



- Expert Verified, Online, **Free**.

20% Discount



Get Unlimited Contributor Access to the all ExamTopics Exams! Take advantage of PDF Files for 1000+ Exams along with community discussions and pass IT Certification Exams Easily.

12 MONTHS

~~\$499.99~~ **\$399.99**

[Buy Now](#)

3 MONTHS

~~\$199.99~~ **\$159.99**

[Buy Now](#)

[Custom View Settings](#)

Topic 1 - Single Topic

Question #1

Topic 1

Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

- A. Threading

- B. Serialization
- C. Dropout Methods
- D. Dimensionality Reduction

Correct Answer: C

Reference:

<https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877>

Please note that there are tons of ways of further improving this result: design of layers and neurons, choosing different initialization and activation schemes, introduction of dropout layers of neurons, early stopping and so on. Furthermore, different types of deep learning models, such as recurrent neural networks might achieve better performance on this task. However, this is not the scope of this introductory post.

Community vote distribution

C (100%)

✉  **henriksoder24** Highly Voted 1 year, 4 months ago

Answer is C.

Bad performance of a model is either due to lack of relationship between dependent and independent variables used, or just overfit due to having used too many features and/or bad features.

- A: Threading parallelisation can reduce training time, but if the selected features are the same then the resulting performance won't have changed.
- B: Serialization is only changing data into byte streams. This won't be useful.
- C: This can show which features are bad. E.g. if it is one feature causing bad performance, then the dropout method will show it, so you can remove it from the model and retrain it.
- D: This would become clear if the model did not fit the training data well. But the question says that the model fits the training data well, so D is not the answer.

upvoted 19 times

✉  **BIGQUERY_ALT** Highly Voted 1 month, 3 weeks ago

anyone from here who took exam after 24-Nov please update what kind of questions you were facing and what is your preparation.

upvoted 5 times

✉  **Ammo21** Most Recent 4 days, 1 hour ago

I am taking the exam on the 31st of January. Can someone please send me the pdf @meemahbacks@gmail.com?

upvoted 1 times

✉  **Ushushusaui_1256** 1 month ago

please send the full version pdf pls i have exam on jan11 @7095965130@gmail.com

pls drop the pdf to this mail

upvoted 1 times

✉  **AceVenturaPD** 1 month, 2 weeks ago

Hi, could somebody please sent me the pdf with all the questions? Mail: acev84411@gmail.com

Thanks!

upvoted 2 times

✉  **Krst1234** 1 month, 2 weeks ago

I have my exam on 14 December. I really appreciate if anyone could share the new full questions PDF with krboddu@gmail.com. Thank You!!

upvoted 2 times

✉  **pss111423** 1 month, 3 weeks ago

I have my exam on 11 December, really appreciate it if anyone could share the new full questions PDF with prash_shimpi@mail.com. Thank You!!

upvoted 1 times

✉  **Fariba** 1 month, 3 weeks ago

Hey, I will have my PDE exam next week. Really appreciate it if anyone could share the new full questions PDF to sypead@yahoo.com. Thank advance.

upvoted 2 times

✉  **Rajkamal1984** 1 month, 3 weeks ago

Hi,
Could any one pls share/send me the latest pdf pls
rrajkamal05@gmail.com

upvoted 2 times

✉  **NV2014** 1 month, 3 weeks ago

I have my exam on 10 December, really appreciate it if anyone could share the new full questions PDF with nalinvarshney@mail.com. Thank You!!

upvoted 3 times

✉  **Takshashila** 1 month, 4 weeks ago

I have my exam on 5 December, really appreciate it if anyone could share the new full questions PDF with agrawalshikhar32@gmail.com. Thank You!!

upvoted 1 times

✉  **BIGQUERY_ALT_ALT** 1 month, 3 weeks ago

Kindly update what questions you faced in the exam and what is your preparation

upvoted 2 times

✉  **karim1321** 2 months ago

I have my exam on 1 December, really appreciate it if anyone could share the new full questions PDF with rifkikarimr@gmail.com. Thank You!!

upvoted 1 times

✉  **roty** 2 months ago

have u received

upvoted 1 times

✉  **andreeviana1** 1 month, 4 weeks ago

Hey
Can you send me the pdf pls
Andre.v@outlook.pt

upvoted 1 times

✉  **navneet3010** 2 weeks, 5 days ago

Hey
Can you send me the pdf at shnavneet000@gmail.com

upvoted 1 times

✉  **jmmoyanokk** 2 months, 1 week ago

Has anyone taken the exam after November 13, 2023? Based on my observation on Google's webpage, they have introduced new sections a topics. Do the old questions still dominate?

upvoted 3 times

✉  **imiu** 2 months, 1 week ago

I did, today, only 2 questions were from here, anyone has the new questions please? I would really appreciate to be posted or sent to my email address irinamiu89@gmail.com, thank you

upvoted 3 times

✉  **Pdpj** 2 months ago

I have my exam tomorrow really appreciate it if anyone could share the new questions with pitu10008@gmail.com, Thank you!!

upvoted 2 times

✉  **roty** 2 months ago

did u got thr pdf

upvoted 1 times

✉  **Mrsq25** 1 month, 2 weeks ago

I have my exam december 26 really appreciate it if anyone could share the new questions with mcastilogarc@gmail.com, Thanks for !!

upvoted 2 times

✉  **himanshu3004** 1 month, 1 week ago

Can someone please share it with me as well at himanshu9xm@gmail.com.

I am writing exam on 25th December.

upvoted 1 times

✉  **chiuhing** 2 months, 2 weeks ago

I will sit for the PDE exam in a week, appreciate if someone can send the full PDF questions to my email (chiuhinggcp@gmail.com), thank you your help!

upvoted 1 times

✉  **rocky48** 2 months, 4 weeks ago

Selected Answer: C

A: Threading parallelisation can reduce training time, but if the selected features are the same then the resulting performance won't have changed

B: Serialization is only changing data into byte streams. This won't be useful.

C: This can show which features are bad. E.g. if it is one feature causing bad performance, then the dropout method will show it, so you can remove it from the model and retrain it.

D: This would become clear if the model did not fit the training data well. But the question says that the model fits the training data well.

So, C is the answer.

upvoted 1 times

✉  **rtcp0st** 3 months, 1 week ago

Selected Answer: C

C. Dropout Methods

Dropout is a regularization technique commonly used in neural networks to prevent overfitting. It helps improve the generalization of the model by randomly setting a fraction of the neurons to zero during each training iteration, which prevents the network from relying too heavily on specific neurons. This, in turn, can lead to better performance on new, unseen data.

upvoted 1 times

✉  **kumarts** 3 months, 1 week ago

Cleared PDE exam today, most of the questions were from here...only 3 questions were new

upvoted 1 times

✉  **Rayjayhelp** 3 months ago

Please help me with the questions

upvoted 1 times

Question #2

Topic 1

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

- A. Continuously retrain the model on just the new data.
- B. Continuously retrain the model on a combination of existing data and the new data.
- C. Train on the existing data while using the new data as your test set.
- D. Train on the new data while using the existing data as your test set.

Correct Answer: B

Community vote distribution

B (93%)

7%

✉️  **serg3d** Highly Voted 3 years, 8 months ago

I think it should be B because we have to use a combination of old and new test data as well as training data
upvoted 36 times

✉️  **dambilwa** 3 years, 7 months ago

Yes - The training set should be shuffled well to represent data across all scenarios
upvoted 4 times

✉️  **jagadamba** Highly Voted 3 years, 7 months ago

B, as we need to train the data with new data, so that it will keep learning, and as well as used for test
upvoted 11 times

✉️  **hpvb** Most Recent 2 months, 3 weeks ago

D
<https://cloud.google.com/automl-tables/docs/prepare#ml-use>
upvoted 1 times

✉️  **hpvb** 2 months, 3 weeks ago

D
<https://datascience.stackexchange.com/questions/12761/should-a-model-be-re-trained-if-new-observations-are-available>
upvoted 1 times

✉️  **rocky48** 2 months, 4 weeks ago

Selected Answer: B

Option A is not recommended because retraining the model on just new data will cause the model to lose the information it has learned from historical data.

Option C and D are not recommended because they are using the new data as test set and this approach will lead to a model that is overfitted and not generalize well to new users.

So answer is B

upvoted 1 times

✉️  **dumpswowner** 3 months ago

I think A is correct answer : I was pass google exam with the help of dumpswowner
upvoted 1 times

✉️  **rajkinz** 3 months ago

Answer is C. It is time sensitive data so latest data should be used for testing.
Reference: <https://cloud.google.com/automl-tables/docs/prepare#ml-use>
upvoted 1 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: B

This approach allows the model to benefit from both the historical data (existing data) and the new data, ensuring that it adapts to changing preferences while retaining knowledge from the past. By combining both types of data, the model can learn to make recommendations that are up-to-date and relevant to users' evolving preferences.

upvoted 1 times

✉  **Victor2087** 4 months ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - singh.bhupi2087@gmail.com. Hoping to hear something good soon :)

Thanks a ton.

upvoted 1 times

✉  **Websurfer** 5 months, 3 weeks ago

Selected Answer: B

train on old and new data

upvoted 1 times

✉  **AsthaGupta** 6 months, 2 weeks ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - aasthagupta019@gmail.com.

Hoping to hear something good soon :)

Thanks a ton.

upvoted 3 times

✉  **Victor2087** 4 months ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - singh.bhupi2087@gmail.com. Hoping to hear something good soon :)

Thanks a ton.

upvoted 1 times

✉  **AmmarFasih** 8 months, 1 week ago

Selected Answer: B

Option B is the right answer. Since the questions states the models needs to be updated since the clothing preference changes. Hence we need the new data to be utilized for training/ updating model.

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: B

Have verified this

upvoted 1 times

✉  **jin0** 11 months, 1 week ago

there are two point first when retraining second what data. I think retraining should be occur when the model could not predict well in this case there is monitoring metric should be needed first but no one said, second what data? in this case I think the answer is A. because when the model could not predict well it means the data variance and bias are changed so, it's no make sense what is combination new data with old data because the data being not be changed is not necessary anymore..

upvoted 1 times

✉  **jin0** 11 months, 1 week ago

And the questions should explain in detail.. whether it's deep learning or tree based machine learning model.. and how large of new dataset is.. I think

upvoted 1 times

✉  **Morock** 11 months, 2 weeks ago

Selected Answer: C

The trend keep changing, so must mix new and old data...

upvoted 1 times

✉️  **samdhimal** 1 year ago

Selected Answer: B

B. Continuously retrain the model on a combination of existing data and the new data.

This approach will help to ensure that the model remains up-to-date with the latest fashion preferences of the users, while also leveraging the historical data to provide context and improve the accuracy of the recommendations. Retraining the model on a combination of existing and new data will help to prevent the model from being overly influenced by the new data and losing its ability to generalize to users with different preferences.

Option A is not recommended because retraining the model on just new data will cause the model to lose the information it has learned from historical data.

Option C and D are not recommended because they are using the new data as test set and this approach will lead to a model that is overfitted and not generalize well to new users.

upvoted 4 times

✉️  **rocky48** 2 months, 4 weeks ago

Nice explanation bro.

upvoted 1 times

✉️  **korntewin** 1 year ago

The answer can be A, if we implement online learning! But for regular model which can't implement online learning (everything with no gradient descent) the answer should be B.

upvoted 1 times

✉️  **testoneAZ** 1 year ago

Correct answer is B

upvoted 1 times

Question #3

Topic 1

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
- C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
- D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Correct Answer: C

Community vote distribution

C (100%)

✉️  **Hanakon**  3 years, 1 month ago

C - based on Google documentation, self-join is an anti-pattern: <https://cloud.google.com/bigquery/docs/best-practices-performance-patterns#self-joins>

upvoted 29 times

✉️  **awssp12345** 2 years, 7 months ago

Super helpful! Thanks. :)

upvoted 1 times

✉️  **[Removed]**  3 years, 10 months ago

Correct: C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per client while also increasing performance due to avoiding self-joins.

upvoted 12 times

✉️  **rocky48**  2 months, 4 weeks ago

Selected Answer: C

Normalization is a technique used to organize data in a relational database to reduce data redundancy and improve data integrity. Breaking the patient records into separate tables (patient and visits) and eliminating self-joins will make the database more scalable and improve query performance. It also helps maintain data integrity and makes it easier to manage large datasets efficiently.

Options A, B, and D may provide some benefits in specific cases, but for a scenario where the project scope has expanded significantly and there are performance issues with self-joins, normalization (Option C) is the most robust and scalable solution.

upvoted 1 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: C

Normalization is a technique used to organize data in a relational database to reduce data redundancy and improve data integrity. Breaking the patient records into separate tables (patient and visits) and eliminating self-joins will make the database more scalable and improve query performance. It also helps maintain data integrity and makes it easier to manage large datasets efficiently.

Options A, B, and D may provide some benefits in specific cases, but for a scenario where the project scope has expanded significantly and there are performance issues with self-joins, normalization (Option C) is the most robust and scalable solution.

upvoted 2 times

✉️  **Victor2087** 4 months ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - singh.bhupi2087@gmail.com. Hoping to hear something good soon :)

Thanks a ton.

upvoted 1 times

✉️  **SB5007** 4 months ago

try here <https://www.passnexam.com/google/professional-data-engineer>

upvoted 2 times

✉️  **ALLYDAN** 4 months ago

Did anyone send this to you, would you please forward it to me aleksdanfan@gmail.com

upvoted 1 times

✉️  **Anushka0712** 5 months, 3 weeks ago

Hi everyone, I am going to appear for this exam on 25th August. It would be a great help if anyone could forward the full set of questions to the email id: mailto:anushka.singh12@gmail.com. Thanks in advance.

upvoted 1 times

✉️  **vaga1** 8 months, 2 weeks ago

Selected Answer: C

"100 times more patient records" immediately brings to create a patient dimensional table to save space on disk if a general relational database is mentioned.

upvoted 1 times

✉️  **maurilio_cardoso_multiedro** 10 months, 3 weeks ago

C - <https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

upvoted 1 times

✉️  **bha11111** 10 months, 3 weeks ago

Selected Answer: C

C- This is correct have verified from different sources

upvoted 1 times

✉️  **Morock** 11 months, 2 weeks ago

Selected Answer: C

Should be C. Basic ER design...

upvoted 1 times

✉️  **GCPpro** 1 year ago

c - is the correct one.

upvoted 1 times

✉️  **testoneAZ** 1 year ago

C should be the correct answer

upvoted 1 times

✉️  **Brilliantyagi** 1 year, 1 month ago

Selected Answer: C

C- Is the correct answer!

upvoted 1 times

✉️  **Arkon88** 1 year, 11 months ago

Selected Answer: C

C - based on Google documentation, self-join is an anti-pattern:

<https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

upvoted 2 times

✉️  **ch1nczyk** 1 year, 11 months ago

Selected Answer: C

Correct

upvoted 1 times

✉️  **samdhimal** 2 years ago

correct answer -> Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.

Avoid self-join at all cost because that's what google says.

Reference:

<https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

upvoted 3 times

✉️  **samdhimal** 1 year ago

Normalizing the database design will help to minimize data redundancy and improve the efficiency of the queries. By separating the patient and visit information into separate tables, the database will be able to handle the increased number of records and generate reports more efficiently, because the self-joins will no longer be required.

Option A is not a good solution because adding more capacity to the server will not address the underlying problem of the database design and it may not be sufficient to handle the increased data volume.

Option B is not a good solution because it limits the flexibility of the queries and reports, and it may not be sufficient to handle the increased data volume.

Option D is not a good solution because partitioning the table into smaller tables may lead to data redundancy and it may not be sufficient to handle the increased data volume.

upvoted 2 times

✉️  **MaxNRG** 2 years, 2 months ago

C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.

A is not correct because adding additional compute resources is not a recommended way to resolve database schema problems.

B is not correct because this will reduce the functionality of the database and make running reports more difficult.

D is not correct because this will likely increase the number of tables so much that it will be more difficult to generate reports vs. the correct option.

<https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

<https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax#explicit-alias-visibility>

upvoted 6 times

You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

- A. Disable caching by editing the report settings.
- B. Disable caching in BigQuery by editing table details.
- C. Refresh your browser tab showing the visualizations.
- D. Clear your browser history for the past hour then reload the tab showing the visualizations.

Correct Answer: A

Reference:

<https://support.google.com/datastudio/answer/7020039?hl=en>

How to tell if report data is cached

You can see if data is coming from the cache by viewing the report and looking in the bottom left corner. When all the charts on the current page are being served from the cache, you'll see a lightning bolt icon along with the time and date of the last update .

Blending and cached data

For a blended data source, the cache will use the setting that satisfies the desired refresh times for all of the data sources included in the blend.

For example, if you blend a Sheets data source having a refresh time of 15 minutes, with a BigQuery data source having a refresh time of 4 hours, the resulting blended data source will have a refresh time of 15 minutes.

Community vote distribution

A (71%)	C (21%)	7%
---------	---------	----

  **Khaled_Rashwan**  11 months, 1 week ago

- A. Disable caching by editing the report settings.

By default, Google Data Studio 360 caches data to improve performance and reduce the amount of queries made to the data source. However, this can cause visualizations to not show data that is less than 1 hour old, as the cached data is not up-to-date.

To resolve this, you should disable caching by editing the report settings. This can be done by following these steps:

Open the report in Google Data Studio 360.

Click on the "File" menu in the top left corner of the screen.

Select "Report settings" from the dropdown menu.

In the "Report settings" window, scroll down to the "Data" section.

Toggle off the "Enable cache" option.

Click the "Save" button to apply the changes.

Disabling caching ensures that the data shown in the visualizations is always up-to-date, but it may increase the query load on the data source and affect the report's performance. Therefore, it's important to consider the trade-off between performance and data accuracy when making this change.

upvoted 6 times

✉️  **rocky48** Most Recent 2 months, 4 weeks ago

Selected Answer: A

A. Disable caching by editing the report settings.

By default, Google Data Studio 360 caches data to improve performance and reduce the amount of queries made to the data source. However, this can cause visualizations to not show data that is less than 1 hour old, as the cached data is not up-to-date.

upvoted 1 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: A

Disabling caching in the report settings will ensure that the visualizations are not using cached data and will reflect the most up-to-date information from your Google BigQuery data source. This will allow your report to show data that is less than 1 hour old. Caching is often used for performance optimization, but it can result in delays in displaying real-time or near-real-time data, so disabling it is the appropriate action in this case.

upvoted 1 times

✉️  **Websurfer** 5 months, 3 weeks ago

Selected Answer: A

disable caching in report setting will get the issue resolved

upvoted 1 times

✉️  **AsthaGupta** 6 months, 2 weeks ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - aasthagupta019@gmail.com.

Hoping to hear something good soon :)

Thanks a ton.

upvoted 1 times

✉️  **Morock** 11 months, 2 weeks ago

Selected Answer: D

The solution from the site is perfect.

upvoted 1 times

✉️  **PolyMoe** 1 year ago

Selected Answer: A

what is relevant here is to uncache Data Studio

upvoted 1 times

✉️  **samdhimal** 1 year ago

A. Disable caching by editing the report settings.

Data Studio 360 uses caching to speed up report loading times. When caching is enabled, Data Studio 360 will only show the data that was present in the data source at the time the report was loaded. To ensure that the visualizations in your report are always up-to-date, you should disable caching by editing the report settings. This will force Data Studio 360 to retrieve the latest data from the data source (in this case BigQuery) every time the report is loaded.

Option B is incorrect as it would only disable caching in BigQuery, but it wouldn't affect the caching in Data Studio 360, so the visualizations would still not show the latest data.

Option C and D will not help as the data is not being updated in Data Studio 360, it's just the cache that needs to be updated.

upvoted 3 times

✉️  **korntewin** 1 year ago

I'm confused, is it possible that the cache is in BigQuery level? and the looker just get the cache from bigquery

upvoted 1 times

✉️  **ejlp** 1 year, 2 months ago

Selected Answer: A

Based on the doc, you can refresh the report using refresh button on the report, not the browser's refresh button. So the answer is A.

upvoted 1 times

✉️  **maksi** 1 year, 2 months ago

On my opinion it's C, because Data Studio doesn't support the Real-Time dashboard updates, so it means that if caching will be disabled, us will be forced to update the dashboard manually, otherwise report will be stuck on the data from the last update. According to documentation <https://support.google.com/looker-studio/answer/7020039?hl=en#zippy=%2Cin-this-article>, if we want to keep the data fresh we need to set caching with minimum value of 15 minites - it means that data in the report will be updated automatically wvery 15 minutes, if cache will be disabled completely then the report will be stuck until we will manually update it. So, tbh for me it doesn't make sense to disable cache.

upvoted 1 times

✉️  **viks1122** 1 year, 3 months ago

C.

Since the data is not always stale. When it is, click on refresh button. Document also says the same

Refresh report data manually

Report editors can refresh the cache at any time:

View or edit the report.

In the top right, click More options. and then click RefreshRefresh data .

This refreshes the cache for every data source added to the report.

upvoted 1 times

✉️  **beowulf_kat** 1 year, 2 months ago

Refreshing the web browser does not refresh the data behind the viz's in Data Studio. You have to click the 'refresh data source' button.

upvoted 3 times

✉️  **nicholascz** 1 year, 3 months ago

Selected Answer: A

<https://support.google.com/looker-studio/answer/7020039?hl=en#zippy=%2Cin-this-article>

upvoted 1 times

✉️  **kennyloo** 1 year, 3 months ago

A. should be correct. after disabled the cache, it will retrieve data every time.

upvoted 1 times

✉️  **max_c** 1 year, 3 months ago

Selected Answer: C

Same question from a Cloud guru and answer was C. The wording is slightly different in the documentation but still, the idea is that you can trigger a manual refresh

<https://support.google.com/datastudio/answer/7020039?hl=en#zippy=%2Cin-this-article:~:text=automatic%20cache%20refreshes.-,Refresh%20report%20data%20manually,-Report%20editors%20can>

upvoted 3 times

✉️  **Lestrang** 1 year ago

The option is to refresh the browser tab not inside data studio itself. incorrect.

upvoted 1 times

✉️  **AWSandeep** 1 year, 4 months ago

Selected Answer: A

A. Disable caching by editing the report settings.

upvoted 4 times

Question #5

Topic 1

An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

- A. Use federated data sources, and check data in the SQL query.
- B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
- C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
- D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Correct Answer: D

Community vote distribution

D (100%)

 **[Removed]** Highly Voted 3 years, 10 months ago

Agreed: D

upvoted 25 times

 **Radhika7983** Highly Voted 3 years, 2 months ago

The answer is D. An ETL pipeline will be implemented for this scenario. Check out handling invalid inputs in cloud data flow

<https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow>

ParDos . . . and don'ts: handling invalid inputs in Dataflow using Side Outputs as a "Dead Letter" file

upvoted 13 times

 **jkhong** 1 year, 1 month ago

The sources you've provided cannot be accessed. Here is an updated best practice. https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-testing#use_dead_letter_queues

upvoted 5 times

 **RT_G** Most Recent 2 months, 3 weeks ago

Selected Answer: D

All other options only alert or error out bad data. As the question requires, option D sends bad data to the dead letter table for further analysis while valid data is loaded to the table

upvoted 1 times

 **rocky48** 2 months, 4 weeks ago

Selected Answer: D

Option A is incorrect because federated data sources do not provide any data validation or cleaning capabilities and you'll have to do it on the SQL query, which could slow down the performance.

Option B is incorrect because Stackdriver monitoring can only monitor the performance of the pipeline, but it can't handle corrupted or incorrectly formatted data.

Option C is incorrect because using gcloud CLI and setting max_bad_records to 0 will ignore the corrupted or incorrectly formatted data and continue the load process, this will lead to incorrect analysis.

Answer D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

upvoted 1 times

 **rtcpost** 3 months, 1 week ago

Selected Answer: D

Google Cloud Dataflow allows you to create a data pipeline that can preprocess and transform data before loading it into BigQuery. This approach will enable you to handle problematic rows, push them to a dead-letter table for later analysis, and load the valid data into BigQuery

Option A (using federated data sources and checking data in the SQL query) can be used but doesn't directly address the issue of handling corrupted or incorrectly formatted rows.

Options B and C are not the best choices for handling data quality and error issues. Enabling monitoring and setting max_bad_records to 0 in BigQuery may help identify errors but won't store the problematic rows for further analysis, and it might prevent loading any data with issues, which may not be ideal.

upvoted 1 times

✉  **NeoNitin** 4 months, 2 weeks ago

ans D ,thank you exam topic , connect me if need any help

neonitin6@gmail.com

upvoted 1 times

✉  **NeoNitin** 4 months, 3 weeks ago

D , Thank you Exam topic : Passed the exam in august and I can say examtopic is help me lot, topic 1 is enough for the exam, just last week I received welcome kit from google for PDE exam one google cloud cup. if you need any help reach out to me neonitin6@therategoogledotcom

upvoted 1 times

✉  **vaga1** 8 months, 2 weeks ago

Selected Answer: D

Agreed: D

upvoted 1 times

✉  **odiez3** 10 months, 1 week ago

D because you need Transform the data

upvoted 1 times

✉  **Morock** 11 months, 2 weeks ago

Selected Answer: D

D. The question is asking pipeline, then let's build a pipeline.

upvoted 3 times

✉  **vaga1** 8 months, 2 weeks ago

I agree. There are not much information on what to do. Every answer is valid except B. But in strictly technical terms only D generates a real pipeline.

upvoted 1 times

✉  **samdhimal** 1 year ago

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

By running a Cloud Dataflow pipeline to import the data, you can perform data validation, cleaning and transformation before it gets loaded in BigQuery. Dataflow allows you to handle corrupted or incorrectly formatted rows by pushing them to another dead-letter table for analysis. This way, you can ensure that only clean and correctly formatted data is loaded into BigQuery for analysis.

upvoted 2 times

✉  **samdhimal** 1 year ago

Option A is incorrect because federated data sources do not provide any data validation or cleaning capabilities and you'll have to do it on SQL query, which could slow down the performance.

Option B is incorrect because Stackdriver monitoring can only monitor the performance of the pipeline, but it can't handle corrupted or incorrectly formatted data.

Option C is incorrect because using gcloud CLI and setting max_bad_records to 0 will ignore the corrupted or incorrectly formatted data and continue the load process, this will lead to incorrect analysis.

upvoted 5 times

✉  **hamza101** 6 months, 1 week ago

for Option C I think when setting max_bad_records to 0 this will prevent the loading to be achieved since the condition will cut off the loading if we have at least 1 corrupted row

upvoted 1 times

✉  **Besss** 1 year, 3 months ago

Selected Answer: D

Agreed: D

upvoted 1 times

✉  **Dip1994** 1 year, 6 months ago

The correct answer is D

upvoted 1 times

✉  **Arkon88** 1 year, 11 months ago

Selected Answer: D

Correct - D (as we need to create Pipeline) which is possible via 'D'

upvoted 1 times

👤 **MaxNRG** 2 years, 2 months ago

Looks like D, with C you will not import anything, stackdriver alerts will not help you with this and with federated resources you won't know what happened with those bad records. D is the most complete one.
<https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow>
upvoted 3 times

👤 **anji007** 2 years, 3 months ago

Ans: D

upvoted 1 times

👤 **nickozz** 2 years, 4 months ago

D seems to be correct. explained here how combined with Pub/Sub, this can be achieved. <https://cloud.google.com/pubsub/docs/handling-failures>
upvoted 1 times

Question #6

Topic 1

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and serves millions of users. How should you design the frontend to respond to a database failure?

- A. Issue a command to restart the database servers.
- B. Retry the query with exponential backoff, up to a cap of 15 minutes.
- C. Retry the query every second until it comes back online to minimize staleness of data.
- D. Reduce the query frequency to once every hour until the database comes back online.

Correct Answer: B

Community vote distribution

B (100%)

👤 **Radhika7983** Highly Voted 3 years, 2 months ago

Correct answer is B. App engine creates applications that use Cloud SQL database connections effectively. Below is what is written in Google Cloud documentation.

If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database.

This retry only makes sense when first connecting, or when first grabbing a connection from the pool. If errors happen in the middle of a transaction, the application must do the retrying, and it must retry from the beginning of a transaction. So even if your pool is configured properly the application might still see errors if connections are lost.

reference link is <https://cloud.google.com/sql/docs/mysql/manage-connections>

upvoted 51 times

👤 **Ilamaste** Highly Voted 3 years, 6 months ago

<https://cloud.google.com/sql/docs/mysql/manage-connections#backoff>

upvoted 12 times

✉️  **RT_G** Most Recent 2 months, 3 weeks ago

Selected Answer: B

Retries with exponential backoff seems like the most efficient option in this scenario
upvoted 1 times

✉️  **rocky48** 2 months, 4 weeks ago

Selected Answer: B

Correct answer is B
upvoted 1 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: B

Exponential backoff is a commonly used technique to handle temporary failures, such as a database server becoming temporarily unavailable. This approach retries the query, initially with a short delay and then with increasingly longer intervals between retries. Setting a cap of 15 minutes ensures that you don't excessively burden your system with constant retries.

Option C (retrying the query every second) can be too aggressive and may lead to excessive load on the server when it comes back online.

Option D (reducing the query frequency to once every hour) would result in significantly stale data and a poor user experience, which is generally not desirable for a weather app.

Option A (issuing a command to restart the database servers) is not a suitable action for a frontend component and might not address the issue effectively. Database server restarts should be managed as a part of the infrastructure and not initiated by the frontend.

upvoted 2 times

✉️  **Victor2087** 4 months ago

Hello Everyone, I am going to take this exam. It would be a great help if anyone could forward the full deck of questions to this email id - singh.bhupi2087@gmail.com. Hoping to hear something good soon :)

Thanks a ton.

upvoted 1 times

✉️  **AceCloud** 3 months, 1 week ago

Hey Victor2087, If you took the exam, Can you please let us know how much portion of questions in the exam you got from these dumps?
upvoted 3 times

✉️  **gudguy1a** 4 months, 3 weeks ago

Selected Answer: B

good answer, good answer @radhika7983.
upvoted 1 times

✉️  **abhi11393** 5 months, 2 weeks ago

Hi everyone, I am going to appear for this exam on 1st September. It would be a great help if anyone could forward the full set of questions to this email id: abhishek11393@gmail.com. Thanks in advance.

upvoted 1 times

✉️  **Anushka0712** 5 months, 3 weeks ago

Hi everyone, I am going to appear for this exam on 25th August. It would be a great help if anyone could forward the full set of questions to this email id: mailto:anushka.singh12@gmail.com. Thanks in advance.

upvoted 1 times

✉️  **Datardp** 7 months, 3 weeks ago

B is answer

upvoted 1 times

✉️  **vaga1** 8 months, 2 weeks ago

Selected Answer: B

I agree with the exponential backoff technique, even though I do not see why 15 minutes should be a desired choice.
upvoted 1 times

✉️  **vaga1** 8 months, 2 weeks ago

I guess that when you have failed after 15 minutes, your app must go through a serious review before being used again, since it is not able to provide the updated results as quickly as desired.

upvoted 1 times

✉  **yafsong** 1 year, 1 month ago

Truncated exponential backoff is a standard error-handling strategy for network applications. In this approach, a client periodically retries a failed request with increasing delays between requests

upvoted 4 times

✉  **hiromi** 1 year, 2 months ago

Selected Answer: B

B is right

upvoted 1 times

✉  **shiv14** 1 year, 11 months ago

Selected Answer: B

According to the documentation

upvoted 1 times

✉  **samdhimal** 2 years ago

correct answer -> Retry the query with exponential backoff, up to a cap of 15 minutes.

If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database.

Reference:

<https://cloud.google.com/sql/docs/mysql/manage-connections#backoff>

upvoted 2 times

✉  **samdhimal** 1 year ago

Exponential backoff with a cap is a common technique used to handle temporary failures, such as database outages. In this approach, the frontend will retry the query with increasing intervals (e.g., 1s, 2s, 4s, 8s, etc.) up to a maximum interval (in this case, 15 minutes), this will help to avoid overwhelming the database servers with too many requests at once, and minimize the impact of the failure on the users.

Option A, is not recommended because it's not guaranteed that restarting the database servers will fix the problem, it could be a network configuration problem and it could cause more downtime.

Option C is not recommended because it could cause too many requests to be sent to the server, overwhelming the database and causing more downtime.

Option D is not recommended because reducing the query frequency too much would result in stale data, and users will not receive the most up-to-date information.

upvoted 2 times

✉  **deep_ROOT** 2 years ago

B is Correct; this question appeared in Cloud Architect exam also

upvoted 1 times

✉  **MaxNRG** 2 years, 2 months ago

B, <https://cloud.google.com/sql/docs/mysql/manage-connections#backoff>

backoff is a standard error handling strategy for network applications in which a client periodically retries a failed request with increasing delays between requests. Clients should use truncated exponential backoff for all requests to Cloud Storage that return HTTP 5xx and 429 response codes, including uploads and downloads of data or metadata.

upvoted 2 times

- A. Linear regression
- B. Logistic classification
- C. Recurrent neural network
- D. Feedforward neural network

Correct Answer: A

Community vote distribution

A (100%)

 **Radhika7983** Highly Voted 3 years, 3 months ago

Correct answer is A. A tip here to decide when a liner regression should be used or logistics regression needs to be used. If you are forecastir that is the values in the column that you are predicting is numeric, it is always liner regression. If you are classifying, that is buy or no buy, yes no, you will be using logistics regression.

upvoted 51 times

 **Anirkent** 3 years ago

Liner Regression is correct but this is one aspect of the question, how does it relates to resource constrained machines? or that could be j a distraction?

upvoted 7 times

 **muzammilnxs** 2 years, 12 months ago

Neural Networks(Feed Forward or Recurrent) require resource intensive machines(i.e GPU's) whereas Linear regression can be done on ordinary CPU's

upvoted 22 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A

Description: Forecasting and Liner regression is used for predicting housing price

upvoted 9 times

 **Fazan456** Most Recent 3 weeks ago

Selected Answer: A

Here, due to budget constraints, we're utilizing a single resource-constrained virtual machine, operating in a minimal resource environment. Linear regression emerges as the appropriate algorithm. It's a lightweight predictive model that suits our resource limitations

upvoted 1 times

 **RT_G** 2 months, 3 weeks ago

Selected Answer: A

Linear regression will be used since the prediction requires forecasting prices involving numeric values and is computationally less resource intensive

upvoted 1 times

 **rocky48** 2 months, 4 weeks ago

Selected Answer: A

Correct answer is A

upvoted 1 times

 **rtcpost** 3 months, 1 week ago

Selected Answer: A

Linear regression is a simple and resource-efficient algorithm for predicting continuous values like housing prices. It's computationally lightwe and well-suited for single machines with limited resources. It doesn't require the extensive computational power or specialized hardware that more complex algorithms like neural networks (options C and D) might need.

Option B (Logistic classification) is used for binary classification tasks, not for predicting continuous values like housing prices, so it's not the right choice in this context.

upvoted 1 times

 **Anushka0712** 5 months, 3 weeks ago

Hi everyone, I am going to appear for this exam on 25th August. It would be a great help if anyone could forward the full set of questions to th email id: mailtoanushka.singh12@gmail.com. Thanks in advance.

upvoted 1 times

✉  **AmmarFasih** 8 months, 1 week ago

Selected Answer: A

Correct Answer is A. Since linear regression is used to predict a numeric value. While logistic regression is used to classify among the binary scenario.

Further option C and D are advance ML options and not cost and resource effective for the current situation.

upvoted 1 times

✉  **Zosby** 11 months, 1 week ago

predict housing prices = linear regression

upvoted 2 times

✉  **JJJJim** 11 months, 2 weeks ago

Selected Answer: A

must be A.

Though C can do it, linear regression is the better practice.

upvoted 1 times

✉  **lukas_xls** 1 year, 1 month ago

Selected Answer: A

Must be A

upvoted 1 times

✉  **rowan_** 1 year, 5 months ago

A for sure. B is for classification. Neural nets can accomplish the task but they take WAY too many resources

upvoted 2 times

✉  **samdhimal** 2 years ago

correct answer -> Linear Regression

Linear regression is a statistical method that allows to summarize and study relationships between two continuous (quantitative) variables: On variable, denoted X, is regarded as the independent variable. The other variable denoted y is regarded as the dependent variable. Linear regression uses one independent variable X to explain or predict the outcome of the dependent variable y.

Whenever you are told to predict some future value of a process which is currently running, you can go with a regression algorithm.

Reference:

<https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60>

upvoted 4 times

✉  **samdhimal** 1 year ago

Linear regression is a simple and computationally efficient algorithm that can be used to predict a continuous target variable based on one more input variables. It is particularly well-suited for resource-constrained environments, as it requires minimal computational resources and can be run on a single virtual machine.

Linear regression is a good fit for this problem as it is a supervised learning algorithm that can be used for regression problems, and it's not computationally expensive.

Option B is not recommended as Logistic classification is a supervised learning algorithm that is used for classification problems, not regression problems.

Option C and D are not recommended as Recurrent Neural Network (RNN) and Feedforward Neural Network (FNN) are computationally expensive and may require significant computational resources and memory to run on a single virtual machine.

upvoted 2 times

✉  **MaxNRG** 2 years, 2 months ago

A as Supervised learning using Regression can help build a model to predict house prices.

Option B is wrong as Classification would not help to solve the problem.

Options C & D are wrong as they would need more resources.

upvoted 3 times

✉  **anji007** 2 years, 3 months ago

Ans: A

upvoted 1 times

✉  **StefanoG** 2 years, 4 months ago

Ok the right answer is A, but the question is why? Then:

- B not because we are make forecasting and not classifying
- C and D not because this solution need more nodes, then more VM.

Right?

upvoted 2 times

✉  **sumanshu** 2 years, 6 months ago

Vote for A

upvoted 1 times

Question #8

Topic 1

You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

- A. Include ORDER BY DESK on timestamp column and LIMIT to 1.
- B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
- C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
- D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Correct Answer: D

Community vote distribution

D (100%)

✉  **[Removed]**  3 years, 10 months ago

Answer: D

Description: Row Number equals 1 with partitioning will ensure only one record is fetched per partition

upvoted 23 times

✉  **[Removed]**  3 years, 10 months ago

Correct D

upvoted 13 times

✉  **RT_G**  2 months, 3 weeks ago

Selected Answer: D

D ensures data is partitioned by the unique id and only one record is picked thereby ensuring results are de-duplicated

upvoted 1 times

✉  **rtcp0st** 3 months, 1 week ago

Selected Answer: D

This approach will assign a row number to each row within a unique ID partition, and by selecting only rows with a row number of 1, you will ensure that duplicates are excluded in your query results. It allows you to filter out redundant rows while retaining the latest or earliest records based on your timestamp column.

Options A, B, and C do not address the issue of duplicates effectively or interactively as they do not explicitly remove duplicates based on the unique ID and event timestamp.

upvoted 1 times

✉  **NeoNitin** 4 months, 3 weeks ago

answer: D ,,, Thank you Exam topic : Passed the exam in august and I can say examtopic is help me lot, topic 1 is enough for the exam, just week I received welcome kit from google for PDE exam one google cloud cup. if you need all question any help reach out to me neonitin6attherategoogledotcom

upvoted 2 times

✉  **abhi11393** 5 months, 2 weeks ago

Hi everyone, I am going to appear for this exam on 1st September. It would be a great help if anyone could forward the full set of questions to this email id: abhishek11393@gmail.com. Thanks in advance.

upvoted 1 times

✉  **Zosby** 11 months, 1 week ago

Correct D

upvoted 1 times

✉  **Morock** 11 months, 2 weeks ago

Selected Answer: D

Row number gives the unique number ranking based on target column.

upvoted 3 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: D

It's the only valid option, try it your self with examples in QB.

upvoted 1 times

✉  **Ender_H** 1 year, 4 months ago

I personally don't think any answer is correct,

D is the closest one but it's missing a "ORDER BY timestamp DESC" to ensure to get only the latest record based in the timestamp

upvoted 8 times

✉  **Davijde13** 1 year ago

The question mention only duplicated data and nothing about taking only the latest ones. Therefore I assume there is no need to always take the latest, we should ensure we take only one record for each ID.

upvoted 3 times

✉  **Mamta072** 1 year, 7 months ago

Ans is D as Row number is the clause to fetch unique record from duplicate

upvoted 1 times

✉  **Arkon88** 1 year, 11 months ago

Answer: D

upvoted 1 times

✉  **samdhimal** 2 years ago

correct answer -> Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

You can use the ROW_NUMBER() to turn non-unique rows into unique rows and then delete the duplicate rows.

Reference:

https://www.mysqltutorial.org/mysql-window-functions/mysql-row_number-function/

upvoted 3 times

✉  **samdhimal** 1 year ago

When you are using BigQuery streaming inserts, there is no guarantee that data will only be sent once. However, you can use the ROW_NUMBER window function to ensure that duplicates are not included while interactively querying data. By using a PARTITION BY clause on the unique ID column, you can assign a unique number to each row within a result set, based on the order specified in the timestamp column. Then, a WHERE clause can be used to select only the row with the number 1. This will return the first row for each unique ID based on the timestamp column, which will ensure that duplicates are not included in your query results.

upvoted 4 times

✉  **samdhimal** 1 year ago

Option A is not recommended because it will only return the first row based on the timestamp column, it doesn't consider the unique ID you could have multiple rows with the same timestamp, and you will get one of them arbitrarily.

Option B is not recommended because it's used for aggregation, it doesn't return the first row for each unique ID based on the timestamp column.

Option C is not recommended because it's used for comparing rows, it doesn't return the first row for each unique ID based on the timestamp column.

upvoted 2 times

✉ **nofaruccio** 2 years, 2 months ago

Sorry, but IMHO no response is correct, because, in addition to making the ID field unique, it occurs consider the record with most recent timestamp

upvoted 1 times

✉ **MaxNRG** 2 years, 2 months ago

D is correct because it will just pick out a single row for each set of duplicates.

A is not correct because this will just return one row.

B is not correct because this doesn't get you the latest value, but will get you a sum of the same event over time which doesn't make too much sense if you have duplicates.

C is not correct because if you have events that are not duplicated, it will be excluded.

upvoted 6 times

✉ **anji007** 2 years, 3 months ago

Ans: D

upvoted 1 times

✉ **sumanshu** 2 years, 10 months ago

Vote for D.

Explanation:

<https://www.youtube.com/watch?v=ysArdMlmULo&list=PLQMsfKRZZviSLraRoqXulcMKFvIXQkHdA&index=3>

upvoted 11 times

✉ **awssp12345** 2 years, 7 months ago

very helpful. :)

upvoted 1 times

Question #9

Topic 1

Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

```
# Syntax error : Expected end of statement but got "-" at [4:11]
SELECT age
FROM
    bigquery-public-data.noaa_gsod.gsod
WHERE
    age != 99
    AND_TABLE_SUFFIX = '1929'
ORDER BY
    age DESC
```

Which table name will make the SQL statement work correctly?

- A. 'bigquery-public-data.noaa_gsod.gsod'
- B. bigquery-public-data.noaa_gsod.gsod*
- C. 'bigquery-public-data.noaa_gsod.gsod'*
- D. 'bigquery-public-data.noaa_gsod.gsod'*

Correct Answer: D

Reference:

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

Filtering selected tables using _TABLE_SUFFIX

To restrict a query so that it scans only a specified set of tables, use the `_TABLE_SUFFIX` pseudo column in a `WHERE` clause with a condition that is a constant expression.

