Level: Intermediate

Google Cloud Certified Professional Data Engineer

← Back to the **Course**

Final Test

Completed on **Sat, 06 Jul 2024**

1st	51/55	92.73%
Attempt	Marks Obtained	Your Score

0h 40m 7s	PASS
Time Taken	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	15	13	2	0
2	Store the data	10	9	1	0
3	Ingest and process the data	13	12	1	0
4	Prepare and use data for analysis	12	12	0	0
5	Maintaining and Automating Data Workloads	5	5	0	0
Total	All Domains	55	51	4	0

Review the Answers

Filter By All Questions

Question 1

Correct

Domain: Design Data Processing Systems

A regional auto dealership is migrating its business applications to Google Cloud. The CTO of this company asked their data engineer to find the possible ways you can ingest data into BigQuery?

A. Batch Ingestion & Streaming Ingestion

B. Data Transfer Service

C. Query Materialization

D. Partner Integrations

E. All of the above right

Explanation:

Correct Answer: E

Option E is CORRECT because these all are the possible ways to load data into BigQuery. Lets understand these one by one -

1. Batch Ingestion - This involves ingesting large, bounded datasets that don't have to be processed in real-time.

To implement batch ingestion, one can use Google Cloud Storage, Cloud Dataflow, Cloud Dataproc, Cloud Data Fusion etc.

1. Stream Ingestion - This involves ingesting large, unbounded data that is processed in real-time.

To implement Stream Ingestion, one can use Apache Kafka to BQ connector, Cloud Pub/Sub, Cloud Dataflow, Cloud Dataproc etc.

1. Data Transfer Service (DTS) - This is a fully managed service to load data from external cloud storage providers such as Amazon S3, Google SaaS applications such as Google Ads, and transferring data from data warehouse technologies such as Teradata etc.
2. Partner Integrations - These are the data integration alternatives from Google Cloud Partners.

This includes, Confluent, Informatica, snapLogic, Talend and many more.

1. Query Materialization - This is the best way to simplify extract, transform and load patterns in BigQuery. Using federated queries in BigQuery, one can persist their analysis results in Big Query to derive any insights.

Please find the image attached below having details about all the possible ways to ingest data into BigQuery.

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

https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion

Question 2

Correct

Domain: Ingest and process the data

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

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

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

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

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

Explanation:

Answer: D.

Description:

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

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

AutoML Vision allows you to sort the images by how "confused" the model is, by the true label and its predicted label. Look through these images and make sure they're labeled correctly.

Consider adding more images to any labels with low quality.

You may need to add different types of images (e.g. wider angle, higher or lower resolution, different points of view).

Consider removing labels altogether if you don't have enough training images.

Remember that machines can't read your label name; it's just a random string of letters to them. If you have one label that says "door" and another that says "door_with_knob" the machine has no way of figuring out the nuance other than the images you provide it.

Augment your data with more examples of true positives and negatives. Especially important examples are the ones that are close to the decision boundary (i.e. likely to produce confusion, but still correctly labeled).

Specify your own TRAIN, TEST, VALIDATION split. The tool randomly assigns images, but near-duplicates may end up in TRAIN and VALIDATION which could lead to overfitting and then poor performance on the TEST set.

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

Source(s):

Cloud AutoML Vision – Evaluating Models:

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

Question 3 Correct

Domain: Maintaining and Automating Data Workloads

Your team is tasked with monitoring and troubleshooting data processes in a Google Cloud environment. Which tool or service should you utilize to ensure effective monitoring of planned usage and identify potential issues proactively?

- A. Leverage Cloud Monitoring to monitor resource utilization and performance metrics in real-time right
- B. Utilize Stackdriver Debugger to debug and troubleshoot data processing errors in real-time
- C. Implement Cloud Logging to capture and analyze logs from various data processing services
- D. Employ Cloud Trace to trace and profile data processing workflows for performance optimization

Explanation:

Correct Answer: A

Option A is CORRECT as Cloud Monitoring allows for monitoring resource utilization and performance metrics in real-time, enabling effective monitoring of planned usage and proactive identification of potential issues.

Option B is incorrect because while Stackdriver Debugger can debug and troubleshoot errors in real-time, it may not provide comprehensive monitoring of planned usage and performance metrics.

Option C is incorrect because while Cloud Logging captures and analyzes logs, it may not offer real-time monitoring of resource utilization and performance metrics.

Option D is incorrect because while Cloud Trace traces and profiles workflows, it may not provide monitoring of planned usage and proactive issue identification.

Reference:

https://cloud.google.com/monitoring

Question 4 Correct

Domain: Prepare and use data for analysis

Your organization operates an e-commerce platform that receives a continuous stream of real-time transaction data from various sources. This data is ingested into Google BigQuery for analysis by data

analysts. Recently, the query performance for analyzing this data has degraded, leading to longer processing times. What strategic step should you take to optimize query performance in BigQuery without compromising real-time data ingestion?

A. Utilize sharding techniques to distribute the data across multiple tables based on transaction types

B. Implement partitioning in BigQuery based on transaction timestamps to organize the data into manageable segments right

C. Integrate caching mechanisms to store and retrieve frequently accessed query results

D. Reconfigure the schema of the transaction data to denormalize tables and reduce join operations during queries

Explanation:

Correct Answer: B

Option B is CORRECT as implementing partitioning in BigQuery based on transaction timestamps allows for efficient query processing by organizing the data into partitions based on time intervals. This approach enhances query performance without affecting real-time data ingestion processes.

Option A is incorrect because while sharding techniques may help distribute data, they may not directly address the query performance issue related to longer processing times.

Option C is incorrect because while caching mechanisms can improve performance for recurring queries, they may not effectively optimize query processing for real-time data analysis.

Option D is incorrect because denormalizing tables may reduce join operations but may not specifically target the query performance degradation observed in analyzing real-time transaction data.

Reference:

https://cloud.google.com/bigquery/docs/query-performance

Question 5

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 right

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/

Question 6

Correct

Domain: Ingest and process the data

A Data Engineering team has developed a data pipeline using Google Cloud Dataflow for real-time analytics. To ensure continuous operation, the team needs a robust CI/CD (Continuous Integration and Continuous Deployment) solution. What Google Cloud service should be integrated into the workflow for automated testing, building, and deployment?

A. Google Cloud Composer for its visual workflow orchestration capabilities

B. Google Cloud Build for its fully managed CI/CD platform, supporting automated testing, building, and deployment right

C. Google Cloud Dataprep for its visual data preparation and CI/CD automation features

D. Google Cloud Data Fusion for its data integration and CI/CD automation capabilities

Explanation:

Correct Answer: B

Option B is CORRECT for integrating CI/CD capabilities seamlessly into the Google Cloud Dataflow workflow. Google Cloud Build is a fully managed CI/CD platform that supports automated testing, building, and deployment.

