference: <https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces>

NEW QUESTION 73

- (Exam Topic 6)

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

- A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
- B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
- C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
- D. Write an application that uses Cloud Datastore client libraries to read all the entitie
- E. Treat each entity as a BigQuery table row via BigQuery streaming inser
- F. Assign an export timestamp for each export, and attach it as an extra column for each ro
- G. Make sure that the BigQuery table is partitioned using the export timestamp column.
- H. Write an application that uses Cloud Datastore client libraries to read all the entitie
- I. Format the exported data into a JSON fil
- J. Apply compression before storing the data in Cloud Source Repositories.

Answer: CE

NEW QUESTION 77

- (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 82

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 85

- (Exam Topic 6)

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

- A. Increase the cluster size with more non-preemptible workers.
- B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
- C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
- D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

Answer: D

Explanation:

Reference <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex>

NEW QUESTION 89

- (Exam Topic 6)

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

- A. PigLatin using Pig
- B. HiveQL using Hive
- C. Java using MapReduce
- D. Python using MapReduce

Answer: D

NEW QUESTION 93

- (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 98

- (Exam Topic 6)

You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.

What should you do?

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

Answer: B

NEW QUESTION 99

- (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 104

- (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 106

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 107

- (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 109

- (Exam Topic 6)

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of your Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

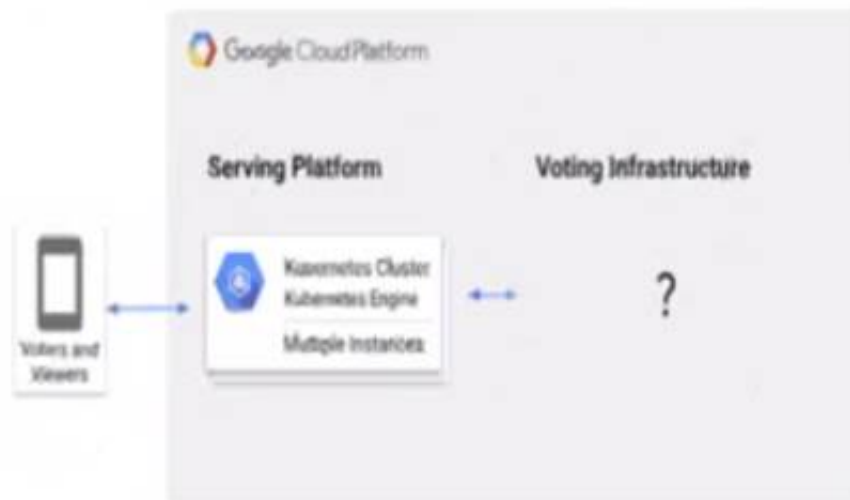
- A. Review Key Visualizer metric
- B. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- C. Review Key Visualizer metric
- D. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- E. Monitor the latency of write operation
- F. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
- G. Monitor storage utilization
- H. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
- I. Monitor latency of read operation
- J. Increase the size of the Cloud Bigtable cluster if read operations take longer than 100 ms.

Answer: AC

NEW QUESTION 112

- (Exam Topic 6)

A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and that all votes are processed. You must display partial results while voting is open. After voting ends you need to count the votes exactly once while optimizing cost. What should you do?



- A. Create a Memorystore instance with a high availability (HA) configuration
- B. Write votes to a Pub Sub topic and have Cloud Functions subscribe to it and write votes to BigQuery
- C. Write votes to a Pub/Sub topic and load into both Bigtable and BigQuery via a Dataflow pipeline. Query Bigtable for real-time results and BigQuery for later analysis. Shutdown the Bigtable instance when voting concludes.
- D. Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas

Answer: C

NEW QUESTION 116

- (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 120

- (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 123

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 125

- (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 127

- (Exam Topic 6)

You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- A. Call the Cloud Natural Language API from your application
- B. Process the generated Entity Analysis as labels.
- C. Call the Cloud Natural Language API from your application
- D. Process the generated Sentiment Analysis as labels.
- E. Build and train a text classification model using TensorFlow
- F. Deploy the model using Cloud Machine Learning Engine
- G. Call the model from your application and process the results as labels.
- H. Build and train a text classification model using TensorFlow
- I. Deploy the model using a Kubernetes Engine cluster
- J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 130

- (Exam Topic 6)

You need to give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID. This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline. There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system. How should you design your pipeline to minimize that backpressure?

