





Level: Intermediate

## Google Cloud Certified Professional Data Engineer

← Back to the **Course**

Practice Test 1

Completed on **Sat, 06 Jul 2024**

2nd
Attempt

46/50
Marks Obtained

92.00%
Your Score

1h 11m 15s
Time Taken

PASS
Result

Share this Report in Social Media Share

Download Report

### Domain wise Quiz Performance Report

| No. | Domain | Total Question | Correct | Incorrect | Unattempted |
|---|---|---|---|---|---|
| 1 | Design Data Processing Systems | 19 | 18 | 1 | 0 |
| 2 | Store the data | 12 | 10 | 2 | 0 |
| 3 | Ingest and process the data | 12 | 11 | 1 | 0 |
| 4 | Prepare and use data for analysis | 7 | 7 | 0 | 0 |
| Total | All Domains | 50 | 46 | 4 | 0 |

### Review the Answers

Filter By All Questions

Question 1 Correct

**Domain:** Design Data Processing Systems

An environment safety facility receives thousands of events every 60 seconds from its sensors assembled in different sectors monitoring air pollution in the region. Scientists want to access and query the data for observation and daily reporting. Due to current funding state, their budget is limited and they seek a cost-effective, highly available and ACID-compliant solution supports SQL querying.

Which approach would you recommend for such scenario?

**A. Use BigQuery to store and query the event data. Enable streaming on BigQuery for data to be loaded in real-time.**

**B. Use Pub/Sub to stream events and ingest data into Bigquery using Bigquery Subscription**

**C. Use Cloud SQL to load events into a relational database and allow access to scientists to query.**

**D. Use BigQuery to store and query event data. Batch load the data to BigQuery using its API.**

## Explanation:

**Correct Answer: D**

BigQuery supports both batch & streaming data. However, due to mentioned budget restrictions, the solution would choose the cheaper approach, which is batching data to BigQuery. Batching data to BigQuery is free of charge. Streaming data on the other hand is charged by size.

So, answer D is correct.

**Answer A is incorrect**: Enabling streaming on BigQuery is not free. Since the budget is limited and the scenario seeks cost-effective solution, this approach is not recommended over batching.

**Answer B is incorrect: Pub/Sub** is a good option for streaming data from sensors to BigQuery, but it cannot be used to store and query the data directly. To store and query the data, the scientists would need to use another service, such as BigQuery or Cloud SQL

**Answer C is incorrect:** Cloud SQL needs administration and not easily scalable. Cloud SQL does not provide batching tools. BigQuery is a better approach for such scenario.

**Source(s):**

Bigquery: Streaming data: https://cloud.google.com/bigquery/streaming-data-into-bigquery

Bigquery: Batch data: https://cloud.google.com/bigquery/batch

Bigquery Pricing: https://cloud.google.com/bigquery/pricing

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

Ask our Experts

Did you like this **Question?**

Question 2 Correct

**Domain:** Design Data Processing Systems

You have a dataflow pipeline read a CSV file daily at 6 am, applies the needed cleansing & transformation on it, and then loads it to BigQuery. Occasionally, the CSV file might be modified within the day due to human error or incomplete data. This causes you to manually re-run dataflow pipeline again. Is there a way to ensure that the Dataflow pipeline automatically re-runs when the CSV file is modified due to human error or incomplete data, thus achieving a more efficient solution?

**A. Use Cloud Scheduler to re-run dataflow after 6 am. Check what is the average time the file is modified and schedule based on it**

**B. Create a Cloud Function that is triggered when an object is modified in a Google Cloud Storage (GCS) bucket. The Cloud Function will then trigger a Dataflow pipeline to reprocess the data in the modified CSV file**

**C. Use Cloud Composer to rerun dataflow and reprocess the file. Create a custom sensor to detect file conditions if changed**

**D. Use a compute engine to schedule a cron job to run every 10 minutes to check if the file was modified to rerun dataflow**

## Explanation:

**Correct Answer: B**

This approach is correct because it utilizes a Cloud Function as a trigger for Google Cloud Storage (GCS) bucket object modifications. When an object is modified, such as the CSV file, the Cloud Function is activated. Subsequently, the Cloud Function triggers a Dataflow pipeline, ensuring automatic reprocessing of the data in the modified CSV file. This method efficiently addresses the requirement for automated data reprocessing upon file modification.

**Answer A is incorrect:** Guessing what time scheduler should rerun dataflow is inefficient.

**Answer C is incorrect:** Cloud Composer is an overkill for the given issue. The above issue can be easily automated using Cloud Function.

**Answer D is incorrect:** Using a Compute Engine VM is unnecessary because Cloud Composer can orchestrate Dataflow pipeline's failure.

**Source(s):**

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

Cloud Functions: https://cloud.google.com/functions/

Cloud Function Triggers: https://cloud.google.com/functions/docs/calling/storage

Ask our Experts

Did you like this **Question?**

## Question 3

Correct

**Domain:** Design Data Processing Systems

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?

**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**

## Explanation:

Answer: A.

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 scenario.

**Answers B & C are wrong** because these are not scalable solutions.

**Answer D is wrong** because Dataflow cannot ingest event streams. It needs Pub/Sub service to do so.

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

## Question 4

Correct

**Domain:** Design Data Processing Systems

A social media platform stores various details of their platform users such as session login time, URLs visited, activities on platform and other logs. With GDPR (General Data Protection Regulation) compliance to be officially implemented, the platform now allows users to download their activity logs from their profile settings which they can click a button to call an API to generate a full report.

Recently, users are complaining timeouts after 60 seconds of requesting to download their activity logs at peak hours when the platform has the most traffic. They have to try for several minutes or even hours for the API to return their report available for download.

How can you solve this issue?

A. Increase timeout for API at peak times to 120 seconds. If it keeps failing, try increasing the timeout until the issue is resolved.

B. Build a Dataflow pipeline to generate daily reports of users' activity logs. Users can download those daily reports whenever they want to.

C. Migrate data source to Cloud Spanner for horizontal scaling to avoid query timeouts.

D. Use Pub/Sub to receive requests for activity logs from users. Deploy a Cloud Function with a Pub/Sub trigger to generate the reports and store them in a GCS bucket. Then, send temporary download links to the users via email.

## Explanation:

Answer: D.

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.

Pub/sub is a good product to de-couple a system's components so they communicate with each other asymmetrically. From the scenario shown here, instead of directly calling the API to export required report which puts great loads on the API and hence the timeouts faced by users. Instead, the platform can "publish" messages to a "topic" related to exporting activity log reports sending the required parameters such as user ID and custom settings such as date range and what data to export. The API can be switched to be a "subscriber" which receives the messages sent and processes each message asymmetrically to generate the report, then sends the download link to the user's mailbox when ready.

Hence, answer D is correct.

**Answer A is incorre**ct: Increasing timeout isn't a scalable solution and it may keep occurring eventually when more and more users join the platform.

**Answer B is incorect**: While this would solve the timeout issues, generating daily reports for users can be costly as more users join, knowing that requesting activity log reports are a non-frequent action and this costs both compute and storage resources. This solution also doesn't provide flexibility with what parameters the report is generated on such as date range and other custom metrics.

**Answer C is incorrect**: This solution has several issues. First, we're assuming the data source is a relational database, which can be unlikely since NoSQL databases better perform for massive log input which uses the user ID as a key to reach the data. Second, Cloud Spanner isn't a cheap solution for a service not frequently used.

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

Question 5 Correct

**Domain:** Design Data Processing Systems

A company decided to migrate its on-premise data infrastructure to the cloud mainly for the high availability of cloud services and to lower the high costs of storing data on-premise. The infrastructure uses HDFS to store data, process, and transform 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?

**A. Use Dataproc to process the data. Store data in Google Storage**

**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. To process your data quickly and affordably, use a Dataproc cluster with preemptible VMs. Then, store the processed data in Google Cloud Storage, and use object lifecycle management to automatically manage the data's lifecycle**

## Explanation:

**Correct Answer: D**

**Option A is incorrect:** Here normal dataproc clusters with non-preemptible workers are used, which is not a cost-effective solution as the main reason to migrate from on-premise to cloud is to deploy infrastructure when required.

**Option B is incorrect:** Dataflow is serverless which may not suit DevOps requirements 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 Correct:** 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.

**Source(s):**

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

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

Google Cloud Platform Migration Best Practices (Hadoop):

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

Google Cloud Storage Object Lifecycle Management:

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

Ask our Experts

Did you like this **Question?**

## Question 6

Correct

**Domain:** Design Data Processing Systems

Data analysts are switching to use Apache Spark to perform experiments on the data before applying

the changes to production. Those experiments are not critical, but they will be conducted on big data sets. As a data engineer, the head of data asked you to prepare the tech stack required to be used by data analysts to run their Spark scripts and experiment on with taking into consideration the cost of the stack used.

Which of the following tech stack is suggested?

A. Launch a Dataproc cluster in high-availability mode with using high-memory worker machine types.

B. Launch a Dataproc cluster in standard mode with using high-CPU worker machine types.

C. Launch a Dataproc cluster in standard mode with using high-memory worker machine types.

D. Advice the data analysts to use Dataprep for their data manipulation.

## Explanation:

Answer: C.

