

Professional-Data-Engineer Dumps

Google Professional Data Engineer Exam

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



NEW QUESTION 1

- (Exam Topic 1)

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristics support this method? (Choose two.)

- A. There are very few occurrences of mutations relative to normal samples.
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.
- D. You expect future mutations to have similar features to the mutated samples in the database.
- E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AD

Explanation:

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. https://en.wikipedia.org/wiki/Anomaly_detection

NEW QUESTION 2

- (Exam Topic 1)

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

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

Answer: B

NEW QUESTION 3

- (Exam Topic 1)

Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

```
# Syntax error : Expected end of statement but got "-" at [4:11] SELECT age
```

```
FROM
```

```
bigquery-public-data.noaa_gsod.gsod WHERE
```

```
age != 99
```

```
AND_TABLE_SUFFIX = '1929' ORDER BY
```

```
age DESC
```

Which table name will make the SQL statement work correctly?

- A. 'bigquery-public-data.noaa_gsod.gsod'
- B. bigquery-public-data.noaa_gsod.gsod*
- C. 'bigquery-public-data.noaa_gsod.gsod'*
- D. 'bigquery-public-data.noaa_gsod.gsod**'

Answer: D

NEW QUESTION 4

- (Exam Topic 1)

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

- A. Load data into different partitions.
- B. Load data into a different dataset for each client.
- C. Put each client's BigQuery dataset into a different table.
- D. Restrict a client's dataset to approved users.
- E. Only allow a service account to access the datasets.
- F. Use the appropriate identity and access management (IAM) roles for each client's users.

Answer: BDF

NEW QUESTION 5

- (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 6

- (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 7

- (Exam Topic 1)

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

- A. Run a local version of Jupiter on the laptop.
- B. Grant the user access to Google Cloud Shell.
- C. Host a visualization tool on a VM on Google Compute Engine.
- D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Answer: B

NEW QUESTION 8

- (Exam Topic 1)

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

- A. Update the current pipeline and use the drain flag.
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Answer: D

NEW QUESTION 9

- (Exam Topic 1)

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

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

Answer: B

NEW QUESTION 10

- (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 10

- (Exam Topic 1)

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

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

Answer: BCD

NEW QUESTION 11

- (Exam Topic 1)

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

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

Answer: C

NEW QUESTION 14

- (Exam Topic 1)

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

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

Answer: A

Explanation:

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

NEW QUESTION 19

- (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 21

- (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 22

- (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 23

- (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 24

- (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 26

- (Exam Topic 3)

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

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

Answer: A

NEW QUESTION 29

- (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 34

- (Exam Topic 3)

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

- A. Create a table called tracking_table and include a DATE column.
- B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
- C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
- D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Answer: B

NEW QUESTION 39

- (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 42

- (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 45

- (Exam Topic 4)

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

A. Manually configure the index in your index config as follows:

```
Indexes:
```

```
-kind: Movie
```

```
  Properties:
```

```
    -name: actors
```

```
    name: date_released
```

```
-kind: Movie
```

```
  Properties:
```

```
    -name: tags
```

```
    name: date_released
```

B. Manually configure the index in your index config as follows:

```
Indexes:
```

```
-kind: Movie
```

```
  Properties:
```

```
    -name: actors
```

```
    -name: tags
```

```
-name: date_published
```

C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'

D. Set the following in your entity options: exclude_from_indexes = 'date_published'

- A. Option A
- B. Option B.
- C. Option C
- D. Option D

Answer: A

NEW QUESTION 50

- (Exam Topic 4)

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

- A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
- B. Add a new column called FullName to the Users tabl
- C. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- D. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
- E. Use BigQuery to export the data for the table to a CSV fil
- F. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullNam
- G. Run a BigQuery load job to load the new CSV file into BigQuery.

Answer: C

NEW QUESTION 53

- (Exam Topic 4)

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

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

Answer: A

NEW QUESTION 54

- (Exam Topic 5)

Which of these is NOT a way to customize the software on Dataproc cluster instances?

- A. Set initialization actions
- B. Modify configuration files using cluster properties
- C. Configure the cluster using Cloud Deployment Manager
- D. Log into the master node and make changes from there

Answer: C

Explanation:

You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.

You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster. When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>]

Reference: <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties>

NEW QUESTION 59

- (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 61

- (Exam Topic 5)

Which action can a Cloud Dataproc Viewer perform?

- A. Submit a job.
- B. Create a cluster.
- C. Delete a cluster.

D. List the jobs.

Answer: D

Explanation:

A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.

Reference: https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

NEW QUESTION 66

- (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 70

- (Exam Topic 5)

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

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

Answer: B

Explanation:

In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.