Option A is incorrect. Google Cloud Composer is an orchestration service but may not provide the same level of CI/CD automation features as Google Cloud Build.

Option C is incorrect. Google Cloud Dataprep is more focused on visual data preparation and does not offer the comprehensive CI/CD capabilities needed for a Dataflow workflow.

Option D is incorrect. Google Cloud Data Fusion is designed for data integration and may not provide the same level of CI/CD automation as Google Cloud Build.

Reference:

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

Question 7 Correct

Domain: Design Data Processing Systems

A company uses BigTable to store their web service's activity logs. Data is later aggregated and enriched by a Dataflow pipeline that stores enriched data to BigQuery for analysis and visualization by both business and security analysis.

It was noticed that the performance of BigTable cluster is not as per the expectations when it was considered for activity log storage. The cluster uses HDD which is a possible reason for lower performance. You decided to use SSD storage for BigTable. How would you achieve this?

A. You can change storage type on the fly from HDD to SDD. Data will be moved to a new storage type. The instance will be inaccessible by this time until the migration is complete.

B. You can change storage type on the fly from HDD to SDD. Data will be moved to a new storage

type. The instance will be in read-only mode by this time until the migration is complete.

C. You cannot change the BigTable storage type on the fly. You need to launch a new BigTable cluster with SSD storage and use Dataproc to export data from the existing BigTable cluster to the new one.

D. You cannot change the BigTable storage type on the fly. You need to launch a new BigTable cluster with SSD storage and use Dataflow to export data from the existing BigTable cluster to the new one. right

Explanation:

Correct Answer: D

You can change cluster IDs only by deleting and recreating the cluster. Also, you cannot change the instance ID or 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 explanation above, the best solution is using Dataflow to migrate data from the old BigTable cluster to the new one.

All other options are incorrect based on the above explanation.

Option C is incorrect as it will require you to create a spark pipeline to move data from one Bigtable instance to another as compared to Dataflow where you can use a ready-made template

Options A and B are incorrect as Bigtable does not support the storage type migration from SSD to HDD

Source(s):

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

Question 8 Incorrect Marked for review

Domain: Design Data Processing Systems

A financial services company which offers credit card and loan package services uses BigQuery as a data warehouse to store clients details in the denormalized structure. Data analysts are experimenting on Apache Spark for more data transformation and enrichment and after a few presentations, the head of data decided to move forward and use Apache Spark. As the data engineer, you are assigned to provide the required tech stack. What would you do?

A. Create a Dataproc cluster. Install Dataproc's BigQuery connector on the cluster using initialization actions. Dataproc temporarily loads data from BigQuery to Google Storage. If failed, create a Python script to clear all the temporary files on the GCS bucket after the job fails to reduce the manual effort right

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

C. Create a Dataproc cluster. Export data from BigQuery to Google Storage in JSON format. Dataproc cluster reads data from Google Storage using a connector. You need to manually delete data files after Dataproc is done.

D. Create a Dataproc cluster. Export data from BigQuery to Google Storage in CSV format. Dataproc cluster reads data from Google Storage using a connector. Dataproc cluster deletes data from Google Storage after Dataproc is done.

Explanation:

Correct Answer: A

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 the BigQuery connector to the Dataproc cluster using initialization actions. When a Dataproc spark job reads from Big Query, it writes the BigQuery table's content temporarily to Google Storage using the Dataproc cluster's assigned bucket. If the job completes successfully, temporary files are automatically deleted from the cluster. If the job fails, run your Python script to delete the temp directory in order to avoid any human error.

Option B is incorrect: If the job fails, you need to delete temp files manually.

Options C and D are incorrect: Dataproc can read from BigQuery by installing the connector. No need to export data from BigQuery to Google Storage manually.

Source(s):

BigQuery Connector: https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery

Initialization Actions:

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

Question 9 Correct

Domain: Store the data

An organization has TBs of data getting loaded into Google Cloud Storage from different source systems. The Data Engineering team is concerned about potential performance bottlenecks in their data lake on Google Cloud i.e Google Cloud Storage. What should be a key consideration when configuring performance monitoring?

- A. Utilizing automated anomaly detection tools right
- B. Focusing solely on storage performance
- C. Monitoring only during peak usage hours
- D. Excluding monitoring for data processing workloads

Explanation:

Correct Answer: A

Option A is CORRECT because utilizing automated anomaly detection tools enables proactive identification of performance issues.

Option B is incorrect because focusing solely on storage performance may neglect other critical aspects.

Option C is incorrect because monitoring only during peak usage hours may cause performance issues at other times.

Option D is incorrect because excluding monitoring for data processing workloads may result in undetected bottlenecks.

Reference:

https://cloud.google.com/stackdriver

Question 10 Correct Marked for review

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. right
- 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.

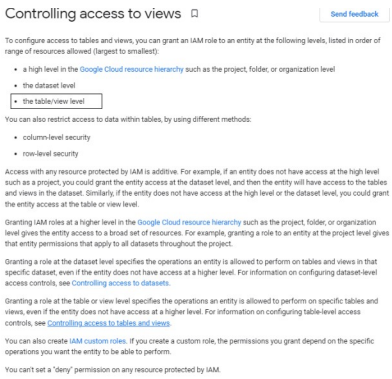
Explanation:

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

Question 11

Correct Marked for review

Domain: Store the data

Your Organization is dealing with a large dataset with varying access patterns. The dataset includes historical records that are rarely accessed but still need to be retained. The organization wants to optimize storage costs while ensuring data availability. Which storage strategy in BigQuery would you recommend for this scenario?

A. Use BigQuery native storage with the "Automatically detect" schema option

B. Use BigQuery native storage with table partitioning by date right

C. Use BigQuery federated queries with Cloud Storage for infrequently accessed data

D. Use BigQuery federated queries with Cloud Bigtable for frequently accessed data

Explanation:

Correct Answer: B

Option B is CORRECT because using BigQuery native storage with table partitioning by date allows for efficient querying and cost optimization, especially when dealing with large historical datasets.

Option A is incorrect because using the "Automatically detect" schema option is more relevant for scenarios where the schema of files occasionally changes.

Option C is incorrect because federated queries with Cloud Storage may introduce higher latency

and costs compared to native storage, especially for infrequently accessed data.

Option D is incorrect because Cloud Bigtable is not the most suitable option for optimizing storage costs in BigQuery. It's typically used for high-throughput, low-latency access to key-value data.

Reference:

https://cloud.google.com/bigquery/docs/querying-partitioned-table

Question 12 Correct Marked for review

Domain: Store the data

A Retail company is loading data into BigQuery from various sources with different schemas. What is the recommended approach to handle varying schemas?

- A. Utilize BigQuery nested and repeated fields to handle diverse schema structures right
- B. Standardize all source schemas to a single format before loading them into BigQuery
- C. Utilize BigQuery schema auto-detection for each load job
- D. Convert all source schemas to Avro format before loading them into BigQuery

Explanation:

Correct Answer: A

Option A is CORRECT because utilizing BigQuery nested and repeated fields allows for flexibility in handling varying schema structures.