The `_TABLE_SUFFIX` pseudo column contains the values matched by the table wildcard. For example, the previous sample query, which scans all tables from the 1940s, uses a table wildcard to represent the last digit of the year:

```
FROM
`bigquery-public-data.noaa_gsod.gsod194*`
```



Community vote distribution

D (75%)

B (25%)

✉ **Ender_H** Highly Voted 1 year, 4 months ago

None, the actual `'bigquery-public-data.noaa_gsod.gsod'` with back ticks at the beginning and at the end.

upvoted 25 times

✉ **jitvimal** 1 month ago

yes, I see from another source that actually ans D has to be backtick. Probably a problem when this web do data ingestion.

upvoted 1 times

✉ **Davijde13** 1 year ago

I suspect there has been some typo with copy-paste of the option D

upvoted 7 times

✉ **Jasar** 1 year, 2 months ago

yes it was the same , i hope im the real test we dont face any wrrors like that

upvoted 1 times

✉ **RT_G** Most Recent 2 months, 3 weeks ago

Selected Answer: D

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

The wildcard table name contains the special character (*), which means that you must enclose the wildcard table name in backtick () character. For example, the following query is valid because it uses backticks:

```
#standardSQL
/* Valid SQL query */
SELECT
max
FROM
`bigquery-public-data.noaa_gsod.gsod*`
WHERE
max != 9999.9 # code for missing data
AND _TABLE_SUFFIX = '1929'
ORDER BY
max DESC
upvoted 1 times
```

✉ **RT_G** 2 months, 3 weeks ago

Selected Answer: D

Agree with others - Option D

upvoted 1 times

✉  **axantroff** 3 months ago

Selected Answer: D

D. 'bigquery-public-data.noaa_gsod.gsod*' is the right answer with 1 typo
upvoted 1 times

✉  **rtcpost** 3 months, 1 week ago

Selected Answer: D

Option D (assuming to have backticks)

Refer: <https://cloud.google.com/bigquery/docs/querying-wildcard-tables>
The following query is NOT valid because it isn't properly quoted with backticks:
``

```
#standardSQL
/* Syntax error: Expected end of statement but got "-" at [4:11] */
SELECT
max
FROM
# missing backticks
bigquery-public-data.noaa_gsod.gsod*
WHERE
max != 9999.9 # code for missing data
AND _TABLE_SUFFIX = '1929'
ORDER BY
max DESC
``
```

upvoted 1 times

✉  **vaga1** 8 months, 2 weeks ago

Selected Answer: D

let's forget the fact that in BQ is used ` instead than ' which retrieves an error in any case. ` is called backquote, backtick, or left quote while ' simply an apostrophe. Let's consider ' to be ` in every answer, since moderators could have not been aware of such when they had received a question.

upvoted 1 times

✉  **vaga1** 8 months, 2 weeks ago

Who used BQ knows that the backquote is necessary only for the project name, while it can be used for the whole string, and necessary o when the project name contains special (special in this specific context) characters.

- is a special character. so
'bigquery-public-data'.noaa_gsod.gsod1940
would have worked too.

The question now turns out to be
'bigquery-public-data'.noaa_gsod.gsod*
still works or due to the * presence we need to write
'bigquery-public-data.noaa_gsod.gsod*'
?

I personally do not remember, and I do not have a BQ at my disposal at the moment.

But I know for sure that

'bigquery-public-data.noaa_gsod.gsod*'
works while
'bigquery-public-data'.noaa_gsod.gsod*
is not in the options.

upvoted 1 times

✉  **Pavaan** 8 months, 3 weeks ago

Answer is 'D'

Reference : <https://cloud.google.com/bigquery/docs/wildcard-table-reference>

Enclose table names with wildcards in backticks

The wildcard table name contains the special character (*), which means that you must enclose the wildcard table name in backtick () character
upvoted 3 times

✉  **Melampus** 9 months, 1 week ago

Selected Answer: B

bigquery-public-data.noaa_gsod.gsod* works

upvoted 2 times

✉  **hkhnhan** 10 months, 2 weeks ago

Selected Answer: B

should be B, the backtick at D answer is wrong ' instead of `
upvoted 1 times

✉  **hkhnhan** 10 months, 2 weeks ago

should be B, the backtick at D answer is wrong ' instead of `
upvoted 1 times

✉  **Zosby** 11 months, 1 week ago

D is correct

upvoted 1 times

✉  **priluft** 1 year, 4 months ago

Selected Answer: D

D. 'bigquery-public-data.noaa_gsod.gsod*'
upvoted 2 times

✉  **AWSandeep** 1 year, 4 months ago

Selected Answer: D

D. 'bigquery-public-data.noaa_gsod.gsod*'
upvoted 2 times

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

- A. Disable writes to certain tables.
- B. Restrict access to tables by role.
- C. Ensure that the data is encrypted at all times.
- D. Restrict BigQuery API access to approved users.
- E. Segregate data across multiple tables or databases.
- F. Use Google Stackdriver Audit Logging to determine policy violations.

Correct Answer: BDF

Community vote distribution

BDE (54%)	BDF (42%)	4%
-----------	-----------	----

 **samdhimal** Highly Voted 2 years ago

correct option -> B. Restrict access to tables by role.

Reference: <https://cloud.google.com/bigquery/docs/table-access-controls-intro>

correct option -> D. Restrict BigQuery API access to approved users.

Only approved users will have access which means other users will have minimum amount of information required to do their job.
Reference: <https://cloud.google.com/bigquery/docs/access-control>

correct option -> F. Use Google Stackdriver Audit Logging to determine policy violations.

Reference: <https://cloud.google.com/bigquery/docs/table-access-controls-intro#logging>

A. Disable writes to certain tables. ---> Read is still available(not minimal access)

C. Ensure that the data is encrypted at all times. ---> Data is encrypted by default.

E. Segregate data across multiple tables or databases. ---> Normalization is of no help here.

upvoted 36 times

✉  **samdhimal** 1 year ago

I was WRONG. I am not sure why so many upvotes lol.

I think this is the correct answer:

- B. Restrict access to tables by role.
- D. Restrict BigQuery API access to approved users.
- E. Segregate data across multiple tables or databases.

Restrict access to tables by role: You can use BigQuery's access controls to restrict access to specific tables based on user roles. This allows you to ensure that users can only access the data they need to do their job.

Restrict BigQuery API access to approved users: By using Cloud Identity and Access Management (IAM) you can control who has access to the BigQuery API, and what actions they are allowed to perform. This will help to ensure that only authorized users can access the data.

Segregate data across multiple tables or databases: You can use multiple tables or databases to separate different types of data, so that users can only access the data they need. This will prevent users from seeing data they shouldn't have access to.

upvoted 21 times

✉  **samdhimal** 1 year ago

Option A is incorrect because disabling writes to certain tables would prevent users from updating the data which is not in line with the goal of providing access to the minimum amount of information required to do their jobs.

Option C is incorrect because while data encryption is important for security it doesn't specifically help with providing users access to the minimum amount of information required to do their jobs.

Option F is incorrect because while Google Stackdriver Audit Logging can help to determine policy violations it does not help to enforce the access controls and segregation of data.

upvoted 5 times

✉  **[Removed]** 11 months, 2 weeks ago

There is no database in Bigquery, only datasets. I would pick it if it says "tables and datasets".

upvoted 4 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Correct: BDF

bigquery.tables.create Create new tables.

bigquery.tables.delete Delete tables.

bigquery.tables.export Export table data out of BigQuery.

bigquery.tables.get Get table metadata.

To get table data, you need bigquery.tables.getData.

bigquery.tables.getData Get table data. This permission is required for querying table data.

To get table metadata, you need bigquery.tables.get.

bigquery.tables.list List tables and metadata on tables.

bigquery.tables.setCategory Set policy tags in table schema.

bigquery.tables.update

Update table metadata.

To update table data, you need bigquery.tables.updateData.

bigquery.tables.updateData

Update table data.

To update table metadata, you need bigquery.tables.update.

upvoted 21 times

✉  **MaxNRG** Most Recent 1 month, 2 weeks ago

Selected Answer: BDE

You want to enforce this requirement with Google BigQuery -> BDE

upvoted 1 times

✉  **LaxmanTiwari** 1 month, 1 week ago

when u are planning to appear in exam ?

upvoted 1 times

✉  **RT_G** 2 months, 3 weeks ago

Selected Answer: BDF

BDF. We are fairly unanimous with options B and D. I'm going with F because it does help identifying policy violations which is also one aspect to be considered when designing access controls. Option D only indicates segregating into multiple tables and databases which may or may not help with controlling access leaving it open-ended for the architect to decide.

upvoted 1 times

✉️  **rocky48** 2 months, 4 weeks ago

Selected Answer: BDF

In Google BigQuery, you can organize and segregate data across multiple tables within the same dataset, but you cannot directly segregate data into separate databases. BigQuery uses a flat namespace structure where data is organized into datasets and tables within those datasets. Datasets are the highest level of organization within BigQuery.

So I'm sticking with BDF

upvoted 1 times

✉️  **rtcpost** 3 months, 1 week ago

Selected Answer: BDE

B. Restrict access to tables by role: You can define roles in BigQuery and grant specific permissions to these roles to control who can access particular tables.

D. Restrict BigQuery API access to approved users: You can control access to the BigQuery API and, consequently, to the underlying data by ensuring that only approved users or services can make API requests.

E. Segregate data across multiple tables or databases: You can separate data into different tables or databases based on user access requirements, which allows you to limit users' access to specific data sets.

These approaches, when used together, can help you enforce data access controls in a regulated environment. Options A, C, and F are also important considerations but are not direct methods for enforcing fine-grained access control to specific data.

upvoted 4 times

✉️  **RheaZhang** 3 months, 1 week ago

Selected Answer: BDE

B. Restrict access to tables by role.

D. Restrict BigQuery API access to approved users.

E. Segregate data across multiple tables or databases.

upvoted 1 times

✉️  **AnonymousPanda** 5 months, 1 week ago

Selected Answer: BDF

BDF as per other answers

upvoted 2 times

✉️  **nescafe7** 6 months ago

Selected Answer: BDF

Regarding E or F, opinions seem to be divided into two parts.

I think E is insufficient because it seems that appropriate conditions must be additionally described for table or dataset separation.

F is also emphasized in Google's official textbook. You need to ensure that it is operating well as set up through monitoring.

So, BDF!

upvoted 3 times

✉️  **Liting** 6 months, 1 week ago

Selected Answer: DEF

Why B is correct? Access control can only be applied on dataset and views, not on partitions and tables. => So it is not possible to restrict access to table, but only to dataset. Can someone help me understand why in this scenario B is correct?

upvoted 2 times

✉️  **FP77** 6 months ago

I was thinking the same thing. I thought dataset access gave you access to all tables within it, and that you couldn't restrict access on the table level.

upvoted 1 times

✉️  **KK0202** 7 months ago

Selected Answer: BDF

Option E says "...or databases". The data housing service in question is BigQuery and the context is to design that supports BigQuery access delegation. Seems random to include moving to another database as an option. If it did not mention databases and stopped at just tables, then E would also be the right option

upvoted 2 times

✉ Oleksandr0501 9 months, 1 week ago

B. Restrict access to tables by role: This approach can be used to control access to tables based on user roles. Access controls can be set at project, dataset, and table level, and roles can be customized to provide granular access controls to different groups of users.

D. Restrict BigQuery API access to approved users: This approach involves using IAM (Identity and Access Management) to control access to the BigQuery API. Access can be granted or revoked at the project or dataset level, and policies can be customized to control access based on user roles, IP addresses, and other factors.

E. Segregate data across multiple tables or databases: This approach involves breaking down large datasets into smaller, more manageable tables or databases. This helps to ensure that individual users have access only to the minimum amount of information required to do their job and reduces the risk of data breaches or policy violations.

upvoted 2 times

✉ juliosb 10 months, 2 weeks ago

Selected Answer: BDE

F won't avoid undesired access, only detect after it already happened.

E makes it easier to control access.

upvoted 5 times

✉ jin0 11 months, 1 week ago

I think BDF.

Segregating the table is needed to try to distribute into a few of dataset not only I said but it looks like complicated.

upvoted 1 times

✉ SidneyHod 11 months, 3 weeks ago

Selected Answer: BDE

I don't choose stack driver because it suits more for audit

upvoted 5 times

✉ Nirca 1 year ago

Selected Answer: BDF

answer F -> audit is not part of the question; so I would not go for it

answer E -> Yes - due to segregation you can add different rights to different data/users

upvoted 3 times

✉ desertlotus1211 1 year ago

It's not about an audit... it's about violation for least privileges. this will aid in correcting

upvoted 1 times

✉ Jackalski 1 year, 1 month ago

Selected Answer: BDE

answer F -> audit is not part of the question; so I would not go for it

answer E -> Yes - due to segregation you can add different rights to different data/users

Question #11

Topic 1

You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

⇒ No interaction by the user on the site for 1 hour

Has added more than \$30 worth of products to the basket

▪

⇒ Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- A. Use a fixed-time window with a duration of 60 minutes.
- B. Use a sliding time window with a duration of 60 minutes.
- C. Use a session window with a gap time duration of 60 minutes.
- D. Use a global window with a time based trigger with a delay of 60 minutes.

Correct Answer: C

Community vote distribution

 **vetaal** Highly Voted  1 year, 2 months ago

There are 3 windowing concepts in dataflow and each can be used for below use case
 1) Fixed window
 2) Sliding window and
 3) Session window.

Fixed window = any aggregation use cases, any batch analysis of data, relatively simple use cases.

Sliding window = Moving averages of data

Session window = user session data, click data and real time gaming analysis.

The question here is about user session data and hence session window.

Reference:

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

upvoted 20 times

 **RT_G** Most Recent  2 months, 3 weeks ago

Selected Answer: C

Session window since the question specifically talks about a specific user for a fixed duration.

upvoted 1 times

 **rocky48** 2 months, 3 weeks ago

Selected Answer: C

Session window = user session data, click data and real time gaming analysis.

upvoted 1 times

 **rtcp0st** 3 months, 1 week ago

Selected Answer: C

C. Use a session window with a gap time duration of 60 minutes.

A session window with a gap time duration of 60 minutes is appropriate for capturing user sessions where there has been no interaction on the site for 1 hour. It allows you to group user activity within a session, and when the session becomes inactive for the defined gap time, you can evaluate whether the user added more than \$30 worth of products to the basket and has not completed a transaction.

Options A and B (fixed-time window and sliding time window) might not capture the specific session-based criteria of inactivity and user interaction effectively.

Option D (global window with a time-based trigger) is not suitable for capturing user sessions and checking inactivity based on a specific time duration. It's more appropriate for cases where you need a single global view of the data.

upvoted 1 times

 **imran79** 3 months, 3 weeks ago

The basket abandonment system needs to determine if a user hasn't interacted with the site for 1 hour, has added products worth more than \$30, and hasn't completed a transaction. Therefore, the pipeline should account for periods of user activity and inactivity. A session-based windowing approach is appropriate here.

The right choice is:

C. Use a session window with a gap time duration of 60 minutes.

Session windows group data based on periods of activity and inactivity. If there's no interaction for the duration of the gap time (in this case, 60 minutes), a new window is started. This would help identify users who haven't interacted with the site for the specified duration, fulfilling the requirement for the basket abandonment system.

upvoted 2 times

 **MikkelRev** 4 months ago

Selected Answer: C

Session windows can divide a data stream representing user activity

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#session-windows>

upvoted 1 times

👤 **Chesternut999** 10 months, 3 weeks ago

Selected Answer: C

C - The best option for this use case.

upvoted 2 times

👤 **bha11111** 10 months, 3 weeks ago

Selected Answer: C

Session window is used for these type of scenario

upvoted 2 times

👤 **samdhimal** 11 months, 4 weeks ago

C. Use a session window with a gap time duration of 60 minutes.

A session window would be the most appropriate option to use in this case, as it would allow you to group events into sessions based on time gaps. In this case, the gap time of 60 minutes could be used to define a session, and if there is no interaction from the user for 60 minutes, a session would be created. By using a session window, you can track the behavior of the user during each session, including the products added to the basket, and determine if the conditions for sending a message have been met (i.e., the user has added more than \$30 worth of products to the basket and has not completed a transaction).

upvoted 2 times

👤 **kennyloo** 1 year, 3 months ago

Only C is feasible for this question

upvoted 1 times

👤 **AWSandeep** 1 year, 4 months ago

Selected Answer: C

C. Use a session window with a gap time duration of 60 minutes.

upvoted 1 times

Question #12

Topic 1

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data.

Which three steps should you take? (Choose three.)

- A. Load data into different partitions.
- B. Load data into a different dataset for each client.
- C. Put each client's BigQuery dataset into a different table.
- D. Restrict a client's dataset to approved users.
- E. Only allow a service account to access the datasets.
- F. Use the appropriate identity and access management (IAM) roles for each client's users.

Correct Answer: BDF

Community vote distribution

BDF (83%)

BEF (17%)

👤 **[Removed]** Highly Voted 3 years, 10 months ago
agreed B,D,F

upvoted 45 times

✉ **saurabh1805** Highly Voted 3 years, 5 months ago

My vota also goes for B,D,F

upvoted 12 times

✉ **philli1011** Most Recent 4 days, 22 hours ago

My Vote is BDF.

I was thinking BEF but the question shows that the Big Query warehouse will be accessed by both direct users and other applications, as preferred by each customer.

upvoted 1 times

✉ **SoloLeveling** 4 days, 22 hours ago

Selected Answer: BDF

agreed B,D,F

upvoted 1 times

✉ **RT_G** 2 months, 3 weeks ago

Selected Answer: BDF

Agree with others

upvoted 1 times

✉ **rtcp0st** 3 months, 1 week ago

Selected Answer: BDF

B. Load data into a different dataset for each client: Organize the data into separate datasets for each client. This ensures data isolation and simplifies access control.

D. Restrict a client's dataset to approved users: Implement access controls by specifying which users or groups are allowed to access each client's dataset. This restricts data access to approved users only.

F. Use the appropriate identity and access management (IAM) roles for each client's users: Assign IAM roles based on client-specific requirements to manage permissions effectively. IAM roles help control access at a more granular level, allowing you to tailor access to specific users or groups within each client's dataset.

These steps ensure that each client's data is separated, and access is controlled based on client-specific requirements. Options A, C, and E, while important in other contexts, are not sufficient on their own to ensure client data isolation and access control in a multi-client environment.

upvoted 4 times

✉ **imran79** 3 months, 3 weeks ago

the answers are B, D, and F.

To ensure that clients cannot see each other's data and have appropriate access, you would want to:

Segregate the data by client.

Restrict access to each client's data.

Use proper identity and access management techniques.

upvoted 2 times

✉ **suku2** 4 months, 2 weeks ago

Selected Answer: BDF

B. Load data into a different dataset for each client.

D. Restrict a client's dataset to approved users.

F. Use the appropriate identity and access management (IAM) roles for each client's users.

upvoted 3 times

✉ **Chi_Wang** 4 months, 2 weeks ago

Selected Answer: BDF

B,D,F is the answer

upvoted 2 times

✉ **elitedea** 10 months, 3 weeks ago

BDF is right

upvoted 4 times

✉  **samdhimal** 11 months, 4 weeks ago

Selected Answer: BDF

- B. Load data into a different dataset for each client.
- D. Restrict a client's dataset to approved users.
- F. Use the appropriate identity and access management (IAM) roles for each client's users.

By loading each client's data into a separate dataset, you ensure that each client's data is isolated from the data of other clients. Restricting access to each client's dataset to only approved users, as specified in D, further enhances data security by ensuring that only authorized users can access the data. By using appropriate IAM roles for each client's users, as specified in F, you can grant different levels of access to different clients and their users, ensuring that each client has only the level of access required for their specific needs.

upvoted 4 times

✉  **Nirca** 1 year ago

Selected Answer: BDF

B, D, F!

C - is technically wrong . tables are being logically stored in a single dataset.

A - Partitioning data is for improving performance. once you SQL (select) the table, you can not control the data being selected for the developer

upvoted 2 times

✉  **jin0** 11 months, 1 week ago

For C. What if thinking about that there are tables by clients? such as customer_clients_a table and giving IAM from each table to users??.

upvoted 1 times

✉  **Nirca** 1 year ago

B, D, F!

C - is technically wrong . tables are being logically stored in a single dataset.

A - Partitioning data is for improving performance. once you SQL (select) the table, you can not control the data being selected for the developer

upvoted 1 times

✉  **DeeData** 1 year, 1 month ago

Please why is DEF not correct?

upvoted 2 times

✉  **Kyr0** 1 year, 1 month ago

Selected Answer: BDF

Agree BDF

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: BEF

Why no E? E has a lot a sense to me, they have external analytical tools, and the best practice is to give access to external row service accounts and not throw user level.

upvoted 4 times

✉  **cloudyy** 11 months, 1 week ago

I hesitated over this too, but the question talks about direct access query so that's the reason for not choosing E.

upvoted 4 times

✉  **ler_mp** 1 year, 1 month ago

Yes, I also wonder why not E instead of D

upvoted 2 times

✉  **VincentMenzel** 6 months ago

Because the client might want a mixture of SAs and user accounts. Maybe they have a Big Data Team that wants to run queries and access the data with their account. Also SAs do not help with segregating the data

upvoted 1 times

✉  **lalli117** 1 year, 4 months ago

agreed, B,D,F

upvoted 1 times

Question #13

Topic 1

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.

Which Google database service should you use?

- A. Cloud SQL
- B. BigQuery
- C. Cloud Bigtable
- D. Cloud Datastore

Correct Answer: A

Community vote distribution

D (60%)	A (33%)	8%
---------	---------	----

✉  **DeepakKhattar**  3 years, 8 months ago

Initially, thinking D is the best answer but when question is re-read, A seems to be correct answer for following reasons

1. Is payment TRANSACTION -- DB should be able to perform full blown transaction (updating inventory, sales info etc, though not specified), not just ATOMIC which DataStore provides
 2. Its point-of-sale application, not ONLINE STORE where HIGH number of concurrent users ordering stuff.
 3. User Base could grow exponentially - again more users does not mean concurrent users and more processing power. Its only about storage.
 4. Do not want to Manage infrastructure scaling. - Cloud SQL can scale in terms of storage.
 5. CloudStore is poor selection for OLTP application
- Each property is index - so higher latency

Not sure, during exam 2 min is enough to think on various point..

I may be wrong or wrong path ... lets brainstrom..

upvoted 73 times

✉  **Blobby** 2 years, 5 months ago

Can't online be considered PoS? CloudSQL does have constraints for scaling and Google seem to specifically be selling Datastore for transactional use cases so going with D:

<https://cloud.google.com/datastore/docs/concepts/transactions>

upvoted 3 times

✉  **Blobby** 2 years, 4 months ago

Based on a re-read of the above comments and other later questions agree with A.
pls ignore my first answer.

upvoted 4 times

✉  **canon123** 2 years, 2 months ago

CloudSQL does not auto scale.

upvoted 10 times

✉  **BigQuery** 2 years, 1 month ago

<https://cloud.google.com/architecture/elastically-scaling-your-mysql-environment#objectives>

Please read. It can be configured for autoscaling.

upvoted 3 times

✉  **nkunwar** 1 year, 4 months ago

Cloud SQL doesn't AUTO SCALE, you need to manually edit, Please show where does it says AUTO SCALING

upvoted 1 times

✉  **hendrixlives** 2 years, 1 month ago

That link explains how to set MySQL autoscaling with Google Compute Engine instances (you install and manage MySQL on the VM). This can not be applied to Cloud SQL (managed service). In Cloud SQL, only the storage can be automatically increased, and changing the Cloud SQL instance size requires a manual edit of the instance type.

upvoted 4 times

✉  **MisuLava** 1 year, 3 months ago

yes, and that is ok since this is a point of sale. an exponential increase in number of clients still means reduced parallel processing (how many customers can buy in the very same time) so an increase in memory and CPU is very unlikely to be necessary. yes, an exponential increase in the number of customers means more memory, and more storage, which in Cloud SQL increases automatically.

upvoted 3 times

✉  **jvg637** Highly Voted 3 years, 10 months ago

D seems to be the right one. Cloud SQL doesn't automatically scale
upvoted 37 times

✉  **BigQuery** 2 years, 1 month ago

Cloud SQL does scale automatically. THERE IS A SETTING WHERE YOU DEFINE INCREASE MEMORY SPACE WHEN IT REACHED 70%

https://cloud.google.com/sql/docs/features#features_3

Here it say's

- > Fully managed SQL Server databases in the cloud.
- > Custom machine types with up to 624 GB of RAM and 96 CPUs.
- > Up to 64 TB of storage available, with the ability to automatically increase storage size as needed.

upvoted 2 times

✉  **hendrixlives** 2 years, 1 month ago

Storage scale is automatic (e.g. you begin with a 50GB disk and it grows automatically as needed), but the instance size (CPU/memory) will be the same. The question states that the user base may increase exponentially. Even if you have enough disk space to store all your user data, the increase in users will cause problems if your instance (CPU/memory) is too small, since the instance will not be able to process all the queries at the required speed.

upvoted 4 times

✉  **imsaikat50** 1 year, 1 month ago

I believe the key point is it's a POS, not an e-commerce. Keeping that in mind, exponential user increase in POS might not mean concurrent user increase, which could be a huge consideration in case of it being e-commerce.

I would rather go with 'Cloud SQL' as the best answer.

upvoted 2 times

✉  **TVH_Data_Engineer** Most Recent 1 month, 2 weeks ago

Selected Answer: D

Cloud Datastore (now part of Google Cloud Firestore in Datastore mode) is designed for high scalability and ease of management for applications. It is a NoSQL document database built for automatic scaling, high performance, and ease of application development. It's serverless, meaning it handles the scaling, performance, and management automatically, fitting your requirement of not wanting to manage infrastructure scaling.

Cloud SQL, while a fully-managed relational database service that makes it easy to set up, manage, and administer your SQL databases, is not automatically scalable as Datastore. It's better suited for applications that require a traditional relational database.

upvoted 1 times

✉  **RT_G** 2 months, 3 weeks ago

Selected Answer: D

D - Because Datastore supports massive scaling and ACID transactions which are two primary considerations in this scenario.

upvoted 1 times

✉  **rocky48** 2 months, 3 weeks ago

Selected Answer: D

B - not an option

C - lack of ACID transactions

A - lack of resource automatic scalability

D - (correct, IMHO) support ACID, suitable for OLTP and scalable enough

upvoted 1 times

✉  **axantroff** 3 months ago

Selected Answer: D

B - not an option

C - lack of ACID transactions

A - lack of resource automatic scalability

D - (correct, IMHO) support ACID, suitable for OLTP and scalable enough

upvoted 1 times

✉  **imran79** 3 months, 3 weeks ago

For a point-of-sale application where you anticipate exponential growth and want to ensure seamless scalability without managing the infrastructure scaling, the best choice among the provided options is:

D. Cloud Datastore

Cloud Datastore is a NoSQL database service that's built for web, mobile, and IoT applications. It provides automatic scaling, high performance and ease of application development, making it a suitable choice for applications where the user base could grow exponentially and where you don't want to manually handle infrastructure scaling.

upvoted 2 times

✉  **gudguy1a** 4 months, 3 weeks ago

Selected Answer: C

Bigtable is the better answer.

Neither Cloud SQL & Datastore (Firestore) can scale to exponential level....

upvoted 1 times

✉  **GCP_PDE_AG** 5 months, 3 weeks ago

A seems to be correct.

Use case here is "process payment transactions in a point-of-sale application". So this is OLTP. Datastore is not a relational database, and it is an effective solution for analytic data. If you need a relational database with full SQL support for an online transaction processing (OLTP) system consider Cloud SQL.

<https://cloud.google.com/datastore/docs/concepts/overview>

upvoted 3 times

✉  **pulse008** 5 months, 3 weeks ago

Selected Answer: D

D is correct because you do not want to manage scaling in this case. Cloud SQL can also scale but it needs involvement unlike Datastore which is completely automatic.

upvoted 2 times

✉  **bhavaneesh** 5 months, 3 weeks ago

Do you by any chance have the contributor access - I need questions 205-209 :(Please

upvoted 1 times

✉  **alihabib** 5 months, 3 weeks ago

Its D, though Datastore is NoSQL, but it supports ACID transaction concepts.

upvoted 1 times

✉  **yash12** 6 months ago

Correct Answer is D, Cloud Datastore is used for PoS

<https://cloud.google.com/blog/topics/developers-practitioners/your-google-cloud-database-options-explained>

upvoted 2 times

✉  **Vipul1600** 6 months, 1 week ago

Datastore automatically scales and is highly optimized for ACID transactions.

upvoted 1 times

✉  **hpvb** 7 months, 1 week ago

cloud sql --> incorrect

<https://cloud.google.com/sql#section-8> --> you will need to manually do scalability

Scalability --> Easily scale up as your data grows—add processor cores, RAM and storage, and scale out by adding read replicas to handle increasing read traffic. Read replicas support high availability, can have their own read replicas, and can be located across regions and platforms.

big query , big table --> doesn't support ACID

Cloud Datastore --> CORRECT (<https://cloud.google.com/datastore/>) check features section

ACID transactions

Ensure the integrity of your data by executing multiple datastore operations in a single transaction with ACID characteristics, so all the grouped operations succeed or all fail.

Fully managed

Datastore is fully managed, which means Google automatically handles sharding and replication in order to provide you with a highly available and consistent database.

upvoted 3 times

✉  **hpvb** 7 months, 1 week ago

BigQuery --> OLAP , DataWareHouse
Cloud BigTable --> NOSQL, Doesn't support ACID.
Cloud DataStore --> NO SQL
upvoted 2 times

✉  **tal_** 7 months, 3 weeks ago

Bard said it's cloud sql:
"If Cloud Spanner is not an option, Cloud SQL is a good choice for your point-of-sale application. It offers a managed service, flexible pricing, security, and global availability."
upvoted 2 times

✉  **biswa_b** 8 months ago

Selected Answer: A

It should be cloud SQL
upvoted 3 times

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

- A. There are very few occurrences of mutations relative to normal samples.
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.
- D. You expect future mutations to have similar features to the mutated samples in the database.
- E. You already have labels for which samples are mutated and which are normal in the database.

Correct Answer: BC

Community vote distribution

AD (49%)	AC (43%)	8%
----------	----------	----

✉  **jvg637** Highly Voted 3 years, 10 months ago

I think that AD makes more sense. D is the explanation you gave. In the rest, A makes more sense, in any anomaly detection algorithm it is assumed a priori that you have much more "normal" samples than mutated ones, so that you can model normal patterns and detect patterns that are "off" that normal pattern. For that you will always need the no. of normal samples to be much bigger than the no. of mutated samples.

upvoted 70 times

✉  **BigQuery** 2 years, 1 month ago

Guys its A & C.

Anomaly detection has two basic assumptions:

- >Anomalies only occur very rarely in the data. (a)
- >Their features differ from the normal instances significantly. (c)

link -> <https://towardsdatascience.com/anomaly-detection-for-dummies-15f148e559c1#:~:text=Unsupervised%20Anomaly%20Detection%20for%20Univariate%20%26%20Multivariate%20Data.&text=Anomaly>

0detection%20has%20two%20basic,from%20the%20normal%20instances%20significantly.

upvoted 17 times

✉  **szefco** 2 years, 1 month ago

I don't agree on C. Anomaly detection assumes "Their features differ from the NORMAL INSTANCES significantly" and in the C option you have:

"You expect future mutations to have different features from the MUTATED SAMPLES IN THE DATABASE".

IMHO Answer D fits better: "D. You expect future mutations to have similar features to the mutated samples in the database." - in other words: Expect future anomalies to be similar to the anomalies that we already have in database

upvoted 26 times

✉  **jvg637** Highly Voted 3 years, 10 months ago

A instead of B:

"anomaly detection (also outlier detection[1]) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data

upvoted 21 times

✉  **tdum76000** Most Recent 1 month, 1 week ago

Selected Answer: AC

As A is a good answer, i'd like to give my point of view on the second right answer. I initially thought D was the correct one, as you would normally train your model to detect mutations seen in the training dataset. But the goal of unsupervised learning is to detect unidentified patterns. If you were sure the mutations would always look the same, you'd rather use supervised learning and labels the "normal" and "mutated" tissues, which would result in better performances in my point of view.

upvoted 1 times

✉  **spicebits** 2 months, 3 weeks ago

Selected Answer: AC

Unsupervised anomaly detection is best for scenarios without labels or when the anomalies are unknown or ever-changing

upvoted 1 times

✉️  **rocky48** 2 months, 3 weeks ago

Selected Answer: AD

A. There are very few occurrences of mutations relative to normal samples. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying rare events or anomalies in large amounts of data. By training the algorithm on the normal tissue samples in the database, it can then identify new tissue samples that have different features from the normal samples and classify them as mutated.

D. You expect future mutations to have similar features to the mutated samples in the database. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying patterns or anomalies in the data. By training the algorithm on the mutated tissue samples in the database, it can then identify new tissue samples that have similar features and classify them as mutated.

upvoted 1 times

✉️  **axantroff** 3 months ago

Selected Answer: AC

AC; might also be interesting - <https://towardsdatascience.com/unsupervised-learning-for-anomaly-detection-44c55a96b8c1> as comments below

upvoted 2 times

✉️  **imran79** 3 months, 3 weeks ago

For unsupervised anomaly detection to be effective, it works best when anomalies (or mutations in this case) are rare compared to normal instances. Moreover, if future mutations are expected to have different features from those currently in the database, an unsupervised method would be beneficial since it doesn't rely on previously seen patterns of mutations.

The two characteristics that support the use of an unsupervised anomaly detection method in this scenario are:

- A. There are very few occurrences of mutations relative to normal samples.
- C. You expect future mutations to have different features from the mutated samples in the database.

upvoted 2 times

✉️  **gaurav0480** 5 months, 1 week ago

A is definitely true. Next comes the tricky difference between C & D. We can in fact even use supervised learning for case D where future mutations are similar to mutations in the training sample given that we have enough samples in the anomalous class then over-sample the anomalous class and under-sample the other class. Therefore I am inclined to choose C instead of D.

upvoted 1 times

✉️  **Mark_86** 6 months, 1 week ago

Selected Answer: AC

A & C

100% sure, as you would only use unsupervised learning if you cannot supervise your algorithm. The other answers imply that you have enough expectedly similar mutations to supervise on.

upvoted 1 times

✉️  **Mathew106** 6 months, 1 week ago

Answers B and C are both dumb, sorry to say. There are different approaches to anomaly detection. Some expect different features from the training dataset anomalies and some don't. If you cluster the training data and assign an anomaly label to any data point in an anomaly cluster then you expect them to have similar features. If you disregard the anomaly clusters and you simply set a rule saying "a data point is an anomaly if it's further away from X than the clusters 1,2,3 with healthy tissues, then you don't care about having similar features, as long as they are not similar to the healthy tissues.

upvoted 1 times

✉️  **azmiozgen** 6 months, 3 weeks ago

Selected Answer: AD

D should be correct. You expect future samples will correlate with the training samples. That's the whole point of learning procedure. If you do not expect that they have similar features, then why would you use features in the training samples in the first place? A is also correct, since anomaly labels would be seen rarely.

upvoted 4 times

✉️  **cchen8181** 8 months, 2 weeks ago

Selected Answer: AC

I would choose A and C.

Not B because mutations should be rare.

Not D because mutations can be unpredictable and if D were true it would point to supervised learning.

Not E since it would point to supervised learning.

upvoted 2 times

✉  **despee** 9 months ago

Selected Answer: AC

Guys if there are equals in the DB, it becomes a classification problem not an anomaly detection.

upvoted 3 times

✉  **momosoundz** 8 months, 4 weeks ago

agree!! :)

upvoted 1 times

✉  **juliosb** 10 months, 2 weeks ago

Selected Answer: AC

The question is more about why *unsupervised* *anomaly detection*.

A explains the *anomaly detection*

C explains why *unsupervised*

If the mutations were like the database you could simply do supervised learning.

upvoted 3 times

✉  **shabfat** 10 months, 2 weeks ago

I think it should be D instead of C, because for a good clustering you want the intra cluster distance to be low --> that would imply you have similar mutations.

upvoted 2 times

✉  **charline** 10 months, 3 weeks ago

Selected Answer: AD

D. You expect future mutations to have similar features to the MUTATED samples in the database.

upvoted 3 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: AD

A and D makes more sense

upvoted 3 times

✉  **Aaronn14** 11 months ago

AC. C rather than D because if the relation between mutation and features does not change, then there is no reason to use unsupervised learning. E.g., consider logistic regression vs clustering algorithm based on L2 norm. Logistic regression will pick up only those features, which are associated with a mutation in your particular sample. If a feature, not associated with a mutation is a training sample, is 5 sigmas outside its mean, logistic regression will not detect anything. On the other hand, clustering will pick up any observation where any feature is 5 sigmas outside its mean.

upvoted 3 times

data. How can you adjust your application design?

- A. Re-write the application to load accumulated data every 2 minutes.
- B. Convert the streaming insert code to batch load for individual messages.
- C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
- D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Correct Answer: D

Community vote distribution

D (72%)	B (26%)	2%
---------	---------	----

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: D

Description: The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issue. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written to storage

upvoted 54 times

 **msaqib934** 1 year, 9 months ago

How to estimate average latency speed.

upvoted 1 times

 **deavid** 1 year, 3 months ago

Cloud Monitoring?

upvoted 1 times

 **[Removed]** Highly Voted 3 years, 10 months ago

D - speaks about near real-time approach. None other.

upvoted 22 times

 **philli1011** Most Recent 4 days, 20 hours ago

Answer: D

I agree with the first part of the D answer, but for the second part, I don't know how they came about the 2 mins, is it from a calculation?

upvoted 1 times

 **imran79** 3 months, 3 weeks ago

A. Re-write the application to load accumulated data every 2 minutes.

By accumulating data and performing a batch load every 2 minutes, you can reduce the potential inconsistency caused by streaming inserts. While this introduces a slight delay, it provides a more consistent approach than streaming each individual message. This method can still meet the near real-time requirement, and the slight delay is often acceptable in scenarios where data consistency is paramount.

upvoted 2 times

 **Nirca** 4 months ago

Selected Answer: B

BBBBB is the only option

upvoted 1 times

 **ckanaar** 4 months, 1 week ago

I'd argue that this question became outdated with the introduction of the BigQuery Storage Write API:

<https://cloud.google.com/bigquery/docs/write-api>

upvoted 3 times

 **axantroff** 3 months ago

Good point

upvoted 1 times

 **NeoNitin** 4 months, 3 weeks ago

D, Thank you Exam topic : Passed the exam in august and I can say examtopic is help me lot, topic 1 is enough for the exam, just last week I received welcome kit from google for PDE exam one google cloud cup. if you need all question any help reach out to me neonitin6attherategoogledotcom

upvoted 2 times

👤 **klughund** 6 months ago

Streaming inserts in BigQuery are not immediately available to be queried, which is causing the weak consistency you're observing. A better approach is to batch the data and load it at regular intervals. Loading the data every two minutes is still relatively real-time, and it should help solve the consistency problem.

Answer A.

upvoted 2 times

👤 **NeoNitin** 6 months ago

All the options aim to address the challenge of strong consistency in the data and potential missing data that may occur with streaming insert. Each approach has its pros and cons, so the best choice depends on the specific needs and requirements of the application. It's like having different strategies for keeping track of all the fun things the kids do and say on the playground, making sure nothing gets left behind!

upvoted 1 times

👤 **WillemHendr** 7 months, 3 weeks ago

Streaming Inserts is marked as Legacy now.

<https://cloud.google.com/bigquery/docs/streaming-data-into-bigquery#dataavailability>

The documentation is hinting on it can take up to 90 minutes to process the buffered data.

This question is testing if you are aware of the possible long times the buffer can build up.

upvoted 2 times

👤 **izekc** 9 months, 1 week ago

Selected Answer: B

In my experience, estimation in D is not a technical solution. it is just a guess solution.

You might still get caught when loading get higher and easily take twice as long latency, then problem occur again.

So for a more permanent solution, you should definitely go with B

upvoted 3 times

👤 **bha11111** 10 months, 3 weeks ago

Selected Answer: D

1st line of question requires near real time queries so D is the best option as batch load is never near real time

upvoted 3 times

👤 **musumusu** 11 months, 1 week ago

Answer: D

What to learn or look for

1. In-Flight data = (Real Time data, i.e still in streaming pipeline and not landed in BigQuery)

2. Dataflow (assume in best case) streaming pipeline is running to send data to Bigquery.

Why not option B: change streaming to batch upload is not business requirement, we have to stuck to streaming and real time analysis.

Option D: make bigquery run after waiting for sometime (twice here), How will you do it?

- there is not setting in BQ to do it, right!. So, adjust it in your pipeline (dataflow)

- For example, add Fixed window, and you want to execute aggregation query after 2 min.

Code

```
``pipeline.apply(...)  
.apply(Window.<TableRow>into(FixedWindows.of(Duration.standardMinutes(2))))  
.apply(BigQueryIO.writeTableRows())  
.to("my_dataset.my_table")  
``
```

upvoted 3 times

👤 **techtitan** 11 months, 2 weeks ago

Selected Answer: D

near realtime

upvoted 3 times

👤 **donbigi** 11 months, 4 weeks ago

Selected Answer: D

the answer is AD

upvoted 2 times

👤 **korntewin** 1 year ago

Selected Answer: D

The streaming mode may be in pending mode or buffered mode where the streaming data is not immediately available before committing or flushing. Thus, we need to wait before the data will be available. Or else we need to switch to committed mode (which is not present in the choices).

upvoted 2 times

👤 **PrashantGupta1616** 1 year, 1 month ago

Selected Answer: D

D make more sense
upvoted 2 times

Question #16

Topic 1

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- A. Use Google Stackdriver Audit Logs to review data access.
- B. Get the identity and access management (IAM) policy of each table
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Correct Answer: A

Community vote distribution

A (90%)

10%

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Correct: A.
B - Wrong, permissions are on dataset level
C & D - won't say anything to influence the access level audit.
upvoted 28 times

👤 **Mathew106** 6 months, 2 weeks ago

While A is correct, B is possible today.
<https://cloud.google.com/bigquery/docs/control-access-to-resources-iam>
upvoted 5 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A
Description: First we need to know who is accessing what then we can create suitable policies. Stackdriver is used to track access logs for Bigquery
upvoted 18 times

👤 **philli1011** Most Recent 4 days, 20 hours ago

A is the answer.
But recently, I think Dataplex is used for data governance .
upvoted 1 times

✉️  **RT_G** 2 months, 3 weeks ago

Selected Answer: A

A - Since the question is to discover what everyone is doing. Also the question has indicated that no security policies have been implemented upvoted 1 times

✉️  **rocky48** 2 months, 3 weeks ago

Selected Answer: A

A. Use Google Stackdriver Audit Logs to review data access.

Reviewing the audit logs provides visibility into who is accessing your data, when they are doing so, and what actions they are taking within BigQuery. This is crucial for understanding current data usage and potential security risks.

Option B (getting the IAM policy of each table) is important but more focused on controlling access rather than discovering what everyone is currently doing.

Option C (using Stackdriver Monitoring to see query slots usage) can help with monitoring and optimizing your BigQuery usage but doesn't provide a comprehensive view of what users are doing with the data.

Option D (using the Google Cloud Billing API) is more related to tracking billing information rather than understanding what users are doing with the data.

upvoted 2 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: A

To begin securing your data warehouse in Google BigQuery and gain insights into what everyone is doing with the datasets, the first step you should take is:

A. Use Google Stackdriver Audit Logs to review data access.

Reviewing the audit logs provides visibility into who is accessing your data, when they are doing so, and what actions they are taking within BigQuery. This is crucial for understanding current data usage and potential security risks.

Option B (getting the IAM policy of each table) is important but more focused on controlling access rather than discovering what everyone is currently doing.

Option C (using Stackdriver Monitoring to see query slots usage) can help with monitoring and optimizing your BigQuery usage but doesn't provide a comprehensive view of what users are doing with the data.