**Answer C is correct:** The data sets are big in size and hence high memory machine is the choice.

**Answer A is incorrect:** Since the scenario states non-critical experiments will be conducted by data analysts, Dataproc cluster used can be in standard mode.

**Answer B is incorrect**: Since the scenario states non-critical experiments, there is no need for high-CPU worker machine types.

**Answer D is incorrect**: Dataprep does not provide Apache Spark job transformation. Dataprep is best for visual exploration and manual cleaning and preparation of data for analysis and machine learning.

**Source(s)**:

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

Ask our Experts

Did you like this **Question?**

Question 7 Correct Marked for review

**Domain:** Design Data Processing Systems

You have several Data Studio reports reading from BigQuery. Those reports are used to visualize several metrics for marketing team. Data visualized is updated only once a day. You notice that reports running queries on BigQuery are not free and they cost for each query. You want to control and minimize the costs caused by frequent queries coming from Data Studio dashboards.

What should you do?

**A. Enable caching on reports for reading from BigQuery. No need to change the credentials.**

**B. Grant owner credentials for the reports on BigQuery datasets and enable caching.**

**C. Configure reports data sources to update data every 24 hours only.**

**D. Export data as CSV files to Google Storage every 24 hours and change reports data source to read from those files.**

## Explanation:

Answer: A.

BigQuery writes all query results to a table. The table is either explicitly identified by the user (a destination table), or it is a temporary, cached results table. Temporary, cached results tables are maintained per-user, per-project. There are no storage costs for temporary tables, but if you write query results to a permanent table, you are charged for storing the data.

When you run a query, a temporary, cached results table is created in a special dataset referred to as an "anonymous dataset". Unlike regular datasets which inherit permissions from the IAM resource hierarchy model (project and organization permissions), access to anonymous datasets is restricted to the dataset owner. The owner of an anonymous dataset is the user who ran the query that produced the cached result.

When an anonymous dataset is created, the user that runs the query job is explictly given `bigquery.dataOwner` access to the anonymous dataset. `bigquery.dataOwner` access gives only the user who ran the query job full control over the dataset. This includes full control over the cached results tables in the anonymous dataset. If you intend to share query results, do not use the cached results stored in an anonymous dataset. Instead, write the results to a named destination table.

Though the user that runs the query has full access to the dataset and the cached results table, using

them as inputs for dependent jobs is strongly discouraged.

The names of anonymous datasets begin with an underscore. This hides them from the datasets list in the GCP Console and the classic BigQuery web UI. You can list anonymous datasets and audit anonymous dataset access controls by using the CLI or the API.

**Answer B is incorrect:** There is no need to enable owner credentials for responsive caching.

**Answer C is incorrect:** Maximum period for caching in Data Studio is 12 hours.

**Answer D is incorrect:** This is a very cumbersome option and there is no need to export to Google Storage when better options are available.

**Source(s):**

Data Studio – Manage Data Freshness: https://cloud.google.com/bigquery/docs/cached-results

Ask our Experts

Did you like this **Question?**

Question 8 Correct

**Domain:** Design Data Processing Systems

You are using BigQuery as the data warehouse. Data analysts & scientists run queries to get data from BigQuery. When you checked the billing costs for the previous month, you noticed a spike in running queries on BigQuery despite the caching is enabled. You tried to find out the reason for the spike by reading some of the queries data analysts and scientists are running on BigQuery.

Which of the following can be the reason for increased bigquery costs? (Select THREE)

A. Queries use current_timestamp function

B. SELECT queries with asterisk (*)

C. Queries select from authorized views on archive tables

D. Querying multiple tables using a wildcard

## Explanation:

**Correct Answers: A, B and D**

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

If the query uses non-deterministic functions; for example, date and time functions such as CURRENT_TIMESTAMP() and NOW(), and other functions such as CURRENT_USER() return different values depending on when a query is executed

**Reference:**

Using cached query results | BigQuery | Google Cloud

Ask our Experts

Did you like this **Question?**

Question 9 Correct Marked for review

**Domain:** Design Data Processing Systems

You are deploying a Tensorflow model built by the data science team to the cloud. Based on the requirements provided by data scientists, the model should be able to return the output as soon as possible to minimize the latency of serving predictions. Input will be passed as JSON.
Which of the following approaches are best for this scenario?

A. Use Google Kubernetes Engine to deploy the model. Use online prediction to pass input data to the model hosted in cloud.

B. Use Google Kubernetes Engine to deploy the model. Use batch prediction to pass input data to the model hosted in cloud.

C. Use Vertex AI to deploy the model. Use batch prediction to pass input data to the model hosted in cloud.

D. Use Vertex AI to deploy the model. Use online prediction to pass input data to the model hosted in cloud.

## Explanation:

**Answer: D**

| 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:<br>• Legacy Editor or Viewer<br>• AI Platform Admin or Developer | Accounts with the following IAM roles can request batch predictions:<br>• Legacy Editor<br>• 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. |

Vertex AI 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.

Online prediction passes input as a JSON string and returns the output as soon as possible.

**Options A and B are incorrect**: GKE is not a recommended option to deploy the model.

**Option C is incorrect:** Batch prediction doesn't support returning the output as soon as possible. Input is passed indirectly as one or more URIs of files in Google Storage.

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

## Question 10

Correct Marked for review

**Domain:** Design Data Processing Systems

You have a massively multiplayer online (MMO) game which sends events from each player every 10 seconds. Events contain stats about the player session's state (play, idle, off) as well as ping duration. You want to use Dataflow for windowing. The purpose is to aggregate events and extracting stats to detect how many players are currently online and what is the average ping duration for each server in a time window of 30 seconds.

Which windowing function you should choose to design the pipeline?

A. Tumbling window

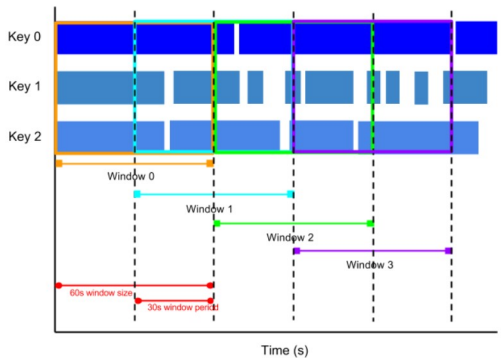
B. Hopping window

C. Session window

D. Global window

## 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.

Sliding-time window is the windowing function recommended for this scenario.

Option A, C and D are incorrect as the other window mechanisms, we would not be able to capture all the data points because of the "period".

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

## Question 11

Correct

**Domain:** Design Data Processing Systems

An e-payment service allows users to purchase online and transfer money securely. They log into the website to perform the transactions and they log out. The website needs to check if their sessions are idle for 10 minutes, means they did not perform any action or they opened a new link within the website. In case of idle session, the website ends their session for security purposes.

You need to build a Dataflow pipeline to aggregate session events received from the website and detect sessions idle more than 10 minutes to get their sessions expired.

Which windowing function you should choose to design the pipeline?

**A. Tumbling window with duration of 10 minutes**

**B. Hopping window with a duration of 10 minutes**

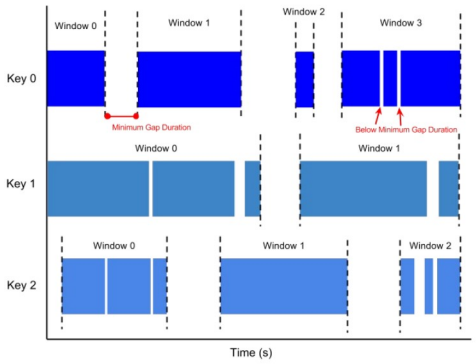
**C. Session window with a time gap duration of 10 minutes**

**D. Global window with time-based trigger of 10 minutes**

## 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

Ask our Experts

Did you like this **Question?**

Question 12 Correct

**Domain:** Design Data Processing Systems

You have several Dataflow pipelines streaming data for transformation and further analysis. At one point of the transformation, there is a need for two pipelines to share data among pipeline instances. You need to modify the architecture to allow data sharing between different pipelines.

How should this requirement be met in Google Cloud?

**A. Enable data sharing option when creating Dataflow pipeline.**

**B. Grant pipeline instances the right IAM roles to access other pipelines instances for data sharing.**

**C. Use Google Storage to share data with other pipeline instances.**

**D. Data sharing among Dataflow pipelines is only possible if instances reside in same region.**

## Explanation:

Answer: C.

There is no Cloud Dataflow-specific cross pipeline communication mechanism for sharing data or processing context between pipelines. You can use durable storage like Cloud Storage or an in-memory cache like App Engine to share data between pipeline instances.

**Answer A is incorrect:** Dataflow doesn't have a cross pipeline communication mechanism for sharing data between pipelines.

**Answer B is incorrect:** This approach is not recommended, if possible. Use Google Storage to share data between pipelines.

**Answer D is incorrect:** Sharing data is not possible unless using a reliable data storage such as Google Storage.

**Source(s):**

Dataflow – FAQ: https://cloud.google.com/dataflow/docs/resources/faq

Ask our Experts

Did you like this **Question?**

Question 13 Correct Marked for review

**Domain:** Design Data Processing Systems

You are building a data pipeline using Google Dataflow SDK. This pipeline is going to perform operations on data using conditional and for loops creating a branch pipeline.

Which of the following concepts should be used to achieve this?

