Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

NEW QUESTION 1
- (Exam Topic 1)
You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

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

Answer: B

NEW QUESTION 2
- (Exam Topic 1)
You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

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

Answer: C

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

NEW QUESTION 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)
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 5
- (Exam Topic 1)
Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

A. Check the dashboard application to see if it is not displaying correctly.
B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Answer: B

NEW QUESTION 6
- (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 7
- (Exam Topic 1)
You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

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

Answer: B

NEW QUESTION 8
- (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 9
- (Exam Topic 1)
Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

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

Answer: C

NEW QUESTION 10
- (Exam Topic 1)
You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP typ
B. Reload the data.
C. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each 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 10
- (Exam Topic 1)
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

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

E. Reinforcement learning to predict the location of a transaction.
F. Unsupervised learning to predict the location of a transaction.

Answer: BCD

NEW QUESTION 14
- (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 18
- (Exam Topic 1)
You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

A. Disable caching by editing the report settings.
B. Disable caching in BigQuery by editing table details.
C. Refresh your browser tab showing the visualizations.
D. Clear your browser history for the past hour then reload the tab showing the virtualizations.

Answer: A

Explanation:
Reference https://support.google.com/datastudio/answer/7020039?hl=en

NEW QUESTION 22
- (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 23
- (Exam Topic 3)
You need to compose visualization for operations teams with the following requirements:
- Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)
- The report must not be more than 3 hours delayed from live data.
- The actionable report should only show suboptimal links.
- Most suboptimal links should be sorted to the top.
- Suboptimal links can be grouped and filtered by regional geography.
- User response time to load the report must be <5 seconds.

You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.
B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.
D. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.

Answer: B

NEW QUESTION 27
- (Exam Topic 4)
You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

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

Answer: A

NEW QUESTION 31
- (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 34
- (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 37
- (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 42
- (Exam Topic 5)
Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

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

Answer: D

Explanation:
When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.
Processing time triggers. These triggers operate on the processing time – the time when the data element is processed at any given stage in the pipeline.
Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.
Reference: https://beam.apache.org/documentation/programming-guide/#triggers

NEW QUESTION 44

- (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 45
- (Exam Topic 5)
What are two of the characteristics of using online prediction rather than batch prediction?

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

Answer: BD

Explanation:
Online prediction
Optimized to minimize the latency of serving predictions.
Predictions returned in the response message. Batch prediction
Optimized to handle a high volume of instances in a job and to run more complex models. Predictions written to output files in a Cloud Storage location that you specify.
Reference:
https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION 47
- (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 48
- (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 51
- (Exam Topic 5)
Cloud Bigtable is Google's Big Data database service.

A. Relational
B. mySQL
C. NoSQL
D. SQL Server

Answer: C

Explanation:
Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail. It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.
Reference: https://cloud.google.com/bigtable/

NEW QUESTION 54
- (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 58
- (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 61
- (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 66
- (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 68
- (Exam Topic 5)
Which Java SDK class can you use to run your Dataflow programs locally?

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

Answer: B

Explanation:
DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

NEW QUESTION 72
- (Exam Topic 5)
Why do you need to split a machine learning dataset into training data and test data?

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

Answer: B

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

NEW QUESTION 77
- (Exam Topic 5)
When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and .

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

Answer: A

Explanation:
At a minimum, you must specify four values when creating a new cluster with the projects.regions.clusters.create operation:
The project in which the cluster will be created The region to use
The name of the cluster
The zone in which the cluster will be created
You can specify many more details beyond these minimum requirements. For example, you can
also specify the number of workers, whether preemptible compute should be used, and the network settings. Reference:
https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

NEW QUESTION 79
- (Exam Topic 5)
Which of the following is NOT a valid use case to select HDD (hard disk drives) as the storage for Google Cloud Bigtable?

A. You expect to store at least 10 TB of data.
B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.
C. You need to integrate with Google BigQuery.
D. You will not use the data to back a user-facing or latency-sensitive application.

