EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 205 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 205
Topic #: 1
[All Professional Data Engineer Questions]

You have a data processing application that runs on Google Kubernetes Engine (GKE). Containers need to be launched with their latest available configurations from a container registry. Your GKE nodes need to have GPUs, local SSDs, and 8 Gbps bandwidth. You want to efficiently provision the data processing infrastructure and manage the deployment process. What should you do?

A. Use Compute Engine startup scripts to pull container images, and use gcloud commands to provision the infrastructure.

B. Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images.
Most Voted

C. Use GKE to autoscale containers, and use gcloud commands to provision the infrastructure.

D. Use Dataflow to provision the data pipeline, and use Cloud Scheduler to run the job.

Hide Answer

Suggested Answer: *C*

Community vote distribution

B (100%)

by Atnafu at *Nov. 30, 2022, 11:10 p.m.*

Comments

vamgcp 2 weeks ago
Selected Answer: B
B is correct
upvoted 1 times

whorillo 3 months, 3 weeks ago
Selected Answer: B
B is correct
upvoted 1 times

charline 5 months, 3 weeks ago
Selected Answer: B
b is ok
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 204 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 204
Topic #: 1
[All Professional Data Engineer Questions]

You want to create a machine learning model using BigQuery ML and create an endpoint for hosting the model using Vertex AI. This will enable the processing of continuous streaming data in near-real time from multiple vendors. The data may contain invalid values. What should you do?

A. Create a new BigQuery dataset and use streaming inserts to land the data from multiple vendors. Configure your BigQuery ML model to use the "ingestion" dataset as the framing data.

B. Use BigQuery streaming inserts to land the data from multiple vendors where your BigQuery dataset ML model is deployed.

C. Create a Pub/Sub topic and send all vendor data to it. Connect a Cloud Function to the topic to process the data and store it in BigQuery.

D. Create a Pub/Sub topic and send all vendor data to it. Use Dataflow to process and sanitize the Pub/Sub data and stream it to BigQuery.
Most Voted

Hide Answer

Suggested Answer: *A*

Community vote distribution

D (100%)

by Atnafu at *Nov. 30, 2022, 11:08 p.m.*

Comments

vamgcp 2 weeks ago
Selected Answer: D
Option D -Dataflow provides a scalable and flexible way to process and clean the incoming data in real-time before loading it into BigQuery.
upvoted 1 times

AzureDP900 7 months, 1 week ago
D. Create a Pub/Sub topic and send all vendor data to it. Use Dataflow to process and sanitize the Pub/Sub data and stream it to BigQuery.
upvoted 1 times

odacir 8 months ago
Selected Answer: D
D is the best option to sanitize the data to its D
upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 203 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 203
Topic #: 1
[All Professional Data Engineer Questions]

A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard-32) takes two days to complete training. The model has custom TensorFlow operations that must run partially on a CPU. You want to reduce the training time in a cost-effective manner. What should you do?

A. Change the VM type to n2-highmem-32.

B. Change the VM type to e2-standard-32.

C. Train the model using a VM with a GPU hardware accelerator. Most Voted

D. Train the model using a VM with a TPU hardware accelerator.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by gudiking at *Nov. 29, 2022, 2:15 p.m.*

Comments

AzureDP900 7 months, 1 week ago
C. Train the model using a VM with a GPU hardware accelerator.
upvoted 1 times

jkhong 7 months, 3 weeks ago
Selected Answer: C
Cost effective - among the choices, it is cheaper to have a temporary accelerator instead of increasing our VM cost for an indefinite amount of time
D -> TPU accelerator cannot support custom operations
upvoted 2 times

Atnafu 8 months, 1 week ago
C
https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus:~:text=Models%20with%20a%20significant%20number%20of%20custom%20TensorFlow%20operations%20that%20must%20run%20at%20least%20partially%20on%20CPUs
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 202 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 202
Topic #: 1
[All Professional Data Engineer Questions]

Your platform on your on-premises environment generates 100 GB of data daily, composed of millions of structured JSON text files. Your on-premises environment cannot be accessed from the public internet. You want to use Google Cloud products to query and explore the platform data. What should you do?

A. Use Cloud Scheduler to copy data daily from your on-premises environment to Cloud Storage. Use the BigQuery Data Transfer Service to import data into BigQuery.

B. Use a Transfer Appliance to copy data from your on-premises environment to Cloud Storage. Use the BigQuery Data Transfer Service to import data into BigQuery.

C. Use Transfer Service for on-premises data to copy data from your on-premises environment to Cloud Storage. Use the BigQuery Data Transfer Service to import data into BigQuery. Most Voted

D. Use the BigQuery Data Transfer Service dataset copy to transfer all data into BigQuery.

Hide Answer

Suggested Answer: *A*

Community vote distribution

C (89%) 11%

by ducc at *Sept. 3, 2022, 4:15 a.m.*

Comments

cetanx 1 month, 1 week ago

Selected Answer: C

"Your on-premises environment cannot be accessed from the public internet" statement suggests that inbound traffic from internet is NOT allowed however, it doesn't mean that outbound internet connectivity from on-prem resources is not possible. Any on-prem system with outbound internet access can copy/transfer the CSV files.

CSV files are located on a filesystem, therefore you cannot copy them with BQ Transfer Service.

Leaving only possible option;
first copy CSVs to cloud storage
then run BQ Transfer Service

pls refer to https://cloud.google.com/bigquery/docs/dts-introduction#supported_data_sources

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 201 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 201
Topic #: 1
[All Professional Data Engineer Questions]

You need to migrate a Redis database from an on-premises data center to a Memorystore for Redis instance. You want to follow Google-recommended practices and perform the migration for minimal cost, time and effort. What should you do?

A. Make an RDB backup of the Redis database, use the gsutil utility to copy the RDB file into a Cloud Storage bucket, and then import the RDB file into the Memorystore for Redis instance. Most Voted

B. Make a secondary instance of the Redis database on a Compute Engine instance and then perform a live cutover.

C. Create a Dataflow job to read the Redis database from the on-premises data center and write the data to a Memorystore for Redis instance.

D. Write a shell script to migrate the Redis data and create a new Memorystore for Redis instance.

Hide Answer

Suggested Answer: *B*

Community vote distribution

A (100%)

by ducc at *Sept. 3, 2022, 4:12 a.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: A

A. Make an RDB backup of the Redis database, use the gsutil utility to copy the RDB file into a Cloud Storage bucket, and then import the RDB file into the Memorystore for Redis instance.

The import and export feature uses the native RDB snapshot feature of Redis to import data into or export data out of a Memorystore for Redis instance. The use of the native RDB format prevents lock-in and makes it very easy to move data within Google Cloud or outside of Google Cloud. Import and export uses Cloud Storage buckets to store RDB files.

Reference:
https://cloud.google.com/memorystore/docs/redis/import-export-overview

upvoted 8 times

vamgcp Most Recent 2 weeks ago

Selected Answer: A

Option A

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 200 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 200
Topic #: 1
[All Professional Data Engineer Questions]

Government regulations in the banking industry mandate the protection of clients' personally identifiable information (PII). Your company requires PII to be access controlled, encrypted, and compliant with major data protection standards. In addition to using Cloud Data Loss Prevention (Cloud DLP), you want to follow
Google-recommended practices and use service accounts to control access to PII. What should you do?

A. Assign the required Identity and Access Management (IAM) roles to every employee, and create a single service account to access project resources. Most Voted

B. Use one service account to access a Cloud SQL database, and use separate service accounts for each human user.

C. Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users.

D. Use Cloud Storage to comply with major data protection standards. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group. Most Voted

Hide Answer

Suggested Answer: *C*

Community vote distribution

D (53%) A (47%)

by ducc at *Sept. 3, 2022, 4:04 a.m.*

Comments

NicolasN Highly Voted 8 months ago

Selected Answer: A

✅[A] is the only acceptable answer.
❌[B] rejected (no need to elaborate)
❌[C] and [D] rejected. Why should we be obliged to use Cloud Storage? Other storage options in Google Cloud aren't compliant with "major data protection standards"?
==========
❗[D] has another rejection reason, the following quotes:
🔶From <https://cloud.google.com/iam/docs/service-accounts>: "You can add service accounts to a Google group, then grant roles to the group. However, adding service accounts to groups is not a best practice. Service accounts are used by applications, and each application is likely to have its own access requirements"
🔶From <https://cloud.google.com/iam/docs/best-practices-service-accounts#groups>: "Avoid using groups for granting service accounts access to resources"

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 199 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 199
Topic #: 1
[All Professional Data Engineer Questions]

You are using BigQuery and Data Studio to design a customer-facing dashboard that displays large quantities of aggregated data. You expect a high volume of concurrent users. You need to optimize the dashboard to provide quick visualizations with minimal latency. What should you do?

