Level: Intermediate

Google Cloud Certified Professional Data Engineer

← Back to the **Course**

Practice Test 3

Completed on **Fri, 05 Jul 2024**

2nd Attempt	49/50 Marks Obtained	98.00% Your Score
0h 17m 0s Time Taken	PASS Result	

Share this Report in Social Media Share

Download Report

Domain wise Quiz Performance Report

No.	Domain	Total Question	Correct	Incorrect	Unattempted
1	Design Data Processing Systems	20	20	0	0
2	Store the data	16	16	0	0
3	Ingest and process the data	1	1	0	0
4	Prepare and use data for analysis	6	6	0	0
5	Other	7	6	1	0
Total	**All Domains**	**50**	**49**	**1**	**0**

Review the Answers

Filter By All Questions

Question 1 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

Ask our Experts

Did you like this **Question?**

Question 2 Correct

Domain: Design Data Processing Systems

You have a system-generated log file required to be later uploaded to Google Storage in the data lake. Since the data is only accessed occasionally a few times a year by the development team for debugging and log analysis. You are looking for a cheaper storage option for log files than the standard class. Which of the following is suitable?

A. Cloud Storage Nearline

B. Cloud Storage Coldline right

C. Google Storage Archive

D. Google Storage Snowline

Explanation:

Correct Answer: B

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

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

Options C is incorrect: Archive storage is the lowest-cost, highly durable storage service for data archiving, online backup, and disaster recovery. Unlike the "coldest" storage services offered by other Cloud providers, your data is available within milliseconds, not hours or days. Archive storage is the best choice for data that you plan to access less than once a year. However, since it's mentioned in the question that data needs to be accessed a few times a year.

Option D is incorrect: There is no storage class of Snowline on Google Cloud Storage.

Source(s):

Google Storage Classes:

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

Ask our Experts

Did you like this **Question?**

Question 3 Correct

Domain: Design Data Processing Systems

You have a Dataflow pipeline that streams data to be stored to BigQuery. It is required to deploy a new version of the same Dataflow pipeline which has code changes of the data transformations in it. Now you want to deploy the new pipeline without losing any data and with minimal downtime. How would you achieve this?

A. Create a new Dataflow pipeline with the new version, then switch the data stream to the new pipeline.

B. Deploy the new job using update, jobName, and region parameters to update the running job

right

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

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

Explanation:

Correct Answer: B

In order the update the code for a running job without losing the data we should use the Google recommended practice to update a streaming job using the parameters provided by Google.

Link: https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

Option A is incorrect: as it will lead to data duplication and you can not create pipeline versions in Dataflow.

Option C is incorrect: as it wants to drain the pipeline which will take time and will add latency to the end results also, it requires a brief downtime.

Option D is incorrect: Using the 'Cancel' option will lead to losing in-flight data.

Ask our Experts

Did you like this **Question?**

Question 4

Correct

Domain: Design Data Processing Systems

You are assigned to take care of a data pipeline that runs several steps such as ingesting, processing, and loading data to the data warehouse. The steps are triggered using the App Engine cron job. Each step is triggered 30 minutes apart from the previous one.

As you manage the pipeline, you notice that when one step fails, the rest of the steps are triggered and fail in sequence because their input is missing. This creates unnecessary alarms and you must manually stop all remaining steps before resolving the incident. How do you solve this problem?

A. Use cloud Workflows to orchestrate your cron jobs right

B. Using a Cloud Function, write a script to check if the job was done. If not, it checks the last successful step and re-runs the failed steps.

C. Use Cloud Composer to orchestrate the pipeline's steps.

D. Use Cloud Scheduler to schedule the pipeline steps, with setting a time period of 60 minutes between each step.

Explanation:

Correct Answer: A

Option A is correct because: Cloud Workflows is a fully managed serverless workflow orchestration service that simplifies the process of connecting and automating tasks across Google Cloud. It allows you to define, execute, and manage workflows in a declarative and repeatable manner, eliminating the need for manual intervention and complex code.

Option B is incorrect: as approach still does not fix the root issue which is to find a fault-tolerant and resilient tool to handle pipeline failures.

Option C is incorrect: Cloud Composer is a fully managed workflow orchestration service built on Apache Airflow. Cloud Composer is built specifically to schedule and monitor workflows and take required actions. It requires having a coding knowledge of creating complex DAG's (Directed Acyclic Graphs), according to your requirement which will add an overhead of creating DAG's. Also just for the sake of orchestration, it will be a very costly solution as compared to Cloud Workflows.

Option D is incorrect: Cloud Scheduler is also a cron job product from Google which relies on App Engine. This approach is not different from the current scenario.

Source(s):

https://cloud.google.com/workflows

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

Ask our Experts

Did you like this **Question?**

Question 5

Correct

Domain: Design Data Processing Systems

A system with Apache Kafka installed on Compute Engine virtual machines on Google Cloud is used to stream data feeds for a social media platform. Data include posts, comments, clicks and other behavior events on the platform's pages.

As the platform expands to reach more users by day, it is noticed by the engineering team that the flow of data feeds is slowing down and the time gap between last feed in data sink and current time expands. After some investigation, it is observed that the bottleneck is from Kafka cluster that is unable to stream feeds incoming in a timely fashion. How would you solve this?

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

B. Scale-out Kafka cluster by doubling the number of VMs used.

C. Use Pub/Sub to ingest and stream data feeds instead of using Apache Kafka. right

D. Spin up a new Kafka cluster and distribute data feed evenly between the two clusters.

Explanation:

Correct Answer: C

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

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

Options B & D are incorrect: These are not scalable solutions and eventually, this issue will arise again. Pub/Sub is a fully managed service that is scalable without user action.

Source(s):

Google Pub/Sub:

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

Ask our Experts

Did you like this **Question?**

Question 6

Correct

Domain: Design Data Processing Systems

Your team built an online service which allows users to upload a set of images to convert them into a single PDF file. The service became popular and more users continue to submit their images on the web page. This led to more requests to be processed by backend instances and in return more pending requests are facing timeouts calling the API. If the team uses Google Cloud as the main platform, how would you solve this issue?

A. Build a Dataflow pipeline to process requests sequentially.

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

C. Use Cloud SQL to store incoming requests. Backend instances can read the requests in order

and mark each as complete when done.

- D. Use Pub/Sub to handle users' requests when submitted using the service web page. deploy your backend service to App Engine and make it subscribed to Pub/Sub topic to scale your application according to workloads right

Explanation:

Correct Answer: D

Cloud Pub/Sub is a service to ingest event streams at any scale. It's scalable and reliable for stream analytics and event-driven computing systems.

Pub/Sub is a good product to de-couple a system's components so they communicate with each other asymmetrically. From the scenario shown here, instead of directly calling the API, the platform can "publish" messages to a "topic. The API can be switched to be a "subscriber" which receives the requests submitted and processes each message asymmetrically.

App Engine's Flexible environment supports auto-scaling based on CPU utilization, allowing it to automatically adjust the number of instances running your backend API based on the incoming message traffic from the Pub/Sub topic. This ensures that your API can handle fluctuating workloads efficiently without manual intervention.

Option A is incorrect: Dataflow does not help in solving the increasing API requests.

Option B is incorrect: Increasing timeout isn't a scalable solution and it may continue occurring eventually when more and more users are using the service.

Option C is incorrect: The problem with this solution is, writing and reading records and managing each record's status may require more development to be done making sure no two instances are processing the same request. Managing Cloud SQL can be another issue required to be handled by the team whereas Pub/Sub is a fully managed service does not require maintenance.

Source(s):

Google Pub/Sub:

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

Ask our Experts

Did you like this **Question?**

Question 7 Correct

Domain: Design Data Processing Systems

The management team of your company decides to migrate their on-premise stack to Google Cloud. You have been assigned the task of migrating on-premise HDFS data and Apache Hive clusters. DevOps team requires that they should be able to still manage the components used in Google Cloud. What is the best approach for this scenario?

A. Move on-premise HDFS data to Google Storage. Use Dataproc to process the data.

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

C. Move on-premise HDFS data to Dataproc's persistent disks. Use Dataproc to process the data.

D. Move on-premise HDFS data to Dataproc's persistent disks. Build a Dataflow pipeline. Execute the code in Spark framework provided by Dataflow.