Answer: C

Explanation:
For example, if you plan to store extensive historical data for a large number of remote-sensing devices and then use the data to generate daily reports, the cost savings for HDD storage may justify the performance tradeoff. On the other hand, if you plan to use the data to display a real-time dashboard, it probably would not make sense to use HDD storage—reads would be much more frequent in this case, and reads are much slower with HDD storage.
Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

NEW QUESTION 82

- (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 83
- (Exam Topic 5)
What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

A. Include multiple time series values within the row key
B. Keep the row keep as an 8 bit integer
C. Keep your row key reasonably short
D. Keep your row key as long as the field permits

Answer: C

Explanation:
A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

NEW QUESTION 86
- (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 89
- (Exam Topic 5)
You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?

A. Both batch and streaming
B. BigQuery cannot be used as a sink
C. Only batch
D. Only streaming

Answer: A

Explanation:
When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts
Reference: https://cloud.google.com/dataflow/model/bigquery-io

NEW QUESTION 94
- (Exam Topic 5)
Does Dataflow process batch data pipelines or streaming data pipelines?

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

Answer: B

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

NEW QUESTION 98
- (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 103
- (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 107
- (Exam Topic 5)
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

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

Answer: C

Explanation:
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,
If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from
your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.
Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

NEW QUESTION 109
- (Exam Topic 5)
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output.
Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

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

Answer: B

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

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

NEW QUESTION 113
- (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 114
- (Exam Topic 5)
When a Cloud Bigtable node fails, is lost.

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

Answer: B

Explanation:
A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.
Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:
Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud
Bigtable simply updates the pointers for each node.
Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.
When a Cloud Bigtable node fails, no data is lost Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION 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 117
- (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 119
- (Exam Topic 5)
What is the recommended action to do in order to switch between SSD and HDD storage for your Google Cloud Bigtable instance?

A. create a third instance and sync the data from the two storage types via batch jobs
B. export the data from the existing instance and import the data into a new instance

C. run parallel instances where one is HDD and the other is SDD
D. the selection is final and you must resume using the same storage type

Answer: B

Explanation:
When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud Platform Console to change the type of storage that is used for the cluster.
If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can write
a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another. Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd–

NEW QUESTION 123
- (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 128
- (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 129
- (Exam Topic 5)
Which of the following statements about Legacy SQL and Standard SQL is not true?

A. Standard SQL is the preferred query language for BigQuery.
B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
C. One difference between the two query languages is how you specify fully-qualified table names (i.
D. table names that include their associated project name).
E. You need to set a query language for each dataset and the default is Standard SQL.

Answer: D

Explanation:
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.
Standard SQL has been the preferred query language since BigQuery 2.0 was released.
In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.
Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

NEW QUESTION 131
- (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 135
- (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 136
- (Exam Topic 6)
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this?
Choose 2 answers.

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

Answer: CE

NEW QUESTION 137
- (Exam Topic 6)
You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.
What should you do?

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

Answer: B

NEW QUESTION 140
- (Exam Topic 6)
An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application.
They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?

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

Answer: C

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

NEW QUESTION 145
- (Exam Topic 6)
You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to
BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

A. Publisher throughput quota is too small.

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

Answer: CD

NEW QUESTION 149
- (Exam Topic 6)
You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

- Each department should have access only to their data.
- Each department will have one or more leads who need to be able to create and update tables and provide them to their team.
- Each department has data analysts who need to be able to query but not modify data. How should you set access to the data in BigQuery?

A. Create a dataset for each departmen
B. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.
C. Create a dataset for each departmen
D. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.
E. Create a table for each departmen
F. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.
G. Create a table for each departmen
H. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

Answer: D

NEW QUESTION 152
- (Exam Topic 6)
You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

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

Answer: D

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

NEW QUESTION 154
- (Exam Topic 6)
You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

A. Export the records from the database as an Avro fil
B. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
C. Export the records from the database as an Avro fil
D. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
E. Export the records from the database into a CSV fil
F. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storag
G. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
H. Export the records from the database as an Avro fil
I. Create a public URL for the Avro file, and thenuse 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 159
- (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 162
- (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 166
- (Exam Topic 6)
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?