A. Use BigQuery BI Engine with materialized views. Most Voted
B. Use BigQuery BI Engine with logical views.
C. Use BigQuery BI Engine with streaming data.
D. Use BigQuery BI Engine with authorized views.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (96%) 4%

by ducc at *Sept. 3, 2022, 4:02 a.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: A

A. Use BigQuery BI Engine with materialized views.

upvoted 8 times

vamgcp Most Recent 2 weeks ago

Selected Answer: A

Materialized views are precomputed query results that are stored in memory, allowing for faster retrieval of aggregated data. When you create a materialized view in BigQuery, it stores the results of a query as a table, and subsequent queries that can leverage this materialized view can be significantly faster compared to computing them on the fly.

upvoted 1 times

phidelics 1 month, 3 weeks ago

Selected Answer: A

periodically cache the results for perfomance

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 198 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 198
Topic #: 1
[All Professional Data Engineer Questions]

You are implementing workflow pipeline scheduling using open source-based tools and Google Kubernetes Engine (GKE). You want to use a Google managed service to simplify and automate the task. You also want to accommodate Shared VPC networking considerations. What should you do?

A. Use Dataflow for your workflow pipelines. Use Cloud Run triggers for scheduling.

B. Use Dataflow for your workflow pipelines. Use shell scripts to schedule workflows.

C. Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the host project.

D. Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the service project. Most Voted

Hide Answer

Suggested Answer: *A*

Community vote distribution

D (93%) 7%

by ducc at *Sept. 3, 2022, 4 a.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: D

D. Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the service project.

Shared VPC requires that you designate a host project to which networks and subnetworks belong and a service project, which is attached to the host project. When Cloud Composer participates in a Shared VPC, the Cloud Composer environment is in the service project.

Reference:
https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc

upvoted 8 times

vamgcp Most Recent 1 week, 6 days ago

Please correct if I am wrong.. I think it is Option C coz I feel Option D is incorrect because placing the Cloud Composer resources in the service project would not allow you to access resources in the host project.

upvoted 1 times

Ender_H 2 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 197 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 197
Topic #: 1
[All Professional Data Engineer Questions]

You are designing a system that requires an ACID-compliant database. You must ensure that the system requires minimal human intervention in case of a failure.
What should you do?

A. Configure a Cloud SQL for MySQL instance with point-in-time recovery enabled.

B. Configure a Cloud SQL for PostgreSQL instance with high availability enabled. Most Voted

C. Configure a Bigtable instance with more than one cluster.

D. Configure a BigQuery table with a multi-region configuration.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

by ducc at *Sept. 3, 2022, 3:57 a.m.*

Comments

NicolasN Highly Voted 9 months ago

Selected Answer: B

We exclude [C[as non ACID and [D] for being invalid (location is configured on Dataset level, not Table).
Then, let's focus on "minimal human intervention in case of a failure" requirement in order to eliminate one answer among [A] and [B].
Basically, we have to compare point-in-time recovery with high availability. It doesn't matter whether it's about MySQL or PostgreSQL since both databases support those features.
- Point-in-time recovery logs are created automatically, but restoring an instance in case of failure requires manual steps (described here: https://cloud.google.com/sql/docs/mysql/backup-recovery/pitr#perform-pitr)
- High availability, in case of failure requires no human intervention: "If an HA-configured instance becomes unresponsive, Cloud SQL automatically switches to serving data from the standby instance." (from https://cloud.google.com/sql/docs/postgres/high-availability#failover-overview)
So answer [B] wins.

upvoted 29 times

Mcloudgirl 9 months ago

Your explanation is perfect, thanks

upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 196 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 196
Topic #: 1
[All Professional Data Engineer Questions]

You have 15 TB of data in your on-premises data center that you want to transfer to Google Cloud. Your data changes weekly and is stored in a POSIX-compliant source. The network operations team has granted you 500 Mbps bandwidth to the public internet. You want to follow Google-recommended practices to reliably transfer your data to Google Cloud on a weekly basis. What should you do?

A. Use Cloud Scheduler to trigger the gsutil command. Use the -m parameter for optimal parallelism.

B. Use Transfer Appliance to migrate your data into a Google Kubernetes Engine cluster, and then configure a weekly transfer job.

C. Install Storage Transfer Service for on-premises data in your data center, and then configure a weekly transfer job. Most Voted

D. Install Storage Transfer Service for on-premises data on a Google Cloud virtual machine, and then configure a weekly transfer job.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by ducc at *Sept. 3, 2022, 3:57 a.m.*

Comments

zellck Highly Voted 8 months, 1 week ago

Selected Answer: C

C is the answer.

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#storage-transfer-service-for-large-transfers-of-on-premises-data
Like gsutil, Storage Transfer Service for on-premises data enables transfers from network file system (NFS) storage to Cloud Storage. Although gsutil can support small transfer sizes (up to 1 TB), Storage Transfer Service for on-premises data is designed for large-scale transfers (up to petabytes of data, billions of files).

upvoted 5 times

Prudvi3266 Most Recent 3 months, 2 weeks ago

Selected Answer: C

C is the Answer as we need weekly run Storage transfer service has the feature to schedule.

upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 195 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 195
Topic #: 1
[All Professional Data Engineer Questions]

Your company wants to be able to retrieve large result sets of medical information from your current system, which has over 10 TBs in the database, and store the data in new tables for further query. The database must have a low-maintenance architecture and be accessible via SQL. You need to implement a cost-effective solution that can support data analytics for large result sets. What should you do?

A. Use Cloud SQL, but first organize the data into tables. Use JOIN in queries to retrieve data.

B. Use BigQuery as a data warehouse. Set output destinations for caching large queries. Most Voted

C. Use a MySQL cluster installed on a Compute Engine managed instance group for scalability.

D. Use Cloud Spanner to replicate the data across regions. Normalize the data in a series of tables.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

by ducc at *Sept. 3, 2022, 3:56 a.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: B

B. Use BigQuery as a data warehouse. Set output destinations for caching large queries.
upvoted 7 times

AzureDP900 Most Recent 7 months, 1 week ago

B. Use BigQuery as a data warehouse. Set output destinations for caching large queries. Most Voted
upvoted 2 times

zellck 8 months, 1 week ago

Selected Answer: B

B is the answer.
upvoted 3 times

TNT87 10 months, 4 weeks ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 194 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 194
Topic #: 1
[All Professional Data Engineer Questions]

An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and call the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

A. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to the Python API. Most Voted

B. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic.

C. Write an application that makes a queue in a NoSQL database.

D. Use Cloud Composer to subscribe to a Pub/Sub topic and call the Python API.

Hide Answer

Suggested Answer: *D*

Community vote distribution

A (92%) 8%

by PhuocT at *Sept. 2, 2022, 7:58 p.m.*

Comments

lucaluca1982 4 months, 1 week ago
A and D are both good. I go for A because we have high volume and easy to scale and optmize cost
upvoted 1 times

musumusu 5 months, 2 weeks ago
Answer A:
assume, Company wants to buy immediately in same second if stock goes down or up.
Somehow, it is connected to PubSub as SINK connector, then immediately there is PUSH to subcriber (cloud function) that is connected to their python API (internal application) that makes the purchase.
upvoted 1 times

AzureDP900 7 months, 1 week ago
A. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to the Python API.
upvoted 1 times

zellck 8 months, 1 week ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 193 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 193
Topic #: 1
[All Professional Data Engineer Questions]

An aerospace company uses a proprietary data format to store its flight data. You need to connect this new data source to BigQuery and stream the data into
BigQuery. You want to efficiently import the data into BigQuery while consuming as few resources as possible. What should you do?

A. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source.

B. Use a standard Dataflow pipeline to store the raw data in BigQuery, and then transform the format later when the data is used.

C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format.

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format. Most Voted

Hide Answer

Suggested Answer: *B*

Community vote distribution

D (76%) B (24%)

by ducc at *Sept. 3, 2022, 3:52 a.m.*

Comments

beanz00 Highly Voted 9 months, 1 week ago
This has to be D. How could it even be B? The source is a proprietary format. Dataflow wouldn't have a built-in template to ead the file. You will have to create something custom.
upvoted 14 times

devaid Highly Voted 9 months, 3 weeks ago
Selected Answer: D
For me it's clearly D
It's between B and D, but read B, store raw data in Big Query? Use a Dataflow pipeline just to store raw data into Big Query, and transform later? You'd need to do another pipeline for that, and is not efficient.
upvoted 10 times

knith66 Most Recent 1 week, 3 days ago
Between B and D. Firstly transformation is not mentioned in the question, So B is less probable. Then Efficient import is mentioned in the question, Converting to Avro will consume less space. I am going with D
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 192 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 192
Topic #: 1
[All Professional Data Engineer Questions]

You are implementing a chatbot to help an online retailer streamline their customer service. The chatbot must be able to respond to both text and voice inquiries.
You are looking for a low-code or no-cade option, and you want to be able to easily train the chatbot to provide answers to keywords. What should you do?

A. Use the Cloud Speech-to-Text API to build a Python application in App Engine.

B. Use the Cloud Speech-to-Text API to build a Python application in a Compute Engine instance.

C. Use Dialogflow for simple queries and the Cloud Speech-to-Text API for complex queries.

D. Use Dialogflow to implement the chatbot, defining the intents based on the most common queries collected. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (85%) C (15%)

by PhuocT at *Sept. 2, 2022, 7:54 p.m.*

Comments

PhuocT Highly Voted 11 months, 1 week ago

Selected Answer: D

D is correct:
https://cloud.google.com/dialogflow/es/docs/how/detect-intent-tts#:~:text=Dialogflow%20can%20use%20Cloud%20Text,to%2Dspeech%2C%20or%20TTS.

upvoted 9 times

Lanro Most Recent 6 days, 19 hours ago

Selected Answer: D

Low-code or no-cade requirement makes it easy to decide.

upvoted 1 times

zellck 8 months, 1 week ago

Selected Answer: D

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 191 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 191
Topic #: 1
[All Professional Data Engineer Questions]

You are developing a new deep learning model that predicts a customer's likelihood to buy on your ecommerce site. After running an evaluation of the model against both the original training data and new test data, you find that your model is overfitting the data. You want to improve the accuracy of the model when predicting new data. What should you do?

A. Increase the size of the training dataset, and increase the number of input features.

B. Increase the size of the training dataset, and decrease the number of input features. Most Voted

C. Reduce the size of the training dataset, and increase the number of input features.

D. Reduce the size of the training dataset, and decrease the number of input features.

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (94%) 6%

by ducc at *Sept. 3, 2022, 3:46 a.m.*

Comments

John_Pongthorn Highly Voted 10 months, 2 weeks ago

Selected Answer: B

There 2 parts and they are relevant to each other
1. Overfit is fixed by decreasing the number of input features (select only essential features)
2. Accuracy is improved by increasing the amount of training data examples.

upvoted 7 times

John_Pongthorn 10 months, 2 weeks ago

https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

upvoted 2 times

NeoNitin Most Recent 1 week, 1 day ago

Increase the size of the training dataset: By adding more diverse examples of customers and their buying behavior to the training data, the model will have a broader understanding of different scenarios and be better equipped to generalize to new customers.

Increase the number of input features: Providing the model with more relevant information about customers can help it identify meaningful

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 190 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 190
Topic #: 1
[All Professional Data Engineer Questions]

You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues, including mismatched data types, such as STRINGs and
INT64s in the same column, and inconsistent formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?

A. Use Data Fusion to transform the data before loading it into BigQuery. Most Voted

B. Use Data Fusion to convert the CSV files to a self-describing data format, such as AVRO, before loading the data to BigQuery.

C. Load the CSV files into a staging table with the desired schema, perform the transformations with SQL, and then write the results to the final destination table.

D. Create a table with the desired schema, load the CSV files into the table, and perform the transformations in place using SQL.

Hide Answer

Suggested Answer: *D*

Community vote distribution

A (88%) 12%

by AWSandeep at *Sept. 2, 2022, 11:15 p.m.*

Comments

phidelics 1 month, 3 weeks ago

Selected Answer: A

Keyword: Data Pipeline

upvoted 1 times

mialll 3 months ago

Selected Answer: A

same as @saurabhsingh4k

upvoted 1 times

Adswerve 3 months, 3 weeks ago

Selected Answer: C

C is the right answer. Do ELT in BigQuery. Data Fusion is not the right too for this job.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 189 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 189
Topic #: 1
[All Professional Data Engineer Questions]

You need to migrate 1 PB of data from an on-premises data center to Google Cloud. Data transfer time during the migration should take only a few hours. You want to follow Google-recommended practices to facilitate the large data transfer over a secure connection. What should you do?

A. Establish a Cloud Interconnect connection between the on-premises data center and Google Cloud, and then use the Storage Transfer Service. Most Voted

B. Use a Transfer Appliance and have engineers manually encrypt, decrypt, and verify the data.

C. Establish a Cloud VPN connection, start gcloud compute scp jobs in parallel, and run checksums to verify the data.

D. Reduce the data into 3 TB batches, transfer the data using gsutil, and run checksums to verify the data.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (59%) B (41%)

by AWSandeep at *Sept. 2, 2022, 11:13 p.m.*

Comments

knith66 1 week, 3 days ago

Selected Answer: A

Dedicated Interconnect provides direct physical connections between your on-premises network and Google's network. Dedicated Interconnect enables you to transfer large amounts of data between networks, which can be more cost-effective than purchasing additional bandwidth over the public internet. https://cloud.google.com/network-connectivity/docs/interconnect/concepts/dedicated-overview

upvoted 1 times

knith66 1 week, 3 days ago

This link has additional clarity
https://cloud.google.com/network-connectivity/docs/interconnect/concepts/terminology

upvoted 1 times

vaga1 1 month ago

Selected Answer: B

1 PB and "few hours". It is clearly referring to Transfer Appliance

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 188 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 188
Topic #: 1
[All Professional Data Engineer Questions]

Your startup has a web application that currently serves customers out of a single region in Asia. You are targeting funding that will allow your startup to serve customers globally. Your current goal is to optimize for cost, and your post-funding goal is to optimize for global presence and performance. You must use a native
JDBC driver. What should you do?

A. Use Cloud Spanner to configure a single region instance initially, and then configure multi-region Cloud Spanner instances after securing funding. Most Voted

B. Use a Cloud SQL for PostgreSQL highly available instance first, and Bigtable with US, Europe, and Asia replication after securing funding.

C. Use a Cloud SQL for PostgreSQL zonal instance first, and Bigtable with US, Europe, and Asia after securing funding.

D. Use a Cloud SQL for PostgreSQL zonal instance first, and Cloud SQL for PostgreSQL with highly available configuration after securing funding.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (70%) D (30%)

by AWSandeep at *Sept. 2, 2022, 11:06 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: A

A. Use Cloud Spanner to configure a single region instance initially, and then configure multi-region Cloud Spanner instances after securing funding.

When you create a Cloud Spanner instance, you must configure it as either regional (that is, all the resources are contained within a single Google Cloud region) or multi-region (that is, the resources span more than one region).

You can change the instance configuration to multi-regional (or global) at anytime.

upvoted 8 times

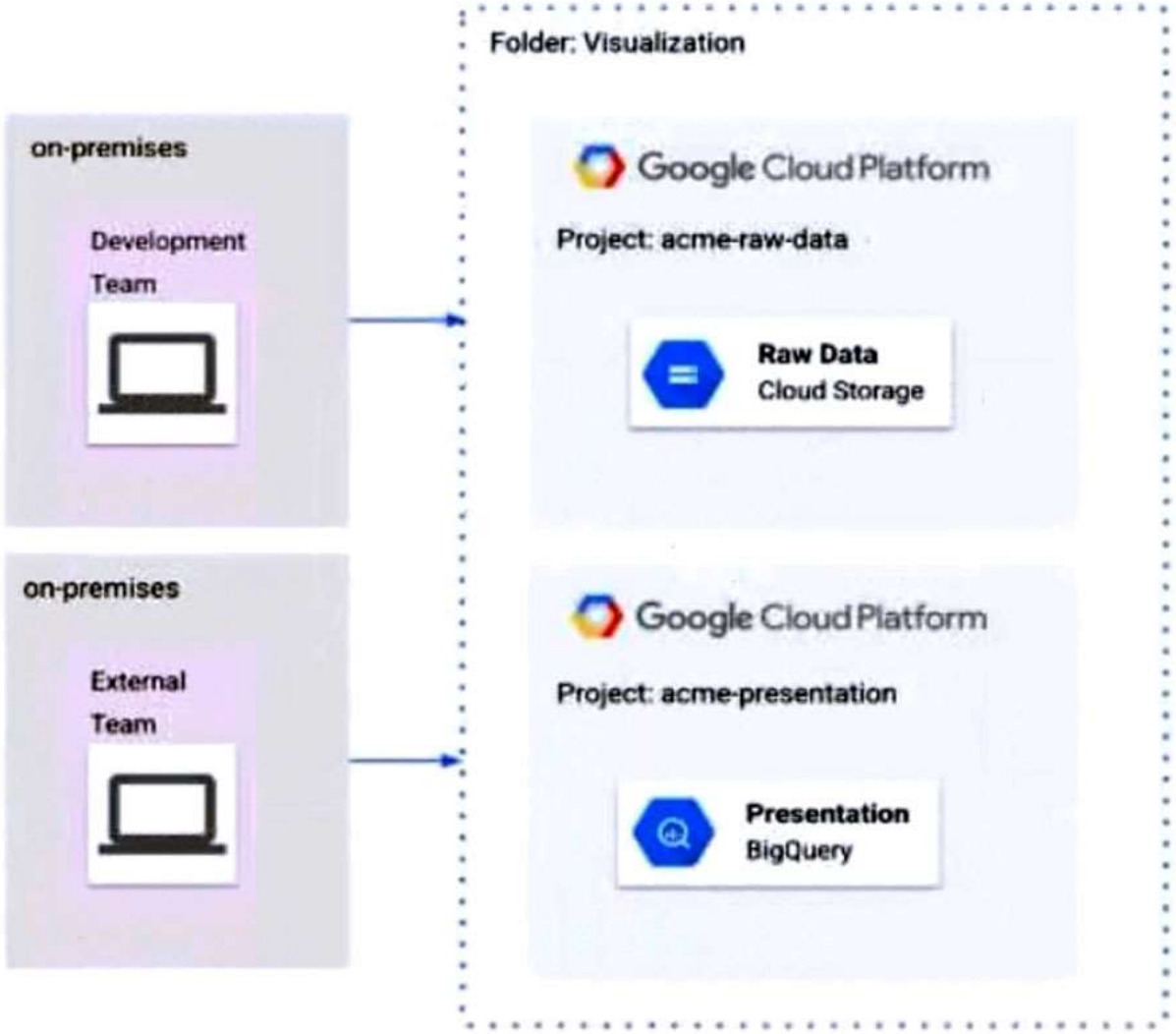
izekc Most Recent 3 months, 1 week ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 187 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 187
Topic #: 1
[All Professional Data Engineer Questions]

The Development and External teams have the project viewer Identity and Access Management (IAM) role in a folder named Visualization. You want the
Development Team to be able to read data from both Cloud Storage and BigQuery, but the External Team should only be able to read data from

BigQuery. What should you do?



A. Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project.

B. Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data project that deny all ingress traffic from the External Team CIDR range.

C. Create a VPC Service Controls perimeter containing both projects and BigQuery as a restricted API. Add the External Team users to the perimeter's Access Level. Most Voted

D. Create a VPC Service Controls perimeter containing both projects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter's Access Level. Most Voted

Hide Answer

Suggested Answer: *C*

Community vote distribution

D (81%) C (19%)

by AWSandeep at *Sept. 2, 2022, 11:02 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: D

D. Create a VPC Service Controls perimeter containing both projects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter's Access Level.

Reveal Solution

upvoted 14 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 186 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 186
Topic #: 1
[All Professional Data Engineer Questions]

Your new customer has requested daily reports that show their net consumption of Google Cloud compute resources and who used the resources. You need to quickly and efficiently generate these daily reports. What should you do?

A. Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user. Most Voted

B. Filter data in Cloud Logging by project, resource, and user; then export the data in CSV format.

C. Filter data in Cloud Logging by project, log type, resource, and user, then import the data into BigQuery.

D. Export Cloud Logging data to Cloud Storage in CSV format. Cleanse the data using Dataprep, filtering by project, resource, and user.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (72%) D (17%) 11%

by AWSandeep at *Sept. 2, 2022, 10:58 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

A. Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user.

You cannot import custom or filtered billing criteria into BigQuery. There are three types of Cloud Billing data tables with a fixed schema that must further drilled-down via BigQuery views.

Reference:
https://cloud.google.com/billing/docs/how-to/export-data-bigquery#setup

upvoted 6 times

vaga1 Most Recent 2 months, 3 weeks ago

Selected Answer: A

B, C, D do no generate a daily scalable solution.

upvoted 2 times

Siant_137 3 months, 1 week ago

Selected Answer: C

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 185 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 185
Topic #: 1
[All Professional Data Engineer Questions]

You issue a new batch job to Dataflow. The job starts successfully, processes a few elements, and then suddenly fails and shuts down. You navigate to the
Dataflow monitoring interface where you find errors related to a particular DoFn in your pipeline. What is the most likely cause of the errors?

A. Job validation
B. Exceptions in worker code Most Voted
C. Graph or pipeline construction
D. Insufficient permissions

Hide Answer

Suggested Answer: *D*

Community vote distribution
B (100%)

by AWSandeep at *Sept. 2, 2022, 10:47 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: B

B. Exceptions in worker code

While your job is running, you might encounter errors or exceptions in your worker code. These errors generally mean that the DoFns in your pipeline code have generated unhandled exceptions, which result in failed tasks in your Dataflow job.

Exceptions in user code (for example, your DoFn instances) are reported in the Dataflow monitoring interface.

Reference (Lists all answer choices and when to pick each one):
https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#Causes
upvoted 5 times

zellck Highly Voted 8 months, 1 week ago

Selected Answer: B

B is the answer.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 184 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 184
Topic #: 1
[All Professional Data Engineer Questions]

You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API. Following Google's best practices, you have both a staging and a production table for the data. How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging.

B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging.

C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours. Most Voted

D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes.

Hide Answer

Suggested Answer: *A*

Community vote distribution

C (82%) Other

by AWSandeep at *Sept. 2, 2022, 10:36 p.m.*

Comments

NicolasN Highly Voted 9 months ago

Selected Answer: C

[C]
I found the correct answer based on a real case, where Google's Solutions Architect team decided to move an internal process to use BigQuery.
The related doc is here: https://cloud.google.com/blog/products/data-analytics/moving-a-publishing-workflow-to-bigquery-for-new-data-insights
upvoted 13 times

NicolasN 9 months ago

The interesting excerpts:
"Following common extract, transform, load (ETL) best practices, we used a staging table and a separate production table so that we could load data into the staging table without impacting users of the data. The design we created based on ETL best practices called for first deleting all

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 183 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 183
Topic #: 1
[All Professional Data Engineer Questions]

You are using Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in. How should you design your row key and tables to ensure that you can access the data with the simplest query?

A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design. Most Voted

B. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design. Most Voted

C. For each index, have a separate table and use a timestamp as the row key design.

D. For each index, have a separate table and use a reverse timestamp as the row key design.

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (49%) A (47%) 4%

by AWSandeep at *Sept. 2, 2022, 10:25 p.m.*

Comments

John_Pongthorn Highly Voted 10 months, 2 weeks ago

This is special case , plese Take a look carefully the below link and read at last paragraph at the bottom of this comment, let everyone share idea, We will go with B, C
https://cloud.google.com/bigtable/docs/schema-design#time-based

Don't use a timestamp by itself or at the beginning of a row key, because this will cause sequential writes to be pushed onto a single node, creating a hotspot.