Option D (using the Google Cloud Billing API) is more related to tracking billing information rather than understanding what users are doing with the data.

upvoted 3 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: A

A. Use Google Stackdriver Audit Logs to review data access.

Reviewing the audit logs provides visibility into who is accessing your data when they are doing so, and what actions they are taking within BigQuery. This is crucial for understanding current data usage and potential security risks.

upvoted 1 times

✉️  **imran79** 3 months, 3 weeks ago

A. Use Google Stackdriver Audit Logs to review data access.

Stackdriver Audit Logs provide detailed logs on who accessed what resources and when, including data in BigQuery. Reviewing these logs will give you insight into which users and service accounts are accessing datasets, what operations they are performing, and when these accesses occur. This would be a crucial first step in understanding current usage and subsequently in crafting a security policy.

upvoted 1 times

✉️  **suku2** 4 months, 2 weeks ago

Selected Answer: A

Stackdriver audit logs is where we will view which datasets are being accessed by whom

upvoted 1 times

✉  **NeoNitin** 6 months ago

A. Use Google Stackdriver Audit Logs to review data access.

In this scenario, you have been asked to secure the data warehouse in Google BigQuery. To do that, you first need to understand what everyone is doing with the data, i.e., who is accessing it and what actions they are performing. Google Stackdriver Audit Logs can provide you with a detailed record of all the data access and actions taken by users in Google BigQuery. It's like having a logbook that keeps track of who enters the library, which books they read, and what they do with the books.

C just give how many people accessing the same dataset at given time

C. Another tool you have is called "Stackdriver Monitoring." It helps you see how many people are using the library at the same time. It's like knowing how many readers are in the library at any given moment.

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: A

In order to take a decision you need to analyze the access logs

upvoted 1 times

✉  **Aparna_22** 11 months, 1 week ago

In exam topics in come to study for my PDE exam which is on 26 feb please suggest me do we need to do all questions to pass the exam

upvoted 2 times

✉  **Subhajeetpal** 11 months, 1 week ago

Mine is also on 26 FEB

upvoted 1 times

✉  **Akshat90** 11 months ago

Kindly Let us know if you get the questions from Exam topics please

upvoted 1 times

✉  **AshokPalle** 11 months, 1 week ago

Mine is tomorrow

upvoted 1 times

✉  **niketd** 11 months, 2 weeks ago

"Discover what everyone is doing" will happen through Audit logs, hence correct answer is A

upvoted 1 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: A

"...to secure the data warehouse" is to list all tables/views/Mviews VS. who is accessing these objects. Slot info is not relevant.

upvoted 1 times

✉  **fedebos8** 1 year, 2 months ago

Selected Answer: A

A is correct.

upvoted 1 times

✉  **nkunwar** 1 year, 4 months ago

Selected Answer: A

Audit log ..logs activities against resources , is the best place to discover about activities against BQ

upvoted 1 times

✉  **John_Pongthorn** 1 year, 4 months ago

Selected Answer: A

A for sure

upvoted 1 times

minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Correct Answer: D

Community vote distribution

D (100%)

 **[Removed]**  3 years, 10 months ago

Answer: D

Description: Dataproc is used to migrate Hadoop and Spark jobs on GCP. Dataproc with GCS connected through Google Cloud Storage connector helps store data after the life of the cluster. When the job is high I/O intensive, then we need to create a small persistent disk.
upvoted 55 times

 **[Removed]**  3 years, 10 months ago

Correct : D

upvoted 17 times

 **rtcpost**  3 months, 1 week ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

Google Cloud Dataproc is a managed Hadoop and Spark service that allows you to easily create and manage Hadoop clusters in the cloud. By using the Google Cloud Storage connector, you can persist data in Google Cloud Storage, which provides durable storage beyond the cluster lifecycle. This approach ensures data is retained even if the cluster is terminated, and it allows you to reuse your existing Hadoop jobs.

Option B (Creating a Dataproc cluster that uses persistent disks for HDFS) is another valid choice. However, using Google Cloud Storage for data storage and processing is often more cost-effective and scalable, especially when migrating to the cloud.

Options A, C, and E do not take full advantage of Google Cloud's services and the benefits of cloud-native data storage and processing with Google Cloud Storage and Dataproc.

upvoted 2 times

 **imran79** 3 months, 3 weeks ago

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

Here's why:

Cloud Dataproc allows you to run Apache Hadoop jobs with minimal management. It is a managed Hadoop service.

Using the Google Cloud Storage (GCS) connector, Dataproc can access data stored in GCS, which allows data persistence beyond the life of the cluster. This means that even if the cluster is deleted, the data in GCS remains intact. Moreover, using GCS is often cheaper and more durable than using HDFS on persistent disks.

upvoted 1 times

 **suku2** 4 months, 2 weeks ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
Dataproc clusters can be created to lift and shift existing Hadoop jobs
Data stored in Google Cloud Storage extends beyond the life of a Dataproc cluster.
upvoted 1 times

 **kshehadyx** 4 months, 2 weeks ago

Correct D

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: D

Hadoop --> Dataproc Persistent storage after the processing --> GCS
upvoted 2 times

✉  **samdhimal** 1 year ago

Selected Answer: D

D Seems right. Cloud storage can be used to achieve data storage even after the life of cluster.
upvoted 1 times

✉  **korntewin** 1 year ago

Selected Answer: D

The answer is D! Dataproc have no need for use to manage the infra and cloudstorage also no need for us to manage too!
upvoted 1 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
upvoted 1 times

✉  **assU2** 1 year, 2 months ago

Seems like it is D. <https://cloud.google.com/dataproc/docs/concepts/dataproc-hdfs>
Never saw they mentioned persistent disks, although they are not deleted with the clusters...
upvoted 1 times

✉  **assU2** 1 year, 2 months ago

although:

By default, when no local SSDs are provided, HDFS data and intermediate shuffle data is stored on VM boot disks, which are Persistent Disks
upvoted 1 times

✉  **assU2** 1 year, 2 months ago

and it says that only VM Boot disks are deleted when the cluster is deleted.
upvoted 2 times

✉  **achafill** 1 year, 3 months ago

Selected Answer: D

Correct Answer : D
upvoted 1 times

✉  **nkunwar** 1 year, 4 months ago

Selected Answer: D

Dataproc cluster set up will be ephemeral to run HDFS Jobs and can be killed after Job execution killing persistent storage with cluster
upvoted 1 times

✉  **crisimenjivar** 1 year, 5 months ago

Answer: D
upvoted 1 times

✉  **Asheesh1909** 1 year, 7 months ago

Isn't it A and D both dataflow for reusable jobs and gcs for data persistence?
upvoted 1 times

✉  **kmaiti** 1 year, 9 months ago

Selected Answer: D

Two key points:
Managed hadoop cluster - dataproc
Persistent storage: GCS (dataproc uses gcs connector to connect to gcs)
upvoted 2 times

✉  **deep_ROOT** 2 years ago

Selected Answer: D

This question is from Practice Test by Google, they gave D as right answer
upvoted 2 times

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

- A. Supervised learning to determine which transactions are most likely to be fraudulent.
- B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
- C. Clustering to divide the transactions into N categories based on feature similarity.
- D. Supervised learning to predict the location of a transaction.
- E. Reinforcement learning to predict the location of a transaction.
- F. Unsupervised learning to predict the location of a transaction.

Correct Answer: BCE

Community vote distribution

BCD (86%)

7%

 **jvg637** Highly Voted 3 years, 10 months ago

BCD makes more sense to me. Its for sure not unsupervised, since locations are in the data already. Reinforcement also doesn't fit, as there n and no interactions with data from the observer.

upvoted 68 times

 **sergio6** 2 years, 5 months ago

D make sense, but i have a doubt: location is a discrete value (no regression), so a multiclass classification model should be applied ... to predict locations?

upvoted 4 times

 **hellofrnds** 2 years, 3 months ago

yes. multiclass classification model should be applied

upvoted 4 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: B, C, D

Description: Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features

upvoted 41 times

 **TVH_Data_Engineer** Most Recent 2 months, 1 week ago

Options B, E, and F are not as suitable for the given scenario:

B. Unsupervised learning to determine which transactions are most likely to be fraudulent.

Unsupervised learning, while useful for anomaly detection, might not be as effective for fraud detection without labeled data indicating which transactions are fraudulent.

E. Reinforcement learning to predict the location of a transaction.

Reinforcement learning is more suitable for scenarios where an agent learns to make decisions through trial and error, which doesn't seem to align with predicting transaction locations.

F. Unsupervised learning to predict the location of a transaction.

Unsupervised learning typically doesn't involve predicting specific values (like location) without labeled data for training.

In summary, A, C, and D are the most appropriate machine learning applications for investigating the provided bank transactions dataset.

upvoted 2 times

✉  **rocky48** 2 months, 3 weeks ago

Selected Answer: BCD

Answer: BCD

upvoted 1 times

✉  **Waqasghaloo** 4 months, 2 weeks ago

Location is already given as attribute so what value is served with predicting location?

upvoted 1 times

✉  **youare87** 5 months, 3 weeks ago

A, B: Data features without the definition of fraudulent, so we can not obtain the answer even if using the unsupervised learning.

C: Kmeans solve this.

D: logistic regression. Just put the location into target.

E: Give the positive reward when the model predicts correct location.

F: Same as C. Use all features but locations, and use similarity to predict new data.

upvoted 1 times

✉  **xiaofeng_0226** 5 months, 3 weeks ago

Selected Answer: BCD

Absolutely

upvoted 1 times

✉  **Dip1994** 5 months, 4 weeks ago

Selected Answer: BCD

makes more sense

upvoted 1 times

✉  **Mark_86** 6 months, 1 week ago

Selected Answer: BCD

BCD make sense and does not require anything that is not given in the question data.

upvoted 1 times

✉  **hpvb** 7 months, 1 week ago

should be BCD.

E doesn't make sense because reinforcement learning is used only when you want to reach an optimal solution to a problem. Like optimized solution for reaching point A to point B and etc. You don't need reinforcement learning to predict a location.

upvoted 1 times

✉  **[Removed]** 9 months ago
NO UFRONT PAYMENT!!

GET CERTIFIED.
100%PASS GUARANTEED.

1. COMPTIA (network+ security+)

2: PASS ANY MICROSOFT EXAM AND PAY AFTER PASS RESULTS

3: IAPP Certifications
(CIPP/E CIPM, CIPT)

4: ISACA certifications (CISA,CISM/ CRISC)

5: EC-COUNCIL Certification (CEH , CCISO)

6: PMI (PMP/CAPM/ACP/PBA ,RMP)

8: CIA,IFRS, CERTIFICATIONS

9: ACCA,CFA,ICAEW certifications

12. APICS CERTIFICATIONS, CSCP, CPIM, CLTD

Book for online proctor exam and we'll remotely take the exam for you. Pay us after confirmation of PASSED results
ITTCA.org

WhatsApp +1(409)223 7790
upvoted 1 times

✉  **Jarek7** 9 months ago

Selected Answer: BCE

I'd go for BCE instead of BCD, assuming that location is geographical location or the geographical location can be found from location using some side input.

With so limited features (there is no even transaction date/time given!) and so huge and variant label as location it is impossible to get any convergence in supervised learning(D).

Reinforced learning(E) with reinforcement inversely proportional to the distance squared between predicted and the real location could get some reasonable results.

upvoted 2 times

✉  **budgier** 9 months ago

According to GPT A,B,C
upvoted 1 times

✉  **FP77** 5 months, 1 week ago

Well, GPT is stupid then
upvoted 4 times

✉  **juliobs** 10 months, 2 weeks ago

Selected Answer: BCD

BCD. E does not make sense.
upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: BCD

BCD is correct
upvoted 1 times

✉  **betterForGo** 10 months, 3 weeks ago

I would like to choose ABC.
upvoted 1 times

Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

- A. Put the data into Google Cloud Storage.
- B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Correct Answer: A

Community vote distribution

A (100%)

 **[Removed]**  3 years, 10 months ago

Answer: A

Description: First rule of dataproc is to keep data in GCS

upvoted 45 times

 **[Removed]**  3 years, 10 months ago

A - Google recommendation.

B - Just talks about use of PVM for Dataproc clusters and no mention of the storage.

A & B both provides cost effective approach. But B lacks completeness in the option.

Should go with A.

upvoted 14 times

 **Vullibabu**  3 weeks, 3 days ago

You are most of the people looking at like for like migration would require 50TB persistent storage but missing to look at CIO concern about c of block storage...considering CIO concern the option here is cloud storage... moreover that is recommended as well ..

upvoted 1 times

 **imran79** 3 months, 3 weeks ago

Option A: Put the data into Google Cloud Storage.

This is the best option. Google Cloud Dataproc is designed to work well with Google Cloud Storage. Using GCS instead of Persistent Disk can save money, and GCS offers advantages such as higher durability and the ability to share data across multiple clusters.

upvoted 1 times

 **emmylou** 3 months, 4 weeks ago

I have seen this question in other places and I believe that you store the older data in Cloud Storage and retain processing data in persistent c D

upvoted 1 times

 **NeoNitin** 4 months, 3 weeks ago

A,, Thank you Exam topic : Passed the exam in august and I can say examtopic is help me lot, topic 1 is enough for the exam, just last week I received welcome kit from google for PDE exam one google cloud cup. if you need all question any help reach out to me neonitin6attherategoogledotcom

upvoted 2 times

 **hxy8** 4 months, 3 weeks ago

Answer: D

Question: A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. which means Persistent is still required.

upvoted 1 times

 **suku2** 4 months, 2 weeks ago

Google Cloud Storage is designed for 11 9's availability. So it is also kind of persistent storage. Also, it is a Google product, hence recommended.

<https://cloud.google.com/storage/docs/availability-durability#key-concepts>

upvoted 1 times

✉  **GHOST1985** 5 months, 2 weeks ago

the question is talking about block storage , GCS is object storage !

upvoted 1 times

✉  **hjava** 6 months ago

Selected Answer: A

GCS is cost-effective and also Google's recommendation!

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: A

Minimize cost then GCS

upvoted 1 times

✉  **Nirca** 1 year ago

Selected Answer: A

A - is the right answer.

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: A

A - dataproc - storage - cost effective is cloud storage

upvoted 1 times

✉  **devaid** 1 year, 3 months ago

Selected Answer: A

Cloud Storage

upvoted 1 times

✉  **sankar_s** 1 year, 8 months ago

Selected Answer: A

Cloud Storage is google recommended one

upvoted 1 times

✉  **anji007** 2 years, 3 months ago

Ans: A

B: Wrong eVM wont solve the problem of larger storage prices.

C: May be, but nothing mentioned in terms of what to tune in the question, also this is like-for-like migration so tuning may not be part of the migration.

D: Again, this is like-for-like so need to define what is hot data and which is cold data, also persistent disk costlier than cloud storage.

upvoted 6 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'A"

upvoted 2 times

✉  **sumanshu** 2 years, 6 months ago

A is correct because Google recommends using Cloud Storage instead of HDFS as it is much more cost effective especially when jobs are running.

B is not correct because this will decrease the compute cost but not the storage cost.

C is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

D is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

upvoted 6 times

✉  **anudeepgupta42** 2 years, 9 months ago

A, Moving the data to GCS will reduce the cost of running the dataproc clusters all the time \

upvoted 1 times

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom

HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

- A. The message body for the sensor event is too large.
- B. Your custom endpoint has an out-of-date SSL certificate.
- C. The Cloud Pub/Sub topic has too many messages published to it.
- D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

Correct Answer: B

Community vote distribution

D (88%)

12%

 **jvg637** Highly Voted 3 years, 10 months ago

The Answer should be D. The custom endpoint is not acknowledging the message, that is the reason for Pub/Sub to send the message again and again. Not B.

upvoted 83 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer D.

<https://cloud.google.com/pubsub/docs/troubleshooting#duplicates>

upvoted 49 times

 **philli1011** Most Recent 4 days, 19 hours ago

D should be the answer. If acknowledgement is not received back to pub/sub , pub/sub may resend messages.

upvoted 1 times

 **rtcp0st** 3 months, 1 week ago

Selected Answer: D

In Google Cloud Pub/Sub, when you use a push subscription, messages are delivered to the specified endpoint (in this case, your custom HTTPS endpoint). The acknowledgement deadline is the time given to the endpoint to acknowledge that it has received and processed the message. If the acknowledgement is not received within the deadline, Pub/Sub may consider the message as unacknowledged and may attempt redelivery, which can lead to duplicate messages.

You should ensure that your custom HTTPS endpoint acknowledges messages within the acknowledgement deadline to prevent duplicate messages from being sent. Additionally, it's essential to handle messages in an idempotent way, so even if duplicates do occur, the action taken by your endpoint doesn't have unintended consequences.

upvoted 2 times

 **imran79** 3 months, 3 weeks ago

The correct answer is:

D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

upvoted 2 times

 **emmylou** 3 months, 4 weeks ago

if there were an out of date certificate then nothing would get through. D

upvoted 1 times

 **FP77** 6 months ago

Selected Answer: D

It should be D

<https://cloud.google.com/pubsub/docs/troubleshooting#duplicates>

upvoted 3 times

✉  **itsmynickname** 6 months, 3 weeks ago

I mistakenly answered by D, but it's indeed B. Here is the explanation: <https://www.youtube.com/watch?v=KObjkda4ZfY>
upvoted 1 times

✉  **marek_skopowski** 6 months, 3 weeks ago

And where exactly in this video it's stated that this is caused by the invalid cert?
upvoted 2 times

✉  **dgteixeira** 7 months, 3 weeks ago

Selected Answer: D

The correct answer is D, because it's how Pub/Sub works.
Documentation here: <https://cloud.google.com/pubsub/docs/troubleshooting#duplicates>
upvoted 3 times

✉  **boca_2022** 9 months ago

Selected Answer: D

D for sure
upvoted 2 times

✉  **juliosb** 10 months, 2 weeks ago

Selected Answer: D

D for sure
upvoted 2 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: D

d is correct
upvoted 1 times

✉  **niketd** 11 months, 1 week ago

Selected Answer: D

No acknowledgment -> Answer B. Moderators please update your answer
upvoted 2 times

✉  **GCPpro** 1 year ago

D is the correct answer.
upvoted 2 times

✉  **AzureDP900** 1 year ago

D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.
upvoted 1 times

✉  **connorscion** 1 year, 1 month ago

D, with a push delivery method if not acknowledge the sensor will keep sending the message.
upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: D

without ack message deliver multiple time.
upvoted 1 times

Question #21

Topic 1

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiently?

- A. Assign global unique identifiers (GUID) to each data entry.
- B. Compute the hash value of each data entry, and compare it with all historical data.
- C. Store each data entry as the primary key in a separate database and apply an index.

D. Maintain a database table to store the hash value and other metadata for each data entry.

Correct Answer: D

Community vote distribution

A (58%)

D (31%)

11%

 **dg63**  3 years, 6 months ago

The best answer is "A".

Answer "D" is not as efficient or error-proof due to two reasons

1. You need to calculate hash at sender as well as at receiver end to do the comparison. Waste of computing power.
2. Even if we discount the computing power, we should note that the system is sending inventory information. Two messages sent at different times can denote same inventory level (and thus have same hash). Adding sender time stamp to hash will defeat the purpose of using hash as now retried messages will have different timestamp and a different hash.

if timestamp is used as message creation timestamp than that can also be used as a UUID.

upvoted 64 times

 **emmylou** 3 months, 4 weeks ago

If you add a unique ID aren't you by definition not getting a duplicate record. Honestly I hate all these answers.

upvoted 2 times

 **retax** 3 years, 3 months ago

If the goal is to ensure at least ONE of each pair of entries is inserted into the db, then how is assigning a GUID to each entry resolving the duplicates? Keep in mind if the 1st entry fails, then hopefully the 2nd (duplicate) is successful.

upvoted 13 times

✉  **ralf_cc** 2 years, 7 months ago

A - In D, same message with different timestamp will have different hash, though the message content is the same.

upvoted 10 times

✉  **omakin** 2 years, 6 months ago

Strong Answer is A - in another question on the gcp sample questions: the correct answer to that particular question was "You are building a new real-time data warehouse for your company and will use BigQuery streaming inserts. There is no guarantee that data only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?"

This means you need a "uniqueid" and timestamps to properly dedupe a data.

upvoted 8 times

✉  **Tanzu** 2 years ago

U need a uniqueid but in this scenario, there is none. So u have to calculate by hashing w/ some of the fields in the dataset.

A is assigning guid in processing side will not solve the issue. Cause u will assign diff. ids...

upvoted 1 times

✉  **cetanx** 1 year ago

Answer - D

Key statement is "Transmitted data includes a payload of several fields and the timestamp of the transmission."

So the timestamp is appended to message while sending, in other words that field is subject to change if message is retransmitted. However, adding a GUID doesn't help much because if message is transmitted twice you will have different GUID for both messages but they will be the same/duplicate data.

You can simply calculate a hash based on not all data but from a select of columns (with the payload of several fields AND definitely by excluding the timestamp). By doing so, you can assure a different hash for each message.

upvoted 2 times

✉  **MaxNRG** 2 years ago

agreed, the key here is "payload of several fields and the timestamp"

upvoted 1 times

✉  **MaxNRG** 2 years ago

"payload of several fields and the timestamp of the transmission"

upvoted 2 times

✉  **BigDataBB** 1 year, 12 months ago

Hi Max, I also think that the hash value would be wrong because the timestamp is part of payload and is not written that the hash value is generated without the ts; but it is also not written if GUID is linked or not with sending. Often this is a point where the answer is vague. Because don't specify if the GUID is related to the data or to the send.

upvoted 1 times

✉  **MarcoDipa** 2 years, 1 month ago

Answer is D. Using Hash values we can remove duplicate values from a database. Hash values will be same for duplicate data and thus can be easily rejected. Obviously you won't check hash for timestamp.

D is better than B because maintaining a different table will reduce cost for hash computation for all historical data

upvoted 5 times

✉  **Mathew106** 6 months, 2 weeks ago

Why can't it be A, where the GUID is a hash value? Why do we need to store the hash with the metadata in a separate database to do the deduplication?

upvoted 1 times

✉  **[Removed]**  3 years, 10 months ago

Answer: D

Description: Using Hash values we can remove duplicate values from a database. Hash values will be same for duplicate data and thus can be easily rejected.

upvoted 24 times

✉  **stefanop** 1 year, 9 months ago

Hash values for same data will be the same, but in this case data contains also the timestamp

upvoted 2 times

✉  **DGames** 1 year, 1 month ago

While calculating Hash value we exclude the timestamp.

upvoted 1 times

TVH_Data_Engineer Most Recent 1 month, 1 week ago

Selected Answer: D

To deduplicate the data most efficiently, especially in a cloud environment where the data is sent periodically and re-transmissions can occur, recommended approach would be:

D. Maintain a database table to store the hash value and other metadata for each data entry.

This approach allows you to quickly check if an incoming data entry is a duplicate by comparing hash values, which is much faster than comparing all fields of a data entry. The metadata, which includes the timestamp and possibly other relevant information, can help resolve any ambiguities that may arise if the hash function ever produces collisions.

upvoted 1 times

JustQ 2 months, 1 week ago

B. Compute the hash value of each data entry, and compare it with all historical data.

Explanation:

Efficiency: Hashing is a fast and efficient operation, and comparing hash values is generally quicker than comparing the entire payload or using other methods.

Space Efficiency: Storing hash values requires less storage space compared to storing entire payloads or using global unique identifiers (GUIDs).

Deduplication: By computing the hash value of each data entry and comparing it with historical data, you can easily identify duplicate transmissions. If the hash value matches an existing one, it indicates that the payload is the same.

upvoted 1 times

steghe 2 months, 3 weeks ago

I thought the answer was A 'cos it's more efficient. But I read the answer with more attention: GUID is given "at each data entry". It's not said that a GUID was given from publisher. If GUID is given in data entry (subscriber), two equal messages can have different GUID.

D is not complete 'cos it's not so precise about hash field that are used.

I'm in doubt on this answer :-(

upvoted 2 times

rocky48 2 months, 3 weeks ago

Selected Answer: A

Answer : A

"D" is not as efficient or error-proof due to two reasons

1. You need to calculate hash at sender as well as at receiver end to do the comparison. Waste of computing power.
2. Even if we discount the computing power, we should note that the system is sending inventory information. Two messages sent at different times can denote same inventory level (and thus have same hash). Adding sender time stamp to hash will defeat the purpose of using hash as now retried messages will have different timestamp and a different hash.

If timestamp is used as message creation timestamp than that can also be used as a UUID.

upvoted 1 times

rtcpost 3 months, 1 week ago

Selected Answer: D

D. Maintain a database table to store the hash value and other metadata for each data entry.

Storing a database table with hash values and metadata is an efficient way to deduplicate data. When new data is transmitted, you can calculate the hash of the payload and check whether it already exists in the database. This approach allows for efficient duplicate detection without the need to compare the new data with all historical data. It's a common and scalable technique used to ensure data consistency and avoid processing the same data multiple times.

Options A (assigning GUIDs to each data entry) and C (storing each data entry as the primary key) can work, but they might be less efficient than using hash values when dealing with a large volume of data. Option B (computing the hash value of each data entry and comparing it with all historical data) can be computationally expensive and slow, especially if there's a significant amount of historical data to compare against.

Storing hash values in a table allows for fast and efficient deduplication.

upvoted 1 times

alihabib 5 months, 4 weeks ago

Why not D ? Generate a Hash for payload entry and maintain the value as metadata. Do the validation check on Dataflow..... A GUID will generate 2 different entries for same payload entry, it will not tackle duplication check

upvoted 2 times

✉️  **Hungry_guy** 5 months, 4 weeks ago

Answer is B - although the time stamp is diff for each transmission - the hash value is computed for the payload, not for the timestamp - which is just an added field for transmission. So, has val remains the same for all transmissions of the same data - which is what we can use for comparison.

So, much more efficient to just directly compare the hash values with the historical data - to check and remove duplicates - instead of again wasting space storing stuff - in option D

upvoted 2 times

✉️  **Mark_86** 6 months, 1 week ago

Selected Answer: D

This question is formulated very badly.

From the way that A is formulated, you would not deduplicate but rather the duplicates would have the same GUID.

Then we have D, which is storing the information (assuming the hash is created without the timestamp). B is doing it right away. D only alludes to the actual deduplication. But it would be more efficient.

upvoted 1 times

✉️  **boca_2022** 9 months ago

Selected Answer: A

A is best choice. D doesn't make sense.

upvoted 2 times

✉️  **FP77** 5 months, 1 week ago

A is incorrect. how can you find duplicates if you assign a unique id to every record? The answer is either B or D. I first selected B, but rechecked through the answers D may be better.

upvoted 2 times

✉️  **Melampos** 9 months, 1 week ago

Selected Answer: D

you cannot deduplicate data adding a random guid, with guid row is distinct than others

upvoted 1 times

✉️  **juliobs** 10 months, 2 weeks ago

Hard question.

It's a *proprietary* system. Who guarantees we can even add a GUID?

But if you can, it's definitely more efficient than calculating hashes (ignoring timestamp).

upvoted 4 times

✉️  **tibuenoc** 11 months ago

Selected Answer: A

As Dg63 wrote.

upvoted 2 times

Question #22

Topic 1

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a

Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks.

What should you do?

- A. Run a local version of Jupiter on the laptop.
- B. Grant the user access to Google Cloud Shell.
- C. Host a visualization tool on a VM on Google Compute Engine.
- D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Correct Answer: B

Community vote distribution

D (98%)

2%

✉  **Rajokkiyam** Highly Voted 3 years, 10 months ago
Answer should be D.
upvoted 47 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago
Answer: D
Description: Datalab provides Jupyter for this kind of work
upvoted 14 times

✉  **GCanteiro** Most Recent 1 week, 5 days ago
Selected Answer: D
D sounds good for me
upvoted 1 times

✉  **TVH_Data_Engineer** 1 month, 2 weeks ago
Selected Answer: A
Hash Value for Deduplication: By computing a hash value for each data entry, you create a unique identifier based on the content of the data. This allows you to efficiently identify duplicates, as entries with identical content will have the same hash value.

Storing Hash Value and Metadata: Maintaining a database table that includes the hash value and other relevant metadata (like the timestamp transmission) allows for quick lookups and comparisons. This way, when new data is received, you can check if an entry with the same hash value already exists.
Assign global unique identifiers (GUID) to each data entry: While GUIDs are unique, they do not inherently identify duplicate content. Two transmissions of the same data would have different GUIDs.
upvoted 1 times

✉  **axantroff** 2 months, 1 week ago
Selected Answer: D
D sounds good for me
upvoted 1 times

✉  **RT_G** 2 months, 3 weeks ago
Selected Answer: D
Agree with D
upvoted 1 times

✉  **rtcpost** 3 months, 1 week ago
Selected Answer: D
D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Google Cloud Datalab is a powerful interactive tool for data exploration, analysis, and machine learning. By deploying it to a VM on Google Compute Engine, you can provide her with a robust and scalable environment where she can work with large datasets, create labeled datasets, and perform data analyses efficiently.

Option A (running a local version of Jupyter on her laptop) might not be sufficient for very large datasets, and her laptop's limited power could be a bottleneck.

Option B (granting access to Google Cloud Shell) is useful for running command-line tools but may not provide the interactive and visual capabilities she needs.

Option C (hosting a visualization tool on a VM on Google Compute Engine) is beneficial for visualization tasks but does not cover the full spectrum of data analysis and machine learning tasks that Google Cloud Datalab offers.
upvoted 3 times

✉  **gudguy1a** 4 months, 3 weeks ago
Selected Answer: D
D - as it is a FULL set up, not a shell that is needed...
upvoted 1 times

✉  **sergiomujica** 5 months ago
Nowadays it should be similar to D, deploy a Vertex workbench
upvoted 1 times

✉  **yash12** 5 months, 2 weeks ago

As per Options , Correct Answer should be D. ie Datalab
However Datalab is no longer used in GCP (Deprecated in Sep2022), It is Vertex AI or Deep Learning VM Images
upvoted 1 times

✉  **HeoMaTo** 5 months, 2 weeks ago

Selected Answer: D

I think.
Answer is D
upvoted 1 times

✉  **Acocado** 6 months, 1 week ago

Datalab is deprecated. This question should appear in the exam.
upvoted 2 times

✉  **Acocado** 6 months, 1 week ago
typo- should NOT appear in the exam
upvoted 6 times

✉  **axantroff** 3 months ago

Good point - <https://cloud.google.com/datalab/deprecation-notice>. Google recommends using Vertex AI Workbench instead
upvoted 1 times

✉  **dgteixeira** 7 months, 3 weeks ago

Selected Answer: D

Should be D, because Cloud shell alone does not provide access to what they need.
Nowadays is Vertex AI, but still, correct answer is D
upvoted 2 times

✉  **Maurilio_Cardoso** 8 months ago

Selected Answer: D

Google Cloud Datalab is now Vertex AI. So, letter D make more sense.
upvoted 3 times

✉  **boca_2022** 9 months ago

Selected Answer: D

B doesn't make sense at all
upvoted 1 times

✉  **abi01a** 9 months, 2 weeks ago

Do you ever get clearly wrong answer like this one ever reversed? I dont understand how D the most voted option at 100% not flagged as the correct answer.
upvoted 2 times

✉  **juliobs** 10 months, 2 weeks ago

Selected Answer: D

D of course
upvoted 1 times

analyze these very large datasets in real time. What should you do?

- A. Send the data to Google Cloud Datastore and then export to BigQuery.
- B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
- C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
- D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Correct Answer: B

Community vote distribution

B (100%)

 **[Removed]**  3 years, 10 months ago

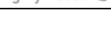
Answer: B

upvoted 30 times

 **[Removed]** 3 years, 10 months ago

<https://cloud.google.com/blog/products/iot-devices/quick-and-easy-way-set-end-end-iot-solution-google-cloud-platform>

upvoted 9 times

 **[Removed]**  3 years, 10 months ago

Answer: B

Description: Pubsub for realtime, Dataflow for pipeline, Bigquery for analytics

upvoted 24 times

 **axantroff**  2 months, 1 week ago

Selected Answer: B

In short, B is less complex and more recommended other than D

upvoted 1 times

 **rtcpost** 3 months, 1 week ago

Selected Answer: B

B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.

Here's why this approach is preferred:

Google Cloud Pub/Sub allows for efficient ingestion and real-time data streaming.

Google Cloud Dataflow can process and transform the streaming data in real-time.

Google BigQuery is a fully managed, highly scalable data warehouse that is well-suited for real-time analysis and querying of large datasets.

upvoted 1 times

 **GCP_PDE_AG** 5 months, 3 weeks ago

Obviously B.

upvoted 1 times

 **Maurilio_Cardoso** 8 months ago

Selected Answer: B

PubSub for queue in real time, Dataflow for processing (pipeline) and Bigquery for analyses.

upvoted 2 times

 **bha11111** 10 months, 3 weeks ago

Selected Answer: B

B is correct

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

GCP recommend best practice for streaming data pipeline as option B - pub/sub, dataflow & Bigquery
upvoted 1 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: B

B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
upvoted 1 times

✉  **gitaexams** 1 year, 2 months ago

B aris tqve yleeb
upvoted 1 times

✉  **deavid** 1 year, 3 months ago

Selected Answer: B

B of course
upvoted 1 times

✉  **Dip1994** 1 year, 5 months ago

B is the correct answer
upvoted 1 times

✉  **noob_master** 1 year, 7 months ago

Selected Answer: B

Answer: B

Deafult ETL streaming process: Pub/Sub + Dataflow + Bigquery.
upvoted 1 times

✉  **nexus1_** 1 year, 7 months ago

Definitely B
upvoted 1 times

✉  **vw13** 1 year, 9 months ago

Selected Answer: B

B is the only option for real time process & analysis
upvoted 1 times

✉  **devric** 1 year, 10 months ago

The most appropriate is B but BQ can't solve Analyzing data in RT.
upvoted 1 times

✉  **samdhimal** 2 years ago

correct answer is Cloud Pub/Sub ---> Dataflow ---> Bigquery

Collected messages containing temperature values will be published to a topic on Cloud Pub/Sub, Messages will be read in streaming mode Cloud Dataflow, a simplified stream and batch data processing solution, Cloud Datastore will save data to be displayed directly into the UI of App Engine application, while BigQuery will act as a data warehouse that will enable the execution of more in depth analysis.

Reference:

<https://cloud.google.com/blog/products/iot-devices/quick-and-easy-way-set-end-end-iot-solution-google-cloud-platform>
upvoted 2 times

Question #24

Topic 1

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.

- B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.
- C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
- D. Add two columns to the table CLICK_STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.
- E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Correct Answer: D

Community vote distribution

E (69%)

C (28%)

3%

✉  **jvg637**  3 years, 10 months ago

"E" looks better. For D, the database will be double in size (which increases the storage price) and the user has to spend some more days reloading all the data.

upvoted 31 times

✉  **jkhong** 1 year, 1 month ago

Also D doesn't make sense since we're filtering IS_NEW to true to only consider future data, which disregards our previously loaded data

upvoted 5 times

✉  **assU2** 1 year, 2 months ago

"You want to minimize the migration effort without making future queries computationally expensive." Nothing about storage price.

upvoted 4 times

✉  **[Removed]**  3 years, 10 months ago

E - more simple and reasonable. Also recommended if not concerned about cost but simplicity.

https://cloud.google.com/bigquery/docs/manually-changing-schemas#changing_a_columns_data_type

upvoted 22 times

✉  **Tanzu** 2 years ago

Due to the hard limitations of bq, Not E is the simple answer by the way!

upvoted 1 times

✉  **TVH_Data_Engineer**  1 month, 2 weeks ago

Selected Answer: C

A view in Google BigQuery is a virtual table defined by a SQL query. By creating a view that casts the DT column as a TIMESTAMP, you can transform the data format without altering the underlying data in the CLICK_STREAM table. This means you don't have to reload any data, thereby minimizing migration effort.

upvoted 1 times

✉  **axantroff** 2 months, 1 week ago

Selected Answer: E

Good point about the logical views and the desire to reduce costs. I would vote for E

upvoted 1 times

✉  **mk_choudhary** 3 months, 1 week ago

The best way to minimize the migration effort without making future queries computationally expensive is to create a view and reference it instead of the table. This is because views are materialized when they are queried, so they do not incur any additional overhead. So the answer is (C).

upvoted 1 times

✉  **brokeasspanda** 2 months, 3 weeks ago

C doesn't say materialized view, there's a difference with a regular view so it'll be slower and more expensive on every call to that view.

upvoted 1 times

✉️  **rtcpost** 3 months, 1 week ago

Selected Answer: E

Option "E"

It avoids the need to delete and recreate the entire CLICK_STREAM table, which is time-consuming and requires reloading all data.

It allows you to use a simple query to cast the existing DT column as TIMESTAMP values and store the results in a new table, NEW_CLICK_STREAM.

You can gradually migrate to the new data format, and your future queries will be able to utilize the TIMESTAMP data type for more efficient processing.

upvoted 2 times

✉️  **sergiomujica** 5 months ago

Option D duplicates, not a good solution

upvoted 1 times

✉️  **NeoNitin** 6 months ago

E. E. You can use a special command to change the time on the old cards to the better type "TIMESTAMP" and create a new box called "NEW_CLICK_STREAM." From now on, you'll look at the new box whenever you want to know the time. It's like having a new and better box keep things tidy and organized.

So, the best way to change the time on the little cards to the better type "TIMESTAMP" is option E. It's like using magic to create a new box a making sure everything is still easy to find and work with. It's a clever way to keep track of time and make your website even better!

upvoted 1 times

✉️  **tal_** 7 months, 2 weeks ago

Selected Answer: E

they asked to "change its data type"

upvoted 1 times

✉️  **mimosoundz** 8 months, 3 weeks ago

Selected Answer: E

it's E. computationally less expensive than running a view every time

upvoted 1 times

✉️  **boca_2022** 9 months ago

Selected Answer: E

E is best option

upvoted 1 times

✉️  **techtitan** 11 months, 2 weeks ago

Selected Answer: E

its C vs E. E is better because C will try to do a cast operation everytime query is run making it computationally expensive.

upvoted 5 times

✉  **Jackalski** 1 year, 1 month ago

Selected Answer: C

option E - includes already effort for C (building query with cast) however adding steps on rebuilding the table with magic word "NEW" (how often you have NEW tables on your side - have you wonder when NEW becomes OLD? not to mentioned that it kept previous table as well) A - not consider to "reload" which took few days .
D - not consider to "reload" which took few days ... and have duplicates
B - almost good one but why kept having two columns with same data

question was about minimal effort on migration - no ideal answer exists on this list (no cleanup etc)
so I vote on C

upvoted 4 times

✉  **Simhamed2015** 3 months, 2 weeks ago

BigQuery's views are logical views, not materialized views. Because views are not materialized, the query that defines the view is run each time the view is queried. Queries are billed according to the total amount of data in all table fields referenced directly or indirectly by the top-level query. For more information, see query pricing.

B would be very expensive, since we will be charged each time the view is queried.

upvoted 1 times

✉  **Kiroo** 8 months, 3 weeks ago

That was my thought initially but note "making future queries not computationally expensive" so creating a view you will need to always provide the value because of that I would go with E

upvoted 1 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: E

We are dealing here with "comma-separated values (CSV)" not "application" that sensitive to SQL fix. E is the best way to implement stand-alone use case like this. For applications that are sensitive for DDL & DML - another method might be useful

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: D

D- Make more sense because then E because we are creating new table again need to overhead to update everywhere table name in application

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: E

If I do this in my work, E is my answer 100% Otherwise is to complicate.

upvoted 1 times

Question #25

Topic 1

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

- A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
- B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
- C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
- D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Correct Answer: B

Community vote distribution

D (100%)

✉  **ivq637** Highly Voted 3 years, 10 months ago

I would choose D.

A and B are wrong since don't notify anything to the monitoring tool.

C has no filter on what will be notified. We want only some tables.

upvoted 47 times

 **MaxNRG** Highly Voted 2 years, 2 months ago

D as the key requirement is to have notification on a particular table. It can be achieved using advanced log filter to filter only the table logs and create a project sink to Cloud Pub/Sub for notification.

Refer GCP documentation - Advanced Logs Filters: <https://cloud.google.com/logging/docs/view/advanced-queries>

A is wrong as advanced filter will help in filtering. However, there is no notification sends.

B is wrong as it would send all the logs and BigQuery does not provide notifications.

C is wrong as it would send all the logs.

upvoted 14 times

 **axantroff** Most Recent 2 months, 1 week ago

Selected Answer: D

Good point by MaxNRG about reducing the number of logs sending to Pub/Sub

upvoted 2 times

 **ruben82** 3 months ago

Theoretically Pub/Sub could filter logs to forward the right ones to the correct topic. <https://cloud.google.com/pubsub/docs/subscription-message-filter>

So C could be accepted, but it's better if filtering is performed earlier, so in this case D is more performing

upvoted 1 times

 **rtcpost** 3 months, 1 week ago

Selected Answer: D

D. Using the Stackdriver API, create a project sink with an advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

This approach allows you to set up a custom log sink with an advanced filter that targets the specific table and then export the log entries to Google Cloud Pub/Sub. Your monitoring tool can subscribe to the Pub/Sub topic, providing you with instant notifications when relevant events occur without being inundated with notifications from other tables.

Options A and B do not offer the same level of customization and specificity in targeting notifications for a particular table.

Option C is almost correct but doesn't mention the use of an advanced log filter in the sink configuration, which is typically needed to filter the logs to a specific table effectively. Using the Stackdriver API for more advanced configuration is often necessary for fine-grained control over filtering.

upvoted 1 times

 **suku2** 4 months, 2 weeks ago

Selected Answer: D

D makes sense.

upvoted 1 times

 **NeoNitin** 4 months, 3 weeks ago

D, : Thank you Exam topic : Passed the exam in August and I can say examtopic is help me a lot, topic 1 is enough for the exam, just last week received welcome kit from Google for PDE exam one Google Cloud Cup. If you need any question, any help, reach out to me neonitin6@therategoogledotcom

upvoted 1 times

 **cseashok** 4 months ago

Hi, am planning to write GCP PDE certification in upcoming month. Can you pls help me with the topics and preparation you did? pls provide your mail id.

upvoted 1 times

 **GCP_PDE_AG** 5 months, 3 weeks ago

D should be the answer

upvoted 1 times

 **Mathew106** 6 months, 1 week ago

Selected Answer: D

A and B mention nothing about notifications and C would push all data. It's D.

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: D

D makes sense

upvoted 1 times

✉  **Jackalski** 1 year, 1 month ago

Selected Answer: D

"advanced log filter" is the key word here, all other options push all data ...

upvoted 2 times

✉  **Jasar** 1 year, 2 months ago

Selected Answer: D

D is the best choice

upvoted 1 times

✉  **alecuba16** 1 year, 9 months ago

Selected Answer: D

Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

upvoted 4 times

✉  **devric** 1 year, 10 months ago

Selected Answer: D

D. Option B doesn't make sense

upvoted 2 times

✉  **samdhimal** 2 years ago

correct answer -> Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Option C is also most likely right answer but it doesn't have the filter. We don't want all the tables. We only want one. So the correct answer is

Logging sink - Using a Logging sink, you can direct specific log entries to your business logic. In this example, you can use Cloud Audit logs or Compute Engine which use the resource type gce_firewall_rule to filter for the logs of interest. You can also add an event type GCE_OPERATION_DONE to the filter to capture only the completed log events. Here is the Logging filter used to identify the logs. You can try the query in the Logs Viewer.

Pub/Sub topic – In Pub/Sub, you can create a topic to which to direct the log sink and use the Pub/Sub message to trigger a cloud function.

Reference: <https://cloud.google.com/blog/products/management-tools/automate-your-response-to-a-cloud-logging-event>

upvoted 3 times

✉  **santoshindia** 2 years ago

Selected Answer: D

explained by MaxNRG

upvoted 3 times

✉  **medeis_jar** 2 years ago

Selected Answer: D

as explained by MaxNRG

upvoted 3 times

Question #26

Topic 1

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

- A. Grant the consultant the Viewer role on the project.
- B. Grant the consultant the Cloud Dataflow Developer role on the project.
- C. Create a service account and allow the consultant to log on with it.
- D. Create an anonymized sample of the data for the consultant to work with in a different project.