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

Answer: B

NEW QUESTION 167
- (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 171
- (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 175
- (Exam Topic 6)
Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

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

Answer: A

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

NEW QUESTION 178
- (Exam Topic 6)
You work for an advertising company, and you’ve developed a Spark ML model to predict click-through rates at advertisement blocks. You’ve been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

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

Answer: C

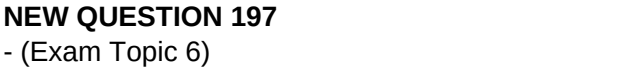
Explanation:
https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

NEW QUESTION 183
- (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 datafor 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: D

NEW QUESTION 188
- (Exam Topic 6)
You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country
You check the query plan for the query and see the following output in the Read section of Stage:1:



What is the most likely cause of the delay for this query?

A. Users are running too many concurrent queries in the system
B. The [myproject:mydataset.mytable] table has too many partitions
C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Answer: A

NEW QUESTION 192
- (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 DATEType.
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 197
- (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

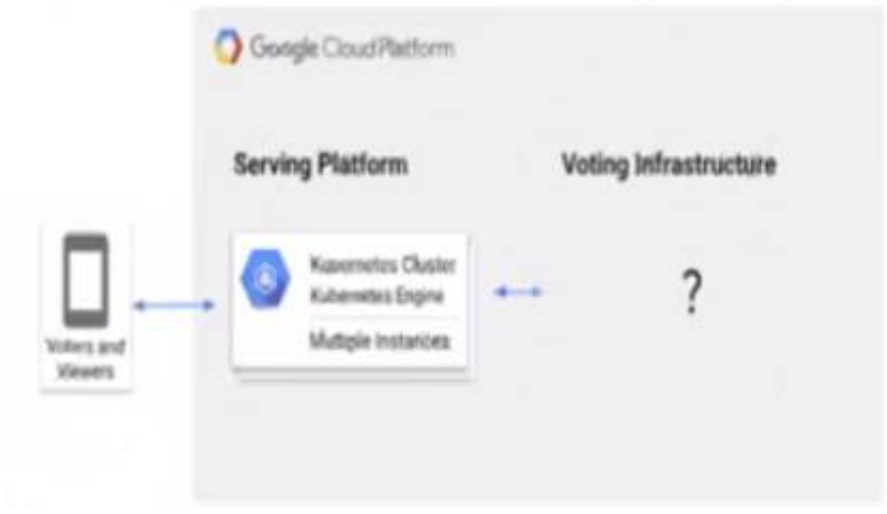
Answer: D

NEW QUESTION 200
- (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 201
- (Exam Topic 6)
A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and Hal all votes are processed. You must display partial results write voting is open. After voting doses you need to count the votes exactly once white optimizing cost. What should you do?



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

Answer: C

NEW QUESTION 202
- (Exam Topic 6)
An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

A. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.
B. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.
C. Write an application that makes a queue in a NoSQL database
D. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic

Answer: C

NEW QUESTION 204
- (Exam Topic 6)
Your company is migrating its on-premises data warehousing solution to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply daily updates from transactional database sources Your company wants to use BigQuery to improve its handling of CDC and to optimize the performance of the data warehouse Source system changes must be available for query m near-real time using tog-based CDC streams You need to ensure that changes in the BigQuery reporting table are available with minimal latency and reduced overhead. What should you do? Choose 2 answers

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

Answer: BD

NEW QUESTION 206
- (Exam Topic 6)
You are 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 210
- (Exam Topic 6)
You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:
- Executing the transformations on a schedule
- Enabling non-developer analysts to modify transformations
- Providing a graphical tool for designing transformations

What should you do?

A. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis
B. Load each month’s CSV data into BigQuery, and write a SQL query to transform the data to a standard schem

C. Merge the transformed tables together with a SQL query
D. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformatio
E. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
F. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe.Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

Answer: A

Explanation:
you can use dataprep for continuously changing target schema
In general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema information can include:
Names of columns
Order of columns Column data types Data type format Example rows of data
A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator (or schema tag) is displayed.
https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049

NEW QUESTION 213
- (Exam Topic 6)
Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).
What should you do?

A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
B. Add a try… catch block to your DoFn that transforms the data, extract erroneous rows from logs.
C. Add a try… catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
D. Add a try… catch block to your DoFn that transforms the data, use a sideOutput to create a PCollectionthat can be stored to PubSub later.

Answer: C

NEW QUESTION 215
- (Exam Topic 6)
You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

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

Answer: A

NEW QUESTION 219
- (Exam Topic 6)
Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of datA. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

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

Answer: B

NEW QUESTION 220
- (Exam Topic 6)
Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

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

Answer: B

NEW QUESTION 221
- (Exam Topic 6)
You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

A. Store and process the entire dataset in BigQuery.
B. Store and process the entire dataset in Cloud Bigtable.

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

Answer: C

NEW QUESTION 223
- (Exam Topic 6)
Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst's projec
B. Restrict access to Stackdriver Logging via Cloud IAM roles.
C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' project
D. Restrict access to the Cloud Storage bucket.
E. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit log
F. Restrict access to the project with the exported logs.
G. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit log
H. Restrict access to the project that contains the exported logs.

Answer: D

NEW QUESTION 226
- (Exam Topic 6)
You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.
What should you do?

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

Answer: B

NEW QUESTION 227
- (Exam Topic 6)
You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

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

Answer: B

NEW QUESTION 229
- (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 234
- (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 235
- (Exam Topic 6)
Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing
on-premises applications every day. What should they do?

A. Execute gsutil rsync from the on-premises servers.
B. Use Cloud Dataflow and write the data to Cloud Storage.
C. Write a job template in Cloud Dataproc to perform the data transfer.
D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Answer: B

NEW QUESTION 236
- (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's support for external data sources to query.

Answer: AE

NEW QUESTION 240
- (Exam Topic 6)
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

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

Answer: A

NEW QUESTION 244
- (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 248
- (Exam Topic 6)
You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.
B. Create a new view with BigQuery that does not include a column which city information.
C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2 3, 4, or 5, and then use that number to represent the city in the model.
D. Use TensorFlow to create a categorical variable with a vocabulary lis
E. Create the vocabulary file and upload that as part of your model to BigQuery ML.

Answer: C

NEW QUESTION 253
- (Exam Topic 6)
You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDF
B. Mount the Hive tables locally.
C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluste

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

Answer: BC

NEW QUESTION 257
- (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 259
- (Exam Topic 6)
You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service About 70% of customer requests are simple requests that are solved within 10 intents The remaining 30% of inquiries require much longer, more complicated requests Which intents should you automate first?

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

Answer: A

NEW QUESTION 262
- (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 267
- (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 272
- (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 277
- (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 281
- (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 284
- (Exam Topic 6)
Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

A. Migrate the workload to Google Cloud Dataflow
B. Use pre-emptible virtual machines (VMs) for the cluster
C. Use a higher-memory node so that the job runs faster
D. Use SSDs on the worker nodes so that the job can run faster

Answer: A

NEW QUESTION 285
- (Exam Topic 6)
You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?

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

Answer: A

NEW QUESTION 288

- (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 289
- (Exam Topic 6)
You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20 TB in size. Which database should you choose?

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

Answer: A

NEW QUESTION 294
- (Exam Topic 6)
Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Answer: B

NEW QUESTION 295
- (Exam Topic 6)
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

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

Answer: C

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

NEW QUESTION 296
- (Exam Topic 6)
You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?

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

Answer: A

NEW QUESTION 298

.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

Professional-Data-Engineer Practice Exam Features:

* Professional-Data-Engineer Questions and Answers Updated Frequently

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

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

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