Question 1Skipped

Data analysts in your company use BigQuery as their structured database supporting SQL for querying. Recently, they started to study about machine learning and what possibilities they can achieve if they can apply some machine learning algorithms such as binary classification on existing data stored in BigQuery for customer behavior prediction. You are to assist preparing the required tools and resources for data analysts to apply machine learning algorithms on data. What should you do?

A. Use Vertex AI to build and deploy your binary classification model

Correct answer

B. Use BigQuery ML by building the binary classification model by specifying the model type and labels when creating the model.

C. Use Apache Spark MLLib with Dataproc to build the model. Use Dataproc-BigQuery connector to allow the cluster reading data.

D. Use Cloud ML Engine to build the classification model(s) using TensorFlow, allowing the model's service account permission to read the data from BigQuery.

Overall explanation

Answer: B.

Description:

BigQuery ML enables users to create and execute machine learning models in BigQuery using standard SQL queries. BigQuery ML democratizes machine learning by enabling SQL practitioners to build models using existing SQL tools and skills. BigQuery ML increases development speed by eliminating the need to move data.

BigQuery ML empowers data analysts to use machine learning through existing SQL tools and skills. Analysts can use BigQuery ML to build and evaluate ML models in BigQuery. Analysts no longer need to export small amounts of data to a spreadsheets or other applications, and analysts no longer need to wait for limited resources from a data science team.

Source(s):

Introduction to BigQuery ML:

https://cloud.google.com/bigquery-ml/docs/bigqueryml-intro

Domain

Ingest and process the data

Question 2Skipped

A coach line bus service company wants to predict how many passengers they expect to book for tickets on their buses for the upcoming months. This helps the company to know how many buses they need to be in service for maintenance and fuel and how many drivers to be available. The company has data sets of all booked tickets since its launch in 1968 and it allows private sharing of the data if this helps the prediction process.

You will build the machine learning model for the coach line company. Which technique you will use to predict the number of passengers in the next months?

Correct answer

A. Regression.

B. Association.

C. Classification.

D. Clustering.

Overall explanation

Answer: A.

A regression problem is a problem which its output variable is of continuous value. Problems which finds out about variables such as weights, prices or age are considered regression problems.

A classification problem is a problem which the output variable is a category. Examples of classification problems are finding a passenger's nationality, detect if a patient is diagnosed with a disease or if an applicant is qualified for a job interview.

Regression and classification are supervised learning problems. It means, the machine learns from past experiences by training it on a labeled data set. A training set is a set of rows with input and output parameters. The machine then learns from the training set and improves its parameters for better detection.

Association is a rule-learning technique for discovering interesting relations between variables in large data sets. Example of association rules is discovering regularities between products in large-scale transaction data recorded by point-of-sales for a retail chain store.

Clustering is an unsupervised learning method. An unsupervised learning is a method to find references between input data without labeled output. The purpose is to find meaningful structure between the input sets with similar features and group them. Clustering is the method of grouping data points share similarities and separating dissimilar points to other groups. Examples of clustering applications are customer segmentation (new, frequent, loyal, ..), city land value and detecting anomalies in network traffic.

From the explanation above, the technique to help solving the scenario is Answer A: Regression.

Domain

Ingest and process the data

Question 3Skipped

A company has over 25TB of data in Avro format stored in on-premise disks. You are migrating the tech stack used to Google Cloud. The current data pipeline built on-premise does the required data transformation and enrichment using Apache Spark. You decide to use Dataproc for data processing. When the migration was approved by the management, one of the base requirements was for data to be highly available and cross-zone durability should be guaranteed. What should you do?

Correct answer

A. Use Google Storage to store data. Allow Dataproc cluster to access data from Google Storage.

B. Use BigQuery to store data. Install Dataproc-BigQuery connector to access data.

C. Use Dataproc cluster's HDFS namenodes to store data.

D. Use BigTable to store data. Use Dataproc-BigTable connector to access data.

Overall explanation

Answer: A.

Description:

When you want to move Hadoop & Spark workloads from an on-premises environment to Google Cloud Platform (GCP), It's recommended to use Dataproc to run Apache Spark & Hadoop clusters.

Cloud Storage is a good option if:

- Your data in ORC, Parquet, Avro, or any other format will be used by different clusters or jobs, and you need data persistence if the cluster terminates.
- You need high throughput and your data is stored in files larger than 128 MB.
- You need cross-zone durability for your data.
- You need data to be highly available—for example, you want to eliminate HDFS NameNode as a single point of failure.

Source(s):

Migrating Apache Spark Jobs to Cloud Dataproc:

https://cloud.google.com/solutions/migration/hadoop/ migrating-apache-spark-jobs-to-cloud-dataproc

Domain

Design Data Processing Systems

Question 4Skipped

You want to build a system which uses a machine learning, image recognition model to detect customers' faces entering a retail shop, and based on the knowledge base it will return whether the customer is a new, returning or loyal customer. You are building the model using AutoML Vision. After training the model and testing it, you find the model's accuracy is lower due to overfitting. How can you solve this?

A. Images used should be taken from the same exact angle and resolution.

B. Instead of manually splitting samples to training and testing sets, allow AutoML Vision to split the sample set.

C. Samples used for training should be covering true positives only.

Correct answer

D. Images used should be taken from different angles, resolutions and points of view.

Overall explanation

Answer: D.

Description:

Google Cloud provies a machine learning service called AutoML to quickly build models for you. AutoML Vision is one of its products which you can start with a training set as little as a dozen photo samples and AutoML takes care of the rest.

While iterating on your model, if the model's quality levels are not up to expectations, you can go back to earlier steps to improve the quality:

- AutoML Vision allows you to sort the images by how "confused" the model is, by the true label and its predicted label. Look through these images and make sure they're labeled correctly.
- Consider adding more images to any labels with low quality.
- You may need to add different types of images (e.g. wider angle, higher or lower resolution, different points of view).
- Consider removing labels altogether if you don't have enough training images.
- Remember that machines can't read your label name; it's just a random string of letters to them. If you have one label that says "door" and another that says "door_with_knob" the machine has no way of figuring out the nuance other than the images you provide it.
- Augment your data with more examples of true positives and negatives. Especially important examples are the ones that are close to the decision boundary (i.e. likely to produce confusion, but still correctly labeled).
- Specify your own TRAIN, TEST, VALIDATION split. The tool randomly assigns images, but near-duplicates may end up in TRAIN and VALIDATION which could lead to overfitting and then poor performance on the TEST set.

Once you've made changes, train and evaluate a new model until you reach a high enough quality level.

Source(s):

Cloud AutoML Vision – Evaluating Models:

https://cloud.google.com/vision/automl/docs/evaluate

Domain

Ingest and process the data

Question 5Skipped

You are preparing a dataset as a training set for a Machine Learning Model. You have the following columns chosen as features for the model:

- Zip code
- Age Group
- Income

Which feature type each column is:

A. 3 continuous.

Correct answer

B. 2 categorical, 1 continuous.

C. 2 continuous, 1 categorical.

D. 3 categorical.

Overall explanation

Correct Answer: B

In machine learning, features are two types: Categorical and Continuous.

Categorical features are features with finite values. For example Country, education level, age group, and marital status.

Continuous features are features with numeric values in a continuous range. For example Income, latitude & longitude, and time.

For zip code, while it's represented as numeric values, it's considered categorical because it represents regions, which means, it marks each region with a number.

Domain

Ingest and process the data

Question 6Skipped

You need to deploy a machine learning model built by data science team in the firm you work for. As a data engineer, you will be responsible of monitoring the health and traffic of the hosted model on the cloud. Some jobs could fail due to several reasons and you should be able to alert data scientists of such failed jobs.

Which of the following approaches is best to implement on Google Cloud?

Correct answer

A. Use Vertex AI to host the model. Use Cloud Monitoring to monitor the status of jobs for 'failed' status.

B. Use Google Kubernetes Engine to host the model. Use Stackdriver to monitor the status of jobs for 'failed' status.

C. Use AutoML Vision to host the model. Use Stackdriver to monitor the status of jobs for 'failed' status.

D. Use Google Kubernetes Engine to host the model. Use Stackdriver to monitor the status of operations for 'error' status.

Overall explanation

Correct Answer: A

Google Kubernetes Engine is a managed, production-ready environment for deploying containerized applications. It brings our latest innovations in developer productivity, resource efficiency, automated operations, and open-source flexibility to accelerate your time to market.

Vertex AI is a unified machine learning (ML) platform that enables developers and data scientists to build, train, deploy, and manage ML models. It provides a wide range of tools and services for ML development, including:

- AutoML: AutoML allows users to train ML models without writing any code. It provides pre-built algorithms and workflows for common ML tasks, such as image classification, text classification, and object detection.
- Custom training: Vertex AI also provides tools for custom ML training, including support for popular ML frameworks such as TensorFlow, PyTorch, and scikit-learn.
- Model deployment: Vertex AI provides tools for deploying ML models to production, including support for serving models through APIs, web applications, and mobile devices.
- Model Management: Vertex AI provides tools for managing ML models throughout their lifecycle, including versioning, monitoring, and retraining.

Cloud Monitoring is a service that provides a unified view of the performance, health, and availability of your cloud resources. It collects metrics, logs, and events from your resources and provides you with tools to analyze and visualize this data.

In this scenario, You should use Vertex AI to deploy the model to the cloud. So, answers B, C & D are incorrect.

References:

Vertex AI : https://cloud.google.com/vertex-ai

Google AutoML: https://cloud.google.com/automl/

Google Machine Learning Engine: (now called Cloud AI Platform) https://cloud.google.com/ml-engine/

Google Kubernetes Engine: https://cloud.google.com/kubernetes-engine/

Domain

Ingest and process the data

Question 7Skipped

You want to launch a Cloud Machine Learning Engine cluster to deploy a deep neural network model built by Tensorflow by data scientists of your company. Reviewing the standard tiers available by Google ML Engine, you could not find a tier that suits the requirements data scientists need for the cluster. Google allows you to specify custom cluster specification.

Which of the following specifications you are allowed to set? (Choose 2)

Correct selection

A. workerCount

Correct selection

B. parameterServerCount

C. masterCount

D. workerMemory

Overall explanation

Answers: A & B.

The Custom tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:

-
    - You must set TrainingInput.masterType to specify the type of machine to use for your master node. This is the only required setting. See the machine types described below.
    - You may set TrainingInput.*workerCount* to specify the number of workers to use. If you specify one or more workers, you must also set TrainingInput.workerType to specify the type of machine to use for your worker nodes.
    - You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use. If you specify one or more parameter servers, you must also set TrainingInput.*parameterServerType* to specify the type of machine to use for your parameter servers.

From the explanation, specifications can be set from the answers are workerCount &

parameterServerCount.

Source(s):

Specifying Machine Types or Scale Tiers: https://cloud.google.com/ml-engine/docs/tensorflow/machine-types

Domain

Ingest and process the data

Question 8Skipped

Currently, anyone can access and modify the data sets, as well as creating & deleting data sets. The security team wants to restrict access of users on BigQuery and assign the minimum roles for each team based on their task requirements. You have the following teams:

-
    -
        - Data scientists: They should have read & write access to data sets. They may create & delete data sets.

- Data analysts: They have read access to data sets only.
- Development team: They need to create jobs to run queries for updating the website's stats and product details.

What roles are recommended for each team?

Correct answer

A. Assign 'roles/bigquery.dataOwner' role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.user'role to the development team.

B. Assign 'roles/bigquery.admin' role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.dataViewer' role to development team.

C. Assign owner role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.dataViewer' role to development team.

D. Assign admin role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.user' role to development team.

Overall explanation

Answer: A.

Here is the table of BigQuery roles and each role's permissions:

Capability	dataViewer	dataEditor	dataOwner	metadataViewer	user	jobUser	admin
List/get projects	✓	✓	✓	✓	✓	✓	✓
List tables	✓	✓	✓	✓	✓	⊘	✓
Get table metadata	✓	✓	✓	✓	⊘	⊘	✓
Get table data	✓	✓	✓	⊘	⊘	⊘	✓
Create tables	⊘	✓	✓	⊘	⊘	⊘	✓
Modify/delete tables	⊘	✓	✓	⊘	⊘	⊘	✓
Get dataset metadata	✓	✓	✓	✓	✓	⊘	✓
Create new datasets	⊘	✓	✓	⊘	✓	⊘	✓
Modify/delete datasets	⊘	⊘	✓	⊘	Self-created datasets	⊘	✓
Create jobs/queries	⊘	⊘	⊘	⊘	✓	✓	✓
Get jobs	⊘	⊘	⊘	⊘	⊘	Self-created jobs	Any jobs
List jobs	⊘	⊘	⊘	⊘	Any jobs (jobs from other users are	⊘	Any jobs

From the table, we can assign each group the right role:

- Data scientists: They need to read & write to tables as well as creating and deleting datasets. So, owner & admin roles are the required roles. Since it's best practice to grant the role with least permissions, we use owner role for this group.
- Data analysts: They need read access to data only. Viewer role is the permission allows them to read data from datasets.
- Development team: They need to create jobs on BigQuery. From the table, we see that user & jobUser roles may be suitable permissions. However, for developers, jobUser can be too tight for developers since they need to list tables and jobs. The answers list user role as a possible answer so we'll stick to user role for this group.

