

[Google Discussions](#)

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

[Go to Exam](#)

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 49 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 49

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.

You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (Choose two.)

- A. Introduce data compression for each file to increase the rate file of file transfer.
- B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
- C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
- D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
- E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.



[Show Suggested Answer](#)

by [deleted] at March 21, 2020, 8:48 a.m.

Comments

Type your comment...

Submit

  **Toto2020** Highly Voted  4 years, 4 months ago

E cannot be: Transfer Service is recommended for 300mbps or faster
<https://cloud.google.com/storage-transfer/docs/on-prem-overview>

Bandwidth is not an issue, so B is not an answer

Cloud Storage loading gets better throughput the larger the files are. Therefore making them smaller with compression does not seem a solution. -m option to do parallel work is recommended. Therefore A is not and C is an answer.
<https://medium.com/@duhroach/optimizing-google-cloud-storage-small-file-upload-performance-ad26530201dc>

That leaves D as the other option. It is true you cannot user tar directly with gsutil, but you can load the tar file to Cloud Storage, move the file to a Compute Engine instance with Linux, use tar to split files and copy them back to Cloud Storage. Batching many files in a larger tar will improve Cloud Storage throughput.

So, given the alternatives, I think answer is CD

   upvoted 55 times

  **awssp12345** 3 years, 10 months ago



This should be the correct answer.

   upvoted 3 times

  **musumusu** 2 years, 2 months ago

50mbps is so slow, why you think bandwidth is ok! For parallel upload you need good internet ?

   upvoted 1 times

  **Booqq** 2 years, 2 months ago

normally the solutions are Google Cloud Services based, as it's a vendor exam

   upvoted 2 times



  **Jarek7** 2 years ago

They have 20.000 files 4kb each per hour, so bandwidth needed for it is far below 1mbps. 50mbps is enough to upload all day generated data in about 5 minutes.

   upvoted 2 times

  **vholti** 3 years, 6 months ago

D is incorrect. gsutil with -m option uses multiprocessing/multithreading. It means it will copy the file in parallel. The benefit of multiprocessing/multithreading is significantly high when working with large number of files, instead of file size. The important point of multiprocessing/multithreading is sending multiple files in parallel. Hence file size doesn't give impact to gsutil with -m option. Gsutil with -m option doesn't split a big file into multiple chunks and transfer it in parallel. So in my opinion the answer is A and C.

   upvoted 4 times

  **vholti** 3 years, 6 months ago

Here is the docs which support my opinion:

<https://cloud.google.com/storage/docs/gsutil/addlhelp/TopLevelCommandLineOptions>

   upvoted 3 times

  **Mathew106** 1 year, 9 months ago

We have small files of 4KB and no issues with bandwidth. It's not an issue that -m does not split files. Our problem is with total volume.

   upvoted 1 times

  **Mathew106** 1 year, 9 months ago

As far as I understand compression is not something we want here because bandwidth is not an issue and compressed files will need to be decompressed on the cloud. On top of that if we want to load those files later in BigQuery to create the report we know that we cannot load compressed csv files in parallel.

gsutil makes the most sense because it will be used to load all new files in parallel.

I answered D as well because I thought that none of the others made sense and D is the only one that mentions

creating the bucket on GCS and perhaps migrating data that is missed during the update in the architecture.

So D to create the bucket, C to update the process and move the data to the bucket, then D to move any lost data during the update.

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **Mathew106** 1 year, 9 months ago

Typo, I meant E in my post. C and E, not C and D.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **[Removed]** Highly Voted 5 years, 1 month ago

Should be AC

👍 ↩ 🚩 upvoted 35 times

🗨️ 👤 **GeeBeeEI** 4 years, 5 months ago

support this with a link....

👍 ↩ 🚩 upvoted 3 times

🗨️ 👤 **gcppde** 4 years, 2 months ago

Here you go: <https://cloud.google.com/storage-transfer/docs/overview#gsutil>

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **tavva_prudhvi** 3 years ago

This link does support for C, but what about A? any supported links?

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **BhupiSG** 4 years, 1 month ago

Thank you! From this doc:

Follow these rules of thumb when deciding whether to use gsutil or Storage Transfer Service:

Transfer scenario Recommendation

Transferring from another cloud storage provider Use Storage Transfer Service.

Transferring less than 1 TB from on-premises Use gsutil.

Transferring more than 1 TB from on-premises Use Transfer service for on-premises data.

Transferring less than 1 TB from another Cloud Storage region Use gsutil.

Transferring more than 1 TB from another Cloud Storage region Use Storage Transfer Service.

👍 ↩ 🚩 upvoted 8 times

🗨️ 👤 **iooj** Most Recent 9 months, 1 week ago

Selected Answer: CD

C - because gsutil is recommended for transferring less than 1 TB from on-premises

C excludes E;

bandwidth is not a problem due to a simple math, so we exclude B;

4 KB file is compressed enough, so we exclude A;

D - works fine because even with -m flag we can send tars in parallel.

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **zohra-khouy.f** 1 year, 2 months ago

Selected Answer: AC

AC is the answer

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **spicebits** 1 year, 6 months ago

How can C and E be the answer? They are solving the same problem with different approaches. If you pick C then E can not be an answer. If you pick E then C can not be an answer. This question also seems a bit dated because of gcloud storage cli which is much more performant than gsutil. I would pick C&D as the combination makes the most sense given the choices.