Explanation:

Correct Answer: B

The main objective of a company to move to the Cloud is to save costs, It's the Google recommended best practice to use an ephemeral dataproc cluster i.e. spin the cluster when required and destroy it when the work is done with Preemtible VMs.

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

Ephemeral Dataproc cluster: This is a type of Dataproc cluster that is designed to be short-lived. It is a good option for workloads that need to be processed quickly and then shut down.

Preemptible VMs: These are VMs that are available at a discounted price because they can be reclaimed by Google at any time. They are a good option for workloads that can be interrupted without losing data.

Object lifecycle management: This is a feature of Google Cloud Storage that allows you to automatically manage the lifecycle of your data. For example, you can set up a policy to automatically delete old data or to move data to a different storage class.

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

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

Option D: Due to reasons discussed in the options A & C.

Source(s):

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

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

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

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

Ask our Experts

Did you like this **Question?**

Question 8 Correct

Domain: Design Data Processing Systems

Your company uses Google Cloud as its main cloud platform. The data science team is working on building a binary classification model and they choose Apache Spark MLLib to build the model. The training set to be used is over 30GB in size.

As the data engineer, data science team lead asked you to spin up an Apache Spark cluster for data scientists to experiment on. He informed you that the cluster's local HDFS data is not critical and performance is not an issue until the deployment phase. Which of the following stack will you use?

A. Train data scientists to use Cloud ML Engine for building their classification model.

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

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

D. Launch a Dataproc cluster in standard mode using high-memory worker machine types. right

Explanation:

Correct Answer: D

Cost-Effectiveness: Standard mode provides a balance between cost and performance, making it an ideal choice for exploratory experiments where resources are not heavily utilized.

Scalability: Standard mode clusters can be easily scaled up or down based on the experimental requirements, ensuring that resources are allocated efficiently.

Flexibility: High-memory worker machine types offer ample memory resources, allowing them to handle the data processing demands of various experiments.

Ease of Use: Standard mode clusters are simple to set up and manage, making them suitable for quick experimentation and prototyping.

Option A is incorrect: Cloud ML Engine is used to deploy the model after it is built. You cannot implement the model using ML Engine.

Option B is incorrect: The scenario states non-critical experiments will be conducted by data scientists, Dataproc cluster used can be in standard mode.

Option C is incorrect: Same as B, the scenario states non-critical experiments, there is no need for high-CPU worker machine types.

Source(s):

Compute Engine Machine Types:

https://cloud.google.com/compute/docs/machine- types#standard_machine_types

Cloud Dataprep:

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

Ask our Experts

Did you like this **Question?**

Question 9 Correct

Domain: Design Data Processing Systems

What is the keyword in BigQuery standard SQL used when selecting from multiple tables with wildcard by their suffices?

A. _WILDCARD_SUFFIX

B. _TABLES_SUFFIX

C. _SUFFIX

D. _TABLE_SUFFIX right

Explanation:

Answer: D.

Description:

To restrict the query so that it scans an arbitrary set of tables, use the *_TABLE_SUFFIX* pseudo column in the WHERE clause. The *_TABLE_SUFFIX* pseudo column contains the values matched by the table wildcard.

Answer D is the correct one. Other answers are incorrect because such keywords do not exist.

Source(s):

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

Ask our Experts

Did you like this **Question?**

Question 10

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

Ask our Experts

Did you like this **Question?**

Question 11

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.

Ask our Experts

Did you like this **Question?**

Question 12 Correct

Domain: Design Data Processing Systems

A system with over 4,000 cameras was installed to collect images of main roads in the city in order to detect traffic density. A model built using TensorFlow processes the images coming from the cameras and based on the number of vehicles covering the road it returns the traffic size as a number between 0 (no traffic) and 1 (heavy traffic).

This system is used by map navigation apps to provide traffic details for navigators and recommend the best route based on their current location and destination. For better data quality usage, it was advised to aggregate traffic density by the time frame of 60 seconds with road number as the dimension. This will provide the average traffic density for every minute on each road. If the Dataflow pipeline is being developed to aggregate traffic stats based on this scenario, which time window should be used?

A. Tumbling Window

B. Hopping Window right

C. Session window

D. Global window

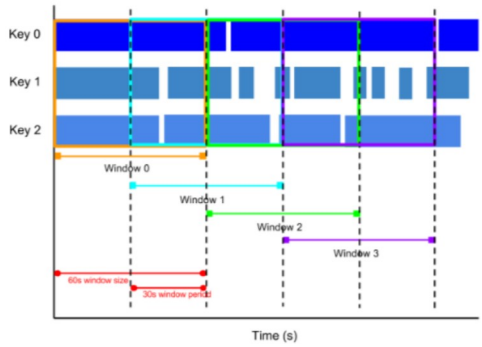
Explanation:

Correct Answer: B

Here we have 4000 cameras installed at different locations in the city. This means, that due to network latency, there may be a lag in receiving the data from the time scheduled to transfer the data. Some Cameras will be able to send the data on time and some require a longer time. This means we need to keep open windows for a longer time.

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

Hopping window is the windowing function recommended for this scenario.



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

Ask our Experts

Did you like this **Question?**

Question 13

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.

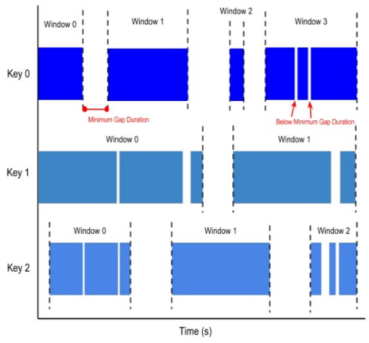
C. Session Window with a time gap of 10 minutes. right

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

Ask our Experts

Did you like this **Question?**

Question 14 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

Ask our Experts

Did you like this **Question?**

Question 15

Correct

Domain: Design Data Processing Systems

Data science team decided to use BigQuery ML for their experimental predictive model scanning session logs generated from the web application that the development team has built and currently maintaining. To use the predictive model in BigQuery, data scientists run the following query to get predictions for the period from July 2017 to August 2018:

```
SELECT *

FROM ML.EVALUATE(MODEL `bqml.predictive_model`, (

SELECT *

FROM `ml-project.predictive_data_table.sessions_*`

WHERE _TABLE_SUFFIX BETWEEN '20170701' AND '20180801'))
```

As a data engineer, you are asked to build a solution which runs this query every 24 hours, scanning session logs for the last 30 days. Results should be written in Google Storage for several data ingestion

and processing tools are able to read prediction output. What would you do?

A. Write a script to run the query in BigQuery, get the results and write them to Google Storage. Deploy the script using Docker to Compute Engine VM instance.

B. Schedule the query to run every 24 hours and export results to Google Storage using BigQuery's scheduled queries tool. right

C. Use Dataproc to run the query on BigQuery. Write the results returned to Google Storage.

D. Build a Dataflow pipeline with scheduler to run the query in BigQuery and export the results to Google Storage.

Explanation:

Correct Answer: B

Using big query's scheduled query functionality you can easily schedule the query to run every 24 hours and export the results to the Cloud storage bucket.

Option A is incorrect: This solution is unnecessary since Dataflow can be a good alternative for this scenario as in option D.

Option C is incorrect: Dataproc does not have a BigQuery connector installed by default, and Dataproc is built to use Hadoop products (Hive, Spark, ..), not for such scenarios.

Option D is incorrect: It is not Google's recommended approach for the given use case. Also, it will require an additional effort to create a dataflow pipeline and schedule it to run every 24 hour, adding the cost of Dataflow and Cloud Scheduler as compared to Scheduled queries which is available at no cost

Source(s):

BigQuery - Exporting Table Data:

https://cloud.google.com/bigquery/docs/exporting-data#export_table_data

BigQuery ML:

https://cloud.google.com/bigquery-ml/docs/bigqueryml-web-ui-start

Ask our Experts

Did you like this **Question?**

Question 16

Correct

Domain: Design Data Processing Systems

Your company is using multiple Google Cloud projects. Since maintaining and managing bills for these projects is becoming very complicated with time, management decided to unify the company's

projects into one and migrate all existing resources to a single project.