A. ParDo

B. PCollection

C. PTransform

D. Pipeline

## Explanation:

Answer: C.

A transform represents a processing operation that transforms data. A transform takes one or more PCollections as input, performs an operation that you specify on each element in that collection, and produces one or more PCollections as output. A transform can perform nearly any kind of processing operation, including performing mathematical computations on data, converting data from one format to another, grouping data together, reading and writing data, filtering data to output only the elements you want, or combining data elements into single values.

**Source(s):**

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

beam-programming-model

Ask our Experts

Did you like this **Question?**

Question 14 Correct

**Domain:** Design Data Processing Systems

A multinational company has multiple Google Storage buckets in different regions around the world. Each branch has its own set of buckets in the region nearest to them to avoid latencies.

However, this led to a problem for the analytics team to reach and do the necessary reports on the data using BigQuery since they need to create the tables in the same region either to import the data or create external tables to access the data in different regions. The head of data decided to sync the data daily from different Google Storage buckets scattered in different regions to a single multi-regional bucket to do the necessary data analysis and reporting.

Which service could help with this approach?

**A. Appliance Transfer Service**

**B. gsutil**

**C. Storage Transfer Service**

**D. Dataflow**

## Explanation:

Answer: C.

Storage Transfer Service allows you to quickly import *ONLINE* data into Cloud Storage. You can also set up a repeating schedule for transferring data, as well as transfer data within Cloud Storage, from one bucket to another.

Transfer Appliance is an *OFFLINE* secure, high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, answer C is correct, while answer A is incorrect.

**Answer B is incorrect:** gsutil tool is good for programmatic usage by developers and may be useful to copy and move megabytes/gigabytes of data. Not so practical for Terabytes of data. It's also not reliable data transfer technique as it is related to the machine's connectivity with Google Cloud.

**Answer D is incorrect**: Dataflow as a solution may be viable, but you need to build a pipeline to migrate data from bucket to another. Storage Transfer Service can do it without the extra effort.

**Source(s):**

Google Cloud Storage Transfer Service: https://cloud.google.com/storage-transfer/docs/

Google Appliance Transfer Service: https://cloud.google.com/transfer-appliance/

Ask our Experts

Did you like this **Question?**

Question 15 Correct

**Domain:** Design Data Processing Systems

A company is migrating its current infrastructure from on-premise to Google cloud. It stores over 280TB of data on its on-premise HDFS servers. You were tasked to move data from HDFS to Google Storage in a secure and efficient manner. Which of the following approaches are best to fulfill this task?

A. Install Google Storage gsutil tool on servers and copy the data from HDFS to Google Storage.

B. Use Cloud Data Transfer Service to migrate the data to Google Storage.

C. Import the data from HDFS to BigQuery. Then, export the data to Google Storage in AVRO format.

D. Use Transfer Appliance Service to migrate the data to Google Storage.

## Explanation:

Answer: D.

Storage Transfer Service allows you to quickly import *ONLINE* data into Cloud Storage. You can also set up a repeating schedule for transferring data, as well as transfer data within Cloud Storage, from one bucket to another.

Transfer Appliance is an *OFFLINE* secure, high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, answer D is the correct one, while B is incorrect.

**Answer A is incorrect:** gsutil tool is good for programmatic usage by developers and may be useful to copy and move megabytes/gigabytes of data. Not so practical for Terabytes of data. It's also not reliable data transfer technique as it is related to the machine's connectivity with Google Cloud.

**Answer C is incorrect**: In order to migrate to BigQuery, you need to migrate data to Google Storage. This is a useless approach as the main challenge is migrating data from HDFS to Google Storage and BigQuery won't help solving it.

**Source(s):**

Google Cloud Storage Transfer Service: https://cloud.google.com/storage-transfer/docs/

Google Appliance Transfer Service: https://cloud.google.com/transfer-appliance/

Migrate HDFS to Google Storage: https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-data

Ask our Experts

Did you like this **Question?**

## Question 16

Correct

**Domain:** Prepare and use data for analysis

You have a MySQL database installed on a Google Cloud Compute Engine virtual machine. MySQL database is used for a WordPress website. You want to monitor the database instance's application performance and availability by showing CPU, uptime, and number of connections. How can you achieve this?

A. Cloud Monitoring can detect and receive application performance stats from virtual machines automatically.

B. Install Cloud Monitoring MySQL plugin. No extra steps are required.

C. Install Cloud Monitoring MySQL plugin. Create a user for the Cloud Monitoring plugin in MySQL database with the required permissions to run the SHOW_STATUS command.

D. Install Prometheus on a different virtual machine. Allow Prometheus access to MySQL database to collect data. Create a dashboard on Prometheus to monitor database's performance and availability.

## Explanation:

Answer: C.

Description:

Cloud Monitoring is a tool from Google to monitor and manage services, containers, applications, and infrastructure. Cloud Monitoring 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). Cloud Monitoring doesn't require extensive integration and it does not lock developers into using a particular cloud provider.

Cloud Monitoring has a plugin to monitor on-premise and manually installed MySQL databases. After installing the plugin, it will detect the database by:

Searching instance names for "mysql"

Checking for ports opened to *3306* via firewall rules

For MySQL collection, you must add a user with a password to MySQL that can run the "SHOW STATUS" command. This user and password are referred to as STATS_USER and STATS_PASS in the config file installed.

**Answer A is incorrect**: Cloud Monitoring will receive performance and availability stats from Compute Engine virtual machine. However, Cloud Monitoring will not receive any from the MySQL database installed on the virtual machine.

**Answer B is incorrect**: You need to meet the prerequisites mentioned on the plugin page (See source). So, you need to create a user which can run the SHOW STATUS command.

**Answer D is incorrect**: No need to use Prometheus for this scenario. Cloud Monitoring is a good

alternative.

**Source(s):**

Cloud Monitoring – MySQL Plugin: https://cloud.google.com/monitoring/agent/plugins/mysql

Ask our Experts

Did you like this **Question?**

## Question 17

Correct

**Domain:** Prepare and use data for analysis

A company uses BigQuery as its main data warehouse. Data stored in Google Storage is being transformed and enriched using a Dataflow pipeline, to be later loaded into BigQuery. More than 80 different datasets exist in BigQuery with each dataset containing between 20-50 tables, all stored in a single project. Data analysts access BigQuery for their reporting tasks, while data scientists are using BigQuery ML (Machine Learning) by creating forecast models. Since BigQuery is used by a wide range of employees, the CTO wants to control the costs of running queries scanning GBs of data from users who frequently trigger such queries.
How can you achieve this?

**A. Set project-level quotas on BigQuery by setting a fixed size limit to be used monthly.**

**B. Set monthly flat-rate pricing for BigQuery.**

**C. Set user-level custom quotas to all users with access to BigQuery**

**D. Separate datasets to different projects to benefit from monthly free tier.**

## Explanation:

**Answer: C**

Description:

If you have multiple BigQuery projects and users, you can manage costs by requesting a user-level 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.

**Option A is incorrect**: Setting a project-level quota is not the best approach for this scenario because this will not set user limit quotas and when the project reaches the limit set it will disallow all users from running queries. Note that, as stated, all datasets reside in a single cloud project.

**Option B is incorrect**: Flat-rate can be a possible approach. However, BigQuery does not provide flexible flat-rate pricing and the cheapest is (Monthly flat-rate: $2,000 for 100 slots, Annual flat-rate $1700 for 100 slots), which may not be a desirable option for small-medium businesses.

Ref.: https://cloud.google.com/bigquery/pricing#flat-rate_analysis_pricing

**Option D is incorrect**: Separating datasets to different projects will lead to more work from data engineers to maintain access among different projects in case users need to join tables from different datasets together. This solution is possible for testing and development projects, as well as small-scale dataset usage, but for this scenario, setting quotas is more efficient.

**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

Ask our Experts

Did you like this **Question?**

## Question 18

Correct Marked for review

**Domain:** Design Data Processing Systems

You have been using BigTable instance with HDD as storage type. You want to increase the performance of the instance by changing the storage type to SSD. You want to make sure the data will

not be lost. How can you achieve that?

A. Export the data to Cloud Storage in Avro format using Dataflow template and import data into new BigTable instance using Dataflow GCS Avro to BigTable

B. From Google Cloud console UI, you can switch the storage type from HDD to SDD. Data will be moved to new storage type. Instance will be inaccessible by this time until migration is complete.

C. From Google Cloud console UI, you can switch the storage type from HDD to SDD. Data will be moved to new storage type. Instance will be in read-only mode by this time until migration is complete.

D. From the Google Cloud console UI, you can switch the storage type from HDD to SDD. Data will be moved to a new storage type. Instance will be in write-only mode by this time until the migration is complete.

## Explanation:

Answer: A.

You can change cluster IDs only by deleting and recreating the cluster. Also, you cannot change the instance ID or **the instance's storage type**, and you cannot downgrade a production instance to a development instance. To change any of these settings, 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 delete the old instance.

From the description above, the best solution is using Dataflow to migrate data from the old BigTable instance to the new one.

All other answers are incorrect based on the description.