Option B is incorrect because standardizing all source schemas may not be practical, especially when dealing with diverse data sources.

Option C is incorrect because schema auto-detection may not handle complex and nested structures effectively.

Option D is incorrect because converting to Avro format does not inherently address diverse schema structures.

Reference:

https://cloud.google.com/bigquery/docs/nested-repeated

Question 13 Correct

Domain: Design Data Processing Systems

As a GCP data engineer, you were asked to use Cloud shell to load the Big query dataset with the data available in Google Cloud Storage.

Google Cloud Storage path: gs://whizlabs-public-dataset/department/cloudcertcount.csv

Big query Table: whizlabs_dataset.cloud_cert_dataset

Data format : CSV

Schema : Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

Can you suggest which one of the following is correct? (select one)

- A. bq load --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER right
- B. bq extract --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER
- C. bq load --schema Name:STRING,CERTIFICATION:STRING,AGE:INTEGER --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv
- D. bq load --target_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

Explanation:

Correct Answer: A

Options A is CORRECT because the current syntax is :
bq load --source_format = [format] [destination dataset].[table] [source_path] [schema]

bq load --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

bq load command is to load data into a Big Query table.

Option B is incorrect because the bq extract command is used to export table data to Google Cloud Storage.

All the other options are invalid.

For more information on the **Big Query-bq load**, please visit the below URL:

https://cloud.google.com/bigquery/docs/reference/bq-cli-reference#bq_load

https://cloud.google.com/bigquery/docs/reference/bq-cli-reference#bq_extract

Question 14 Incorrect Marked for review

Domain: Design Data Processing Systems

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

Which of the following approaches is recommended?

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

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

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

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

Explanation:

Answer: B.

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

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

following table:

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

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

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

Source(s):

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

Question 15 Correct Marked for review

Domain: Design Data Processing Systems

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

- A. Use Google Storage to store data. Allow Dataproc cluster to access data from Google Storage. right
- B. Use BigQuery to store data. Install Dataproc-BigQuery connector to access data.
- C. Use Dataproc cluster's HDFS namenodes to store data.
- D. Use BigTable to store data. Use Dataproc-BigTable connector to access data.

Explanation:

Answer: A.

Description:

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

Cloud Storage is a good option if:

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

Source(s):

Migrating Apache Spark Jobs to Cloud Dataproc:

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

Question 16

Correct Marked for review

Domain: Ingest and process the data

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

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

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

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

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

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

Explanation:

Correct Answer: A

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

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

AutoML: AutoML allows users to train ML models without writing any code. It provides pre-built algorithms and workflows for common ML tasks, such as image classification, text classification, and object detection.

Custom training: Vertex AI also provides tools for custom ML training, including support for popular ML frameworks such as TensorFlow, PyTorch, and scikit-learn.

Model deployment: Vertex AI provides tools for deploying ML models to production, including support for serving models through APIs, web applications, and mobile devices.

Model Management: Vertex AI provides tools for managing ML models throughout their lifecycle, including versioning, monitoring, and retraining.

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

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

incorrect.

References:

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

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

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

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

Question 17

Correct

Domain: Store the data

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

A. Increase persistent disk size for master node.

B. Increase persistent disk size for worker nodes.

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

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

Explanation:

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

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

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

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

Source(s):

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

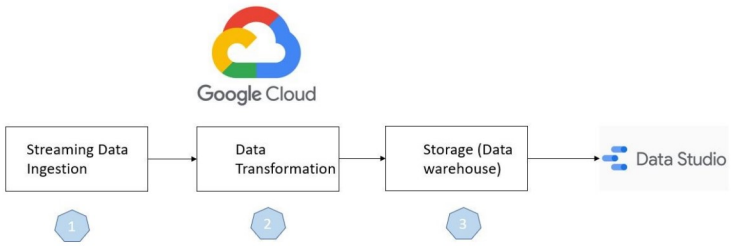
Question 18

Correct

Domain: Design Data Processing Systems

You need to design a data architecture to bring together all your data at any scale and provide insights into all your users.

You need to complete the below architecture.



Which of the following can be used in place of number 3 i.e. Storage (Data Warehouse)?

A. Cloud Pub/Sub

B. Cloud SQL

C. Big Query right

D. Cloud Spanner

Explanation:

Correct Answer: C

Option C is CORRECT because Big Query is recommended as a SQL-complaint analytics data warehouse.

This is a serverless, highly scalable, and cost-effective multi-cloud SQL-complaint data warehouse.

Option A is incorrect because Cloud Pub/Sub is recommended for Streaming data ingestion.

Option B is incorrect because Cloud SQL is a fully-managed database service but not a data warehouse service.

You can use Cloud SQL with MySQL, PostgreSQL, or SQL Server

Option D is incorrect because Cloud Spanner is a distributed SQL database management and storage service.

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

https://cloud.google.com/bigquery/docs

Question 19 Correct

Domain: Ingest and process the data

A data engineering team is implementing a CI/CD (Continuous Integration and Continuous Deployment) pipeline for their data processing workflows on Google Cloud. The team needs a service that can automate the testing, building, and deployment of their data pipelines. Which Google Cloud service is best suited for achieving this CI/CD automation?

A. Google Cloud Dataprep

B. Google Cloud Composer

C. Google Cloud Dataflow

D. Google Cloud Build right

Explanation:

Correct Answer: D

Option D is CORRECT because Google Cloud Build is a fully managed CI/CD platform that automates the testing, building, and deployment of applications, including data pipelines. It integrates seamlessly with other Google Cloud services, making it suitable for CI/CD automation in data engineering workflows.

Option A is incorrect because Google Cloud Dataprep is a visual data preparation tool and is not specifically designed for CI/CD automation.

Option B is incorrect. While Google Cloud Composer is an orchestration service, it is focused on workflow automation and may not provide the same level of CI/CD automation capabilities as Google Cloud Build.

Option C is incorrect. Google Cloud Dataflow is a fully managed service for data processing but does not directly address CI/CD automation for testing and deployment.

Reference:

https://cloud.google.com/build/docs/automating-builds/create-manage-triggers

Question 20 Correct

Domain: Prepare and use data for analysis

Your organization maintains a vast repository of sales data stored in Google BigQuery, partitioned into monthly tables named SALES_yyyymm. Analysts frequently query these tables to generate reports spanning multiple months. However, some queries that cover extended date ranges fail due to hitting the limit of 1,000 tables. How can you effectively address this challenge while optimizing query performance?

A. Aggregate the data into quarterly or yearly tables to reduce the number of tables queried

B. Implement query caching to store and retrieve results from previous queries

C. Use clustered tables to organize and optimize data storage and query execution right

D. Repartition the tables based on a combination of date and region to distribute the query load

Explanation:

Correct Answer: C

Option C is CORRECT as using clustered tables helps optimize query performance by organizing data based on related columns, such as date, which can significantly reduce the number of tables scanned during queries. This approach directly addresses the issue of hitting the table limit while also enhancing query efficiency.

