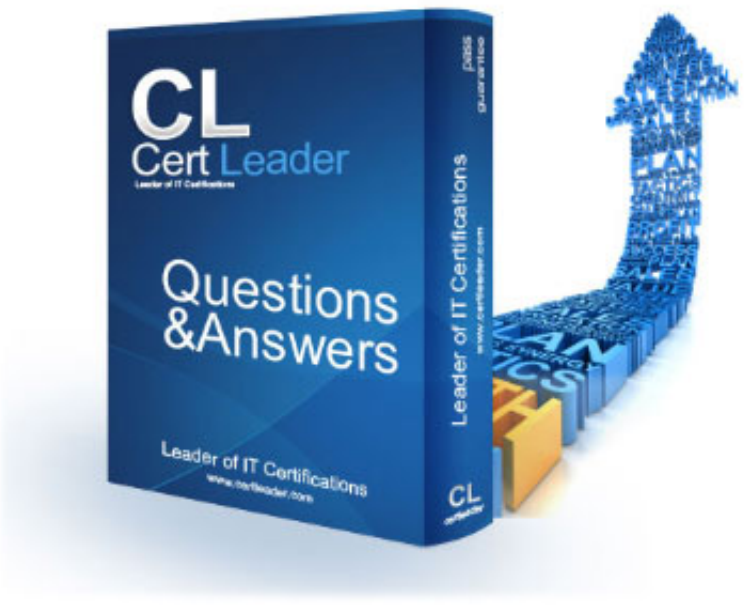
Professional-Data-Engineer Dumps

Google Professional Data Engineer Exam

https://www.certleader.com/Professional-Data-Engineer-dumps.html



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

Answer: D

NEW QUESTION 3
- (Exam Topic 1)
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 4
- (Exam Topic 1)
Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

A. Disable writes to certain tables.
B. Restrict access to tables by role.
C. Ensure that the data is encrypted at all times.
D. Restrict BigQuery API access to approved users.
E. Segregate data across multiple tables or databases.
F. Use Google Stackdriver Audit Logging to determine policy violations.

Answer: BDF

NEW QUESTION 5
- (Exam Topic 1)
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

A. Re-write the application to load accumulated data every 2 minutes.
B. Convert the streaming insert code to batch load for individual messages.
C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Answer: D

Explanation:
The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written tio storage

NEW QUESTION 6
- (Exam Topic 1)
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are

collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
B. The performance issue should be resolved over time as the site of the BigDate cluster is increased.
C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Answer: A

NEW QUESTION 7
- (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 8
- (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 9
- (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 10
- (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 10
- (Exam Topic 3)
MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone
B. The number of workers
C. The disk size per worker
D. The maximum number of workers

Answer: A

NEW QUESTION 11
- (Exam Topic 4)
Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.
You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (choose two.)

A. Introduce data compression for each file to increase the rate file of file transfer.
B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
D. Assemble 1,000 files into a tape archive (TAR) fil
E. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
F. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premices data to the designated storage bucket.

Answer: CE

NEW QUESTION 12
- (Exam Topic 4)
You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

A. Change the processing job to use Google Cloud Dataproc instead.
B. Manually start the Cloud Dataflow job each morning when you get into the office.
C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Answer: C

NEW QUESTION 13
- (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 tabl
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 fil
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 FullNam
G. Run a BigQuery load job to load the new CSV file into BigQuery.

Answer: C

NEW QUESTION 16
- (Exam Topic 4)
Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.
B. Rewrite the job in Apache Spark.
C. Increase the size of the Hadoop cluster.
D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Answer: A

NEW QUESTION 20
- (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 25
- (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 30
- (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 34
- (Exam Topic 5)
If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

A. 1 continuous and 2 categorical
B. 3 categorical
C. 3 continuous
D. 2 continuous and 1 categorical

Answer: D

Explanation:
The columns can be grouped into two types—categorical and continuous columns:
A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.
A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. $14,084) is a continuous column.
Year of birth and income are continuous columns. Country is a categorical column.
You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.
Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

NEW QUESTION 36
- (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 39
- (Exam Topic 5)
In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) .