If you usually retrieve the most recent records first, you can use a reversed timestamp in the row key by subtracting the timestamp from your programming language's maximum value for long integers (in Java, java.lang.Long.MAX_VALUE). With a reversed timestamp, the records will be ordered from most recent to least recent.

upvoted 9 times

Mcloudgirl 9 months ago

I agree, based on the docs, B. Leading with a non-reversed timestamp will lead to hotspotting, reversed is the way to go.

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 182 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 182
Topic #: 1
[All Professional Data Engineer Questions]

You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center. Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB.
Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time. What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

A. Storage Transfer Service for the migration; Pub/Sub and Cloud Data Fusion for the real-time updates

B. BigQuery Data Transfer Service for the migration; Pub/Sub and Dataproc for the real-time updates

C. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates Most Voted

D. gsutil for both the migration and the real-time updates

Hide Answer

Suggested Answer: *B*

Community vote distribution

C (100%)

by AWSandeep at *Sept. 2, 2022, 9:55 p.m.*

Comments

zellck Highly Voted 8 months, 1 week ago

Selected Answer: C

C is the answer.

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil_for_smaller_transfers_of_on-premises_data
The gsutil tool is the standard tool for small- to medium-sized transfers (less than 1 TB) over a typical enterprise-scale network, from a private data center to Google Cloud.

upvoted 8 times

musumusu 5 months, 2 weeks ago

what is wrong with A, there is no cost constraint

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 181 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 181
Topic #: 1
[All Professional Data Engineer Questions]

You need to give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID. This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline. There will be tens of thousands of messages per second and that can be multi-threaded. and you worry about the backpressure on the system. How should you design your pipeline to minimize that backpressure?

A. Call out to the service via HTTP.

B. Create the pipeline statically in the class definition.

C. Create a new object in the startBundle method of DoFn.

D. Batch the job into ten-second increments. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (91%) 9%

by ducc at *Sept. 4, 2022, 3:19 a.m.*

Comments

John_Pongthorn Highly Voted 10 months, 1 week ago

Selected Answer: D

D: I have insisted on this choice all aling.
please read find the keyword massive backpressure
https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1

if the call takes on average 1 sec, that would cause massive backpressure on the pipeline. In these circumstances you should consider batching these requests, instead.

upvoted 12 times

NicolasN 9 months ago

Thanks for sharing, you found exactly the same problem!
The document defitely proposes batching for this scenario.

I'm quoting another part from the same example that would be useful for a similar question with different conditions:
- If you're using a client in the DoFn that has heavy instantiation steps, rather than create that object in each DoFn call:

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 180 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 180
Topic #: 1
[All Professional Data Engineer Questions]

You are migrating an application that tracks library books and information about each book, such as author or year published, from an on-premises data warehouse to BigQuery. In your current relational database, the author information is kept in a separate table and joined to the book information on a common key. Based on Google's recommended practice for schema design, how would you structure the data to ensure optimal speed of queries about the author of each book that has been borrowed?

A. Keep the schema the same, maintain the different tables for the book and each of the attributes, and query as you are doing today.

B. Create a table that is wide and includes a column for each attribute, including the author's first name, last name, date of birth, etc.

C. Create a table that includes information about the books and authors, but nest the author fields inside the author column. Most Voted

D. Keep the schema the same, create a view that joins all of the tables, and always query the view.

Hide Answer

Suggested Answer: *D*

Community vote distribution

C (100%)

by AWSandeep at *Sept. 2, 2022, 9:22 p.m.*

Comments

musumusu Highly Voted 5 months, 2 weeks ago
C
if data is time based or sequential, find partition and cluster option
if data is not time based,
always look for denomalize / nesting option.
upvoted 5 times

AzureDP900 Most Recent 7 months, 1 week ago
C. Create a table that includes information about the books and authors, but nest the author fields inside the author column.
upvoted 1 times

zellck 8 months, 1 week ago
Selected Answer: C
C is the answer.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 179 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 179
Topic #: 1
[All Professional Data Engineer Questions]

You are building a real-time prediction engine that streams files, which may contain PII (personal identifiable information) data, into Cloud Storage and eventually into BigQuery. You want to ensure that the sensitive data is masked but still maintains referential integrity, because names and emails are often used as join keys.
How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the PII data is not accessible by unauthorized individuals?

A. Create a pseudonym by replacing the PII data with cryptogenic tokens, and store the non-tokenized data in a locked-down button. Most Voted

B. Redact all PII data, and store a version of the unredacted data in a locked-down bucket.

C. Scan every table in BigQuery, and mask the data it finds that has PII.

D. Create a pseudonym by replacing PII data with a cryptographic format-preserving token. Most Voted

Hide Answer

Suggested Answer: *B*

Community vote distribution

D (55%) A (39%) 6%

by AWSandeep at *Sept. 2, 2022, 9:18 p.m.*

Comments

cetanx 1 month, 2 weeks ago

Selected Answer: B

I've also asked to GPT but I had to remind the hard condition "names and emails are often used as join keys".
It changed the answer to "B" after 3rd iteration.

masking all PII data may not satisfy the requirement of using names and emails as join keys, as the data is obfuscated and cannot be used for accurate join operations.

In this approach, you would redact or remove the sensitive PII data, such as names and emails, from the dataset that will be used for real-time processing and analysis. The redacted data would be stored in the primary dataset to ensure that sensitive information is not accessible.

Additionally, you would create a copy of the original dataset with the PII data still intact, but this copy would be stored in a locked-down bucket with restricted access. This ensures that authorized individuals who need access to the unredacted data for specific purposes, such as join operations, can retrieve it from the secured location.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 178 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 178
Topic #: 1
[All Professional Data Engineer Questions]

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using
SideInputs to join data. You noticed that the pipeline is taking longer to complete than expected; what should you do to expedite the Dataflow job?

A. Switch to compressed Avro files.

B. Reduce the batch size.

C. Retry records that throw an error.

D. Use CoGroupByKey instead of the SideInput. Most Voted

Hide Answer

Suggested Answer: *C*

Community vote distribution

D (81%) Other

by AWSandeep at *Sept. 2, 2022, 9:02 p.m.*

Comments

John_Pongthorn Highly Voted 10 months, 2 weeks ago

Selected Answer: D

D: it is most likely.
There are a lot of reference doc to tell about comparison between them
https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-testing#choose_correctly_between_side_inputs_or_cogroupbykey_for_joins

https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-2

https://stackoverflow.com/questions/58080383/sideinput-i-o-kills-performance

upvoted 14 times

zellck Highly Voted 8 months, 1 week ago

Selected Answer: D

D is the answer.

https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 177 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 177
Topic #: 1
[All Professional Data Engineer Questions]

You want to rebuild your batch pipeline for structured data on Google Cloud. You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run. To expedite development and pipeline run time, you want to use a serverless tool and SOL syntax. You have already moved your raw data into Cloud Storage. How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

A. Convert your PySpark commands into SparkSQL queries to transform the data, and then run your pipeline on Dataproc to write the data into BigQuery.

B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated quenes from BigQuery for machine learning.

C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table. Most Voted

D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery.

Hide Answer

Suggested Answer: *D*

Community vote distribution

C (84%) A (16%)

by AWSandeep at *Sept. 2, 2022, 8:30 p.m.*

Comments

devaid Highly Voted 10 months ago

Selected Answer: C

The question is C but not because the SQL Syntax, as you can perfectly use SparkSQL on Dataproc reading files from GCS. It's because the "serverless" requirement.

upvoted 8 times

MoeHaydar Most Recent 3 weeks, 6 days ago

Selected Answer: C

Note: Dataproc by itself is not serverless
https://cloud.google.com/dataproc-serverless/docs/overview

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 176 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 176
Topic #: 1
[All Professional Data Engineer Questions]

You have uploaded 5 years of log data to Cloud Storage. A user reported that some data points in the log data are outside of their expected ranges, which indicates errors. You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Import the data from Cloud Storage into BigQuery. Create a new BigQuery table, and skip the rows with errors.