Answer B is incorrect: admin role is too broad for data scientists. Viewer role doesn't allow developers to create jobs.

Answer C is incorrect: Same as B, viewer role doesn't allow developers to create jobs.

Answer D is incorrect: Same as B, admin role is too broad for data scientists.

Source(s):

BigQuery – Access Control: https://cloud.google.com/bigquery/docs/access-control

Domain

Prepare and use data for analysis

Question 9Skipped

Marketing team in your company wants to access certain tables in BigQuery. These tables are stored along with other tables considered sensitive and it's not approved to be accessed by marketing team. You need to restrict marketing team's role to only read from the tables they are allowed to.

Which of the following actions will you take?

Correct answer

A. Assign marketing team's roles as viewers on these specific tables. They won't be able to select from other tables in the dataset.

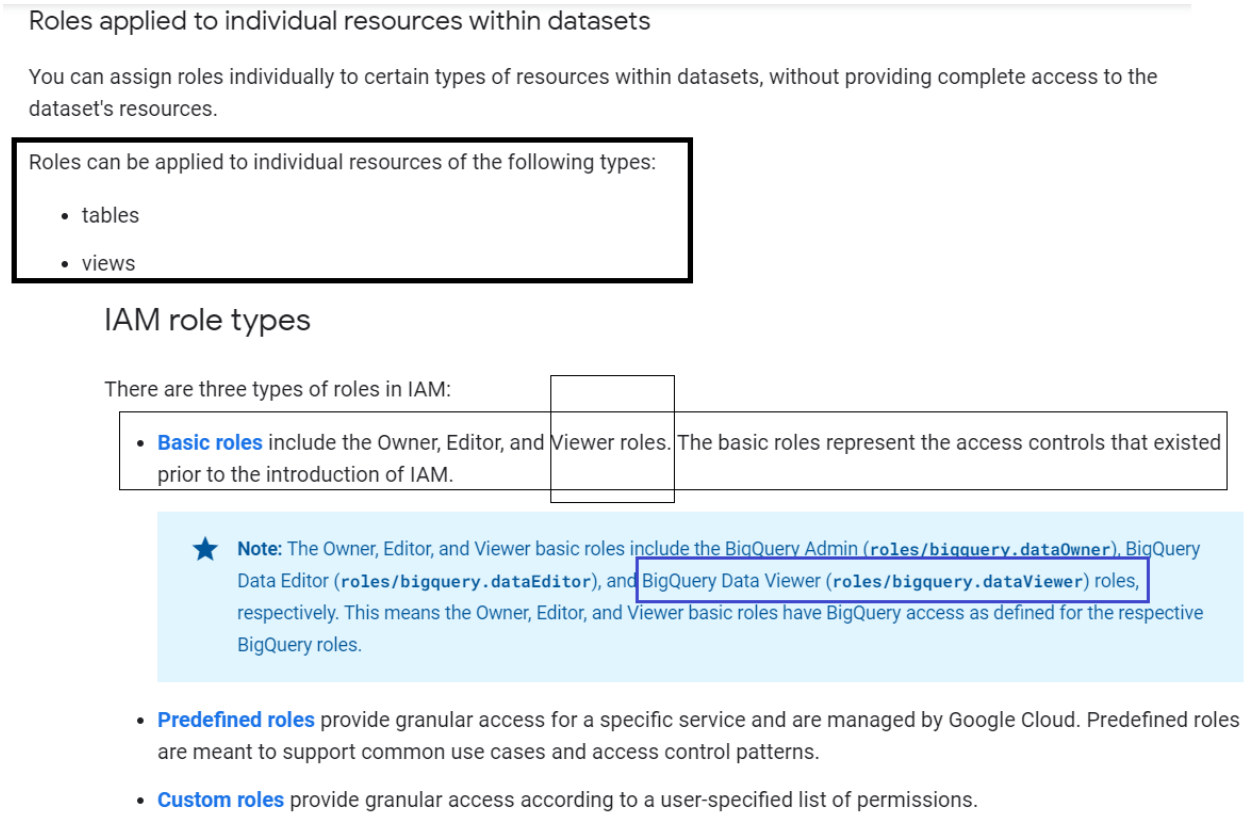
B. Create authorized views on tables marketing team wants to read from on the same dataset tables reside in. Grant viewer role to marketing team on the views.

C. Create a new dataset in BigQuery. Grant viewer role to marketing team on the new dataset. Copy the tables from the current dataset to the new one.

D. Create a new dataset in BigQuery. Create authorized views on tables marketing team wants to read from. Grant viewer role to marketing team on new dataset.

Overall explanation

Answer: A



Ref:

https://cloud.google.com/bigquery/docs/access-control#resource_level

Domain

Prepare and use data for analysis

Question 10Skipped

You have about 20TB of data which is not accessed and the data team decided to archive them in the cloud. The team is looking for a storage solution that is highly available with minimum costs. On the other hand, the data may be accessed a few times a year for reconciliation purposes. Reconciliation frequency is per month.

Which of the following choices best satisfy data team's requirements?

A. Google Storage Standard.

Correct answer

B. Google Storage Nearline.

C. Google Storage Coldline.

D. BigQuery.

Overall explanation

Correct Answer: B

Nearline Storage is a low-cost, highly durable storage service for storing infrequently accessed data. Nearline Storage is a better choice than Standard Storage in scenarios where slightly lower availability, a 30-day minimum storage duration, and costs for data access are acceptable trade-offs for lowered at-rest storage costs.

Nearline Storage is ideal for data you plan to read or modify on average once per month or less. For example, if you want to continuously add files to Cloud Storage and plan to access those files once a month for analysis, Nearline Storage is a great choice.

Coldline Storage is a very low-cost, highly durable storage service for storing infrequently accessed data. Coldline Storage is a better choice than Standard Storage or Nearline Storage in scenarios where slightly lower availability, a 90-day minimum storage duration, and higher costs for data access are acceptable trade-offs for lowered at-rest storage costs.

Coldline Storage is ideal for data you plan to read or modify at most once a quarter. Note, however, that for data being kept entirely for backup or archiving purposes, Archive Storage is more cost-effective, as it offers the lowest storage costs.

Option A is incorrect: Google Storage Standard is not a cheap option for storing archive data. There are other options to store archive data cheaper.

Option C is incorrect: The scenario mentions the need to access the data several times a year. Hence, Coldline is not a suitable option because Coldline is to plan archive data accessed at most once a year.

Option D is incorrect: BigQuery does not have a cheaper cost for storing data than Nearline for the first 90 days until it's moved to long-term storage.

Source(s):

Google Storage Classes: https://cloud.google.com/storage/docs/storage-classes

Domain

Design Data Processing Systems

Question 11Skipped

You have a complex data pipeline which has a combination of shell scripts, python code and spark jobs. These task are scheduled by cron jobs to run. The problem with this approach is, in case of failure, the whole pipeline breaks and failure control with stopping next tasks from running after a certain task fails and re-running the pipeline again is difficult and messy. You want a solution which can manage the pipeline's different jobs to be failure-resilient, scalable and easy to monitor.

What approach is best for this scenario?

A. Use Cloud Composer to orchestrate the pipeline workflow.

B. Use Dataproc for Apache Spark jobs and migrate all other tasks to use Apache Spark instead.

C. Use Cloud Scheduler to schedule pipeline's tasks.

Correct answer

D. Use Google Workflows to orchestrate the pipeline

Overall explanation

Answer: D

You can easily orchestrate the pipeline using Workflows without having any programming knowledge, it's a low-code solution and apparently cost-effective as compared to Cloud Composer.

Option A is incorrect: In order to migrate the existing cron jobs to Cloud Composer, you first have to be familiar with Apache Airflow, and then you have to code the workflow which will require development efforts and it will further delay the process. Also, it is a very costly solution.

Option B is incorrect: Dataproc does not orchestrate workflows and migrating all tasks can be a difficult job.

Option C is incorrect: Using Cloud Scheduler does not solve the core problem which is how to manage the workflow.

Source(s):

https://cloud.google.com/workflows/docs/overview

Domain

Design Data Processing Systems

Question 12Skipped

A dairy products company is using sensors installed around different areas in its farms to monitor employees activities and detect any intruders. Apache Kafka cluster is used to gather the events coming from sensors. Recently, Kafka cluster is becoming a bottleneck causing lag in receiving sensor events. Turns out sensors are sending more frequent events and due to the company expanding with more farms, more sensors are installed and this will cause extra load on the cluster. What is the most resilient approach to solve this issue?

Correct answer

A. Use pub/sub to ingest and stream sensor events.

B. Scale out Kafka cluster to withstand the continuously flowing event stream.

C. Spin up a new Kafka cluster and distribute sensors even streams between the two clusters.

D. Deploy Confluent's Managed Apache Kafka Cluster from the marketplace to scale the cluster according to workload

Overall explanation

Correct Answer: A

Description: Cloud Pub/Sub is a service to ingest event streams at any scale. It's scalable and reliable for stream analytics and event-driven computing systems. So it's the most reliable Google product for such a scenario.

Option B is incorrect as this is not a scalable solution and eventually, this issue will arise again. Pub/Sub is a fully managed service that is scalable without user action.

Option C is incorrect as this is not a scalable solution and eventually, this issue will arise again

Option D is incorrect as this solution includes Confluent's Managed Apache Kafka Cluster, which is a third-party service on the GCP. Moreover, it will cost more as compared to Pub/Sub. Also, the scaling will take some time, unlike Pub/Sub.

Reference:

Google Pub/Sub: https://cloud.google.com/pubsub/docs/overview

Domain

Prepare and use data for analysis

Question 13Skipped

You are building a streaming data pipeline for a VOD (Video-on-demand) service company. It receives event data from its player app sending details of what users are watching, video state (play, pause, loading), and other metrics that can be derived from the device used such as OS, brand, and screen resolution.

The event data collected should be analyzed by most recent data for quality check and further action in case of streaming issues. How can you ingest the stream data?

A. Use Cloud Pub/Sub to ingest the events and attach a unique ID to every event in the publisher.

B. Use Cloud Pub/Sub to ingest the events and attach timestamp to every event in the publisher.

Correct answer

C. Use Cloud Pub/Subb to ingest the events and store them in BigQuery using a Pub/Sub Bigquery subscription.

D. Launch a compute engine and install Apache Kafka to ingest the event stream.

Overall explanation

Answer: C.

Here the requirement is "The event data collected should be analyzed by most recent data for quality check and further action in case of streaming issues."
To accomplish this, we need to store the ingested data in BigTable/BigQuery like data storage and do analysis on the basis of ingested data and its time.
As we know, The Pub/Sub service adds the following fields to the message:

- A message ID unique to the topic
- A timestamp for when the Pub/Sub service receives the message

Option A: Use Cloud Pub/Sub to ingest the events and attach a unique ID to every event in the publisher.
MessageId will serve as Unique Id but where to store data to do analysis is not clear. Hence this is not the correct solution
B: Use Cloud Pub/Sub to ingest the events and attach a timestamp to every event in the publisher.

