Professional-Data-Engineer Dumps

Google Professional Data Engineer Exam

https://www.certleader.com/Professional-Data-Engineer-dumps.html
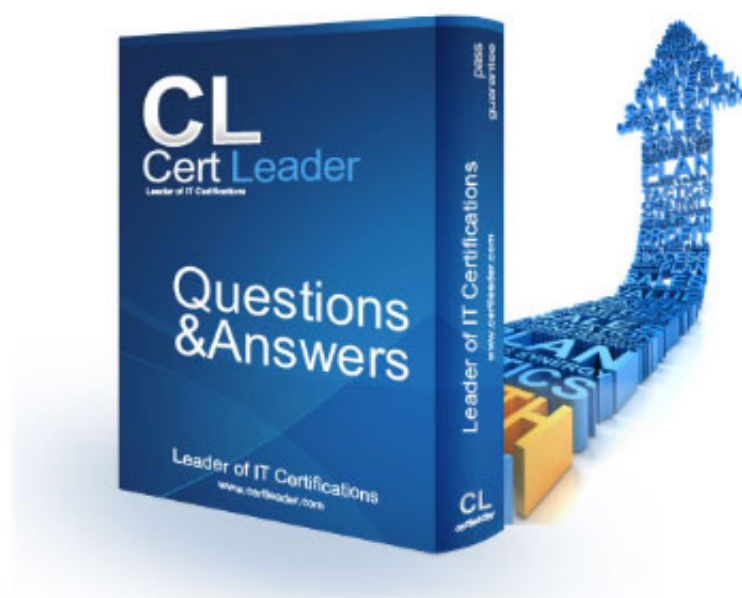
NEW QUESTION 1
- (Exam Topic 1)
Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

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

Answer: C

NEW QUESTION 2
- (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: D

NEW QUESTION 3
- (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 typ
B. Reload the data.
C. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numericvalues from the column TS for each ro
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 typ
H. Reload all data in append mod
I. For each appended row, set the value of IS_NEW to tru
J. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.
K. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP value
L. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP typ
M. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now o
N. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

NEW QUESTION 4
- (Exam Topic 1)
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

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

Answer: BCE

NEW QUESTION 5
- (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 6
- (Exam Topic 1)
Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

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

Answer: D

NEW QUESTION 7
- (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: D

NEW QUESTION 8
- (Exam Topic 1)
You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

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

Answer: B

NEW QUESTION 9
- (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 10
- (Exam Topic 1)
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being 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 PCollexction represents asingle row in the table.

Answer: D

NEW QUESTION 10
- (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 14
- (Exam Topic 1)
Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

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

Answer: B

NEW QUESTION 19
- (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 24
- (Exam Topic 2)
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?

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

Answer: B

NEW QUESTION 29
- (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 33
- (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 34
- (Exam Topic 3)
You need to compose visualizations for operations teams with the following requirements: Which approach meets the requirements?

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

Answer: C

NEW QUESTION 38
- (Exam Topic 4)
Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.
B. The CSV data has invalid rows that were skipped on import.
C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Answer: B

NEW QUESTION 40
- (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 42
- (Exam Topic 5)
When you design a Google Cloud Bigtable schema it is recommended that you .

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

Answer: C

Explanation:
All operations are atomic at the row level. For example, if you update two rows in a table, it's possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

NEW QUESTION 47
- (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 52
- (Exam Topic 5)
Which of these are examples of a value in a sparse vector? (Select 2 answers.)

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

Answer: CD

Explanation:
Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).
[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1. [0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s. Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION 57
- (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 58
- (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 62
- (Exam Topic 5)
Which methods can be used to reduce the number of rows processed by BigQuery?

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

Answer: A

Explanation:
If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.
If you use the LIMIT clause, BigQuery will still process the entire table. Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION 67
- (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 70
- (Exam Topic 5)
Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?

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

Answer: BD

Explanation:
The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster: Processing only—Since preemptibles can be reclaimed at any time, preemptible workers do not store data.
Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.
No preemptible-only clusters—To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters.
Persistent disk size—As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS.
The managed group automatically re-adds workers lost due to reclamation as capacity permits. Reference:
https://cloud.google.com/dataproc/docs/concepts/preemptible-vms

NEW QUESTION 72
- (Exam Topic 5)
When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

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

Answer: A

Explanation:
When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

NEW QUESTION 74
- (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 79

- (Exam Topic 5)
To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

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

Answer: A

Explanation:
gcloud ml-engine local train - run a Cloud ML Engine training job locally
This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.
This is especially useful in the case of testing distributed models, as it allows you to validate that you are
properly interacting with the Cloud ML Engine cluster configuration. Reference: https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train

NEW QUESTION 80
- (Exam Topic 5)
What are two methods that can be used to denormalize tables in BigQuery?

