Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

NEW QUESTION 1
- (Exam Topic 1)
You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

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

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

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

Answer: C

NEW QUESTION 2
- (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 3
- (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 4
- (Exam Topic 1)
You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users’ privacy?

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

Answer: C

NEW QUESTION 5
- (Exam Topic 1)
Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

A. Create a Google Cloud Dataflow job to process the data.
B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Answer: D

NEW QUESTION 6
- (Exam Topic 1)
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

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

Answer: A

NEW QUESTION 7
- (Exam Topic 1)
You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

A. Linear regression
B. Logistic classification
C. Recurrent neural network
D. Feedforward neural network

Answer: A

NEW QUESTION 8
- (Exam Topic 1)
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.
The data scientists have written the following code to read the data for a new key features in the logs. BigQueryIO.Read
.named(“ReadLogData”)
.from(“clouddataflow-readonly:samples.log_data”)
You want to improve the performance of this data read. What should you do?

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

Answer: D

NEW QUESTION 9
- (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 10
- (Exam Topic 1)
Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

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

Answer: C

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

NEW QUESTION 10
- (Exam Topic 2)
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?

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

Answer: B

NEW QUESTION 13
- (Exam Topic 3)
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?

A. Rowkey: date#device_idColumn data: data_point
B. Rowkey: dateColumn data: device_id, data_point
C. Rowkey: device_idColumn data: date, data_point
D. Rowkey: data_pointColumn data: device_id, date
E. Rowkey: date#data_pointColumn data: device_id

Answer: D

NEW QUESTION 15
- (Exam Topic 3)
MJTelco is building a custom interface to share data. They have these requirements:

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

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

Answer: C

NEW QUESTION 16
- (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 18
- (Exam Topic 4)
Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

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

Answer: B

NEW QUESTION 22
- (Exam Topic 4)
Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.
You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (choose two.)

A. Introduce data compression for each file to increase the rate file of file transfer.
B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.

D. Assemble 1,000 files into a tape archive (TAR) fil
E. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
F. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premices data to the designated storage bucket.

Answer: CE

NEW QUESTION 26
- (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 30
- (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 33
- (Exam Topic 5)
You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.
Tom,555 X street Tim,553 Y street Sam, 111 Z street
Which operation is best suited for the above data processing requirement?

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

Answer: A

Explanation:
In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.
Reference: https://cloud.google.com/dataflow/model/par-do

NEW QUESTION 37
- (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 38
- (Exam Topic 5)
How would you query specific partitions in a BigQuery table?

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

Answer: C

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

NEW QUESTION 39
- (Exam Topic 5)
What are two methods that can be used to denormalize tables in BigQuery?

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

Answer: B

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

NEW QUESTION 44
- (Exam Topic 5)
What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?

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

Answer: D

Explanation:
Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Tri

NEW QUESTION 45
- (Exam Topic 5)
What are the minimum permissions needed for a service account used with Google Dataproc?

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

Answer: D

Explanation:
Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging.
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION 46
- (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 47
- (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 49
- (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 51
- (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 55
- (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 59
- (Exam Topic 5)
Which of the following are examples of hyperparameters? (Select 2 answers.)

A. Number of hidden layers
B. Number of nodes in each hidden layer
C. Biases
D. Weights

Answer: AB

Explanation:
If model parameters are variables that get adjusted by training with existing data, your hyperparameters are the variables about the training process itself. For example, part of setting up a deep neural network is deciding how many "hidden" layers of nodes to use between the input layer and the output layer, as well as how many nodes each layer should use. These variables are not directly related to the training data at all. They are configuration variables. Another difference is that parameters change during a training job, while the hyperparameters are usually constant during a job.
Weights and biases are variables that get adjusted during the training process, so they are not hyperparameters. Reference: https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview

NEW QUESTION 62
- (Exam Topic 5)
To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

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

Answer: A

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

NEW QUESTION 67
- (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 70
- (Exam Topic 5)
If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?

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

Answer: B

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

NEW QUESTION 72
- (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 74
- (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 78
- (Exam Topic 5)
Which of these operations can you perform from the BigQuery Web UI?

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

Answer: B

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

NEW QUESTION 79
- (Exam Topic 5)
Which of the following statements is NOT true regarding Bigtable access roles?