B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage. Skip the rows with errors.

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage. Most Voted

D. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by PhuocT at *Sept. 2, 2022, 7:15 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: C

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.

You can't filter out data using BQ load commands. You must imbed the logic to filter out data (i.e. time ranges) in another decoupled way (i.e. Dataflow, Cloud Functions, etc.). Therefore, A and B add additional complexity and deviates from the Data Lake design paradigm. D is wrong as the question strictly implies that the existing data set needs to be retained for compliance.

upvoted 8 times

AzureDP900 Most Recent 7 months, 1 week ago

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 175 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 175
Topic #: 1
[All Professional Data Engineer Questions]

Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model. You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days. Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

A. Denormalize the data.

B. Shard the data by customer ID.

C. Materialize the dimensional data in views.

D. Partition the data by transaction date. Most Voted

Hide Answer

Suggested Answer: *A*
Reference:
https://cloud.google.com/architecture/dw2bq/dw-bq-schema-and-data-transfer-overview

Community vote distribution
D (88%) 12%

by PhuocT at *Sept. 2, 2022, 7:10 p.m.*

Comments

waiebdi Highly Voted 5 months, 3 weeks ago

Selected Answer: D
D is the right answer because it does not increase storage costs.
A is not correct because denormalization typically increases the amount of storage needed.
upvoted 7 times

vamgcp Most Recent 1 week, 6 days ago

Selected Answer: D
Option D - BigQuery supports partitioned tables, where the data is divided into smaller, manageable portions based on a chosen column (e.g., transaction date). By partitioning the data based on the transaction date, BigQuery can efficiently query only the relevant partitions that contain data for the past 30 days, reducing the amount of data that needs to be scanned.Partitioning does not increase storage costs. It organizes existing data in a more structured manner, allowing for better query performance without any additional storage expenses.
upvoted 1 times

Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 174 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 174
Topic #: 1
[All Professional Data Engineer Questions]

You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app. You have reviewed old chat logs and tagged each conversation for intent based on each customer's stated intention for contacting customer service. About 70% of customer requests are simple requests that are solved within 10 intents. The remaining 30% of inquiries require much longer, more complicated requests. Which intents should you automate first?

A. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests. Most Voted

B. Automate the more complicated requests first because those require more of the agents' time.

C. Automate a blend of the shortest and longest intents to be representative of all intents.

D. Automate intents in places where common words such as 'payment' appear only once so the software isn't confused.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by AWSandeep at *Sept. 4, 2022, 10:50 p.m.*

Comments

vamgcp 1 week, 6 days ago

Selected Answer: A

Option A : : By automating the intents that cover a significant majority (70%) of customer requests, you target the areas with the highest volume of interactions. This helps reduce the load on live agents, enabling them to focus on more complicated and time-consuming inquiries that require their expertise.

upvoted 1 times

Takshashila 1 month, 3 weeks ago

Selected Answer: A

A is the answer.

upvoted 1 times

zellck 8 months, 1 week ago

Selected Answer: A

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 173 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 173
Topic #: 1
[All Professional Data Engineer Questions]

You are designing a pipeline that publishes application events to a Pub/Sub topic. Although message ordering is not important, you need to be able to aggregate events across disjoint hourly intervals before loading the results to BigQuery for analysis. What technology should you use to process and load this data to
BigQuery while ensuring that it will scale with large volumes of events?

A. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

B. Schedule a Cloud Function to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.

C. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.

D. Create a streaming Dataflow job that reads continually from the Pub/Sub topic and performs the necessary aggregations using tumbling windows. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by AWSandeep at *Sept. 2, 2022, 7:51 p.m.*

Comments

Atnafu Highly Voted 7 months, 3 weeks ago

D

TUMBLE=> fixed windows.
HOP=> sliding windows.
SESSION=> session windows.

upvoted 9 times

vamgcp Most Recent 1 week, 6 days ago

We can use TUMBLE(1 HOUR) to create hourly windows, where each window contains events from a specific hour.

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 172 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 172
Topic #: 1
[All Professional Data Engineer Questions]

You are analyzing the price of a company's stock. Every 5 seconds, you need to compute a moving average of the past 30 seconds' worth of data. You are reading data from Pub/Sub and using DataFlow to conduct the analysis. How should you set up your windowed pipeline?

A. Use a fixed window with a duration of 5 seconds. Emit results by setting the following trigger:
AfterProcessingTime.pastFirstElementInPane().plusDelayOf (Duration.standardSeconds(30))

B. Use a fixed window with a duration of 30 seconds. Emit results by setting the following trigger:
AfterWatermark.pastEndOfWindow().plusDelayOf (Duration.standardSeconds(5))

C. Use a sliding window with a duration of 5 seconds. Emit results by setting the following trigger:
AfterProcessingTime.pastFirstElementInPane().plusDelayOf (Duration.standardSeconds(30))

D. Use a sliding window with a duration of 30 seconds and a period of 5 seconds. Emit results by setting the following trigger:
AfterWatermark.pastEndOfWindow () Most Voted

Hide Answer

Suggested Answer: *B*

Community vote distribution

D (100%)

by AWSandeep at *Sept. 2, 2022, 7:48 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: D

D. Use a sliding window with a duration of 30 seconds and a period of 5 seconds. Emit results by setting the following trigger:
AfterWatermark.pastEndOfWindow ()
Reveal Solution

upvoted 7 times

vamgcp Most Recent 1 week, 6 days ago

Selected Answer: D

Option D: Sliding Window: Since you need to compute a moving average of the past 30 seconds' worth of data every 5 seconds, a sliding window is appropriate. A sliding window allows overlapping intervals and is well-suited for computing rolling aggregates.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 171 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 171
Topic #: 1
[All Professional Data Engineer Questions]

You work for a large real estate firm and are preparing 6 TB of home sales data to be used for machine learning. You will use SQL to transform the data and use
BigQuery ML to create a machine learning model. You plan to use the model for predictions against a raw dataset that has not been transformed. How should you set up your workflow in order to prevent skew at prediction time?

A. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data. Most Voted

B. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. Before requesting predictions, use a saved query to transform your raw input data, and then use ML.EVALUATE.

C. Use a BigQuery view to define your preprocessing logic. When creating your model, use the view as your model training data. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

D. Preprocess all data using Dataflow. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any further transformations on the input data.

Hide Answer

Suggested Answer: *B*

Community vote distribution

A (96%) 4%

by AWSandeep at *Sept. 2, 2022, 7:44 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: A

A. When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps. At prediction time, use BigQuery's ML.EVALUATE clause without specifying any transformations on the raw input data.

Using the TRANSFORM clause, you can specify all preprocessing during model creation. The preprocessing is automatically applied during the prediction and evaluation phases of machine learning.

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

upvoted 9 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 170 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 170
Topic #: 1
[All Professional Data Engineer Questions]

You are updating the code for a subscriber to a Pub/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. Your subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?

A. Set up the Pub/Sub emulator on your local machine. Validate the behavior of your new subscriber logic before deploying it to production.

B. Create a Pub/Sub snapshot before deploying new subscriber code. Use a Seek operation to re-deliver messages that became available after the snapshot was created. Most Voted

C. Use Cloud Build for your deployment. If an error occurs after deployment, use a Seek operation to locate a timestamp logged by Cloud Build at the start of the deployment.

D. Enable dead-lettering on the Pub/Sub topic to capture messages that aren't successfully acknowledged. If an error occurs after deployment, re-deliver any messages captured by the dead-letter queue.

Hide Answer

Suggested Answer: *C*
Reference:

Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 169 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 169
Topic #: 1
[All Professional Data Engineer Questions]

You are migrating a table to BigQuery and are deciding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID, and the city and state in which the store is located. You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state, city, and individual store. How would you model this table for the best query performance?