As we know, Pub/Sub adds timestamp is every message, storing the data to do analysis, is not mentioned. Hence it will not be going to serve our objectives.
C: Use Cloud Pub/Sub to ingest the events and store them to BigTable without any enrichment. Pub/Sub publisher automatically adds a timestamp to messages before publishing to subscribers.
Storing the ingested data without enrichment will help us to do multiple analysis requirements. Pub/Sub contains a message Id that is unique and the Timestamp of each message will help our analysis to the extent of an individual message.
Hence this is a possible answer.
D: Launch a compute engine and install Apache Kafka to ingest the event stream.
Apache Kafka is used for stream processing, website activity tracking, metrics collection and monitoring, log aggregation, real-time analytics, CEP, ingesting data into Spark, ingesting data into Hadoop, CQRS, replay messages, error recovery, and guaranteed distributed commit log for in-memory computing.
But it does not store data for future use. Additionally, we need to Compute engine which s IaaS is not good if have PaaS as BigTable Service.
Hence this is also not a good solution

Fields	
`data`	`string (bytes format)` The message data field. If this field is empty, the message must contain at least one attribute. A base64-encoded string.
`attributes`	`map (key: string, value: string)` Attributes for this message. If this field is empty, the message must contain non-empty data. This can be used to filter messages on the subscription. An object containing a list of `"key": value` pairs. Example: `{ "name": "wrench", "mass": "1.3kg", "count": "3" }`.
`messageId`	`string` ID of this message, assigned by the server when the message is published. Guaranteed to be unique within the topic. This value may be read by a subscriber that receives a `PubsubMessage` via a `subscriptions.pull` call or a push delivery. It must not be populated by the publisher in a `topics.publish` call.
`publishTime`	`string (Timestamp format)` The time at which the message was published, populated by the server when it receives the `topics.publish` call. It must not be populated by the publisher in a `topics.publish` call. A timestamp in RFC3339 UTC "Zulu" format, with nanosecond resolution and up to nine fractional digits. Examples: `"2014-10-02T15:01:23Z"` and `"2014-10-02T15:01:23.045123456Z"`.
`orderingKey`	`string` If non-empty, identifies related messages for which publish order should be respected. If a `Subscription` has `enableMessageOrdering` set to `true`, messages published with the same non-empty `orderingKey` value will be delivered to subscribers in the order in which they are received by the Pub/Sub system. All `PubsubMessages` published in a given `PublishRequest` must specify the same `orderingKey` value.

Ref. URL: https://cloud.google.com/pubsub/docs/reference/rest/v1/PubsubMessage

Domain

Design Data Processing Systems

Question 14Skipped

Data analysts are using Google Data Studio to build dashboards reading data from BigQuery as a data source. The CTO wants to minimize the costs of BigQuery queries run by dashboards. You suggested enabling predictive (pre-fetch) caching.

Which of the following will minimize the costs?

A. Restrict data fetch to be once every 24 hours and make sure Data Studio report has view credentials on the BigQuery dataset.

B. Enable pre-fetch caching for the report and make sure Data Studio report has view credentials on the BigQuery dataset.

Correct answer

C. Enable pre-fetch caching for the report and make sure Data Studio report is an owner on the BigQuery dataset.

D. Restrict data fetch to be once every 24 hours and make sure Data Studio report is an owner on the BigQuery dataset.

Overall explanation

Answer C.

The predictive (pre-fetch) cache analyzes the dimensions, metrics, and filter controls contained in the report, and predicts the possible queries. Data Studio then executes those queries in the background and stores the responses in the predictive cache. When a query can't be answered by the responsive cache, Data Studio tries to answer it using this predicted data. The predictive cache is limited in size, so it's possible your report can issue queries not already contained in the cache. If the query can't be answered by the predictive cache, Data Studio requests the data from the underlying data set.

Limitation: The predictive (pre-fetch) cache is only active for data sources that use **owner's credentials** to access the underlying data.

Answer A is incorrect: Data Studio caching maximum period is 12 hours.

Answer B is incorrect: As stated in the description, owner credentials should be granted on data sets.

Answer D is incorrect: Same as A, data fetch caching maximum period is 12 hours.

Source(s):

Data Studio – Manage Data
Freshness: https://support.google.com/datastudio/answer/7020039?hl=en

Domain

Design Data Processing Systems

Question 15Skipped

You have deployed a Tensorflow machine learning model using Cloud Machine Learning Engine. The model should be able to handle high volume of instances in a job to run complex models. The model should also write the output to Google Storage.

Which of the following approaches is recommended?

A. Use online prediction when using the model. Batch prediction supports asynchronous requests.

Correct answer

B. Use batch prediction when using the model. Batch prediction supports asynchronous requests.

C. Use batch prediction when using the model to return the results as soon as possible.

D. Use online prediction when using the model to return the results as soon as possible.

Overall explanation

Answer: B.

Online prediction	Batch prediction
Optimized to minimize the latency of serving predictions.	Optimized to handle a high volume of instances in a job and to run more complex models.
Can process one or more instances per request.	Can process one or more instances per request.
Predictions returned in the response message.	Predictions written to output files in a Cloud Storage location that you specify.
Input data passed directly as a JSON string.	Input data passed indirectly as one or more URIs of files in Cloud Storage locations.
Returns as soon as possible.	Asynchronous request.
Accounts with the following IAM roles can request online predictions: • Legacy Editor or Viewer • AI Platform Admin or Developer	Accounts with the following IAM roles can request batch predictions: • Legacy Editor • AI Platform Admin or Developer
Runs on the runtime version and in the region selected when you deploy the model.	Can run in any available region, using any available runtime version. Though you should run with the defaults for deployed model versions.
Runs models deployed to AI Platform.	Runs models deployed to AI Platform or models stored in accessible Google Cloud Storage locations.
Can serve predictions from a TensorFlow SavedModel or a custom prediction routine (beta).	Can serve predictions from a TensorFlow SavedModel.

AI Platform provides two ways to get predictions from trained models: *online prediction* (sometimes called HTTP prediction), and *batch prediction*. In both cases, you pass input data to a cloud-hosted machine-learning model and get inferences for each data instance. The differences are shown in the following table:

Batch prediction can handle high volume of instances in a job to run complex models. It also writes the output to Google Storage by specified location.

Answer A & D are incorrect: Online prediction doesn't support handling high volume of instances per job and doesn't write output to Google Storage.

Answer C is incorrect: Batch prediction doesn't return the output as soon as possible, it supports asynchronous requests.

Source(s):

Online vs. Batch Prediction: https://cloud.google.com/ml-engine/docs/tensorflow/online-vs-batch- prediction

Domain

Design Data Processing Systems

Question 16Skipped

A weather station facility which receives events from sensors installed on different pods distributed around the region return the current weather temperature based on sensor's location.

You are asked to build a pipeline to aggregate the incoming events to get the average temperature every 60 seconds for each region.

A. Tumbling window with a duration of 60 seconds.

Correct answer

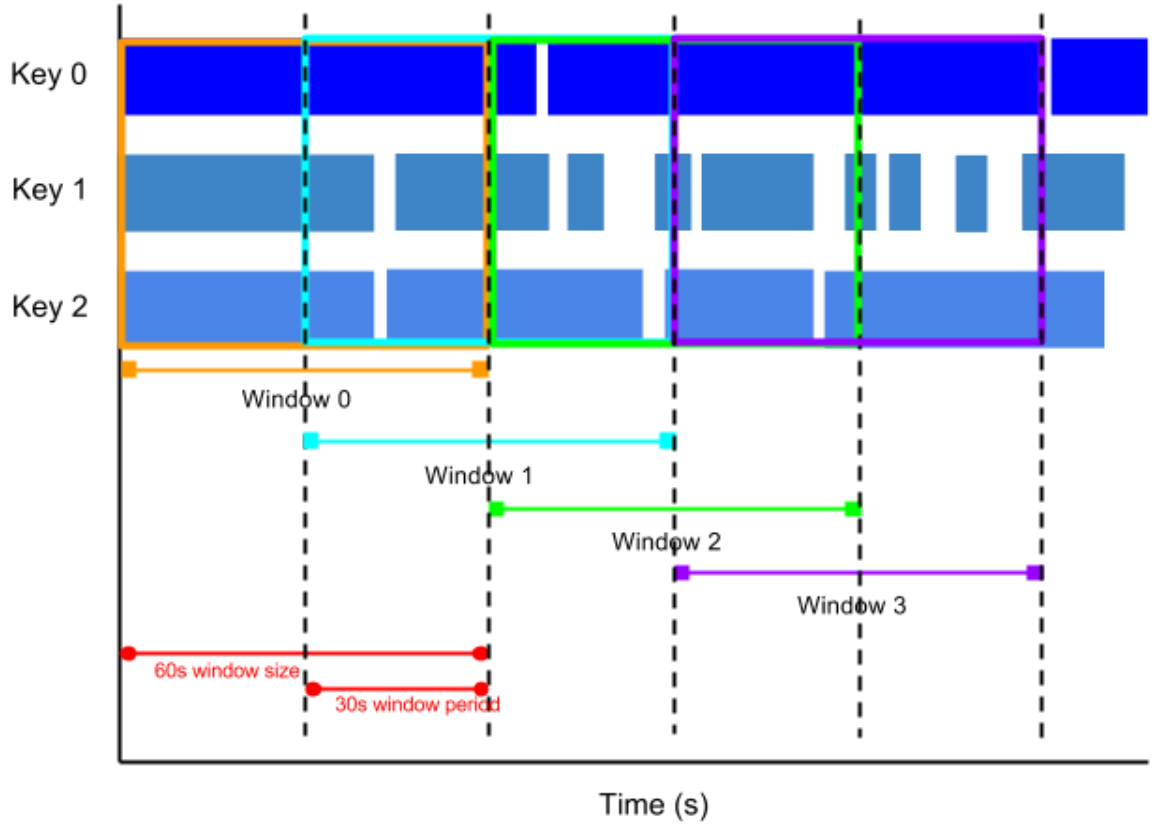
B. Hopping window with a duration of 60 seconds.

C. Session window with a time gap duration of 60 seconds.

D. Global window with the time-based trigger of 60 seconds.

Overall explanation

Answer: B.



A sliding time window uses time intervals in the data stream to define bundles of data. However, with sliding time windowing, the windows overlap. Each window might capture five minutes worth of data, but a new window starts every ten seconds. The frequency with which sliding windows begin is called the period. Therefore, our example would have a window size of five minutes and a period of ten seconds.

In this scenario, we want to get the average temperature per region by aggregating all events coming from sensors of that region. Sensor ID can be the key row. Sliding-time window aggregates event data in time window of 5 minutes, with sliding window of 60 seconds.

Source(s):

Windowing Functions: https://cloud.google.com/dataflow/model/windowing#windowing-functions

---

Note:

- 5 minutes is just an assumption to explain windowing through an example, where, for a time interval of compute a running average of the past five minutes' worth of data is updated every ten seconds.
- Sliding window function produce an output only when an event occurs. Every window will have at least one event and the window continuously moves forward. In fixed window, it doesn't have overlap and an event cannot belong to more than one fixed window. Each segment represents a distinct time segment.

1. The source of data is in different region and their data ingestion would not be same.
2. In sliding time Windows, the Window overlaps and window may be more that 1 Window..
3. In fixed Time Window, is good where streaming of data has limited source,

I think if you understand the windowing concept well and each of the 4 windowing functions, then sliding time window will appear as straight forward choice.

Domain

Design Data Processing Systems

Question 17Skipped

A gaming app allows up to 32 players to compete in battle royale mode in a single gaming session. Recently, players are sending feedback complaining some users are idle and not competing in the session which breaks the experience for them. The development team decided to end the session for players who are idle for more than 60 seconds to solve this problem.

Gaming app sends events every second contain player's state (active, idle, pending) and other details. You want to build a Dataflow pipeline which aggregates these events so idle players can be detected in the time frame specified by development team.

Which windowing function you should choose to design the pipeline?