Correct Answer: C

Community vote distribution

D (78%)

B (22%)

 **jvg637** Highly Voted 3 years, 10 months ago

The Answer should be B. The Dataflow developer role will not provide access to the underlying data.

upvoted 75 times

 **VincentMenzel** 6 months ago

I'm not sure how you expect the consultant to implement a pipeline without having access to any data that is being processed. Having test data is a prerequisite.

upvoted 2 times

 **ThorstenStaerk** 9 months, 3 weeks ago

and now? For seeing test data, (D) would be right. And the system tells me (C) is the right answer. What shall I click in the exam?

upvoted 4 times

 **cleroy** 3 years, 10 months ago

Remember he's an external consultant. You need to create a service account for him, you can't grant before that... I think C is correct in this case.

upvoted 5 times

 **Rajuuu** 3 years, 6 months ago

Service account is between applications and non human entry.

upvoted 34 times

 **Tanzu** 2 years ago

u can enable a service account as user so that externals can use to login.

but the problem is service account is about login. not the minimum resources to do the dataflow related staffs. so C is not enough !.

so the answer should be B.

if the question was about "doing the 1st thing", then yeah may be creating a service account could be the 1st thing.

upvoted 3 times

 **Tanzu** 2 years ago

u can enable a service account as user so that externals can use to login

upvoted 1 times

 **willymac2** 1 year, 7 months ago

The answer should be D.

You do not need any DataFlow permission to implement a pipeline.

If needed, you can test using the DirectRunner which runs locally:

https://cloud.google.com/dataflow/docs/concepts/access-control#example_role_assignment

upvoted 2 times

 **willymac2** 1 year, 7 months ago

Sorry I did a wrong copy/paste on the link, I wanted to send:

https://cloud.google.com/dataflow/docs/concepts/security-and-permissions#security_and_permissions_for_local_pipelines

<https://cloud.google.com/dataflow/docs/guides/setting-pipeline-options#LocalExecution>

upvoted 2 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Answer: B

Description: Provides the permissions necessary to execute and manipulate Dataflow jobs.

upvoted 17 times

👤 MaxNRG Most Recent 1 month, 2 weeks ago

Selected Answer: B

B as the Dataflow developer role would help provide the third-party consultant access to create and work on the Dataflow pipeline. However, does not provide access to view the data, thus maintaining user's privacy.

Refer GCP documentation - Dataflow roles:

<https://cloud.google.com/dataflow/docs/concepts/access-control#roles>

Option A is wrong as it would not allow the consultant to work on the pipeline.

Option C is wrong as the consultant cannot use the service account to login.

Option D is wrong as it does not enable collaboration.

upvoted 2 times

👤 Jconnor 1 month, 3 weeks ago

C and A will not maintain user's privacy so out. B without data will be enough. D will give a good sample data, maintain privacy and the consultant will help creating the dataflow pipe for the project as requested. so D.

upvoted 1 times

👤 axantroff 2 months, 1 week ago

Selected Answer: D

I follow the corresponding logic choosing between B and D:

Yes, with the Dataflow Developer role it is possible to execute and manipulate Dataflow jobs, but do we need to execute it? Based on my understanding we only need to ask for help to write it down. Is it possible without having access to test the data? I don't think so. At the same time, we need to perform an anonymization on it. So the answer D is more appropriate for me

upvoted 1 times

👤 rocky48 2 months, 3 weeks ago

Selected Answer: D

By creating an anonymized sample of the data, you can provide the consultant with a realistic dataset that doesn't contain sensitive or private information. This way, the consultant can work on the project without direct access to sensitive data, reducing privacy risks.

Options A and B involve granting the consultant access to the project, which may expose sensitive data, even if they have limited permissions.

Option C involves creating a service account, but it doesn't address the need to anonymize the data or provide a separate, safe environment for the consultant to work with.

Option D provides a controlled environment that allows the consultant to work effectively while maintaining data privacy.

upvoted 1 times

👤 rtcpost 3 months, 1 week ago

Selected Answer: D

D. Create an anonymized sample of the data for the consultant to work within a different project.

By creating an anonymized sample of the data, you can provide the consultant with a realistic dataset that doesn't contain sensitive or private information. This way, the consultant can work on the project without direct access to sensitive data, reducing privacy risks.

Options A and B involve granting the consultant access to the project, which may expose sensitive data, even if they have limited permissions.

Option C involves creating a service account, but it doesn't address the need to anonymize the data or provide a separate, safe environment for the consultant to work with.

Option D provides a controlled environment that allows the consultant to work effectively while maintaining data privacy.

upvoted 1 times

👤 imran79 3 months, 3 weeks ago

D. Creating an anonymized sample of the data for the consultant to work with in a different project is the safest option. This way, the consultant can develop and test the transformation logic without accessing the real, sensitive data.

upvoted 1 times

👤 ruben82 3 months ago

The question says "with coding a complex transformation", so I don't think that a sample of data is enough. I think that the most suitable way is C, 'cos with a service account you can handle access fine-grained

upvoted 1 times

navioshi 4 months, 1 week ago

I think C would be correct, as the question says external consultants want to do some work and how we can maintain the 'external consultant user privacy. Question didn't mention about the company user data or customer information.

upvoted 2 times

NeoNitin 4 months, 3 weeks ago

D : Thank you Exam topic : Passed the exam in august and I can say examtopic is help me lot, topic 1 is enough for the exam, just last week I received welcome kit from google for PDE exam one google cloud cup. if you need all question any help reach out to me neonitin6attherategoogledotcom

upvoted 2 times

hxy8 4 months, 3 weeks ago

Answer: C

upvoted 1 times

madhu15 5 months, 2 weeks ago

Dataflow Developer
(roles/dataflow.developer)

Provides the permissions necessary to execute and manipulate Dataflow jobs.

upvoted 1 times

marek_skopowski 6 months, 3 weeks ago

Unfortunately it's the Service Account answer: "The developer who creates and examines jobs needs the roles/iam.serviceAccountUser role."
<https://cloud.google.com/dataflow/docs/concepts/access-control#example>

upvoted 2 times

itsmynickname 6 months, 3 weeks ago

None of the answer convinced me. there is no information about the service account (the granted permissions). Accessing to a project doesn't mean accessing to its data, so it is not necessary to oblige the external consultant to work in another project, still giving sample (anonymized) data to the developer is important. Giving Dataflow permissions is not necessary as it is possible to run the apache beam job locally using DirectRunner (if we keep performance tuning problematic aside).. Having viewer role on the project, hell no!

upvoted 2 times

baht 7 months, 3 weeks ago

Selected Answer: D

The answer is D. With only an anonymized sample of the data, the consultant can work on the project.

upvoted 1 times

boca_2022 9 months ago

Selected Answer: D

It should be D

upvoted 1 times

Oleksandr0501 9 months, 1 week ago

D. Create an anonymized sample of the data for the consultant to work with in a different project.

As the project involves sensitive user data, it is important to protect users' privacy. Granting the consultant the Viewer or Cloud Dataflow Developer role would give them too much access to the data. Creating a service account and allowing the consultant to log on with it would provide access to the data as well.

Creating an anonymized sample of the data for the consultant to work with in a different project would allow them to complete their work without exposing sensitive user data. This way, the consultant can still work on the complex transformation in a controlled environment without putting user data at risk.

upvoted 2 times

Question #27

Topic 1

You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.

- B. Combine highly co-dependent features into one representative feature.
- C. Instead of feeding in each feature individually, average their values in batches of 3.
- D. Remove the features that have null values for more than 50% of the training records.

Correct Answer: B

Community vote distribution

B (88%)

13%

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: B

Description: Best Choice out of given options.

upvoted 32 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Should be B

upvoted 17 times

 **axantroff** Most Recent 2 months, 1 week ago

Selected Answer: B

I am not into ML, to be honest, so I will rely on community opinion and choose B

upvoted 1 times

 **rtcp0st** 3 months, 1 week ago

Selected Answer: B

B. Combine highly co-dependent features into one representative feature.

Combining highly correlated features into a single representative feature can reduce the dimensionality of your dataset, making the training process faster while preserving relevant information. This approach often helps eliminate redundancy in the input data.

Option A (eliminating features that are highly correlated to the output labels) can be counterproductive, as you want to maintain features that are informative for your prediction task. Removing features that are correlated with the output may reduce model accuracy.

Option C (averaging feature values in batches of 3) is not a common technique for reducing dimensionality, and it could lead to loss of important information.

Option D (removing features with null values for more than 50% of training records) can help reduce the dimensionality and may be useful if you have a large number of features with missing data, but it may not necessarily address co-dependency among features.

upvoted 1 times

 **suku2** 4 months, 2 weeks ago

Selected Answer: B

B. Combine highly co-dependent features into one representative feature.

This is the best choice.

upvoted 1 times

 **WillemHendr** 8 months, 3 weeks ago

"D" is wrong, and very dangerous. For instance, it might represent modern measurements only installed in <50% of weather stations, but very very precise and valuable.

Nulls are not a problem for models, out-of-the-box or with transformations models can handle nulls just fine.

upvoted 1 times

jin0 11 months, 1 week ago

Selected Answer: D

wrong question. there are two answers B, D

upvoted 1 times

jin0 11 months, 1 week ago

B. Combine highly co-dependent features into one representative feature.

-> Explainable feature should be dependent from each other feature. especially not deep learning. so, in this case normally eliminated or combined

D. Remove the features that have null values for more than 50% of the training records.

-> it's too large null data in the feature. normally the feature should be removed because it's too hard to fill up replacing data

upvoted 2 times

AzureDP900 1 year ago

Answer is Combine highly co-dependent features into one representative feature.

A: correlated to output means that feature can contribute a lot to the model. so not a good idea.

C: you need to run with almost same number, but you will iterate twice, once for averaging and second time to feed the averaged value.

D: removing features even if it 50% nulls is not good idea, unless you prove that it is not at all correlated to output. But this is nowhere so can remove.

upvoted 1 times

jin0 11 months, 1 week ago

But, if there are null data more than 50% then, it should be eliminated because there are two ways to train the model. first, remove record containing having null but in this case there are too many records should be removed and second, replace null to other data but in this case cause of it's too large data having null then it's literally hard to replace. so normally the feature having too many null data should be removed. So, there are two answers in this question B, D I think

upvoted 1 times

Thasni 1 year, 2 months ago

I have a doubt, instead of combining highly correlated features why can't we remove correlated features which may give much more simplified dataset?

upvoted 2 times

noob_master 1 year, 7 months ago

Selected Answer: B

Answer: B

Data that is co-dependent is highly correlated is some kind of redundant information in some cases. If the features x_1 , x_2 and x_3 are $x_2 = x_1 + 1$ and $x_3 = 2 \cdot x_1$, for example, x_2 and x_3 are redundant because can be explained with x_1 feature, so can be excluded from the model. Other option to group these features. There is a lot of ways to resolve, but the main idea is to use data engineering in co-dependent features to reduce the number of features in the model.

upvoted 2 times

Ishiske 1 year, 7 months ago

Selected Answer: B

This method is called Data Engineering, that you combine two or more values to get a custom info, this will avoid that the model read an extra column on the training and probably increase its accuracy.

upvoted 1 times

Yad_datatonic 1 year, 8 months ago

Answer: B

upvoted 1 times

alecuba16 1 year, 9 months ago

Selected Answer: B

Co-dependent -> correlated -> correlated info = already present info in other variable.

upvoted 2 times

👤 **pamepadero** 1 year, 11 months ago

Trying to find a reason why it is B and not D, found this and it seems the answer is D.

<https://cloud.google.com/architecture/data-preprocessing-for-ml-with-tf-transform-pt1>

Feature selection. Selecting a subset of the input features for training the model, and ignoring the irrelevant or redundant ones, using filter or wrapper methods. This can also involve simply dropping features if the features are missing a large number of values.

upvoted 6 times

👤 **Dayashankar_H_A** 1 year, 8 months ago

Yes. But nearly 50% of the non-null data still seems to be a lot to ignore.

upvoted 1 times

👤 **exnaniantwort** 2 years ago

Selected Answer: B

B

null values can have many meanings and need different approach to handle, otherwise it causes inaccurate model, so not D

upvoted 4 times

👤 **ZIMARAKI** 2 years ago

Selected Answer: B

For me the best option is B

upvoted 1 times

👤 **aet2123** 2 years, 1 month ago

Selected Answer: B

B is the correct option

upvoted 1 times

Question #28

Topic 1

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.

The data scientists have written the following code to read the data for a new key features in the logs.

```
BigQueryIO.Read
    .named("ReadLogData")
    .from("clouddataflow-readonly:samples.log_data")
```

You want to improve the performance of this data read. What should you do?

- A. Specify the TableReference object in the code.
- B. Use .fromQuery operation to read specific fields from the table.
- C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
- D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table.

Correct Answer: D

Community vote distribution

B (86%)

14%

👤 **arthur2385**  1 year, 4 months ago

B BigQueryIO.read.fromQuery() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs)

upvoted 11 times

👤 **maxdataengineer** Highly Voted 1 year, 3 months ago

Since we want to be able to analyze data from a new ML feature (column) we only need to check values from that column. By doing a `fromQuery(SELECT featureColumn FROM table)` we are optimizing costs and performance since we are not checking all columns.

[https://cloud.google.com/bigquery/docs/best-practices-costs#avoid_select_](https://cloud.google.com/bigquery/docs/best-practices-costs#avoid_select_upvoted 6 times)

👤 **maxdataengineer** 1 year, 3 months ago

The answer is B

upvoted 2 times

👤 **cetanx** 8 months, 2 weeks ago

According to Chat GPT, it is also B

In general, if your "primary goal is to reduce the amount of data read and transferred", and the downstream processing mainly focuses a subset of fields, using `.fromQuery` to select specific fields would be a good choice.

On the other hand, if you need to simplify downstream processing and optimize resource utilization, transforming data into `TableRow` objects might be more suitable.

upvoted 1 times

👤 **MaxNRG** Most Recent 1 month, 2 weeks ago

Selected Answer: B

B as `BigQueryIO.read.from()` directly reads the whole table from BigQuery.

This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from.

This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from BigQuery).

`BigQueryIO.read.fromQuery()` executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs) <https://stackoverflow.com/questions/54413681/bigqueryio-read-vs-fromquery>

upvoted 1 times

👤 **axantroff** 2 months, 1 week ago

Selected Answer: B

B works for me

upvoted 1 times

👤 **pue_dev_anon** 2 months, 1 week ago

Selected Answer: B

We are trying to optimize reading each row is not optimal, we want columns

upvoted 1 times

👤 **rtcp0st** 3 months, 1 week ago

Selected Answer: B

B. Use the `.fromQuery` operation to read specific fields from the table.

Using the `.fromQuery` operation allows you to specify the exact fields you need to read from the table, which can significantly improve performance by reducing the amount of data that needs to be processed. This is particularly important when dealing with large and growing datasets.

Option A (specifying the `TableReference` object) provides information about the table but doesn't inherently improve the performance of reading specific fields.

Option C (using Google BigQuery `TableSchema` and `TableFieldSchema` classes) is related to specifying the schema of the data but doesn't directly address improving the performance of reading specific fields.

Option D (calling a transform that returns `TableRow` objects) is more about how the data is processed after it's read, not how it's initially read from BigQuery.

upvoted 2 times

👤 **emmylou** 4 months, 1 week ago

When I have a different answer than the "Correct Answer", I run it through AI and it keeps saying ExamTopics is wrong. Is there any way to know if I am going to pass or fail this exam?

upvoted 2 times

👤 **axantroff** 2 months, 1 week ago

AI is just a LLM model, not a silver bullet at all

upvoted 1 times

✉  **suku2** 4 months, 2 weeks ago

Selected Answer: B

Since the requirement is to read the data for a *new* key features in the logs, it makes sense to select limited columns, which are required rather than using .from() method which exports the entire BigQuery table.

B makes sense here.

upvoted 1 times

✉  **gudguy1a** 4 months, 3 weeks ago

Selected Answer: B

SHOULD be B.

Not quite sure how D is the correct answer (Red herring....?) when you want to improve the query, which is .fromQuery and NOT transform an PCollection....

upvoted 1 times

✉  **odiez3** 6 months ago

Answer Is D, imagine that you dont have permission on BQ AND you cant see the table info or anything else about the table you only are working with dataflow the only way is transform the data using apache beam

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: B

I have seen people explain why B is not right because it doesn't optimize performance but only cost, which is not true, or because fromQuery still not performant.

I think it's B because no other option is more performant, even if you claim it's not good.

As for option D, the transform given by the description is already a transform that provides as output a PCollection of TableRow objects. So how would that be any different?

<https://beam.apache.org/releases/javadoc/2.1.0/org/apache/beam/sdk/io/gcp/bigquery/BigQueryIO.html>

upvoted 1 times

✉  **theseawillclaim** 6 months, 2 weeks ago

Why should it be D?

"fromQuery()" allows us to read only the columns we want, I see no point in using a Transform for each row of a "SELECT *", which, moreover is a bad BQ Practice.

upvoted 1 times

✉  **avg_uchiha** 8 months, 1 week ago

Correct answer should be B. BigQuery is a columnar storage, so reducing the number of fields being selected should improve performance.

upvoted 1 times

✉  **jkh_goh** 1 year ago

Selected Answer: B

Does BigQuery have a pCollections? I thought it's unique to Apache Beam i.e. Cloud Dataflow

upvoted 1 times

✉  **kelvintoys93** 1 year, 2 months ago

Guys, how is B the answer? Like all the justifications given here, BigQueryIO.read.fromQuery() is time consuming and the question asked for a better performance solution.

upvoted 4 times

✉  **Lestrang** 1 year ago

That part is the docs trying to explain the side effects of using it, however, the part that is important to us is the fact that it reads from a query. "Read" reads the whole table. If we specify a query we can say select col1 only, which makes it all more efficient.

upvoted 2 times

✉  **gcm7** 1 year, 3 months ago

Selected Answer: B

reading only relevant cols

upvoted 6 times

✉  **deavid** 1 year, 3 months ago

Selected Answer: D

Answer is D, apparently.

upvoted 1 times

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

- A. Use a row key of the form <timestamp>.
- B. Use a row key of the form <sensorid>.
- C. Use a row key of the form <timestamp>#<sensorid>.
- D. Use a row key of the form >#<sensorid>#<timestamp>.

Correct Answer: D

Community vote distribution

D (100%)

✉️  [Removed]  3 years, 10 months ago

Description: Best practices of bigtable states that rowkey should not be only timestamp or have timestamp at starting. It's better to have sensor and timestamp as rowkey

upvoted 33 times

✉️  [Removed]  3 years, 10 months ago

Answer D

upvoted 19 times

✉️  axantroff  2 months, 1 week ago

Selected Answer: D

Looks like D is the best option

Reference: <https://cloud.google.com/bigtable/docs/schema-design#time-based>

upvoted 1 times

✉️  rtcpost 3 months, 1 week ago

Selected Answer: D

D. Use a row key of the form <sensorid>#<timestamp>.

By using the sensor ID as the prefix in the row key, you can achieve better distribution of data across Bigtable tablets. This can help balance the workload and prevent hotspots in the table. Additionally, placing the timestamp after the sensor ID allows you to perform range scans for a specific sensor and retrieve data efficiently within a time frame.

Option C (using a row key of the form <timestamp>#<sensorid>) can work for some use cases but may not be as efficient for range scans when you want to retrieve data for a specific sensor within a time range.

Option A (using a row key of the form <timestamp>) may lead to hotspots and inefficient range scans because it doesn't consider sensor IDs.

Option B (using a row key of the form <sensorid>) is not optimal because it doesn't allow for efficient time-based filtering and could lead to uneven data distribution in Bigtable.

upvoted 2 times

✉  **AzureDP900** 1 year ago

D is right

Best practices of bigtable states that rowkey should not be only timestamp or have timestamp at starting. It's better to have sensorid and timestamp as rowkey.

Reference:

<https://cloud.google.com/bigtable/docs/schema-design>

upvoted 1 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: D

#<sensorid>#<timestamp> -----> low cardinality # high cardinality

This is current Bigtable Best Practice (to avoid Hotspots on the inserts)

upvoted 5 times

✉  **maxdataengineer** 1 year, 3 months ago

Selected Answer: D

Discard:

A -> timestamp unique id could not be unique in the case that sensors transmit data at the same time.

B -> sensorid repeated id for messages coming from the same sensor

C -> a bad performance choice

D -> BEST CHOICE. Each time BigTable looks for data in a table it does a scan and sort operations. By starting each unique id by sensorid it make it easier to group and sort data since it has the lowest cardinality

<https://cloud.google.com/bigtable/docs/schema-design#general-concepts>

upvoted 1 times

✉  **John_Pongthorn** 1 year, 4 months ago

as I look at <https://cloud.google.com/bigtable/docs/schema-design#row-keys>

asia#india#bangalore

asia#india#mumbai

they didn't have # ahead of this first value.

asia#india#bangalore OR #asia#india#bangalore

Are both valid?

upvoted 2 times

✉  **crisimenjivar** 1 year, 5 months ago

ANSWER: D

upvoted 1 times

✉  **som_420** 1 year, 7 months ago

Selected Answer: D

Answer is D

upvoted 1 times

✉  **samdhimal** 2 years ago

A. Use a row key of the form <timestamp>.

---> Incorrect, because google says don't use a timestamp by itself or at the beginning of a row key.

B. Use a row key of the form <sensorid>.

---> Incorrect, because google says include a timestamp as part of your row key.

C. Use a row key of the form <timestamp>#<sensorid>.

---> Incorrect, because google says don't use a timestamp by itself or at the beginning of a row key.

D. Use a row key of the form >#<sensorid>#<timestamp>.

---> Correct answer, because of option A,B,C reasons.

- Timestamp isn't by itself, neither at the beginning.

- Timestamp is included.

Reference: <https://cloud.google.com/bigtable/docs/schema-design#row-keys>

upvoted 9 times

✉  **anji007** 2 years, 3 months ago

Ans: D

upvoted 2 times

⊕  **sumanshu** 2 years, 7 months ago

Vote for 'D' - Store multiple delimited values in each row key. (But avoid starting with Timestamp)

"Row keys to avoid"

<https://cloud.google.com/bigtable/docs/schema-design>

upvoted 9 times

⊕  **sumanshu** 2 years, 6 months ago

A is not correct because this will cause most writes to be pushed to a single node (known as hotspotting)

B is not correct because this will not allow for multiple readings from the same sensor as new readings will overwrite old ones.

C is not correct because this will cause most writes to be pushed to a single node (known as hotspotting)

D is correct because it will allow for retrieval of data based on both sensor id and timestamp but without causing hotspotting.

upvoted 7 times

⊕  **naga** 2 years, 11 months ago

Correct D

upvoted 2 times

⊕  **NamitSehgal** 3 years, 1 month ago

Should be D

Reverse of timestamp even better but no options for that.

Also changing sensor ID if they are in sequential to hash or changing data to bits even better.

Idea is not to use timestamp or sequential ID as first key.

upvoted 3 times

⊕  **Tanzu** 2 years ago

reverse TS or hashing is not always first choice or better. never.

upvoted 1 times

⊕  **Radhika7983** 3 years, 2 months ago

The correct answer is D.

Refer to the link <https://cloud.google.com/bigtable/docs/schema-design> for Big table schema design.

C is not the right answer because

Timestamps

If you often need to retrieve data based on the time when it was recorded, it's a good idea to include a timestamp as part of your row key. Use the timestamp by itself as the row key is not recommended, as most writes would be pushed onto a single node. For the same reason, avoid placing a timestamp at the start of the row key.

For example, your application might need to record performance-related data, such as CPU and memory usage, once per second for a large number of machines. Your row key for this data could combine an identifier for the machine with a timestamp for the data (for example, machine_4223421#1425330757685).

upvoted 3 times

⊕  **arghya13** 3 years, 2 months ago

Question #30

Topic 1

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations.

The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

- A. Add a node to the MySQL cluster and build an OLAP cube there.
- B. Use an ETL tool to load the data from MySQL into Google BigQuery.
- C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
- D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Correct Answer: C

Community vote distribution

B (58%)

D (41%)

✉  **HectorLeon2099**  3 years, 1 month ago

It is a GOOGLE exam. The answer won't be on-premise or OLAP cubes even if it is the easiest. The answer is B
upvoted 105 times

✉  **Tanzu** 2 years ago

choose dataproc over hadoop cluster
chose bigquery over all..

there is no special customer requirement that gonna drive us to hadoop or dataproc.

upvoted 8 times

✉  **cetanx** 1 year ago

Answer - B

mysql dump: This utility creates a logical backup and a flat file containing the SQL statements that can be run again to bring back the database to the state when this file was created. So this file can easily be processed by an ETL tool and loaded into BQ.

upvoted 2 times

✉  **ThorstenStaerk** 9 months, 3 weeks ago

So, you are saying that B takes the backup data from the nightly dumps? How can you be sure?

upvoted 2 times

✉  **cetanx** 8 months, 2 weeks ago

I agree that B sounds like running ETL directly on the database. It doesn't say anything explicitly about using dumps.

However, by leveraging the Dataproc JDBC Connector, one can perform various operations such as querying, joining, filtering, or aggregating data from your SQL databases within your Dataproc jobs. This can be particularly useful when you want to combine data from multiple sources or perform complex data transformations before processing the data further.

So with D, you can run your analysis from a separate cloud-sql instance created from the dump and without affecting the production database.

upvoted 1 times

✉  **[Removed]**  3 years, 10 months ago

Answer: D

Description: Easy and it won't affect processing

upvoted 39 times

✉  **[Removed]** 2 years, 3 months ago

27 up vote for a wrong ans!!

Why do you need dataproc for MySQL dump?!

upvoted 31 times

✉  **StefanoG** 2 years, 4 months ago

Google Cloud Dataproc is not an analytic tool

upvoted 9 times

✉  **dambilwa** 3 years, 7 months ago

Agreed- Option[D] is most appropriate in this scenario

upvoted 5 times

✉  **StefanoG** 2 years, 4 months ago

So I vote for B

upvoted 6 times

✉  **TVH_Data_Engineer**  2 months, 1 week ago

Selected Answer: B

Based on these considerations, option B is likely the best approach. By using an ETL tool to load data from MySQL into Google BigQuery, you can leverage BigQuery's strengths in handling large-scale analytics workloads without impacting the performance of the operational databases. This option provides a clear separation of operational and analytical workloads and takes advantage of BigQuery's fast analytics capabilities.

upvoted 1 times

✉  **axantroff** 2 months, 1 week ago

Selected Answer: B

Do not spend much time on in - just B

upvoted 1 times

✉️  **rocky48** 2 months, 3 weeks ago

Selected Answer: B

Answer is B - Use an ETL tool to load the data from MySQL into Google BigQuery.

- * Google BigQuery is a serverless, highly scalable data warehouse that can handle large-scale analytics workloads without impacting your MySQL cluster's performance.
- * Using an ETL (Extract, Transform, Load) tool to transfer data from MySQL to BigQuery allows you to maintain a separate analytics environment, ensuring that your operational database remains unaffected.

Option C (connecting an on-premises Apache Hadoop cluster to MySQL and performing ETL) introduces complexity and may not be as scalable as a cloud-based solution.

Option D (mounting backups to Google Cloud SQL and processing the data using Google Cloud Dataproc) could be an option for historical data analysis but might not be the best choice for real-time analytics while the MySQL cluster is under heavy load. Additionally, the backups need to be restored and processed, which might introduce some delay.

upvoted 2 times

✉️  **mk_choudhary** 3 months, 1 week ago

It's GOOGLE exam where choosing the GCP service shall be first preference.

Now notice the problem statement "perform analytics with minimal impact on operations"

BigQuery is right option for analytic as well as Cloud SQL does provide easy export to GCS where we can query from BigQuery without loading into BQ to save storage cost.

upvoted 2 times

✉️  **rtcp0st** 3 months, 1 week ago

Selected Answer: B

B. Use an ETL tool to load the data from MySQL into Google BigQuery.

- * Google BigQuery is a serverless, highly scalable data warehouse that can handle large-scale analytics workloads without impacting your MySQL cluster's performance.

- * Using an ETL (Extract, Transform, Load) tool to transfer data from MySQL to BigQuery allows you to maintain a separate analytics environment, ensuring that your operational database remains unaffected.

Option C (connecting an on-premises Apache Hadoop cluster to MySQL and performing ETL) introduces complexity and may not be as scalable as a cloud-based solution.

Option D (mounting backups to Google Cloud SQL and processing the data using Google Cloud Dataproc) could be an option for historical data analysis but might not be the best choice for real-time analytics while the MySQL cluster is under heavy load. Additionally, the backups need to be restored and processed, which might introduce some delay.

upvoted 2 times

✉️  **melligeri** 3 months, 1 week ago

Selected Answer: B

The question clearly says there is load on MySQL already so doing analytics on it is bad idea. Its bad to run analytics on MySQL but still a better option to run etl with it to load it to BigQuery.

upvoted 1 times

✉️  **imran79** 3 months, 3 weeks ago

B. Use an ETL tool to load the data from MySQL into Google BigQuery. This way, analytics is entirely separated from the operational database and BigQuery is well-suited for large-scale analytics.

upvoted 2 times

✉️  **emmylou** 3 months, 3 weeks ago

The correct answer is to build a read replica :- but since we can't do that then migrating to BigQuery will have to suffice.

upvoted 2 times

✉️  **Fotofilico** 3 months, 3 weeks ago

thanks! :3

upvoted 1 times

✉️  **Nirca** 4 months ago

Selected Answer: D

Answer is Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

A: OLAP on MySQL performs poorly.

B: ETL consumes lot of MySQL resources, to read the data, as per question MySQL is under pressure already.

C: Similar to B.

D: By mounting backup can avoid reading from MySQL, data freshness is not an issue as per the question (and is not mentioned in the question)

upvoted 1 times

✉  **emmylou** 4 months, 1 week ago

Wouldn't the correct answer be to create read replica and do analytics off of that?

upvoted 1 times

✉  **boraxer1** 4 months, 3 weeks ago

Selected Answer: D

Answer is Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

A: OLAP on MySQL performs poorly.

B: ETL consumes lot of MySQL resources, to read the data, as per question MySQL is under pressure already.

C: Similar to B.

D: By mounting backup can avoid reading from MySQL, data freshness is not an issue as per the question (and is not mentioned in the question)

Reference:

<https://cloud.google.com/blog/products/data-analytics/genomics-data-analytics-with-cloud-pt2>

upvoted 1 times

✉  **FP77** 6 months ago

Selected Answer: B

Why overcomplicate things by using Dataproc? I choose B

upvoted 3 times

✉  **FP77** 6 months ago

I meant hadoop cluster

upvoted 1 times

✉  **theseawillclaim** 6 months, 2 weeks ago

I think it might be "C" because "B" mentions a random ETL tool, while C uses a more GCP-specific solution.

However, terrible question.

upvoted 1 times

✉  **Jarek7** 9 months ago

Stupid answer options. I'd mount a backup and use Dataflow to import into BQ.

upvoted 3 times

✉  **izekc** 9 months, 3 weeks ago

Selected Answer: D

D. minimal impact is the key hit

upvoted 2 times

Question #31

Topic 1

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

- A. Update the current pipeline and use the drain flag.
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Correct Answer: D

Community vote distribution

A (51%)

B (28%)

D (15%)

5%

 **VishalB** Highly Voted 3 years, 6 months ago

Correct Option : A

Explanation:-This option is correct as the key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option. Drain options would cause Dataflow to stop any new processing, but would also allow the existing processing to complete

upvoted 77 times

✉  **BigQuery** 2 years, 1 month ago

To all the New Guys Here. Please don't get confused with all the people's fight over here. Just google the question and you will get the cor ans in many website. Still I recommend to refer this website for question. for this Particular problem ans is A. Reason is here --> <https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python>
have time to read the full page when to use Update using Json mapping and when to use Drain. (you will have question following for Drain option though).

Thumb rule is this,

If any major change to windowing transformation (like completely changing window fn from fixed to sliding) in Beam/Dataflow/you want t stop pipeline but want inflight data --> use Drain option.

For all other use cases and Minor changing to windowing fn (like just changing window time of sliding window) --> Use Update with Jso mapping.

In this case it is Code change to new version. so, Update with Json mapping. Simple as that.

All the Best Guys.

upvoted 30 times

✉  **BigQuery** 2 years, 1 month ago

SORRY I MEANT TO SAY ANS IS 'B'. In this case it is Code change to new version. so, Update with Json mapping.

upvoted 6 times

✉  **anji007** 1 year, 10 months ago

Its clearly mentioned in the question that pipeline in compatible, if it is so you can not update with JSON mapping. Only way is to stop the pipeline with Drain and replace it with a new one. So the closest answer is A only.

upvoted 5 times

✉  **maxdataengineer** 1 year, 3 months ago

JSON Mapping is a way to solve compatibility issues when updating

upvoted 1 times

✉  **maxdataengineer** 1 year, 3 months ago

As you said, Drain stops the pipeline but it does not solve the compatibility issue. The pipeline will not be able to be updated which is the core problem of the question.

upvoted 3 times

✉  **assU2** 1 year, 2 months ago

You do not want to lose any data when making this update - is the core problem. You are doing it ANYWAY.

upvoted 2 times

✉  **VishalB** 3 years, 6 months ago

Option C & D are incorrect as Cancel Option will lead to loose the data

Option B is very Close, since the new Code make pipeline incompatible by providing transform mapping JSON file you can handle this

upvoted 3 times

✉  **Tanzu** 2 years ago

There are 5 update scenarios in a job in update-a-pipeline context.

a- changing transform name (requires mapping) , adding a new step (no need for mapping)

b- windowing or triggering (only for minor changes, otherwise don't do that)

c- coders (don't do that)

d- schema (adding or required to nullable is possible) other scenarios not possible

e- stateful operations

none of them are relevant, here. cause there is no specific detail, secondly incompatible w. new pipeline.

and mostly if in compatible only a has a solve. but not for all cases.

so, drain == no data loss (ingesting, buffered and in-flight data) is the only scenario.

upvoted 5 times

✉  **sergio6** 2 years, 3 months ago

A is incorrect because updating pipeline does not include any drain flag

upvoted 2 times

✉  **Tanzu** 2 years ago

two steps.. 1st drain the job w/ sdk or console. then, update the pipeline. cause it is OK to update a job while in draining

upvoted 2 times

✉  **maxdataengineer** 1 year, 3 months ago

Yes but the compatibility problem will still be there, stopping the pipeline does not solve that

upvoted 1 times

✉️  **sergio6** 2 years, 4 months ago

C and D are incorrect because canceling the old pipeline can cause data loss

<https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline>

A is incorrect because updating pipeline does not include any drain flag

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

upvoted 6 times

✉️  **Tanzu** 2 years ago

drain is in the guide ...stopping-a-pipeline. Just ...updating-a-pipeline is not enough to evaluate this question.

that's why drainning is not a flag in a pipeline update. it is a process about how to stop a pipeline w/o data loss !

data in dataflow is in 3 stages. ingestion data, buffered data and in-flight data which is processing by old pipeline.

upvoted 1 times

✉️  **sergio6** 2 years, 3 months ago

B is correct: Update the current pipeline and provide the transform mapping JSON object.

Dataflow always performs a compatibility check between the old and new job and without the mapping (necessary as old and new are incompatible) it would give an error and the old job would continue to be executed

<https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#Mapping>

<https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#CCheck>

upvoted 5 times

✉️  **Tanzu** 2 years ago

new pipeline is incompatible means, compatibility check will fail. so you wil not be able to update as new pipeline.

that's why B cannot be valid answer here in this context.

upvoted 2 times

✉️  **maxdataengineer** 1 year, 3 months ago

B is a way to solve compatibility issues

upvoted 2 times

✉️  **[Removed]**  3 years, 10 months ago

Correct B - https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#preventing_compatibility_breaks

upvoted 24 times

✉️  **[Removed]** 3 years, 10 months ago

Changing the pipeline graph without providing a mapping. When you update a job, the Dataflow service attempts to match the transforms your prior job to the transforms in the replacement job in order to transfer intermediate state data for each step. If you've renamed or removed any steps, you'll need to provide a transform mapping so that Dataflow can match state data accordingly.

upvoted 4 times

✉️  **arnabbis4u** 3 years, 9 months ago

The job can be incompatible for reasons other than transformation changes. Since it is clearly mentioned that the change job is incompatible, I think we have to create a new job and D should be correct.

upvoted 9 times

✉️  **sergio6** 2 years, 3 months ago

Canceling the job will cause data loss, against the requirement

upvoted 2 times

✉️  **philli1011**  3 days, 21 hours ago

Option: C

using draining will stop the subscription totally while allowing the existing data to complete processing. While the pipeline is stopped, will lose streaming data. The best option is to create a new pipeline that is connected to the same subscription, then we can apply drain to the old pipeline and end it. That way we will capture all the streaming data.

upvoted 1 times

MaxNRG 1 month, 2 weeks ago

Selected Answer: A

Drain flag: This flag allows the pipeline to finish processing all existing data in the Pub/Sub subscription before shutting down. This ensures no data is lost during the update.

Current pipeline: Updating the current pipeline minimizes disruption and avoids setting up entirely new infrastructure.

Incompatible changes: Even with incompatible changes, the drain flag ensures existing data is processed correctly.

upvoted 1 times

MaxNRG 1 month, 2 weeks ago

While other options might work in some cases, they have drawbacks:

B. Transform mapping JSON: This option is mainly for schema changes and doesn't guarantee data completion before shutdown.

C. New pipeline, same subscription: This risks duplicate processing of data if both pipelines run concurrently.

D. New pipeline, new subscription: This loses the current pipeline's state and potentially data, making it impractical for incompatible changes.

Therefore, the most reliable and data-safe approach is to update the current pipeline with the drain flag for seamless transition and data integrity.

Remember, always test updates in a staging environment before deploying to production.

upvoted 1 times

TVH_Data_Engineer 1 month, 2 weeks ago

Selected Answer: C

Same Cloud Pub/Sub Subscription: By using the same Cloud Pub/Sub subscription for the new pipeline, you ensure that no messages are lost during the transition. Pub/Sub manages message delivery, ensuring that unacknowledged messages (those that haven't been processed by the old pipeline) will be available for the new pipeline to process.

Creating a New Pipeline: Since the update makes the new pipeline incompatible with the current version, it's necessary to create a new pipeline. Attempting to update the current pipeline in place (options A and B) would not be feasible due to compatibility issues.

Cancel the Old Pipeline: Once the new pipeline is up and running and processing messages, you can safely cancel the old pipeline. This ensures a smooth transition with no data loss.

upvoted 1 times

JOKKUNO 1 month, 3 weeks ago

In order to make an update to a Google Cloud Dataflow streaming pipeline without losing any data, the recommended approach is:

A. Update the current pipeline and use the drain flag.

Explanation:

The drain flag is designed to allow the current pipeline to finish processing any remaining data before shutting down. This helps ensure that no data is lost during the update process.

By updating the current pipeline and using the drain flag, you allow the pipeline to complete its current processing before the update takes effect, minimizing the risk of data loss.

This approach is a safe way to transition from the old version to the new version without interrupting data processing.

upvoted 1 times

axantroff 2 months, 1 week ago

I would vote for A because of the structure of the exam, but there are other options worth considering as well

upvoted 1 times

RT_G 2 months, 3 weeks ago

Selected Answer: C

My answer is C. Chatted with ChatGPT and narrowed down on this option. Let me know your thoughts on this perspective.

Option C - By using the existing subscription, you can ensure that the data flow remains uninterrupted, and there is no loss of data during the transition from the old pipeline to the new one.

Creating a new pipeline that uses the same Cloud Pub/Sub subscription allows for a seamless transition without any interruptions to the data flow. This approach ensures that the new pipeline can continue to consume data from the same subscription as the old pipeline, thereby maintaining data continuity throughout the update process.

upvoted 1 times

rocky48 2 months, 3 weeks ago

Selected Answer: A

Correct Option : A

Explanation:-This option is correct as the key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option.

upvoted 1 times

✉  **mk_choudhary** 3 months, 1 week ago

It should be B

Drain will stop the existing job only and it does not suffice the updated schema.

In order to bring updated schema into effect, updated JSON mapping need to be applied.

upvoted 1 times

✉  **Simhamed2015** 3 months, 2 weeks ago

The two cores of this question are: 1- Don't lose data ← Drain, is perfect for this because you process all buffer data and stop reviving messages; normally this message is alive for 7 days of retry, so when you start a new job you will receive all without lose any data. 2- Incompatible new code ← mapping solve some incompatibilities like change name of a ParDO but no a version issue. So launch a new job with the new code, it's the only option.

So, option is A.

upvoted 1 times

✉  **imran79** 3 months, 3 weeks ago

The best choice is D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

upvoted 1 times

✉  **Mathew106** 6 months, 2 weeks ago

Selected Answer: A

Answer is A

<https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline#drain>

upvoted 1 times

✉  **itsmynickname** 6 months, 3 weeks ago

You don't want to lose data, so drain dude, drain!

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

Option C is the correct answer because creating a new pipeline with the same Cloud Pub/Sub subscription and canceling the old pipeline all for a seamless transition without any data loss. The new pipeline can consume from the same subscription as the old pipeline and continue processing data, while the old pipeline can be safely stopped without impacting data ingestion. This approach ensures continuity of data processing while updating the pipeline code.

upvoted 1 times

✉  **izekc** 9 months, 3 weeks ago

Selected Answer: A

A is correct since it is the first step to prevent data loss

upvoted 1 times

✉  **vivek123etl** 10 months, 3 weeks ago

<https://cloud.google.com/blog/products/gcp/managing-streaming-pipelines-in-google-cloud-dataflow-just-got-better>

upvoted 1 times

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

- A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
- B. The performance issue should be resolved over time as the size of the BigTable cluster is increased.
- C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
- D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Correct Answer: A

Community vote distribution

A (100%)

 **IsaB** Highly Voted 3 years, 4 months ago

I hate it when I read the question, than I think oh easy and I KNOW the answer, then I look at the choices and the answer I thought of is just not there at all... and I realize I absolutely have no idea :D

upvoted 48 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Correct A

upvoted 23 times

 **[Removed]** 3 years, 10 months ago

<https://cloud.google.com/bigtable/docs/performance#troubleshooting>

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

upvoted 18 times

 **JOKKUNO** Most Recent 1 month, 3 weeks ago

Improving performance in Google Cloud Bigtable involves optimizing the schema design to distribute the load efficiently across the clusters. Given the scenario, the best option would be:

- A. Redefine the schema by evenly distributing reads and writes across the row space of the table.

Explanation:

Distributing reads and writes evenly across the row space helps prevent hotspots and ensures that the load is spread evenly, avoiding performance bottlenecks.

Google Cloud Bigtable's performance is influenced by how well the data is distributed across the tablet servers, and evenly distributing the load can lead to better performance.

This approach aligns with best practices for designing scalable and performant Bigtable schemas.

upvoted 1 times

 **axantroff** 2 months, 1 week ago

Selected Answer: A

The comment from hilel_eth totally makes sense to me. I would go with A

upvoted 1 times

✉  **hkris909** 5 months, 2 weeks ago

Guys, how relevant are these questions, as of Aug 14, 2023 Could we clear the PDE exam with these set of questions?
upvoted 7 times

✉  **roty** 1 month, 2 weeks ago

HEY DID U CLEAR THE EXAM
upvoted 1 times

✉  **FP77** 6 months ago

Selected Answer: A

A is the only one that makes sense and is correct
upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

I understand why it could be A. But why not B also? Is it because of the typo saying BigDate instead of BigTable?
upvoted 1 times

✉  **Adswerve** 9 months, 3 weeks ago

Selected Answer: A

A to avoid hot-spotting <https://cloud.google.com/bigtable/docs/schema-design>
upvoted 2 times

✉  **Brilliantyagi** 1 year, 1 month ago

Selected Answer: A

A -
Make sure you're reading and writing many different rows in your table. Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Bigtable nodes, performance will suffer.