Reference: <https://cloud.google.com/dataflow/model/programming-model>

NEW QUESTION 73

- (Exam Topic 5)

What are two of the characteristics of using online prediction rather than batch prediction?

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

Answer: BD

Explanation:

Online prediction

Optimized to minimize the latency of serving predictions.

Predictions returned in the response message. Batch prediction

Optimized to handle a high volume of instances in a job and to run more complex models. Predictions written to output files in a Cloud Storage location that you specify.

Reference:

https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION 75

- (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 79

- (Exam Topic 5)

Cloud Bigtable is Google's Big Data database service.

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

Answer: C

Explanation:

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

Reference: <https://cloud.google.com/bigtable/>

NEW QUESTION 80

- (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 83

- (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 86

- (Exam Topic 5)

Which is not a valid reason for poor Cloud Bigtable performance?

- A. The workload isn't appropriate for Cloud Bigtable.
- B. The table's schema is not designed correctly.
- C. The Cloud Bigtable cluster has too many nodes.
- D. There are issues with the network connection.

Answer: C

Explanation:

The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.

Reference: <https://cloud.google.com/bigtable/docs/performance>

NEW QUESTION 91

- (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/Triggers>

NEW QUESTION 96

- (Exam Topic 5)

Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.

SELECT person FROM `project1.example.table1` WHERE city = "London" How would you correct the error?

- A. Add ", UNNEST(person)" before the WHERE clause.
- B. Change "person" to "person.city".
- C. Change "person" to "city.person".
- D. Add ", UNNEST(city)" before the WHERE clause.

Answer: A

Explanation:

To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference:

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_results

NEW QUESTION 100

- (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 102

- (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 105

- (Exam Topic 5)

When you design a Google Cloud Bigtable schema it is recommended that you .

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

Answer: C

Explanation:

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

NEW QUESTION 107

- (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 108

- (Exam Topic 5)

Which of the following IAM roles does your Compute Engine account require to be able to run pipeline jobs?

- A. dataflow.worker
- B. dataflow.compute
- C. dataflow.developer
- D. dataflow.viewer

Answer: A

Explanation:

The dataflow.worker role provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline

Reference: <https://cloud.google.com/dataflow/access-control>

NEW QUESTION 113

- (Exam Topic 5)

Which of the following is NOT one of the three main types of triggers that Dataflow supports?

- A. Trigger based on element size in bytes
- B. Trigger that is a combination of other triggers
- C. Trigger based on element count
- D. Trigger based on time

Answer: A

Explanation:

There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way

Reference: <https://cloud.google.com/dataflow/model/triggers>

NEW QUESTION 114

- (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 117

- (Exam Topic 5)

When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and .

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

Answer: A

Explanation:

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

NEW QUESTION 118

- (Exam Topic 5)

Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

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

Answer: A

Explanation:

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

NEW QUESTION 121

- (Exam Topic 5)

Which of the following is NOT a valid use case to select HDD (hard disk drives) as the storage for Google Cloud Bigtable?

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

Answer: C

Explanation:

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

NEW QUESTION 122

- (Exam Topic 5)

Which of these statements about BigQuery caching is true?

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

Answer: D

Explanation:

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

NEW QUESTION 126

- (Exam Topic 5)

What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

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

Answer: C

Explanation:

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

NEW QUESTION 131

- (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 134

- (Exam Topic 5)

What are two of the benefits of using denormalized data structures in BigQuery?

- A. Reduces the amount of data processed, reduces the amount of storage required
- B. Increases query speed, makes queries simpler
- C. Reduces the amount of storage required, increases query speed
- D. Reduces the amount of data processed, increases query speed

Answer: B

Explanation:

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINS on large tables, but with a denormalized data

structure, you don't have to use JOINS, since all of the data has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses.

Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.

Reference:

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION 138

- (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 139

- (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 144

- (Exam Topic 5)

Which of these are examples of a value in a sparse vector? (Select 2 answers.)

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

Answer: CD

Explanation:

Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).

[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1. [0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s. Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION 146

- (Exam Topic 5)

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

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

Answer: C

Explanation:

It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,

If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.

It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

Reference: <https://cloud.google.com/bigtable/docs/creating-compute-instance>

NEW QUESTION 147

- (Exam Topic 5)

By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?

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

Answer: B

Explanation:

Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections

Reference: <https://cloud.google.com/dataflow/model/pcollection>

NEW QUESTION 151

- (Exam Topic 5)

Which of these numbers are adjusted by a neural network as it learns from a training dataset (select 2 answers)?

