

Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam



NEW QUESTION 1

- (Exam Topic 1)

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

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

Answer: C

NEW QUESTION 2

- (Exam Topic 1)

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

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

Answer: A

NEW QUESTION 3

- (Exam Topic 1)

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.

Which Google database service should you use?

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

Answer: A

NEW QUESTION 4

- (Exam Topic 1)

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

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

Answer: B

NEW QUESTION 5

- (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 6

- (Exam Topic 1)

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

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

Answer: BCE

NEW QUESTION 7

- (Exam Topic 1)

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

- A. Disable writes to certain tables.
- B. Restrict access to tables by role.
- C. Ensure that the data is encrypted at all times.
- D. Restrict BigQuery API access to approved users.
- E. Segregate data across multiple tables or databases.
- F. Use Google Stackdriver Audit Logging to determine policy violations.

Answer: BDF

NEW QUESTION 8

- (Exam Topic 1)

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

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

Answer: C

NEW QUESTION 9

- (Exam Topic 1)

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

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

Answer: D

NEW QUESTION 10

- (Exam Topic 1)

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

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

Answer: C

Explanation:

Reference

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

NEW QUESTION 10

- (Exam Topic 1)

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

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

Answer: A

NEW QUESTION 13

- (Exam Topic 1)

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and serves millions of users. How should you design the frontend to respond to a database failure?

- A. Issue a command to restart the database servers.

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

Answer: B

NEW QUESTION 14

- (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 16

- (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 19

- (Exam Topic 4)

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.

You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

- A. Redis
- B. HBase
- C. MySQL
- D. MongoDB
- E. Cassandra
- F. HDFS with Hive

Answer: BDF

NEW QUESTION 21

- (Exam Topic 4)

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

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

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

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

Answer: A

NEW QUESTION 26

- (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 28

- (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 33

- (Exam Topic 5)

What are all of the BigQuery operations that Google charges for?

- A. Storage, queries, and streaming inserts
- B. Storage, queries, and loading data from a file
- C. Storage, queries, and exporting data
- D. Queries and streaming inserts

Answer: A

Explanation:

Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.

Reference: <https://cloud.google.com/bigquery/pricing>

NEW QUESTION 34

- (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 38

- (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 40

- (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 42

- (Exam Topic 5)

Which of the following statements about Legacy SQL and Standard SQL is not true?

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

Answer: D

Explanation:

You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.

Standard SQL has been the preferred query language since BigQuery 2.0 was released.

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.

Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

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

NEW QUESTION 46

- (Exam Topic 5)

Which Google Cloud Platform service is an alternative to Hadoop with Hive?

- A. Cloud Dataflow
- B. Cloud Bigtable
- C. BigQuery
- D. Cloud Datastore

Answer: C

Explanation:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.

Google BigQuery is an enterprise data warehouse. Reference: https://en.wikipedia.org/wiki/Apache_Hive

NEW QUESTION 50

- (Exam Topic 5)

What are two methods that can be used to denormalize tables in BigQuery?

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

Answer: B

Explanation:

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

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

NEW QUESTION 52

- (Exam Topic 5)

Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?

- A. Use K-means Clustering to detect faces in the pixels.
- B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
- C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
- D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: C

Explanation:

Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.

In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.

A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.

Feature engineering is difficult to perform on raw image data.

K- means Clustering is an unsupervised learning method used to categorize unlabeled data. Reference: <https://deeplearning4j.org/neuralnet-overview>

NEW QUESTION 53

- (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 56

- (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 61

- (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 66

- (Exam Topic 5)

When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?

- A. 500 TB
- B. 1 GB
- C. 1 TB
- D. 500 GB

Answer: C

Explanation:

Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

Reference: https://cloud.google.com/bigtable/docs/overview#title_short_and_other_storage_options

NEW QUESTION 68

- (Exam Topic 5)

To give a user read permission for only the first three columns of a table, which access control method would you use?

- A. Primitive role
- B. Predefined role

- C. Authorized view
- D. It's not possible to give access to only the first three columns of a table.