A. Tumbling window with a duration of 60 seconds.

B. Hopping window with a duration of 60 seconds.

Correct answer

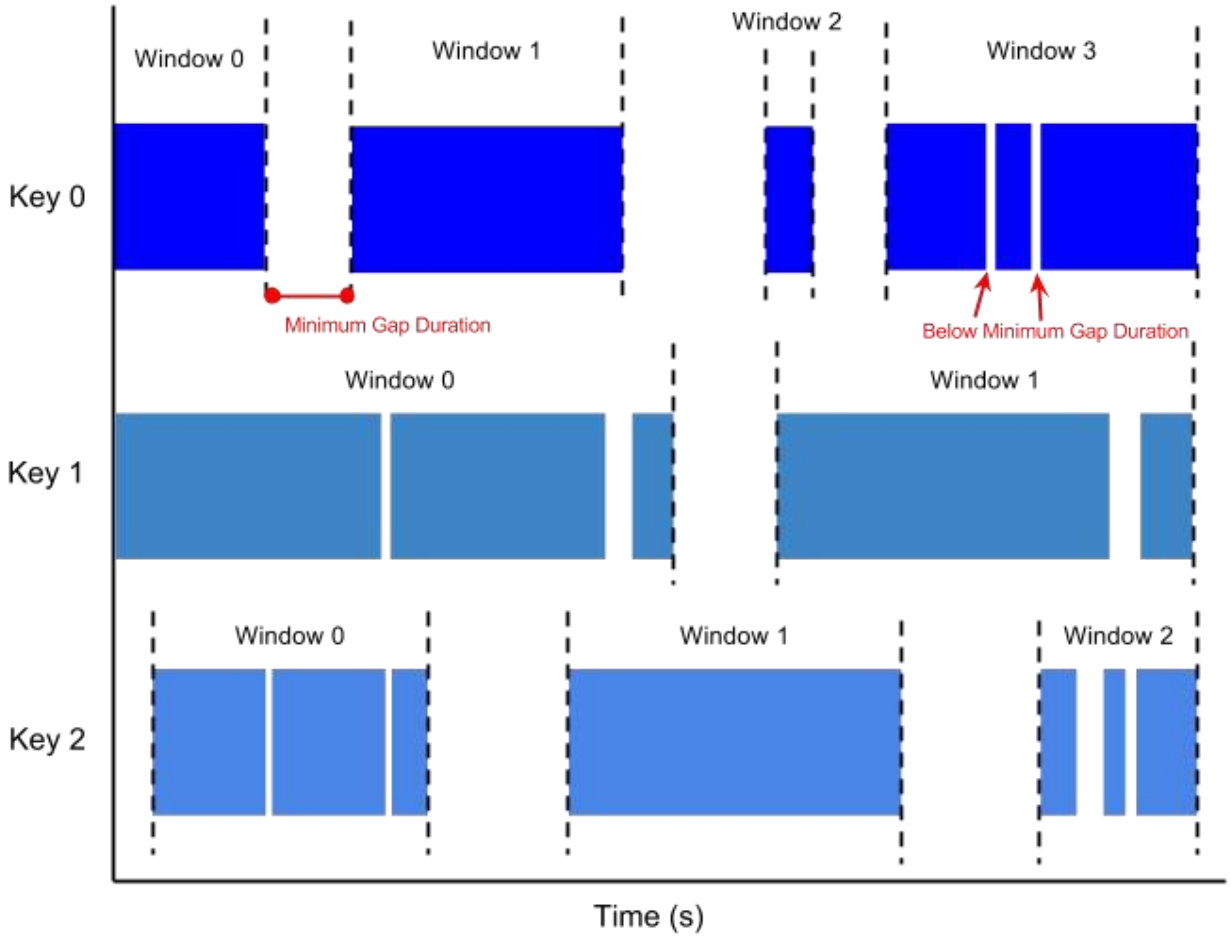
C. Session window with a time gap duration of 60 seconds.

D. Global window with the time-based trigger of 60 seconds.

Overall explanation

Answer: C.

A session window function defines windows around areas of concentration in the data. Session windowing is useful for data that is irregularly distributed with respect to time; for example, a data stream representing user mouse activity may have long periods of idle time interspersed with high concentrations of clicks. Session windowing groups the high concentrations of data into separate windows and filters out the idle sections of the data stream. Note that session windowing applies **on a per-key basis:** That is, grouping into sessions **only** takes into account data that has the same key. Each key in your data collection will therefore be grouped into disjoint windows of differing sizes.



For this scenario, per-session window is the function to choose to build Dataflow pipeline.

Source(s):

Windowing Functions: https://cloud.google.com/dataflow/model/windowing#windowing-functions

Domain

Design Data Processing Systems

Question 18Skipped

Your company uses Google Cloud as its main platform. The lead architect decided to migrate the architecture's relational databases to Cloud Spanner for horizontal scaling and multi-regional availability. Upon using Cloud Spanner for a while after migration and monitoring its performance, it was reported Cloud Spanner instance's performance is not as expected as in the planning phase. What could be a possible reason for this? (Choose 2).

A. UUID is used as primary keys for the tables.

Correct selection

B. Primary keys are monotonically increased.

C. Primary keys are randomly generated 16-byte alphanumericals.

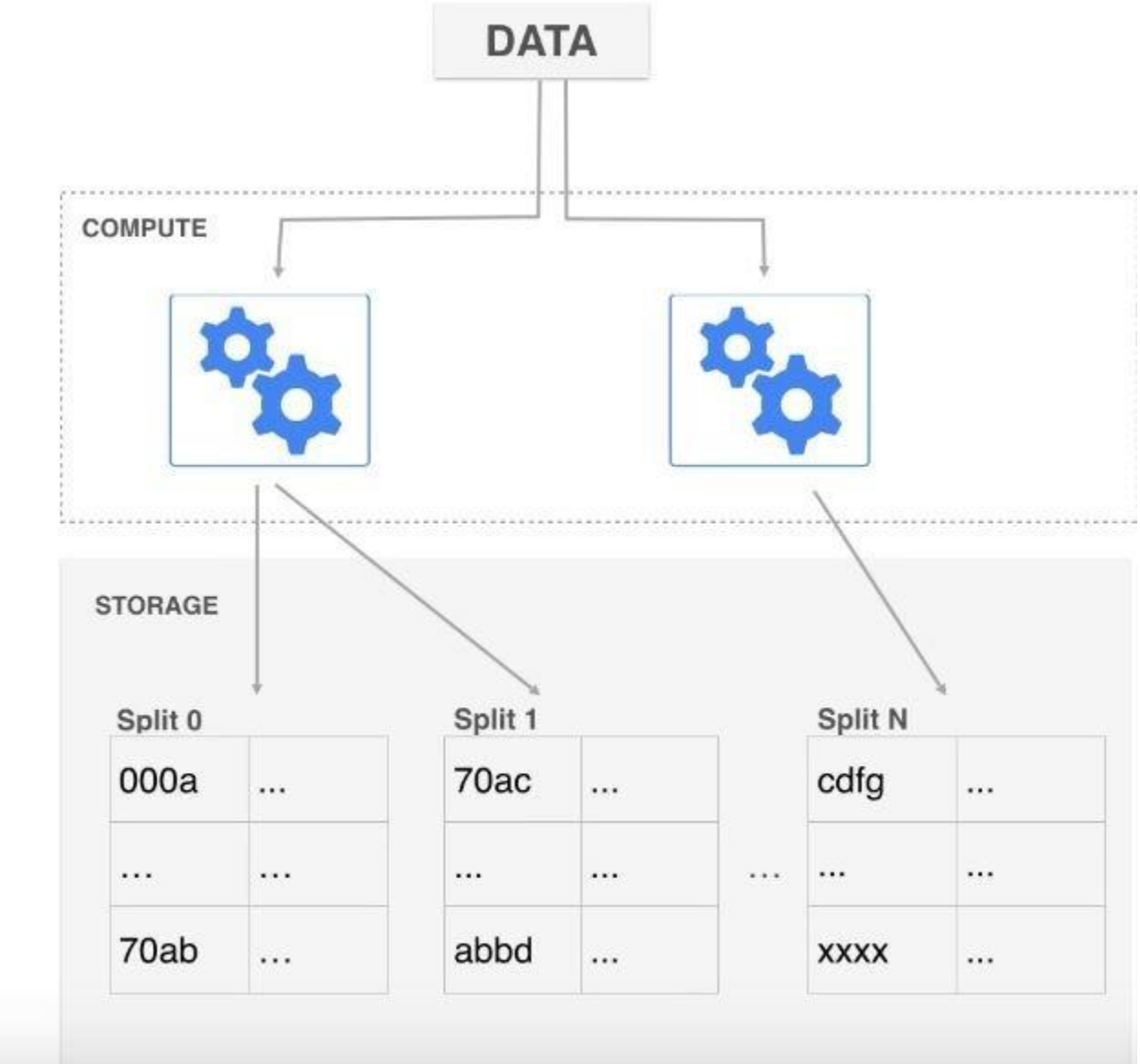
Correct selection

D. Primary key used is a combination of timestamp and original primary key as #timestamp-pk.

Overall explanation

Answer: B & D.

Description:



Cloud Spanner uses compute nodes to read and write data. Tables data is stored lexicographically by primary key. Data is distributed among multiple storage “splits.”

This is the reason why choosing the right primary key for Cloud Spanner is important for performance. If primary key is monotonic, this leads to storing table data to one storage split, which in return leads to compute nodes to hit the same storage split for reading & writing. A good primary key is a key which helps evenly distributing data among different storage splits.

Source(s):

Choosing the Right Primary Keys (TIL about Cloud Spanner):

https://www.youtube.com/watch? v=FFTHQt_KFNM

Domain

Store the data

Question 19Skipped

You have a Dataflow pipeline which streams data to be stored to BigTable after it has been transformed and enriched. Development team needs to modify the transformation code based on client's needs. The pipeline is in production which keeps streaming and any interruption to the pipeline may lead to data loss or unexpected output.

How can you make sure the pipeline can be stopped without any consequences?

A. Turn off Dataflow pipeline with 'cancel' option.

B. Create a new Dataflow pipeline with the new transformation code, then switch data stream to the new pipeline.

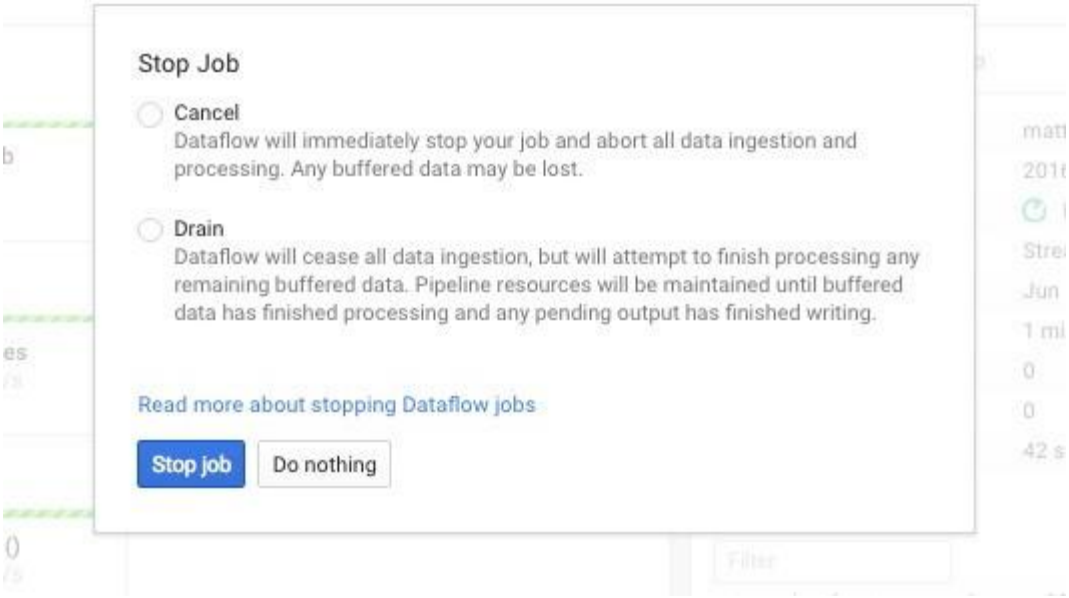
C. Transfer Dataflow pipeline to write data to Google Storage. Perform the needed changes then transfer pipeline back to write to BigTable and re-process the data written in Google Storage.

Correct answer

D. Turn off Dataflow pipeline with 'drain' option.

Overall explanation

Answer: D.



Using the *Drain* option to stop your job tells the Cloud Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources. When all pending processing and write operations are complete, the Cloud Dataflow service will clean up the GCP resources associated with your job.

Because *Cancel* option immediately halts processing, you may lose any "in-flight" data. "In-flight" data refers to data that has been read but is still being processed by your pipeline.

Answer A is incorrect: Using 'Cancel' option will lead to losing in-flight data.

Answer B & C are incorrect: Those options are not necessary.

Source(s):

Dataflow – Stopping a running pipeline: https://cloud.google.com/dataflow/docs/guides/stopping-a- pipeline

Domain

Design Data Processing Systems

Question 20Skipped

You are using Dataflow SDK to analyze data related to customer segmentation. You need to extract certain fields from the data files to be processed for further transformation.

Which operation is used to perform the operation required?

Correct answer

A. ParDo

B. PCollection

C. Transform

D. Pipeline

Overall explanation

Answer: A.

ParDo is the core parallel processing operation in the Apache Beam SDKs, invoking a user-specified function on each of the elements of the input PCollection. ParDo collects the zero or more output elements into an output PCollection. The ParDo transform processes elements independently and possibly in parallel.