A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
C. You can configure access control only at the project level.
D. To give a user access to only one table in a project, you must configure access through your application.

Answer: B

Explanation:
For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:
Read from, but not write to, any table within the project.
Read from and write to any table within the project, but not manage instances. Read from and write to any table within the project, and manage instances.
Reference: https://cloud.google.com/bigtable/docs/access-control

NEW QUESTION 83
- (Exam Topic 5)
Cloud Bigtable is a recommended option for storing very large amounts of _____?

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

Answer: C

Explanation:
Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data.

A single value in each row is indexed; this value is known as the row key. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION 88
- (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 89
- (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 91
- (Exam Topic 5)
Which methods can be used to reduce the number of rows processed by BigQuery?

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

Answer: A

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

NEW QUESTION 95
- (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 98
- (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 100
- (Exam Topic 6)
You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

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

Answer: B

NEW QUESTION 103
- (Exam Topic 6)
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

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

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

Answer: A

NEW QUESTION 105
- (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 108
- (Exam Topic 6)
You are building a teal-lime prediction engine that streams files, which may contain Pll (personal identifiable information) data, into Cloud Storage and eventually into BigQuery You want to ensure that the sensitive data is masked but still maintains referential Integrity, because names and emails are often used as join keys How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the Pll data is not accessible by unauthorized individuals?

A. Create a pseudonym by replacing the Pll data with cryptogenic tokens, and store the non-tokenized data in a locked-down button.
B. Redact all Pll data, and store a version of the unredacted data in a locked-down bucket
C. Scan every table in BigQuery, and mask the data it finds that has Pll
D. Create a pseudonym by replacing Pll data with a cryptographic format-preserving token

Answer: A

NEW QUESTION 113
- (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 115
- (Exam Topic 6)
Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS) in a colocation facility. The duster is designed to support peak usage on the system, however, many jobs are batch n nature, and usage of the cluster fluctuates quite dramatically.
Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will still executing the migration in time?

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

Answer: D

NEW QUESTION 120
- (Exam Topic 6)
You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

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

Answer: B

NEW QUESTION 122
- (Exam Topic 6)
You need to choose a database for a new project that has the following requirements:
- Fully managed
- Able to automatically scale up
- Transactionally consistent
- Able to scale up to 6 TB
- Able to be queried using SQL Which database do you choose?

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

Answer: C

NEW QUESTION 124
- (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 129
- (Exam Topic 6)
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

A. Use Cloud Dataproc to run your transformation

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

Answer: B

NEW QUESTION 133
- (Exam Topic 6)
You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

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

Answer: D

NEW QUESTION 134
- (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 136
- (Exam Topic 6)
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

A. Cloud Scheduler
B. Cloud Dataflow
C. Cloud Functions
D. Cloud Composer

Answer: A

NEW QUESTION 140
- (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 142
- (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 145
- (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 147
- (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 150
- (Exam Topic 6)
You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage.Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.

Answer: A

NEW QUESTION 154
- (Exam Topic 6)
You 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 156
- (Exam Topic 6)
An aerospace company uses a proprietary data format to store its night data. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?

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

Answer: D

NEW QUESTION 158
- (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 159
- (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 164
- (Exam Topic 6)
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

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

Answer: D

NEW QUESTION 168
- (Exam Topic 6)
You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

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

Answer: D

NEW QUESTION 169
- (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 171
- (Exam Topic 6)
You are implementing security best practices on your data pipeline. Currently, you are manually executing
jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery.
How should you securely run this workload?

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

Answer: B

NEW QUESTION 174
- (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 176
- (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 178
- (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 179
- (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 183
- (Exam Topic 6)
You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.
You have the following requirements:
- You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- You will extract topics and sentiment from the posts.
- You must store the raw posts for archiving and reprocessing.
- You will create dashboards to be shared with people both inside and outside your organization.

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

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

Answer: D

NEW QUESTION 184

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

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

Answer: B

NEW QUESTION 186
- (Exam Topic 6)
You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

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

Answer: A

NEW QUESTION 190
- (Exam Topic 6)
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

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

Answer: C

NEW QUESTION 193
......