Answer: C

Explanation:

An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view. When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.
Reference: <https://cloud.google.com/bigquery/docs/views#authorized-views>

NEW QUESTION 73

- (Exam Topic 5)

Which of the following are feature engineering techniques? (Select 2 answers)

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

Answer: CD

Explanation:

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

NEW QUESTION 77

- (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 82

- (Exam Topic 5)

If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

- A. 1 continuous and 2 categorical
- B. 3 categorical
- C. 3 continuous
- D. 2 continuous and 1 categorical

Answer: D

Explanation:

The columns can be grouped into two types—categorical and continuous columns:
A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.
A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. \$14,084) is a continuous column.
Year of birth and income are continuous columns. Country is a categorical column.
You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.
Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

NEW QUESTION 83

- (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 85

- (Exam Topic 5)

When a Cloud Bigtable node fails, is lost.

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

Answer: B

Explanation:

A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.

Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:

Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node.

Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.

When a Cloud Bigtable node fails, no data is lost Reference: <https://cloud.google.com/bigtable/docs/overview>

NEW QUESTION 87

- (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 88

- (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 92

- (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 97

- (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 99

- (Exam Topic 5)

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

```
SELECT person FROM `project1.example.table1` WHERE city = "London"
```

How would you correct the error?

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

Answer: A

Explanation:

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

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

NEW QUESTION 102

- (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 105

- (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 108

- (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: DE

NEW QUESTION 111

- (Exam Topic 6)

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

- A. Use bq load to load a batch of sensor data every 60 seconds.
- B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
- C. Use the INSERT statement to insert a batch of data every 60 seconds.
- D. Use the MERGE statement to apply updates in batch every 60 seconds.

Answer: C

NEW QUESTION 114

- (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 119

- (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 121

- (Exam Topic 6)

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

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

Answer: D

NEW QUESTION 122

- (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 124

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 127

- (Exam Topic 6)

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

- A. Create a new view over events using standard SQL
- B. Create a new partitioned table using a standard SQL query
- C. Create a new view over events_partitioned using standard SQL
- D. Create a service account for the ODBC connection to use for authentication
- E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connectionand shared “events”

Answer: AE

NEW QUESTION 130

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 134

- (Exam Topic 6)

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account. What should you do?

- A. Convert your batch BQ queries into interactive BQ queries.
- B. Create an additional project to overcome the 2K on-demand per-project quota.
- C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.
- D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Answer: C

Explanation:

Reference <https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery>

NEW QUESTION 139

- (Exam Topic 6)

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

- A. cron
- B. Cloud Composer
- C. Cloud Scheduler
- D. Workflow Templates on Cloud Dataproc


Answer: D

NEW QUESTION 142

- (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 143

- (Exam Topic 6)

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

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

Answer: CE

NEW QUESTION 144

- (Exam Topic 6)

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

- A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.
- B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
- C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
- D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Answer: D

NEW QUESTION 145

- (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 146

- (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 148

- (Exam Topic 6)

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

- A. Transform text files to compressed Avro using Cloud Dataflo
- B. Use BigQuery for storage and query.
- C. Transform text files to compressed Avro using Cloud Dataflo
- D. Use Cloud Storage and BigQuerypermanent linked tables for query.
- E. Compress text files to gzip using the Grid Computing Tool

- F. Use BigQuery for storage and query.
- G. Compress text files to gzip using the Grid Computing Tool
- H. Use Cloud Storage, and then import into Cloud Bigtable for query.

Answer: D

NEW QUESTION 152

- (Exam Topic 6)

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

- ▶ You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- ▶ You will extract topics and sentiment from the posts.
- ▶ You must store the raw posts for archiving and reprocessing.
- ▶ You will create dashboards to be shared with people both inside and outside your organization.