Option A is incorrect because aggregating data into larger intervals may simplify queries but does not directly resolve the table limit problem.

Option B is incorrect because query caching can improve performance for recurring queries but does not address the underlying issue of hitting table limits.

Option D is incorrect because repartitioning tables based on date and region may improve data organization but may not efficiently address the issue of hitting the table limit for queries spanning multiple months.

Reference:

https://cloud.google.com/bigquery/docs/query-performance

Question 21

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 very few number of cells.

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

- [x] C. The schema is not designed for the instance to evenly read and write data across the tables. right

- [x] D. The instance uses HDD storage type. right

- [x] E. The instance was scaled up recently. right

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

Explanation:

Answer: C, D & E.

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

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

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

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

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

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

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

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

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

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

Source(s):

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

Question 22 Correct Marked for review

Domain: Prepare and use data for analysis

A multinational retail corporation wants to share analytical insights derived from its Google BigQuery datasets with various regional teams across the organization. They aim to provide aggregated data views while safeguarding the privacy of individual customer information. Additionally, they seek to optimize resource allocation by ensuring that analysis costs are attributed to the respective regional teams. What approach should the corporation take to fulfill these objectives effectively?

A. Establish separate Google BigQuery datasets for each regional team and grant access based on role-specific Identity and Access Management (IAM) policies

B. Implement Data Studio reports with restricted access controls, enabling regional teams to view and interact with aggregated insights tailored to their needs

C. Develop a custom web application hosted on Google Cloud Platform (GCP), integrating OAuth authentication to enforce user permissions and deliver personalized analytical dashboards

D. Create authorized views in Google BigQuery to expose aggregated data summaries to regional teams while restricting access to underlying user-level data based on role-based access controls right

Explanation:

Correct Answer: D

Option D is CORRECT as creating authorized views in Google BigQuery allows the organization to share aggregated data summaries while controlling access to underlying user-level data. This approach ensures privacy protection and efficient resource allocation by attributing analysis costs to the respective regional teams.

Option A is incorrect because maintaining separate datasets for each regional team may lead to data duplication and increased management complexity.

Option B is incorrect because while Data Studio reports offer visualization capabilities, they may not provide sufficient control over data access and privacy protection.

Option C is incorrect because developing a custom web application introduces additional complexity and may require significant development effort, whereas creating authorized views in BigQuery offers a more streamlined solution for sharing analytical insights.

Reference:

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

Question 23 Correct

Domain: Prepare and use data for analysis

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

Which of the following is recommended?

- A. Cloud Dataprep right
- B. Dataproc
- C. Cloud Composer
- D. Cloud Function

Explanation:

Answer: A.

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

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

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

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

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

Source(s):

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

Question 24 Correct

Domain: Ingest and process the data

Your organization is dealing with an ever-growing dataset stored in Google Cloud Storage. The data engineering team needs to implement a scalable solution for efficiently processing this data in batch mode. Which Google Cloud service is well-suited for this scenario, providing a fully managed and scalable environment for running Apache Spark and Hadoop clusters?

A. Google Cloud Dataflow

B. Google Cloud Composer

C. Google Cloud Dataprep

D. Google Cloud Dataproc right

Explanation:

Correct Answer: D

Option D is CORRECT because Google Cloud Dataproc is a fully managed service for running Apache Spark and Apache Hadoop clusters. It allows for scalable and efficient processing of large datasets in batch mode.

Option A is incorrect. While Google Cloud Dataflow can handle both batch and stream processing, it is more focused on the data processing aspect and may not provide the same level of fine-tuned control over Spark and Hadoop clusters as Dataproc.

Option B is incorrect. Google Cloud Composer is an orchestration service and may not directly address the efficient processing of large datasets in a batch-mode scenario.

Option C is incorrect. Google Cloud Dataprep is a visual data preparation tool and is not designed for running Apache Spark and Hadoop clusters for large-scale batch processing.

Reference:

https://cloud.google.com/dataproc/docs/concepts/overview

Question 25

Incorrect Marked for review

Domain: Store the data

You are using BigQuery as the data warehouse. Different departments are using BigQuery to read data. Upon checking the billing costs, you notice that there is a spike in running queries on BigQuery despite caching is enabled. You started scanning through the queries run on BigQuery trying to find

out if some queries are not cached.

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

- [x] A. SELECT queries without asterisk (*). wrong
- [ ] B. Queries select from authorized views on archive tables.
- [x] C. Queries multiple tables use wildcard. right
- [ ] D. Jobs use destination tables. right

Explanation:

Answers: C & 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.

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

Source(s):

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

Question 26

Correct

Domain: Design Data Processing Systems

A multinational company has multiple Google Cloud projects used by tech teams residing in different countries around the world. Each project is designed to perform data ingestion, storage, processing, cleansing, and transformation based on the country's data such as currency and language. Dataflow pipelines are built to perform ETL/ELT processing for every project.

However, for a certain need, it was required for more than one pipeline to share the same data source while each pipeline does a part of processing implemented by different tech teams. To mitigate this issue, both pipelines should be able to share data among their different phases. Which of the following would help to achieve this?

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

B. If Dataflow instances reside in the same region, data sharing among pipelines is possible. Otherwise, a storage option should be considered.

C. Enable data sharing option while creating Dataflow pipelines.

D. Use Google Storage to share data with other pipeline instances. right

Explanation:

Correct Answer: D

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.

Option A is incorrect: This approach is not recommended. Use Google Storage to share data between pipelines.

Option B is incorrect: Sharing data is not possible unless using a reliable data storage such as Google Storage.

Option C is incorrect: Dataflow doesn't have a cross pipeline communication mechanism for sharing data between pipelines.

Source(s):

Dataflow – FAQ:

https://cloud.google.com/dataflow/docs/resources/faq

Question 27 Correct

Domain: Design Data Processing Systems

An online bank system allows its clients to log into their accounts to check their balance, transfer money, enable and disable debit & credit cards, print transaction logs and offers many other online services. As a rule for security measurements, session logs for each client is recorded with events incoming every 10 seconds from the client's web browser including details about session ID, timestamp, current page, and network IP address. These logs are stored in BigTable for further

aggregation and analysis.

The online system should detect if the client is idle for more than 600 seconds. In case of the idle session, the system should automatically log out and the client is obligated to enter his credentials again to log in. In order to detect that the client is idle within the time window of 600 seconds, data in BigTable should be aggregated and transformed accordingly for the server-side system to deactivate all tokens linked to sessions considered idle. You are using Dataflow to build a data pipeline to aggregate the data.

Which time window should be applied for this scenario?