- A. Weights
- B. Biases
- C. Continuous features
- D. Input values

Answer: AB

Explanation:

A neural network is a simple mechanism that's implemented with basic math. The only difference between the traditional programming model and a neural network is that you let the computer determine the parameters (weights and bias) by learning from training datasets.

Reference:

<https://cloud.google.com/blog/big-data/2016/07/understanding-neural-networks-with-tensorflow-playground>

NEW QUESTION 155

- (Exam Topic 5)

Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

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

Answer: AB

Explanation:

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

Reference: <https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html>

NEW QUESTION 157

- (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 158

- (Exam Topic 5)

Google Cloud Bigtable indexes a single value in each row. This value is called the .

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

Answer: C

Explanation:

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key.

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

NEW QUESTION 159

- (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 163

- (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 166

- (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 169

- (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 171

- (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 173

- (Exam Topic 5)

When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a proxy.

- A. HTTPS
- B. VPN
- C. SOCKS
- D. HTTP

Answer: C

Explanation:

When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.

Reference: <https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces>

NEW QUESTION 176

- (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 178

- (Exam Topic 6)

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

- A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.
- B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.
- C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.
- D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

Answer: B

NEW QUESTION 180

- (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 183

- (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 department
- 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 department
- 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 department
- 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 department
- 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 187

- (Exam Topic 6)

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

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

Answer: D

Explanation:

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

NEW QUESTION 188

- (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 190

- (Exam Topic 6)

You need ads data to serve AI models and historical data for analytics. Longtail and outlier data points need to be identified. You want to cleanse the data in near-real time before running it through AI 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 for processing, and BigQuery to create views for desired datasets.
- C. Use Dataflow to identify longtail and outlier 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 193

- (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 table.
- 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 198

- (Exam Topic 6)

You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL. What should you do?

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

Answer: B

NEW QUESTION 203

- (Exam Topic 6)

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

- A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
- B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
- C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
- D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Answer: D

NEW QUESTION 205

- (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 accessed, controlled, encrypted, and compliant with major data protection standards. In addition to using Cloud Data Loss Prevention (Cloud DLP), 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 protected 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 standards.
- D. Use one service account shared by all users.
- E. Use Cloud Storage to comply with major data protection standards.
- F. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group.

Answer: D

NEW QUESTION 207

- (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 211

- (Exam Topic 6)

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

- A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
- B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
- C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
- D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API.
- E. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Answer: D

NEW QUESTION 214

- (Exam Topic 6)

You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

- A. Increase the number of max workers
- B. Use a larger instance type for your Cloud Dataflow workers
- C. Change the zone of your Cloud Dataflow pipeline to run in us-central1
- D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data
- E. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery
- F. Create a temporary table in Cloud Spanner that will act as a buffer for new data
- G. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Answer: AB

NEW QUESTION 215

- (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 219

- (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 222

- (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 DDL
- B. Partition the tables by a column containing a TIMESTAMP or DATETIME.
- 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 day

F. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Answer: C

NEW QUESTION 223

- (Exam Topic 6)

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

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

Answer: D

NEW QUESTION 224

- (Exam Topic 6)

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products or features of the platform. What should you do?

- A. Export the information to Cloud Stackdriver, and set up an Alerting policy
- B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
- C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs
- D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Answer: B

NEW QUESTION 226

- (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 227

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 230

- (Exam Topic 6)

You've migrated a Hadoop job from an on-prem cluster to Dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffling operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.

What should you do?

- A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- B. Switch to TFRecords formats (approx 200MB per file) instead of parquet files.
- C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
- D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Answer: A

NEW QUESTION 235

- (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 237

- (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 cluster
- 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 archive
- F. Locate the bottleneck and adjust cluster resources.
- G. Use Cloud Dataflow to run your transformation
- H. Monitor the job system lag with Stackdriver
- 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 238

- (Exam Topic 6)

You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- A. Call the Cloud Natural Language API from your application
- B. Process the generated Entity Analysis as labels.
- C. Call the Cloud Natural Language API from your application
- D. Process the generated Sentiment Analysis as labels.
- E. Build and train a text classification model using TensorFlow
- F. Deploy the model using Cloud Machine Learning Engine
- G. Call the model from your application and process the results as labels.
- H. Build and train a text classification model using TensorFlow
- I. Deploy the model using a Kubernetes Engine cluster
- J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 241

- (Exam Topic 6)

You need to give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID. This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline. There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system. How should you design your pipeline to minimize that backpressure?

- A. Call out to the service via HTTP
- B. Create the pipeline statically in the class definition
- C. Create a new object in the startBundle method of DoFn
- D. Batch the job into ten-second increments

Answer: A

NEW QUESTION 245

- (Exam Topic 6)

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? Choose 2 answers.