One of the projects to be migrated contains several Google Storage buckets with the total estimated file size of 25TB. This data is required to be moved to the newly created project. You need to find a secure and efficient method to migrate data. What Google Cloud product is best for this task?

A. gsutil command

B. Storage Transfer Service right

C. Appliance Transfer Service

D. Dataproc

Explanation:

Correct Answer: B

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 to 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 data center. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, option B is correct, while option C is incorrect.

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

Option D is incorrect: Dataproc may help by reading from source buckets and writes into the destination buckets, but this requires data in source buckets to be used by Hadoop/Apache tools (Partitioned, optimized file formats such as ORC, ..).

Source(s):

Google Cloud Storage Transfer Service:

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

Google Appliance Transfer Service:

https://cloud.google.com/transfer-appliance/

Ask our Experts

Did you like this **Question?**

Question 17

Correct

Domain: Design Data Processing Systems

A startup web-hosting company is using its on-premise servers to host websites and web applications for their clients. The company offers two plans: Free tier and premium tier. The free tier allows clients to host for one year for free with hosting storage up to 50GB. As the startup grows and gains more clients, it was harder to manage more on-premise servers and disks, so the founders decided to move their existing stack to the cloud.

Choosing Google Cloud, one of the challenges required is moving over 100TB of hosting content to Google Storage. The existing network bandwidth can be dedicated to uploading data is approximately 20mb/s. What is the best approach for this?

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

B. Install gsutil command tool on servers and copy the data from on-premise disks to Google Storage.

C. Move on-premise data to HDFS. Launch a Dataproc cluster with 20 worker nodes. Write a script for Dataproc to read from on-premise HDFS and write to Google Storage using a connector.

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

Explanation:

Correct 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 data center. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, option D is the correct one, while option **A is incorrect.**

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

Option C is incorrect: Using Dataproc is not straightforward because data needs to be moved to HDFS and, based on the scenario, files are not optimized for HDFS ecosystem.

Source(s):

Google Cloud Storage Transfer Service:

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

Google Appliance Transfer Service:

https://cloud.google.com/transfer-appliance/

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

Ask our Experts

Did you like this **Question?**

Question 18

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

Ask our Experts

Did you like this **Question?**

Question 19 Correct

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.
- 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

Ask our Experts

Did you like this **Question?**

Question 20

Correct

Domain: Design Data Processing Systems

Your company signed a contract with a retail chain store to handle its data processing applications and tech stack. One of the several applications to be implemented is building an ETL pipeline to ingest the chain store's daily purchase transaction logs to be processed and stored for analysis and reporting; visualize the chain's purchase details for the head management.

Daily transaction logs will be available at 2 am when the day is over and logs are exported to a Google Storage bucket partitioned by date in format (yyyy-mm-dd). Dataflow pipeline should run every day at 3:00 am to ingest and process the logs. Which of the following Google products would help?

A. Cloud Function

B. Compute Engine

C. Cloud Scheduler right

D. Kubernetes Engine

Explanation:

Correct Answer: C

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

Option A: Cloud Functions

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

Hence this is not a correct solution.

Option B: Compute Engine

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

Hence this is not a correct solution.

Option D: Kubernetes Engine

Kubernetes Engine (GKE) is a fully managed service for deploying, managing, and scaling containerized applications on the Google Cloud Platform. It eliminates the need to manage virtual machines and the underlying infrastructure, making it a popular choice for running containerized applications at scale.

Hence it can not be used to trigger the Dataflow pipeline, so option D is incorrect.

Source(s):

Cloud Scheduler:

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

Ask our Experts

Did you like this **Question?**

Question 21 Correct

Domain: Store the data

A banking system is linked to over 400 ATM machines distributed around a region. Each ATM machine sends event data about machine's current state (active, standby, maintenance, ..), bank note balance, current activity type (withdrawal, check balance, ..) and other stats related to business and security purposes. Due to the dynamic attribute structure sent by ATM machines, JSON-formatted events are sent to the centralized system.

The head office's chief data officer wants event data received from ATM machines to be stored in a data warehouse after required cleansing and transformation for the analytics team to fetch reports using SQL-syntax. Which of the following is the best to achieve this?

A. Store event data to Google Storage after converting data to ORC format. Launch a Dataproc cluster to create external tables using Apache Hive on data residing in Google Storage.

B. Load the data to BigQuery with enabling the "auto-detect" option. right

C. Import the data to BigTable. Choose a key combination which allows the best performance while fetching event data based on Google's recommendations.

D. Build a Dataflow pipeline to read JSON data and transform it into a structured format like CSV. Then, load the data to BigQuery.

Explanation:

Correct Answer: B

Schema Auto-detection: Schema auto-detection is available when you load data into BigQuery, and when you query an external data source. When auto-detection is enabled, BigQuery starts the inference process by selecting the file in the data source and scanning up to 100 rows of data to use as a representative sample. BigQuery then examines each field and attempts to assign a data type to that field based on the values in the sample. BigQuery makes the best effort to attempt to automatically infer the schema for CSV and JSON files.

Option A is incorrect as It requires two conversions before getting stored in the table and unnecessary usage of cloud services like Dataproc which is not required

Option C is incorrect as BigTable does not support SQL query syntax for the dashboard

Option D is incorrect as the question mentions the dynamic JSON schema used by the ATM machines to send the events data, so in the Dataflow pipeline you are converting it to CSV which requires have static schema to be able to be ingested in Bigquery directly.

Source(s):

BigQuery – Auto-detect schema:

https://cloud.google.com/bigquery/docs/schema-detect

Ask our Experts

Did you like this **Question?**

Question 22 Correct

Domain: Store the data

As a data engineer, you are assigned to assist data analysts with data modeling to maintain and re-design the existing tables in BigQuery. Data analysts have provided you with the new table schema to be applied to existing datasets. From the list of to-do for modifying the schema, the most notable changes are renaming columns and changing the data type for others, as well as adding "REQUIRED" constraints to some critical columns. How would you achieve this?

- A. Create a new table schema in BigQuery. Insert data from existing dataset to the new one using INSERT SELECT statement. right
- B. Create authorized views on BigQuery tables with new column names and data types.
- C. You can modify columns names and data types and apply constraints using ALTER command.
- D. Export data to Google Storage. Create new tables with the updated schema. Import data to the new schema from Google Storage.

Explanation:

Correct Answer: A

In BigQuery, the following schema modifications are unsupported and require manual

workarounds:

• Changing a column's name.

• Changing a column's data type.

• Changing a column's mode (aside from relaxing REQUIRED columns to NULLABLE). • Deleting a column.

Option B is incorrect: You can cast a column to different data types while creating a view, but cannot apply constraints since views are only for projection.

Option C is incorrect: You can change a column's name but you can not change the data type in bigquery. Also, you cannot apply constraints once the fields are created.

Option D is incorrect: No need to export data to Google Storage. You can use INSERT SELECT to move data from existing tables to new ones.

Source(s):

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

Ask our Experts

Did you like this **Question?**

Question 23 Correct

Domain: Store the data

A system is expected to receive over 15,000 content delivery logs every minute from different web & mobile apps. Logs are received in JSON format. Due to logs being generated by different apps, each developed by a different team, logs do not have a fixed structure and may hold different attributes. Which of the following is a recommended storage option?

A. Cloud SQL

B. Cloud Spanner

C. BigTable right

D. Datastore

Explanation:

Correct Answer: C

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

Options A and B are incorrect: Cloud SQL & Spanner are relational database services. They are not recommended for JSON-format log data with a flexible schema.

Option D is incorrect: Datastore can be a potential choice since it's a NoSQL database. However, Datastore is not built for storing huge data volumes as required in this scenario. Datastore is designed for web applications of small scale.

Source(s):

BigTable vs Datastore:

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

Ask our Experts

Did you like this **Question?**

Question 24

Correct

Domain: Store the data

You receive event data related to on-premise servers holding information about the servers CPU load, memory, disk space, I/O reads and writes, and another application performance stats every 60 seconds. These events are stored in Google Storage. It is decided to use the data to monitor the on-premise architecture. Metrics should be extracted from these events in time-series base for further calculations based on the timeline. Which of the following is best to achieve this?