A. Tumbling Window with a duration of 10 minutes.

B. Hopping Window with a duration of 10 minutes.

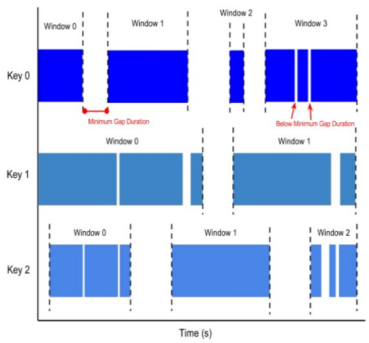
C. Session Window with a time gap of 10 minutes. right

D. Global Window with a time-based trigger of 10 minutes.

Explanation:

Correct Answer: C

A session window function defines windows around the 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 the 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, the Session window is the function to choose to build a Dataflow pipeline.

Source(s):

https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines

Question 28 Correct

Domain: Prepare and use data for analysis

You are working on a project that requires predicting customer churn based on various features stored in a BigQuery dataset. Which technique in BigQuery Machine Learning (BQML) would you use to train a binary classification model for this task?

A. Linear Regression

B. k-Means Clustering

C. Logistic Regression right

D. Neural Network

Explanation:

Correct Answer: C

Option C is CORRECT because logistic regression is a supervised learning algorithm used for binary classification tasks like predicting customer churn.

Option A is incorrect because linear regression is used for predicting continuous numerical values and is not suitable for binary classification tasks.

Option B is incorrect because k-Means clustering is an unsupervised learning algorithm used for clustering and is not suitable for binary classification tasks.

Option D is incorrect because while neural networks can perform binary classification, logistic regression is a more straightforward and efficient choice for this task, especially within the context of BQML.

Reference:

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

Question 29 Correct

Domain: Maintaining and Automating Data Workloads

You oversee data pipelines running on BigQuery, Dataflow, and Dataproc. Your responsibility includes performing health checks, monitoring their behavior, and promptly notifying the team managing the pipelines in case of failure. Additionally, you need to operate seamlessly across multiple projects and prefer utilizing managed products or platform features. What is the recommended approach to fulfill these requirements?

- A. Export the relevant information to Cloud Monitoring and configure an Alerting policy to notify the team in case of pipeline failures right
- B. Deploy a Virtual Machine in Compute Engine running Airflow, and export the information to Cloud Monitoring for monitoring purposes
- C. Export pipeline logs to BigQuery and configure an App Engine application to monitor the logs and send email alerts upon detecting failures
- D. Develop a custom App Engine application leveraging GCP APIs to ingest logs, detect failures, and send email notifications accordingly

Explanation:

Correct Answer: A

Option A is CORRECT as exporting relevant information to Cloud Monitoring and configuring an Alerting policy aligns with the requirement to perform health checks, monitor behavior, and notify the team promptly in case of pipeline failures. This approach utilizes managed products and features for effective monitoring across multiple projects.

Option B is incorrect because while using Airflow on a Virtual Machine in Compute Engine is a viable option for workflow management, it may not offer the same level of integration and ease of use as Cloud Monitoring for monitoring data pipelines.

Option C is incorrect because while exporting logs to BigQuery allows for centralized log storage, setting up an App Engine application to monitor logs and send alerts may introduce additional complexity and maintenance overhead.

Option D is incorrect because while developing a custom App Engine application can provide flexibility, it involves more manual effort and may not offer the same level of integration and ease

of use as Cloud Monitoring for monitoring data pipelines.

Reference:

https://cloud.google.com/monitoring

Question 30 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 right

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

Question 31

Correct

Domain: Prepare and use data for analysis

Your organization utilizes a dataset in BigQuery for extensive analysis. Now, you intend to grant access to the same dataset for third-party companies while keeping data sharing costs low and ensuring data currency. Which solution should you choose?

- A. Utilize Analytics Hub to manage data access and provide third-party companies with access to the dataset right
- B. Implement Cloud Scheduler to regularly export the data to Cloud Storage and grant third-party companies access to the bucket
- C. Create a separate dataset in BigQuery containing the relevant data for sharing and grant access to third-party companies for the new dataset

D. Develop a Dataflow job to periodically read the data and write it to the appropriate BigQuery dataset or Cloud Storage bucket for third-party usage

Explanation:

Correct Answer: A

Option A is CORRECT because leveraging Analytics Hub allows centralized management of data access, ensuring security and control while providing third-party companies with access to the dataset. This solution helps maintain low data-sharing costs and ensures data currency.

Option B is incorrect because while using Cloud Scheduler for data export may provide regular updates, it may not offer the same level of control and access management as Analytics Hub for third-party companies.

Option C is incorrect because creating a separate dataset could lead to redundancy and increased management overhead. It may also not provide the necessary control and access management features offered by Analytics Hub.

Option D is incorrect because although Dataflow can automate data movement, it may not be the most efficient solution for managing data access and ensuring currency for third-party companies.

Reference:

https://cloud.google.com/analytics-hub

Question 32

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. right

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

Question 33 Correct

Domain: Ingest and process the data

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

Which approach will you consider to build the project?

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

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

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

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

Explanation:

Answer B.

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

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

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

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

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

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

Source(s):

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

Question 34

Correct

Domain: Store the data

As part of implementing a data pipeline on Google Cloud, you are tasked with designing storage for a substantial dataset of 20 TB in CSV format. The objective is to minimize the cost of querying aggregate values for multiple users who will interact with the data in Cloud Storage using various query engines. In light of these requirements, what storage service and schema design would you recommend?

A. Leverage Cloud Bigtable for storage and utilize the HBase shell on a Compute Engine instance to query the Cloud Bigtable data

B. Opt for Cloud Bigtable as the storage solution and establish permanent tables in BigQuery for efficient querying

C. Utilize Cloud Storage for storage and establish permanent tables in BigQuery for seamless query execution right

D. Use Cloud Storage for storage and create temporary tables in BigQuery for on-demand query processing

Explanation:

Correct Answer: C

Option C is CORRECT, as Cloud Storage provides a cost-effective and scalable storage solution for large datasets, and linking it as permanent tables in BigQuery allows for efficient querying of aggregate values with the flexibility to use multiple query engines.

Option A is incorrect as it introduces unnecessary complexity by using Cloud Bigtable with the HBase shell, which may not be well-suited for analytical queries.

Option B is incorrect as it might not be the most cost-effective solution for storing large amounts of text files.

Option D is incorrect as it proposes linking Cloud Storage as temporary tables in BigQuery, which may not be ideal for long-term storage and query performance.

Reference:

https://cloud.google.com/bigquery

Question 35 Correct

Domain: Prepare and use data for analysis

You are employed at a financial institution that facilitates online customer registrations. Upon registration, customer user data is sent to Pub/Sub before being ingested into BigQuery. Due to security concerns, you decide to redact customers' Government-issued Identification Numbers while allowing customer service representatives to view the original values when necessary. What approach should you take to achieve this?