- A. Use Cloud Deployment Manager to automate access provision.
- B. Introduce resource hierarchy to leverage access control policy inheritance.
- C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
- D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
- E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access
- F. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

Answer: AC

NEW QUESTION 250

- (Exam Topic 6)

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_YYYYMMDD. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you

discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

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

Answer: A

NEW QUESTION 255

- (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 schema
- C. Merge the transformed tables together with a SQL query
- D. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation
- 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 260

- (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 265

- (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 PCollection that can be stored to PubSub later.

Answer: C

NEW QUESTION 269

- (Exam Topic 6)

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period.

However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

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

Answer: B

NEW QUESTION 273

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 274

- (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 project
- 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 project 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 275

- (Exam Topic 6)

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Cloud Dataflow pipeline to use local execution
- B. Increase the maximum number of Cloud Dataflow workers by setting `maxNumWorkers` in `PipelineOptions`
- C. Increase the number of nodes in the Cloud Bigtable cluster
- D. Modify your Cloud Dataflow pipeline to use the `Flatten` transform before writing to Cloud Bigtable
- E. Modify your Cloud Dataflow pipeline to use the `CoGroupByKey` transform before writing to Cloud Bigtable

Answer: BC

NEW QUESTION 277

- (Exam Topic 6)

You work for a bank. You have a labelled dataset that contains information on already granted loan applications 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 279

- (Exam Topic 6)

Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

- A. Cloud Bigtable
- B. Google BigQuery
- C. Google Cloud Storage
- D. Google Cloud Datastore

Answer: A

Explanation:

Reference: <https://cloud.google.com/bigtable/docs/schema-design-time-series>

NEW QUESTION 282

- (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 284

- (Exam Topic 6)

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

- A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
- B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
- C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
- D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Answer: C

NEW QUESTION 289

- (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 292

- (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 efficiently import the data into BigQuery where consuming as few resources as possible. What should you do?

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

Answer: D

NEW QUESTION 295

- (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 298

- (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 storage
- B. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
- C. Use Cloud Bigtable for storage
- D. Link as permanent tables in BigQuery for query.
- E. Use Cloud Storage for storage
- F. Link as permanent tables in BigQuery for query.
- G. Use Cloud Storage for storage

H. Link as temporary tables in BigQuery for query.

Answer: A

NEW QUESTION 302

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 304

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 307

- (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 311

- (Exam Topic 6)

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.

What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- C. Create a service account and grant dataset access to that account
- D. Use the service account's private key to access the dataset
- E. Create a dummy user and grant dataset access to that user
- F. Store the username and password for that user in a file on the file system, and use those credentials to access the BigQuery dataset

Answer: C

NEW QUESTION 314

- (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 315

- (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 query
- 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 query
- 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 BigtableIO
- 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 318

- (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 Dataproc
- B. Call the model from your application.
- C. Build and train a classification model with Spark MLlib to generate labels
- D. Build and train a second classification model with Spark MLlib to filter results to match customer preference
- E. Deploy the models using Cloud Dataproc
- F. Call the models from your application.
- G. Build an application that calls the Cloud Video Intelligence API to generate labels
- 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 labels
- 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 323

- (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 327

- (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 328

- (Exam Topic 6)

You have uploaded 5 years of log data to Cloud Storage. A user reported that some data points in the log data are outside of their expected ranges, which indicates errors. You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

- A. Import the data from Cloud Storage into BigQuery. Create a new BigQuery table, and skip the rows with errors.
- B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage. Skip the rows with errors.
- C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.
- D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage.

Answer: D

NEW QUESTION 333

- (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 337

- (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 342

- (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 347

- (Exam Topic 6)

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on nonkey columns. What should you do?

- A. Use Cloud SQL for storag
- B. Add secondary indexes to support query patterns.
- C. Use Cloud SQL for storag
- D. Use Cloud Dataflow to transform data to support query patterns.
- E. Use Cloud Spanner for storag
- F. Add secondary indexes to support query patterns.
- G. Use Cloud Spanner for storag
- H. Use Cloud Dataflow to transform data to support query patterns.

Answer: D

Explanation:

Reference: <https://cloud.google.com/solutions/data-lifecycle-cloud-platform>

NEW QUESTION 348

- (Exam Topic 6)

You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

- A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size
- B. Increase the see of your parquet files to ensure them to be 1 GB minimum
- C. Switch to TFRecords format (appr 200 MB per We) instead of parquet files
- D. Switch from HDDs to SSD
- E. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

Answer: A

NEW QUESTION 349

- (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 353

- (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 355

- (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 357

.....

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>