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **Maurilio_Cardoso** 1 year, 11 months ago

@hendrixlives arguments are correct. The approach between the resources in use and how to optimize the ingestion must be balanced.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **Kiroo** 1 year, 11 months ago

Selected Answer: CD

C is correct without an doubt

I was in doubt between D and E

A and B does not seems correct because it states that the bandwidth is not fully utilized .

Now D and E

If the bandwidth was higher the F would be good

if the bandwidth was higher the E would be good

D even if it seems that will not make difference because tar files does not have compression transmit one file instead of 1000 is significantly faster so I would choose

C and D

   upvoted 2 times

  **Oleksandr0501** 2 years ago

Selected Answer: CD

CD , i guess. Liked explanation in discussion.
changing from BC to CD

   upvoted 1 times

  **Kart87** 2 years ago

Guys. need a help. anyone appeared for the exam very recently (Apr 2023)? preparing all the questions from here, would be enough?

   upvoted 2 times

  **Jarek7** 2 years ago

Selected Answer: AC

It seems that Google would like AC. A is not necessary - it doesn't make a significant change - small files do not compress well and bandwidth is so big that file size is not an issue - the issue is 0,2s latency. The biggest benefit is that we can simply enable compression from gsutils parameters, it will not add any implementation complexity.

For me C solo is ok and D solo might be even better, but more complex. C and D cannot be mixed - they exclude each other. C is more simple and uses Google service so it seems to be desired answer. And it makes sense if they want us to select 2 actions we have to make - If we go for C we can also get some benefit from A, if we go for D there is no other answer we can select and it is much more complex in implementation than AC(which is by far good enough).

   upvoted 3 times

  **patiwwb** 1 year, 6 months ago

Yes the 2 are excluding each other. So it's AC

   upvoted 1 times

  **musumusu** 2 years, 2 months ago

Answer: B&C

A: files are 4kb, no need for compression

B: more files to be transmitted per unit time with 100mbps or get 5g network (~200 mbps)

C: gsutil parallel ingestion will reduce time

D: TAR is not a good compression and slower in transfer even slower than csv. speed is 50mbps so don't go with it.

E: Storage Transfer service needs good internet and used for large size of data and for on premises storage, this one is regular ingestion.

   upvoted 3 times

  **manigcp** 2 years, 2 months ago

-- From ChatGPT --

B. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.

A. Introduce data compression for each file to increase the rate of file transfer.

Reasoning:

B. Redesigning the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel will help improve the rate at which the data is transferred to the cloud. This is because gsutil allows for parallel transfers, thereby utilizing the available bandwidth more efficiently and reducing the time required to transfer the data.

A. Introducing data compression for each file will also help improve the rate of file transfer. This is because compressed data takes up less space and can be transferred faster, thereby reducing the time required to transfer the data.


   upvoted 1 times

  **manigcp** 2 years, 2 months ago

Why not option D?

Option D, which involves assembling 1000 files into a TAR file and then transmitting it, may not be an effective solution for the current situation. While TAR archives can help reduce the number of files that need to be transmitted, disassembling the TAR archive in the cloud after receiving it could increase the time required to process the data. This could make it difficult to meet the goal of making reports with data from the previous day available by 10:00 a.m. each day.

Furthermore, compressing the TAR archive could increase the time required to create the archive, and may not provide a significant improvement in terms of transfer time, as the individual CSV files are already small in size. This makes it less effective compared to the other options of parallel transfers and data compression.

   upvoted 1 times

  **Jarek7** 2 years ago

Il wouldnt agree, the main issue here is latency 0,2s and 20000 files per hour - it is even beyond possible transfer

without paralelisation or file merging. Compression and sending 1000 files at once resolves the issue. Just as option C. But they don't make any sense together. I think they exclude D because of additional complexity - compression and then decompression is much more difficult than using gsutil. Thus we go for C. If we need one more then only A makes some sense, but I wouldn't go for it. We have enough bandwidth for this size of file. We just need get rid of latency, by paralelization.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **Jarek7** 2 years ago

OK. AC seems to be right as we can simply enable the compression by gsutil options.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **Leeeeee** 2 years, 5 months ago

Selected Answer: CD

<https://cloud.google.com/storage/docs/parallel-composite-uploads>

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **abhineet1313** 3 years ago

A is incorrect as rate of file transfer is not an issue, system is not able to handle current load itself, compression will make it even faster

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **alecuba16** 3 years ago

Selected Answer: DE

Multiple small files transfer is a bad practice. You should always use some aggregation strategy like tar or zip multiple files. A is discarded because talks about compressing a single file. B is discarded because the bandwidth is not the problem.

Option C could be , but multi threading has a limit. Then the best option is D or use some google on prem mirroring service like E.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **tavva_prudhvi** 2 years, 10 months ago

E is wrong, as Bandwidth already low, so storage Transfer service will not help here

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **Jojo9400** 3 years, 1 month ago

E is wrong Google Cloud Storage Transfer Service (online) != Transfer Appliance(on-premise)

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **OmJanmeda** 3 years, 1 month ago

Selected Answer: CD

CD is correct option

👍 ↩ 🚩 upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses



