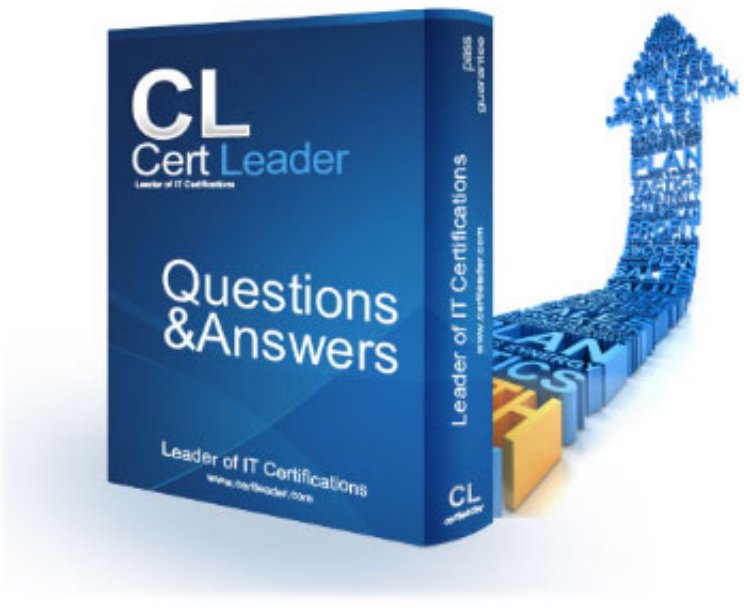
Professional-Data-Engineer Dumps

Google Professional Data Engineer Exam

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



NEW QUESTION 1
- (Exam Topic 1)
Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

A. Issue a command to restart the database servers.
B. Retry the query with exponential backoff, up to a cap of 15 minutes.
C. Retry the query every second until it comes back online to minimize staleness of data.
D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:
https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

NEW QUESTION 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)
Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

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

Answer: A

NEW QUESTION 4
- (Exam Topic 1)
You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.
B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer: D

Explanation:
https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

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 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 7
- (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 8
- (Exam Topic 1)
You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

A. The message body for the sensor event is too large.
B. Your custom endpoint has an out-of-date SSL certificate.
C. The Cloud Pub/Sub topic has too many messages published to it.
D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

Answer: B

NEW QUESTION 9
- (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 10
- (Exam Topic 1)
You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time. What should you do?

A. Send the data to Google Cloud Datastore and then export to BigQuery.
B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Answer: B

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 15
- (Exam Topic 2)
Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?

A. Export the data into a Google Sheet for virtualization.
B. Create an additional table with only the necessary columns.
C. Create a view on the table to present to the virtualization tool.
D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

Answer: C

NEW QUESTION 19
- (Exam Topic 2)
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Answer: C

NEW QUESTION 20
- (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 21
- (Exam Topic 3)
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)

A. Ensure all the tables are included in global dataset.
B. Ensure each table is included in a dataset for a region.
C. Adjust the settings for each table to allow a related region-based security group view access.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Adjust the settings for each dataset to allow a related region-based security group view access.

Answer: BD

NEW QUESTION 22
- (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 26

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

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

Answer: C

NEW QUESTION 27
- (Exam Topic 4)
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

- The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
- The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

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

Answer: A

NEW QUESTION 31
- (Exam Topic 5)
Scaling a Cloud Dataproc cluster typically involves .

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

Answer: A

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

NEW QUESTION 33
- (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)
How can you get a neural network to learn about relationships between categories in a categorical feature?

A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column

Answer: D

Explanation:
There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.
Both of these problems can be solved by representing a categorical feature with an embedding

column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).
You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.
Reference:
https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-dee

NEW QUESTION 40
- (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 45
- (Exam Topic 5)
In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) .

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

Answer: C

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

NEW QUESTION 47
- (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 52
- (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 57
- (Exam Topic 5)

Which of the following is not possible using primitive roles?

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

Answer: C

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

NEW QUESTION 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 62
- (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 65
- (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 67
- (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 68
- (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 69
- (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 71
- (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 74
- (Exam Topic 5)
Cloud Dataproc is a managed Apache Hadoop and Apache service.