<https://cloud.google.com/bigtable/docs/performance#troubleshooting>

upvoted 6 times

✉  **hilel_eth** 1 year, 1 month ago

Selected Answer: A

A good way to improve read and write performance in a database system like Google Cloud Bigtable is to redefine the schema of the table so that reads and writes are evenly distributed across the row space of the table. This can help reduce bottlenecks in processing capacity and improve efficiency in table management. In addition, by evenly distributing read and write operations, it can prevent the accumulation of operations in one part of the table, which can improve the overall performance of the system.

upvoted 3 times

✉  **Pime13** 1 year, 6 months ago

Selected Answer: A

<https://cloud.google.com/bigtable/docs/keyvis-overview#what-is-keyvis>

To accomplish these goals, Key Visualizer can help you complete the following tasks:

Check whether your reads or writes are creating hotspots on specific rows
upvoted 4 times

✉  **Arkon88** 1 year, 11 months ago

Selected Answer: A

A is correct
<https://cloud.google.com/bigtable/docs/performance#troubleshooting>

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

upvoted 3 times

✉️  **samdhimal** 2 years ago

correct answer -> Redefine the schema by evenly distributing reads and writes across the row space of the table.

Make sure you're reading and writing many different rows in your table. Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Bigtable nodes, performance will suffer.

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

Reference: <https://cloud.google.com/bigtable/docs/performance#troubleshooting>

upvoted 2 times

✉️  **MaxNRG** 2 years, 2 months ago

A as the schema needs to be redesigned to distribute the reads and writes evenly across each table.

Refer GCP documentation - Bigtable Performance:

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

The table's schema is not designed correctly. To get good performance from Cloud Bigtable, it's essential to design a schema that makes it possible to distribute reads and writes evenly across each table. See Designing Your Schema for more information.

<https://cloud.google.com/bigtable/docs/schema-design>

Option B is wrong as increasing the size of cluster would increase the cost.

Option C is wrong as single row key for frequently updated identifiers reduces performance

Option D is wrong as sequential IDs would degrade the performance.

A safer approach is to use a reversed version of the user's numeric ID, which spreads traffic more evenly across all of the nodes for your Cloud Bigtable table.

upvoted 11 times

✉️  **anji007** 2 years, 3 months ago

Ans: A

upvoted 1 times

✉️  **sumanshu** 2 years, 6 months ago

Vote for A

upvoted 3 times

✉️  **timolo** 2 years, 6 months ago

Correct is A: <https://cloud.google.com/bigtable/docs/performance#troubleshooting>

Make sure you're reading and writing many different rows in your table. Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Bigtable nodes, performance will suffer.

upvoted 4 times

Question #33

Topic 1

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud

Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

- A. Check the dashboard application to see if it is not displaying correctly.
- B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
- C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
- D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Correct Answer: B

Community vote distribution

B (74%)

C (17%)

9%

✉️ [Removed] Highly Voted 3 years, 10 months ago

Answer: C

Description: Stackdriver can be used to check the error like number of unack messages, publisher pushing messages faster
upvoted 35 times

✉️ tprashanth 3 years, 6 months ago

B.

Stack driver monitoring is for performance, not logging of missing data.

upvoted 24 times

✉️ jkhong 1 year, 1 month ago

Please refer to this PubSub specific Monitoring metrics https://cloud.google.com/pubsub/docs/monitoring#monitoring_the_backlog
upvoted 1 times

✉️ mikey007 3 years, 6 months ago

<https://cloud.google.com/pubsub/docs/monitoring>

upvoted 2 times

✉️ ritinhabb 1 year, 7 months ago

Exactly!

upvoted 1 times

✉️ snamburi3 3 years, 2 months ago

All messages are being published to Cloud Pub/Sub successfully, so Stackdriver might not help.

upvoted 10 times

✉️ kubosuke 2 years, 4 months ago

messages sent successfully to Topic, but not Subscription.

in this case, if Dataflow cannot handle messages correctly it might not return acknowledgments to the Pub/Sub, and these errors can be seen from Monitoring.

https://cloud.google.com/pubsub/docs/monitoring#monitoring_exp

upvoted 12 times

✉️ Tanzu 2 years ago

to be more precise, first to publisher,

- then forwards to topic, and persistency for a while
- then forwards to subscribe,
- then to subscription..
- then acknowledgement happens

so in every step, there is possibly for errors.

upvoted 1 times

✉️ jkhong 1 year, 1 month ago

PubSub doesn't forward from subscriber to subscription. A topic sends it over to subscription first, then to subscriber

upvoted 2 times

✉️ [Removed] 3 years, 10 months ago

this will help us understand the reason, when we know that the data is not reaching subscriber then there is no point in checking it with dummy data

upvoted 12 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Should be B

upvoted 25 times

👤 [Removed] 3 years, 10 months ago

confused with D as well.

upvoted 1 times

👤 Tanzu 2 years ago

push or pull is about how target will handle the messages. pull mode gives flexibility when to get messages , so considering if target (or client) is slow, then it can make predictable choices.

dataflow is serverless. so if you need to awake it when necessary, you should use push mechanism. or leverage cloud composer/airflow and listen to pub/sub to trigger the dataflow.

upvoted 3 times

👤 Rajokkiyam 3 years, 10 months ago

Push or Pull guarantees the message to be delivered at-least once. So it doesn't make any difference.

upvoted 6 times

👤 gopinath_k 2 years, 10 months ago

Push needs Https endpoint

upvoted 3 times

👤 jvg637 3 years, 10 months ago

pushing is only for a https endpoint. So Dataflow just can pull messages

upvoted 4 times

👤 rtcpost Most Recent 3 months, 1 week ago

Selected Answer: B

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

* By running a fixed dataset through the Cloud Dataflow pipeline, you can determine if the problem lies within the data processing stage. This allows you to identify any issues with data transformation, filtering, or processing in your pipeline.

* Analyzing the output from this fixed dataset will help you isolate the problem and confirm whether it's related to data processing or the dashboard application.

upvoted 1 times

👤 ruben82 3 months ago

You must know what kind of data causes errors. I think, the first step is to get erroneous data and then test with sample of it.

upvoted 1 times

👤 imran79 3 months, 3 weeks ago

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output. If this results in the expected output, then the problem might be with the dashboard application (Option A), and that should be checked next.

upvoted 1 times

👤 WillemHendr 7 months, 3 weeks ago

Selected Answer: B

"...to find the missing messages"

Up to that remark, Monitoring was a valid option as well. But missing messages cannot be found with monitoring.

It is simply not possible to find the exact missing message. I read this remark as a test if you know what is, and what isn't possible with monitoring.

upvoted 3 times

👤 izekc 9 months, 3 weeks ago

Selected Answer: B

here is to determine next step. Not better way to optimize the workload. So B is the correct next step

upvoted 2 times

👤 Jarek7 9 months ago

B is not the next step. The next step is between pub/sub and dataflow(C). B will not help with it at all. However it could show the issue if in the pipeline or the view. But also it could not show it - you have no idea why some messages are not shown, so most probably it wouldn't give you any info. Definitely next step is to check if the issue is between pubsub and dataflow. Then you could go with B.

upvoted 1 times

✉  **Adswerve** 9 months, 3 weeks ago

Selected Answer: D

Pull subscription is the correct one. Push subscription means Dataflow cannot keep up with the topic.

upvoted 1 times

✉  **Jarek7** 9 months ago

It could be the issue. But C would reveal it if this is the real issue - if you will not check stackdriver, you cannot be sure if you really resolved the issue, as even if it seems to be working properly after switch to pull you cannot be sure if it is because of some other temporal factor.

upvoted 1 times

✉  **midgoo** 11 months, 1 week ago

Selected Answer: B

If the Dataflow does not have the expected output, it is either wrong at the input or at the pipelines. The chance that the issue is at the input (PubSub) is very low. For this case, it is likely the pipelines got some mistakes (e.g. JSON parsing failed). So we should follow B to debug the pipelines (using snapshot as test dataset for example)

upvoted 4 times

✉  **ploer** 12 months ago

Selected Answer: B

The most efficient solution would be to run a fixed dataset through the Cloud Dataflow pipeline and analyze the output (Option B). This will allow you to determine if the issue is with the pipeline or with the dashboard application. By analyzing the output, you can see if the messages are being processed correctly and determine if there are any discrepancies or missing messages. If the issue is with the pipeline, you can then debug and make any necessary updates to ensure that all messages are processed correctly. If the issue is with the dashboard application, you can then focus on resolving that issue. This approach allows you to isolate and identify the root cause of the missing messages in a controlled and efficient manner.

upvoted 7 times

✉  **Lestrang** 1 year ago

Selected Answer: B

I've just skimmed over the Stackdriver docs, yes guys, it helps you check the number and age of messages that were not received/acknowledged, excellent, hurrah.

So first off, C will not give us the missing messages, it will give us the count and age.
that means that C is inherently incorrect.

Additionally, will knowledge of the number of messages make resolving the problem any easier? No, it is just confirming what we already know.

Meanwhile, approach B, will allow us to see HOW and WHY it is missing some messages, which is the step that proceeds the fix.

upvoted 6 times

✉  **PolyMoe** 1 year ago

Selected Answer: C

Here is ChatGPT answer :

It's always a good practice to start by checking the logs and monitoring tools to see if there is any indication of an issue with the messages being published to Cloud Pub/Sub. In this case, you should use Google Stackdriver Monitoring to investigate if the missing messages have been published or not. You can also run a fixed dataset through the Cloud Dataflow pipeline to see if the pipeline is processing the messages correctly. If there is no issue found on the Cloud Pub/Sub and Cloud Dataflow, then you can check the dashboard application to see if it is not displaying the messages correctly. As a last resort, you can switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

upvoted 2 times

✉  **Jarek7** 9 months ago

I'm tired with these responses about what chatGPT says. Most probably you've used the free 3.5 version which is absolute disaster regarding being all knowing oracle. BTW in this case I wouldn't believe even GPT4. It is a difficult question that needs a specific knowledge and experience which might be not available in the GPT training data. You cannot use any GPT up to 4 as an argument in such cases.

upvoted 2 times

✉  **axantroff** 2 months, 1 week ago

Exactly. Sometimes it is total garbage

upvoted 1 times

✉  **Lestrang** 1 year ago

I provided it with the question as input but added the metrics available in Stackdriver, here is the response:

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

If messages are being published successfully to Cloud Pub/Sub but are missing in the dashboard, the issue is likely to be with the Cloud Dataflow pipeline that processes the messages. To find the root cause of the problem, you should run a fixed dataset through the pipeline and analyze the output. This will allow you to see if the pipeline is correctly processing all messages, and identify any processing errors that might be causing messages to be lost. The output can be compared to the expected results to identify any discrepancies and resolve the issue.

upvoted 4 times

✉  **desertlotus1211** 1 year ago

The question is really not asking for a solution to the problem, per se - but more of what would the next step in the situation to triage the issue

Answer would be B over D. Answer D would be the recommended solution IF the question asked to rectify/fixed the issue.

Thoughts?

upvoted 3 times

✉  **izekc** 9 months, 3 weeks ago

Agree with u

upvoted 1 times

✉  **Prakzz** 1 year, 1 month ago

Selected Answer: D

D. Dataflow must PULL the data to process it in real-time. Missing messages in the dashboard, means that the Pub/Sub to Dataflow was misconfigured as PUSH.

upvoted 3 times

✉  **hasoweh** 1 year ago

Pull will lead to latency as new data will not be streamed upon arrival, but instead will only be passed on when Dataflow makes a pull request. So if data comes in at time 0:01 but pull requests are only happening every 10 seconds, we have a 9-second delay. Push will automatically pass the data to any subscribers as soon as the data comes, and thus is closer to real-time.

upvoted 1 times

✉  **Krish6488** 1 year, 1 month ago

Selected Answer: B

To me, B sounds more logical for the below reason.

Option C would have been ideal because any debugging starts with checking the logs, however the option says, check stackdriver for missing messages. Had it been, check stackdriver to figure out the number of undelivered messages, C would have been more suitable. Given the slight bit of dodginess in option C, I would go with B

upvoted 3 times

✉  **Nirca** 1 year, 1 month ago

Selected Answer: B

Why checking Pub/Sub again when this is already verified to be fine according to the question. Shouldn't you be checking the next stage in the flow which is Dataflow?

Option - B

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

Answer - B. Because already we know message is missing so better to test with fixed dataset and check code .

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

C

<https://cloud.google.com/pubsub/docs/monitoring#:~:text=the%20specific%20metrics.-,Monitor%20message%20backlog,information%20about%20this%20metric%2C%20see%20the%20relevant%20section%20of%20this%20document.,-Create%20alerting%20policies>

upvoted 1 times

Question #34

Topic 1

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- ⇒ Databases
 - 8 physical servers in 2 clusters
 - SQL Server ` user data, inventory, static data
 - 3 physical servers
 - Cassandra ` metadata, tracking messages
- 10 Kafka servers ` tracking message aggregation and batch insert
 - ⇒ Application servers ` customer front end, middleware for order/customers
- 60 virtual machines across 20 physical servers
 - Tomcat ` Java services
 - Nginx ` static content
 - Batch servers
 - ⇒ Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) ` SQL server storage

- Network-attached storage (NAS) image storage, logs, backups

⇒ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

⇒ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

Build a reliable and reproducible environment with scaled party of production.

▪

⇒ Aggregate data in a centralized Data Lake for analysis

⇒ Use historical data to perform predictive analytics on future shipments

⇒ Accurately track every shipment worldwide using proprietary technology

⇒ Improve business agility and speed of innovation through rapid provisioning of new resources

⇒ Analyze and optimize architecture for performance in the cloud

⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

⇒ Handle both streaming and batch data

⇒ Migrate existing Hadoop workloads

⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.

⇒ Use managed services whenever possible

⇒ Encrypt data flight and at rest

⇒ Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO's tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.

Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to

BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

A. Store the common data in BigQuery as partitioned tables.

B. Store the common data in BigQuery and expose authorized views.

C. Store the common data encoded as Avro in Google Cloud Storage.

D. Store the common data in the HDFS storage for a Google Cloud Dataproc cluster.

Correct Answer: B

Community vote distribution

C (58%)

B (26%)

D (16%)

✉  **JOKKUNO** 1 month, 3 weeks ago

Given the scenario described for Flowlogistic's requirements and technical environment, the most suitable option for storing common data that is used by both Google BigQuery and Apache Hadoop/Spark workloads is:

C. Store the common data encoded as Avro in Google Cloud Storage.

upvoted 1 times

✉  **rtcp0st** 3 months, 1 week ago

Selected Answer: C

C. Store the common data encoded as Avro in Google Cloud Storage.

This approach allows for interoperability between BigQuery and Hadoop/Spark as Avro is a commonly used data serialization format that can be read by both systems. Data stored in Google Cloud Storage can be accessed by both BigQuery and Dataproc, providing a bridge between the two environments. Additionally, you can set up data transformation pipelines in Dataproc to work with this data.

upvoted 3 times

✉  **nescafe7** 6 months ago

Selected Answer: D

To simplify the question, Apache Hadoop and Spark workloads that cannot be moved to BigQuery can be handled by DataProc. So the correct answer is D.

upvoted 2 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: B

B is the right answer. Common data will lie in BigQuery but will be accessible via the views with SQL in Hadoop workloads.

upvoted 2 times

✉  **midgoo** 11 months, 1 week ago

Selected Answer: B

C should be the correct answer. However, please note that Google just released the BigQuery Connector for Hadoop, so if they ask the same question today, B will be the correct answer.

A could be correct too, but I cannot see why it has to be partitioned

upvoted 2 times

✉  **res3** 6 months, 4 weeks ago

If you check the <https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery>, it unloads the BQ data to GCS, utilizes it, and then deletes it from the GCS. Storing common data twice (at BQ and GCS) will not be the best option compared to 'C' (using GCS as the main common dataset).

upvoted 2 times

✉  **korntewin** 1 year ago

Selected Answer: C

I would vote for C as it can be used for analysis with Bigquery. Furthermore, Hadoop workload can also be transferred to dataproc connected to GCS.

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

Answer B looks ok, because in question they want to store common data which can be used by both workloads, and using BigQuery as the primary analytical tool would be the best option and easy to analyze common data.

upvoted 1 times

✉  **kelvintoy93** 1 year, 2 months ago

"Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data" - BigQuery can't take unstructured data so A and B are out.

Storing data in HDFS storage is never recommended unless latency is a requirement, so D is out.

That leaves us with GCS. Answer is C

upvoted 3 times

✉  **tunstila** 1 year ago

I thought you can now store unstructured data in BigQuery via the object tables announced during Google NEXT 2022... If that's possible, does that make B a better choice?

upvoted 1 times

✉  **drunk_goat82** 1 year, 2 months ago

Selected Answer: C

BigQuery can use federated queries to connect to the avro data in GCS while running spark jobs on it. If you duplicate the date you have to manage both data sets.

upvoted 2 times

✉  **wan2three** 1 year, 2 months ago

A

They wanted BigQuery. And connector is all you need to perform Hadoop or spark. Hadoop migration can be done using dataproc.

upvoted 1 times

✉  **wan2three** 1 year, 2 months ago

Also apparently they want all data at one place and want bigQ

upvoted 2 times

✉  **gudiking** 1 year, 2 months ago

Selected Answer: C

C as it can be used as an external table from BigQuery and with the Cloud Storage Connector it can be used by the Spark workloads (running Dataproc)

upvoted 1 times

✉  **solar_maker** 1 year, 2 months ago

Selected Answer: C

C, as both capable of AVRO, but the customer does not know what they want to do with the data yet.

upvoted 1 times

✉  **Leelas** 1 year, 2 months ago

Selected Answer: D

In Technical requirements it Was clearly mentioned that they need to Migrate existing Hadoop Cluster for which Data Proc Cluster is a replacement.

upvoted 1 times

✉  **vishal0202** 1 year, 4 months ago

C is ans...avro data can be accessed by spark as well

upvoted 4 times

✉  **ducc** 1 year, 4 months ago

Selected Answer: C

The answer is C

upvoted 3 times

Question #35

Topic 1

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand into. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- ⇒ Databases
 - 8 physical servers in 2 clusters
 - SQL Server " user data, inventory, static data
 - 3 physical servers
 - Cassandra " metadata, tracking messages
 - 10 Kafka servers " tracking message aggregation and batch insert
- ⇒ Application servers " customer front end, middleware for order/customers
 - 60 virtual machines across 20 physical servers
 - Tomcat " Java services
 - Nginx " static content
 - Batch servers
 - ⇒ Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) " SQL server storage
 - Network-attached storage (NAS) image storage, logs, backups
 - ⇒ 10 Apache Hadoop /Spark servers
 - Core Data Lake
 - Data analysis workloads
 - ⇒ 20 miscellaneous servers
 - Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled parity of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- ⇒ Handle both streaming and batch data
- ⇒ Migrate existing Hadoop workloads
- ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.
- ⇒ Use managed services whenever possible

- ⇒ Encrypt data flight and at rest
- ⇒ Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO's tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment. Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

- A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
- B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Correct Answer: C

Community vote distribution

A (90%)

10%

✉  **jvg637** Highly Voted 3 years, 10 months ago

I would say A.
I think Pub/Sub can't directly send data to Cloud SQL.
upvoted 38 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A
upvoted 15 times

✉  **JOKKUNO** Most Recent 1 month, 3 weeks ago

Given the requirements for ingesting data from global sources, processing and querying in real-time, and storing the data reliably for the real-time inventory tracking system, the most suitable combination of Google Cloud Platform (GCP) products is:
A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
Explanation:
Cloud Pub/Sub: It is a messaging service that allows you to asynchronously send and receive messages between independent applications.
Cloud Dataflow: It can handle both streaming and batch data, making it suitable for real-time processing of data from various sources.
Cloud Storage: Cloud Storage can be used to store the processed and analyzed data reliably. It provides scalable, durable, and globally accessible object storage, making it suitable for storing large volumes of data.
upvoted 1 times

✉️  **rtcpost** 3 months, 1 week ago

Selected Answer: A

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage

Here's why this combination is suitable:

Cloud Pub/Sub: It is used for ingesting real-time data from various global sources. It's a messaging service that can handle large volumes of data and is highly scalable.

Cloud Dataflow: It's a stream and batch data processing service that allows you to process and analyze the data in real-time. It can take data from Pub/Sub and perform transformations or aggregations as needed.

Cloud Storage: It provides reliable storage for the data. You can store the processed data in Cloud Storage for further analysis, and it is a scalable and durable storage solution.

Option B is not ideal because Local SSDs are not a suitable storage option for persisting data that needs to be reliably stored. Option C includes Cloud SQL, which is not typically used for ingesting and processing real-time data. Option D includes Cloud Load Balancing, which is not relevant to the use case of ingesting and processing data for the inventory tracking system.

upvoted 2 times

✉️  **Vipul1600** 6 months, 1 week ago

Since Cloud SQL is a fully managed service & Dataflow is serverless hence we should opt for Dataflow as it is a thumb rule for Google that we should choose a serverless product over a fully managed service.

upvoted 1 times

✉️  **Mathew106** 6 months, 2 weeks ago

Selected Answer: A

The technical requirements mention that the pipeline should handle both streaming and batch data. The solution should include DataFlow and not Cloud SQL. The answer is A.

upvoted 2 times

✉️  **niketd** 11 months, 2 weeks ago

Selected Answer: A

Pub/Sub to scale streaming data, Dataflow to process both structured and unstructured data and cloud storage to store common data

upvoted 1 times

✉️  **PolyMoe** 1 year ago

Selected Answer: A

Option B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD is not a good option as Local SSD is not a scalable solution and could not handle large amounts of data

Option C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage is not a good option as Cloud SQL is a relational database and is not suitable for real-time processing and querying large amounts of data

Option D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage is not a good option as Cloud Load Balancing is used for distributing traffic across multiple instances, it doesn't handle data processing and storage.

upvoted 1 times

✉️  **PolyMoe** 1 year ago

Selected Answer: A

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage is the best combination of GCP products for the use case described.

Cloud Pub/Sub can be used to ingest data from a variety of global sources, as it allows for easy integration with external systems through its publish-subscribe messaging model.

Cloud Dataflow can be used to process and query the data in real-time, as it is a fully managed service for creating data pipelines that can handle both batch and streaming data.

Cloud Storage can be used to store the data reliably, as it is a fully managed object storage service that can handle large amounts of data and is highly durable and available.

upvoted 1 times

✉️  **jkh_goh** 1 year ago

Selected Answer: A

Answer is A. Cloud Dataflow for batch + streaming, Cloud Pub/Sub for streaming ingestion, Cloud Storage for long term data storage.

upvoted 1 times

✉️  **Jay_Krish** 1 year, 2 months ago

Are scenario based questions still in the latest exam?? Are these still relevant?

upvoted 2 times

✉  **kastuarr** 1 year, 3 months ago

Selected Answer: C

Existing inventory data is in SQL, data ingested from Kafka will need to update inventory at some point. Existence of SQL in current estate indicates SQL must be present in the Cloud estate

upvoted 2 times

✉  **Dhams1** 1 year, 4 months ago

This site make me feel that it intends to make users to be involved in discussion by suggesting wrong answer

upvoted 9 times

✉  **Megmang** 1 year, 5 months ago

Selected Answer: A

Answer is clearly option A.

upvoted 3 times

✉  **rytizzle** 1 year, 6 months ago

why are there so many incorrect answers? it's so hard to study this way

upvoted 6 times

✉  **ratnesh99** 1 year, 6 months ago

Answer A : because Cloud Sql not suitable for Global

upvoted 1 times

✉  **CedricLP** 1 year, 9 months ago

Selected Answer: A

Only A can manage a lot's of data.

Target is Cloud Storage (obviously not SSD)

Input is Pub/Sub to replace Kafka

Cloud SQL + Storage has no sense in this context

upvoted 2 times

Question #36

Topic 1

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads

☞ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

☞ Databases

8 physical servers in 2 clusters

- SQL Server ` user data, inventory, static data

3 physical servers
- Cassandra ` metadata, tracking messages
10 Kafka servers ` tracking message aggregation and batch insert
⇒ Application servers ` customer front end, middleware for order/customs
60 virtual machines across 20 physical servers
- Tomcat ` Java services
- Nginx ` static content
- Batch servers
⇒ Storage appliances
- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) ` SQL server storage
- Network-attached storage (NAS) image storage, logs, backups
⇒ 10 Apache Hadoop /Spark servers
- Core Data Lake
- Data analysis workloads
⇒ 20 miscellaneous servers
- Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled panty of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- Handle both streaming and batch data
- - ⇒ Migrate existing Hadoop workloads
 - ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.
 - ⇒ Use managed services whenever possible
 - ⇒ Encrypt data flight and at rest
 - ⇒ Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO's tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment. Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very

technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?

- A. Export the data into a Google Sheet for virtualization.
- B. Create an additional table with only the necessary columns.
- C. Create a view on the table to present to the virtualization tool.
- D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

Correct Answer: C

Community vote distribution

C (68%)

B (32%)

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

upvoted 39 times

 **Radhika7983** Highly Voted 3 years, 2 months ago

Answer is C. A logical view can be created with only the required columns which is required for visualization. B is not the right option as you will create a table and make it static. What happens when the original data is updated. This new table will not have the latest data and hence view is the best possible option here.

upvoted 22 times

 **jin0** 11 months, 1 week ago

I don't think so because in question they worried about spending money for query but, using view could not make money safe because logical view scan all of the data in the table. so, for saving money for query then Answer B is more suitable

upvoted 2 times

 **rocky48** Most Recent 2 months, 3 weeks ago

Selected Answer: C

Answer: C

upvoted 2 times

 **rtcp0st** 3 months, 1 week ago

Selected Answer: C

C. Create a view on the table to present to the virtualization tool.

Creating a view in BigQuery allows you to define a virtual table that is a subset of the original data, containing only the necessary columns or filtered data that the sales team requires for their reports. This approach is cost-effective because it doesn't involve exporting data to external tools or creating additional tables, and it ensures that the sales team is working with the specific data they need without running expensive queries on the full dataset. It simplifies the data for non-technical users while keeping the data in BigQuery, which is a powerful and cost-effective data warehousing solution.

Options A (exporting to Google Sheet) and B (creating an additional table) might introduce data redundancy and maintenance overhead, and they don't provide the same level of control and security as creating a view. Option D (IAM roles) doesn't address the issue of simplifying the data for the sales team; it's more focused on access control.

upvoted 4 times

 **Mathew106** 6 months, 1 week ago

Selected Answer: C

C. You won't pay for storage for the view, and it will only include the necessary columns. Even if we assume that we don't talk about a materialized view, a logical view query can use the cache as much as a table query. So a new table does not have any benefit over a view, even if the view is logical.

upvoted 2 times

 **abi01a** 9 months, 2 weeks ago

The answer is C

upvoted 2 times

✉  **kplam** 9 months, 2 weeks ago

Answer is C

upvoted 3 times

✉  **lucaluka1982** 10 months, 1 week ago

Selected Answer: B

B it is more cost-effective and efficient approach to handle reports

upvoted 1 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 3 times

✉  **Booqq** 10 months, 4 weeks ago

C — view is better than another table to keep data consistent

upvoted 3 times

✉  **JJJJim** 11 months, 3 weeks ago

Selected Answer: C

Answer is C, creating view tables can easy and flexible to do the most cost-efftive way.

upvoted 2 times

✉  **PolyMoe** 1 year ago

Selected Answer: C

The appropriate solution is C, creating a view on the table, by selecting the relevant columns only (and not by creating another, static, table)

upvoted 2 times

✉  **dconesoko** 1 year ago

Selected Answer: B

Providing that the question was explicit in B and D about the selection of the appropriate columns it quite intriguing that it did not mention the selection of the appropriate column for the view. We can definitely build a view which might just present the same data or something much complex, thus i vote for B

upvoted 1 times

✉  **GCPpro** 1 year ago

C is the correct answer

upvoted 1 times

✉  **Isaga** 1 year ago

Selected Answer: C

I mean C

upvoted 1 times

✉  **noonting** 1 year ago

Selected Answer: B

The answer is B. Because it is not specified as a materialized view

upvoted 1 times

✉  **ler_mp** 1 year ago

But you would want a logical view for this, not materialized view

upvoted 1 times

✉  **Rodolfo_Marcos** 1 year, 1 month ago

A materialized view would be the best choice since it contains real-time streaming data and is also cost-effective since it only changes the de from query updates. This is possible by using smart caching.

upvoted 2 times

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

⇒ Databases

8 physical servers in 2 clusters

- SQL Server " user data, inventory, static data

3 physical servers

- Cassandra " metadata, tracking messages

10 Kafka servers " tracking message aggregation and batch insert

⇒ Application servers " customer front end, middleware for order/customs

60 virtual machines across 20 physical servers

- Tomcat " Java services

- Nginx " static content

- Batch servers

⇒ Storage appliances

- iSCSI for virtual machine (VM) hosts

- Fibre Channel storage area network (FC SAN) " SQL server storage

- Network-attached storage (NAS) image storage, logs, backups

⇒ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

⇒ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled parity of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- ⇒ Handle both streaming and batch data
- ⇒ Migrate existing Hadoop workloads
- ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.
- ⇒ Use managed services whenever possible
- ⇒ Encrypt data flight and at rest
- ⇒ Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO's tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single

Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in

Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

- A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
- B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub.
- C. Use the NOW () function in BigQuery to record the event's time.
- D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Correct Answer: B

Community vote distribution

B (73%)

D (27%)

✉  **Manue**  2 years, 9 months ago

"However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume."

Sure man, Kafka is not performing, let's use PubSub instead hahaha...

upvoted 34 times

✉  **sfsdeniso** 1 year, 2 months ago

google send via pub sub web indexes
twice a day a whole internet is being sent via pub sub
upvoted 2 times

✉  **ralf_cc** 2 years, 7 months ago

lol this is a vendor exam...
upvoted 7 times

✉️  **[Removed]**  3 years, 10 months ago

Answer: B

upvoted 23 times

✉️  **rtcpost**  3 months, 1 week ago

Selected Answer: B

B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub.

Here's why this approach is the most suitable:

By attaching a timestamp and Package ID at the point of origin (publisher device), you ensure that each message has a clear and consistent timestamp associated with it from the moment it is generated. This provides a reliable and accurate record of when each package-tracking message was created, which is crucial for analyzing the data over time.

This approach allows you to maintain the chronological order of events as they occurred at the source, which is important for real-time reports and historical analysis.

Option A suggests attaching the timestamp in the Cloud Pub/Sub subscriber application. While this can work, it introduces a potential delay as the risk of timestamps not being accurate if there are issues with message processing.

Option C, using the NOW() function in BigQuery, records the time when the data is ingested into BigQuery, which may not reflect the actual time of the event.

upvoted 3 times

✉️  **JJJJim** 11 months, 3 weeks ago

Selected Answer: B

Answer is B, attach the timestamp and ID is necessary to analyze data easily.

upvoted 1 times

✉️  **nidmed** 1 year, 9 months ago

Selected Answer: B

Answer: B

upvoted 4 times

✉️  **Arkon88** 1 year, 11 months ago

Selected Answer: B

we need package ID + Timestamp so B

upvoted 1 times

✉️  **davidqianwen** 2 years ago

Selected Answer: B

Answer: B

upvoted 1 times

✉️  **exnaniantwort** 2 years ago

Selected Answer: B

agree with humza

upvoted 1 times

✉️  **sraakesh95** 2 years ago

Selected Answer: D

<https://cloud.google.com/pubsub/docs/reference/rest/v1/PubsubMessage>

upvoted 4 times

✉️  **[Removed]** 2 years, 3 months ago

D is enough.. we have publish timestamp which is enough for this requirement

upvoted 1 times

✉️  **Tanzu** 2 years ago

there are 2 requirements

1- is about ordering due to historical data analysis

2- what it means to write a single topic and its impact.. why some sentence added here.

1st is primary, 2nd is secondary req. in this context.

So,

- in pub/sub, processTime is filled by server, not publisher. but that does not guarantee the ordering due to latency, pub/sub handling, sensor or any other reasons..

- you need to populate orderingKey field too, so that subscribers can get in ordered.

upvoted 3 times

✉️  **Pinko1497** 1 year, 9 months ago

Also, since this is a International company, adding timestamp on message receiving would help catch local time.

upvoted 1 times

✉️  **Chelseajcole** 2 years, 3 months ago

It is about processing time and event time.. Answer is B.

upvoted 1 times

✉️  **Tanzu** 2 years ago

not just timing, but also package-id .. cause they are sending 1 topic in gcp instead of to many in kafka. that means there must be added some additional critical data too.

upvoted 1 times

✉️  **anji007** 2 years, 3 months ago

Ans: B

A: Adding timestamp as they received is not a better option, messages may not arrive in order at the receiver/ subscriber, could be due to connectivity or network.

B: Timestamp should be added here.

C: Doesn't make sense at all.

D: Ordering should be based on the order how messages are generated at the publisher but not as per order they reach the pub/sub.

upvoted 4 times

✉️  **humza** 2 years, 6 months ago

Answer: B

A. There is no indication that the application can do this. Moreover, due to networking problems, it is possible that Pub/Sub doesn't receive messages in order. This will analysis difficult.

B. This makes sure that you have access to publishing timestamp which provides you with the correct ordering of messages.

C. If timestamps are already messed up, BigQuery will get wrong results anyways.

D. The timestamp we are interested in is when the data was produced by the publisher, not when it was received by Pub/Sub.

upvoted 7 times

✉️  **sumanshu** 2 years, 6 months ago

Vote for B

upvoted 2 times

✉️  **funtoosh** 2 years, 11 months ago

Better if the publisher attached the package ID and Timestamp as packages can come in an Asynchronous fashion.

upvoted 3 times

✉️  **naga** 2 years, 11 months ago

Correct B

upvoted 3 times

👤 **Radhika7983** 3 years, 2 months ago

The answer is B.

JSON representation

```
{  
  "data": string,  
  "attributes": {  
    string: string,  
    ...  
  },  
  "messageId": string,  
  "publishTime": string,  
  "orderingKey": string  
}
```

In the attribute, we can have package id and timestamp.

Question #38

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

- ⇒ Ensure secure and efficient transport and storage of telemetry data
- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

- A. The zone
- B. The number of workers
- C. The disk size per worker
- D. The maximum number of workers

Correct Answer: A

Community vote distribution

D (100%)

✉  **jvg637** Highly Voted 3 years, 10 months ago

D. The maximum number of workers answers to the scale question
upvoted 26 times

✉  **Radhika7983** Highly Voted 3 years, 2 months ago

The correct answer is D. Please look for the details in below
<https://cloud.google.com/dataflow/docs/guides/specifying-exec-params>
We need to specify and set execution parameters for cloud data flow .

Also, to enable autoscaling, set the following execution parameters when you start your pipeline:

--autoscaling_algorithm=THROUGHPUT_BASED

--max_num_workers=N

The objective of autoscaling streaming pipelines is to minimize backlog while maximizing worker utilization and throughput, and quickly react to spikes in load. By enabling autoscaling, you don't have to choose between provisioning for peak load and fresh results. Workers are added as CPU utilization and backlog increase and are removed as these metrics come down. This way, you're paying only for what you need, and the pipeline is processed as efficiently as possible.

upvoted 25 times

✉  **rtcpost** Most Recent 3 months, 1 week ago

Selected Answer: D

D. The maximum number of workers

By increasing the maximum number of workers, you ensure that Cloud Dataflow can scale its compute power to handle the increased data processing load efficiently.

upvoted 2 times

✉  **vaga1** 8 months, 1 week ago

Selected Answer: D

Cloud Dataflow auto-scales, then if it is not scaling it is because it has reached the maximum number of workers that have been set.

upvoted 2 times

✉  **abi01a** 9 months, 2 weeks ago

A is the correct answer. Dataflow is Serverless. Specify your Region, autoscaling and other 'knobing' activities that are 'under the hood' will be taken care for you. Remember the company cannot afford to staff an Operations team to monitor data feeds so rely on ...

upvoted 2 times

✉  **bha11111** 10 months, 3 weeks ago

Selected Answer: D

this is correct

upvoted 2 times

✉  **GCPpro** 1 year ago

D . is the correct answer

upvoted 1 times

✉  **jkh_goh** 1 year ago

Answer A provided is definitely wrong. Who comes up with these answers?

upvoted 1 times

✉  **Ender_H** 1 year, 4 months ago

Selected Answer: D

Correct Answer: D

 A: The zone has nothing to do with scaling computer power.

 B: The key word here is, "Scale its compute power up AS REQUIRED", with this answer, the number of workers would immediately scale to computer power.

 C: we need to scale compute power, not storage

 D: is the correct answer, changing the Number of Maximum workers will allow Dataflow to add up to that number of workers if required.

https://cloud.google.com/dataflow/docs/reference/pipeline-options#resource_utilization

upvoted 2 times

✉  **sraakesh95** 2 years ago

Selected Answer: D

@Radhika7983

upvoted 2 times

✉  **medeis_jar** 2 years ago

Selected Answer: D

The correct answer is D.

<https://cloud.google.com/dataflow/docs/guides/specifying-exec-params>

We need to specify and set execution parameters for cloud data flow .

Also, to enable autoscaling, set the following execution parameters when you start your pipeline:

--autoscaling_algorithm=THROUGHPUT_BASED

--max_num_workers=N

upvoted 2 times

✉  **maurodipa** 2 years, 2 months ago

Answer is A: Dataflow is serverless, so no need to specify neither the number of workers, nor the max number of workers.

<https://cloud.google.com/dataflow>

upvoted 5 times

✉  **Jarek7** 9 months ago

Have you ever use it? You pay for workers processing, so you specify max number of workers. Here is the doc:

<https://cloud.google.com/sdk/gcloud/reference/dataflow/jobs/run>

upvoted 1 times

✉  **anji007** 2 years, 3 months ago

Ans: D

upvoted 1 times

⊕  **sumanshu** 2 years, 6 months ago

Vote for D

upvoted 3 times

⊕  **Lodu_Lalit** 2 years, 10 months ago

D, that's because scalability is directly correlated to max number of workers, size determines the speed of functioning
upvoted 3 times

⊕  **naga** 2 years, 11 months ago

Correct D

upvoted 4 times

⊕  **daghayeghi** 3 years ago

A exactly is correct:

Because Dataflow is scalable and don't need to repeat like cluster in Spark. then the only problem that remain is the zone of data that is important.

upvoted 3 times

⊕  **daghayeghi** 2 years, 10 months ago

D is correct,

it was my mistake, because it said: "as required". then we can increase maxNumWorkers parameter to be used in required situation.

upvoted 2 times

Question #39

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

⇒ Provide reliable and timely access to data for analysis from distributed research workers

Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

- ⇒ Ensure secure and efficient transport and storage of telemetry data
- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualizations for operations teams with the following requirements:

- ⇒ The report must include telemetry data from all 50,000 installations for the most recent 6 weeks (sampling once every minute).
- ⇒ The report must not be more than 3 hours delayed from live data.
- ⇒ The actionable report should only show suboptimal links.
- ⇒ Most suboptimal links should be sorted to the top.
- ⇒ Suboptimal links can be grouped and filtered by regional geography.
- ⇒ User response time to load the report must be <5 seconds.

Which approach meets the requirements?

- Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
- Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
- Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
- Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Correct Answer: C

Community vote distribution

D (83%)

C (17%)

 **[Removed]**  3 years, 10 months ago

Answer: D

upvoted 28 times

👤 **itche_scratche** Highly Voted 3 years, 9 months ago

D; dataflow doesn't connect to datastore, and not really for reporting. BQ, and data studio is a better choice.
upvoted 13 times

👤 **ckanaar** 4 months, 1 week ago

Dataflow does connect to Datastore, D is still the right answer though.
upvoted 1 times

👤 **rtcp0st** Most Recent 3 months, 1 week ago

Selected Answer: D

D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Here's why this option is the most suitable:

Google BigQuery is a powerful data warehouse for processing and analyzing large datasets. It can efficiently handle the telemetry data from a 50,000 installations.

Google Data Studio 360 is designed for creating interactive and visually appealing reports and dashboards.

Using Google Data Studio allows you to connect to BigQuery, calculate the required metrics, and apply filters to show only suboptimal links. It can provide real-time or near-real-time data updates, ensuring that the report is not more than 3 hours delayed from live data.

Google Data Studio can also be used to sort and group suboptimal links and display them based on regional geography.

With the right design, you can ensure that user response time to load the report is less than 5 seconds.

This approach leverages Google's cloud services effectively to meet the specified requirements.

upvoted 2 times

👤 **theseawillclaim** 6 months, 2 weeks ago

Selected Answer: D

Why bother with a custom GAE app when you have Data Studio?

upvoted 3 times

👤 **DGames** 1 year, 1 month ago

Selected Answer: C

Its think answer would be C because of telemetry data and response time is <5 second that force me to think about datastore,

upvoted 2 times

👤 **willymac2** 1 year, 7 months ago

I believe the answer is C.

First requirement is that it must be a visualisation with, so A and B do not work (create a table and a spreadsheet).

Now the second constraint which I believe is important is that the report MUST load in less than 5 seconds. But we do not know how complex the metric computation is, thus I cannot assume that we can compute it when we want to load the report, making me think that it must be pre-computed. Thus option D cannot work as it creates the metric AFTER querying the data (we are also not sure if we can really compute it in query).

upvoted 4 times

👤 **gudguy1a** 4 months, 3 weeks ago

Ummm, sorry @willymac2, but you have to account for size and growth which datastore cannot scale to.

Then, you have to worry about sub-second response time and datastore cannot do that as well as BigQuery...

upvoted 1 times

👤 **Raj0123** 1 year, 8 months ago

Answer D

upvoted 1 times

👤 **CedricLP** 1 year, 9 months ago

Selected Answer: D

DataStudio and BQ are the simplest way to do it

upvoted 1 times

👤 **devric** 1 year, 10 months ago

Selected Answer: D

They also can activate BI Engine feature to improve the response time.

upvoted 1 times

✉  **sraakesh95** 2 years ago

Selected Answer: D

D: Usually when a reporting tool is involved for GCP, DataStudio mostly goes by default due to it's no cost analytics and BigQuery joins it due it's OLAP nature and the wonderful integration provided by GCP for these 2

upvoted 2 times

✉  **medeis_jar** 2 years ago

Selected Answer: D

as explained by JayZeeLee

upvoted 1 times

✉  **JayZeeLee** 2 years, 2 months ago

D.

A and B are incorrect, because Google Sheets are not the best fit to handle large amount of data.

C may work, but it requires building an application which equates to more work.

D is more efficient, therefore a better option.

upvoted 4 times

✉  **wubston** 1 year, 2 months ago

I can't think of a single compelling reason to go with anything but D, given the scope definition in the question brief.

upvoted 1 times

✉  **Chelseajcole** 2 years, 3 months ago

Visualization = Data Studio 360

upvoted 1 times

✉  **Chelseajcole** 2 years, 3 months ago

Next question give you the answer: Question #40. They using Data Studio 360 and Bigquery as source

upvoted 1 times

✉  **anji007** 2 years, 3 months ago

Ans: D

upvoted 1 times

✉  **sumanshu** 2 years, 6 months ago

Vote for D

upvoted 3 times

✉  **zosoabi** 2 years, 8 months ago

just check the next question (#40) to get an idea about correct answer

upvoted 3 times

✉  **BhupiSG** 2 years, 10 months ago

D

Correct

upvoted 2 times

Question #40

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive

hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers

▪

- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

- ⇒ Ensure secure and efficient transport and storage of telemetry data

- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.

Which two actions should you take? (Choose two.)

- A. Ensure all the tables are included in global dataset.
- B. Ensure each table is included in a dataset for a region.
- C. Adjust the settings for each table to allow a related region-based security group view access.
- D. Adjust the settings for each view to allow a related region-based security group view access.
- E. Adjust the settings for each dataset to allow a related region-based security group view access.