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

Answer: B

Explanation:
The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are
dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information.
The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing
records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a
JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which
would be represented as nested, repeated elements.
Reference: https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION 84
- (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 87
- (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 88
- (Exam Topic 5)
Which of the following is not true about Dataflow pipelines?

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

Answer: D

Explanation:
The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms
Reference: https://cloud.google.com/dataflow/model/pipelines

NEW QUESTION 90
- (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 94
- (Exam Topic 5)
How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause
B. Use the EXTRACT(DAY) clause
C. Use the PARTITIONTIME pseudo-column in the WHERE clause
D. Use DATE BETWEEN in the WHERE clause

Answer: C

Explanation:
Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02')
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

NEW QUESTION 98
- (Exam Topic 5)
Scaling a Cloud Dataproc cluster typically involves .

A. increasing or decreasing the number of worker nodes
B. increasing or decreasing the number of master nodes
C. moving memory to run more applications on a single node
D. deleting applications from unused nodes periodically

Answer: A

Explanation:
After creating a Cloud Dataproc cluster, you can scale the cluster by increasing or decreasing the number of worker nodes in the cluster at any time, even when jobs are running on the cluster. Cloud Dataproc clusters are typically scaled to:
1) increase the number of workers to make a job run faster
2) decrease the number of workers to save money
3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage Reference:
https://cloud.google.com/dataproc/docs/concepts/scaling-clusters

NEW QUESTION 101

- (Exam Topic 5)
If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?

A. Unsupervised learning
B. Regressor
C. Classifier
D. Clustering estimator

Answer: B

Explanation:
Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores.
Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.
Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset.
Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.
Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION 102
- (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 103
- (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 105
- (Exam Topic 5)
Which of the following are feature engineering techniques? (Select 2 answers)

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

Answer: CD

Explanation:
Selecting and crafting the right set of feature columns is key to learning an effective model. Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive
bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.
Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.
Reference: https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION 106
- (Exam Topic 5)
Which of these is not a supported method of putting data into a partitioned table?

A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.
B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "$YYYYMMDD".
C. Create a partitioned table and stream new records to it every day.
D. Use ORDER BY to put a table's rows into chronological order and then change the table's type to "Partitioned".

Answer: D

Explanation:
You cannot change an existing table into a partitioned table. You must create a partitioned table from scratch. Then you can either stream data into it every day and the data will automatically be put in the right partition, or you can load data into a specific partition by using "$YYYYMMDD" at the end of the table name.
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION 110
- (Exam Topic 5)
The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

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

Answer: A

Explanation:
The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.
Reference: https://cloud.google.com/bigtable/docs/dataflow-hbase

NEW QUESTION 112
- (Exam Topic 5)
Which action can a Cloud Dataproc Viewer perform?

A. Submit a job.
B. Create a cluster.
C. Delete a cluster.
D. List the jobs.

Answer: D

Explanation:
A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.
Reference: https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

NEW QUESTION 115
- (Exam Topic 5)
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

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

Answer: B

Explanation:
In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference: https://cloud.google.com/dataflow/model/programming-model

NEW QUESTION 116
- (Exam Topic 5)
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

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

Answer: AB

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

NEW QUESTION 119
- (Exam Topic 5)
Which of the following is NOT true about Dataflow pipelines?

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

Answer: A

Explanation:
Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs
Reference: https://cloud.google.com/dataflow/

NEW QUESTION 123
- (Exam Topic 5)
Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

A. Field promotion
B. Randomization
C. Salting
D. Hashing

Answer: A

Explanation:
By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.
Reference:
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

NEW QUESTION 126
- (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 128
- (Exam Topic 5)
Which software libraries are supported by Cloud Machine Learning Engine?

A. Theano and TensorFlow
B. Theano and Torch
C. TensorFlow
D. TensorFlow and Torch

Answer: C

Explanation:
Cloud ML Engine mainly does two things:
Enables you to train machine learning models at scale by running TensorFlow training applications in the cloud.
Hosts those trained models for you in the cloud so that you can use them to get predictions about new data.
Reference: https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does

NEW QUESTION 130
- (Exam Topic 5)
Which of the following is not possible using primitive roles?

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

Answer: C

Explanation:
Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.
Reference: https://cloud.google.com/bigquery/docs/access-control#primitive_iam_roles

NEW QUESTION 131
- (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 135
- (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 138
- (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 143
- (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 146
- (Exam Topic 6)
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.
B. Implement clustering in BigQuery on the package-tracking ID column.
C. Tier older data onto Cloud Storage files, and leverage extended tables.
D. Re-create the table using data partitioning on the package delivery date.