A. VPN connection
B. Special browser
C. SSH tunnel
D. FTP connection

Answer: C

Explanation:
To connect to the web interfaces, it is recommended to use an SSH tunnel to create a secure connection to the master node.
Reference:
https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#connecting_to_the_web_interfaces

NEW QUESTION 43
- (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 45
- (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 47
- (Exam Topic 5)
What is the HBase Shell for Cloud Bigtable?

A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.
B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.
C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.
D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

Answer: B

Explanation:
The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables. The Cloud Bigtable HBase client for Java makes it possible to use the HBase shell to connect to Cloud Bigtable.
Reference: https://cloud.google.com/bigtable/docs/installing-hbase-shell

NEW QUESTION 48
- (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 51
- (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 53
- (Exam Topic 5)
Which of these statements about BigQuery caching is true?

A. By default, a query's results are not cached.
B. BigQuery caches query results for 48 hours.
C. Query results are cached even if you specify a destination table.
D. There is no charge for a query that retrieves its results from cache.

Answer: D

Explanation:
When query results are retrieved from a cached results table, you are not charged for the query. BigQuery caches query results for 24 hours, not 48 hours.
Query results are not cached if you specify a destination table.
A query's results are always cached except under certain conditions, such as if you specify a destination table. Reference:
https://cloud.google.com/bigquery/querying-data#query-caching

NEW QUESTION 58
- (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)
By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?

A. Windows at every 100 MB of data
B. Single, Global Window
C. Windows at every 1 minute
D. Windows at every 10 minutes

Answer: B

Explanation:

Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections
Reference: https://cloud.google.com/dataflow/model/pcollection

NEW QUESTION 66
- (Exam Topic 5)
Which of these operations can you perform from the BigQuery Web UI?

A. Upload a file in SQL format.
B. Load data with nested and repeated fields.
C. Upload a 20 MB file.
D. Upload multiple files using a wildcard.

Answer: B

Explanation:
You can load data with nested and repeated fields using the Web UI. You cannot use the Web UI to:
- Upload a file greater than 10 MB in size
- Upload multiple files at the same time
- Upload a file in SQL format
All three of the above operations can be performed using the "bq" command. Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION 71
- (Exam Topic 5)
Google Cloud Bigtable indexes a single value in each row. This value is called the .

A. primary key
B. unique key
C. row key
D. master key

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

NEW QUESTION 72
- (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 76
- (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 77

- (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 81
- (Exam Topic 5)
Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

A. Hive
B. Pig
C. YARN
D. Spark

Answer: ABD

Explanation:
Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.
Reference: https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

NEW QUESTION 86
- (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 89
- (Exam Topic 6)
You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

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

Answer: C

NEW QUESTION 90
- (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 95
- (Exam Topic 6)
You need ads data to serve Al models and historical data tor analytics longtail and outlier data points need to be identified You want to cleanse the data n near-reel time before running it through Al models What should you do?

A. Use BigQuery to ingest prepare and then analyze the data and then run queries to create views
B. Use Cloud Storage as a data warehouse shell scripts tor processing, and BigQuery to create views tor desired datasets
C. Use Dataflow to identity longtail and outber data points programmatically with BigQuery as a sink
D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery

Answer: A

NEW QUESTION 100
- (Exam Topic 6)
The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

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

Answer: D

NEW QUESTION 105
- (Exam Topic 6)
You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples 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 of batch normalization) to avoid overfitting.
D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

Answer: D

NEW QUESTION 107
- (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 transformMapReduce 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 112
- (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 duster is designed to support peak usage on the system, however, many jobs are batch n 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 servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will 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 115
- (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 118
- (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 in BigQuery and then transform the format later when the data is used.
B. Write a shell 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 119
- (Exam Topic 6)
You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?