In Google Dataflow SDK, ParDo allows for parallel programming. It acts on one item at a time (like a map in MapReduce). ParDo is useful for:

- Filtering and emitting input.
- Type conversion.
- Extracting parts of input and calculating values from different parts of inputs.

Source(s):

Dataflow - Programming Model for Apache Beam: https://cloud.google.com/dataflow/docs/concepts/beam-programming-model

Domain

Design Data Processing Systems

Question 21Skipped

An e-wallet company is designing a relational database solution for their e-payment transactions. Database will face high read/write transactions and accessed from different parts of Europe and may be expanded to other continents in the future. The database should be scalable and able to scale out to meet high demands.

What is the best approach for this scenario?

A. Use Cloud BigTable as a database. For scaling out, monitor CPU utilization and increase nodes when more than 75% of CPU is utilized in a 15-minute timespan.

B. Use Cloud SQL as a database. For scaling out, monitor disk utilization and increase nodes when more than 85% of storage is utilized in a 15-minute timespan.

C. Use Cloud BigTable as a database. For scaling out, monitor memory utilization and increase nodes when more than 80% of memory is utilized in a 15-minute timespan.

Correct answer

D. Use Cloud Spanner as a database. For scaling out, monitor CPU utilization and increase nodes when more than 65% of CPU is utilized in a 15-minute timespan.

Overall explanation

Answer D.

Cloud Spanner is a horizontally scalable, strongly consistent, relational database service. it's built to combine the benefits of relational database structure with non-relational horizontal scale. This delivers high-performance transactions and strong consistency across rows, regions and continents.

While Cloud Spanner does not automatically scale the number of nodes, You may set up notifications for Cloud Spanner using Stackdriver monitoring for CPU utilization alerts based on the threshold set. In this scenario, we need to scale out (increase nodes) when we are notified of high usage.

Answer A & C are incorrect: BigTable is not a relational database. It doesn't meet the scenario's requirements.

Answer B is incorrect: Cloud SQL cannot scale out on the fly and it does not support multi-regional/continental coverage.

Source(s):

Cloud Spanner:

https://cloud.google.com/spanner/

Cloud Spanner scale out: https://cloud.google.com/spanner/docs/instances

Stackdriver monitoring: https://cloud.google.com/spanner/docs/monitoring

Domain

Design Data Processing Systems

Question 22Skipped

An e-payment service generates thousands of gigabytes of logs every month which are streamed to a Dataflow pipeline, transformed, and stored in a data warehouse for further analysis. These raw logs are not accessed once the transformation is done by Dataflow. CTO suggested that log files should be archived after a month of the transformation and after making sure that the data is not required for debugging. Which of the following storage types are recommended?

A. Google Storage Standard

B. Google Storage Nearline

C. Google Storage Coldline

Correct answer

D. Google Storage Archive

Overall explanation

The correct answer is Option D.

Archive storage is the lowest-cost, highly durable storage service for data archiving, online backup, and disaster recovery. Unlike the "coldest" storage services offered by other Cloud providers, your data is available within milliseconds, not hours or days.

Like Nearline storage and Coldline storage, Archive storage has a slightly lower availability than Standard storage. Archive storage also has higher costs for data access and operations, as well as a 365-day minimum storage duration. Archive storage is the best choice for data that you plan to access less than once a year.

Option A is incorrect: Google Storage Standard is not a cheap option for storing archive data. There are other options to store archive data cheaper.

Option B is incorrect: The scenario states the raw logs are to be archived after processed and kept for a month to ensure data is not required anymore. Data archived will most probably won't be
accessed so Nearline storage is not necessary.

Answer C is incorrect: Google Storage Coldline is not apt for archival storage.

Source(s): Google Storage Classes:

https://cloud.google.com/storage/docs/storage-classes

Domain

Design Data Processing Systems

Question 23Skipped

As a solution for a serverless data warehouse, you decided to use BigQuery to store and query data. You built a Dataflow pipeline to read data from Google Storage and import it to BigQuery. You added a few users to access BigQuery for reporting purposes. You want to monitor the activity on BigQuery by getting details about query count and execution time. You want such metrics to appear on a dashboard to be shared later with other stakeholders. What should you do?

A. Build a script to use gcloud command to extract queries execution time and data size scanned every 1 hour. Send the stats to Operation Suite and create a dashboard showing the metrics.

Correct answer

B. Use Cloud Monitoring to create a dashboard and graphs showing query metrics.

C. You need to contact Google Cloud support in order to enable metrics on BigQuery UI.

D. From BigQuery UI, you can view run queries and execution time. You can share it by exporting the stats to a file.

Overall explanation

Answer: B

Operation Suite is a tool from Google to monitor and manage services, containers, applications, and infrastructure. Operation Suite aggregates metrics, logs, and events from infrastructure, giving developers and operators a rich set of observable signals that speed root-cause analysis and reduce mean time to resolution (MTTR). Operation Suite doesn't require extensive integration and it does not lock developers into using a particular cloud provider.

One of the resources Operation Suite supports monitoring is BigQuery. Operation Suite provides a wide set of metrics to create charts and dashboards for better monitoring of BigQuery such as query
execution time, storage and slots allocated for run queries.

Option A is incorrect: There is no need to send BigQuery metrics to Operation Suite BigQuery can automatically send metrics to Operation Suite after enabling API.

Option C is incorrect: You do not need to contact Google Cloud support to enable metrics sent to Operation Suite. You can enable Operation Suite API if you have the required role(s).

Option D is incorrect: Smart predictor does not show you the approximate execution time for the query.

Source(s):
Operation Suite: https://cloud.google.com/products/operations
BigQuery Monitoring Using Operation Suite**:**
https://cloud.google.com/bigquery/docs/monitoring#slots-available

Domain

Prepare and use data for analysis

Question 24Skipped

A company decides to migrate its on-premise data infrastructure to the cloud mainly for high availability of cloud services and to lower the high costs of storing data on-premise. The infrastructure uses HDFS to store data and be processed and transformed using Apache Hive & Spark. The company wants to migrate the infrastructure and DevOps team still wants to administrate the infrastructure in the cloud. As a data architect, which of the following is the approach recommended by Google?

Correct answer

A. Use ephemeral Dataproc cluster with preemptible VMs to process the data and Store data in Google Cloud Storage with an object lifecycle management policy.

B. Build a Dataflow pipeline. Store the data in Google Storage. Use Cloud Compute to launch instances and install the required dependencies for processing the data.

C. Use Dataproc to process the data. Store data in Dataproc's HDFS.

D. Build a Dataflow pipeline. Store the data in persistent disks in HDFS. Execute the code in Spark framework provided by Dataflow.

Overall explanation

Correct Answer: A

The main objective of the use case is to save the cost, It's the Google recommended best practice to use an ephemeral dataproc cluster i.e. spin the cluster when required and destroy it when the work is done with Preemptible VMs.

Also as per Google's recommended best practice, we should use the object lifecycle management policy on the GCS buckets.

- Ephemeral Dataproc cluster: This is a type of Dataproc cluster that is designed to be short-lived. It is a good option for workloads that need to be processed quickly and then shut down.
- Preemptible VMs: These are VMs that are available at a discounted price because they can be reclaimed by Google at any time. They are a good option for workloads that can be interrupted without losing data.
- Object lifecycle management: This is a feature of Google Cloud Storage that allows you to automatically manage the lifecycle of your data. For example, you can set up a policy to automatically delete old data or to move data to a different storage class.

Option B is incorrect: Dataflow is serverless which may not suit DevOps' requirement to fully manage the pipeline and it's unnecessary to use Cloud Compute for installing dependencies.

Option C is incorrect: Dataproc's HDFS is volatile, which means it will be removed when the cluster is deleted. Dataproc clusters can be kept up indefinitely but this may lead to high costs which defeats the purpose of migration.

Option D is incorrect: In addition to what was discussed in answer B, storing data using persistent disks can be only accessible by Compute engines and it's more expensive than storing in Google Storage.

Answer A fulfills the requirements for migrating the on-premise infra to the cloud with high availability, minimum costs, and full control by DevOps.

References:

https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs

https://cloud.google.com/storage/docs/lifecycle

Cloud Dataproc: https://cloud.google.com/dataproc/

Domain

Prepare and use data for analysis

Question 25Skipped

Your team was working on a development BigTable instance for some time experimenting on it to stream event data coming from hundreds of sensors sending events frequently. The team lead considered instead of deleting the instance and losing all events collected since building the pipeline, it would be a better idea to use the instance in production with the required changes to ensure high availability and best performance.

Which of the following approaches is best to satisfy the team lead's requirements, given that you are currently using HDD for development purposes?

Correct answer

A. Export the data from BigTable development instance to Google Storage, launch a new BigTable production instance with SSD storage type, then load the data from Google Storage to the new BigTable instance.

B. Export the data from BigTable development instance to Google Storage, launch a new BigTable production instance with HDD storage type, then load the data from Google Storage to the new BigTable instance.

C. Change BigTable instance type from development to production, scale up number of nodes and ensure the storage type is HDD.

D. Change BigTable instance type from development to production, scale up number of nodes and ensure the storage type is SSD.

Overall explanation

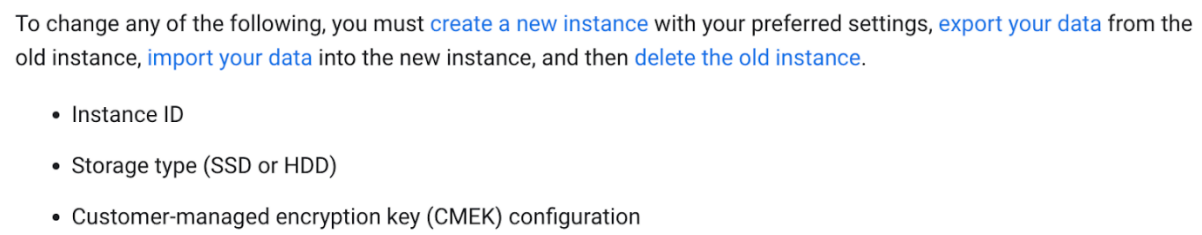
Answer: A.

If you no longer want to use a development instance for development and testing, you can upgrade it to a production instance at any time. Upgrading a development instance is permanent.

A cluster must have enough nodes to support its current workload and the amount of data it stores. Otherwise, the cluster might not be able to handle incoming requests, and latency could go up.

SSD is significantly faster and has a more predictable performance than HDD. In a Cloud Bigtable cluster, SSD storage delivers 6 ms latencies for both reads and writes for 99% of all

requests. By contrast, HDD storage delivers 200 ms read latencies and 50 ms write latencies on the same benchmark.



To change any of the following, you must create a new instance with your preferred settings, export your data from the old instance, import your data into the new instance, and then delete the old instance.

- Instance ID
- Storage type (SSD or HDD)
- Customer-managed encryption key (CMEK) configuration

Option B & C is incorrect: HDD is not recommended for optimal performance. SSD should be used.

Option D is incorrect: Since the storage type of a given Bigtable instance is HDD you can't change the storage disk type directly

---

Notes: Here Option A and Option D are possible solutions. Now we need to select one between these 2.

As we know for BigTable we cannot upgrade HDD to SSD and the question is silent about drive type whether SSD or HDD.

Hence our first approach would be to check if Drive is SSD or HDD. if yes then Option D is a correct answer and if not then Option A will be a possible answer.

Source(s):

BigTable Instances, Clusters & Nodes: https://cloud.google.com/bigtable/docs/instances-clusters-nodes BigTable – SSD vs. HDD: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

Domain

Design Data Processing Systems

Question 26Skipped

You have a system which writes its logs to a file. The system writes to a new log file every hour. You want to merge log files every 24 hours to a single file and upload it to a specific bucket in Google Storage. You created a Dataflow pipeline to process the logs files and you want to run this pipeline everyday at 3am.

Which of the following is the best approach to schedule this task?

Correct answer

A. Use Cloud Scheduler to create a cron job to run the Dataflow pipeline at 3am.

B. Create a Compute Engine VM and create a cron job to run Dataflow pipeline at 3am.

C. Configure Dataflow pipeline as a streaming job to process the data in real time.

D. Use Cloud Functions to run Dataflow pipeline at 3am.

Overall explanation

Answer: A.

Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It allows you to schedule virtually any job, including batch, big data jobs, cloud infrastructure operations, and more. You can automate everything, including retries in case of failure to reduce manual toil and intervention. Cloud Scheduler even acts as a single pane of glass, allowing you to manage all your automation tasks from one place.