A. Partition by transaction time; cluster by state first, then city, then store ID. Most Voted
B. Partition by transaction time; cluster by store ID first, then city, then state.
C. Top-level cluster by state first, then city, then store ID.
D. Top-level cluster by store ID first, then city, then state.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (94%) 6%

by ducc at *Sept. 3, 2022, 7:10 a.m.*

Comments

AWSandeep Highly Voted 11 months ago

Selected Answer: A

A. Partition by transaction time; cluster by state first, then city, then store ID.
upvoted 8 times

Atnafu Highly Voted 7 months, 3 weeks ago

A
Partitioning is obvious
Clustering is already mentioned in the question
past 30 days and to look at purchasing trends by
state,
city, and
individual store
upvoted 5 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 168 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 168
Topic #: 1
[All Professional Data Engineer Questions]

You work for a financial institution that lets customers register online. As new customers register, their user data is sent to Pub/Sub before being ingested into
BigQuery. For security reasons, you decide to redact your customers' Government issued Identification Number while allowing customer service representatives to view the original values when necessary. What should you do?

A. Use BigQuery's built-in AEAD encryption to encrypt the SSN column. Save the keys to a new table that is only viewable by permissioned users.

B. Use BigQuery column-level security. Set the table permissions so that only members of the Customer Service user group can see the SSN column. Most Voted

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

D. Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (55%) B (39%) 5%

by AWSandeep at *Sept. 2, 2022, 7:01 p.m.*

Comments

AWSandeep Highly Voted 11 months, 1 week ago

Selected Answer: B

B. While C and D are intriguing, they don't specify how to enable customer service representatives to receive access to the encryption token.
upvoted 9 times

Lanro Most Recent 6 days, 21 hours ago

Selected Answer: D

I don't see why we should use DLP since we know exactly the column that should be locked or encrypted. On the other hand having a cryptographic representation of SSN helps to aggregate/analyse entries. So I will vote for D, but B is much more easy to implement. Garbage question indeed.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 167 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 167
Topic #: 1
[All Professional Data Engineer Questions]

Your company currently runs a large on-premises cluster using Spark, Hive, and HDFS in a colocation facility. The cluster is designed to accommodate peak usage on the system; however, many jobs are batch in nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more serverless offerings in order to take advantage of the cloud. Because of the timing of their contract renewal with the colocation facility, they have only 2 months for their initial migration. How would you recommend they approach their upcoming migration strategy so they can maximize their cost savings in the cloud while still executing the migration in time?

A. Migrate the workloads to Dataproc plus HDFS; modernize later.

B. Migrate the workloads to Dataproc plus Cloud Storage; modernize later. Most Voted

C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery.

D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery.

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (87%) 13%

by AWSandeep at *Sept. 2, 2022, 6:57 p.m.*

Comments

zellck Highly Voted 8 months, 1 week ago

Selected Answer: B

B is the answer.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview
When you want to move your Apache Spark workloads from an on-premises environment to Google Cloud, we recommend using Dataproc to run Apache Spark/Apache Hadoop clusters. Dataproc is a fully managed, fully supported service offered by Google Cloud. It allows you to separate storage and compute, which helps you to manage your costs and be more flexible in scaling your workloads.

https://cloud.google.com/bigquery/docs/migration/hive#data_migration
Migrating Hive data from your on-premises or other cloud-based source cluster to BigQuery has two steps:

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 166 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 166
Topic #: 1
[All Professional Data Engineer Questions]

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.

B. Implement clustering in BigQuery on the package-tracking ID column. Most Voted

C. Tier older data onto Cloud Storage files and create a BigQuery table using Cloud Storage as an external data source.

D. Re-create the table using data partitioning on the package delivery date.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (75%) D (25%)

by AWSandeep at *Sept. 4, 2022, 10:50 p.m.*

Comments

sdi_studiers 1 month, 3 weeks ago

Selected Answer: D

I vote D
Queries to analyze the package lifecycle will cross partitions when using ingest date. Changing this to delivery date will allow a query to full a package's full lifecycle in a single partition.

upvoted 2 times

zellck 8 months, 1 week ago

Selected Answer: B

B is the answer.

https://cloud.google.com/bigquery/docs/clustered-tables
Clustered tables in BigQuery are tables that have a user-defined column sort order using clustered columns. Clustered tables can improve query performance and reduce query costs.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 165 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 165
Topic #: 1
[All Professional Data Engineer Questions]

You work for a large bank that operates in locations throughout North America. You are setting up a data storage system that will handle bank account transactions. You require ACID compliance and the ability to access data with SQL. Which solution is appropriate?

A. Store transaction data in Cloud Spanner. Enable stale reads to reduce latency.

B. Store transaction in Cloud Spanner. Use locking read-write transactions. Most Voted

C. Store transaction data in BigQuery. Disabled the query cache to ensure consistency.

D. Store transaction data in Cloud SQL. Use a federated query BigQuery for analysis. Most Voted

Hide Answer

Suggested Answer: *C*

Community vote distribution

B (70%) D (27%) 3%

by ducc at *Sept. 3, 2022, 1:19 a.m.*

Comments

devaid Highly Voted 10 months ago

Selected Answer: B

I'd say B as the documentation primarily says ACID compliance for Spanner, not Cloud SQL.
https://cloud.google.com/blog/topics/developers-practitioners/your-google-cloud-database-options-explained
Also, spanner supports read-write transactions for use cases, as handling bank transactions:
https://cloud.google.com/spanner/docs/transactions#read-write_transactions

upvoted 8 times

Jay_Krish 8 months, 2 weeks ago

I wonder if you understood the meaning of ACID. This is an inherent property of any relational DB. Cloud SQL is fully ACID complaint

upvoted 8 times

AzureDP900 7 months, 1 week ago

B is right

upvoted 1 times

juliobs Highly Voted 4 months, 2 weeks ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 164 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 164
Topic #: 1
[All Professional Data Engineer Questions]

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component. In order to train and serve the model, your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Create a new view with BigQuery that does not include a column with city information.

B. Use SQL in BigQuery to transform the state column using a one-hot encoding method, and make each city a column with binary values. Most Voted

C. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file and upload that as part of your model to BigQuery ML.

D. Use Cloud Data Fusion to assign each city to a region that is labeled as 1, 2, 3, 4, or 5, and then use that number to represent the city in the model. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (53%) D (47%)

by AWSandeep at *Sept. 2, 2022, 6:25 p.m.*

Comments

cajica Highly Voted 6 months ago

Selected Answer: D

If we're rigorous, as we should because it's a professional exam, I think option B is incorrect because it's one-hot-encoding the "state" column, if the answer was "city" column, then I'd go for B. As this is not the case and I do not accept an spelling error like this in an official question, I would go for D.

upvoted 7 times

knith66 1 week, 4 days ago

you are right, OHE is mentioned for state in option B, but in option B it is also mentioned to use binary conversion for the city column. an additional method can be used which is applicable for the conversion.

upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 163 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 163
Topic #: 1
[All Professional Data Engineer Questions]

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?

A. Export the information to Cloud Stackdriver, and set up an Alerting policy

B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver

C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs

D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Hide Answer

Suggested Answer: *B*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 6:04 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer : A
upvoted 20 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Monitoring does not only provide you with access to Dataflow-related metrics, but also lets you to create alerting policies and dashboards so you can chart time series of metrics and choose to be notified when these metrics reach specified values.
upvoted 12 times

zellck Most Recent 8 months, 1 week ago
Selected Answer: A
A is the answer.
upvoted 1 times

rbeeraka 1 year, 5 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 162 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 162
Topic #: 1
[All Professional Data Engineer Questions]

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

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

B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket. Manually destroy the key previously used for encryption, and rotate the key once.

C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.

D. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

Hide Answer

Suggested Answer: *B*

Community vote distribution

A (72%) D (28%)

by rickywck at *March 18, 2020, 2:11 a.m.*

Comments

dhs227 Highly Voted 3 years, 4 months ago

The correct answer must be D
A and B can be eliminated immediately since kms generated keys are considered potentially accessible by CSP.
C is incorrect because memory store is essentially a cache service.