A. Use Cloud SQL as a database. Move data from Google Storage to Cloud SQL.

B. Use Dataproc with Apache Hive to do required queries on data.

C. Move data to BigTable. Use tall & narrow tables when designing the schema and row key. right

D. Move data to BigTable. Use short & wide tables when designing the schema and row key.

Explanation:

Correct Answer: C

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

For time series, you should generally use tall and narrow tables. This is for two reasons:

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

Option A is incorrect: Cloud SQL is a relational database. Event data might require a flexible structure. Cloud SQL is not scalable to write thousands of rows in a given second.

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

Option D is incorrect: Wide & short table schema is not optimal for time-series event data.

Source(s):

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

Ask our Experts

Did you like this **Question?**

Question 25 Correct

Domain: Store the data

You receive payment transaction logs from e-wallet apps. Transaction logs have a dynamic structure which differs from the e-wallet apps received from. Logs are required to be stored for further security analysis. Transaction logs are critical and it is expected from data storage to have high performance in order to query the required security metrics to be updated in near-real time. Which of the following approaches should you use?

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

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

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

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

Explanation:

Correct Answer: B

When you create a Cloud Bigtable instance, you choose whether its clusters store data on solid-state drives (SSD) or hard disk drives (HDD):

SSD is significantly faster and has more predictable performance than HDD.
HDD throughput is much more limited than SSD throughput. In a cluster that uses HDD storage, it's easy to reach the maximum throughput before CPU usage reaches 100%. To increase throughput, you must add more nodes, but the cost of the additional nodes can easily exceed your savings from using HDD storage. SSD storage does not have this limitation because it offers much more throughput per node. Individual row reads on HDD are very slow. Because of disk seek time, HDD storage supports only 5% of the read rows per second of SSD storage.

The cost savings from HDD are minimal, relative to the cost of the nodes in your Cloud Bigtable cluster, unless you're storing very large amounts of data.

Option A is incorrect as high-performance data storage is required since HDD are not as performant as SSD

Option C and D are incorrect as both are document databases optimized for storing structured JSON data as compared to Logs data which is unstructured.

Source(s):

Choosing Between SSD and HDD Storage:

https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

Querying Cloud Bigtable Data:

https://cloud.google.com/bigquery/external-data-bigtable

Ask our Experts

Did you like this **Question?**

Question 26

Correct Marked for review

Domain: Store the data

A company uses Apache Hive for querying data of 200TB which resides in Google Storage. Apache Hive is installed in on-premise infrastructure.

A decision was made to stop maintaining the on-premise Hive cluster since it incurs outsourcing charges and to find a solution on Google Cloud to replace the Hive cluster as the data warehouse.

The migration should be done in a short time period and cost should be considered.

Which of the following approaches is the most appropriate for this?

A. Use BigQuery as the new data warehouse. Import data from Google Storage to BigQuery.

B. Use BigQuery as the new data warehouse. Create external tables referencing to data in Google Storage. right

C. Use Dataproc as an alternative to the on-premise Apache Hive cluster.

D. Use Bigtable as the new data warehouse. Import data from Google Storage to Bigtable.

Explanation:

Correct Answer: B

Native tables in BigQuery are tables that import the full data inside Google BigQuery as you do in any

other common database system. In contrast, external tables are tables that do not store the data in Google BigQuery, instead, reference the data from an external source, such as a data lake.

The advantages of creating external tables are that they are fast to create so you skip the part of importing data and no additional monthly billing storage costs are accrued to your account since you only get charged for the data that is stored in the data lake, which is comparatively cheaper than storing it in BigQuery.

Option A is incorrect: Importing data into BigQuery from Google Storage may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue that can be more expensive than Google Storage.

Option C is incorrect: Using Dataproc we are still required to use Apache Hive. Our objective is to decommission Apache Hive. Hence this is not the correct solution

Option D is incorrect: Cloud Bigtable is a NoSQL database service for managing massive amounts of structured and semi-structured data. It offers high performance, scalability, and availability, making it ideal for storing and analyzing time-series data, user profiles, and other large datasets.

Hence BigTable can not replace Apache Hive as a data warehouse, so option D is incorrect

Source(s):

BigQuery - Introduction to external data sources:

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

Ask our Experts

Did you like this **Question?**

Question 27

Correct

Domain: Store the data

A stock market company receives real-time updates from different stock prices in the USA. The company seeks a solution that can use stock price data for real-time analysis. The solution should allow high throughput to allow queries to run and return the required results with minimum latency. The solution should also be scaled out for more performance. Which of the following products is the best solution in this scenario?

- A. BigTable right
- B. BigQuery
- C. Datastore
- D. Cloud Spanner

Explanation:

Correct Answer: A

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

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

In general, a cluster's performance increases linearly as you add nodes to the cluster. For example, if you create an SSD cluster with 10 nodes, the cluster can support up to 100,000 rows per second for a typical read-only or write-only workload, with 6 ms latency for each read or write operation.

Option B is incorrect: BigQuery doesn't provide the high throughput and low latency competent to Bigtable. Moreover, you are unable to increase BigQuery's performance, opposed to Bigtable which you can add more nodes for linear performance return.

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

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

Source(s):

Understanding BigTable Performance:

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

Ask our Experts

Did you like this **Question?**

Question 28 Correct

Domain: Store the data

Your company uses BigQuery as its main data warehouse. There are different users with different departments who access BigQuery and run ad-hoc queries. The CTO noticed a hike in BigQuery costs due to running ad-hoc queries scan high volume of data. Hence, she wants to limit the quota for the departments based on their requirements and business needs. How would you achieve this?

A. Allow access to department leads and managers only to control BigQuery access.

B. Set monthly flat-rate pricing for BigQuery.

C. Set custom quotas for each user with access on BigQuery based on their business requirements. right

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

Explanation:

Correct Answer: C

If you have multiple BigQuery projects and users, you can manage costs by requesting a custom quota that specifies a limit on the amount of query data processed per day.

Creating a custom quota on query data allows you to control costs at the project-level or at the user-level.

• Project-level custom quotas limit the aggregate usage of all users in that project.

• User-level custom quotas are separately applied to each user or service account within a project.

In this scenario, the aim is to control user quotas. So, option C is the best approach.

Option A is incorrect: This is not a sufficient solution to restrict access to management only. This is not the goal of CTO.

Option B is incorrect: Flat-rate can be a possible approach. However, BigQuery does not provide flexible flat-rate pricing and the cheapest is $10,000 for 500 slots, which may not be a desirable option for small to medium businesses.

Option D is incorrect: Setting 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 to run queries.

Source(s):

BigQuery - Creating custom cost controls:

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

BigQuery Pricing - Monthly Flat Rate:

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

Ask our Experts

Did you like this **Question?**

Question 29 Correct

Domain: Store the data

You have Apache Spark jobs running on on-premise machines. The team decided to migrate all on-premise resources to Google Cloud and Dataproc was considered to be used to run Spark jobs on the cloud while data was migrated from on-premise HDFS to Google Storage. Dataproc will read and write data from and to Google Storage using the connector.

After a while, you noticed that Spark jobs running on Dataproc are I/O intensive and this is causing latencies reading and writing data in Storage. How would you solve this?

A. Increase persistent disk size for Dataproc cluster's nodes.

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

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

D. Increase RAM capacity of Dataproc cluster's master node.

Explanation:

Correct Answer: C

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

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 worker nodes alone is not enough. You should move data to local HDFS storage of Dataproc. Increasing size may help to increase HDFS storage.

Options B and D are 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

Ask our Experts

Did you like this **Question?**

Question 30 Correct

Domain: Store the data

You have an on-premise relational database that you want to migrate to Google Cloud. You choose Cloud Spanner for importing the database. You noticed that after migration the queries are taking longer time to run. You want to optimize the performance of the Cloud Spanner. Which of the following

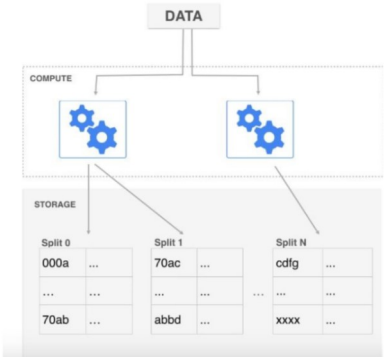
should you have taken into consideration while migrating your tables to Cloud Spanner?