Answer B is incorrect: No need to create a VM when you can use Cloud Scheduler.

Answer C is incorrect: This process is a batch job. It is not required to make it a streaming pipeline.

Answer D is incorrect: You cannot schedule a Cloud Functions unless you use Cloud Scheduler.

Source(s):

Cloud Scheduler: https://cloud.google.com/scheduler/

Domain

Design Data Processing Systems

Question 27Skipped

A company uses Airflow to orchestrate its data pipelines and DAGs (Directed Acyclic Graphs), installed and maintained on-premise by DevOps team. The company wants to

migrate the data pipelines managed in Airflow to Google Cloud. The company is looking for a migration method which can make DAGs available and migrated without extra code modifications so the data pipelines can be available once migrated. Which service should you use?

A. App Engine

B. Cloud Function

C. Dataflow

Correct answer

D. Cloud Composer

Overall explanation

Answer: D.

Description:

Cloud Composer is a fully managed workflow orchestration service built on Apache Airflow. Cloud composer is built specifically to schedule and monitor workflows and take required actions. You can use Cloud Composer to orchestrate dataflow pipeline and create a custom sensor to detect file's condition if any changes occurred, then it triggers the dataflow pipeline to run again.

Source(s):

Cloud Composer:

https://cloud.google.com/composer/

Domain

Design Data Processing Systems

Question 28Skipped

You use BigQuery as the main data warehouse. As you are checking the bill of Google Cloud usage for the past month, you noticed a hike in BigQuery costs. Upon more inspection, you found out it's due to ad-hoc queries scanning large amounts of data. You want to limit the size of data queries process to control costs and avoid excessive bills. What should you do?

A. Set monthly flat-rate pricing for BigQuery.

B. Change to using Google Storage as external table instead of using BigQuery's storage option.

C. Launch a Cloud SQL instance. Replicate BigQuery datasets to Cloud SQL. Allow users with ad-hoc queries to use Cloud SQL instead.

Correct answer

D. Set custom quotas on User-level with access on BigQuery based on their business requirements.

Overall explanation

If you have multiple BigQuery projects and users, you can manage costs by requesting a custom quota that specifies a limit on the amount of query data processed per day. Creating a custom quota on query data allows you to control costs at the project level or at the user level.

- Project-level custom quotas limit the aggregate usage of all users in that project.
- User-level custom quotas are separately applied to each user or service account within a project.

In this scenario, the aim is to control data volume quotas. So, **option D is the best approach**.

- **Option A is incorrect:** Flat-rate can be a possible approach. However, BigQuery does not provide flexible flat-rate pricing and the cheapest is $10,000 for 500 slots, which may not be a desirable option for small-medium businesses.
- **Option B is incorrect**: You are charged for the number of bytes processed whether the data is stored in BigQuery or in an external data source such as Cloud Storage, Google Drive, or Cloud Bigtable. Moving data to Google Storage will not save costs when querying data.
- **Option C is incorrect:** Replicating data from BigQuery to Cloud SQL is not straightforward. You may need to export data to Cloud Storage, then import data to Cloud SQL. This solution can be more costly in the long run depending on data size since you pay for Google Storage costs as well as choosing the Cloud SQL instance type.

Source(s):
BigQuery - Creating custom cost controls:

- https://cloud.google.com/bigquery/docs/custom-quotas#controlling_query_costs_using_bigquery_custom_quotas

BigQuery Pricing - Monthly Flat Rate:

- https://cloud.google.com/bigquery/pricing#monthly-flat-rate

Domain

Store the data

Question 29Skipped

You are using BigQuery as the data warehouse. Different departments are using BigQuery to read data. Upon checking the billing costs, you notice that there is a spike in running queries on BigQuery despite caching is enabled. You started scanning through the queries run on BigQuery trying to find out if some queries are not cached.

Which of the following can be reasons for queries not cached? (Choose 2)

A. SELECT queries without asterisk (*).

B. Queries select from authorized views on archive tables.

Correct selection

C. Queries multiple tables use wildcard.

Correct selection

D. Jobs use destination tables.

Overall explanation

Answers: C & D

Currently, cached results are not supported for queries against multiple tables using a wildcard even if the "Use Cached Result**s"** option is checked. If you run the same wildcard query multiple times, you are billed for each query.

Query results are not cached when a destination table is specified in the job configuration, the GCP Console, the classic web UI, the command line, or the API.

Source(s):

BigQuery - Wildcards: https://cloud.google.com/bigquery/docs/querying-wildcard-tables BigQuery – Cached Results: https://cloud.google.com/bigquery/docs/cached-results

Domain

Store the data

Question 30Skipped

You have a web service which allows users to create personal profiles and upload their photos online. The service will create thumbnails of various sizes and resolutions to be used on the service's different pages and widgets. The service uses Google Storage to store the original photos uploaded and thumbnails generated. The service usually uses the thumbnails for a certain period before generating new ones depends on user's activity and the age of the original photo the thumbnail is created from.

You want to remove all thumbnails created before 90 days. What should you do?

A. Use Cloud Scheduler to scan objects in thumbnails folder within the Google Storage bucket and delete all objects created before 90 days.

B. Use Cloud Function to scan objects in thumbnails folder within the Google Storage bucket and delete all objects created before 90 days.

Correct answer

C. Use object lifecycle management in Google Storage to delete thumbnail objects created before 90 days.

D. Use object lifecycle management in Google Storage to delete thumbnail objects older than the current thumbnail object version last created. Make sure the bucket used for storage has object versioning enabled.

Overall explanation

Answer: C.

Description:

To support common use cases like setting a Time to Live (TTL) for objects, archiving older versions of objects, or "downgrading" storage classes of objects to help manage costs, Cloud Storage offers the Object Lifecycle Management feature. This page describes the feature as well as the options available when using it.

You can assign a lifecycle management configuration to a bucket. The configuration contains a set of rules which apply to current and future objects in the bucket. When an object meets the criteria of one of the rules, Cloud Storage automatically performs a specified action on the object. Here are some example use cases:

- Downgrade the storage class of objects older than 365 days to Coldline Storage.
- Delete objects created before January 1, 2013.
- Keep only the 3 most recent versions of each object in a bucket with versioning enabled.

Answers A & B are incorrect: Cloud Scheduler & Cloud Function are not required since Google Storage already supports lifecycle management to satisfy the requirement.

Answer D is incorrect: The scenario mentions the thumbnails are not needed anymore after 90 days, so there is no need to enable object versioning, which leads to more costs due to keep versions of objects.

Source(s):

https://cloud.google.com/storage/docs/lifecycle https://cloud.google.com/storage/docs/object-versioning

Domain

Store the data

Question 31Skipped

An e-commerce company uses BigQuery as its main data warehouse. One of the tables stores customers details such as name, address, email and phone number. Data team wants to modify the table's schema and add a new column called 'zipcode' which is previously included in address column. You are asked to modify the table's schema and do necessary changes. You need to perform the changes with minimal costs. What should you do?

Correct answer

A. Add a new column called 'zipcode' to customers table. Run an update statement to extract zip code from address column and set it to the new column.

B. Create a materialized view in Bigquery that extracts the zip code from the address as a new column

C. Export table data from BigQuery to Google Storage. Use Dataproc to transform data and extract the zip code from addresses and append it as a new column. Create a new table for customers with new column 'zipcode.' Import transformed data to new table.

D. Create a Dataflow pipeline to read data from BigQuery, extract zip code from address column, then write data to a newly created table in BigQuery with 'zipcode' column.

Overall explanation

Answer: A.

BigQuery allows partial modification on an existing table's schema definition. The following actions are allowed:

- Adding columns to a schema definition.
- Relaxing a column's mode from REQUIRED to NULLABLE.

Answer B is incorrect: BigQuery's views are logical views, not materialized views.

Answer C & D are incorrect: Using Dataproc or Dataflow is not a cheap or simple solution comparing to updating the table directly from BigQuery.

Source(s):

BigQuery – Modifying Table Schemas: https://cloud.google.com/bigquery/docs/managing-table- schemas

BigQuery – Introduction to Views: https://cloud.google.com/bigquery/docs/views-intro

Domain

Store the data

Question 32Skipped

You use BigQuery as your main data warehouse. By time, your tables start to get bigger and selecting from these tables result in scanning many rows which increases the cost of queries running on them. You want to find a way to reduce the costs of queries scanning through your big tables. What should you do? (Choose 2)

A. Use LIMIT when running SELECT statements on the tables.

Correct selection

B. Use partitioning to split data into partitions by columns most used for filtering data.

C. Set BigQuery to limit scanning data to certain size.

Correct selection

D. Use sharding to split data into several tables.

Overall explanation

Answers: B & D.

A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

There are two types of table partitioning in BigQuery:

- Tables partitioned by ingestion time: Tables partitioned based on the data's ingestion (load) date or arrival date.
- Partitioned tables: Tables that are partitioned based on a TIMESTAMP or DATE column.

As an alternative to partitioned tables, you can shard tables using a time-based naming approach such as [PREFIX]_YYYYMMDD. This is referred to as creating *date-sharded* tables.

Source(s):

BigQuery – Introduction to Partitioned Tables: https://cloud.google.com/bigquery/docs/partitioned- tables#partitioning_versus_sharding

Domain

Store the data

Question 33Skipped

An air-quality research facility monitors the quality of the air and alerts of possible high air pollution in a region. The facility receives event data from 25,000 sensors every 60 seconds. Event data is then used for time-series analysis per region. Cloud experts suggested using BigTable for storing event data.

What will you design the row key for each even in BigTable?

A. Use event's timestamp as row key.

Correct answer

B. Use combination of sensor ID with timestamp as sensorID-timestamp.

C. Use combination of sensor ID with timestamp as timestamp-sensorID.

D. Use sensor ID as row key.

Overall explanation

Answer B.

Storing time-series data in Cloud Bigtable is a natural fit. Cloud Bigtable stores data as unstructured columns in rows; each row has a row key, and row keys are sorted lexicographically.

For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).

When Cloud Bigtable stores rows, it sorts them by row key in lexicographic order. There is effectively a single index per table, which is the row key. Queries that access a single row, or a contiguous range of rows, execute quickly and efficiently. All other queries result in a full table scan, which will be far, far slower. A full table scan is exactly what it sounds like—every row of your table is examined in turn.

For Cloud Bigtable, where you could be storing many petabytes of data in a single table, the performance of a full table scan will only get worse as your system grows.

Choosing a row key that facilitates common queries is of paramount importance to the overall performance of the system. Enumerate your queries, put them in order of importance, and then design row keys that work for those queries.

From the description, you need to combine both sensor ID and timestamp in order to fetch data you want fast. So, answers A & D are incorrect.

If you start the row key with timestamp, most recent data will be inserted at the bottom of the table since rows are sorted in lexicographic order. Starting the row key with sensor ID will allow writing all sensor's events together and allow distributing data among nodes.

Source(s):

BigTable – Schema Design for Time Series
Data: https://cloud.google.com/bigtable/docs/schema- design-time-series

Domain

Store the data

Question 34Skipped

Your team is planning to perform tests on Cloud BigTable instance to ensure the performance quality of the BigTable instance to be used in production. Which of the following conditions should be met to consider the performance testing valid? (Choose 3)

Correct selection

A. Use production instance.

B. Use at least 1TB of data.

Correct selection

C. Use at least 300GB of data.

Correct selection

D. Tests should run for at least 10 minutes.

E. Use development instance.

F. Storage type should be HDD.

Overall explanation

Answers: A, C & D.

If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:

1. Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.
2. Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use at least 100 GB of data per node.
3. Stay below the recommended storage utilization per node.
4. Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.
5. Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

Source(s):

Understanding BigTable
Performance: https://cloud.google.com/bigtable/docs/performance

Domain

Store the data

Question 35Skipped

Your company hosts a gaming app which reaches over 30,000 players in a single minute. The app generates event data including information about players state, score, location coordinates and other stats. You need to find a storage solution which can support high read/write throughput with very low latency which doesn't exceed 10 milliseconds to ensure a quality performance experience for the players.

Which of the following is the best option for this scenario?

A. Cloud Spanner

B. BigQuery

Correct answer

C. BigTable

D. Datastore

Overall explanation

Answer: C.