A. Blaze
B. Spark
C. Fire
D. Ignite

Answer: B

Explanation:
Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.
Reference: https://cloud.google.com/dataproc/docs/

NEW QUESTION 78
- (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 81
- (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)
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 87
- (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 88
- (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 92
- (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 94
- (Exam Topic 6)
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

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

Answer: C

NEW QUESTION 95
- (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 99
- (Exam Topic 6)
You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

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

Answer: D

NEW QUESTION 102
- (Exam Topic 6)
You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

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

Answer: C

NEW QUESTION 104
- (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 109
- (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 112
- (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 115
- (Exam Topic 6)
Government regulations in the banking industry mandate the protection of client's personally identifiable information (PII). Your company requires PII to be access controlled encrypted and compliant with major data protection standards In addition to using Cloud Data Loss Prevention (Cloud DIP) you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?

A. Assign the required identity and Access Management (IAM) roles to every employee, and create a single service account to access protect resources
B. Use one service account to access a Cloud SQL database and use separate service accounts for each human user
C. Use Cloud Storage to comply with major data protection standard
D. Use one service account shared by all users
E. Use Cloud Storage to comply with major data protection standard
F. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group

Answer: D

NEW QUESTION 120
- (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 124
- (Exam Topic 6)
You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

A. Create an API using App Engine to receive and send messages to the applications
B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them
C. Create a table on Cloud SQL, and insert and delete rows with the job information
D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Answer: A

NEW QUESTION 129
- (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 130
- (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 132
- (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 136
- (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 138
- (Exam Topic 6)
You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

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

Answer: AC

NEW QUESTION 141
- (Exam Topic 6)
You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.
What should you do?

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
B. Place the MariaDB instances in an Instance Group with a Health Check.
C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
D. Install the StackDriver Agent and configure the MySQL plugin.

Answer: C

NEW QUESTION 143
- (Exam Topic 6)
A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?

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

Answer: C

NEW QUESTION 147
- (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 152
- (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 156
- (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 157
- (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 162
- (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 163
- (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 166
- (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 169
- (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 172
- (Exam Topic 6)
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

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

Answer: C

NEW QUESTION 174
- (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 179
- (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 184
- (Exam Topic 6)
A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: SELECT predicted_label, user_id FROM ML.PREDICT (MODEL ‘dataset.model’, table user_features). How should you create the ML pipeline?

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

Answer: D

NEW QUESTION 187
- (Exam Topic 6)
You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results.Deploy the models using Cloud Datapro
B. Call the model from your application.
C. Build and train a classification model with Spark MLlib to generate label
D. Build and train a second classification model with Spark MLlib to filter results to match customer preference
E. Deploy themodels using Cloud Datapro
F. Call the models from your application.
G. Build an application that calls the Cloud Video Intelligence API to generate label
H. Store data in Cloud Bigtable, and filter the predicted labels to match the user’s viewing history to generate preferences.
I. Build an application that calls the Cloud Video Intelligence API to generate label

J. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Answer: C

NEW QUESTION 192
- (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 193
- (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 194
- (Exam Topic 6)
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

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

Answer: A

NEW QUESTION 195
- (Exam Topic 6)
Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

A. Denormalize the data
B. Shard the data by customer ID
C. Materialize the dimensional data in views
D. Partition the data by transaction date

Answer: C

NEW QUESTION 198
- (Exam Topic 6)
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

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

Answer: A

NEW QUESTION 202
- (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 207
- (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 211
- (Exam Topic 6)
You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.
What should you do?

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

Answer: A

NEW QUESTION 215
- (Exam Topic 6)
You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

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

Answer: B

NEW QUESTION 217
- (Exam Topic 6)
You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?

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

Answer: D

NEW QUESTION 221
- (Exam Topic 6)
You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the
Google-recommended cloud native architecture for this scenario?

A. Edge TPUs as sensor devices for storing and transmitting the messages.
B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

Answer: C

NEW QUESTION 225
- (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 226
- (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 231
.....

Thank You for Trying Our Product

* 100% Pass or Money Back

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

* One year free update

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

* Trusted by Millions

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

* Shop Securely

All transactions are protected by VeriSign!

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

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