**Source(s):**

BigTable - Modifying a Cloud Bigtable Instance: https://cloud.google.com/bigtable/docs/modifying-instance

Ask our Experts

Did you like this **Question?**

Question 19 Correct

**Domain:** Design Data Processing Systems

You use BigQuery as the main data warehouse. You decided to perform advanced data transformation of the data. You want to use Dataproc with Apache Spark to do the transformation. How can you enable Dataproc's access to data in BigQuery?

**A. Install Dataproc's BigQuery connector on the cluster using initialization actions. Dataproc temporarily loads data from BigQuery to Google Storage. If failed, Dataproc deletes temp files before finishing the job.**

**B. Install Dataproc's BigQuery connector on the cluster using initialization actions. Dataproc temporarily loads data from BigQuery to Google Storage. If failed, you need to manually delete temp files.**

**C. Dataproc cannot directly connect to BigQuery. You should export data from BigQuery to Google Storage first. Dataproc can then read data from Google Storage. You need to manually delete data files after Dataproc is done.**

**D. Dataproc can connect to BigQuery if you set the cluster as owner to the dataset.**

## Explanation:

Answer: B.

You can use a BigQuery connector to enable programmatic read/write access to BigQuery. This is an ideal way to process data that is stored in BigQuery. No command-line access is exposed. The BigQuery connector is a Java library that enables Hadoop to process data from BigQuery using abstracted versions of the Apache Hadoop InputFormat and OutputFormat classes.

You can access BigQuery from Dataproc by installing BigQuery connector to Dataproc cluster using initialization actions. When a Dataproc spark job reads from BigQuery, it writes the BigQuery table's content temporarily to Google Storage using Dataproc cluster's assigned bucket. If the job completes successfully, temporary files are automatically deleted from the cluster. If the job fails, you need to delete temp files manually.

**Answer A is incorrect:** If job fails, you need to delete temp files manually.

**Answer C is incorrect**: Dataproc can read from BigQuery by installing the connector. No need to export data from BigQuery to Google Storage manually.

**Answer D:** Dataproc cannot directly read from BigQuery without installing the connector first.

**Source(s):** https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery

https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions

Ask our Experts

Did you like this **Question?**

Question 20 Correct

**Domain:** Design Data Processing Systems

You have a Dataflow pipeline to run and process a set of data files received from a client, for transformation and loading into a data warehouse. This pipeline should run each morning so that metrics can be ready when stakeholders need the latest stats based on data sent the day before. Select the most efficient Google Cloud Service to achieve the requirement.

A. Cloud Functions

B. Compute Engine

C. Cloud Scheduler

D. Cloud Composer

## Explanation:

**Correct Answer: Option C**

The question is asking to suggest the most efficient name of service that can be used to trigger scheduling a dataflow pipeline.

**Option A: Cloud Functions is an incorrect answer**

Cloud Functions can be written in Node.js, Python, Go, Java, NET, Ruby, and PH programming languages, and are executed in language-specific runtimes. This can be invoked by HTTP functions from standard HTTP requests. These HTTP requests wait for the response and support handling of common HTTP request methods like GET, PUT, POST, and DELETE.

**Option B: Compute Engine is an incorrect answer**

Here we can install Apache Airflow on it and
then orchestrate the Dataflow pipeline but it requires too much manual effort and the user has to maintain and upgrade the Airflow Service on the VM instance.

**Option C: Cloud Scheduler is the correct answer**

This GCP service is used to trigger and schedule different types of jobs on the GCP platform.

Hence this is a correct solution.

**Option D: Cloud Composer is the incorrect answer**

Cloud Composer is a fully managed workflow orchestration service, enabling you to create, schedule, monitor, and manage workflows that span across clouds and on-premises data. But here we have to only trigger a dataflow pipeline, so just for triggering one dataflow pipeline we should not use Cloud Composer as it is a costly service. Hence this is not the correct solution since we have to select the most efficient service

**Reference:**

Cloud Composer: https://cloud.google.com/composer/docs/concepts/overview

Ask our Experts

Did you like this **Question?**

## Question 21

Correct

**Domain:** Ingest and process the data

An online retail company uses BigQuery to store its session logs generated by visitors browsing its website. The marketing team wants to know who are the visitors potential to buy from the online store. This is to focus their marketing efforts on those visitors instead of mass marketing over 100,000 visitors a month. You want to build a logistic regression model to predict visitors willing to buy from the company's products using the logs stored in BigQuery. What should you do?

**A. Use Vertex AI to build a model using TensorFlow allowing permission to the engine's service account to read from BigQuery.**

**B. Use Dataproc to build a logistic regression model using Spark MLLib. Install Dataproc-BigQuery connector for the cluster to access session logs.**

C. Use BigQuery ML by building a model and specifying the model type as logistic regression and the labels.

D. After building a logistic regression model using TensorFlow, export session logs data from BigQuery to Google Storage. Deploy the cluster to Vertex AI and allow its service account to read data from Google Storage.

## Explanation:

**Correct Answer: C**

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 spreadsheets or other applications, and analysts no longer need to wait for limited resources from a data science team.

**Option A is incorrect**: Vertex AI is used to deploy models. It does not help to build the models.

**Option B & D are incorrect**: These approaches are not required if BigQuery ML is adequate for the scenario.

**Source(s):**

Introduction to BigQuery ML:

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

BigQuery ML - Machine Learning using SQL in BigQuery:

https://www.youtube.com/watch?v=BanOYQVl30I

Ask our Experts

Did you like this **Question?**

Question 22 Incorrect Marked for review

**Domain:** Design Data Processing Systems

A FinTech company has over 20TB of data in ORC format stored in on-premise disks. The CTO decides to migrate the current infrastructure to Google Cloud. The current data pipeline cleanses and transforms raw data for reporting and further analysis and prediction using Apache Hive & Spark.

Which of the following Google Cloud products you should use?

**A. Dataproc for processing and Cloud Storage for storing data.**

**B. Dataproc for processing, BigQuery for storage, Dataflow for data pipeline.**

**C. App Engine for processing, Google Storage for storing data, Dataflow for data pipeline.**

**D. Dataproc for processing, Dataproc local HDFS for storage, Dataflow for data pipeline.**

## Explanation:

**Answer: A**

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

Ask our Experts

Did you like this **Question?**

Question 23 Incorrect

**Domain:** Store the data

A weather forecasting facility receives events from its 25,000 sensors every 10 seconds. Those events are stored in Google Storage in JSON format. Events can have different attributes based on purpose, location, and brand. The Data Science team wants to apply SQL queries to this data for further transformation and forecasting analysis.

Which of the following approaches satisfy the Data Scientist's request?

**A. Load the data directly to BigQuery with enabling "auto-detect" option** wrong

**B. Build a dataflow pipeline to read JSON data, and transform it and then load the data to BigQuery**

**C. Create an External table in Bigquery**

**D. Use Dataproc cluster and create Hive external clusters on the data for data scientists to query data**

## Explanation:

**Correct Answer: C**

BigQuery external tables store metadata in BigQuery, but their data resides in an external source. BigQuery supports a variety of external data sources, including Cloud Storage, Google Drive, and Amazon S3.

External tables can be used to query data without having to load it into BigQuery. This can be useful for large datasets that are too expensive or time consuming to load, or for datasets that are constantly changing and need to be queried in real time.

To create an external table, you need to specify the following information:

The name of the table

The schema of the table

The location of the data

Once you have created an external table, you can query it using the same SQL statements that you would use to query a regular BigQuery table.

So, answer C is the correct answer.

The other options are complicated and unnecessary approaches for this scenario.

Since this is streaming data from the sensor devices data keeps on adding from time to time, Hence this is a correct solution.

Reference Link - https://cloud.google.com/bigquery/docs/external-tables

Ask our Experts

Did you like this **Question?**

## Question 24

Correct Marked for review

**Domain:** Store the data

A pharmaceutical factory has over 100,000 different sensors generating JSON-format events every 10 seconds to be collected. You need to gather the event data for sensor & time series analysis.

Which database is best used to collect event data?

**A. Google Storage**

**B. Cloud Spanner**

**C. BigTable**

**D. Datastore**

## Explanation:

Answer: C.

Cloud BigTable is a petabyte-scale, fully managed NoSQL database service for large analytical and operational workloads.

**Answer A is incorrect**: Storing data to Google Storage needs further processing to be eligible for time-series analysis using tools such as Apache Hive or Presto.

**Answer B is incorrect**: Cloud Spanner is a relational database service. It is not recommended for JSON-format data that may have changing structure.

**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):**

BigTable vs. Datastore: https://stackoverflow.com/questions/30085326/google-cloud-bigtable-vs-google-cloud-datastore

Ask our Experts

Did you like this **Question?**

Question 25 Correct Marked for review

**Domain:** Store the data

An e-payment company collects its service payment transaction events from its app installed in nearly 200,000 devices. Those events need to be stored for further time-series analysis and fraud detection. Which of the following approaches is recommended to implement?

A. Use Google Storage to store data. Use Dataproc with Apache Hive to do required queries on data.

B. Use Cloud SQL as a database. Make sure you launch a multi-regional instance for higher peformance.

C. Use BigTable as a database. Use short & wide tables when designing the schema and row key.

D. Use BigTable as a database. Use tall & narrow tables when designing the schema and row key.

## Explanation:

Answer: D.

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:

1. Storing one event per row makes it easier to run queries against your data.
2. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum.