A. The current epoch time
B. A concatenation of the product name and the current epoch time
C. A random universally unique identifier number (version 4 UUID)
D. The original order identification number from the sales system, which is a monotonically increasing integer

Answer: C

NEW QUESTION 123
- (Exam Topic 6)
You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?

A. Consume the stream of data in Cloud Dataflow using Kafka I
B. Set a sliding time window of 1 hour every 5 minute
C. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
D. Consume the stream of data in Cloud Dataflow using Kafka I
E. Set a fixed time window of 1 hour.Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
F. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Su
G. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtabl
H. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hou
I. If that number falls below 4000, send an alert.
J. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Su
K. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuer
L. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hou
M. If that number falls below 4000, send an alert.

Answer: C

NEW QUESTION 124
- (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 cluste
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 archiv
F. Locate the bottleneck and adjust cluster resources.
G. Use Cloud Dataflow to run your transformation
H. Monitor the job system lag with Stackdrive
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 job
L. Configure the job to use non-default Compute Engine machine types when needed.

Answer: B

NEW QUESTION 125
- (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 130
- (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 image
F. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer: A

NEW QUESTION 135
- (Exam Topic 6)
A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?

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

Answer: C

NEW QUESTION 136
- (Exam Topic 6)
You need (o 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 141
- (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 145
- (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 146
- (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 Bl Engine with authorized views
D. Use BigQuery Bl Engine with logical reviews

Answer: B

NEW QUESTION 149
- (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 154
- (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)
Your neural network model is taking days to train. You want to increase the training speed. What can you do?

A. Subsample your test dataset.
B. Subsample your training dataset.
C. Increase the number of input features to your model.
D. Increase the number of layers in your neural network.

Answer: D

Explanation:
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION 158
- (Exam Topic 6)
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
B. Create encryption keys in Cloud Key Management Servic
C. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
D. Create encryption keys locall
E. Upload your encryption keys to Cloud Key Management Servic
F. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
G. Create encryption keys in Cloud Key Management Servic
H. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Answer: C

NEW QUESTION 161
- (Exam Topic 6)
You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

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

Answer: A

NEW QUESTION 162
- (Exam Topic 6)
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 location_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 locationed and device_version

Answer: B

NEW QUESTION 166
- (Exam Topic 6)
You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

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

Answer: B

NEW QUESTION 170
- (Exam Topic 6)
You are collecting loT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data tittered by location_id and device_version with the following query:

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

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

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

Answer: C

NEW QUESTION 171
- (Exam Topic 6)
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

A. Create a Cloud Dataproc Workflow Template
B. Create an initialization action to execute the jobs
C. Create a Directed Acyclic Graph in Cloud Composer
D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Answer: C

NEW QUESTION 174
- (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 179
- (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 182
- (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 184
- (Exam Topic 6)
Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

- Single global endpoint
- ANSI SQL support
- Consistent access to the most up-to-date data

What should you do?

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

Answer: B

NEW QUESTION 189
- (Exam Topic 6)
You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.
What should you do?

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

Answer: A

NEW QUESTION 194
- (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 198
- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 200
- (Exam Topic 6)
You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?

A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.
B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.
C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.
D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

Answer: D

NEW QUESTION 203
- (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 207
- (Exam Topic 6)
You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence. To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory
B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS
C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up
D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

Answer: A

NEW QUESTION 208
- (Exam Topic 6)
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

A. Use Cloud Dataprep to find null values in sample source dat
B. Convert all nulls to ‘none’ using a Cloud Dataproc job.
C. Use Cloud Dataprep to find null values in sample source dat
D. Convert all nulls to 0 using a Cloud Dataprep job.
E. Use Cloud Dataflow to find null values in sample source dat
F. Convert all nulls to ‘none’ using a Cloud Dataprep job.
G. Use Cloud Dataflow to find null values in sample source dat
H. Convert all nulls to using a custom script.

Answer: C

NEW QUESTION 211
.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your Professional-Data-Engineer Exam with Our Prep Materials Via below:

https://www.certleader.com/Professional-Data-Engineer-dumps.html