A. Utilize BigQuery's built-in AEAD encryption to encrypt the SSN column. Store the keys in a new table accessible only to authorized users

B. Before loading the data into BigQuery, leverage Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token right

C. Before loading the data into BigQuery, employ Cloud Data Loss Prevention (DLP) to substitute input values with a cryptographic hash

D. Implement BigQuery column-level security. Configure table permissions so that only members of the Customer Service user group can access the SSN column

Explanation:

Correct Answer: B

Option B is CORRECT because using Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token ensures that sensitive information, such as Government-issued Identification Numbers, is protected while still allowing authorized users, like customer service representatives, to access the original values when necessary.

Option A is incorrect because while using BigQuery's built-in AEAD encryption may provide encryption for the SSN column, it does not inherently provide a mechanism for selectively revealing the original values to authorized users.

Option C is incorrect because replacing input values with a cryptographic hash using Cloud Data Loss Prevention (DLP) may irreversibly obscure the original values, which may not be suitable for scenarios where selective access to the original values is necessary.

Option D is incorrect because while implementing BigQuery column-level security may restrict access to the SSN column, it does not address the requirement to redact sensitive information while allowing authorized users to view the original values.

Reference:

https://cloud.google.com/dlp

Question 36

Correct

Domain: Maintaining and Automating Data Workloads

Your organization operates several critical data processes in a Google Cloud environment, each with varying resource requirements. You need to minimize costs while ensuring that enough resources are available for business-critical data processes. Which approach should you take to achieve this balance effectively?

A. Implement resource quotas to limit the consumption of compute resources for non-essential processes

B. Utilize preemptible VM instances to reduce costs while prioritizing resource allocation for critical processes

C. Implement auto-scaling policies to dynamically adjust resource allocation based on workload demands right

D. Utilize reserved instances to guarantee resource availability for business-critical processes

while optimizing costs

Explanation:

Correct Answer: C

Option C is CORRECT as implementing auto-scaling policies allows for dynamically adjusting resource allocation based on workload demands, ensuring optimal resource utilization and cost-efficiency for both critical and non-critical data processes.

Option A is incorrect because while implementing resource quotas may limit resource consumption, it may not dynamically adjust to changing workload demands and could potentially hinder critical processes during peak periods.

Option B is incorrect because while preemptible VM instances offer cost savings, they may not provide the required resource availability for business-critical processes, leading to potential disruptions.

Option D is incorrect because while reserved instances guarantee resource availability, they may not dynamically adapt to changing workload demands and could result in underutilization of resources for non-critical processes.

Reference:

https://cloud.google.com/compute/docs/autoscaler

Question 37 Correct

Domain: Design Data Processing Systems

A system receives water temperature details per minute from 500 sensors installed in the different water sources of the region such as lakes, rivers, streams, and natural springs. As a data engineer, you are asked to find a solution to store data to a data warehouse for further analytics and reporting.

Data analytics team recommends using the SQL-like query based on their expertise. Management seeks a solution which could save storage and loading costs. What would you do if you are informed that real-time data reporting is not crucial and update rate for dashboards can be up to 15 minutes?

- A. Use BigQuery to store and query event data. Batch load the data to BigQuery directly using its API. right

- B. Batch-load data into Google Storage. Launch BigTable with 10 nodes to allow high performance and import data from Google Storage to BigTable.

C. Store data in Google Storage. Use Cloud SQL as the main data warehouse and create users with required permissions for data analysts.

D. Enable streaming data to BigQuery and create users for analysts to use BigQuery for reporting.

Explanation:

Correct Answer: A

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

Option B is incorrect: BigTable does not support SQL querying.

Option C is incorrect: This approach is valid, except, it may cost more since you pay for both storage and transfer costs in Google Storage as well as Cloud SQL instance which in return needs administration and scaling up compared to BigQuery which is serverless.

Option D is incorrect: Streaming data to BigQuery is not free. This will add to the costs of building a data warehouse solution and streaming is unnecessary here because streaming data into BigQuery involves continuously ingesting and processing high volumes of data, which can lead to significant charges. BigQuery charges based on the amount of data processed, not the amount of data stored. This means that the more data you stream into BigQuery, the higher your costs will be. BigQuery needs to process streaming data in real-time or near real-time, which requires more computing resources. This increased computational demand can lead to higher costs

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

Question 38 Correct

Domain: Design Data Processing Systems

The data analytics team in your corporation is using commercial visualization software to build dashboards for management and commercial reporting. The software is integrated with the corporation's data warehouse, which is BigQuery, to fetch the data. The finance team reported a hike in billing costs for Google Cloud since the visualization software is being used. You were asked to find the root cause of this. You found that the BigQuery bill was relatively higher compared to previous months due to querying. You checked that caching is enabled so this should not be due to the queries written by data analysts and used for visualization.

Which of the following are valid reasons for BigQuery not using cached queries? (Choose 2 Options)

A. Queries select from the authorized views on archive tables.

B. Queries use nested fields.

- [x] **C. Queries use now() function.** right
- [x] **D. Queries multiple tables using wildcard table.** right

Explanation:

Correct Answer: C and D

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

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

For the complete list of query cases not cached in BigQuery, check "BigQuery – Using Cached Query Results" below.

Source(s):

BigQuery – Using Cached Query Results: https://cloud.google.com/bigquery/docs/cached-results

BigQuery - Wildcards:

https://cloud.google.com/bigquery/docs/querying-wildcard-tables

BigQuery – Cached Results:

https://cloud.google.com/bigquery/docs/cached-results

Question 39

Correct

Domain: Ingest and process the data

Your team is tasked with building a real-time data processing pipeline for processing clickstream data from a website. The requirements include low-latency processing and the ability to scale dynamically based on incoming traffic. Which Google Cloud service or tool is most suitable for this scenario?

- A. Google Cloud Dataflow right
- B. Apache Kafka
- C. Google Cloud Dataproc
- D. Apache Spark

Explanation:

Correct Answer: A

Option A is CORRECT. Google Cloud Dataflow is a fully managed service for both batch and stream processing, making it suitable for real-time data processing with low latency requirements. It dynamically scales based on incoming traffic, providing an efficient solution.

Option B is incorrect. While Apache Kafka is excellent for message buffering and real-time streaming, it is not a fully managed service and may require additional configuration for dynamic scaling.

Option C is incorrect. Google Cloud Dataproc is designed for running Apache Spark and Apache Hadoop clusters but may not offer the same level of real-time processing as Google Cloud Dataflow.

Option D is incorrect. Apache Spark is a powerful framework, but using it on Google Cloud would require additional infrastructure management and might not provide the same level of managed real-time processing as Google Cloud Dataflow.

Reference:

https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines

Question 40

Correct

Domain: Ingest and process the data

Your organization is designing a highly resilient and fault-tolerant data processing pipeline on Google Cloud Platform. The pipeline must handle large volumes of streaming data with low latency requirements. Additionally, the organization requires a mechanism to ensure exact-once delivery semantics for critical financial transactions.Considering these requirements, which combination of Google Cloud services and features would you recommend for ingesting and processing the streaming financial data?