**Answer A is incorrect:** For this scenario, using BigTable is preferred over storing data in Google Storage as further data partitioning and file formatting is required to use Dataproc with Apache Hive.

**Answer B is incorrect:** Cloud SQL is a relational database. Event data might not have a fixed structure. Cloud SQL is not scalable to write thousands of rows in a given second.

**Answer C is incorrect:** Wide & short table schema is not optimal for time-series event data.

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

Question 26 Correct Marked for review

**Domain:** Store the data

A company wants to use NoSQL database for storing its system logs. The system can generate thousands of logs every minute. Those logs are occasionally read by security team in case of possible anomaly behavior or developers for debugging purposes. Due to system's architecture, system logs are not structured and can be different between different components.

Which database do you suggest be used for this scenario?

A. Use BigTable as a database with HDD storage to store system logs.

B. Use BigTable as a database with SSD storage to store system logs.

C. Use Datastore as a database to store system logs.

D. Use Firebase as a database to store system logs.

## Explanation:

Answer: A.

When you create a Cloud Bigtable instance, you choose whether its clusters store data on solid-state drives (SSD) or hard disk drives (HDD).HDD storage is suitable for use cases that meet the following criteria:

You expect to store at least 10 TB of data.

You will not use the data to back a user-facing or latency-sensitive application.

Your workload falls into one of the following categories:

Batch workloads with scans and writes, and no more than occasional random reads of a small number of rows.

Data archival, where you write very large amounts of data and rarely read that data.

From the scenario, system logs are to be stored to BigTable. This data will be only used for occasional debugging and security anomaly detection. So, using HDD storage type for BigTable is the answer.

**Answer B is incorrect:** The scenario does not require SSD storage type.

**Answer C is incorrect**: Datastore is not built for storing and reading huge data volumes as in this scenario. Datastore is deisgned for web applications of small scale.

**Answer D is incorrect:** Firebase is for mobile and web applications. Not a solution for storing big data.

**Source(s):**

Choosing Between SSD and HDD Storage: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

Ask our Experts

Did you like this **Question?**

Question 27 Incorrect

**Domain:** Store the data

You design a pipeline for your company. You want to find a solution to store event data generated in CSV format. The goal is to query data using SQL over time window.

Which storage and schema design should you use recommended by Google?

**A. Use Google Storage to store event data and use BigQuery to create external tables referencing event data and partitioned by time window.**

**B. Use Google Storage to store event data and use DataPrep jobs to partition data by time windows and load partitioned data into Cloud SQL.**

**C. Use BigTable for storage and design tall and narrow tables adding each event as single row.**

**D. Use BigTable for storage and design short and wide tables adding each event as single row.**

## Explanation:

Answer: A.

The scenario states the goal is to query data using SQL. From the available answers, BigQuery is the best service to meet this requirement. Data can be stored in Google Storage, partitioned by time.

BigQuery can read directly from Google Storage creating external tables partitioned by time.

**Answer B is incorrect**: Dataprep does not support SQL queries.

**Answer C & D are incorrect**: BigTable is a NoSQL database by nature. Nonetheless, it supports SQL to query data. However, Bigtable is used if scaling is a critical issue. For this scenario, data is in CSV format and BigQuery is better structured to handle importing CSV data. While Bigtable requires extra prerequisites.

**Source(s):**

BigQuery - Introduction to external data sources: https://cloud.google.com/bigquery/external-data-sources

Ask our Experts

Did you like this **Question?**

Question 28 Correct Marked for review

**Domain:** Store the data

A company specializes in monitoring and distributing data about road traffic for more than 60 cities. Data is used by navigation apps to notify users of traffic congestion on their destination routes and alert them of road accidents. There are thousands of queries running to write new events and read events for analysis and get the latest stats on road traffic.

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

A. Cloud Spanner

B. BigQuery

C. Datastore

D. BigTable

## Explanation:

Answer: D.

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 C 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

Ask our Experts

Did you like this **Question?**

Question 29 Correct Marked for review

**Domain:** Store the data

Analytics team receives data from different data sources stored in Google Storage. The team wants to query the data for required ETL operations which they will fully take care of using SQL. They want your advice on what is the best approach recommended by Google to do it. What would you suggest?

A. Batch load the data from Google Storage into BigQuery using its batch API, run cleansing and transformation queries on data and insert the transformed rows to another BigQuery table.

B. Batch load the data from Google Storage into BigQuery using its batch API, run cleansing and transformation queries on data and export the data to Google Storage. Launch Dataproc cluster and use Hive to query the transformed data.

C. Create external tables on data using BigQuery, apply the cleansing and transformation queries on data then load the output to an internal BigQuery table for reporting and visualization.

D. Create external tables on data using BigQuery, apply the cleansing and transformation queries on data then load the output to BigTable for reporting and visualization.

## Explanation:

Answer: C.

An external data source (also known as a federated data source) is a data source that you can query directly even though the data is not stored in BigQuery. Instead of loading or streaming the data, you create a table that references the external data source.

Querying an external data source using a temporary table is useful for one-time, ad-hoc queries over external data, or for extract, transform, and load (ETL) processes.

In summary, using external tables in BigQuery is useful for such cases:

Perform ETL operations on data.

Frequently changed data.

Data is being ingested periodically.

**Answer C is the correct** answer based on above explanation and using BigQuery for reporting and visualization is a better approach.

**Answer D is incorrect** because BigTable isn't a practical (and cheap) approach to report and visualize data.

**Answers A & B are incorrect**: Based on Google's best practices, using external tables for ETL is better than loading data to BigQuery.

**Source(s):**

BigQuery external tables: https://cloud.google.com/bigquery/external-data-sources

BigQuery – Define external tables: https://cloud.google.com/bigquery/external-table-definition

Ask our Experts

Did you like this **Question?**

## Question 30

Correct

**Domain:** Store the data

You want to import data into BigQuery. You select a CSV file to upload and enable "automatically detect" for BigQuery to predict the schema of the table. When you checked the data, you found a skew in data and it did not match with the file you uploaded. What is the reason for this?

A. Field separator/delimiter is not ','

B. The file is not UTF-8 encoded, so you have to provide its encoding

C. File is corrupted

D. All of the Above

## Explanation:

**Correct Answer: B**

BigQuery supports UTF-8 encoding for both nested or repeated and flat data. BigQuery supports ISO-8859-1 encoding for flat data only for CSV files.

By default, the BigQuery service expects all source data to be UTF-8 encoded. Optionally, if you have CV files with data encoded in ISO-8859-1 format, you should explicitly specify the encoding when you import your data so that BigQuery can properly convert your data to UTF-8 during the import process.

**Option A is incorrect:** If the delimiter is not ',' then by querying the data you will see that all the records are inserted into the table but all the fields are in one column. Since here you can query the data and find skewness. This can't be the case

**Option C is incorrect:** If the file is corrupted then you won't be able to load it and create the table

**Option D is incorrect:** Since options A and C can't be the reason for the issue

**References:**

BigQuery - Loading data: Ittps://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv

https://cloud.google.com/bigquery/docs/loading-data

Ask our Experts

Did you like this **Question?**

Question 31 Correct

**Domain:** Store the data

You maintain a web service used by over 12,000 different clients. The service uses MySQL as its main data storage hosted on Google Cloud SQL. Since the service is critical and should be highly available at all times, this includes the MySQL database, considerations should be met so the service does not face any outage due to network connectivity or database having temporary outage. The service was not developed by your company so code refactoring is not currently possible. How can you ensure high availability for the web service's database?

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

**B. Enable high availability on Cloud SQL instance by creating a read replica.**

**C. Migrate the database to BigQuery and use it as the main data storage instead.**

**D. Enable high availability on Cloud SQL instance by creating a fail-over replica.**

## Explanation:

Answer: D.

Description:

The failover replica in Cloud SQL is configured with the same database flags, users (including root) and passwords, authorized applications and networks, and databases as the primary instance. If an High-availability-configured instance becomes unresponsive, Cloud SQL automatically switches to serving data from the failover replica. This is called a *failover*.

**Answer A is incorrect**: Scaling up Cloud SQL instance does not ensure high availability.

**Answer B is incorrect**: Read replicas offer a copy of master (primary) instance to read from. In case the master instance is down, read replica will not be updated.

**Answer C is incorrect**: Migrating database to a different solution is not applicable based on the scenario because the service cannot be refactored.

**Source(s):**

Cloud SQL - MySQL High Availability:

https://cloud.google.com/sql/docs/mysql/high-availability

Cloud SQL - MySQL Replication Options:

https://cloud.google.com/sql/docs/mysql/replication

Ask our Experts

Did you like this **Question?**

Question 32 Correct

**Domain:** Store the data

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)

**A. Use development instance for testing.**

**B. Run a heavy pre-test for several minutes before the test starts.**

**C. Scale up the instance just before the test starts.**

**D. Use at least 400GB of data.**

**E. Do not scale up the instance just before the test starts.**

**F. Test should take no longer than 10 minutes.**

## Explanation:

**Answers: B, D & E**

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:

Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.

Use at least 400 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 400

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.

Stay below the recommended storage utilization per node.

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.