Cloud BigTable is a petabyte-scale, fully managed NoSQL database service for large analytical and operational workloads. Under a typical workload, Cloud BigTable delivers highly predictable performance. When everything is running smoothly, a typical workload can achieve the following performance for each node in the Cloud Bigtable cluster, depending on which type of storage the cluster uses:

Storage Type	Reads		Writes	Scans
SSD	10,000 rows per second @ 6 ms	or	10,000 rows per second @ 6 ms	220 MB/s
HDD	500 rows per second @ 200 ms	or	10,000 rows per second @ 50 ms	180 MB/s

In general, a cluster's performance increases linearly as you add nodes to the cluster. For example, if you create an SSD cluster with 10 nodes, the cluster can support up to 100,000

rows per second for a typical read-only or write-only workload, with 6 ms latency for each read or write operation.

Answer A is incorrect: Cloud Spanner does not guarantee the same performance and low latency as BigTable.

Answer B is incorrect: While BigQuery is a potential choice, BigQuery doesn't provide high throughput and low latency as powerful as BigTable.

Answer D is incorrect: Datastore can be a potential choice since it's a NoSQL database. The issue is, Datastore is not built for storing and reading huge data volumes as in this scenario. Datastore is deisgned for web applications of small scale.

Source(s):

Understanding BigTable
Performance: https://cloud.google.com/bigtable/docs/performance

Domain

Store the data

Question 36Skipped

The development team decided to use BigTable to write thousands of incoming stream data. Their choice was based on BigTable's high performance and high throughput and low latency. However, the team is facing less than expected performance from the cluster. You are asked for advice on the reason for BigTable's instance performance issue.

Which of the following can be a reason for the performance issue of BigTable cluster?

Choose 2 Options.

Correct selection

A. Row key used is increased monotonically.

B. Rows are less than 10MB of size.

Correct selection

C. HDD disk is used in the BigTable cluster.

D. Cluster is launched in a region different than where users reside.

Overall explanation

Answer: A & C.

The most common issue for time series in Cloud Bigtable is *hotspotting*. This issue can affect any type of row key that contains a monotonically increasing value. In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting. Because Cloud Bigtable stores adjacent row keys on the same server node, all writes will focus only on one node until that node is full, at which point writes will move to the next node in the cluster.

BigTable instances are best performed with big size of data and SSD.

URN: https://cloud.google.com/bigtable/docs/performance#testing

Answer B is incorrect: You can read and write larger amounts of data per row, but increasing the amount of data per row will also reduce the number of rows per second. So, having rows with small size is better.

Answer D is incorrect: While selecting a region has an impact on network performance, it doesn't affect BigTable's performance when reading and writing data. It impacts the replication of data in the region

---

URN: https://cloud.google.com/bigtable/docs/performance#performance-replication

When you use multi-cluster routing, replication for Bigtable is eventually consistent. As a general rule, it takes longer to replicate data across a greater distance. Replicated clusters in different regions will typically have higher replication latency than replicated clusters in the same region.

Source(s):

Understanding BigTable Performance:

- https://cloud.google.com/bigtable/docs/performance

Domain

Store the data

Question 37Skipped

Data team is looking for a database system which is highly available and supports atomic transactions. Database should have a flexible but semi-structured schema and supports querying using SQL-like language. Solution should be fully managed with no planned downtime.

Which of the following is the best choice for this scenario?

A. Cloud SQL

B. Cloud Spanner

C. BigTable

Correct answer

D. Datastore

Overall explanation

Answer: D.

Cloud Datastore is a highly-scalable NoSQL database for your applications. Cloud Datastore automatically handles sharding and replication, providing you with a highly available and durable database that scales automatically to handle your applications' load. Cloud Datastore provides a myriad of capabilities such as ACID transactions, SQL-like queries and indexes.

Answer A is incorrect: Cloud SQL is a relational database. The scenario requires a flexible semi- structured schema and relational databases are strictly-structured.

Answer B is incorrect: Cloud Spanner is a relational database supports multi-regional and continental scaling-out. The scenario requires a flexible semi-structured schema and relational databases are strictly-structured.

Answer C is incorrect: BigTable does not have a semi-structured schema.

Source(s):

Cloud Datastore: https://cloud.google.com/datastore/

Domain

Store the data

Question 38Skipped

You have an existing data pipeline which uses Apache Spark to transform the data to be exported in CSV format, to be later loaded into MySQL database for further analysis. The CTO decides it's time to migrate the pipeline to Google Cloud. As a data architect, you are tasked to design the new pipeline with minimal changes to the current pipeline architecture for a smoother migration.

Which of the following approaches are best suitable for your CTO's requirement?

A. Use Dataproc for data transformation. Use Google Storage for storing data.

B. Use Dataflow for data transformation. Use BigQuery for storing data.

Correct answer

C. Use Dataproc for data transformation. Use BigQuery for storing data.

D. Use Dataflow for data transformation. Use Google Storage for storing data.

Overall explanation

Answer: C.

For this scenario, it’s recommended to migrate the existing pipeline as-is. So, for transformation, Dataproc should be used to run Apache Spark jobs. Hence, answers B & D are incorrect.

For data warehousing, BigQuery is a good alternative to MySQL. Since BigQuery supports SQL- querying and schema is structured. BigQuery can be also used for analysis and reporting integrating it with Dataprep & Data Studio. So, answer C is correct.

Source(s):

Cloud Dataproc: https://cloud.google.com/dataproc/

Cloud Dataflow: https://cloud.google.com/dataflow/

Option A suggests use of google storage for storing data, the data in google storage will be stored as objects which can't be analyzed directly. It'll have to either imported to My SQL etc.

Now, MySQl is not given in either of options. Hence, we can't assume on our own. CTO requirement is finally to analyze the data.

Domain

Store the data

Question 39Skipped

A company decided to migrate their on-premise hadoop jobs to Google Cloud. As recommended by Google Cloud engineers, Dataproc is used to run Apache Hive jobs. Data residing in on-premise HDFS has been moved to Google Storage and connector was used for Dataproc to read the data. Upon monitoring the performance of Dataproc clusters running Hive jobs, you noticed the jobs are I/O intensive and use local disk to read/write data. This leads to performance issues. How can you solve this problem?

A. Increase persistent disk size for master node.

B. Increase persistent disk size for worker nodes.

C. Increase RAM capacity of Dataproc cluster's worker nodes.

Correct answer

D. Use local HDFS storage of Dataproc cluster nodes instead of Google Storage.

Overall explanation

When you want to move Hadoop & Spark workloads from an on-premises environment to Google Cloud Platform (GCP), It's recommended to use Dataproc to run Apache Spark & Hadoop clusters. Local HDFS storage is a good option if you have workloads that involve heavy I/O. For example, you have a lot of partitioned writes. It is a good option if you also have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation.

Option A is incorrect: Increasing disk size for master node will not help with the performance issue.

Option B is incorrect: Increasing disk size for worker nodes alone is not enough. You should move data to local HDFS storage of Dataproc. Increasing size may help to increase HDFS storage.

Option C is incorrect: Increasing memory will not help fix the issue because the problem is because of intensive disk read/write.

Source(s):

Migrating Apache Spark Jobs to Cloud Dataproc: https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc

Domain

Store the data

Question 40Skipped

Your team decided to use BigTable for storing event data. The engineer responsible of launching and testing the instance has reported a slower performance than expected by Google Cloud documentation. Which of the following could be a factor for the slow performance? (Choose 3)

A. The rows in the tables tested contain very few number of cells.

B. The rows in the tables have small data size.

Correct selection

C. The schema is not designed for the instance to evenly read and write data across the tables.

Correct selection

D. The instance uses HDD storage type.

Correct selection

E. The instance was scaled up recently.

F. The instance has too high number of nodes for the data size tested.

Overall explanation

Answer: C, D & E.

There are several factors that can cause Cloud Bigtable to perform more slowly than expected:

The table's schema is not designed correctly. To get good performance from Cloud BigTable, it's essential to design a schema that makes it possible to distribute reads and writes evenly across each table.

The workload isn't appropriate for Cloud BigTable. If you test with a small amount (< 300 GB) of data, or if you test for a very short period of time (seconds rather than minutes or hours), Cloud BigTable won't be able to balance your data in a way that gives you good performance.

The rows in your Cloud Bigtable table contain large amounts of data. You can read and write larger amounts of data per row, but increasing the amount of data per row will also reduce the number of rows per second.

The rows in your Cloud Bigtable table contain a very large number of cells. It takes time for Cloud Bigtable to process each cell in a row. Also, each cell adds some overhead to the amount of data that's stored in your table and sent over the network.

The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance.

The Cloud Bigtable cluster was scaled up or scaled down recently. After you change the number of nodes in a cluster, it can take up to 20 minutes under load before you see an improvement in the cluster's performance.

The Cloud Bigtable cluster uses HDD disks. In most cases, your cluster should use SSD disks, which have significantly better performance than HDD disks.

The Cloud Bigtable instance is a development instance. Development instance's performance is equivalent to an instance with one single-node cluster, it will not perform as well as a production instance.

There are issues with the network connection. Network issues can reduce throughput and cause reads and writes to take longer than usual.

Source(s):

Understanding BigTable
Performance: https://cloud.google.com/bigtable/docs/performance

Domain

Store the data

Question 41Skipped

An online learning platform wants to generate captions for its videos. The platform offers around 2,500 courses with topics about business, finance, cooking, development & science. The platform allows content with different languages such as French, German,

Turkish and Thai. Thus, this can be very difficult for a single team to caption all available courses and they are looking for an approach which helps do such massive job.

Which product from Google Cloud will you suggest them to use?

Correct answer

A. Cloud Speech-to-Text.

B. Cloud Natural Language.

C. Vertex AI

D. AutoML Vision API.

Overall explanation

Answer: A.

Answer A is correct: Cloud Speech-to-Text is a service to generate captions from videos by detecting speakers language and speech.

Answer B is incorrect: Cloud natural language service is to derive insights from unstructured text revealing meaning of the documents and categorize articles. It won't help extracting captions from videos.

Answer C is incorrect: Machine Learning Engine is a managed service letting developers and scientists build their own models and run them in production. This means, you have to build your own model to generate text from videos which needs much effort and experience to build such model. So, it's not a practical solution for this scenario.

Answer D is incorrect: AutoML Vision API is a service to recognize and derive insights from images by either using pre-trained models or training a custom model based on a set of photographics.

Source(s):

Google NLP: https://cloud.google.com/natural-language/

Google Machine Learning Engine: https://cloud.google.com/ml-engine/

Google Vision API: https://cloud.google.com/vision

Google Speech-to-Text API: https://cloud.google.com/speech-to-text/

Domain

Ingest and process the data

Question 42Skipped

You need to build a machine learning model to recognize different animals for a pet shop. The purpose is to scan the photos on their twitter page and get stats about what pets people like sharing while tagging the pet shop brand the most. Due to cost constraints, the project should be as cost-effective as possible, and that includes work hours dedicated to the project.

Which approach will you consider to build the project?

A. Use Cloud ML Engine API and inspect the descriptions returned by the API. Consider the description with highest score.

Correct answer

B. Use Vision API and inspect the descriptions returned by the API. Consider the description with highest score.

C. Use Vision API and inspect the MID values returned by API to recognize the pets in photos.

D. Use Vision API and inspect the descriptions returned by the API. Consider the description with median score.

Overall explanation

Answer B.

Google AutoML Vision API automates the training of your own custom machine learning models by simply uploading images and training custom image models with an easy-to-use graphical interface.

Google AutoML Vision is recommended in this scenario because you can build an image recognition model quickly with less work time, comparing to building your very own model from scratch.

Answer A is incorrect: Any approach other than using AutoML vision API is not recommended.

When inspecting returned values from Vision API, here, you need to check the output's values descriptions to recognize what type of animal recognized by API. Descriptions with highest score should be considered as they have better prediction.

Answer C is incorrect: MID values are not useful for this scenario.

Answer D is incorrect: API does not provide median scores. It provides a rate how likely the description is accurate.

Source(s):

Google Vision API – Detect Labels: https://cloud.google.com/vision/docs/labels

Domain

Ingest and process the data

Question 43Skipped

You are working on building your own machine learning model and training it. When you tested the model on a testing set, you realized the error rate is very high and the model's output only matched 25% of expected output.

What is the problem you are facing and how can you fix it?

A. The model is underfitting: You need to increase the features and use more training data.

B. The model is underfitting: You need to lower the features and use less training data.

Correct answer

C. The model is overfitting: You need to lower the features and use more training data.

D. The model is overfitting: You need to increase the features and use more training data.

Overall explanation

Answer: C.