Answer: A

NEW QUESTION 150
- (Exam Topic 6)
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud

Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

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

Answer: C

NEW QUESTION 155
- (Exam Topic 6)
You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the
Internet so public initialization actions cannot fetch resources. What should you do?

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

Answer: D

NEW QUESTION 158
- (Exam Topic 6)
You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

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

Answer: D

NEW QUESTION 162
- (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 applicatio
B. Process the generated Entity Analysis as labels.
C. Call the Cloud Natural Language API from your applicatio
D. Process the generated Sentiment Analysis as labels.
E. Build and train a text classification model using TensorFlo
F. Deploy the model using Cloud Machine Learning Engin
G. Call the model from your application and process the results as labels.
H. Build and train a text classification model using TensorFlo
I. Deploy the model using a KubernetesEngine cluste
J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 165
- (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 170
- (Exam Topic 6)
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the

schema design will allow for future growth of the dataset. Which database and data model should you choose?

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

Answer: D

NEW QUESTION 172
- (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 176
- (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 178
- (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: C

NEW QUESTION 182
- (Exam Topic 6)
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

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

Answer: D

NEW QUESTION 187
- (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 188
- (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 189
- (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: A

NEW QUESTION 193
- (Exam Topic 6)
Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

A. Create a file on a shared file and have the application servers write all bid events to that fil
B. Process the file with Apache Hadoop to identify which user bid first.
C. Have each application server write the bid events to Cloud Pub/Sub as they occu
D. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
E. Set up a MySQL database for each application server to write bid events int
F. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
G. Have each application server write the bid events to Google Cloud Pub/Sub as they occu
H. Use a pull subscription to pull the bid events using Google Cloud Dataflo
I. Give the bid for each item to the userIn the bid event that is processed first.

Answer: C

NEW QUESTION 198
- (Exam Topic 6)
You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

A. Re-create the tables using DD
B. Partition the tables by a column containing a TIMESTAMP or DATE Type.
C. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
D. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
E. Write an Apache Beam pipeline that creates a BigQuery table per da
F. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Answer: C

NEW QUESTION 199
- (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 201
- (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 AP
E. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Answer: A

NEW QUESTION 202
- (Exam Topic 6)
Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes
B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project
C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric
D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

Answer: D

NEW QUESTION 206
- (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 you 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 utilizatio
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 of read operations take longer than 100 ms.

Answer: AC

NEW QUESTION 209
- (Exam Topic 6)
You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

A. Denormalize the data as must as possible.
B. Preserve the structure of the data as much as possible.
C. Use BigQuery UPDATE to further reduce the size of the dataset.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro fil
F. Use BigQuery'ssupport for external data sources to query.

Answer: DE

NEW QUESTION 213
- (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 217
- (Exam Topic 6)
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

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

Which pipeline should you use to meet these requirements?

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

Answer: A

NEW QUESTION 222
- (Exam Topic 6)
You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

A. Create a cron schedule in Cloud Dataprep.
B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.
C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.
D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

Answer: C

NEW QUESTION 225
- (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 228
- (Exam Topic 6)
You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

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

Answer: A

NEW QUESTION 230
- (Exam Topic 6)
Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

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

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

Answer: A

NEW QUESTION 232
- (Exam Topic 6)
You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?

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

Answer: A

NEW QUESTION 234
- (Exam Topic 6)
You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples
B. Reduce the number of training examples
C. Use a smaller set of features
D. Use a larger set of features
E. Increase the regularization parameters
F. Decrease the regularization parameters

Answer: ADF

NEW QUESTION 239
- (Exam Topic 6)
You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

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

Answer: A

NEW QUESTION 243
- (Exam Topic 6)
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

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

Answer: C

NEW QUESTION 248
- (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 249
- (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 252
- (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: A

NEW QUESTION 254
- (Exam Topic 6)
You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models
B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Answer: A

NEW QUESTION 259
- (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 263
- (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 tothe 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 268
- (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 273
- (Exam Topic 6)
You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?

A. Assign the users/groups data viewer access at the table level for each table
B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
C. Create authorized views for each team in the same dataset in which the data resides, and assign theusers/groups data viewer access to the authorized views
D. Create authorized views for each team in datasets created for each tea
E. Assign the authorized views data viewer access to the dataset in which the data reside
F. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer: C

NEW QUESTION 276
- (Exam Topic 6)
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

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

Answer: B

NEW QUESTION 277
- (Exam Topic 6)
A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features). How should you create the ML pipeline?

A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.
B. Create an Authorized View with the provided quer
C. Share the dataset that contains the view with the application service account.
D. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the quer
E. Grant the Dataflow Worker role to the application service account.
F. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query.Write the results to Cloud Bigtable using BigtableI
G. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.

Answer: D

NEW QUESTION 279
......

Thank You for Trying Our Product

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

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