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

Ask our Experts

Did you like this **Question?**

Question 33 Correct

**Domain:** Store the data

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 high number of cells.**

**B. The rows in the tables have large data size.**

**C. Test data size is over 300GB.**

**D. The instance uses SSD storage type.**

**E. Heavy pre-test was done before the testing started.**

**F. The instance doesn't have enough nodes.**

## Explanation:

Answers: A, B & F

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

Ask our Experts

Did you like this **Question?**

Question 34 Correct Marked for review

**Domain:** Ingest and process the data

A fast-food chain restaurant wants to detect the different meal photos its customers upload to the different social media platforms tagged with their name in order to know what meals customers like and share the most for better quality analysis.

It asks your advice on developing such solution for them. However, they want it to be available and in production the soonest possible because they expect a high activity on their social media pages by the next public holiday which is coming in 2 weeks and marketing team finds it a great opportunity to receive feedback based on what customers say online.

What is the best approach for this?

**A. Use AutoML Vision to build and train the model by using all the training photos you collected from food-chain's social media pages for better results.**

**B. Use AutoML Vision to build and train the model by using 50-70% of training photos you collected from food-chain's social media pages while the rest of training set is to test and tune the model.**

**C. Use Dataproc to build the model using SparkML. Use 50-70% of training photos you collected to train the model and the rest to test and tune the model. Deploy the model using Vertex AI.**

**D. Use Vertex AI with TensorFlow to build the model. Use all training photos you collected to train the model. Deploy the model using Vertex AI.**

## Explanation:

**Correct Answer: B**

Since you have a very short time to build, train and deploy the model, building your own model can be time-consuming and not in your favor. Google provides a great ML service called AutoML to quickly build models for you. AutoML Vision is one of its products which you can start with a training set of as little as a dozen photo samples and AutoML takes care of the rest.

**Option A is incorrect:** AutoML Vision is the right choice. However, training the model with a whole training set is not the right approach in Machine Learning because you ought to test the model before considering it accurate enough for production. Usually, the training set is split into 70-30% sets, the first

for training while the second is for testing and tuning the model's parameters.

**Option C is incorrect:** Using any approach other than AutoML can be time-consuming and with such tight deadlines, it's not the best approach.

**Option D is incorrect:** Using this approach can also be time-consuming and using the whole training set for training is not a best practice as explained before.

Thus, the best approach for this scenario is Answer B.

**Source(s):**

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

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

Ask our Experts

Did you like this **Question?**

Question 35 Correct

**Domain:** Ingest and process the data

A video-on-demand company wants to generate subtitles for its content on the web. They have over 20,000 hours of content to be subtitled and their current subtitle team cannot catch up with the every-growing video hours the content team keep adding to the website library. They want a solution to automate this as man power can be expensive and may take long time.

Which service of the following can greatly help the automation of video subtitles?

A. Cloud Natural Language.

B. Cloud Speech-to-Text.

C. AutoML Vision API.

D. Vertex AI

## Explanation:

Answer: B.

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

**Answer C 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.

**Answer D 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.

**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/

Ask our Experts

Did you like this **Question?**

## Question 36

Correct

**Domain:** Ingest and process the data

A financial services firm providing products such as credit cards and bank loans receives thousands of online applications from clients applying for their products. Because it takes a lot of effort to scan and check all applications if they meet the minimum requirements for the products they are applying for, they want to build a machine learning model takes application fields like annual income, marital status, date of birth, occupation and other attributes as input and finds out if the applicant is qualified for the product the client applies for.

What is the machine learning technique will help build such model?

A. Regression

B. Classification

C. Clustering

D. Reinforcement learning

## Explanation:

Answer: B.

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.

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.

Reinforcement learning is a technique which a machine takes actions without training sets to reach the highest rewards possible. The agent learns from trial and decides what to do to perform a given task without supervision. The task punishes the agent for a wrong action and rewards it for achieving the task. Examples of reinforcement learning is asking an agent to play a maze game to reach the exit with traps along the way or making an agent play a video game and win a racing game.

From the explanation above, we can see the scenario problem which finding if a client is qualified for a product is a classification problem. So, answer B is correct.

Ask our Experts

Did you like this **Question?**

Question 37 Incorrect

**Domain:** Ingest and process the data

You have built a machine learning model to classify if a customer would buy a certain product when recommended by the company's website. You trained the model with a sample set. Upon testing the model, you found out only 28% of the testing sets are actually true positives and the model isn't very accurate. You figured out the model is over-fitted. How would you solve this?

**A. Increase training data, increase feature parameters & increase L1 regularization.**

**B. Decrease training data, decrease feature parameters & increase L1 regularization.**

**C. Increase training data, decrease feature parameters & increase L1 regularization.**

**D. Increase training data, decrease feature parameters & decrease L1 regularization.**

## Explanation:

Answer: C.

Overfitting happens when a model performs 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.

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.

Increase regularization.

Hence, answer C is correct.

**Source(s):**

Building a serverless Machine learning model: https://cloud.google.com/solutions/building-a-serverless-ml-model

Ask our Experts

Did you like this **Question?**

Question 38 Correct

**Domain:** Ingest and process the data

Data scientists are testing a TensorFlow model on Google Cloud using four NVIDIA Tesla P100 GPUs to test a TensorFlow model. After experimenting with several use cases, they decide to scale up by using a different machine type for testing. As a data engineer, you are responsible of assisting with choosing the right machine type to reach a better model performance. What should you do?

**A. Use TPU machine type for testing the TensorFlow on.**

**B. Scale up machine type by using NVIDIA Tesla V100 GPUs.**

**C. Use 8 NVIDIA Tesla K80 GPUs instead of the current 4 P100 GPUs.**

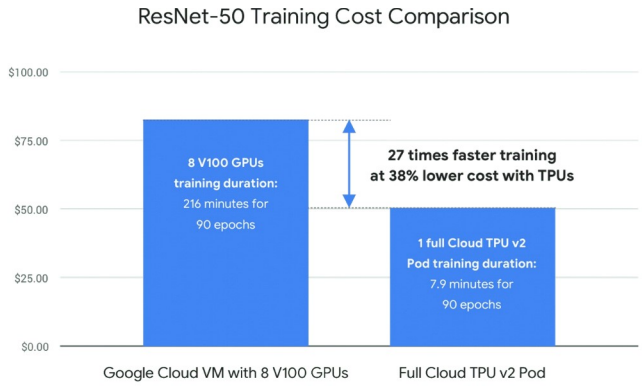
**D. Increase number of Tesla P100 GPUs used until test results return satisfactory performance.**

## Explanation:

Answer: A.

Description:

Google built the Tensor Processing Unit (TPU) in order to make it possible for data scientists to achieve business and research breakthroughs ranging from network security to medical diagnoses. Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail.




As for cost, below is a benchmark for cost comparison between TPU & GPU types:

So, for this scenario, using TPU machine type is the recommended type to build Tensorflow models on.

**Source(s):**

TPU Machine Type:

https://cloud.google.com/tpu/

GPU Machine Type:

https://cloud.google.com/gpu/

https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep- learning

Using GPUs for training models in the cloud: https://cloud.google.com/ml-engine/docs/tensorflow/using-gpus

Ask our Experts

Did you like this **Question?**

## Question 39

Correct Marked for review

**Domain:** Ingest and process the data

You are building a machine learning model to solve a classification problem. The model should identify if a patient has a tumor. Based on statistics, only 1.4% of scanned patients are identified positive for tumor.

You want to make sure the machine learning model is able to correctly identify patients with tumor. What is the technique to examine the effectiveness of the model?

A. Gradient Descent

B. Precision

C. Recall

D. Dropout

## Explanation:

Answer: C.

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.

Dropout is a regularization method to remove random selection of fixed number of units in a neural network layer. More units dropped out, the stronger the regularization.

From the description, answers A & D are unrelated so they are incorrect.

Since very few cases are positively diagnosed with tumor, recall formula should be used to calculate the accuracy of the model. So, answer C is the correct answer.

**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

Dropout Regularization: https://developers.google.com/machine-learning/glossary/

Ask our Experts

Did you like this **Question?**

Question 40 Correct

**Domain:** Ingest and process the data

You are building a machine learning classification model using TensorFlow. You trained the model by using 70% of the total set available for training, validation and testing. After testing the model, AUC returned from the test results was 0.68. The main issue here is due to overfitting. You want to increase the AUC for better accuracy of results. What should you do?

**A. Increase regularization.**

**B. Reduce samples used for training.**

**C. Reduce regularization.**

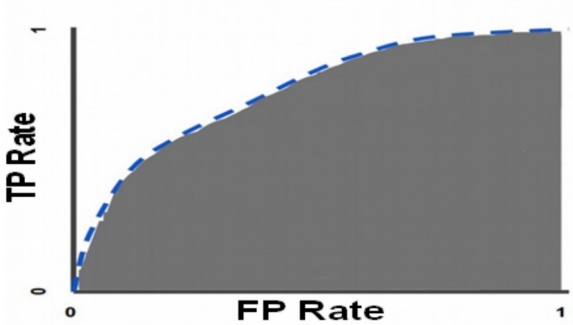
**D. Increase feature parameters.**

## Explanation:

Answer: A.

Description:

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1):




AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

The problem in this scenario is due to overfitting. To solve the overfitting problem, you need to:

Increase the training set.

Decrease features parameters.

Increase regularization.