- A. Call out to the service via HTTP
- B. Create the pipeline statically in the class definition
- C. Create a new object in the startBundle method of DoFn
- D. Batch the job into ten-second increments

Answer: A

NEW QUESTION 133

- (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 138

- (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 141

- (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 BigQuery.
- E. Keep this ratio as 80% warm and 20% active.

Answer: C

NEW QUESTION 144

- (Exam Topic 6)

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Cloud Dataflow pipeline to use local execution
- B. Increase the maximum number of Cloud Dataflow workers by setting `maxNumWorkers` in `PipelineOptions`
- C. Increase the number of nodes in the Cloud Bigtable cluster
- D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
- E. Modify your Cloud Dataflow pipeline to use the `CoGroupByKey` transform before writing to Cloud Bigtable

Answer: BC

NEW QUESTION 149

- (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 `SideInputs` 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 `SideInput`

Answer: B

NEW QUESTION 153

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 155

- (Exam Topic 6)

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

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

Answer: C

NEW QUESTION 156

- (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 158

- (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.
- Performance 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 availabilit
- B. Store the data in HDFS, and perform analysis as needed.
- C. Store the data in BigQuer
- D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- E. Store the data in a regional Cloud Storage bucke
- F. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- G. Store the data in a multi-regional Cloud Storage bucke
- H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: D

NEW QUESTION 162

- (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 166

- (Exam Topic 6)

You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

- A. Use Transfer Appliance to copy the data to Cloud Storage
- B. Use gsutil cp -J to compress the content being uploaded to Cloud Storage
- C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
- D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

Answer: A

NEW QUESTION 169

- (Exam Topic 6)

You have uploaded 5 years of log data to Cloud Storage. A user reported that some data points in the log data are outside of their expected ranges, which indicates errors. You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

- A. Import the data from Cloud Storage into BigQuery. Create a new BigQuery table, and skip the rows with errors.
- B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage. Skip the rows with errors.
- C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.
- D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage.

Answer: D

NEW QUESTION 170

- (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 connector.
- 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 connector.
- 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 connector.
- I. Use a Dataflow job to read from PubSub and write to GCS.

Answer: A

NEW QUESTION 172

- (Exam Topic 6)

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps. You have the following requirements:

- You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- You will extract topics and sentiment from the posts.
- You must store the raw posts for archiving and reprocessing.
- You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

- A. Store the social media posts and the data extracted from the API in BigQuery.
- B. Store the social media posts and the data extracted from the API in Cloud SQL.
- C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
- D. Feed social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Answer: D

NEW QUESTION 174

- (Exam Topic 6)

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Cloud Storage. Your Spark job is a complex analytical workload that consists of many shuffling operations, and initial data are parquet files (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 like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload. What should you do?

- A. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.
- B. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- C. Switch to TFRecords format (approx 200 MB per file) 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 177

- (Exam Topic 6)

You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center. Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB. Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time. What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

- A. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates.
- B. BigQuery Data Transfer Service for the migration, Pub/Sub and Dataproc for the real-time updates.
- C. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates.
- D. gsutil for both the migration and the real-time updates.

Answer: A

NEW QUESTION 178
.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

Professional-Data-Engineer Practice Exam Features:

- * Professional-Data-Engineer Questions and Answers Updated Frequently
- * Professional-Data-Engineer Practice Questions Verified by Expert Senior Certified Staff
- * Professional-Data-Engineer Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * Professional-Data-Engineer Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

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