Correct Answer: BD

Community vote distribution

BE (51%)	BC (31%)	Other
----------	----------	-------

 **[Removed]**  3 years, 10 months ago

Answer: B E

upvoted 46 times

 **shanjin14**  2 years, 8 months ago

C is correct starting 2020, as BigQuery come with table level access control

<https://cloud.google.com/blog/products/data-analytics/introducing-table-level-access-controls-in-bigquery>

upvoted 33 times

 **samstar4180** 2 years, 5 months ago

Yes, the correct answer should be BC - since we can have table-level access and each region represents a table.

upvoted 10 times

 **rocky48**  2 months, 3 weeks ago

Answer : BE

B. Ensure each table is included in a dataset for a region.

This means that you should organize your data in BigQuery into separate datasets, one for each region. Each dataset contains the tables specific to that region. This ensures that data is segregated by region.

E. Adjust the settings for each dataset to allow a related region-based security group view access.

upvoted 2 times

 **rtcp0st** 3 months, 1 week ago

Selected Answer: BE

B. Ensure each table is included in a dataset for a region.

This means that you should organize your data in BigQuery into separate datasets, one for each region. Each dataset contains the tables specific to that region. This ensures that data is segregated by region.

E. Adjust the settings for each dataset to allow a related region-based security group view access.

You should set the access controls at the dataset level in BigQuery. This means configuring access permissions for each dataset based on regional security groups. This way, you can enforce the regional access policy to the data, ensuring that users from different regions can only access the data associated with their region.

Option A is not necessary because you don't need to include all the tables in a global dataset. Segregating data into region-specific datasets is a better approach for enforcing access controls.

Options C and D are not typical actions in BigQuery. Access control and permissions are usually managed at the dataset level, and you can grant access to specific groups at that level.

upvoted 1 times

 **Mathew106** 6 months, 1 week ago

It's B and E or B and C. However, B and E makes some more sense because if you have one dataset for each region and they need to access data for each region then why not allow them access to the whole dataset? What if you want to add other supplementary tables later? If you create that on a table level you would have to add access to every table separately.

Still, I think both are valid because we don't have any extra requirement, but B E makes more sense.

upvoted 4 times

✉  **Jarek7** 9 months ago

Selected Answer: C

The intended answer was for sure BE. If C or D would be the right answers there is absolutely no reason to do B, right? Why should you put e table into separate dataset if you then set the accesss on table/view level? What is more the question is about tables not views, so I have no i why would anybody take D.

The issue is that this question is out of date and now the right answer would be sole C.

upvoted 4 times

✉  **Oleksandr0501** 9 months, 1 week ago

The two actions that should be taken are B and E.

B. Ensure each table is included in a dataset for a region: By creating separate datasets for each region and including only the tables associat with that region, you can enforce the regional access policy.

E. Adjust the settings for each dataset to allow a related region-based security group view access: By adjusting the settings for each dataset to allow only the related region-based security group view access, you can ensure that employees can only view data associated with their regic

A is incorrect because including all tables in a global dataset would not enforce the regional access policy.

C is incorrect because adjusting the settings for each table is not a scalable solution, especially as the number of tables grows.

D is incorrect because adjusting the settings for each view does not ensure that employees can only view data associated with their region.

upvoted 2 times

✉  **sjtesla** 9 months, 2 weeks ago

Selected Answer: BC

B: Location is on dataset level: https://cloud.google.com/bigquery/docs/datasets#dataset_limitations

C: IAM can be set on table level

upvoted 2 times

✉  **Lestrang** 1 year ago

Guys,

there are 2 possible combinations

If you think that each table represents a region, then they should all be in a global dataset and you should apply table access control to them. So A+C

Otherwise you would put each table in a regional dataset, and apply access control to the dataset. Why would you create a dataset for the purpose of controlling regional access, and then only apply the controls to a table inside it? that is not extensible in the future.

Anyway create dataset+access control for dataset (B+E) is also valid.

Which to choose? I dont know.

upvoted 3 times

✉  **PolyMoe** 1 year ago

Selected Answer: BE

First put tables in region-dedicated dataset (B)

Then, ensure access control at dataset level (by creating region-based security groups) (E)

upvoted 2 times

✉  **korntewin** 1 year ago

Selected Answer: AC

I would vote for AC. As we already split the table for each region, why do we need to split the dataset per region? Furthermore, the access control will be provided to the users based on table level anyway.

upvoted 3 times

✉  **korntewin** 1 year ago

Oh, the location should be specified in the dataset level! Then, the dataset should be splitted by region, my bad!

upvoted 1 times

✉  **MisuLava** 1 year, 3 months ago

Selected Answer: BE

if you create table-level access control and grant it to different groups for different tables, what is the point of putting tables in different database and different regions?

So i choose BE

upvoted 7 times

 **svkds** 1 year, 4 months ago

Selected Answer: BC

BigQuery come with table level access control. Since we can have table-level access and each region represents a table, B & C is correct answer.

upvoted 2 times

 **ducc** 1 year, 5 months ago

Selected Answer: BC

I voted for BC

upvoted 1 times

 **NR22** 1 year, 9 months ago

Selected Answer: BC

can apply IAM policies at table level

upvoted 1 times

 **devric** 1 year, 10 months ago

Selected Answer: BE

B & E. If you want to apply access controls over tables and views, you should apply them over datasets.

upvoted 1 times

 **szl0144** 1 year, 10 months ago

Selected Answer: BC

BC are correct

upvoted 1 times

Question #41

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.

⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

-
- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day.

Which schema should you use?

- A. Rowkey: date#device_id Column data: data_point
- B. Rowkey: date Column data: device_id, data_point
- C. Rowkey: device_id Column data: date, data_point
- D. Rowkey: data_point Column data: device_id, date
- E. Rowkey: date#data_point Column data: device_id

Correct Answer: D

Community vote distribution

A (57%)

C (35%)

9%

✉  **itche_scratche**  3 years, 9 months ago
None, rowkey should be Device_Id+Date(reverse)
upvoted 81 times

✉  **maxdataengineer** 1 year, 3 months ago

Actually it depends. You will only have 365×2 dates unique dates at a given time, since is a 2 year history, while most likely have more devices than that. So it will make more sense to start with the date first instead of the device

upvoted 14 times

✉  **Ankit267** 3 years, 6 months ago

A - Key should be less granular item first to more granular item, there are more devices than date key (every 15 min)
upvoted 16 times

✉  **Jlozano** 2 years, 1 month ago

A - "Date#Device_Id" is not the same that "Timestamp#Device_Id". If you want to query historical data, rowkey as "2021-12-09#12345dev" is optimal design. Nevertheless, "2021-12-09:09:10:47:2000#12345device" isn't it. Each record has a date (2021-12-09) and unique device (12345, 12346, 12347...).

upvoted 20 times

✉  **Rajuuu** 3 years, 6 months ago

A is a better option then other ..though not perfect as you mentioned.
upvoted 5 times

✉  **jvg637**  3 years, 10 months ago

think is A, since "The most common query is for all the data for a given device for a given day", rowkey should have info for both device and date.

upvoted 17 times

✉  **michaelkhan3** 2 years, 4 months ago

Google specifically mentions that it's a bad idea to use a timestamp at the start of a rowkey

<https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid>

The answer really should be Device_id#Timestamp but with the answers we were given you would be better off leaving the timestamp out together

upvoted 12 times

✉  **Whoswho** 1 year, 1 month ago

I remember seeing it as well. the answer should be A. (reversed)
upvoted 1 times

✉  **wan2three** 1 year, 2 months ago

but it didn't say can't use date, date and timestamp are different
upvoted 5 times

✉  **FP77** 5 months ago

The date is even worse than timestamp for the problem of hot-spotting
upvoted 1 times

✉  **gise**  1 week, 3 days ago

Selected Answer: C

C. This schema is best suited for historical analysis of device data over time when the most common query is to retrieve all data for a **specific device** on a **given day**.

* **Row Key as `device_id`:** This allows for efficient retrieval of all data points related to a particular device in a single operation. Bigtable sorts data lexicographically by row key, so all data for a single device will be stored together.

* **Column with `date` and `data_point`:**

- Using `date` as a column name or part of the column qualifier allows you to quickly filter and retrieve data for specific date ranges.
- Storing `data_point` as the column value provides the actual data associated with each timestamp.

Example:

With this schema, a query to get all data for `device_12345` on `2023-12-20` would efficiently target the specific row key `device_12345` and filter the relevant columns (with dates around `2023-12-20`).

upvoted 1 times

✉  **JonFrow** 1 month, 2 weeks ago

C - the answer should be the right answer.

Key is "all the data for a given device for a given day"

as in, Device first, and all the data + data points after.

This has nothing to do with Date-based search.

upvoted 1 times

✉️  **rocky48** 2 months, 3 weeks ago

Selected Answer: A

A - Key should be less granular item first to more granular item, there are more devices than date key (every 15 min)
upvoted 1 times

✉️  **imran79** 3 months, 3 weeks ago

the closest match to this in the provided options is:

C. Rowkey: device_id Column data: date, data_point

Thus, option C would be the best choice from the given option

upvoted 1 times

✉️  **kenwilliams** 8 months, 1 week ago

Selected Answer: A

It all comes down to the most common query

upvoted 3 times

✉️  **FP77** 5 months ago

Exactly

"all the data for a given device for a given day"

That's why the answer is C. You start by selecting the device and then the date. This solution is not prone to hot-spotting, yours is.

upvoted 1 times

✉️  **PolyMoe** 1 year ago

Selected Answer: A

A. Rowkey: date#device_id Column data: data_point This schema would allow querying all data for a given device for a given day by looking at the row key, which would be the date followed by the device_id. This would be the most efficient way to access the data as it would be stored in sorted order by date and device_id.

upvoted 2 times

✉️  **GCPpro** 1 year ago

A is the answer.

upvoted 2 times

✉️  **Jackalski** 1 year, 1 month ago

Selected Answer: A

I vote on A, none is the ideal answer as often here needs to distribute data within cluster by date and device. this key will answer most used query - so OK for me

upvoted 2 times

✉️  **slade_wilson** 1 year, 1 month ago

Answer : A

upvoted 2 times

✉️  **DGames** 1 year, 1 month ago

Selected Answer: D

I would with option D because it clearly mention the access pattern - all the data for a given device for a given day.

upvoted 1 times

✉️  **Gudwin** 1 year, 2 months ago

Obv you can't have date as the key, that would mean getting yourself a nice hotspot. The answer is C.

upvoted 3 times

✉️  **ejlp** 1 year, 2 months ago

Selected Answer: A

focus on "for a given device for a given day"

upvoted 1 times

✉️  **NM1212** 1 year, 6 months ago

Out of the given options, only option A gives you the read use case of retrieving data by device by date. All other options will require a full scan of the table which is not ideal.

Even though that option is not write optimized as it will cause hot-spotting, it is still the only possible option for the given use case.

upvoted 5 times

✉  **gcpdata** 1 year, 7 months ago

Selected Answer: C

need to make a key of device id and date, date is a secondary key.

upvoted 2 times

✉  **rr4444** 1 year, 7 months ago

Yup this Q&A is a mess

upvoted 2 times

Question #42

Topic 1

Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

- A. Rewrite the job in Pig.
- B. Rewrite the job in Apache Spark.
- C. Increase the size of the Hadoop cluster.
- D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Correct Answer: A

Community vote distribution

B (100%)

✉  **jvg637** Highly Voted 3 years, 10 months ago

I would say B since Apache Spark is faster than Hadoop/Pig/MapReduce

upvoted 35 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: B

Description: Spark performs in-memory processing and faster, which results in optimization of job's processing time

upvoted 17 times

✉  **axantroff** Most Recent 2 months, 1 week ago

Selected Answer: B

Just a regular Spark. B

upvoted 1 times

□  **DataFrame** 2 months, 2 weeks ago

C. I think it should be C because intent of asking question is to realize the problem of on-prem auto-scaling not the optimization that we achieve using spark in-memory features. Its GCP exam they want to highlight if hadoop cluster commodity hard doesn't increase when data increases then it can create problem unlike GCP. Hence migrate to GCP.

upvoted 1 times

□  **itsmynickname** 6 months, 3 weeks ago

None. Being a GCP exam, it must be either Dataflow or BigQuery :D

upvoted 7 times

□  **KHAN0007** 9 months, 1 week ago

I would like to take a moment to thank you all guys

You guys are awesome!!!

upvoted 3 times

□  **ler_mp** 1 year ago

Wow, a question that does not recommend to use Google product

upvoted 12 times

□  **Whoswho** 1 year, 1 month ago

looks like he's trying to spark the company up.

upvoted 6 times

□  **itsmynickname** 6 months, 3 weeks ago

It seems he's not well paid.

upvoted 1 times

□  **Krish6488** 1 year, 1 month ago

Selected Answer: B

Both Pig & Spark requires rewriting the code so its an additional overhead, but as an architect I would think about a long lasting solution. Resizing Hadoop cluster can resolve the problem statement for the workloads at that point in time but not on longer run. So Spark is the right choice, although its a cost to start with, it will certainly be a long lasting solution

upvoted 2 times

□  **Mamta072** 1 year, 7 months ago

Ans is B . Apache spark.

upvoted 2 times

□  **alecuba16** 1 year, 9 months ago

Selected Answer: B

SPARK > hadoop, pig, hive

upvoted 4 times

□  **kped21** 1 year, 11 months ago

B - Apache Spark

upvoted 1 times

□  **luamail** 9 months, 4 weeks ago

<https://www.ibm.com/cloud/blog/hadoop-vs-spark>

upvoted 2 times

□  **kped21** 2 years ago

B Spark for optimization and processing.

upvoted 1 times

□  **sraakesh95** 2 years ago

Selected Answer: B

B: Spark is suitable for the given operation is much more powerful

upvoted 1 times

□  **medeis_jar** 2 years ago

Selected Answer: B

as explained by pr2web

upvoted 1 times

👤 **pr2web** 2 years, 1 month ago

Selected Answer: B

Ans B:

Spark is a 100 times faster and utilizes memory, instead of Hadoop Mapreduce's two-stage paradigm.

upvoted 1 times

👤 **MaxNRG** 2 years, 2 months ago

B as Spark can improve the performance as it performs lazy in-memory execution.

Spark is important because it does part of its pipeline processing in memory rather than copying from disk. For some applications, this makes Spark extremely fast.

upvoted 1 times

👤 **MaxNRG** 2 years, 2 months ago

With a Spark pipeline, you have two different kinds of operations, transforms and actions. Spark builds its pipeline using an abstraction called a directed graph. Each transform builds additional nodes into the graph but Spark doesn't execute the pipeline until it sees an action. Spark waits until it has the whole story, all the information. This allows Spark to choose the best way to distribute the work and run the pipeline. The process of waiting on transforms and executing on actions is called, lazy execution. For a transformation, the input is an RDD and the output is an RDD. When Spark sees a transformation, it registers it in the directed graph and then it waits. An action triggers Spark to process the pipeline, the output is usually a result format, such as a text file, rather than an RDD.

upvoted 1 times

👤 **MaxNRG** 2 years, 2 months ago

Option A is wrong as Pig is a wrapper and would initiate Map Reduce jobs

Option C is wrong as it would increase the cost.

Option D is wrong Hive is a wrapper and would initiate Map Reduce jobs. Also, reducing the size would reduce performance.

upvoted 3 times

👤 **kastuarr** 1 year, 3 months ago

Wont Option B increase the cost ? Cost of re-writing the job in Spark + Cost of additional memory ?

upvoted 1 times

Question #43

Topic 1

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

- A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
- B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
- D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

Correct Answer: C

Community vote distribution

A (45%)

B (28%)

C (20%)

7%

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Answer will be A because when you create View it does not store extra space and its a logical representation, for rest of the option you need to

write large code and extra processing for dataflow/dataproc

upvoted 64 times

✉  **funtoosh** 2 years, 11 months ago

cannot be 'A' as it clearly says that you need to change the schema and data.

upvoted 17 times

✉  **exnaniantwort** 2 years ago

Your primary task is to "make data available".

Changing the schema is just the request from the member "A member of IT is building an application and ***asks you to modify the schema and data*** in BigQuery". You don't have to follow it if it does not make sense.

upvoted 15 times

✉  **exnaniantwort** 2 years ago

There is always different application requirement to use different format. That way you will just creating more and more redundant columns in different formats. That is tedious.

upvoted 5 times

✉  **YorelNation** 1 year, 5 months ago

A yes, That make a lot of sense and also if you update the table only once with UPDATE if there is a new employee it will not be up to date with the new column, if the app use a view it will be up to date every time it query.

But in any case the cost will not be minimized.

upvoted 2 times

✉  **[Removed]** 3 years, 10 months ago

Because views are not materialized, the query that defines the view is run each time the view is queried. Queries are billed according to the total amount of data in all table fields referenced directly or indirectly by the top-level query

upvoted 8 times

✉  **lgdantas** 3 years, 5 months ago

Wouldn't "total amount of data in all table fields referenced directly or indirectly by the top-level query" be FirstName and LastName?

upvoted 3 times

✉  **lollo1234** 2 years, 8 months ago

You're right, BigQuery bills on number of bytes processed, regardless of them being materialized. If you don't create a new column & use a view instead, you will probably have a small performance hit but query costs would be the same and storage cost wouldn't increase (unlike storing a new column)

upvoted 4 times

✉  **yoshik** 2 years, 4 months ago

You are asked to modify the schema and data. By using a view, the underlined table remains intact.

upvoted 10 times

✉  **HarshKothari21** 1 year, 4 months ago

good catch, yoshik.

upvoted 1 times

✉  **alecuba16** 1 year, 8 months ago

Views are cached the same as regular tables are, so I don't get the point of billing. It will cost the same as query to a regular table.

upvoted 3 times

✉  **ovokpus** 1 year, 2 months ago

the point of billing is extra storage costs for a new concatenated column

upvoted 3 times

✉  **beowulf_kat** 1 year, 2 months ago

I agree that A is correct. Also, I think B is wrong as the UPDATE statement is used to update values in existing columns, not to create a new column.

upvoted 2 times

✉  **ovokpus** 1 year, 2 months ago

Of course, you use UPDATE after creating the new column, that is what the option said

upvoted 2 times

✉  **[Removed]** 11 months, 1 week ago

What happen if there are new employees joining the company, update every single time?

upvoted 1 times

✉️  **[Removed]** 3 years, 10 months ago

Can't be A

upvoted 5 times

✉️  **BhupiSG** Highly Voted 2 years, 10 months ago

Correct: B

BigQuery has no quota on the DML statements. (Search Google - does bigquery have quota for update).

Why not C: This is a one time activity and SQL is the easiest way to program it. DataFlow is way overkill for this. You will need to find an engir who can develop DataFlow pipelines. Whereas, SQL is so much more widely known and easier. One of the great features about BigQuery is its SQL interface. Even for BigQueryML services.

upvoted 46 times

✉️  **lollo1234** 2 years, 8 months ago

I will also add that B would imply changing upstream workloads to write the new field every time a record gets added

upvoted 7 times

✉️  **DGames** 1 year, 1 month ago

But you need to maintain table means regularly you have to execute the update query whenever new data comes.

upvoted 2 times

✉️  **lollo1234** 2 years, 8 months ago

DML statements don't increase costs, but storing a new column does. I see A is correct (also see my comment above)

upvoted 4 times

✉️  **exnaniantwort** 2 years ago

Exactly. Cost is the reason to reject B.

How come so many people vote for this wrong option?

upvoted 1 times

✉️  **ler_mp** 1 year ago

Storage is cheap compared to computation

upvoted 2 times

✉️  **philli1011** Most Recent 1 day, 19 hours ago

Definitely A

upvoted 1 times

✉️  **axantroff** 2 months, 1 week ago

Selected Answer: A

The question might be outdated, but I would like to offer my perspective:

1. Ideally, I would opt for a materialized view to avoid updating pipelines
 2. In 2023, I see no concerns regarding the costs involved in storing denormalized data for analytical needs
 3. Regarding this question I would choose option A, although the concern about extra costs due to recalculations is valid for me
- upvoted 2 times

✉️  **LaxmanTiwari** 1 month, 1 week ago

Did u pass the exam ?

upvoted 1 times

✉️  **steghe** 2 months, 3 weeks ago

Answer should be A 'cos the First request is: make that data available.

upvoted 1 times

✉️  **alihabib** 5 months, 3 weeks ago

Its A "asked to change schema" is a trick to test your skills. Better to make use of MV's if anyhow the application is gonna query repeated MV's will rebuild itself, if query invalidates from cache results

upvoted 1 times

✉️  **nescafe7** 6 months ago

Selected Answer: A

In the case of B, the data pipeline that adds new employee information must also be modified, which is not the correct answer in terms of cost minimization.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: A

It's A. If you add a column to the table, you will be billed every time you query that new column. The same way you would be billed with the view created by A.

B,C and D create a new column. A does not create a new column. It just provides the interface for the application to access the data. B,C and D will have to be rerun to compute the column value of new customers.

A is done only once, costs 0 for storage, and is charged about the same as all the others when it comes to compute because even if you choose B,C and D you would have to query the data in the end anyway.

upvoted 1 times

✉  **autumn2005** 6 months, 1 week ago

Selected Answer: C

modify the schema

upvoted 1 times

✉  **theseawillclaim** 6 months, 2 weeks ago

Can you code a script for a BQ Column? I don't think it's "B", it is pretty tricky

upvoted 1 times

✉  **KC_go_reply** 7 months ago

Selected Answer: A

Everything but A) new view is wrong.

B) sounds okay, but introduces a new column which means more storage, thus increasing cost.

C) Dataflow is obvious overkill for a simple task such as concatenating two strings.

D) Starting up a Dataproc cluster just for string concatenation is super overkill.

upvoted 1 times

✉  **vaga1** 8 months, 1 week ago

Selected Answer: A

if a new field is only necessary for one project, and it is only the concatenation of two existing fields, it is ok to create a view that gets used for specific task.

upvoted 1 times

✉  **Jarek7** 9 months, 1 week ago

Selected Answer: A

I'd go for A.

The main issue with answers B,C,D is that they are just temporary solution. Whenever a new employee comes in (there are 400.000 of them at the moment, so we can expect every day a few new guys) we need to update the fullname table/field again. Additionally each of these answers need twice as much capacity (BigQuery stores data in a columnar format, so optimizing is not possible). Although the price for the needed capacity will be far below 0.01\$/month.

The main argument against A is that compute power costs more than capacity. Please look how BQ is priced:

https://cloud.google.com/bigquery/pricing#query_pricing

In the default On-demand compute pricing it is charged for "the number of bytes processed by each query" so there will be no any difference in computing costs for any option.

Yeah, there is also this argument about modifying schema in the requirements. Lets be professional - it is not a requirement for OUR schema. If you can resolve the issue with 0 change to YOUR schema then you are more than ok. And anyway, from requestor point of view, the schema he uses in his app will be modified as he needed.

upvoted 2 times

✉  **izekc** 9 months, 1 week ago

Selected Answer: A

Should be A.

upvoted 1 times

 **Adswerve** 9 months, 3 weeks ago

Selected Answer: B

I vote B.

A is expensive, the requirements say we need to minimize cost.

B works and meets the requirements. We create an empty column and then UPDATE it to set it to a desired values.

<https://cloud.google.com/bigquery/docs/managing-table-schemas#console>

https://cloud.google.com/bigquery/docs/reference/standard-sql/dml-syntax#update_statement

upvoted 2 times

 **juliosb** 10 months, 2 weeks ago

Selected Answer: A

A is the simplest.

B will not only increase the cost of storage for the *duplicated* data as it would be a pain to keep. Would require a trigger on inserts and upda

upvoted 1 times

 **mcrokz** 11 months ago

I'd choose neither, I'd advise to concatenate on the application level the IT Member is developing than changing schema and needing to worry about everything up or downstream that the schema change could impact

upvoted 3 times

 **itsmynickname** 6 months, 3 weeks ago

Exactly..

upvoted 1 times

Question #44

Topic 1

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property

'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

Indexes:

-kind: Movie

Properties:

-name: actors

name: date_released

-kind: Movie

Properties:

-name: tags

name: date_released

A. Manually configure the index in your index config as follows:

Indexes:

-kind: Movie

Properties:

-name: actors

-name: tags

-name: date_published

B. Manually configure the index in your index config as follows:

C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'

D. Set the following in your entity options: exclude_from_indexes = 'date_published'

Correct Answer: A

Community vote distribution

A (80%)

D (20%)

✉  **Wasss123**  1 year, 4 months ago

Selected Answer: A

Correct answer is A

Read in reference : https://cloud.google.com/datastore/docs/concepts/indexes#index_limits

In this case, you can circumvent the exploding index by manually configuring an index in your index configuration file:

indexes:

- kind: Task

properties:

- name: tags

- name: created

- kind: Task

properties:

- name: collaborators

- name: created

This reduces the number of entries needed to only $(|tags| * |created| + |collaborators| * |created|)$, or 6 entries instead of 9

upvoted 7 times

✉  **jkhong**  1 year, 1 month ago

Selected Answer: A

you can circumvent the exploding index by manually configuring an index in your index configuration file:

https://cloud.google.com/datastore/docs/concepts/indexes#index_limits

upvoted 1 times

✉  **Krish6488** 1 year, 1 month ago

Selected Answer: D

Tempted to go with D as the syntax in Option A seems incorrect. D is still a possible answer because one of the ways to get rid of index errors to remove the entities that are causing the index to explode. In this case its date_released and hence D appears right to me

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: A

Option B & D reject because mention date_publised in question date_released is column

Option C also not correct, I would go with option A.

upvoted 3 times

👤 **Ender_H** 1 year, 4 months ago

Selected Answer: D

Correct Answer D:

This is the way the DB is typically queried:

- movies with actor=<actorname> ordered by date_released
- movies with tag=Comedy ordered by date_released

so it seems that we need indices in actor,tag and date_released for sorting.

✗ A: this would be the correct answer, however, the format is incorrect, the correct format would be '- name: date_released' correctly indented

✗ B: This seems to be unnecessary, since typically actor and tag are not queried together. also, there is a clear indentation issue

✗ C: We don't want to ignore actor and tag, we need those indices.

✓ D: If we leave datastore to automatically create the indices and if we specify that the 'date_released' property needs to be excluded from indices, then we would have less indices (but maybe slower queries when ordering them, but hey, how many 'comedies' there could be in the world)

upvoted 2 times

👤 **Ender_H** 1 year, 4 months ago

Findings for this answer:

Indices, if not defined, will be automatically created:

"By default, a Datastore mode database automatically predefines an index for each property of each entity kind. These single property indexes are suitable for simple types of queries."

source: <https://cloud.google.com/datastore/docs/concepts/indexes>

In the index limits section we see this:

"a Datastore mode database creates an entry in a predefined index for every property of every entity except those you have explicitly declared as excluded from your indexes."

source: https://cloud.google.com/datastore/docs/concepts/indexes#index_limits

upvoted 1 times

👤 **Ender_H** 1 year, 4 months ago

And here is the correct way to configure indices:

<https://cloud.google.com/datastore/docs/tools/indexconfig>

so this would be the best answer:

indexes:

- kind: Movie

properties:

- name: actors

- name: date_released

direction: asc. <This could be left out, it defaults to direction: asc if excluded>

- kind: Movie

properties:

- name: tag

- name: date_released

direction: asc. <This could be left out, it defaults to direction: asc if excluded>

upvoted 3 times

👤 **Hm92730** 1 year, 4 months ago

What do people think about C? The question is asking how to avoid a combinatorial explosion in the number of indexes. It says "You have entities with multiple properties, some of which can take on multiple values". Put this with the below text from the documentation for Datastore indexes, it seems they're looking for "exclude the properties that will cause combinatorial explosion" which would be C.

"The situation becomes worse in the case of entities with multiple properties, each of which can take on multiple values. To accommodate such an entity, the index must include an entry for every possible combination of property values. Custom indexes that refer to multiple properties, each with multiple values, can "explode" combinatorially, requiring large numbers of entries for an entity with only a relatively small number of possible property values."^[1]

[1] https://cloud.google.com/datastore/docs/concepts/indexes#index_limits

upvoted 1 times

soichirokawa 1 year, 4 months ago

B. is correct

To avoid combinatoric explosion of indexes.

"Two queries of the same form but with different filter values use the same index."

<https://cloud.google.com/datastore/docs/concepts/indexes>

upvoted 1 times

Question #45

Topic 1

You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud

Dataflow job to process that log file. You need to make sure the log file is processed once per day as inexpensively as possible. What should you do?

- A. Change the processing job to use Google Cloud Dataproc instead.
- B. Manually start the Cloud Dataflow job each morning when you get into the office.
- C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
- D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Correct Answer: C

Community vote distribution

C (100%)

soichirokawa [Removed] Highly Voted 3 years, 10 months ago

Answer: C

upvoted 22 times

soichirokawa [Removed] Highly Voted 3 years, 10 months ago

Answer: C

Description: Scheduler for adhoc jobs – 3 jobs free and \$0.10 per job

upvoted 12 times

soichirokawa [Most Recent] 2 months, 1 week ago

Selected Answer: C

Service was renamed, but the answer is still - C

upvoted 1 times

soichirokawa imran79 3 months, 3 weeks ago

C. Using the Google App Engine Cron Service to run the Cloud Dataflow job allows you to automate the execution of the job. By creating a cron job, you can ensure that the Dataflow job is triggered exactly once per day at a specified time. This approach is automated, reliable, and fits the requirement of processing the log file once per day.

upvoted 1 times

soichirokawa itsmynickname 6 months, 3 weeks ago

C. For a modern solution, Cloud Scheduler

upvoted 5 times

soichirokawa Maurilio_Cardoso 8 months ago

Selected Answer: C

Currently, Cloud Scheduler takes over the scheduling functions.

upvoted 2 times

soichirokawa jin0 11 months, 1 week ago

I don't understand why that dataflow is used for processing? even though it should be processed once per a day?? is it more suitable for processing by using Dataproc instead?

upvoted 2 times

✉  **captainbu** 1 year ago

Selected Answer: C

C was correct but nowadays you'd schedule a Dataflow job with Cloud Scheduler: <https://cloud.google.com/community/tutorials/schedule-dataflow-jobs-with-cloud-scheduler>

upvoted 5 times

✉  **Ender_H** 1 year, 4 months ago

Selected Answer: C

Correct Answer: C.

✗ A: Dataproc is a managed Apache Spark and Apache Hadoop service, makes no sense to use it

✗ B: This might sound as the cheapest, but is highly error prone, besides, anyone in charge of this has a salary and I doubt is a low one.

✓ C: This is the easiest/fastest/cheapest way to trigger job runs, you can even set retry attempts.
source: <https://cloud.google.com/appengine/docs/flexible/nodejs/scheduling-jobs-with-cron-yaml>.

✗ D: Setting this would be much more expensive than the cron-job

upvoted 2 times

✉  **noob_master** 1 year, 7 months ago

Selected Answer: C

Answer: C

upvoted 1 times

✉  **anji007** 2 years, 3 months ago

Ans: C

upvoted 2 times

✉  **Chelseajcole** 2 years, 3 months ago

I know probably this question is testing on if you know cron.yaml and its function in App Engine. But why B will be more expensive? Human capital cost? Let's say if hiring a person click the button will be cheaper than launch an app engine, should we reconsider B?

upvoted 3 times

✉  **AmirN** 1 year, 7 months ago

Would you rather pay someone \$100,000 a year to click 'run' on jobs all day, or have them automate it and do more cutting edge work? The would be opportunity cost.

upvoted 3 times

✉  **Chelseajcole** 2 years, 4 months ago

Scheduling Jobs with cron.yaml

Free applications can have up to 20 scheduled tasks. Paid applications can have up to 250 scheduled tasks.

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'C'

upvoted 2 times

✉  **naga** 2 years, 11 months ago

Correct C

upvoted 3 times

✉  **Radhika7983** 3 years, 2 months ago

Answer is C. <https://cloud.google.com/appengine/docs/flexible/nodejs/scheduling-jobs-with-cron-yaml>

upvoted 5 times

✉  **haroldbenites** 3 years, 5 months ago

C Correct

upvoted 4 times

You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible.

What should you do?

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery
- C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
- D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Correct Answer: B

Community vote distribution

B (75%)

13%

8%

 **mmarulli** Highly Voted 3 years, 10 months ago

this is one of the sample exam questions that google has on their website. The correct answer is B
upvoted 39 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: B
Description: B is correct because regional storage is cheaper than BigQuery storage.
upvoted 12 times

 **funtoosh** 2 years, 11 months ago

it's not only cheaper but the requirement is that the data keep updating every 30 min and you need to combine the data in bigquery, use external tables to do that is the recommended practice
upvoted 9 times

 **TVH_Data_Engineer** Most Recent 1 month, 2 weeks ago

Selected Answer: A

BigQuery supports partitioned tables, which allow for efficient querying and management of large datasets that are updated frequently. By loading the updated data into a new partition every 30 minutes, you can ensure that only relevant partitions are queried, reducing the amount of data processed and thereby minimizing costs.
What's wrong with B? While creating a federated data source in BigQuery pointing to a Google Cloud Storage bucket is feasible, it might not be the most efficient for data that is updated every 30 minutes. Querying federated data sources can sometimes be more expensive and less performant than querying data stored directly in BigQuery.

upvoted 1 times

 **Melampos** 9 months, 1 week ago

Selected Answer: D

Federated queries let you send a query statement to Cloud Spanner or Cloud SQL databases not to cloud storage
upvoted 1 times

 **sid_is_dis** 7 months, 2 weeks ago

Is you are right about "federated queries", but the option B says about "federated data source". These are different concepts
upvoted 3 times

 **Abhilash_pendyala** 9 months, 2 weeks ago

ChatGPT says partitioned tables is the best approach. The answers here are quite contrasting with that answer. Even I thought it has to be option A, I am so confused now? Any proper straight forward answer?

upvoted 1 times

✉  **musumusu** 11 months, 1 week ago

Answer B:

Uploading data into staging tables/ external tables or federated source in BQ is the best approach. Option A is also good approach, anyone can explain about his part what is wrong about this?

upvoted 1 times

✉  **yoga9993** 11 months ago

we can't implement A, it's because bigquery partition table can only be done minimum in range 1 hour, the requirement said it must be updated every 30 minutes, so A is impossible option as the minimum partition is in hour level

upvoted 7 times

✉  **AzureDP900** 1 year ago

B is right

upvoted 1 times

✉  **Krish6488** 1 year, 1 month ago

Selected Answer: B

Discounting A due to limitations on partitions

Discounting C because datastore does not fit into the nature of data we are talking about and federation between BQ and datastore is an over between B and D, updating the price file on GCS and joining BQ tables and external tables sourcing data from GCS is most cost optimal way this use case

upvoted 2 times

✉  **ler_mp** 1 year ago

D is also overkill for this use case, so I'd pick B

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

Selected Answer: B

Consideration: As cheaply as possible. Make sure data stays up to date.

Initially chose A. But in actuality there is no need to maintain or store past data so storage of past data and partitioning doesn't seem like a key requirement.

Instead we can connect just to a single Cloud Storage file, either by:

- i. replace previous prices with latest prices
- ii. store previous prices in GCS if required to be retained

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

B is most inexpensive approach.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: B

The technical requirement is having frequently access info to join with other BQ data, as cheap as possible. B fits perfectly.

Corner cases for external data sources:

- Avoiding duplicate data in BigQuery storage
- Queries that do not have strong performance requirements
- Small amount of frequently changing data to join with other tables in BigQuery

<https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer>

upvoted 1 times

✉  **assU2** 1 year, 2 months ago

Selected Answer: D

I would say D, regional Google Cloud Storage bucket - cheap.

A - not cheap

B - NoSQL database for your web and mobile applications

C - Federated queries let you send a query statement to Cloud Spanner or Cloud SQL databases

And we need to combine data in DQ with data from bucket

upvoted 1 times

✉  **MisuLava** 1 year, 3 months ago

according to this :
<https://cloud.google.com/bigquery/docs/external-data-sources>
Federated queries don't work with Cloud Storage.
how can it be B ?
upvoted 2 times

✉  **cloudmon** 1 year, 2 months ago

Correct, it cannot be B because BQ federated queries only work with Cloud SQL or Spanner
upvoted 1 times

✉  **gudiking** 1 year, 2 months ago

It seems to me that they do: <https://cloud.google.com/bigquery/docs/external-data-cloud-storage>
upvoted 1 times

✉  **ducc** 1 year, 5 months ago

Selected Answer: B

I voted for B
upvoted 2 times

✉  **FrankT2L** 1 year, 7 months ago

Selected Answer: C

The average prices of these goods are updated every 30 minutes
--> No Cloud Storage
upvoted 1 times

✉  **paulino_gauna** 1 year, 7 months ago

Selected Answer: C

need to update the data
upvoted 2 times

✉  **Yad_datatonic** 1 year, 8 months ago

Selected Answer: B

Answer: B is correct because regional storage is cheaper than BigQuery storage.
upvoted 1 times

Question #47

Topic 1

You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

- ⇒ The user profile: What the user likes and doesn't like to eat
- ⇒ The user account information: Name, address, preferred meal times
- ⇒ The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

- A. BigQuery
- B. Cloud SQL
- C. Cloud Bigtable
- D. Cloud Datastore

Correct Answer: A

Community vote distribution

A (44%)

B (35%)

D (21%)

✉  **jvg637** Highly Voted 3 years, 10 months ago

You want to optimize the data schema + Machine Learning --> Bigquery. So A
upvoted 53 times

✉  **yoshik** 2 years, 4 months ago

BigQuery is a datawarehouse, not a transactional db. You need to store transactional data as a requirement.
upvoted 24 times

✉  **alecuba16** 1 year, 6 months ago

Biqury Supports transactions:
<https://cloud.google.com/bigquery/docs/reference/standard-sql/transactions>
, but indeed is not a good DB for OLTP.

But I would said or CloudSQL or BigQuery

upvoted 3 times

✉  **alexmirmao** 2 years, 3 months ago

In my opinion transactional data doesnt mean transactions they could be grouped so there is no need to write register by register.
upvoted 5 times

✉  **yoshik** 2 years, 3 months ago

In other questions they talk about 'transactional log data' when referring to past transactions, but you could be right, agree. In that c
ok A BigQuery. Nevertheless, the question is formulated ambiguously.
upvoted 5 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: Should be D - Datastore
upvoted 25 times

✉  **cetanx** 7 months, 3 weeks ago

It was also a difficult one for Chat GPT, it did give different answers each time I inquiry more about the question. After a few iterations, we
agreed on "D" :) - because;

In the context of a food ordering service, storing data about what a user likes and doesn't like to eat can potentially involve a varied and
dynamic set of data. Some users might have a long list of food preferences, while others might have only a few. Some users might update
their likes and dislikes frequently, while others rarely or never. This kind of data is a good match for a NoSQL database like Datastore, which
can easily accommodate such variations.

upvoted 5 times

✉  **GeeBeeEl** 3 years, 2 months ago

There is SQLML with BigQuery, you know that?
You cannot optimize a schema in datastore, it is a NoSQL document database built for automatic scaling, high performance, and ease of
application development. It does not work based on schemas!
upvoted 18 times

✉  **BigQuery** 2 years, 1 month ago

BQML is there. But, In question do they want to do ML on BQ?? Its saying just ML Based Company.
upvoted 6 times

✉  **philli1011** Most Recent 1 day, 17 hours ago

C

It says that the database will be used to store the transactions data. BigQuery is not usually characterized as a data storage system. Also a
databased is used for storing transactional Data not a Data Wharehouse.
upvoted 1 times

✉  **philli1011** 1 day, 17 hours ago

My choice is C
It says "DataBase Schema" not DataWarehouse Schema. It didn't mention if the ML is to be done in the DataBase or not, it just states that a
database is to be created.
upvoted 1 times

👤 **Camaro** 1 month, 2 weeks ago

I asked ChatGPT this question.
It first answered Datastore.

I said the question asks us to optimize the data schema and Datastore has no schema.

Then it answered CloudSQL.

I said the question asks about Machine learning aspect.

Then it answered Bigquery.

I said its a food ordering service and must need low latency

Then it answered Bigtable.

GPT is clearly not a good tool to use for the prep please avoid it. Its flawed currently that is DEC2023!

upvoted 3 times

👤 **LaxmanTiwari** 1 month, 1 week ago

Hello Camaro, when u going to take the exam ? am appearing on 23rd Dec and keen to know if this set is still valid ?

upvoted 1 times

👤 **rocky48** 2 months, 3 weeks ago

Selected Answer: A

A. BigQuery -> Most Probably the answer because for analytic it should automatically be big query. But my question is why not others. I've used BigQuery and I know that it allows schema optimization as well as one of its feature is built in machine learning.

More here: <https://cloud.google.com/bigquery/docs/introduction>

B. Cloud SQL -> I am not being about to find the doc for cloud sql where it talks about being able to optimize the schema and if it can be used for machine learning applications.

C. Cloud Bigtable -> Possible answer because there's schema and can be used for machine learning application.

Doc: <https://cloud.google.com/bigtable/docs/overview>

D. Cloud Datastore -> Wrong because Datastore doesn't have a schema.

Doc: <https://cloud.google.com/datastore/docs/concepts/overview>

upvoted 1 times

👤 **A_Nasser** 4 months ago

Selected Answer: D

The right answer is D because Datastore is transactional, scalable, and can deliver an output to ML.

upvoted 1 times

👤 **Mark_86** 5 months, 1 week ago

Selected Answer: B

Although you could argue with the formulation of the question, I did also read that this is about a transactional database which BigQuery is not. Thus I would go with Cloud SQL.

upvoted 1 times

👤 **alihabib** 5 months, 3 weeks ago

Initially I also thought it would be D, but then I re-read & it mentions the whole question is about Data Analytics, hence it has to be A... BigQuery

upvoted 1 times

👤 **NeoNitin** 6 months ago

A sahi hai

upvoted 1 times

👤 **nescafe7** 6 months ago

Selected Answer: B

to store all the transactional data of the product

upvoted 1 times

✉  **[Removed]**  3 years, 10 months ago

Answer: C

Description: Bigquery understands UTF-8 encoding anything other than that will result in data issues with schema
upvoted 26 times

✉  **YAS007**  2 years, 5 months ago

Answer : C :

" If you don't specify an encoding, or if you specify UTF-8 encoding when the CSV file is not UTF-8 encoded, BigQuery attempts to convert the data to UTF-8. Generally, your data will be loaded successfully, but it may not match byte-for-byte what you expect. "
https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv#details_of_loading_csv_data

upvoted 15 times

✉  **samdhimal**  1 year ago

SITUATION:

- Your company is loading comma-separated values (CSV) files into Google BigQuery.
- Data is fully imported successfully.

PROBLEM:

- Imported data is not matching byte-to-byte to the source file. Reason?

upvoted 2 times

✉  **samdhimal** 1 year ago

A. The CSV data loaded in BigQuery is not flagged as CSV.

Since BigQuery support multiple formats it could be that maybe avro or json was selected.

But the file import was successful hence csv was selected. Either manually or it was left as is since the default file type is csv. Lastly, this is WRONG.

B. The CSV data has invalid rows that were skipped on import.

-> Since the data was successfully imported there were no invalid rows. Hence, This is wrong answer too.

upvoted 2 times

✉  **samdhimal** 1 year ago

C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

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

Source: <https://cloud.google.com/bigquery/docs/loading-data>

Default BQ Encoding: UTF-8

This is probably the correct answer because if the csv file encoding was not UTF-8 and instead it was ISO-8859-1 then we would have tell bigquery that or else it will assume it is UTF-8. Hence, Imported data is not matching byte-to-byte to the source file. CORRECT ANSWER!

upvoted 2 times

✉  **samdhimal** 1 year ago

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

-> ETL means Extract, Transform and Load and this is actually very important content for Cloud Data Engineers. Look into it if interested! But getting back to the topic: ETL is usually required when the source format and target format are different. You need to extract source file and the transform it before loading the data to fit the target. This is also not a viable option. Also Data is imported successfully and the question doesn't mention anything regarding ETL.

upvoted 2 times

✉  **medeis_jar** 2 years ago

Selected Answer: C

A is not correct because if another data format other than CSV was selected then the data would not import successfully.