Additional authenticated data (AAD) acts as a "salt", it is not a cipher.

upvoted 35 times

mikey007 2 years, 11 months ago

AAD is bound to the encrypted data, because you cannot decrypt the ciphertext unless you know the AAD, but it is not stored as part of the ciphertext. AAD also does not increase the cryptographic strength of the ciphertext. Instead it is an additional check by Cloud KMS to authenticate a decryption request.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 161 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 161
Topic #: 1
[All Professional Data Engineer Questions]

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a table in BigQuery, and append the new samples for CPU and memory to the table

B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second

C. Create a narrow table in Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second Most Voted

D. Create a wide table in Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (86%) 14%

by madhu1171 at *March 15, 2020, 7:41 p.m.*

Comments

psu Highly Voted 3 years, 3 months ago

Answer C

A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data.

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

https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design

upvoted 30 times

AzureDP900 7 months, 1 week ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 160 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 160
Topic #: 1
[All Professional Data Engineer Questions]

You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20
TB in size. Which database should you choose?

A. Cloud SQL Most Voted
B. Cloud Bigtable
C. Cloud Spanner
D. Cloud Datastore

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 7:43 a.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
A. Cloud SQL (30TB)
upvoted 31 times

dagoat 1 year, 10 months ago
65 TB now in Sept 2021
upvoted 13 times

[Removed] 1 year, 7 months ago
https://cloud.google.com/sql/docs/quotas#storage_limits
upvoted 1 times

vindahake 3 years, 2 months ago
Up to 30,720 GB, depending on the machine type. This looks like correct choice.
https://cloud.google.com/sql/docs/quotas#fixed-limits
upvoted 7 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 159 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 159
Topic #: 1
[All Professional Data Engineer Questions]

You need to choose a database for a new project that has the following requirements:
- Fully managed
- Able to automatically scale up
- Transactionally consistent
- Able to scale up to 6 TB
- Able to be queried using SQL

Which database do you choose?

A. Cloud SQL Most Voted
B. Cloud Bigtable
C. Cloud Spanner
D. Cloud Datastore

Hide Answer

Suggested Answer: *C*

Community vote distribution
A (58%) C (42%)

by [deleted] at *March 22, 2020, 7:42 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Correct: A
It asks for scaling up which can be done in cloud sql, horizontal scaling is not possible in cloud sql
Automatic storage increase
If you enable this setting, Cloud SQL checks your available storage every 30 seconds. If the available storage falls below a threshold size, Cloud SQL automatically adds additional storage capacity. If the available storage repeatedly falls below the threshold size, Cloud SQL continues to add storage until it reaches the maximum of 30 TB.
upvoted 29 times

Rajuuu 3 years, 1 month ago
C:- Cloud SQL is not fully managed as that is one of the requirement.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 158 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 158
Topic #: 1
[All Professional Data Engineer Questions]

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

A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master

B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet

C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter Most Voted

D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role

Hide Answer

Suggested Answer: *D*

Community vote distribution

C (100%)

by rickywck at *March 18, 2020, 1:46 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Correct: C

If you create a Dataproc cluster with internal IP addresses only, 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 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.
upvoted 33 times

AzureDP900 7 months, 1 week ago
Thank you for detailed explanation. C is right
upvoted 1 times

rickywck Highly Voted 3 years, 4 months ago
Should be C:

https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions
upvoted 10 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 157 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 157
Topic #: 1
[All Professional Data Engineer Questions]

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

A. Perform hyperparameter tuning Most Voted

B. Train a classifier with deep neural networks, because neural networks would always beat SVMs

C. Deploy the model and measure the real-world AUC; it's always higher because of generalization

D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

Hide Answer

Suggested Answer: *D*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 7:33 a.m.*

Comments

aadaisme Highly Voted 3 years, 1 month ago
Seems to be A. Preprocessing/scaling should be done with input features, instead of predictions (output)
upvoted 41 times

FARR Highly Voted 2 years, 11 months ago
A
Deep LEarning is not always the best solution
D talks about fudgin the output which is wrong
upvoted 11 times

vaga1 Most Recent 2 months, 4 weeks ago
Selected Answer: A
B,C are simply not true. D is modifing the scoring, making it not realiable anymore. A makes sense, is potentially increasing the model accuracy.
upvoted 1 times

rishu2 3 months ago
Selected Answer: A

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 156 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 156
Topic #: 1
[All Professional Data Engineer Questions]

You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?

A. Deploy a Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

B. Deploy a Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://

D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 7:31 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Correct: A

Ask for cost effective so persistent disk are HDD which are cheaper in comparison to SSD.
upvoted 29 times

[Removed] Highly Voted 3 years, 4 months ago
Confused between A and B. For r/w intensive jobs need to use SSDs. But questions doesnt state anything about the nature of the jobs. So better to start with a default option.
Choose A
upvoted 14 times

baubaumiaomiao 1 year, 7 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 155 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 155
Topic #: 1
[All Professional Data Engineer Questions]

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:
- The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured
- Support for publish/subscribe semantics on hundreds of topics
- Retain per-key ordering

Which system should you choose?

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

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (95%) 5%

by [deleted] at *March 22, 2020, 8:23 a.m.*

Comments

Blobby Highly Voted 1 year, 10 months ago
C in Sept 2021, A back in 2019/20?
ycombinator post from @daghayeghi is May 2018 and it looks like pub/sub now has the missing functionality;
1. key per-key ordering; https://www.youtube.com/watch?v=S2evHtbl4F8
2. replay functionality can be configured beyond the 7 day default; https://cloud.google.com/pubsub/docs/replay-overview
upvoted 20 times

shroffshivangi 1 year, 9 months ago
it still cannot be configured beyond 7 days, so the answer should be A
upvoted 5 times

MarcoDipa 1 year, 7 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 154 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 154
Topic #: 1
[All Professional Data Engineer Questions]

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region. Most Voted

B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.

C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.

D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (57%) B (43%)

by madhu1171 at *March 15, 2020, 7:17 p.m.*

Comments

madhu1171 Highly Voted 3 years, 4 months ago
A should be correct answer
upvoted 29 times

tycho 1 year, 7 months ago
yes A is correct, whe creating ne cloud sql instance there is an option
"Multiple zones (Highly available)
Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost."
upvoted 2 times

[Removed] Highly Voted 3 years, 4 months ago
Correct: A

https://cloud.google.com/sql/docs/mysql/high-availability
upvoted 13 times

wan2three Most Recent 3 weeks ago
Selected Answer: B

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 153 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 153
Topic #: 1
[All Professional Data Engineer Questions]

You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?

A. Consume the stream of data in Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages. Most Voted

B. Consume the stream of data in Dataflow using Kafka IO. Set a fixed time window of 1 hour. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

C. Use Kafka Connect to link your Kafka message queue to Pub/Sub. Use a Dataflow template to write your messages from Pub/Sub to Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Bigtable in the last hour. If that number falls below 4000, send an alert.

D. Use Kafka Connect to link your Kafka message queue to Pub/Sub. Use a Dataflow template to write your messages from Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 8:12 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be A
upvoted 26 times

[Removed] Highly Voted 3 years, 4 months ago
Correct: A

Dataflow can connect with Kafka and sliding window is used for taking averages
upvoted 16 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 152 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 152
Topic #: 1
[All Professional Data Engineer Questions]

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

A. BigQuery

B. Cloud Bigtable

C. Cloud Datastore

D. Cloud SQL for PostgreSQL

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 8:05 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Geospatial and ML functionality is with bigquery
upvoted 21 times

[Removed] Highly Voted 3 years, 4 months ago
Answer : A
upvoted 15 times

musumusu Most Recent 5 months, 2 weeks ago
Answer B: BigTable,
Catchup words: Telemetry (sensor- semi structured data) as data is bigger than 500GB, datastore is not a good option.
GEOJSON , bigquery has geospatical capabilites but still not quick enough for semi structure geojson data.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 151 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 151
Topic #: 1
[All Professional Data Engineer Questions]

You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be closing soon, so a rapid lift-and-shift migration is necessary. However, the data you've been using will be migrated to migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Cloud ML Engine

C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Hide Answer

Suggested Answer: *A*

Community vote distribution

C (100%)

by jvg637 at *March 18, 2020, 9:25 a.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
Correct C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
upvoted 24 times

[Removed] Highly Voted 3 years, 4 months ago
Correct: C

A Cloud Dataproc cluster has the Spark components, including Spark ML, installed.
upvoted 12 times

Suriyajan Most Recent 4 months ago
C
Option C . Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
upvoted 1 times