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.

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

Show Suggested Answer

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

Comments

Type your comment...

Submit

muhusman Highly Voted 2 years ago

Therefore, the correct option is 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.

Option A is incorrect because Cloud Scheduler is not designed for data transfer, but rather for scheduling the execution of Cloud Functions, Cloud Run, or App Engine applications.

Option B is incorrect because Transfer Appliance is designed for large-scale data transfers from on-premises environments to Google Cloud and is not suitable for transferring data on a daily basis.

Option D is also incorrect because the BigQuery Data Transfer Service dataset copy feature is designed for copying datasets between BigQuery projects and not suitable for copying data from on-premises environments to BigQuery.

upvoted 10 times

datapassionate 1 year, 3 months ago

With BigQuery Data Transfer Service we can copy files not only from other BigQuery, but also a bunch of cloud services listed here:
https://cloud.google.com/bigquery/docs/dts-introduction
But you are right. It wont work with on-premises.

upvoted 1 times

cetanx Highly Voted 1 year, 10 months 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

upvoted 7 times

desertlotus1211 Most Recent 1 month, 2 weeks ago

Selected Answer: C

I'm torn on this question. Okay no access from public internet... does that mean they don't have private lines (e.g. Ded/Partner interconnects)?

Poorly worded. IMO it can either be: Answer B or C based on interpretation of Public Internet.

upvoted 1 times

marlon.andrei 3 months, 3 weeks ago

Selected Answer: B

I vote B, because in "Your on-premises environment cannot be accessed from the public internet.", it would only allow data to be extracted internally within the company. So Transfer Appliance is the most appropriate tool.

upvoted 1 times

namesgeo 4 months, 2 weeks ago

Selected Answer: C

Transfer Service for on-premises data is designed specifically for this scenario. It uses a private, secure agent-based approach to move data from on-premises environments to Google Cloud Storage.

upvoted 1 times

namesgeo 4 months, 2 weeks ago

https://cloud.google.com/blog/products/storage-data-transfer/introducing-storage-transfer-service-for-on-premises-data?hl=en

upvoted 1 times

baimus 7 months ago

They don't define "cannot be accessed from the public internet" - does this mean no incoming traffic, or no traffic or any kind regardless of the initiation point? We simply do not know, and so are left guessing. C? Probably, but could be B, just depending.

upvoted 1 times

Takshashila 1 year, 10 months ago

Selected Answer: C

the correct option is C

upvoted 2 times

wjtb 2 years, 1 month ago

I would say B. It is the ONLY option that is possible without data being accessible over the public (unless we assume that a direct interconnect is already set up, which seems farfetched). Also, nowhere does it say how up-to-date the data needs to be that we are querying or how often we need to query, only that the data increases in size by 100gb per day (indicating that its going to be a lot of data)

upvoted 3 times

musumusu 2 years, 2 months ago

Answer C,
What is wrong with B ? Key words = Daily transfer .. so no to transfer appliance,

upvoted 2 times

zellck 2 years, 5 months 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
Storage Transfer Service for on-premises data enables transfers from network file system (NFS) storage to Cloud Storage.

https://cloud.google.com/bigquery/docs/cloud-storage-transfer-overview
The BigQuery Data Transfer Service for Cloud Storage lets you schedule recurring data loads from Cloud Storage buckets to BigQuery.

upvoted 3 times

AzureDP900 2 years, 4 months ago

yes, It is C

upvoted 1 times

Atnafu 2 years, 5 months ago

C
D-no answer because bq transfer service don't support from on-prem

upvoted 1 times

Atnafu 2 years, 5 months ago

B-is not answer because you want transfer appliance for one time bulk transfer but the question is You want to use Google Cloud products to query and explore the platform data.

query and explore is the key

upvoted 1 times

John_Pongthorn 2 years, 7 months ago

Selected Answer: C

Transfer Service for on-premises is optimal for on-premises google (large files (< 1 TB) and bandwidth available and scheduling)
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options
https://cloud.google.com/blog/products/storage-data-transfer/introducing-storage-transfer-service-for-on-premises-data

BigQuery Data Transfer Service is good for gcs to bigquery
https://cloud.google.com/bigquery/docs/cloud-storage-transfer

upvoted 1 times

John_Pongthorn 2 years, 7 months ago

Sorry I am wrong
(large files > 1 TB + bandwidth available on internal IP address communication + daily scheduling)

upvoted 1 times

John_Pongthorn 2 years, 7 months ago

Your on-premises environment cannot be accessed from the public internet.
It signifies that we can apply private connection like Cloud Interconnect https://cloud.google.com/network-connectivity/docs/interconnect/concepts/overview

upvoted 2 times

Wasss123 2 years, 7 months ago

Selected Answer: C

I will go with C

upvoted 3 times

MounicaN 2 years, 7 months ago

I will g with C

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options

upvoted 1 times

John_Pongthorn 2 years, 7 months ago

C is correct, b is suitable for weekly .
https://cloud.google.com/transfer-appliance/docs/4.0/overview

upvoted 2 times

John_Pongthorn 2 years, 7 months ago

C
Your on-premises environment cannot be accessed from the public internet.
It signifies that we can apply private connection like Cloud Interconnect https://cloud.google.com/network-connectivity/docs/interconnect/concepts/overview

upvoted 1 times

TNT87 2 years, 7 months ago

Selected Answer: C

Ans C
https://cloud.google.com/storage-transfer/docs/on-prem-agent-best-practices

upvoted 1 times

HarshKothari21 2 years, 8 months ago

I would go with option C.
You need a service to transfer data from on-premises to cloud storage. so "Transfer service" is the best option & additionally you can easily configure the network so that data flows through private network.

cloud scheduler on other hand is used mostly for automation. You can schedule a service but in my view cannot be used solo to transfer data.

upvoted 1 times

Load full discussion...