B is not correct because the data was fully imported meaning no rows were skipped.

C is correct because this is the only situation that would cause successful import.

D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table

upvoted 6 times

MaxNRG 2 years, 2 months ago

Selected Answer: C

C is correct because this is the only situation that would cause successful import.

A is not correct because if another data format other than CSV was selected then the data would not import successfully.

B is not correct because the data was fully imported meaning no rows were skipped.

D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table https://cloud.google.com/bigquery/docs/loading-data#loading_encoded_data

upvoted 2 times

NicolasN 1 year, 1 month ago

Exactly 

The updated link (Dec. 2022) and the quote:

 <https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv#encoding>

"If you don't specify an encoding, or if you specify UTF-8 encoding when the CSV file is not UTF-8 encoded, BigQuery attempts to convert the data to UTF-8. Generally, your data will be loaded successfully, but it may not match byte-for-byte what you expect."

upvoted 1 times

anji007 2 years, 3 months ago

Ans: C

upvoted 3 times

sumanshu 2 years, 7 months ago

Vote for 'C'

upvoted 3 times

sumanshu 2 years, 6 months ago

A is not correct because if another data format other than CSV was selected then the data would not import successfully.

B is not correct because the data was fully imported meaning no rows were skipped.

C is correct because this is the only situation that would cause successful import.

D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table.

upvoted 2 times

naga 2 years, 11 months ago

Correct C

upvoted 2 times

haroldbenites 3 years, 5 months ago

C is correct

upvoted 3 times

saurabh1805 3 years, 5 months ago

C is correct answer. Refer below link for more information.

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

upvoted 6 times

[Removed] 3 years, 10 months ago

Answer: C

upvoted 10 times

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.

Correct Answer: CE

Community vote distribution

CD (78%)

AC (17%)

4%

 **Toto2020** Highly Voted 3 years, 1 month 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. The -m option to do parallel work is recommended. Therefore A is not an answer 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 use 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 51 times

✉  **vholti** 2 years, 3 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 - 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

✉  **Mathew106** 6 months, 1 week 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 can't 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** 6 months, 1 week ago

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

upvoted 1 times

✉  **vholti** 2 years, 3 months ago

Here is the docs which support my opinion: <https://cloud.google.com/storage/docs/gsutil/addlhelp/TopLevelCommandLineOptions>

upvoted 3 times

✉  **Mathew106** 6 months, 1 week 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

✉  **musumusu** 11 months, 1 week ago

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

upvoted 1 times

✉  **Jarek7** 9 months, 1 week 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 1 times

✉  **Booqq** 11 months ago

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

upvoted 1 times

✉  **awssp12345** 2 years, 6 months ago

This should be the correct answer.

upvoted 2 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Should be AC

upvoted 33 times

👤 GeeBeeEl 3 years, 2 months ago

support this with a link....

upvoted 3 times

👤 gcppde 2 years, 11 months ago

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

upvoted 1 times

👤 BhupiSG 2 years, 10 months 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

👤 tavva_prudhvi 1 year, 9 months ago

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

upvoted 1 times

👤 spicebits Most Recent 2 months, 3 weeks 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 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 1 times

👤 Maurilio_Cardoso 7 months, 3 weeks 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 8 months, 3 weeks 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 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 1 times

👤 Oleksandr0501 8 months, 4 weeks ago

Selected Answer: CD

CD , i guess. Liked explanation in discussion.

changing from BC to CD

upvoted 1 times

👤 Kart87 9 months, 1 week 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** 9 months, 1 week 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 gsutil 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 the desired answer. And it makes sense if they want us to select 2 actions we have to make - If we go 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** 3 months, 2 weeks ago

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

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks 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** 11 months, 3 weeks 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** 11 months, 3 weeks 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** 9 months, 1 week ago

I wouldn't agree, the main issue here is latency 0.2s and 20000 files per hour - it is even beyond possible transfer without parallelization. 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 to get rid of latency, by parallelization.

upvoted 1 times

✉  **Jarek7** 9 months, 1 week ago

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

upvoted 1 times

✉  **Leeeeeee** 1 year, 2 months ago

Selected Answer: CD

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

upvoted 2 times

abhinet1313 1 year, 9 months 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 1 year, 9 months 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 1 year, 7 months ago

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

upvoted 2 times

Jojo9400 1 year, 10 months ago

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

upvoted 1 times

OmJanmeda 1 year, 10 months ago

Selected Answer: CD

CD is correct option

upvoted 1 times

Tanzu 1 year, 11 months ago

20k files * 24 hours = 480k files x2 * 4kilobyte= 3.8gb everyday and must be processed in 10 hours

Question #50

Topic 1

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100

TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID).

However, high availability and low latency are required.

You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

- A. Redis
- B. HBase
- C. MySQL
- D. MongoDB
- E. Cassandra
- F. HDFS with Hive

Correct Answer: BDF

Community vote distribution

BDE (94%)

6%

jvg637 3 years, 10 months ago

BDE. Hive is not for NoSQL

upvoted 36 times

👤 **sergio6** 2 years, 4 months ago

Redis is also NoSQL

upvoted 1 times

👤 **vholti** 2 years, 3 months ago

Redis is limited to 1 TB capacity quota per region. So it doesn't satisfy the requirement.

<https://cloud.google.com/memorystore/docs/redis/quotas>

upvoted 2 times

👤 **ckanaar** 4 months, 1 week ago

Memorystore, Google's managed Redis service is. But OS Redis is not. Though it is hard to find a 100GB RAM machine

upvoted 1 times

👤 **awssp12345** Highly Voted 2 years, 6 months ago

Answer is BDE -

A. Redis - Redis is an in-memory non-relational key-value store. Redis is a great choice for implementing a highly available in-memory cache to decrease data access latency, increase throughput, and ease the load off your relational or NoSQL database and application. Since the question does not ask cache, A is discarded.

B. HBase - Meets reqs

C. MySQL - they do not need ACID, so not needed.

D. MongoDB - Meets reqs

E. Cassandra - Apache Cassandra is an open source NoSQL distributed database trusted by thousands of companies for scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data.

F. HDFS with Hive - Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. HIVE IS NOT A DATABASE.

upvoted 31 times

👤 **[Removed]** 11 months, 1 week ago

HDFS is. Hadoop Distributed File System. HDFS is storage and HIVE is for processing.

upvoted 1 times

👤 **musumusu** Most Recent 11 months, 2 weeks ago

BDE

Faster Database are NoSQL db than SQL, Cassandra is the fastest one in market now than Hbase and then others, in given list MongoDB

upvoted 1 times

👤 **MisuLava** 1 year, 5 months ago

"Which three databases meet your requirements? "

Hive is not a database server.

HBase, Mongo and Cassandra are and meet the criteria.

BDE is the right answer

upvoted 1 times

👤 **sraakesh95** 2 years ago

Selected Answer: BDE

@hendrixlives

upvoted 1 times

👤 **medeis_jar** 2 years ago

Selected Answer: BDE

as explained by hendrixlives

upvoted 1 times

⊕  **hendrixlives** 2 years, 1 month ago

Selected Answer: BDE

BDE:

- A. Redis is a key-value store (and in many cases used as in-memory and non persistent cache). It is not designed for "100TB per year" of high available storage.
- B. HBase is similar to Google Bigtable, fits the requirements perfectly: highly available, scalable and with very low latency.
- C. MySQL is a relational DB, designed precisely for ACID transactions and not for the stated requirements. Also, growth may be an issue.
- D. MongoDB is a document-db used for high volume data and maintains currently used data in RAM, so performance is usually really good. Should also fit the requirements well.
- E. Cassandra is designed precisely for highly available massive datasets, and a fine tuned cluster may offer low latency in reads. Fits the requirements.
- F. HDFS with Hive is great for OLAP and data-warehouse scenarios, allowing to solve map-reduce problems using an SQL subset, but the latency is usually really high (we may talk about seconds, not milliseconds, when obtaining results), so this does not complies with the requirements.

upvoted 13 times

⊕  **MaxNRG** 2 years, 2 months ago

Selected Answer: BEF

Very strange question, seems outdated and irrelevant to me as it doesn't contain any GCP products :)

Anyway, I would choose BEF.

Redis is in-memory key value, not good

HBase yes, excellent case for linear growth and a column-oriented database

mysql not good, too big and no need for transactionality

Mongodb, document db with flexible schema ??

Yes Cassandra, good use case

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis.

https://www.wikiwand.com/en/Apache_Hive

upvoted 1 times

⊕  **hendrixlives** 2 years, 1 month ago

Latency in Hive is usually quite high, and one of the requirements is "low latency"

upvoted 2 times

⊕  **MaxNRG** 2 years ago

agreed on BDE

upvoted 2 times

⊕  **MaxNRG** 2 years ago

good point!

upvoted 1 times

⊕  **anji007** 2 years, 3 months ago

Ans: B, D and E

upvoted 2 times

⊕  **sumanshu** 2 years, 7 months ago

vote for BDE

upvoted 2 times

⊕  **BhupiSG** 2 years, 10 months ago

BEF

B: HBASE is based upon BigTable

E: Cassandra is low latency columnar distributed database like BigTable

F: HDFS is low latency distributed file system and Hive will help with running the queries

upvoted 2 times

⊕  **Manue** 2 years, 9 months ago

Hive is not for low latency queries. It is for analytics.

upvoted 5 times

⊕  **daghayeghi** 2 years, 10 months ago

BDE:

These are NoSQL DB, Hive is not for NoSQL.

upvoted 2 times

✉  **Rayleigh** 2 years, 11 months ago

The answer is ADE, the statement says they require a NoSQL with high availability and low latency, they do not require consistency.
C. it is not NoSQL.
F. it is not NoSQL.
B. it is NoSQL but focused on strong consistency and based on HDFS, you need HDFS for Hbase.
Therefore the answer is ADE

upvoted 1 times

✉  **daghayeghi** 2 years, 11 months ago

BDF:
Redis and Cassandra have only Rowkey and couldn't be indexed, and MySQL isn't NoSQL, Then B D and E is correct answer.

upvoted 1 times

✉  **naga** 2 years, 11 months ago

Correct BDE

upvoted 3 times

✉  **apnu** 3 years, 1 month ago

it should be BDE because Hive is a sql based datawarehouse , it is not a nosql DB

upvoted 3 times

✉  **GeeBeeEl** 3 years, 2 months ago

I agree with HaroldBenites and I like your answer. I did some research and yes, you cannot query HBase by individual fields. See <https://opensource.com/article/19/8/apache-hive-vs-apache-hbase> you cannot query by row 00001 or 00002 etc but not by the field!!! wow

upvoted 2 times

Question #51

Topic 1

You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

- A. Get more training examples
- B. Reduce the number of training examples
- C. Use a smaller set of features
- D. Use a larger set of features
- E. Increase the regularization parameters
- F. Decrease the regularization parameters

Correct Answer: ADF

Community vote distribution

ACE (94%)

6%

✉  **madhu1171** Highly Voted 3 years, 10 months ago

it should be ACE

upvoted 66 times

✉  **[Removed]**  3 years, 10 months ago

Should be ACE

upvoted 19 times

✉  **[Removed]** 3 years, 10 months ago

prevent overfitting: less variables, regularisation, early ending on the training

upvoted 14 times

✉  **TVH_Data_Engineer**  1 month, 2 weeks ago

Selected Answer: ACE

To address the problem of overfitting in training a spam classifier, you should consider the following three actions:

A. Get more training examples:

Why: More training examples can help the model generalize better to unseen data. A larger dataset typically reduces the chance of overfitting the model has more varied examples to learn from.

C. Use a smaller set of features:

Why: Reducing the number of features can help prevent the model from learning noise in the data. Overfitting often occurs when the model is complex for the amount of data available, and having too many features can contribute to this complexity.

E. Increase the regularization parameters:

Why: Regularization techniques (like L1 or L2 regularization) add a penalty to the model for complexity. Increasing the regularization parameter will strengthen this penalty, encouraging the model to be simpler and thus reducing overfitting.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: ACE

100% ACE

We need more data because less data induces overfitting. We need less features to make the problem simpler to learn and not promote learning a very complex function for thousands of features that might not apply to the test data. We also need to use regularization to keep the weights constrained.

upvoted 2 times

✉  **theseawillclaim** 6 months, 2 weeks ago

Selected Answer: ACE

Definitely ACE.

More training data and less variables can prevent the model from being too picky or specific.

upvoted 1 times

✉  **jin0** 11 months, 1 week ago

? why A is answer? even though 'more training example' not 'more dataset example'. I understand that there is dataset same and there is only change the size of training examples size. in this case there are valid and test example should be reduced. isn't it?

upvoted 1 times

✉  **desertlotus1211** 1 year ago

Collect more training data: This will help the model generalize better and reduce overfitting.

Use regularization techniques: Techniques such as L1 and L2 regularization can be applied to the model's weights to prevent them from becoming too large and causing overfitting.

Use early stopping: This involves monitoring the performance of the model on a validation set during training, and stopping the training when performance on the validation set starts to degrade. This helps to prevent the model from becoming too complex and overfitting the training data.

upvoted 1 times

✉  **desertlotus1211** 1 year ago

Regularization is a technique that penalizes the coefficient. In an overfit model, the coefficients are generally inflated. Thus, Regularization adds penalties to the parameters and avoids them weighing heavily.

A & C are correct... the third one --- not sure on

upvoted 1 times

👤 **RoshanAshraf** 1 year ago

Selected Answer: ACE

A -The training data is causing the overfitting for the testing data, so addition of training data will solve this.
C - Larger sets will cause overfitting, so we have to use smaller sets or reduce features
E - Increase the regularization is a method for solving the Overfitting model

upvoted 1 times

👤 **AzureDP900** 1 year ago

Answers are;

A. Get more training examples
C. Use a smaller set of features
E. Increase the regularization parameters

Prevent overfitting: less variables, regularisation, early ending on the training

Reference:

<https://cloud.google.com/bigquery-ml/docs/preventing-overfitting>

upvoted 3 times

👤 **DGames** 1 year, 1 month ago

Selected Answer: ADE

Answer ADE

upvoted 1 times

👤 **MisuLava** 1 year, 3 months ago

Selected Answer: ACE

100% sure ACE

<https://elitedatascience.com/overfitting-in-machine-learning>

upvoted 1 times

👤 **MisuLava** 1 year, 5 months ago

Answer is : ACE

<https://www.ibm.com/cloud/learn/overfitting#:~:text=Overfitting%20is%20a%20concept%20in,unseen%20data%2C%20defeating%20its%20purpose.>

upvoted 1 times

👤 **Noahz110** 1 year, 5 months ago

Selected Answer: ACE

im vote for ACE

upvoted 1 times

👤 **Dip1994** 1 year, 5 months ago

It should be ACE

upvoted 1 times

👤 **sraakesh95** 2 years ago

Selected Answer: ACE

@medeis_jar

upvoted 1 times

👤 **medeis_jar** 2 years ago

Selected Answer: ACE

As MaxNRG wrote:

The tools to prevent overfitting: less variables, regularization, early ending on the training.

- Adding more training data will increase the complexity of the training set and help with the variance problem.
- Reducing the feature set will ameliorate the overfitting and help with the variance problem.
- Increasing the regularization parameter will reduce overfitting and help with the variance problem.

upvoted 4 times

👤 **MaxNRG** 2 years, 1 month ago

Selected Answer: ACE

ACE

The tools to prevent overfitting: less variables, regularization, early ending on the training...

Overfitting means that the classifier knows too well the data and fails to generalize. We should use a smaller number of features to help the classifier generalize, and more examples so that it can have more variety.

The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set.

- Adding more training data will increase the complexity of the training set and help with the variance problem.

- Reducing the feature set will ameliorate the overfitting and help with the variance problem.

- Increasing the regularization parameter will reduce overfitting and help with the variance problem.

https://github.com/mGalarnyk/datasciencecoursera/blob/master/Stanford_Machine_Learning/Week6/AdviceQuiz.md

upvoted 4 times

Question #52

Topic 1

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud

Dataproc cluster, and depositing the results into Google BigQuery.

How should you securely run this workload?

- A. Restrict the Google Cloud Storage bucket so only you can see the files
- B. Grant the Project Owner role to a service account, and run the job with it
- C. Use a service account with the ability to read the batch files and to write to BigQuery
- D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Correct Answer: **B**

Community vote distribution

C (78%)

B (22%)

👤 **digvijay**  3 years, 10 months ago

A is wrong, if only I can see the bucket no automation is possible, besides, also needs launch the dataproc job

B is too much, does not follow the security best practices

C has one point missing...you need to submit dataproc jobs.

In D viewer role will not be able to submit dataproc jobs, the rest is ok

Thus....the only one that would work is B! BUT this service account has too many permissions. Should have dataproc editor, write big query & read from bucket

upvoted 33 times

👤 **dambilwa** 3 years, 7 months ago

Hence - Contextually, Option [C] looks to be the right fit

upvoted 15 times

👤 **retep007** 2 years, 4 months ago

C doesn't need permission to submit dataproc jobs, it's workload SA. Job can be submitted by any other identity

upvoted 5 times

👤 **rickywck** Highly Voted 3 years, 10 months ago

Should be C
upvoted 31 times

👤 **Mathew106** Most Recent 6 months, 1 week ago

Selected Answer: B

We need permissions for submitting dataproc jobs and writing to BigQuery. Project Owner will fix all of that even though it's not a good solution.
The rest won't work at all.

upvoted 1 times

👤 **Adswerve** 9 months, 3 weeks ago

Selected Answer: C

C
Project Owner is too much, violates the principle of least privilege
upvoted 3 times

👤 **PolyMoe** 1 year ago

Selected Answer: C

C. Use a service account with the ability to read the batch files and to write to BigQuery

It is best practice to use service accounts with the least privilege necessary to perform a specific task when automating jobs. In this case, the needs to read the batch files from Cloud Storage and write the results to BigQuery. Therefore, you should create a service account with the ability to read from the Cloud Storage bucket and write to BigQuery, and use that service account to run the job.

upvoted 3 times

👤 **Mkumar43** 1 year ago

Selected Answer: B

B works for the given requirement
upvoted 1 times

👤 **Krish6488** 1 year, 1 month ago

Least privilege principle. Option C. job can be submitted or triggered using a Cron or a composer which uses another SA with different set of privileges
upvoted 2 times

👤 **DGames** 1 year, 1 month ago

Selected Answer: B

B because we need to run job .. option C mentioned permission about read and write nothing mention to run the job . In case project owner to run service account it's similar just running job and doing rest of tasks read and writing as well.
upvoted 2 times

👤 **ThomasChoy** 1 year, 9 months ago

Selected Answer: C

The answer is C because Service Account is the best way to access the BigQuery API if your application can run jobs associated with service credentials rather than an end-user's credentials, such as a batch processing pipeline.
<https://cloud.google.com/bigquery/docs/authentication>
upvoted 2 times

👤 **Bhawantha** 2 years ago

Selected Answer: C

Data owners can't create jobs or queries. -> B out
We need service Account -> D out
Access only granting me does not solve the problem -> A out
The answer is C. (Minimum rights to perform the job)
upvoted 4 times

👤 **medeis_jar** 2 years ago

Selected Answer: C

"taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery"
upvoted 1 times

- 👤 **prasanna77** 2 years, 1 month ago
C should be okay, since he is already a project owner, I guess compute service account created will have access to run the jobs
upvoted 1 times
- 👤 **MaxNRG** 2 years, 1 month ago
Selected Answer: C
C,
Project Owner role to a service account - is too much
upvoted 1 times
- 👤 **JG123** 2 years, 2 months ago
Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?
Ans: C
upvoted 5 times
- 👤 **anji007** 2 years, 3 months ago
Ans: C
See this: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2
upvoted 3 times
- 👤 **Blobby** 2 years, 4 months ago
C as service account invoked to read the data into GCS and write to BQ once transformed via Data Proc. Assumes DataProc can inherit SA authorisation to perform transform and propagate.
B seems to violate key IAM principle enforcing least privilege;
<https://cloud.google.com/iam/docs/recommender-overview>
upvoted 4 times
- 👤 **sumanshu** 2 years, 7 months ago
Vote for 'C'
upvoted 4 times
- 👤 **sumanshu** 2 years, 6 months ago
Vote for B, (though it's too much access) - But C has one accessing missing (i.e Dataproc job execution) Thus B is correct
upvoted 3 times

Question #53

Topic 1

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:

`SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country`

You check the query plan for the query and see the following output in the Read section of Stage:1:

What is the most likely cause of the delay for this query?

- A. Users are running too many concurrent queries in the system
- B. The [myproject:mydataset.mytable] table has too many partitions
- C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
- D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Correct Answer: A

Community vote distribution

D (88%)

12%

✉  **[Removed]**  3 years, 10 months ago

Should be D
upvoted 26 times

✉  **itche_scratche**  3 years, 9 months ago

D; Purple is reading, Blue is writing. so majority is reading.
upvoted 24 times

✉  **squishy_fishy** 2 years, 3 months ago

I have been looking for the color code descriptions for a while. Thank you!
upvoted 1 times

✉  **MaxNRG**  1 month, 2 weeks ago

Selected Answer: D

The most likely cause of the delay for this query is option D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew.

Group by queries in BigQuery can run slowly when there is significant data skew on the grouped columns. Since the query is grouping by country, if most rows have the same country value, all that data will need to be shuffled to a single reducer to perform the aggregation. This can cause a data skew slowdown.

Options A and B might cause general slowness but are unlikely to affect this specific grouping query. Option C could also cause some slowness but not to the degree that heavy data skew on the grouped column could. So D is the most likely root cause. Optimizing the data distribution to reduce skew on the grouped column would likely speed up this query.

upvoted 1 times

✉  **JOKKUNO** 1 month, 2 weeks ago

Data skew is when one or some partitions have significantly more data compared to other partitions. Data-skew is usually the result of operations that require re-partitioning the data, mostly join and grouping (GroupBy) operations. So D.

upvoted 1 times

✉  **PolyMoe** 1 year ago

Selected Answer: D

D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Data skew occurs when one or more values in a column have a disproportionately large number of rows compared to other values in that column. This can cause performance issues when running queries that group by that column, like the one in the question. In this case, if most of the rows in the [myproject:mydataset.mytable] table have the same value in the country column, then the query will need to process a large number of rows with that value, which can cause significant delay.

upvoted 4 times

✉  **AzureDP900** 1 year ago

D is right
upvoted 2 times

✉  **Krish6488** 1 year, 1 month ago

Selected Answer: D

data skewing causing imbalance in data distribution across slots. It also causes errors if the group by column has NULLS. Since option C does not call out the Group by column, D is a closer answer contextually

upvoted 1 times

✉  **Jasar** 1 year, 2 months ago

Selected Answer: A

A is the best option because the color bar show the high number of reads and i think its not a skew because bigquery was build to compute things fast

upvoted 2 times

✉  **arpitagrawal** 1 year, 4 months ago

The query would throw the error because you're using a group by clause on country but not aggregating city or state.
upvoted 11 times

✉  **MisuLava** 1 year, 5 months ago

Selected Answer: D

<https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

upvoted 1 times

✉  **Paul_Oprea** 1 year, 9 months ago

BTW, how is the query even syntactically valid? It has non aggregated columns in the SELECT part of the query. That query will not run in the place, unless I'm missing something.

upvoted 17 times

✉  **Arkon88** 1 year, 11 months ago

Selected Answer: D

D

Image says that average(dark) and maximum(light) have difference in few times, this it is a skew

<https://cloud.google.com/bigquery/query-plan-explanation>

The color indicators show the relative timings for all steps across all stages. For example, the COMPUTE step of Stage 00 shows a bar whose shaded fraction is 21/30 since 30ms is the maximum time spent in a single step of any stage. The parallel input information shows that each stage required only a single worker, so there's no variance between average and slowest timings.

upvoted 1 times

✉  **sraakesh95** 2 years ago

Selected Answer: D

<https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d>

upvoted 3 times

✉  **sraakesh95** 2 years ago

<https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d>

upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: D

D

Colors: Purple is reading, Blue is writing. so the majority is reading.

<https://cloud.google.com/bigquery/docs/best-practices-performance-patterns>

upvoted 4 times

✉  **morpho444** 2 years, 2 months ago

If you read this <https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d> C can't be right because the skewness happen when the column you use for grouping contains lots of null values, here C mentions columns that aren't part of the grouping clause.

D, that's not how data get skewed, it gets skewed due to null values.

A is the only answer here.

upvoted 1 times

✉  **BigQuery** 2 years, 1 month ago

A Can't be answer. Since users whenever running queries facing the problems.

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: D

upvoted 5 times

- A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
- B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
- C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
- D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

Correct Answer: C

Community vote distribution

B (52%)	D (46%)	2%
---------	---------	----

✉  **jvg637**  3 years, 10 months ago

I'd go with B: real-time is requested, and the only scenario for real time (in the 4 presented) is the use of pub/sub with push.
upvoted 60 times

✉  **[Removed]** 2 years, 9 months ago

i would go with option B, Cause option D states "Give the bid for each item to the user in the bid event that is processed first" . The requirement is to get the first bid based on event time not processed first in dataflow.
upvoted 25 times

✉  **ralf_cc** 2 years, 7 months ago

Yep, Pub/Sub doesn't have FIFO yet, B is the one that keeps the right order
upvoted 7 times

✉  **Tanzu** 1 year, 11 months ago

it is not a queue, and that is not a issue :)
upvoted 3 times

✉  **Tanzu** 1 year, 11 months ago

in a distributed environment, you can not handle this problem wit a queue by the way !
upvoted 3 times

✉  **donbigi** 11 months, 4 weeks ago

This approach is not ideal because it requires a custom endpoint to write the bid event information into Cloud SQL. This adds additional complexity and potential points of failure to the architecture, as well as adding latency to the processing of bid events, since the data must be written to both Pub/Sub and Cloud SQL. Additionally, it can be more challenging to ensure that bid events are processed in the order they were received, since the data is being written to multiple databases. Finally, using a single database to store bid events could limit scalability and availability, and can also result in slow query performance.

upvoted 3 times

✉  **AzureDP900** 1 year ago

Agree with B
upvoted 1 times

👤 **Ganshank** Highly Voted 3 years, 9 months ago

D

The need is to collate the messages in real-time. We need to de-dupe the messages based on timestamp of when the event occurred. This can be done by publishing to Pub/Sub and consuming via Dataflow.

upvoted 34 times

👤 **unnamed12355** 10 months, 2 weeks ago

D isn't correct, Pub/Sub can send messages out of order, it is no guarantee that the event with lowest timestamp will be processed first. B is correct

upvoted 3 times

👤 **Tanzu** 1 year, 11 months ago

Yeap, that's why B is the right one. It has pub/sub push, more real time than pub/sub pull. You need to aware at some point, something has to be pulled which adds a latency.

upvoted 1 times

👤 **arpana_naa** Most Recent 1 month ago

Selected Answer: D

pub/sub for entry time stamp + event time
dataflow for processing and dataflow is better for real time

upvoted 1 times

👤 **Nandababy** 1 month, 3 weeks ago

To accurately determine who bid first in a globally distributed auction application, utilizing a push mechanism instead of a pull mechanism is generally considered the more reliable approach.

B should be correct answer.

upvoted 1 times

👤 **Zepopo** 2 months, 1 week ago

Selected Answer: B

key words is "single location in real time"

upvoted 1 times

👤 **rocky48** 2 months, 2 weeks ago

Selected Answer: D

Answer : D

We need to de-dupe the messages based on timestamp of when the event occurred. This can be done by publishing to Pub/Sub and consuming via Dataflow.

D sounds like a complete answer. B does not.

upvoted 2 times

👤 **Nivea007** 3 months, 3 weeks ago

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud.

This approach leverages Google Cloud Pub/Sub for real-time data ingestion and Google Cloud Dataflow for real-time data processing, ensuring that bids are processed as they occur, which aligns with real-time requirements.

It's not B because there is a step involving a custom endpoint that writes data into Cloud SQL. This additional step could introduce some latency and it's important to ensure that the custom endpoint and Cloud SQL database can handle the real-time load effectively.

upvoted 1 times

👤 **patiwwb** 3 months, 2 weeks ago

But D treats the bids according to the processed time. We need to consider event time that's why B is the right answer.

upvoted 1 times

👤 **imran79** 3 months, 3 weeks ago

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

upvoted 1 times

👤 **Nirca** 3 months, 4 weeks ago

Selected Answer: B

B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL. is correct

upvoted 2 times

✉  **DeepakVenkatachalam** 4 months, 1 week ago

Correct Answer is B. option D is based on processing first and not based on event first. so option D cannot be right answer
upvoted 1 times

✉  **np717** 5 months ago

Selected Answer: D

D is the best solution because it is both real-time and scalable. Google Cloud Dataflow can process the bid events in the order in which they occurred and give the bid for each item to the user in the bid event that is processed first.

upvoted 1 times

✉  **NeoNitin** 5 months, 4 weeks ago

B.here is why

Option B is like using special messaging balloons that quickly carry all the bids to a special spot. From there, a super-fast friend checks them tells us who bid first. This way, we find out quickly!

Option D is like having all the bids sent to a special magic box that quickly sends them to a smart computer friend. This friend looks at the bid right away and tells us who should get the toy based on who bid first.

In conclusion, Option B (Cloud Pub/Sub and Cloud SQL) or Option D (Google Cloud Pub/Sub and Cloud Dataflow) are the most suitable choice for real-time processing of bids and determining the first bidder. They offer efficient, scalable, and real-time solutions for handling bid events in a globally distributed auction application. The final decision would depend on factors such as the specific requirements, infrastructure, and expertise of the development team.

upvoted 1 times

✉  **NeoNitin** 5 months, 4 weeks ago

Option B (Cloud Pub/Sub and Cloud SQL):

Advantage: Cloud Pub/Sub can handle real-time data streaming, making it a good choice for quickly receiving bids. Storing bid events in Cloud SQL ensures they are easily accessible and can be analyzed in real-time.

Disadvantage: It might require some additional setup and configuration to connect Cloud Pub/Sub to a custom endpoint and Cloud SQL.

Option D (Google Cloud Pub/Sub and Cloud Dataflow):

Advantage: Cloud Pub/Sub can handle real-time data streaming, similar to Option B. Using Cloud Dataflow can process the bid events quickly and efficiently.

Disadvantage: It might require some additional setup and configuration for Cloud Dataflow.

upvoted 1 times

✉  **knith66** 6 months, 1 week ago

B is correct, As push provides near-real time

upvoted 1 times

✉  **KC_go_reply** 7 months ago

A tough call between B and D. What favors B is that push is the right pattern for a real-time scenario. Also, the timestamp is already contained in the bid event. D gives the bid to the bid event that is processed first. This is not chronologically correct, as Dataflow has not guarantee to process the events in order. Overall, B makes more sense, as it respects the given timestamp and the processing itself will have lower latency

upvoted 1 times

✉  **cchen8181** 8 months, 2 weeks ago

Selected Answer: B

I would choose B.

D is not correct due to giving the bid to the first event processed, there is no guarantee the bids will be processed in the right order when using Pub/Sub.

B allows storing to a central location as requested, and Cloud SQL is a transactional database that can handle the kind of real time OLTP processing described here.

upvoted 1 times

✉  Oleksandr0501 9 months, 1 week ago

Selected Answer: D

Option D is the correct solution.

Having each application server write the bid events to Google Cloud Pub/Sub as they occur allows for real-time collation of the bid events into a single location. Using a pull subscription to pull the bid events using Google Cloud Dataflow ensures that the events are processed in a single order and that the user who bid first is identified. This approach is scalable and provides low-latency processing, making it suitable for a global distributed auction application.

Option B also suggests using Cloud Pub/Sub, but it then pushes the events to a custom endpoint that writes to Cloud SQL. This approach can

Question #55

Topic 1

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

- A. Create a new view over events using standard SQL
- B. Create a new partitioned table using a standard SQL query
- C. Create a new view over events_partitioned using standard SQL
- D. Create a service account for the ODBC connection to use for authentication
- E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

Correct Answer: AE

Community vote distribution

CD (53%)

DE (21%)

CE (21%)

5%

✉  jvg637 Highly Voted 3 years, 10 months ago

C = A standard SQL query cannot reference a view defined using legacy SQL syntax.

D = For the ODBC drivers is needed a service account which will get a standard Bigquery role.

upvoted 50 times

✉  [Removed] Highly Voted 3 years, 10 months ago

Answer: CD

upvoted 9 times

✉  Vullibabu Most Recent 3 weeks, 2 days ago

I think question should be rewrite slightly like which 3 actions should you take rather than 2 ..

Then answer would be A,D and E..No ambiguity then

upvoted 1 times

✉  task_7 3 weeks, 4 days ago

Selected Answer: BD

ODBC connections require standard SQL, not legacy SQL.

Service account for the ODBC connection

upvoted 1 times

✉  Bahubali1988 4 months ago

This dump is full of wrong answers - not sure which one to go for.

upvoted 1 times

✉  alihabib 5 months, 3 weeks ago

CD..... C because, ODBC drivers don't support switch b/w legacy SQL & google SQL, hence better to create a new view from recent partitioned table & D as Google best practice for role binding

upvoted 1 times

✉  **baht** 7 months, 2 weeks ago

the answer is C & D

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

answer: A & D

Confusion here: Legacy SQL vs Standard, BQ supports legacy SQL but ODBC or Most RDBMS connection doesn't support Legacy SQL, so in this case we need to create a new view on existing view or replace the existing one by changing syntax.

For ODBC, you just need a service account to authenticate as its external service connection. Option E is not necessary.

upvoted 2 times

✉  **musumusu** 11 months, 1 week ago

Go for B, create a new view from the table, If you modify the syntax in option A, its also mean you created a new view on table :P

upvoted 1 times

✉  **PolyMoe** 1 year ago

Selected Answer: DE

D. Create a service account for the ODBC connection to use for authentication. This service account will be used to authenticate the ODBC connection, and will be granted specific permissions to access the BigQuery resources.

E. Create a Cloud IAM role for the ODBC connection and shared events. This role will be used to grant permissions to the service account created in step D, and will allow the applications to access the events view in BigQuery.

Creating a new view over events using standard SQL may also be beneficial to improve performance and compatibility with the applications, it is not required for the ODBC connection to work.

upvoted 4 times

✉  **samdhimal** 1 year ago

INFO:

- The majority of the data analyzed is placed in a time-partitioned table named events_partitioned.
- To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data.
- The view is described in legacy SQL.

QUESTION:

Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

-> First and foremost we need to understand the information. So our actual data is stored in events_partitioned table. The organization is currently using view called events to reduce the cost.

-> Since the view called events only has last 14 days of data we cannot use that view.

-> We also cannot use that view because standard SQL is not used to describe the view. In order to connect ODBC we need a view described by standard SQL.

upvoted 3 times

✉  **samdhimal** 1 year ago

A. Create a new view over events using standard SQL

-> Wrong, events view contains only last 14 days of data and also it uses Legacy SQL.

B. Create a new partitioned table using a standard SQL query

-> Partitioned Table is not helpful in this situation. Hence, I am ruling it out.

C. Create a new view over events_partitioned using standard SQL

-> Correct this is exactly what we need.

1. We need to create a new view over events_partitioned.

2. We need to use Standard SQL.

This is a valid option.

D. Create a service account for the ODBC connection to use for authentication.

-> Correct answer because we are required to authenticate before ODBC connection.

E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared events

-> This option is of no use in this scenario

upvoted 1 times

✉  **GCPpro** 1 year ago

CE is the correct answer

upvoted 1 times

✉  **MisuLava** 1 year, 5 months ago

Selected Answer: CD

needed a service account for ODBC drivers
standard SQL vs legacy SQL.

upvoted 1 times

✉  **Smaks** 1 year, 6 months ago

Selected Answer: CE

1. Create Services account from IAM & Admin
2. Add Services account permission Roles as "BigQuery Admin" or any custom Role.
Other options are not related ' to ensure the applications can connect'

upvoted 4 times

✉  **Smaks** 1 year, 6 months ago

typo - D; E

upvoted 4 times

✉  **Arkon88** 1 year, 11 months ago

Selected Answer: CD

As stated by jvg637

C = A standard SQL query cannot reference a view defined using legacy SQL syntax.
D = For the ODBC drivers is needed a service account which will get a standard Bigquery role.
upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: CD

As stated by jvg637

C = A standard SQL query cannot reference a view defined using legacy SQL syntax.
D = For the ODBC drivers is needed a service account which will get a standard Bigquery role.
upvoted 3 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: CD

A standard SQL query cannot reference a view defined using legacy SQL syntax. In order to connect through ODBC connection, we need to use standard SQL. So, we need to create a new view over events_partitioned table using standard SQL which is C. Need service account to connect through ODBC which is option D. Check the links below.

I am not sure about A whether we can create a view over another view which was built using legacy SQL

<https://cloud.google.com/bigquery/docs/views>

<https://cloud.google.com/community/tutorials/bigquery-from-excel>

https://www.simba.com/products/BigQuery/doc/ODBC_InstallGuide/mac/content/odbc/bq/configuring/authenticating/serviceaccount.htm

upvoted 5 times

✉  **MaxNRG** 2 years, 1 month ago

It has to be standard because of this:

Google has collaborated with Magnitude Simba to provide ODBC and JDBC drivers that leverage the power of BigQuery's standard SQL. On what should we build the view, on the events_partitioned, just like the view you had before but in standard SQL. no sense in creating a new partitioned table as B says.

To let it access the data you should access with a service account.

You can configure the driver to authenticate the connection with a Google service account. When you authenticate your connection this way, the driver handles authentication on behalf of the service account, so that an individual user account is not directly involved and no user input is required.

So I think is C and D.

upvoted 2 times

✉  **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?
Ans: C,D

upvoted 8 times

Question #56

Topic 1

You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

- A. Use the TABLE_DATE_RANGE function
- B. Use the WHERE_PARTITIONTIME pseudo column
- C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
- D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD)

Correct Answer: A

Reference:

<https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understand-your-mobile-app?hl=am>

Building complex queries

What if we want to run a query across both platforms of our app over a specific date range? Since Firebase Analytics data is split into tables for each day, we can do this using BigQuery's **TABLE_DATE_RANGE** function. This query returns a count of the cities users are coming from over a one week period:

```
01  SELECT
02      user_dim.geo_info.city,
03      COUNT(user_dim.geo_info.city) as city_count
04  FROM
05  TABLE_DATE_RANGE([firebase-analytics-sample-data:analytics_*],
06  TABLE_DATE_RANGE([firebase-analytics-sample-data:ios_*],
07  GROUP BY
08      user_dim.geo_info.city
09  ORDER BY
10      city_count DESC
```

Community vote distribution

A (100%)

damaldon Highly Voted 1 year, 4 months ago

A. is correct according to this link:
<https://cloud.google.com/bigquery/docs/reference/legacy-sql>
upvoted 9 times

Ajose0 Most Recent 3 months, 2 weeks ago

Selected Answer: A

The recommended action is to use the TABLE_DATE_RANGE function (option A). This function allows you to specify a range of dates to query across multiple tables.

upvoted 1 times

Nirca 3 months, 4 weeks ago

Selected Answer: A

The TABLE_DATE_RANGE function in BigQuery is a table wildcard function that can be used to query a range of daily tables. It takes two arguments: a table prefix and a date range. The table prefix is the beginning of the table names, and the date range is the start and end dates the tables to be queried.

The TABLE_DATE_RANGE function will expand to cover all tables in the dataset that match the table prefix and are within the date range. For example, if you have a dataset that contains daily tables named my_table_20230804, my_table_20230805, and my_table_20230806, you could use the TABLE_DATE_RANGE function to query all of the tables in the dataset between August 4, 2023 and August 6, 2023 as follows:

SELECT *

Question #57

Topic 1

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

- A. They have not assigned the timestamp, which causes the job to fail
- B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
- C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
- D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Correct Answer: C

Community vote distribution

D (77%)

A (15%)

8%

[Removed] Highly Voted 3 years, 10 months ago

Answer: D

Description: Caution: Beam's default windowing behavior is to assign all elements of a PCollection to a single, global window and discard late data, even for unbounded PCollections. Before you use a grouping transform such as GroupByKey on an unbounded PCollection, you must do at least one of the following:

—>>>> Set a non-global windowing function. See Setting your PCollection's windowing function.

Set a non-default trigger. This allows the global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur.

—>>>> If you don't set a non-global windowing function or a non-default trigger for your unbounded PCollection and subsequently use a grouping transform such as GroupByKey or Combine, your pipeline will generate an error upon construction and your job will fail.

So it looks like D

upvoted 64 times

✉  **samdhimal** 1 year ago

Why not C?

Because I think that the most likely cause of the problem is C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created.

In Dataflow, windowing is used to divide the input data into smaller time intervals, called windows. Without a windowing function, all the data may be treated as part of the same window and the pipeline may not be able to process the data correctly. In this specific scenario, the engineers are trying to use windowing and transformation in Google Cloud Dataflow to periodically identify the inputs and their timings during the campaign, so it's likely that they need to use a windowing function to divide the data into smaller time intervals in order to process it correctly. Not applying a windowing function, or applying the wrong one can cause the job to fail.

Someone Clarify? Am I missing an important point?

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

You are missing that the global window is the default window that we typically use for batch processing. The global window by default waits until all data is available before processing it so if you want to use it with streaming you need to set some custom trigger so that you don't wait indefinitely to wait until we aggregate. All in all it doesn't sound right.

<https://www.youtube.com/watch?v=oJ-LueBvOcM>

<https://www.youtube.com/watch?v=MuFA6CSti6M>

upvoted 2 times

✉  **jvg637** Highly Voted 3 years, 10 months ago

Global windowing is the default behavior, so I don't think C is right.

An error can occur if a non-global window or a non-default trigger is not set.

I would say D.

(<https://beam.apache.org/documentation/programming-guide/#windowing>)

upvoted 14 times

✉  **MikkelRev** Most Recent 4 months ago

option B: They have not set the triggers to accommodate the data coming in late, which causes the job to fail.

In a streaming data processing pipeline, it's common to encounter data that arrives late, meaning it arrives after the event time has passed for its associated window. If you don't handle late data appropriately by setting triggers, it can cause issues in your pipeline, including job failures.

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

Selected Answer: A

gpt: The most likely cause of the problem is A, that they have not assigned the timestamp.

In streaming data processing, timestamps are essential for proper windowing and triggering of data. Without timestamps, the system cannot correctly determine which window a particular piece of data belongs to, or when it is safe to trigger processing of a window. If the engineers do not assign timestamps to the data, the Cloud Dataflow job would not be able to process the data correctly, and it would fail.

Option B, not setting triggers to accommodate late data, is also an important consideration for streaming data processing. However, it is less likely to cause the job to fail outright than missing timestamps.

Option C, not applying a global windowing function, and Option D, not applying a non-global windowing function, are also important considerations for windowing in Cloud Dataflow. However, neither of these would cause the job to fail when the pipeline is created. Instead, they would affect the performance and correctness of the data processing.

upvoted 2 times

✉  **emmylou** 2 months, 1 week ago

Which is the moment I decided that AI was nothing to fear

upvoted 1 times

✉  **muhusman** 9 months, 1 week ago

without a correct timestamp, the pipeline still run fine with the default timestamp. The result maybe incorrect but the job will not fail.

upvoted 3 times

✉  **Oleksandr0501** 9 months, 1 week ago

got pt2: For example, if your use case requires you to calculate a running average of values over a fixed time interval, you would likely use a non-global windowing function with a fixed time interval. On the other hand, if you need to perform a computation on the entire stream of data at once, a global windowing function might be more appropriate.