**Source(s):**

Classification: ROC Curve and AUC:

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Ask our Experts

Did you like this **Question?**

Question 41 Correct

**Domain:** Ingest and process the data

You are training a Tensorflow deep neural network model. The model should recognize different type of cars and return the brand and type of the car from the image input. While training, you decided to perform hyper-parameter tuning to optimize the model.

Which of the variables are used for hyperparameter tuning? (Choose 2)

**A. Number of nodes in hidden layers**

**B. Number of features**

**C. Number of hidden layers**

**D. Weight values**

## Explanation:

Answer: A & C.

hyperparameters are the variables govern the training process itself. For example, part of setting up a

deep neural network is deciding **how many hidden layers** of nodes to use between the input layer and the output layer, and **how many nodes each layer** should use. These variables are not directly related to the training data. They are configuration variables. Note that parameters change during a training job, while hyperparameters are usually constant during a job.

From the description, the right answers are A & C.

**Answer B is incorrect:** Feature numbers are set by feature engineering, not hyperparameter tuning.

**Answer D is incorrect:** Weight values are set when training the model.

**Source(s):**

Hyperparameter Tuning: https://cloud.google.com/ml-engine/docs/tensorflow/hyperparameter-tuning- overview

Ask our Experts

Did you like this **Question?**

Question 42 Correct

**Domain:** Ingest and process the data

The data scientists at your company have built a machine learning neural network model using TensorFlow. After several tests on the model, the team decides the model is ready to be deployed for production use. Which of the following services would you use to host the model to Google Cloud?

A. Google Kubernetes Engine

B. Google ML Deep Learning VM

C. Google Container Registry

D. Vertex AI

## Explanation:

**Answer: D**

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.

**Option A is incorrect:** GKE is a service to deploy and scale Docker containers in the cloud. You need to build the docker image for your model if you want to use it, which is not recommended for this scenario.

**Option B is incorrect:** Google ML Deep Learning VM is a service that offers pre-configured virtual machines for deep learning applications. It is not used to deploy ML models to production.

**Option C is incorrect:** Google Container Registry is a service to store, manage, and secure your Docker container images. It does not for deploying machine learning models.

Cloud Machine Learning Engine is a managed service that lets developers and data scientists build and run superior machine learning models in production. Cloud ML Engine offers training and prediction services, which can be used together or individually.

**Option D is correct**: Google Vertex AI Auto ML is the service to use to deploy your machine learning models.

**Source(s):**

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

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

Google ML Deep Learning VM: https://cloud.google.com/deep-learning-vm/

Ask our Experts

Did you like this **Question?**

## Question 43

Correct

**Domain:** Ingest and process the data

The data scientists at your company have successfully built a deep neural network machine learning model to detect car plate numbers entering and exiting a parking lot of a high-rise condominium. The model was built using Tensorflow and the model was exported as SavedModel. As a data engineer, you are assigned to deploy their model. The company is using Google Cloud for its project.

Which approach is best for deploying the detection model?

A. Upload SavedModel object to Google Storage. Use Dataproc with Spark ML to use the model by accessing it using Google Storage Connector.

B. Deploy the model to Google Kubernetes Engine after wrapping SavedModel as docker image and uploading it to Google Container Registry.

C. Deploy the model to ML module in GCP after asking the data science team to convert the model to binary format using PyTorch.

D. Deploy the model exported as SavedModel directly to Vertex AI in GCP.

## Explanation:

**Answer: D**

You can export your SavedModel directly to the Vertex AI. There is no need to take any other approach as it is more cumbersome.

**Answer A is incorrect**: This approach is too complicated and not necessary for this scenario.

**Answer B is incorrect:** There is no need to go through converting SavedModel to a docker image and use GKE. It's not the recommended approach by Google anyway.

**Answer C is incorrect** because there is no need to convert the model to any other format. Cloud ML Engine supports Tensorflow and SavedModel format.

**Source(s):**

Google ML Engine: https://cloud.google.com/ml-engine/docs/tensorflow/prediction-overview

Deploying ML models to Google Cloud:

https://cloud.google.com/ml-engine/docs/tensorflow/deploying-models

Ask our Experts

Did you like this **Question?**

Question 44 Correct Marked for review

**Domain:** Ingest and process the data

You are asked by the data science team to deploy their Tensorflow deep neural network model to the cloud. You choose the ML model in the GCP cloud. Upon checking the available tiers, you suggested choosing a custom tier by launching a cluster with custom specifications to cover the requirements provided to deploy the model.

Which of the following specifications you can set for the ML Model cluster? (Choose TWO)

A. workerCount

B. masterCount

C. masterCPU

☐ D. workerType right

## Explanation:

**Answers: A and D**

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.master type 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 workerCount and workerType.

**References:**

Specifying Machine Types or Scale Tiers:

> https://cloud.google.com/ml-engine/docs/tensorflow/machine-types
>
> https://cloud.google.com/ai-platform/docs/technical-overview

Ask our Experts

Did you like this **Question?**

## Question 45

Correct Marked for review

**Domain:** Prepare and use data for analysis

You launched a Dataproc cluster to perform some Apache Spark jobs. You are looking for a method to securely transfer web traffic data between your machine's web browser and Dataproc cluster.

How can you achieve this?

**A. FTP connection**

B. SSH tunnel

C. VPN connection

D. Incognito mode

## Explanation:

Answer: B.

Some of the core open source components included with Google Cloud Dataproc clusters, such as Apache Hadoop and Apache Spark, provide web interfaces. These interfaces can be used to manage and monitor cluster resources and facilities, such as the YARN resource manager, the Hadoop Distributed File System (HDFS), MapReduce, and Spark. Other components or applications that you install on your cluster may also provide web interfaces.

It is recommended to create an SSH tunnel for a secure connection between your web browser and Dataproc's master node. SSH tunnel supports traffic proxying using the SOCKS protocol. To configure your browser to use the proxy, start a new browser session with proxy server parameters.

**Source(s):**

Dataproc – Cluster Web Interfaces: https://cloud.google.com/dataproc/docs/concepts/accessing/cluster- web-interfaces#connecting_to_the_web_interfaces

Ask our Experts

Did you like this **Question?**

## Question 46

Correct

**Domain:** Prepare and use data for analysis

There is a plan by the data team to migrate the data warehouse to BigQuery. After the migration is done, you are tasked to assign each user the right role to access datasets in BigQuery. You have the following teams need to access the data warehouse:

Data analysts: They need read/write access to data. They should not create or delete datasets.

Data engineers: They are admins in the data warehouse. They need full privileges on data

sets.

Dev team: They need read access only to the datasets. They can list the project's data sets and tables.

How would you assign the roles to each team?

**A. Assign admin role to data engineer group. Assign dataOwner role to data analyst group. Assign dataViewer role to dev team group.**

**B. Assign dataOwner role to data engineer group. Assign dataEditor role to data analyst group. Assign user role to dev team group.**

**C. Assign admin role to data engineer group. Assign dataEditor role to data analyst group. Assign dataViewer role to dev team group.**

**D. Assign dataOwner role to data engineer group. Assign dataEditor role to data analyst group. Assign dataViewer role to dev team group.**

## Explanation:

Answer: C.

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 list, we can assign each group the right role:

Data Analysts: They need read & write access to tables without permissions to create & delete data sets. Thus, editor role is assigned to this group.

Data engineers: It's stated they will be admins on BigQuery. Admin role should be assigned to this gorup.

Dev team: They need read access to data and be able to list tables and datasets. Viewer role should be assigned to this group.

**Answer A is incorrect:** Owner role is too broad for data analysts.

**Answer B is incorrect:** Data engineers need admin role. User role doesn't allow dev team to get tables data.

**Answer D is incorrect:** As B, data engineers need admin role.

**Source(s):**

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

Ask our Experts

Did you like this **Question?**

Question 47 Correct

**Domain:** Prepare and use data for analysis

Your company uses BigQuery as the main data warehouse. A data warehouse is divided into several datasets based on data origin and profile. Data analysts want to access certain data that resides in a dataset considered sensitive and should not be openly available to all users. The security team allows only certain tables with limited columns for data analysts to read from.

Which of the following actions will you take?

A. Create a new dataset in BigQuery. Create authorized views on tables data analysts want to read from. Grant viewer role to data analysts on a new dataset.

B. Create authorized views on tables, the data analysts want to read from on the same dataset tables reside in. Grant viewer role to the Data analysts team on the views.

C. Grant data analysts viewer the role of these specific tables by specifying what columns to be read from.

D. Create a new dataset in BigQuery. Grant viewer role to data analysts on the new dataset. Copy the tables from the current dataset to the new one with only columns allowed.

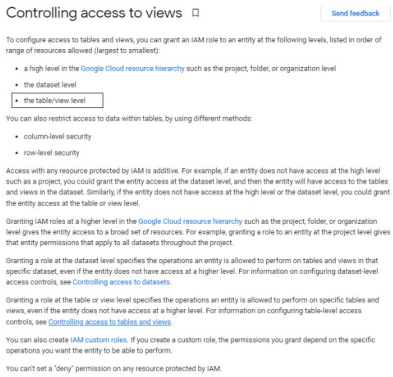
## Explanation:

**Correct Answer - A**

**Option A is correct:** Creating a new dataset is the correct solution. Authorized views should be created in a different dataset from the source data. That way, data owners can give users access to the authorized view without simultaneously granting access to the underlying data. The source data dataset and authorized view dataset must be in the same regional location.