- A. Google Cloud Dataflow with Pub/Sub as the ingestion layer, utilizing Cloud Pub/Sub ordering for exactly-once delivery right
- B. Google Cloud Dataprep with Cloud Storage as the ingestion layer, implementing Cloud Storage object versioning for data consistency
- C. Google Cloud Dataproc with Cloud Pub/Sub as the ingestion layer, leveraging Spark Streaming for low-latency data processing

D. Google Cloud Composer with Cloud Dataflow as the ingestion layer, configuring Apache Beam with at-least-once delivery semantics

Explanation:

CorrectAnswer: A

Option A is the CORRECT choice for ingesting and processing streaming financial data with low latency and ensuring exactly-once delivery semantics. Google Cloud Dataflow, combined with Cloud Pub/Sub as the ingestion layer, allows for efficient processing of streaming data, and Cloud Pub/Sub ordering ensures that messages are processed exactly once.

Option B is incorrect because Google Cloud Dataprep is more suitable for data preparation, and Cloud Storage object versioning is not designed for streaming data ingestion.

Option C is incorrect as Google Cloud Dataproc, while powerful for batch processing, may not meet the low-latency requirements of streaming financial data.

Option D is incorrect because although Google Cloud Composer can orchestrate workflows, using Cloud Dataflow with at-least-once delivery semantics may not meet the requirement for exact-once delivery of critical financial transactions in a streaming pipeline.

Reference:

https://cloud.google.com/dataflow/docs/concepts/streaming-with-cloud-pubsub

Question 41

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

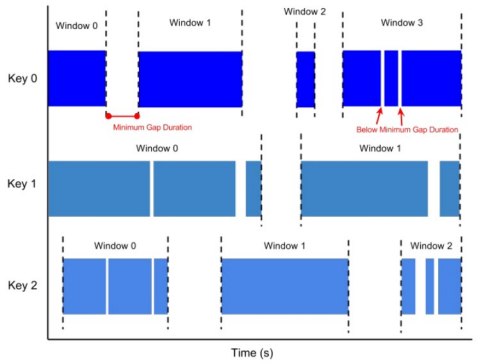
C. Session window with a time gap duration of 10 minutes right

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

Question 42 Correct

Domain: Store the data

You have an on-premise production MySQL database that you have been asked to move to Google Cloud. Users should run SQL queries to fetch data from the database and there are some legacy applications using the database. You are expected to select a cost-effective solution with minimum downtime. Which of the following Google Cloud products is the best for this scenario?

A. Cloud Storage

B. Cloud Spanner

C. Cloud SQL right

D. Cloud Datastore

Explanation:

Correct Answer: C

Cloud SQL is a fully managed database service that makes it easy to set up, maintain, manage, and administer your relational PostgreSQL, MySQL, and SQL Server databases in the cloud.

Here since it's a production instance, minimum downtime is also required a lift, and the shift migration approach is best suitable to this use case

Hence option C is correct

Option A is incorrect: Google Storage is blob storage. It does not work as an RDMS.

Option B is incorrect: Cloud Spanner is a fully managed, relational database service for global application data. It offers strong consistency, high availability, and horizontal scalability, making it ideal for mission-critical applications that require high performance and data integrity across multiple regions.

Here we need to modify the legacy applications so that they can connect with Cloud Spanner. Moreover, we have to transform the queries of the users to be compatible with Cloud Spanner.

Considering all the above factors along with minimum downtime and cost parameters option B is incorrect

Option D is incorrect: Datastore is a schemaless NoSQL database. Migration is from a structured SQL database so Datastore is not a viable choice.

Source(s):

Cloud SQL:

https://cloud.google.com/sql/

Question 43 Correct

Domain: Maintaining and Automating Data Workloads

When organizing workloads based on business requirements, what consideration is essential for determining whether to use interactive or batch query jobs?

A. Ensuring scalability and resource availability for query processing

B. Analyzing data access patterns to optimize query performance right

C. Leveraging dynamic scaling for on-demand resource provisioning

D. Utilizing persistent data clusters for stability and reliability

Explanation:

Correct Answer: B

Option B is CORRECT because analyzing data access patterns allows for optimizing query performance based on actual usage patterns, ensuring efficient resource utilization for both interactive and batch query jobs.

Option A is incorrect because while scalability and resource availability are important factors, they may not directly optimize query performance without considering actual data access patterns.

Option C is incorrect because while dynamic scaling allows for on-demand resource provisioning, it may not directly optimize query performance without considering actual data access patterns.

Option D is incorrect because while persistent data clusters offer stability and reliability, they may not directly optimize query performance without considering actual data access patterns.

Reference:

https://cloud.google.com/bigquery/docs/running-queries

Question 44 Correct

Domain: Prepare and use data for analysis

Your organization has implemented Looker Studio for data visualization and analysis. As part of your role, you need to design a complex dashboard that integrates multiple data sources and provides real-time insights to senior management. Which approach should you take to ensure optimal

performance and usability of the dashboard?

A. Aggregate data at the source before loading it into Looker Studio to minimize query complexity

B. Utilize Looker's caching mechanisms to store frequently accessed data and reduce query processing time right

C. Implement incremental loading techniques in Looker to handle large datasets efficiently

D. Utilize Looker's native SQL Runner feature to directly query the underlying database for real-time results

Explanation:

Correct Answer: B

Option B is CORRECT because leveraging Looker's caching mechanisms helps store frequently accessed data, reducing query processing time and enhancing dashboard performance and usability.

Option A is incorrect because aggregating data at the source before loading it into Looker may limit the flexibility of analyses and visualization options available in Looker Studio.

Option C is incorrect because Looker handles incremental loading automatically, and manually implementing such techniques may not be necessary for optimal performance.

Option D is incorrect because using Looker's native SQL Runner for real-time querying bypasses Looker's caching mechanisms, potentially leading to increased query processing time and performance issues.

Reference:

https://docs.looker.com/dashboard-guide/dashboard-best-practices/query-optimization

Question 45 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. right
- 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

Question 46 Correct

Domain: Prepare and use data for analysis

Your team is developing an application on Google Cloud aimed at automatically generating subject labels for users' blog posts. Due to competitive pressure and limited developer resources, you need to implement this feature quickly, with no prior experience in machine learning. What approach should you take?

- A. Integrate the Cloud Natural Language API into your application and process the generated Entity Analysis results as labels right
- B. Utilize the Cloud Natural Language API within your application and process the generated Sentiment Analysis as labels
- C. Develop and train a text classification model using TensorFlow, deploy it using Cloud Machine

Learning Engine, and call the model from your application to process the results as labels

D. Create and train a text classification model using TensorFlow, deploy it using a Kubernetes Engine cluster, and call the model from your application to process the results as labels

Explanation:

Correct Answer: A

Option A is CORRECT because leveraging the Cloud Natural Language API allows for quick implementation of subject labels without the need for machine learning expertise. By processing the generated Entity Analysis results, the application can efficiently generate subject labels for blog posts.