- A. Use version 4 UUID as primary keys for your tables. Keep the original primary keys for legacy queries. right
- B. Make sure your tables primary keys are monotonically increased.
- C. Use UNIX timestamp as a primary key for your tables. Keep the original primary keys for legacy queries.
- D. Use the combination of a timestamp and a primary key (pk) as #timestamp-pk.

Explanation:

Correct Answer: A

Cloud Spanner uses compute nodes to read and write data. The data of tables is stored lexicographically by primary key. Data is distributed among multiple storage "splits."



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

Option A suggests using version 4 UUID as a primary key. A version 4 UUID or a universally unique identifier is a 128-bit number used to identify information in computer systems. An example of a UUID is "81c96908-6a8f-46b2-bc16-3ee4c5376182" which consists of 32 hex characters.

UUIDs generate widely unique and diverse keys which allow potential primary keys, so it can be a good choice to consider to generate primary keys for records in Cloud Spanner.

Option B is incorrect: As mentioned, monotonic primary keys will lead to performance issues while reading & writing data to only one storage split.

Option C is incorrect: UNIX timestamps are monotonic since only right-most digits are changing while left-most digits will be the same in almost all cases, which leads to a performance issue.

Option D is incorrect: Combining timestamp with primary key as #timestamp-pk will lead to having PK combination with the left-most characters being the same for all primary keys.

Source(s):

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

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

Universally Unique Identifier (UUID):

https://en.wikipedia.org/wiki/Universally_unique_identifier

Ask our Experts

Did you like this **Question?**

Question 31

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/

Ask our Experts

Did you like this **Question?**

Question 32 Correct

Domain: Store the data

You have raw data related to retail chain products and purchase records, stored as CSV files in Google Storage. Analytics team wants to use the data for extracting useful statistics for management. They want to run a simple ETL pipeline which analytics will use SQL to do the required transformation and table joins. What is the best approach to achieve this?

- A. Create external tables on data using BigQuery. Run transformation queries on data then load the output to a BigQuery table for reporting and visualization. right
- B. Create external tables on data using BigQuery. Run transformation queries on data then load the output to BigTable for reporting and visualization
- C. Import the data from Google Storage to BigQuery. Run transformation queries on data and insert the transformed records to a CSV file in Google Cloud Storage and use BigQuery external table for reporting and visualization.
- D. Import the data from Google Storage to BigQuery. Run transformation queries on data and export the data to Google Storage. Launch Dataproc cluster and use Hive to query the transformed data

Explanation:

Correct Answer: A

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

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

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

Perform ETL operations on data.

Frequently changed data.

Data is being ingested periodically.

Option B is incorrect because BigTable's schema-less design, while advantageous for certain applications, makes it challenging to perform complex SQL queries and analytics.BigTable is not optimized for SQL query processing, leading to slower query performance compared to relational databases designed for analytics workloads.BigTable lacks native integration with popular SQL-based

data analysis tools, making it less convenient for data exploration and analysis.

Option C is incorrect: Based on Google's best practices, we should use Bigquery native tables for visualization and analytics purposes.

Option D is incorrect: Based on Google's best practices, in the current use case Dataproc should not be used for reporting and visualization

Source(s):

BigQuery external tables:

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

BigQuery – Define external tables:

https://cloud.google.com/bigquery/external-table-definition

Ask our Experts

Did you like this **Question?**

Question 33 Correct

Domain: Store the data

You imported a CSV file to BigQuery using API. You checked the data and found that data is skewed and not properly aligned by the table's columns. What could be the possible issue?

A. You need to explicitly specify the delimiter of the file before loading the data.

B. The file's encoding is not UTF-8. You need to convert the file's encoding and load it again.

C. The file's encoding is not UTF-8. You need to explicitly specify the encoding while loading data. right

D. File size may exceed 1GB. You need to break the file into smaller files.

Explanation:

Correct Answer: C

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

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

Option A is incorrect: While you may explicitly define the delimiter, BigQuery can detect the delimiter.

Option B is incorrect: There is no need to manually convert the file to UTF-8. BigQuery can convert it for you.

Option D is incorrect: There is no limit in data size while loading data to BigQuery.

Source(s):

BigQuery – Loading data:

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

Ask our Experts

Did you like this **Question?**

Question 34 Correct

Domain: Store the data

You are building an ETL pipeline to process over 25GB of data every day. You decide to use Dataproc for custom Apache Spark jobs for data cleansing and transformation. Jobs running on Dataproc require dependencies that are not installed in Dataproc by default while launching a cluster. Security measures in your company don't allow resources to connect to the internet and Dataproc cannot install these dependencies online. What would you do in this situation?

A. Launch a Compute Engine instance and use it as NAT instance to install the dependencies from.

B. Store required dependencies in Google Storage. Install the dependencies to Dataproc nodes using initialization actions. right

C. Launch a Compute Engine instance. Download the dependencies on a persistent disk using VM

instance. Stop the VM instance and use persistent disk for Dataproc cluster to install the dependencies from.

D. While launching Dataproc cluster, you may provide the URIs with which you can download the dependencies from. Dependencies will be installed before the cluster is up.

Explanation:

Correct Answer: B

You can create a Cloud Dataproc cluster with *internal IP addresses* only. However, attempts to access the Internet in an initialization action will fail unless you have configured routes to direct the traffic through a NAT or a VPN gateway. Without having access to the Internet, you can enable Private Google Access, and place job dependencies in Cloud Storage; cluster nodes can download the dependencies from Cloud Storage from internal IPs.

Options A & C are incorrect: There is no need to use a Compute Engine instance as answer B provides a better and cheaper solution.

Option D is incorrect: This is not possible because the cluster will not be able to access the internet if it was set with internal IP addresses only. This also does not comply with security measures which do not allow installing dependencies from the internet.

Source(s):

Initialization actions:

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

Ask our Experts

Did you like this **Question?**

Question 35

Correct

Domain: Store the data

A team of data engineers wants to import a large set of data that contains clickstream logs to Bigtable. They want to test Bigtable before considering it as their final decision. Which of the following conditions should be taken into consideration for successful testing? (Choose 3 options)

- [x] A. You need to wait at least 20 minutes if you scaled up the instance before running the test. right
- [ ] B. Scale up the instance just before the test starts.
- [x] C. Run a heavy pre-test for several minutes before the test starts. right
- [x] D. Don't use less than 300GB of test data. right
- [ ] E. The test should take no longer than 10 minutes.
- [ ] F. Use the development instance for testing.

Explanation:

Correct Answer: A, C, and D

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

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

Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use at least 100 GB of data per node.

Stay below the recommended storage utilization per node.

Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.

Run your test for at least 10 minutes. This step lets Cloud Bigtable optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

Source(s):

Understanding BigTable Performance:

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

Ask our Experts

Did you like this **Question?**

Question 36 Correct

Domain: Store the data

A company uses Bigtable to store session logs generated from its web service. The security team queries these logs to detect possible DDoS (Distributed Denial of Service) attacks coming from specific regions. Lately, the security team informed that they face a slow performance as opposed to what was expected when Bigtable was considered as data storage for session logs. Which of the following are the possible reasons? (Select TWO)

- [x] **A. The rows in the tables have a large data size.** right
- [ ] **B. Data is over 300GB in size.**
- [x] **C. The instance doesn't have enough nodes.** right
- [ ] **D. The instance uses SDD storage type.**

Explanation:

Correct Answer: A and C

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 contain large amounts of data. You can read and write a larger amount 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 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 the 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. The performance of a development instance 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.

Reference:

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

Additional info:

The followings are supported logic:-

Option C: The instance doesn't have enough nodes

There is enough scope for an increase in traffic and a No. of Nodes is not enough to cope with such a huge No. of traffic.

Option A: The rows in the tables have a large data size

This is also possible that "The rows in your Cloud Bigtable contain large amounts of data"

Option D: The instance uses SDD storage type

This is of course not possible for the slow performance.

Ask our Experts

Did you like this **Question?**

Question 37 Correct

Domain: Other

National weather agency wants to collect temperature data in few geographic locations. For this purpose the agency is deploying 100000 internet of things devices. You need to process, store and analyze these very large datasets in real time. How should the system be designed in Google Cloud?

A. Deploy Confluent's Managed Apache Kafka Cluster from the marketplace, create a streaming job in Google Cloud Dataflow to ingest data from Kafka Cluster, and store the data in Google BigQuery.

B. Send the data to Google Cloud Pub/Sub, create streaming job in Google Cloud Dataflow to ingest data from Cloud Pub/Sub and store the data in Google BigQuery. right

C. Send the data to Cloud Storage and create Apache Spark job in Cloud Dataproc for further analysis.

D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Explanation:

Correct Answer is B

Managed Service: Pub/Sub is a fully managed service, meaning that Google takes care of all the infrastructure and maintenance tasks, such as provisioning, scaling, and updating the cluster. This can save you a significant amount of time and effort, especially if you don't have a lot of experience with managing Kafka clusters.

Cost-Effectiveness: Pub/Sub is a highly cost-effective solution, especially for low-volume workloads. You only pay for the data that you ingest and egress, and there are no upfront costs or minimum fees. This can be a significant advantage over Confluent's Managed Apache Kafka Cluster, which has a fixed monthly fee, regardless of your usage.

Integration with Google Cloud: Pub/Sub is tightly integrated with other Google Cloud services, such as Dataflow and BigQuery. This makes it easy to create streaming pipelines that ingest data from Pub/Sub, process it in Dataflow, and store it in BigQuery.

In addition to these three reasons, Pub/Sub also offers several other advantages over Confluent's Managed Apache Kafka Cluster, such as:

Scalability: Pub/Sub can automatically scale to handle spikes in traffic, so you don't have to worry about provisioning or managing additional capacity.

Availability: Pub/Sub is highly available, so you can be confident that your data will be delivered even if there are outages or failures.

Security: Pub/Sub is a secure service that supports role-based access control and encryption.

Considering all the above aspects option A is incorrect

Option B is correct. As requirement is to ingest, transform and store the data so the ideal stack in Google Cloud for this purpose is Pub/Sub - Dataflow - Big Query for IOT data. Refer GCP documentation –

https://cloud.google.com/solutions/iot-overview#ingestion;

Option C is incorrect - As Cloud Storage is not an ideal ingestion service and DataProc is not a data warehousing solution.

Option D is incorrect - As Cloud SQL is not a data warehousing solution.

Ask our Experts

Did you like this **Question?**

Question 38 Correct

Domain: Other

Which of these is the correct statement for Javascript UDF in BigQuery.

A. The user cannot reference a table in UDF.

B. The output of a single row processed by Javascript UDF should be 5MB or less. right

C. Filtering down data before passing to a Javascript UDF does not have any performance benefits.

D. BigQuery only provides persistent UDF that can be used across multiple queries. Creating UDF only for a single query is not yet available.

Explanation:

Correct Answer - B

Option A is incorrect. "

A maximum number of unique UDF plus table references per query — 1000 After full expansion, each UDF can reference up to 1000 combined unique tables and UDFs.

Option B is correct. Data processed for a single row by Javascript UDF in BiqQuery should be 5MB or less. BigQuery processing environment for Javascript UDF has limited memory available per query.

Option C is incorrect. Filtering down the data before being passed to Javascript UDF will have performance benefits as query execution will be faster.

Option D is incorrect. Bigquery provides both persistent and temporary UDF. Temporary UDF can only be used in a single query.

Refer GCP Documentation -

https://cloud.google.com/bigquery/quotas#udf_limits

Ask our Experts

Did you like this **Question?**

Question 39 Correct

Domain: Other

You have a table company which includes a nested column called “financial_group” inside a column called “department”. While querying below query in Big Query an error is received:

SELECT department From `project1.branch.company` WHERE financial_group=””CORPORATE_OPERATION””;

How can the error be corrected?

A. Change “department” to “department.financial_group”.

B. Add “, UNNEST(department)” before the WHERE clause. right

C. Change “department” to “financial_group.department”.

D. Add “, UNNEST(financial_group)” before the WHERE clause.

Explanation:

Correct Answer - B

Option B is correct. The department column needs to be UNNEST for the nested financial_group field to

be used directly in the WHERE clause.

Option A is incorrect as Changing "department" to "department.financial_group" would not resolve the issue as it still attempts to directly access the 'financial_group' property within the 'department' column, which is nested and requires flattening using UNNEST.

Option C is incorrect as Changing "department" to "financial_group.department" would attempt to access the 'department' property within the 'financial_group' property, which is not the correct structure of the nested data. The correct approach is to flatten the 'financial_group' column and then access the 'department' property within it.

Option D is incorrect as Adding ", UNNEST(financial_group)" before the WHERE clause would not flatten only 'financial_group' field, and it won't be accessible directly in the query. Hence, this would not resolve the specific requirement of filtering based on the 'financial_group' property within the nested 'department' column.

Reference:

Please refer to GCP documentation for more information - https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

Ask our Experts

Did you like this **Question?**

Question 40

Correct

Domain: Other

Which of the syntax/statement for wildcard tables is true?

- [x] A. Wildcard tables are available only in standard sql. right
- [ ] B. Wildcard tables WHERE clause supports _TABLE_PREFIX and _TABLE_SUFFIX pseudo column.
- [ ] C. Below wildcard table syntax is correct - 'bigquery-public-data.noaa_gsod.gsod*'
- [x] D. TABLE_DATE_RANGE() function can be used as a table wildcard to query multiple tables in legacy SQL right

Explanation:

Correct answer is A, D.

Refer GCP documentation - https://cloud.google.com/bigquery/docs/reference/standard-sql/wildcard-table-reference

Option A is correct. Wildcard tables feature is only available in standard SQL. Users can query multiple tables using the wildcard character "*" which represents one or more tables.

Option B is incorrect. Wildcard table WHERE clause only supports _TABLE_SUFFIX pseudo column. Users can use this column in WHERE clause to filter the query.

Option C is incorrect. As the wildcard table name contains "*" character, the table name should be enclosed only with backticks(`). In the option above table, the name is enclosed with a single quote (') which will give an error.

Option D is correct. User can query multiple table in legacy SQL using TABLE_DATE_RANGE() and TABLE_DATE_RANGE_STRICT() function.

Ask our Experts

Did you like this **Question?**

Question 41

Correct

Domain: Other

You work in banking related organization where data is stored in Google Cloud Spanner. Data schema is already designed in spanner. One of the bank divisions need to use this data in a heavily manner and would query few of the tables based on certain fixed columns which are not primary key for the table. The division is facing slowness while getting query result. Which of the statement is true and will solve the problem in cost effective manner?

A. Adding additional nodes to Cloud Spanner will provide more CPU and RAM. Thus query latency will be reduced.

B. Adding secondary index for the columns that are frequently queried will avoid full table scan

and thus query latency will be reduced. right

C. User can only add indexes at the time of table creation. Creating secondary index to an existing table while the database continue to serve traffic is not yet available.

D. At the time of table creation all the column which are defined as primary key is already indexed by Cloud Spanner. There is no concept of secondary index in Cloud Spanner. Creating another table will required columns and indexes will solve the problem.

Explanation:

Correct answer is B.

Refer GCP documentation https://cloud.google.com/spanner/docs/secondary-indexes#query-with-index

Option A is incorrect. Adding additional nodes will add more cost to the solution. Also, without correct indexes on the columns full table scan will be done. Adding nodes will not help to solve full table scan problem.

Option B is correct. We can add secondary index at the time of table creation and also for an existing table while the database continues to serve traffic. Adding secondary index will make that column more efficient and data look up in that column will be much faster. It will also avoid full table scan.

Option C is incorrect. User can add secondary table to an existing table without bringing the database down.

Option D is incorrect. User can create secondary index in cloud spanner. There is no need of creating another table or adding nodes to Cloud Spanner.

Ask our Experts

Did you like this **Question?**

Question 42

Correct

Domain: Other

You work as a data engineer in a bank that is migrating from an on-premise database to Google Cloud. You need to select a database in Google Cloud that should cater to below needs:

Database should be relational and should support SQL features.

Should be ACID compliant

Applying a DML should be fast and should not bring the database down.

Should cater growing need for data storage

Should be fast and atomic for transaction handling

Below are the options available. Select the most appropriate one.

A. Cloud SQL

B. Cloud Datastore

C. Cloud Spanner right

D. Alloy DB

Explanation:

Correct answer is C.

Option A is incorrect. As one of the requirement in the question is to consider the growing need for data storage, Cloud SQL is not the correct database as it supports up to 10 TB (Shared Core) for storage per database server. Cloud SQL does vertical scalability and not horizontal scalability.

Option B is incorrect. As the requirement states that the database should be relational, Cloud Datastore is NOSQL database.

Option C is correct. Cloud Spanner should be the choice of a database based on the requirement. Spanner is a relational database with horizontal scaling. Also, it is ACID-compliant and can handle the transaction with strong consistency. Users can add nodes to spanner to increase the database size. Also, implementing a DML/bulk update does not bring the database down. Spanner takes care of all.

Option D is incorrect. Alloy DB checks all the requirements of the given use case and seems the best fit for it, except that it is not ACID-compliant

Ask our Experts

Did you like this **Question?**

Question 43 Incorrect Marked for review

Domain: Other

A social networking company has TB's of data hosted on Cloud BigTable. This data is used by a web application that reads and writes to Bigtable. Currently, the BigTable instance is created in 1 cluster. The data analytics team of the company wants to run some batch jobs that will read the data and produce some analysis. However, it has been observed running the batch jobs alongside web application requests has slowed down things for the application users. Select the approach which will help resolve the problem.

A. Add nodes to Cloud Bigtable instance.

B. Stop the instance, add a new cluster to the instance and enable replication in-app profile. With single cluster routing batch jobs and application, traffic can be routed to different clusters.

C. Add a new cluster without stopping the instance, Cloud Bigtable will replicate the data to the new cluster. With single cluster routing batch jobs and applications, traffic can be routed to different clusters. right

D. Add a new cluster without stopping the instance, Cloud Bigtable will replicate the data to the new cluster. With multi-cluster routing batch jobs and applications, traffic can be routed to different clusters. wrong

Explanation:

Correct Answer - C

Option A is incorrect. In this use case batch jobs submitted will only read the data while the application will read as well as write the data. Also, application requests are latency-sensitive. Adding nodes to the cluster will not solve the issue as both batch jobs and application requests will be catered by the same instance. The more appropriate solution would be to separate the cluster for batch job and application traffic.

Option B is incorrect. For adding a cluster to an existing instance users do not have to stop the instance. BigTable provides the flexibility to add the cluster while the instance is operational.

Once the cluster is added to the instance, BigTable will start the replication and will complete in some time.

Option C is correct. The additional cluster can be added to an existing instance without any downtime. Creating an additional clusters will route the traffic from batch job and application to different clusters based on app profile. By using single cluster routing, users can route the traffic manually in case of

failover.

Option D is incorrect. Multi-cluster routing automatically routes requests to the nearest cluster in an instance. If the cluster becomes unavailable, traffic automatically fails over to the nearest cluster that is available. Bigtable considers clusters in a single region to be equidistant, even though they are in different zones.

For this use case, single cluster routing is an ideal solution as you can define the cluster where the request should be diverted based on the nature of the request. Refer to GCP documentation for more details:

https://cloud.google.com/bigtable/docs/replication-settings

Ask our Experts

Did you like this **Question?**

Question 44

Correct

Domain: Ingest and process the data

You are building a model using TensorFlow. Upon training the model, the results show that the model could return 73% true positives. When you tested the model with a set derived from real data. You noticed a decrease in true positive returns to 65%. You need to tune the model for better prediction. What would you do? (Choose 2 options)

A. Increase feature parameters.

- [x] B. Increase regularization. right
- [x] C. Decrease feature parameters. right

D. Decrease regularization.

Explanation:

Correct Answer: B and C

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

set.

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

Increase the number of examples, the more data a model is trained with, the more use cases the model can be training on and better improves its predictions.

Tune hyperparameters which are related to number and size of hidden layers (for neural networks), and regularization, which means using techniques to make your model simpler such as using dropout method to remove neuron networks or adding "penalty" parameters to the cost function.

Remove features by removing irrelevant features. Feature engineering is a wide subject and feature selection is a critical part of building and training a model. Some algorithms have built-in feature selection, but in some cases, data scientists need to cherry-pick or manually select or remove features for debugging and finding the best model output.

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

Increase the training set.

Decrease features parameters.

Increase regularization.

Source(s):

Building a serverless Machine learning model:

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

Ask our Experts

Did you like this **Question?**

Question 45

Correct

Domain: Prepare and use data for analysis

The security team in your company asked to apply the following rules:

Data on-premise and on the cloud should be encrypted at all times.

Encryption is done using 256-bit AES keys provided by the security team.

Keys should be rotated every 72 days.

Should not store keys on a cloud platform

Should only use the keys provided by the Security team

You use Google Storage to store raw and transformed data. As per the rules above, data should be encrypted when written to Google Storage. As a data engineer, what would you do to satisfy the security team's requirements?

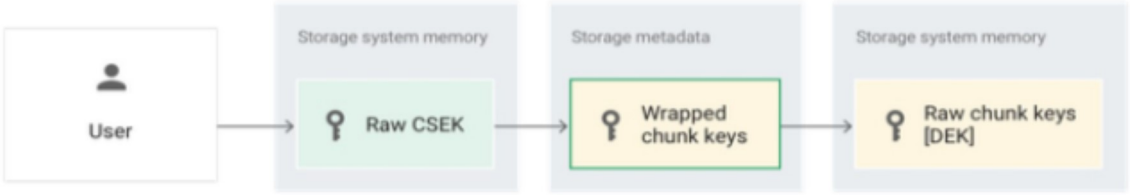
- A. Supply the encryption key provided by the security team and reference it as part of the API service calls to encrypt data in Cloud Storage. right
- B. Upload encryption key provided by the security team to Cloud Key Management Service (KMS) and use the key to encrypt data while writing to Google Storage.
- C. Create symmetric keys using Cloud Key Management Service (KMS) and use them to encrypt data while writing to Google Storage. Create new keys every 72 days.
- D. Create asymmetric keys using Cloud Key Management Service (KMS) and use them to encrypt data while writing to Google Storage. Create new keys every 72 days.

Explanation:

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

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

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



Option B is incorrect: As per one of the requirements of the security team, you should not store keys on cloud platforms. Here option B refers to store encryption keys on KMS on a cloud platform

Options C and D are incorrect: As per one of the requirements of the security team, only encryption keys provided by the Security team should be used for encrypting the data. Hence the keys generated by KMS can't be used to encrypt the data.

Source(s):

Customer-Supplied Encryption Keys:

https://cloud.google.com/security/encryption-at-rest/customer-supplied-encryption-keys/

Ask our Experts

Did you like this **Question?**

Question 46

Correct

Domain: Prepare and use data for analysis

Data analysts in your company are looking for a tool that can be used to check the data tables uploaded by showing review of the data rows and some of the useful stats on columns such as missing cells, expected data type and possible pattern. The tool should provide a user-friendly and easy-to-use UI for data analysts. Which of the following products would you recommend for this scenario?

A. Dataflow

B. BigQuery

C. Dataproc

D. Cloud Dataprep right

Explanation:

Correct Answer: D

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

Option A is incorrect: Using Dataflow means building a workflow pipeline which can be complicated and time-consuming for this scenario.

Option B is incorrect: BigQuery does not provide the stats required in the scenario by default. You need to run and schedule queries to return such stats.

Option C is incorrect: Dataproc is a fully managed and highly scalable service for running Apache Hadoop, Apache Spark, Apache Flink, Presto, and 30+ open-source tools and frameworks on the Google Cloud Platform (GCP). It enables users to easily create and manage clusters of virtual machines (VMs) that are pre-configured with the necessary software and tools for big data processing and analytics. Dataproc can handle both batch and streaming data workloads, and it can be integrated with other GCP services, such as BigQuery and Cloud Storage, for seamless data processing and analysis.

So, Dataproc is used for data processing and it is required to develop Spark code for the use case. Here in the question, it is clearly stated to select a user-friendly tool

Source(s):

Cloud Dataprep:

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

Ask our Experts

Did you like this **Question?**

Question 47 Correct

Domain: Prepare and use data for analysis

You have a compute engine virtual instance hosting your Wordpress blog on the cloud. You've scheduled daily snapshots for your VM's persistent disk. One day, you faced an issue and the blog crashed due to incompatible PHP configuration while trying to upgrade to a newer version, so you need to recover the persistent disk from a snapshot for the previous day. How would you achieve this?