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

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

Answer: D

NEW QUESTION 156

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 158

- (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 163

- (Exam Topic 6)

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

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

Answer: DE

NEW QUESTION 164

- (Exam Topic 6)

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of data. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

- A. Encrypted on Cloud Storage with user-supplied encryption key
- B. A separate decryption key will be given to each authorized user.
- C. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
- D. In Cloud SQL, with separate database user names to each use

E. The Cloud SQL Admin activity logs will be used to provide the auditability.

F. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Answer: B

NEW QUESTION 167

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 169

- (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 173

- (Exam Topic 6)

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query – -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

- A. Create a separate table for each ID.
- B. Use the LIMIT keyword to reduce the number of rows returned.
- C. Recreate the table with a partitioning column and clustering column.
- D. Use the bq query - -maximum_bytes_billed flag to restrict the number of bytes billed.

Answer: B

NEW QUESTION 175

- (Exam Topic 6)

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

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

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

Answer: A

NEW QUESTION 176

- (Exam Topic 6)

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

- A. Cloud Spanner
- B. Cloud Bigtable
- C. Cloud Firestore
- D. Cloud SQL

Answer: D

NEW QUESTION 179

- (Exam Topic 6)

You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are likely to cause delays within a region.

You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?

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

Answer: A

NEW QUESTION 182

- (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 of 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 185

- (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 187

- (Exam Topic 6)

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

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

Answer: C

NEW QUESTION 189

- (Exam Topic 6)

An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

- A. Create and share an authorized view that provides the aggregate results.
- B. Create and share a new dataset and view that provides the aggregate results.
- C. Create and share a new dataset and table that contains the aggregate results.
- D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Answer: D

Explanation:

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

NEW QUESTION 193

- (Exam Topic 6)

You are designing a cloud-native historical data processing system to meet the following conditions:

- ▶ The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
- ▶ A streaming data pipeline stores new data daily.
- ▶ Performance is not a factor in the solution.
- ▶ The solution design should maximize availability.

How should you design data storage for this solution?

- A. Create a Cloud Dataproc cluster with high availabilit

- B. Store the data in HDFS, and perform analysis as needed.
- C. Store the data in BigQuery
- D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
- E. Store the data in a regional Cloud Storage bucket
- F. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
- G. Store the data in a multi-regional Cloud Storage bucket
- H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: C

NEW QUESTION 197

- (Exam Topic 6)

You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

- A. Set the BigQuery dataset to be regional
- B. In the event of an emergency, use a point-in-time snapshot to recover the data.
- C. Set the BigQuery dataset to be regional
- D. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup
- E. In the event of an emergency, use the backup copy of the table.
- F. Set the BigQuery dataset to be multi-regional
- G. In the event of an emergency, use a point-in-time snapshot to recover the data.
- H. Set the BigQuery dataset to be multi-regional
- I. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup
- J. In the event of an emergency, use the backup copy of the table.

Answer: B

NEW QUESTION 201

- (Exam Topic 6)

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

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

Answer: A

NEW QUESTION 204

- (Exam Topic 6)

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

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

Answer: B

NEW QUESTION 209

- (Exam Topic 6)

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

- A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
- B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
- C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Function
- D. Integrate the package tracking applications with this function.
- E. Use TensorFlow to create a model that is trained on your corpus of images
- F. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer: A

NEW QUESTION 210

- (Exam Topic 6)

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

- A. Use gcloud kms keys create to create a symmetric key
- B. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
- C. Use gcloud kms keys create to create a symmetric key
- D. Then use gcloud kms encrypt to encrypt each archival file with the key

- E. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket
- F. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.
- G. Specify customer-supplied encryption key (CSEK) in the .boto configuration file
- H. Use gsutil cp to upload each archival file to the Cloud Storage bucket
- I. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
- J. Specify customer-supplied encryption key (CSEK) in the .boto configuration file
- K. Use gsutil cp to upload each archival file to the Cloud Storage bucket
- L. Save the CSEK in a different project that only the security team can access.

Answer: B

NEW QUESTION 211

.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questions and Answers in PDF Format

Professional-Data-Engineer Practice Exam Features:

- * Professional-Data-Engineer Questions and Answers Updated Frequently
- * Professional-Data-Engineer Practice Questions Verified by Expert Senior Certified Staff
- * Professional-Data-Engineer Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * Professional-Data-Engineer Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The Professional-Data-Engineer Practice Test Here](#)