Option B is incorrect because while Sentiment Analysis could provide insights into the sentiment of blog posts, it may not be suitable for generating subject labels.

Option C is incorrect because building and deploying a custom text classification model using TensorFlow and Cloud Machine Learning Engine would require significant time and expertise, which contradicts the requirement for a quick implementation.

Option D is incorrect because deploying a TensorFlow model using a Kubernetes Engine cluster would also require considerable effort and resources, making it unsuitable for the scenario described.

Reference:

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

Question 47

Correct

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. right

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

Question 48

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. right

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/

Question 49

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. right

B. Reduce samples used for training.

C. Reduce regularization.

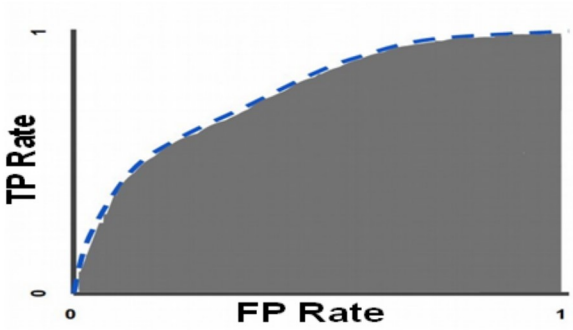
D. Increase feature parameters.

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

Question 50 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)

- [x] A. workerCount right
- [ ] B. masterCount
- [ ] C. masterCPU
- [x] 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

Question 51

Correct

Domain: Maintaining and Automating Data Workloads

Your organization operates critical data processes in a Google Cloud environment. You need to decide between persistent or job-based data clusters for processing large-scale data workloads efficiently. What should you consider when making this decision?

A. Evaluate the cost-effectiveness of persistent clusters based on long-term utilization trends

B. Implement job-based clusters to ensure scalability and resource optimization for irregular workloads right

C. Analyze the availability of resources in persistent clusters to ensure uninterrupted data processing

D. Use persistent clusters to avoid the overhead of cluster initialization and termination

Explanation:

Correct Answer: B

Option B is CORRECT as implementing job-based clusters allows for scalability and resource optimization, especially for irregular workloads, ensuring efficient utilization of resources without incurring unnecessary costs.

Option A is incorrect because while evaluating the cost-effectiveness of persistent clusters is important, it may not address the scalability and optimization needs for sporadic workloads.

Option C is incorrect because analyzing resource availability in persistent clusters may ensure uninterrupted processing but may not offer the flexibility needed for varying workloads.

Option D is incorrect because opting for persistent clusters to avoid initialization and termination overhead may lead to underutilization of resources and increased costs for sporadic workloads.

Reference:

https://cloud.google.com/dataproc/docs/concepts/compute

Question 52 Correct

Domain: Store the data

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

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

A. Cloud SQL

B. Cloud Spanner

C. BigTable

D. Datastore right

Explanation:

Answer: D.

Cloud Datastore is a highly-scalable NoSQL database for your applications. Cloud Datastore

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

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

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

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

Source(s):

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

Question 53 Correct Marked for review

Domain: Prepare and use data for analysis

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

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

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

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

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

Explanation:

Answer: B

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

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

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

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

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

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

Question 54

Incorrect Marked for review

Domain: Ingest and process the data

You want to build a machine learning model to recognize images for Thai cuisine restaurants. You are provided with several image samples for each dish and its name. You used AutoML Vision to build the model. You split the samples into training and test sets. You uploaded the training set to Google Cloud with labels and build the model on AutoML vision. When you tested the newly built model with the test, the confusion matrix shows high false positives with the model confused between different labels. How can you fix the model's accuracy? (Select TWO)

- [ ] A. Remove all images with bad quality. wrong
- [ ] B. Sort images by how “confused” the model is and check if they are labeled correctly. right
- C. If you have a very low training set, consider removing labels altogether. right
- D. Let AutoML Vision decide which images to be considered for training or testing.

Explanation:

Correct Answers - B and C

Description:

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

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

AutoML Vision allows you to sort the images by how “confused” the model is, by the true label and its predicted label. Look through these images and make sure they're labeled correctly.

Consider adding more images to any labels with low quality.

You may need to add different types of images (e.g. wider angle, higher or lower resolution, different points of view).

Consider removing labels altogether if you don't have enough training images.

Remember that machines can’t read your label name; it's just a random string of letters to them. If you have one label that says "door" and another that says "door_with_knob" the machine has no way of figuring out the nuance other than the images you provide it.

Augment your data with more examples of true positives and negatives. Especially important examples are the ones that are close to the decision boundary (i.e. likely to produce confusion but still correctly labeled).

Specify your own TRAIN, TEST, VALIDATION split. The tool randomly assigns images, but near-duplicates may end up in TRAIN and VALIDATION which could lead to overfitting and then poor performance on the TEST set.

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

Source(s):

Cloud AutoML Vision – Evaluating Models:

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

Question 55 Correct

Domain: Design Data Processing Systems

Your company is in a highly regulated industry. You are working on a new project where several data feed from company will be sent to a third party. After start of few weeks of the project your data analyst raise a concern that data feed might have Personal identifiable information(PII) as well. You need to quickly identify whether the outgoing feed has PII information. If yes, you need to take appropriate action. How will you work through this problem in quicker and efficient way?

A. Create a spark job in cloud data proc to read all the data in the feed and search for particular pattern. Using spark will give high performance, also using spark streaming can perform an operation on the feed instantly. If PII information is identified, mask that information with '#'.

B. Create cloud dataflow job to read outgoing data feed first and mask and identified PII information with '#'. Cloud dataflow will give the flexibility to use the same code for batch as well as stream processing.

C. Use cloud data loss prevention api to identify any PII information and de-identify the same by masking it with '#". right

D. Use cloud data prep to figure out the pattern of PII information in the data feed. Create dataprep recipe to mask or delete the PII information from the data.

Explanation:

Correct answer is C.

Option A is incorrect. As we need to find a solution which should be quicker and efficient, creating spark job from scratch will take time and effort. For this purpose, we can use Google Cloud managed data loss prevention API.

Option B is incorrect. Creating dataflow job is not an ideal solution as it will take time and effort.

Option C is correct. Cloud data loss prevention api can detect PII by classifying the data using more

than 90 predefined detectors to identify patterns. Also, it gives the flexibility to mask the data. Refer GCP documentation – https://cloud.google.com/dlp/

Option D is incorrect. The ideal solution to quickly identify the PII data is using data loss prevention API. Using cloud data prep would first of all take time and also it might not remove the PII data as efficiently as data loss prevention API.

Finish Review