- A. Create a replacement instance directly by selecting the snapshot from the list of daily snapshots available. right
- B. Create a new compute instance with the same exact machine type as the one in production which the snapshots were created from before. Create a persistent disk using the snapshot to be restored from. Attach the persistent disk to the compute engine instance.
- C. You need to create a persistent disk from the snapshot to be restored from. Then, create a new compute engine instance and attach it to the restored persistent disk.
- D. Export one of the snapshots to be used for recovery to Google Storage. Create a new compute engine instance, then using gsutil tool, copy the snapshot to the instance's persistent disk to be restored.

Explanation:

Correct Answer: A

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

Option B is incorrect because it requires you to manually create a new compute instance with the same machine type and attach the persistent disk to it. This can be a time-consuming and error-prone process also it doubles the resource consumption which is not a best practice

Option C is incorrect because it does not take advantage of the fact that snapshots are already stored in a highly available and durable storage location. By creating a new persistent disk from the snapshot, you are creating a copy of the snapshot, which can increase your storage costs.

Option D is incorrect because it requires you to export the snapshot to Google Storage, which can be a

time-consuming and network-intensive process. Additionally, you need to use the gsutil tool to copy the snapshot to the instance's persistent disk, which can be a complex and error-prone process.

Source(s):

Restoring and Deleting Persistent Disk Snapshots:

https://cloud.google.com/compute/docs/disks/restore-and-delete-snapshots

Ask our Experts

Did you like this **Question?**

Question 48 Correct

Domain: Prepare and use data for analysis

Your company's managed system uses a Datastore as a NoSQL database for data storage. In the last meeting, your team decided to make a mandatory daily backup for the Datastore for future recovery in case of data loss or uncontrolled manipulation.

You are responsible for implementing a way to make daily backups of the data in the datastore. Which of the following options is suitable for the requirement of the question?

- A. Use Cloud Scheduler to schedule daily backups specifying which Storage bucket and path location to be exported to. right
- B. Import data to BigQuery, run a query to select data for the past 24 hours and export the results to Google Storage.
- C. Create a BigTable cluster to be used as a backup database. Import whole data from Datastore to BigTable every day.
- D. Using gcloud command, run a cron job to export Datastore data to Google Storage. You may import the data later using gcloud when needed.

Explanation:

Correct Answer: A

Option A is correct because this is the most efficient and straightforward approach for implementing daily Datastore backups. The Google Cloud Scheduler provides a user-friendly interface for scheduling

backups, specifying the Storage bucket and path location for data export. This eliminates the need for manual scripting or complex data processing, making it the most suitable choice for daily Datastore backups.

Option B is incorrect as importing data to BigQuery and then exporting the results to Google Storage is an inefficient and time-consuming approach for daily backups. It involves unnecessary data movement and processing, making it less suitable for a daily backup routine.

Option C is incorrect as creating a BigTable cluster and importing the entire Datastore data into it every day is an overly complex and resource-intensive solution for daily backups. BigTable is designed for high-throughput data operations, not for backup purposes.

Option D is incorrect as while using gcloud commands to export Datastore data to Google Storage is feasible, it requires manual intervention to schedule the cron job. It's not as straightforward as using the Google Cloud console's built-in backup scheduling feature.

Reference:

https://cloud.google.com/datastore/docs/export-import-entities

Ask our Experts

Did you like this **Question?**

Question 49 Correct

Domain: Prepare and use data for analysis

A company wants to migrate its on-premise MySQL relational database to the cloud. The company is looking for a cost-effective solution that grants the same capabilities as their on-premise database. The company mentions that the performance is not an issue and latency is acceptable. However, their main demand is high availability with minimum costs. What would you do in this situation?

A. Use BigTable with 10-node cluster.

B. Use Cloud Spanner multi-regional instance with multiple nodes.

C. Use Cloud SQL with read replicas.

D. Use Cloud SQL with fail-over replicas. right

Explanation:

Correct Answer: D

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

Option A is incorrect: BigTable is a NoSQL table.

Option B is incorrect: Cloud Spanner can be an expensive approach while the scenario is seeking a cost-efficient alternative for their on-premise database.

Option C is incorrect: A read replica is a copy of the master that reflects changes to the master instance in almost real time. The main purpose of reading replicas is for additional read capacity for analytics. Read replicas are NOT for failure recovery in case the primary database is out of service.

Source(s):

Cloud SQL - MySQL High Availability:

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

Cloud SQL - MySQL Replication Options:

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

Ask our Experts

Did you like this **Question?**

Question 50

Correct

Domain: Prepare and use data for analysis

A company uses BigQuery as its main data warehouse. The CTO wants to monitor BigQuery's usage among users of different departments. Since BigQuery runs queries by allocating slots for performance. She wants to know how many slots are used daily for further billing considerations. How would you fulfill CTO's requirements?

A. From BigQuery UI, you can view slot allocation. You can share it by exporting the stats to a file.

B. You need to contact Google Cloud support in order to enable slot utilization metrics on BigQuery UI.

C. Use Cloud Monitoring to view the slot utilization chart on the dashboard. right

D. When you write a query in BigQuery text box, the smart predictor will show how much data will be scanned and the number of slots will be allocated for the query.

Explanation:

Correct Answer: C

CTO wants "how many slots are used daily"

This means we need a Daily slot utilization report

Slot utilization is derived by dividing the total number of slot-milliseconds (`total_slot_ms` from `INFORMATION_SCHEMA.JOBS_BY_ORGANIZATION`) consumed by all jobs on a given day by the number of milliseconds in a day (1000 * 60 * 60 * 24).

```
SELECT
  TIMESTAMP_TRUNC(jbo.creation_time, DAY) AS usage_date,
  jbo.reservation_id,
  jbo.project_id,
  jbo.job_type,
  jbo.user_email,
  -- Aggregate total_slots_ms used for all jobs on this day and divide
  -- by the number of milliseconds in a day. Most accurate for days with
  -- consistent slot usage
  SAFE_DIVIDE(SUM(jbo.total_slot_ms), (1000 * 60 * 60 * 24)) AS average_daily_slot_usage
FROM
  `region-{region_name}`.INFORMATION_SCHEMA.JOBS_BY_ORGANIZATION jbo
GROUP BY
  usage_date,
  jbo.project_id,
  jbo.job_type,
  jbo.user_email,
  jbo.reservation_id
ORDER BY
  usage_date ASC
```

However, this has not been given in any of the options.

Option C is correct::

Cloud Monitoring is a crucial tool for monitoring and managing BigQuery slot allocation and utilization. It provides comprehensive insights into how your BigQuery slots are being used, enabling you to optimize resource allocation and avoid costly overprovisioning or underutilization.

Key Metrics for BigQuery Slot Monitoring:

1. Slots Allocated: This metric indicates the total number of slots currently allocated to your project. It helps you understand the overall slot usage and identify potential bottlenecks.
2. Slots Utilized: This metric represents the number of slots actively being used by your BigQuery queries. It helps you assess the actual demand for slots and identify periods of high or low utilization.
3. Slot Utilization Ratio: This metric measures the percentage of slots that are being utilized, providing a relative measure of slot usage efficiency. It helps you identify underutilized periods and potential opportunities to optimize resource allocation.
4. Slot Wait Time: This metric measures the average time a query waits before it can be assigned slots. It indicates the level of contention for slots and potential resource constraints.

Option A is incorrect: "BigQuery UI, you can view slot allocation"

Here we need Slot Utilization and not Slot Allocation.

Option B is incorrect: You do not need to contact Google Cloud support. We can do it ourselves. We have multiple ways.

Option D is incorrect: Smart predictor does not show you the number of slots that will be allocated for the query. Additionally, We need Slot Utilization Report and not Slot Allocation Report.

Source(s):

Stackdriver:

https://cloud.google.com/stackdriver/

BigQuery Monitoring Using Stackdriver:

https://cloud.google.com/bigquery/docs/monitoring#slots- available

Ask our Experts

Did you like this **Question?**

Finish Review