Overfitting happens when a model goes well on a training set, generating only a small error, while giving wrong output for the test set. This happens because the model is only picking up specific features input found in the training set instead of picking out general features of the given training set.

The opposite of overfitting is *underfitting*. Underfitting occurs when there is still room for improvement on the test data. This can happen for a number of reasons: If the model is not powerful enough, is over-regularized, or has simply not been trained long enough. This means the network has not learned the relevant patterns in the training data.

To solve overfitting, the following would help improving the model's quality:

-
    - Increase the number of examples, the more data a model is trained with, the more use cases the model can be training on and better improves its predictions.
    - Tune hyperparameters which is related to number and size of hidden layers (for neural networks), and regularization, which means using techniques to make your model simpler such as using dropout method to remove neuron networks or adding “penalty” parameters to the cost function.
    - Remove features by removing irrelevant features. Feature engineering is a wide subject and feature selection is a critical part of building and training a model. Some algorithms have built- in feature selection, but in some cases, data scientists need to cherry-pick or manually select or remove features for debugging and finding the best model output.

From the brief explanation, to solve the overfitting problem in the scenario, you need to:

-
    - Increase the training set.
    - Decrease features parameters.

Hence, answer C is correct.

Answer A & B are incorrect: The problem in this scenario is not underfitting.

Answer D is incorrect: You should work on decreasing the features to solve overfitting, not increasing them.

Source(s):

Overfitting and
underfitting: https://www.tensorflow.org/tutorials/keras/overfit_and_underfit

Domain

Ingest and process the data

Question 44Skipped

You are building a machine learning model to solve a binary classification problem. The model is going to predict the likelihood of a customer to be using a fraudulent credit card when purchasing online.

Since there is a very small fraction of purchase transactions are proved to be fraudulent, more than 99% of the purchase transactions are valid. You want to make sure the machine learning model is able to identify the fraudulent transactions.

What is the technique to examine the effectiveness of the model?

A. Gradient Descent

Correct answer

B. Recall

C. Feature engineering

D. Precision

Overall explanation

Answer: B.

Precision is the formula to check how accurate the model is when most of the output are positives. In other words, if most of the output is yes.

Recall: is the formula to check how accurate the model is when most of the output are negatives. In other words, if most of the output is no.

Gradient Descent is an optimization algorithm to find the minimal value of a function. Gradient descent is used to find the minimal minimal RMSE or cost function.

Feature Engineering is the process of deciding which data is important for the model.

From the description, answers A & C are incorrect. It leaves us with B & D.

Since the scenario mentions very little likelihood a transaction can be fraudulent. There are more “no” than “yes” means more negative than positive. Hence, to calculate the effectiveness of the model, you should use recall formula.

Source(s):

Precision & Recall: https://developers.google.com/machine-learning/crash-course/classification/ precision-and-recall

Gradient Descent: https://en.wikipedia.org/wiki/Gradient_descent

Feature Engineering: https://cloud.google.com/ml-engine/docs/tensorflow/data-prep

Domain

Ingest and process the data

Question 45Skipped

A Kafka cluster is receiving event data from outsourced sensors. The cluster is installed in a Compute Engine instance and it writes events to Google Storage. Due to the new security rules in the company, data written to Google Storage should be encrypted. Security team wants to be sure encryption key used is provided by them using on-premise vault and no keys generated by third-parties are used.

What should you do to follow security team's rules?

Correct answer

A. Reference the encryption key provided by security team when calling API service when writing data to Google Storage to encrypt the data.

B. Store the encryption key provided by security team in Compute Engine instance and reference it when calling API service when writing data to Google Storage to encrypt the data.

C. Store the encryption key provided by security team in Cloud Key Management Service (KMS) and reference it when calling API service when writing data to Google Storage to encrypt the data.

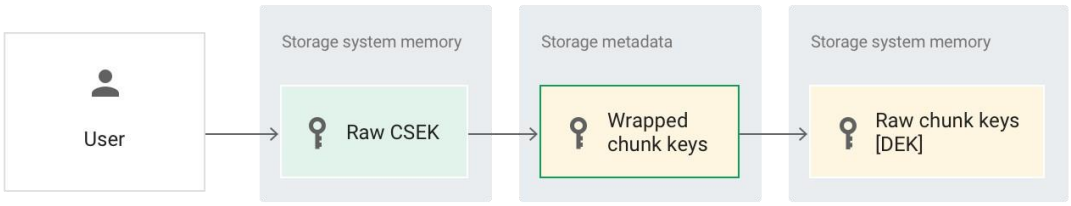
D. Create encryption keys using Cloud Key Management Service (KMS) and reference it when calling API service when writing data to Google Storage to encrypt the data.

Overall explanation

Answer: A.

Customer-Supplied Encryption Keys (CSEK) are a feature in Google Cloud Storage and Google Compute Engine. If you supply your own encryption keys, Google uses your key to protect the Google- generated keys used to encrypt and decrypt your data.

When you use Customer-Supplied Encryption Keys in Cloud Storage, you provide a raw CSEK as part of an API call. This key is transmitted from the Google front end to the storage system’s memory. This key is used as the key encryption key in Google Cloud Storage for your data.



The raw CSEK is used to unwrap wrapped chunk keys, to create raw chunk keys in memory. These are used to decrypt data chunks stored in the storage systems. These keys are used as the data encryption keys (DEK) in Google Cloud Storage for your data.

Answer B & C are incorrect: Security team does not recommend storing encryption key in the cloud and using on-premise safe storage instead.

Answer D is incorrect: Security team doesn't allow using generated keys from KMS.

Source(s):

Customer-Supplied Encryption Keys: https://cloud.google.com/security/encryption-at-rest/customer- supplied-encryption-keys/

Domain

Prepare and use data for analysis

Question 46Skipped

You have a PostgreSQL database instance on Cloud SQL. The database is used in production by the different micro-services writing and reading to it based on each micro-service's needs. Data analysts connect to the instance to run their analysis and reporting SQL queries which adds more load to the database with intensive I/O operations on it. You want to find a solution which can avoid intensive I/O operations on production database and allow data analysts to run their queries without interruptions may lead to late reports to management. What should you do?

A. Scale up Cloud SQL instance from standard to high CPU machine type.

Correct answer

B. Use Cloud SQL read replicas to replicate production instance. Share endpoint with data analysts to replace with existing endpoint in their SQL clients to use the replica.

C. Create a snapshot from production instance. Create a new Cloud SQL instance from the snapshot. Use Cloud Scheduler cron job to create a snapshot and a new instance daily.

D. Import data from Cloud SQL instance to BigQuery. Create necessary users on BigQuery dataset created for data analysts to run their SQL queries on.

Overall explanation

Answer: B.

Description:

Cloud SQL provides the ability to replicate a master instance to one or more read replicas. A read replica is a copy of the master that reflects changes to the master instance in almost real time. You create a replica to offload read requests or analytics traffic from the master. You can create multiple read replicas for a single master instance. By default, the replica is created with the same number of CPUs and the same amount of memory as the master instance. You can increase these values for the replica, but you cannot decrease them.

Answer A is incorrect: Scaling up instance by CPU will not solve the core issue which is analysts using production instance for intensive reads.

Answer C & D are incorrect: While both approaches are viable, they require maintenance and implementation to be done by engineers. Google Cloud provides a fully-managed service called read replica can handle replicating production instance in near real-time manner.

Source(s):

Cloud SQL Instance Settings:

https://cloud.google.com/sql/docs/mysql/instance-settings

Cloud SQL PostgreSQL – Creating Read Replicas:

https://cloud.google.com/sql/docs/postgres/replication/create-replica

Domain

Store the data

Question 47Skipped

You receive a daily comma-separated (CSV) file which should be imported to BigQuery. You need to scan the file in case of incomplete or improperly aligned column values which will cause importing to BigQuery fail.

What should you do to handle invalid inputs?

A. Import data to BigQuery, then run queries to check if data skew exists among table fields.

B. Import file to BigQuery with setting –max_bad_records.

C. Use Google Stackdriver to monitor import status and create an alert for failed imports.

Correct answer

D. Trigger the Cloud Function on the GCS bucket and process the CSV file, insert the valid records in the Bigquery table, and push the invalid records to the error table in Bigquery for later analysis

Overall explanation

Answer: D.

A better way to solve this problem would be to have a dead letter file where all of the failing inputs are written for later analysis and reprocessing. We can use a side output in Dataflow to accomplish this goal.

Answer A is incorrect: This won't avoid the fact data will fail loading into BigQuery.

Answer B is incorrect: Setting –max_bad_records option won't help processing the file, and if number of errors exceed the value set for this option, it will return an error and cause the job to fail.

Answer C is incorrect: Stackdriver doesn't have a native support for detecting BigQuery import failures. Also, this does not help fixing corrupted file.

Source(s):

Handling Invalid Inputs in Dataflow: https://cloud.google.com/blog/products/gcp/handling-invalid- inputs-in-dataflow

BigQuery – Loading CSV Data from Google Storage: https://cloud.google.com/bigquery/docs/loading- data-cloud-storage-csv

Domain

Prepare and use data for analysis

Question 48Skipped

The data analysts in your company want to prepare data sets for reporting to upper management. While the current data pipeline does part of data modeling to the data sets, data analysts still want to perform extra data profiling on data such as detecting duplicates, count null values and other profiling techniques. They ask your advice on what tool to use.

Which of the following is recommended?

Correct answer

A. Cloud Dataprep

B. Dataproc

C. Cloud Composer

D. Cloud Function

Overall explanation

Answer: A.

Cloud Dataprep is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning.

Because Cloud Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

With automatic schema, datatype, possible joins, and anomaly detection, you can skip time-consuming data profiling and focus on data analysis.

Answer B is incorrect: Dataproc is a complicated service for data profiling comparing to Dataprep.

Answer C & D are incorrect: Cloud Function and Cloud Composer don't directly help with data modeling and profiling without coding and pipeline design.

Source(s):

Cloud Dataprep: https://cloud.google.com/dataprep/

Domain

Prepare and use data for analysis

Question 49Skipped

You are in need to restore a snapshot of a compute engine instance you have previously scheduled for regular daily snapshots. Which of the following are the steps you should do to perform the restoration?

Correct answer

A. You can simply create a replacement instance directly by selecting the snapshot from the list of snapshots available.

B. You need to create a persistent disk from the snapshot of your choice. Create a new compute engine instance and attach the persistent disk to it.

C. Create a new compute instance with the same exact machine type as the one the snapshot was created from. Create a persistent disk using the snapshot to be restored from. Attach the persistent disk to the compute engine instance.

D. Export snapshot to Google Storage. Create new compute engine instance, then using gsutil tool, copy the snapshot to the instance's persistent disk to be restored.

Overall explanation

Answer: A.

Google Cloud supports easy snapshot restoration to a persistent disk as well resotring a book disk snapshot to create a new VM instance. You can simply create a replacement instance directly by selecting the snapshot from the list of snapshots available.

Source(s):

Restoring and Deleting Persistent Disk Snapshots: https://cloud.google.com/compute/docs/disks/restore-and-delete-snapshots

Domain

Prepare and use data for analysis

Question 50Skipped

Your data team is using BigQuery as their main data warehouse. There is no formal security policy implemented to track users activity on the data warehouse. A new security policy to be implemented which states any activity on cloud resources should be tracked and logged and BigQuery is one of these resources.

What action should be taken to log the data warehouse's activities?

A. Restrict users access to BigQuery's tables using Identity & Access Management (IAM).

B. You can list all query history from BigQuery UI.

Correct answer

C. Use Google Audit Logs to capture all the activities for Bigquery and review data warehouse access.

D. Enable caching on BigQuery to allow auditing users activity.

Overall explanation

Answer: C.

Cloud Audit Logs are a collection of logs provided by Google Cloud Platform that provide insight into operational concerns related to your use of Google Cloud services. BigQuery automatically sends audit logs to Stackdriver Logging.

All remaining activities and events are reported as part of the main activity stream. Events such as job insertions and completions are reported in this stream, as are other resource events such as dataset creation. Those are called “Admin activity.”

Answer A is incorrect: You cannot restrict BigQuery access by table level. The lowest level is the dataset.

Answer B is incorrect: BigQuery UI’s query history has a limit of 1,000 cumulative jobs.

Answer D is incorrect: Enabling cache has nothing to do with audit logging in BigQuery.

Source(s):

Audit Logs: https://cloud.google.com/bigquery/docs/reference/auditlogs/

Domain

Prepare and use data for analysis