**Option C is incorrect: To grant access to a column of a table, one needs to** update the table schema to set a policy tag on a column. In this option, there is no policy. Hence this is also the wrong answer.

**Option D is incorrect:** Creating a new dataset is the wrong approach.



Controlling access to views

Send feedback

To configure access to tables and views, you can grant an IAM role to an entity at the following levels, listed in order of range of resources allowed (largest to smallest):

- a high level in the Google Cloud resource hierarchy such as the project, folder, or organization level
- the dataset level
- the table/view level

You can also restrict access to data within tables, by using different methods:

- column-level security
- row-level security

Access with any resource protected by IAM is additive. For example, if an entity does not have access at the high level such as a project, you could grant the entity access at the dataset level, and then the entity will have access to the tables and views in the dataset. Similarly, if the entity does not have access at the high level or the dataset level, you could grant the entity access at the table or view level.

Granting IAM roles at a higher level in the Google Cloud resource hierarchy such as the project, folder, or organization level gives the entity access to a broad set of resources. For example, granting a role to an entity at the project level gives that entity permissions that apply to all datasets throughout the project.

Granting a role at the dataset level specifies the operations an entity is allowed to perform on tables and views in that specific dataset, even if the entity does not have access at a higher level. For information on configuring dataset-level access controls, see Controlling access to datasets.

Granting a role at the table or view level specifies the operations an entity is allowed to perform on specific tables and views, even if the entity does not have access at a higher level. For information on configuring table-level access controls, see Controlling access to tables and views.

You can also create IAM custom roles. If you create a custom role, the permissions you grant depend on the specific operations you want the entity to be able to perform.

You can't set a "deny" permission on any resource protected by IAM.

**Reference:**

https://cloud.google.com/bigquery/docs/share-access-views

Ask our Experts

Did you like this **Question?**

## Question 48

Correct

**Domain:** Prepare and use data for analysis

You are writing highly-confidential data related to customers' personally identifiable information (PII). The security team is concerned about how secure the network connection between the instances and Google Storage buckets. Security team proposes to use encryption keys generated by security team.

Those keys will be rotated every 30 days for more security.

As a data engineer, what should you do to satisfy security team's requirement?

A. Upload encryption key provided by security team to Cloud Key Management Service (KMS) and use the key to encrypt data when writing to Google Storage.

B. Create symmetric keys using Cloud Key Management Service (KMS) and use those to encrypt data when writing to Google Storage. Create new keys every 30 days.

C. Create asymmetric keys using Cloud Key Management Service (KMS) and use those to encrypt data when writing to Google Storage. Create new keys every 30 days.

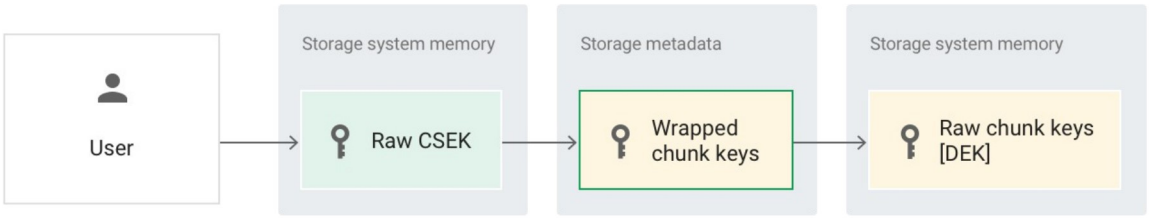
D. Supply the encryption key provided by security team and reference it as part of the API service calls to encrypt data in Cloud Storage. right

## Explanation:

Answer: D.

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 A is incorrect:** Security team does not recommend storing encryption key in the cloud.

**Answer B & C are 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/

Ask our Experts

Did you like this **Question?**

Question 49 Correct

**Domain:** Prepare and use data for analysis

You work as a data engineer in an organization with a large amount of data. They use Google Big Table to store their web service's activity logs for faster retrieval and update. What will happen if the BigTable node fails?

A. Data will be lost

B. Recover data from Cloud Storage when the node comes back online

C. Data will not be lost

D. Data will be transferred automatically to new node

## Explanation:

**Correct Answer: C**

**Option C is CORRECT** because data is never stored in Big Tables nodes.

Recovery of the Bigtable node from the failure is very fast because metadata information only needs to be replicated to the new node.

**Option A is incorrect** because storage and compute are separate so data will not be lost.

**Option B is incorrect** because Big Table does not store its data in Cloud storage.

For more information on the **Cloud BigTable**, please visit the below URL:

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

Ask our Experts

Did you like this **Question?**

## Question 50

Correct

**Domain:** Store the data

You are working as a Data Engineer and your company has asked you to use Big Query.

The requirement is to fetch the data from a table named "customer" having 1000 columns, and you can skip the following columns as those are not needed for the solution if needed -

A11, B11, B6, C7, D9, E1, P1, Q2, R2, Z5

You were asked to consider the solution to be cost-effective. (Single Option)

**A. Use Select * FROM `customer`**

**B. Use Select * EXCLUDE (A11, B11, B6 , C7, D9, E1, P1,Q2, R2,Z5) FROM `customer`**

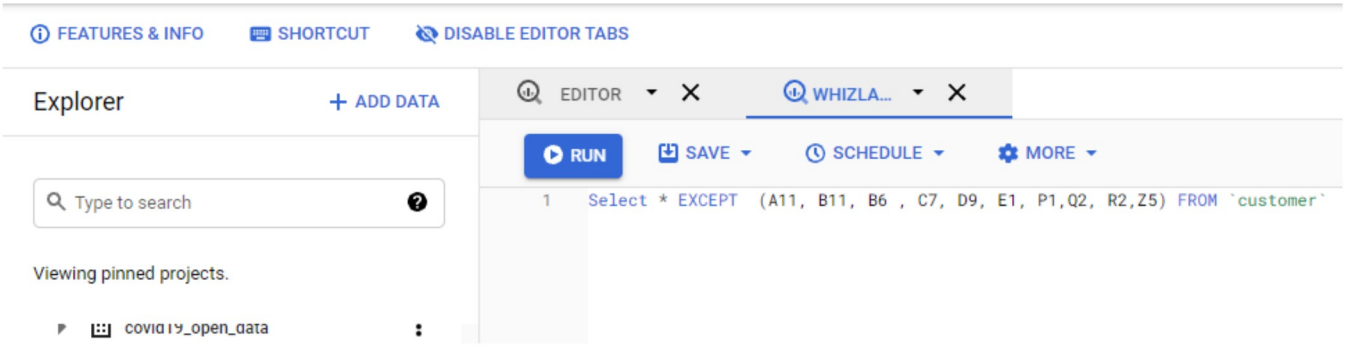
**C. Use Select * EXCEPT (A11, B11, B6 , C7, D9, E1, P1,Q2, R2,Z5) FROM `customer`**

**D. Use Select A11, B11, B6 , C7, D9, E1, P1,Q2, R2,Z5 FROM `customer`**

## Explanation:

**Correct Answer: C**

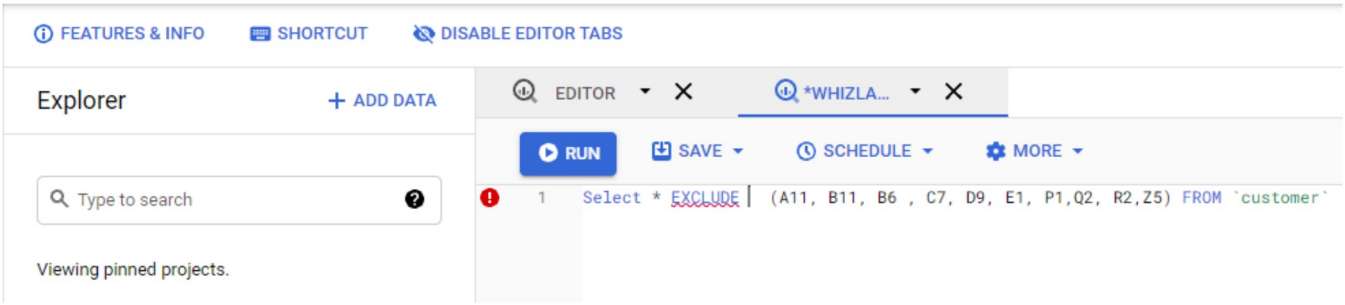
**Option C is CORRECT** because we can use EXCEPT to skip or remove the unwanted columns, this is useful as well in terms of cost-saving.




Note: This will not delete your columns, this is only for the SELECT to view the table.

**Option A is incorrect** because this query will fetch all the columns and this is not a cost-effective solution.

**Option B is incorrect** because EXCLUDE is not a valid option.




**Option D is incorrect** because this query will select only (A11, B11, B6, C7, D9, E1, P1, Q2, R2, Z5) and in the requirement, we need to skip these columns.

For more information on the **Big Query- SQL EXCEPT**, please visit the below URL:

https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax#select_except

Ask our Experts

Did you like this **Question?**

Finish Review

Java Certifications

Docker Certified Associate

**Legal**

Privacy Policy
Terms of Use
EULA
Refund Policy
Programs Guarantee

**Support**

Contact Us
Discussions
FAQs

Need help? Please or +91 6364678444