So, the choice of windowing function should be based on the specific requirements of the data processing task at hand, and it may or may not be important to apply a non-global windowing function when the pipeline is created.

question says that we need to identify streaming input, time, so non-global needed, let it be done...

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

- is it important to apply a non-global windowing function when the pipeline is created?

GPT: It is important to choose the appropriate windowing function for your data processing needs, but whether it should be a global or non-global windowing function depends on the requirements of your specific use case.

A global windowing function considers all data elements within a bounded time interval as a single window, whereas a non-global windowing function divides the data stream into smaller windows based on specified criteria (such as a fixed time interval or a number of elements).

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

okay, so D maybe

upvoted 1 times

✉  **lucaluca1982** 10 months, 2 weeks ago

what about A? This can cause the job to fail

upvoted 1 times

✉  **midgoo** 11 months, 1 week ago

Selected Answer: D

A: note that without a correct timestamp, the pipeline still run fine with the default timestamp. The result maybe incorrect but the job will not fail.
D: For unbound collection, this will fail if any aggregation function is done.

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

Answer: A

All Streaming Insert failed, because there is no TimeStamp added, otherwise there is already a DEFAULT global windowing function and can execute without assigning any windowing function.

I mean first there should be Timestamp in the data, then according to our aggregation outcome either its full time (global) or batch/chunks time aggregation(non global) will be performed.

upvoted 1 times

👤 DipT 1 year, 1 month ago

Selected Answer: D

<https://beam.apache.org/documentation/programming-guide/#windowing>

Beam's default windowing behavior is to assign all elements of a PCollection to a single, global window and discard late data, even for unbounded PCollections. Before you use a grouping transform such as GroupByKey on an unbounded PCollection, you must do at least one of the following:

Set a non-global windowing function. See Setting your PCollection's windowing function.

Set a non-default trigger. This allows the global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur.

upvoted 1 times

👤 Ray0506 1 year, 4 months ago

Selected Answer: D

Answer is D

upvoted 1 times

👤 TOXICcharlie 1 year, 4 months ago

Selected Answer: D

Correct answer is D. C does not make sense because for unbounded source like Pub/Sub, the global functions are applied by default. The reason for failure would be they are using specific aggregations that require non-global window functions, e.g. tumbling or hopping windows.

upvoted 2 times

👤 FrankT2L 1 year, 7 months ago

Selected Answer: C

C is the answer.

<https://beam.apache.org/documentation/programming-guide/#windowing-bounded-collections>

8.2.4. The single global window

By default, all data in a PCollection is assigned to the single global window, and late data is discarded. If your data set is of a fixed size, you can use the global window default for your PCollection (not our case because we are streaming).

You can use the single global window if you are working with an unbounded data set (e.g. from a streaming data source) but use caution when applying aggregating transforms such as GroupByKey and Combine. The single global window with a default trigger generally requires the entire data set to be available before processing, which is not possible with continuously updating data. To perform aggregations on an unbounded PCollection that uses global windowing, you should specify a non-default trigger for that PCollection.

upvoted 1 times

👤 gingercat 1 year, 9 months ago

Why A won't cause an error?

upvoted 3 times

👤 910 1 year, 10 months ago

Selected Answer: D

Answer: D

Description: Caution: Beam's default windowing behavior is to assign all elements of a PCollection to a single, global window and discard late data, even for unbounded PCollections. Before you use a grouping transform such as GroupByKey on an unbounded PCollection, you must do at least one of the following:

-->>>> Set a non-global windowing function. See Setting your PCollection's windowing function.

Set a non-default trigger. This allows the global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur.

-->>> If you don't set a non-global windowing function or a non-default trigger for your unbounded PCollection and subsequently use a grouping transform such as GroupByKey or Combine, your pipeline will generate an error upon construction and your job will fail.

So it looks like D

upvoted 2 times

👤 medeis_jar 2 years ago

Selected Answer: D

D.

<https://beam.apache.org/documentation/programming-guide/#windowing>

upvoted 3 times

👤 JG123 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: D

upvoted 3 times

anji007 2 years, 3 months ago

From Beam documentation:

"If you do apply GroupByKey or CoGroupByKey to a group of unbounded PCollections without setting either a non-global windowing strategy or trigger strategy, or both for each collection, Beam generates an IllegalStateException error at pipeline construction time."

Question #58

Topic 1

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

- A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
- B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
- C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
- D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Correct Answer: A

Community vote distribution

B (100%)

SteelWarrior [Highly Voted] 3 years, 4 months ago

Should go with B. Two reasons, it is a cleaner approach with single job to handle the calibration before the data is used in the pipeline. Second doing this step in later stages can be complex and maintenance of those jobs in the future will become challenging.

upvoted 55 times

Yiouk 2 years, 5 months ago

B. different MR jobs execute in series, adding 1 more job makes sense in this case.

upvoted 7 times

[Removed] [Highly Voted] 3 years, 10 months ago

Answer: A

Description: My take on this is for sensor calibration you just need to update the transform function, rather than creating a whole new mapred job and storing/passing the values to next job

upvoted 20 times

Jphix 2 years, 8 months ago

It's B. A would involving changing every single job (notice it said jobS, plural, not a single job). If that is computationally intensive, which it is you're repeating a computationally intense process needlessly several times. SteelWarrior and YuriP are right on this one.

upvoted 11 times

jin0 [Most Recent] 11 months ago

What kinds of sensor calibrations exists? I don't understand how computation in pipeline would be expense due to calibration being omitted..

upvoted 1 times

✉  **samdhimal** 1 year ago

B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.

This approach would ensure that sensor calibration is systematically carried out every time the ETL process runs, as the new MapReduce job would be responsible for calibrating the sensors before the data is processed by the other steps. This would ensure that all data is calibrated before being analyzed, thus avoiding the omission of the sensor calibration step in the future.

It also allows you to chain all other MapReduce jobs after this one, so that the calibrated data is used in all the downstream jobs.

upvoted 1 times

✉  **samdhimal** 1 year ago

Option A is not ideal, as it would be time-consuming to modify all the transformMapReduce jobs to apply sensor calibration before doing anything else, and there is a risk of introducing bugs or errors.

Option C is not ideal, as it would rely on users to apply sensor calibration themselves, which would be inefficient and could introduce inconsistencies in the data.

Option D is not ideal, as it would require a lot of simulation and testing to develop an algorithm that can predict the variance of data output accurately and it may not be as accurate as calibrating the sensor directly.

upvoted 1 times

✉  **DipT** 1 year, 1 month ago

Selected Answer: B

It is much cleaner approach

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

Best approach is calibration will be separate job because if we need to tune the calibration later also it would be to maintain without worries about all other jobs.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: B

Should be B. My reason, this is like an Anti corruption layer, and that's a good practice,

C- , if you modify your transformMapReduce will be harder to test and debug, so it's a bad practice.

C the idea de introduce manual operation is an anti pattern and has a lot of problems

D It's overkilling, a don't have sense in this scenario.

upvoted 1 times

✉  **ZIMARAKI** 2 years ago

Selected Answer: B

SteelWarrior explanation is correct :)

upvoted 3 times

✉  **lord_ryder** 2 years ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

✉  **hendrixlives** 2 years, 1 month ago

Selected Answer: B

SteelWarrior's answer is correct

upvoted 1 times

✉  **anji007** 2 years, 3 months ago

Ans: B

Adding a new job in the beginning of chain makes more sense than updating existing chain of jobs.

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'B' (introduce new job) over 'A', (instead of modifying existing job)

upvoted 5 times

👤 **YuriP** 3 years, 5 months ago

Should be B. It's a Data Quality step which has to go right after Raw Ingest. Otherwise you repeat the same step unknown (see "job_s_" in A) number of times, possibly for no reason, therefore extending ETL time.

upvoted 5 times

Question #59

Topic 1

An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application. They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?

- A. BigQuery
- B. Cloud SQL
- C. Cloud BigTable
- D. Cloud Datastore

Correct Answer: B

Reference:

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

Cloud SQL

Fully managed relational database service for MySQL, PostgreSQL, and SQL Server. Run the same relational databases you know with their rich extension collections, configuration flags and developer ecosystem, but without the hassle of self management.

[Try Cloud SQL free](#)

[Contact sales](#)

- ✓ Reduce maintenance cost with fully managed [MySQL](#), [PostgreSQL](#) and [SQL Server](#) databases
- ✓ Ensure business continuity with reliable and secure services backed by 24/7 SRE team
- ✓ Automate database provisioning, storage capacity management, and other time-consuming tasks
- ✓ Database observability made easy for developers with Cloud SQL Insights
- ✓ Easy integration with existing apps and Google Cloud services like GKE and BigQuery

Community vote distribution

B (70%)

A (28%)

2%

👤 **PolyMoe** Highly Voted 1 year ago

Selected Answer: B

B. Cloud SQL would be the most appropriate choice for the online retailer in this scenario. Cloud SQL is a fully-managed relational database

service that allows for easy management and analysis of data using SQL. It is well-suited for applications built on Google App Engine and can handle the transactional workload of an e-commerce application, as well as the analytical workload of a BI tool.

upvoted 9 times

✉  **Mathew106** Most Recent 6 months, 1 week ago

Selected Answer: B

Cloud SQL seems to fit the best. It supports transactions and can be used to run queries and do analytics.

BigQuery is good for the analysis part but it's not good for managing transactions. If the question needed a database just to store the data for analysis it would be ok. But if we want to update single transactions or add them row by row, then it's not good. BigQuery is not made to support an application. It's a DW.

BigTable is can not carry transactions over multiple rows and is better for large scale analytics jobs. Also we should pick it for use-cases with high throughput/low latency requirements. Seems redundant.

upvoted 2 times

✉  **Siddhesh05** 9 months, 2 weeks ago

Selected Answer: A

Big Query because of analysis

upvoted 4 times

✉  **izekc** 9 months, 3 weeks ago

Selected Answer: C

Should be bigtable

upvoted 1 times

✉  **juliobs** 10 months, 2 weeks ago

Selected Answer: A

I think BigQuery makes sense here. It works for transactions too.

upvoted 3 times

✉  **juliobs** 10 months ago

I just did a session with an official trainer from Google that said BigTable is better.

upvoted 2 times

✉  **Fotofilico** 3 months, 1 week ago

I'm an official trainer from Google and I can say that my best two options for this scenario would be Cloud SQL and BigQuery in that order. Also we can consider datastore since we're using it with a web app, but it's another topic.

upvoted 1 times

✉  **Aaronn14** 10 months, 4 weeks ago

A. "They want to use only a single database for this purpose" is a key requirement. You can use BigQuery for transactions, though it is not efficient. You can not use CloudSQL for analytics. So it is probably BQ.

upvoted 3 times

✉  **ninjatech** 11 months ago

Transactional Data need to be written first by application before it could be analysed so cloudsqll.

upvoted 2 times

✉  **samdhimal** 1 year ago

Both BigQuery and Cloud Bigtable are valid options for this use case, but BigQuery is better suited for this specific scenario where the retailer needs to manage and analyze large amounts of data from multiple datasets using a BI tool.

BigQuery is a fully-managed, cloud-native data warehouse that enables super-fast SQL queries using the processing power of Google's infrastructure. It can handle large, complex datasets and is well-suited for both transactional and analytical workloads. It can also handle data from multiple datasets and can be integrated with other Google Cloud services, such as Dataflow, Dataproc and Looker for BI analysis.

While Cloud Bigtable is also a good option for this use case as it is a highly scalable and performant NoSQL database that is well-suited for handling large amounts of data and high-write loads. It is not as good as BigQuery for analytical workloads and it may not be as well-suited for this specific scenario where the retailer needs to manage and analyze large amounts of data from multiple datasets using a BI tool.

upvoted 2 times

✉  **jin0** 11 months ago

Bigquery is a OLAP. So it could be not a answer I think.

upvoted 1 times

✉  **samdhimal** 1 year ago

Cloud SQL and Cloud Datastore are also good options for certain use cases, but they may not be as well-suited for this specific scenario where the retailer needs to manage and analyze large amounts of data from multiple datasets using a BI tool.

upvoted 1 times

✉  **desertlotus1211** 1 year ago

The Community is choosing Answer B - Cloud SQL, as per the question.

However when they explain - they're speaking about BQ[????]

So is it BigQuery or Cloud SQL?

upvoted 2 times

✉  **DipT** 1 year, 1 month ago

Selected Answer: B

<https://cloud.google.com/bigquery/docs/partitioned-tables>

upvoted 1 times

✉  **DipT** 1 year, 1 month ago

Selected Answer: B

It needs support for transaction so cloud sql is the choice of database and with Bigquery we can still analyze cloud sql data via federated query

<https://cloud.google.com/bigquery/docs/reference/legacy-sql>

upvoted 4 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: B

Most important part of question is transaction (RDBMS) strong ACID property database. Second part analysis of data, yes possible using any tool its possible with RDBMS db.

upvoted 1 times

✉  **zellick** 1 year, 1 month ago

Selected Answer: B

B is the answer.

<https://cloud.google.com/bigquery/docs/cloud-sql-federated-queries>

BigQuery Cloud SQL federation enables BigQuery to query data residing in Cloud SQL in real time, without copying or moving data. Query federation supports both MySQL (2nd generation) and PostgreSQL instances in Cloud SQL.

upvoted 4 times

✉  **sjesla** 9 months, 2 weeks ago

Agreed. Two catches here: transactional, and BI tool. Although BigQuery nowadays can handle everything, if we specifically deal with questions highlighting transactional data, I believe to differentiate services, we should choose what they primarily mean to be .

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: B

C and D are not able to work with BI directly, so discard.

A: It's the best option for BI for awful for transactions

B: it's the best option for transaction, and works for BI, so this must be the answer

upvoted 1 times

✉️  **Leeeeee** 1 year, 2 months ago

Selected Answer: B

BigQuery for Analytics and BI

upvoted 1 times

✉️  **Leelas** 1 year, 2 months ago

Selected Answer: B

Cloud Sql is Used to store Transactional Data and supports Sql Transactions. Where as Big Query is used for Analytics.

upvoted 2 times

✉️  **Zion0722** 1 year, 2 months ago

Cloud SQL supports transactions as well as analysis through a BI tool. Firestore/Datastore does not support SQL syntax typically needed to c analysis done by a BI tool. BigQuery is not suitable for transactional use case. BigTable does not support SQL.

It's A.

upvoted 2 times

Question #60

Topic 1

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

- A. Convert all daily log tables into date-partitioned tables
- B. Convert the sharded tables into a single partitioned table
- C. Enable query caching so you can cache data from previous months
- D. Create separate views to cover each month, and query from these views

Correct Answer: A

Community vote distribution

B (94%)

6%

✉️  **[Removed]**  3 years, 10 months ago

should be B

https://cloud.google.com/bigquery/docs/creating-partitioned-tables#converting_date-sharded_tables_into_ingestion-time_partitioned_tables

upvoted 38 times

✉️  **jin0** 11 months ago

is it already partitioned? there is a table [table]_yyyymmdd it seems to partitioned by date from log files. but I confuse why D. is not a ans if there is only reason to fail from query that exceeding 1,000 tables then I think creating views could be solution because querying views containing under 1,000 tables by a view could be queried.

upvoted 1 times

✉  **Rajuuu** 3 years, 6 months ago

The above link does mention about shard ing benefits but only about partition tables.

A is correct.

upvoted 5 times

✉  **g2000** 3 years ago

keyword is single

upvoted 6 times

✉  **Chelseajcole** 2 years, 4 months ago

you are right.

Partitioning versus sharding

Table sharding is the practice of storing data in multiple tables, using a naming prefix such as [PREFIX]_YYYYMMDD.

Partitioning is recommended over table sharding, because partitioned tables perform better. With sharded tables, BigQuery must maintain a copy of the schema and metadata for each table. BigQuery might also need to verify permissions for each queried table. practice also adds to query overhead and affects query performance.

If you previously created date-sharded tables, you can convert them into an ingestion-time partitioned table.

upvoted 10 times

✉  **vholti** 2 years, 3 months ago

The question mentions tables are sharded. So B is more appropriate answer I think.

<https://cloud.google.com/bigquery/docs/creating-partitioned-tables#convert-date-sharded-tables>

upvoted 4 times

✉  **Tanzu** 1 year, 11 months ago

<https://cloud.google.com/bigquery/docs/partitioned-tables> provides that info you are looking for. Shortly, partitioning performs better th sharding (PREFIX_yymmdd). and it is easy and supported that you can convert sharded tables into ingestion-time partitioned table. So, B is only option and better one.

upvoted 3 times

✉  **[Removed]**  3 years, 10 months ago

Answer: B

Description: Google says that when you have multiple wildcard tables, best option is to shard it into single partitioned table. Time and cost efficient

upvoted 25 times

✉  **lgdantas** 3 years, 5 months ago

Can you please share the reference?

upvoted 2 times

✉  **Tumri** 2 years, 5 months ago

https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

upvoted 7 times

✉  **Oleksandr0501**  9 months, 1 week ago

Selected Answer: B

gpt: Thank you for your feedback and additional information. You are correct that partitioned tables have a limit of 4,000 partitions, so partitioning tables by date could potentially run into this limit in the future. In this case, option B, converting sharded tables into a single partitioned table, could be a reasonable solution to avoid exceeding the maximum number of tables in BigQuery.

As you mentioned, sharded tables require additional metadata and permissions verification, which can impact query performance. Converting sharded tables into a single partitioned table can improve performance and reduce query overhead.

Therefore, based on the information provided, option B seems to be the most appropriate solution for avoiding the limit of 1,000 tables in BigQuery and optimizing query performance.

upvoted 1 times

✉  **luks_skywalker** 10 months, 1 week ago

The question seems pretty badly written. One important thing to remember is that partitioned tables also have a limit of 4000 partitions (https://cloud.google.com/bigquery/docs/partitioned-tables#ingestion_time), so moving everything to one table would just delay the problem. However, option A is not clear on how it will be done. One table per year with daily partitions? Best solution as no limit will be reached. One table per day? Then we have the same 1000 tables problem.

All things considered I'll stick to B, simply because the problem will definitely be solved for the next few years, so I'd say it's a reasonable solution.

upvoted 2 times

👤 **PolyMoe** 1 year ago

Selected Answer: B

Answer is B.

Table sharding is the practice of storing data in multiple tables, using a naming prefix such as [PREFIX]_YYYYMMDD. Partitioning is recommended over table sharding, because partitioned tables perform better. With sharded tables, BigQuery must maintain a copy of the schema and metadata for each table. BigQuery might also need to verify permissions for each queried table. This practice also adds to query overhead and affects query performance.

In answer A. we still are creating tableS (even though partitioned). So we still facing the issue of max 1000 tables. In B. we have only ONE table (partitioned)

upvoted 2 times

👤 **samdhimal** 1 year ago

Why not A?

By converting all daily log tables into date-partitioned tables, you can take advantage of partition pruning to limit the number of tables that need to be scanned during a query. Partition pruning allows BigQuery to skip scanning partitions that are not within the date range specified in the query, thus reducing the number of tables that need to be scanned and can help to avoid reaching the 1,000 table limit.

A Seems like the correct answer but I can be wrong...

upvoted 2 times

👤 **RoshanAshraf** 1 year ago

Selected Answer: B

B. Convert the sharded tables into a single partitioned table

It was a sharded Table (format is the HINT here); converting to partition table is the option.

Also as per GCP its recommended to use Partition over Sharding

upvoted 1 times

👤 **korntewin** 1 year ago

Selected Answer: A

I chose option A. From all the comments I have seen, there are various things that are misunderstood.

1. Option A is a single table with multiple shards! Google does recommend to use partition rather than shard as it has a better performance (https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard)

2. Option B is a single table with single partition! Single partition is a no for large table

upvoted 1 times

👤 **DipT** 1 year, 1 month ago

Selected Answer: B

<https://cloud.google.com/bigquery/docs/partitioned-tables>

upvoted 1 times

👤 **DGames** 1 year, 1 month ago

Selected Answer: B

Option A - already doing same loading data in separate table daily and reached 1000 table limit.

Option B - Use wild card to query the data

Option C & D - make no sense

upvoted 1 times

👤 **odacir** 1 year, 1 month ago

its B.

A - Even if you have 100+ partitioned tables, you still have the limit of less than 1000 tables. So this doesn't work for this problem.

C It's a no sense. Cache its 24h for every table that has been query in the last 24 and has no changes. Also, cache is not support with wildcard multiple tables.

D Will not work because it's a recursive issue. You still will have 100+ tables, beam query

B will work, you materialize in only one table, so will be working perfectly.

upvoted 1 times

👤 **Nirca** 1 year, 3 months ago

Selected Answer: B

Convert MANY sharded tables into a single ONE (partitioned) table

upvoted 2 times

👤 **rr000** 1 year, 5 months ago

selecting for daily/monthly data from one single partition will be very expensive. I think A is the best answer

upvoted 1 times

✉  **Preemptible_cerebrus** 1 year, 7 months ago

Selected Answer: B

C'mon, how much time are you going to take to partition every single table you have? second point and the most important, you have a table every SINGLE DAY "LOGS_YYYYMMDD" partitioning every table will end on scanning all the records of each table when you query them by c ranges using the wildcards, there will be no difference on time-partitioning each table versus consuming them as described.

upvoted 2 times

✉  **AmirN** 1 year, 7 months ago

If you follow option A, you will end up with the same amount of tables, e.g 1500 tables, though they will all be partitioned, which is not helpful Option B takes all the sharded tables and makes one large partitioned table.

upvoted 1 times

✉  **rrr000** 1 year, 5 months ago

Partitions are not tables. The issue is not performance. It is the limit imposed by bq regarding how many tables you can query.

upvoted 1 times

✉  **mihaioff** 1 year, 9 months ago

Selected Answer: B

It's B

https://cloud.google.com/bigquery/docs/creating-partitioned-tables#converting_date-sharded_tables_into_ingestion-time_partitioned_tables

upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: B

Partitioning > table sharding:

https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

upvoted 2 times

Question #61

Topic 1

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud

Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google

BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

- A. Migrate the workload to Google Cloud Dataflow
- B. Use pre-emptible virtual machines (VMs) for the cluster
- C. Use a higher-memory node so that the job runs faster
- D. Use SSDs on the worker nodes so that the job can run faster

Correct Answer: A

Community vote distribution

 **jvg637** Highly Voted 3 years, 10 months ago

B. (Hadoop/Spark jobs are run on Dataproc, and the pre-emptible machines cost 80% less)
upvoted 45 times

 **rickywck** Highly Voted 3 years, 10 months ago

I think the answer should be B:

<https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms>
upvoted 16 times

 **theseawillclaim** Most Recent 6 months, 2 weeks ago

I believe it might be "B", but what if the job is mission critical?
Pre-emptible VMs would be of no use.
upvoted 1 times

 **abi01a** 9 months, 1 week ago

I believe Exam Topics ought to provide brief explanation or supporting link to picked correct answers such as this one. Option A may be correct from the view point that Dataflow is a Serverless service that is fast, cost-effective and the fact that Preemptible VMs though can give large price discount may not always be available. It will be great to know the reason(s) behind Exam Topic selected option.

upvoted 7 times

 **samdhimal** 1 year ago

B. Use pre-emptible virtual machines (VMs) for the cluster

Using pre-emptible VMs allows you to take advantage of lower-cost virtual machine instances that may be terminated by Google Cloud after a short period of time, typically after 24 hours. These instances can be a cost-effective way to handle workloads that can be interrupted, such as batch processing jobs like the one described in the question.

Option A is not ideal, as it would require you to migrate the workload to Google Cloud Dataflow, which may cause additional complexity and would not address the issue of cost optimization.

Option C is not ideal, as it would require you to use a higher-memory node which would increase the cost.

Option D is not ideal, as it would require you to use SSDs on the worker nodes which would increase the cost.

Using pre-emptible VMs is a better option as it allows you to take advantage of lower-cost virtual machine instances and handle workloads that can be interrupted, which can help to optimize the cost of the cluster.

upvoted 3 times

 **Rodolfo_Marcos** 1 year ago

What is happening with this test "correct answer" a lot of times it doesn't make any sense. As this one... Clear it's B

upvoted 1 times

 **DipT** 1 year, 1 month ago

Selected Answer: B

Using preemptible machines are cost effective , and because is suitable for a job mentioned here as it is fault tolerant .

upvoted 1 times

 **DGames** 1 year, 1 month ago

Selected Answer: B

User Pre-emptible VM machine and save process cost, and question want simple solution.

upvoted 1 times

 **odacir** 1 year, 1 month ago

Selected Answer: B

A- Data flow it's not cost-effective in comparison with dataproc

B- Preemptible VM instances are available at much lower price—a 60-91% discount—compared to the price of standard, so this is the answer
C and D are more expensive.

upvoted 1 times

 **Remi2021** 1 year, 4 months ago

Selected Answer: B

B is right way to go

upvoted 1 times

✉  **FrankT2L** 1 year, 8 months ago

Selected Answer: B

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, you may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost

<https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms>

upvoted 1 times

✉  **Remi2021** 1 year, 10 months ago

B is the right answer. Examtopics update your answers or make your site free again.

upvoted 4 times

✉  **OmJanmeda** 1 year, 10 months ago

Selected Answer: B

B is right answer.

My experience is not good with Examtopics, so many wrong answers.

upvoted 4 times

✉  **Yaa** 1 year, 12 months ago

Selected Answer: B

B should be the right answer.

I am amazed that almost 60% of the marked answers on the site are wrong.

upvoted 2 times

✉  **byash1** 2 years ago

Ans : B,

here we are checking on reducing cost, so pre-emptible machines are best choice

upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: B

"this workload can run in approximately 30 minutes on a 15-node cluster,"
so you need performance for only 30 mins -> preemptible VMs

<https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms>

upvoted 4 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: B

A is not valid, for apache spark jobs dataproc is the best choice.

C and D are not correct, that might speed up the job or not.

For sure if we use pre-emptible machines this will be cheaper and since we don't have severe time restriction...that's the one. B

upvoted 1 times

Question #62

Topic 1

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period.

However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

- A. Set a single global window to capture all the data.
- B. Set sliding windows to capture all the lagged data.
- C. Use watermarks and timestamps to capture the lagged data.
- D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Correct Answer: B

Community vote distribution

C (88%)

6%

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description: A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

upvoted 44 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

upvoted 18 times

 **MikkelRev** Most Recent 4 months ago

Selected Answer: C

option C: Use watermarks and timestamps to capture the lagged data.

upvoted 1 times

 **MikkelRev** 4 months ago

option C: Use watermarks and timestamps to capture the lagged data.

upvoted 1 times

 **samdhimal** 1 year ago

C: Use watermarks and timestamps to capture the lagged data.

Watermarks are a way to indicate that some data may still be in transit and not yet processed. By setting a watermark, you can define a time period during which Dataflow will continue to accept late or out-of-order data and incorporate it into your processing. This allows you to maintain a predictable time period for processing while still allowing for some flexibility in the arrival of data.

Timestamps, on the other hand, are used to order events correctly, even if they arrive out of order. By assigning timestamps to each event, you can ensure that they are processed in the correct order, even if they don't arrive in that order.

upvoted 8 times

 **samdhimal** 1 year ago

Option A: Set a single global window to capture all the data is not a good idea because it may not allow for late or out-of-order data to be processed.

Option B: Set sliding windows to capture all the lagged data is not suitable for the case where you want to process the data over a predictable time period. Sliding windows are used when you want to process data over a period of time that is continuously moving forward, not a fixed period.

Option D: Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data is a good practice but not a complete solution, because it only ensures that data is ordered correctly, but it does not account for data that may arrive late.

upvoted 4 times

✉  **desertlotus1211** 1 year ago

Answer is C:

There is no such thing as a sliding windows using by dataflow.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

The naming in Apache Beam is: Fixed, Sliding, Session

In Dataflow it's: Tumbling, Hopping, Session.

I was very confused at first too when I saw "hopping" in a question.

upvoted 1 times

✉  **DeeData** 1 year ago

I highly doubt, DataFlow windowing is divided into three(3) types:

1. Fixed
2. Sliding
3. Session

upvoted 1 times

✉  **AzureDP900** 1 year ago

Answer is Use watermarks and timestamps to capture the lagged data.

A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: C

Watermark is use for late date,

upvoted 2 times

✉  **[Removed]** 1 year, 2 months ago

Watermark doesn't solve the out-of-order data problem. It only solves the problem of late data. However, with D, you can use the timestamps to solve both problems (for instance, if you're storing incoming data in a table, you can easily insert any late data to its correct place a time-partitioned table using the timestamp of the element)

upvoted 3 times

✉  **ovokpus** 1 year, 2 months ago

with watermarks, when the late data arrives, it goes into its rightful window and gets in order

upvoted 1 times

✉  **ovokpus** 1 year, 2 months ago

C even says watermarks AND timestamps.

upvoted 1 times

✉  **FrankT2L** 1 year, 8 months ago

Selected Answer: B

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, you may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost

<https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms>

upvoted 1 times

✉  **FrankT2L** 1 year, 8 months ago

delete this answer. The answer belongs to another question

upvoted 7 times

✉  **Tanzu** 1 year, 11 months ago

Selected Answer: A

That's why we have watermarks in apache beam.

upvoted 1 times

👤 **VishalBule** 1 year, 11 months ago

Answer is C Use watermarks and timestamps to capture the lagged data.

A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

upvoted 1 times

👤 **medeis_jar** 2 years ago

Selected Answer: C

"Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data."

upvoted 3 times

👤 **MaxNRG** 2 years, 1 month ago

Selected Answer: C

A is a direct No, if data don't have timestamp, we'll only have the processing time and not the "event time".

B is not either, sliding windows are not for this. Hopping/sliding windowing is useful for taking running averages of data, but not to process late data.

D looks correct but has one concept missing, the watermark to know if the process time is ok with the event time or not. I'm not 100% sure is incorrect. If, since we have a "predictable time period", might be this will do. I mean, if our dashboard is shown after the last input data has arrived (single global window), this should be ok. We'd have a "perfect watermark". Anyway we'd need triggering .

upvoted 4 times

👤 **MaxNRG** 2 years, 1 month ago

C is, I think, the correct answer: Watermark is different from late data. Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data.

I'll try to explain: Late data is inherent to Beam's model for out-of-order processing. What does it mean for data to be late? The definition and its properties are intertwined with watermarks that track the progress of each computation across the event time domain. The simple intuition behind handling lateness is this: only late input should result in late data anywhere in the pipeline.

So, is not easy to decide between C and D. If you ask me I'd say C since for D we ought to make some suppositions.

upvoted 3 times

👤 **MaxNRG** 2 years, 1 month ago

https://docs.google.com/document/d/12r7frmxNickxB5tbpuEh_n35_IJeVZn1peOrBrhhP6Y/edit#heading=h.7a03n7d5mf6g

<https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102>

<https://www.oreilly.com/library/view/streaming-systems/9781491983867/>

https://docs.google.com/presentation/d/1In5KndBTiskEOGa1QmYSCq16YWO9Dtmj7ZwzjU7SsW4/edit#slide=id.g19b6635698_3_4

upvoted 2 times

👤 **Jlozano** 2 years, 1 month ago

Selected Answer: C

"Expert Verified" but >50% questions have random answer. "Sliding window" really? Please, this can be fixed easily with our most voted answer. Of course, the correct answer is C.

upvoted 4 times

👤 **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: C

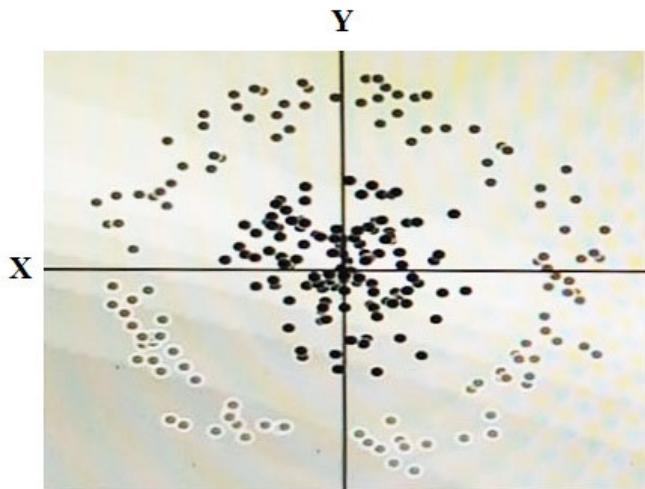
upvoted 4 times

👤 **anji007** 2 years, 3 months ago

Ans: C

upvoted 2 times

You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm. To do this you need to add a synthetic feature. What should the value of that feature be?



- A. $X^2 + Y^2$
- B. X^2
- C. Y^2
- D. $\cos(X)$

Correct Answer: D

Community vote distribution

A (76%)

B (24%)

 **jvg637** Highly Voted 3 years, 10 months ago

For fitting a linear classifier when the data is in a circle use A.

upvoted 39 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A

upvoted 14 times

✉  **Mathew106** Most Recent 6 months, 1 week ago

It's not obvious to me it is A.

As others said, $\cos(X)$ does ignore the Y value. But answer A does not seem good either. The differences seem minimal.

If you do A then you have the following issues. If you take elements in the bottom right or the top left of the circle, they will all have the same value, ZERO. Not only that, they will actually have the same value with the elements in the middle of the circle which are completely black. Moreover, elements on the extreme right and extreme left will have different values ($-x_{max}$ and $+x_{max}$).

However, if you use a $\cos(x)$ then the elements in the beginning

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Nevermind I did not understand that $X2$ and $Y2$ meant X^2 and Y^2 . Answer is A because that gives the distance from the circle. Circle radii = $\sqrt{X^2 + Y^2}$. So even though it's not a perfect answer, it makes sense.

upvoted 2 times

✉  **samdhimal** 1 year ago

A. X^2+Y^2

The synthetic feature that should be added in this case is the squared value of the distance from the origin (0,0). This is equivalent to X^2+Y^2 . If adding this feature, the classifier will be able to make more accurate predictions by taking into account the distance of each data point from the origin.

X^2 and Y^2 alone will not give enough information to classify the data because they do not take into account the relationship between X and Y.

D. $\cos(X)$ is not a suitable option because it does not take into account the Y coordinate.

upvoted 3 times

✉  **GCPpro** 1 year ago

A is the correct answer as graph of circle is $x^2 + y^2$

upvoted 2 times

✉  **desertlotus1211** 1 year ago

Answer is A:

The answer reflects 'x' to the 2nd power + 'y' the 2nd power.

I guess they can't use carrots in the exam answers!

upvoted 1 times

✉  **AzureDP900** 1 year ago

A is right

Reference:

<https://medium.com/@sachinkun21/using-a-linear-model-to-deal-with-nonlinear-dataset-c6ed0f7f3f51>

upvoted 1 times

✉  **DipT** 1 year, 1 month ago

Selected Answer: A

<https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture>

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: A

linear circle X^2+Y^2 <https://www.stat.cmu.edu/~cshalizi/dm/20/lectures/08/lecture-08.html>

upvoted 2 times

✉  **mvww11** 1 year, 7 months ago

If the shape was a circle, it would be $(x^2 + y^2)$. But I think that a quadric curve will do a better job of separating the two classes, so it would be (x^2)

upvoted 2 times

✉  **gabrysave** 1 year, 8 months ago

Answer: A.

X^2+Y^2 is the equation of a circle.

upvoted 1 times

👤 **diagniste** 1 year, 9 months ago

Selected Answer: A

C'est A

upvoted 2 times

👤 **Tanzu** 1 year, 11 months ago

Selected Answer: A

only A is draw a circle

upvoted 1 times

👤 **sraakesh95** 2 years ago

Selected Answer: A

Equation of circle as represented in the question

upvoted 1 times

👤 **moumou** 2 years ago

$F(x)$ as A B C will have always a positive values as result, for A will need a third dimension Z to represent data, only $D:\cos(x)$ can be present in the shown classification. this is a math question

upvoted 1 times

👤 **NR22** 1 year, 9 months ago

A B C will only have positive values

imaginary numbers $(i + j)$: am I a joke to you?

upvoted 1 times

👤 **medeis_jar** 2 years ago

Selected Answer: A

The 2 variables that make a circle in <http://playground.tensorflow.org> are $x1^2$ and $x2^2$.

$\sin(x)$ or $\cos(x)$ would just make horizontal stripes.

To do this you'd use those 2 variables, learning rate 0,3 for example, classification type, no regularization needed and any activation function will work fine.

upvoted 5 times

👤 **MaxNRG** 2 years, 1 month ago

Question #64

Topic 1

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
- D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the file system, and use those credentials to access the BigQuery dataset

Correct Answer: C

Community vote distribution

C (100%)

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Correct: C

upvoted 27 times

✉️  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description: Service Account are best option when granting access from tools/applications

upvoted 15 times

✉️  **samdhimal** Most Recent 1 year ago

C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset.

Creating a service account and granting dataset access to that account is the most secure way to access BigQuery from an IT application. Service accounts are designed for use in automated systems and do not require user interaction, eliminating the need for individual users to authenticate to BigQuery. Additionally, by using the private key of the service account to access the dataset, you can ensure that the authentication process is secure and that only authorized users have access to the data.

upvoted 2 times

✉️  **samdhimal** 1 year ago

Option A: Create groups for your users and give those groups access to the dataset, is not the best option because it still requires users to authenticate to BigQuery

Option B: Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request is not the best option because it still requires users to authenticate to BigQuery.

Option D: Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the file system, and use those credentials to access the BigQuery dataset is not a secure option because it involves storing sensitive information in a file on the file system, which can be easily accessed by unauthorized users.

upvoted 2 times

✉️  **DGames** 1 year, 1 month ago

Selected Answer: C

Service account approach is secure in GCP to communicate with between service or application

upvoted 1 times

✉️  **NicolasN** 1 year, 2 months ago

Selected Answer: C

[C]

The reason in <https://cloud.google.com/bigquery/docs/data-governance#identity>

"Users of BigQuery might be humans, but they might also be nonhuman applications that communicate using a BigQuery client library or the REST API. These applications should identify themselves using a service account, the special type of Google identity intended to represent a nonhuman user."

upvoted 1 times

✉️  **Tanzu** 1 year, 11 months ago

Selected Answer: C

such kind of questions are always service account oriented. and sa can be used as a user ..not just for machine2machine

users may or may not enter their credentials app login window. that's not main point by the way

upvoted 2 times

✉️  **rcruz** 1 year, 12 months ago

Selected Answer: C

Correct: C

upvoted 2 times

✉️  **anji007** 2 years, 3 months ago

Ans: C

upvoted 2 times

✉️  **daghayeghi** 2 years, 10 months ago

C:

It says "do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset", then C is the best choice.

upvoted 3 times

✉  **ave4000** 3 years, 2 months ago

Granting access to the app through a service account would mean all of the users that access the app have access to the BQ. Question was filter it out, so I believe each user would have to be added to a group that does or doesn't have access to the dataset.

upvoted 4 times

✉  **squishy_fishy** 2 years, 3 months ago

The answer is C.

When access data through application, Google recommendation is using service account.

upvoted 1 times

✉  **kavs** 3 years, 2 months ago

Yes A seems to be right

upvoted 4 times

✉  **kavs** 3 years, 2 months ago

It says individually don't want to authorise service account could be right too

upvoted 3 times

✉  **Ankush_j** 3 years, 4 months ago

Ans is C, Service account is best for secure data

upvoted 3 times

✉  **haroldbenites** 3 years, 5 months ago

C is correct

upvoted 3 times

✉  **Rajuuu** 3 years, 6 months ago

Correct C.

upvoted 4 times

Question #65

Topic 1

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

- A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataproc job.
- B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.
- C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.
- D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 0 using a custom script.

Correct Answer: C

Community vote distribution

B (100%)

✉  **jvg637**  3 years, 10 months ago

real-valued can not be null N/A or empty, have to be "0", so it has to be B.

upvoted 39 times

✉  **[Removed]**  3 years, 10 months ago

Should be B

upvoted 16 times

✉  **Nandababy**  1 month, 2 weeks ago

Why not D? keyword is Monitor, B would replace all empty fields and also cause unintended bias.

upvoted 1 times

✉  **Nandababy** 1 month, 2 weeks ago

However, Sergiomujica is right. If we need to prepare data using a casual method then its B "Dataprep".

upvoted 1 times

✉  **sergiomujica** 4 months, 3 weeks ago

The questions says "You need to prepare data using a casual method ", thats dataprep and values should be 0 so the right answer is B

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: B

No brainer. We need a real value and Dataprep is made for this. Dataflow is mainly for pre-processing before BigQuery ingests the data.

upvoted 1 times

✉  **theseawillclaim** 6 months, 2 weeks ago

Selected Answer: B

Dataprep is made for this kind of stuff, no reason to use a streaming service such as Dataflow.

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

Selected Answer: B

gpt:Cloud Dataprep is a data preparation service that can be used to transform, clean and shape data in a visually interactive way. It provides easy-to-use interface to find and replace null values.

Cloud Dataflow is a fully-managed service for executing data processing pipelines, which allows for parallel execution of data processing task. However, it requires more expertise to set up and operate than Cloud Dataprep, and is usually used for more complex data processing needs.

Therefore, option B is the most suitable approach for the given requirements.

upvoted 1 times

✉  **samdhimal** 1 year ago

Seems to me like Answers are both B and D.

B is faster to implement while D takes time.

Doesnt mean that it's wrong though. I m not sure why everyone has picked just B. Why not D? D works and does the same job. And also having custom script provides more flexibility and control over the data processing tasks and it allows you to handle missing values in a more flexible and efficient way.

upvoted 2 times

✉  **AmmarFasih** 8 months, 1 week ago

A simple rule. Wheneve any service is available by GCP for a task, always recommend to use GCP service over any other.

upvoted 1 times

✉  **rajm893** 8 months, 2 weeks ago

The "casual way" or easy way to convert to 0 is using Dataprep job rather than using the custom script.

upvoted 1 times

✉  **GCPpro** 1 year ago

B is the correct answer.

upvoted 1 times

✉  **AzureDP900** 1 year ago

Answer is Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.

Key phrases are "casual method", "need to replace null with real values", "logistic regression". Logistic regression works on numbers. Null needs to be replaced with a number. And Cloud dataprep is best casual tool out of given options.

upvoted 3 times

✉️  **DGames** 1 year, 1 month ago

Selected Answer: B

real value 0

upvoted 1 times

✉️  **byash1** 2 years ago

Selected Answer: B

It is B

upvoted 2 times

✉️  **medeis_jar** 2 years ago

Selected Answer: B

Dataprep + real value (0)

upvoted 1 times

✉️  **MaxNRG** 2 years, 1 month ago

Selected Answer: B

Dataprep is the tool. A or B.

Since they need to have a real-valued cannot be null N/A or empty, have to be "0", so it has to be B.

upvoted 3 times

✉️  **anji007** 2 years, 3 months ago

Ans: B

Dataprep suites this, so none of dataflow options qualify as answer. Then 0 can be real-value than a "~none".

upvoted 2 times

✉️  **sumanshu** 2 years, 7 months ago

Vote for 'B'

upvoted 4 times

✉️  **Ral17** 2 years, 4 months ago

why not D?

upvoted 1 times

✉️  **szefco** 2 years, 1 month ago

Because why writing custom script when you can do it easily directly in Dataprep? D technically would also work, but B is just simpler t preferred option.

upvoted 3 times

✉️  **Sush12** 2 years, 11 months ago

It is B as dataprep has a feature to convert values as desired

upvoted 2 times

Question #66

Topic 1

You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
- B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.

- C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Correct Answer: C

Community vote distribution

B (100%)

 **SonuKhan1** Highly Voted 2 years, 3 months ago

Dear Admin, almost every answer is incorrect . Please check the comments and update your website.
upvoted 47 times

 **[Removed]** Highly Voted 3 years, 10 months ago

correct: B
upvoted 21 times

 **[Removed]** 3 years, 10 months ago

<https://cloud.google.com/security/encryption-at-rest/>
upvoted 4 times

 **tprashanth** 3 years, 6 months ago

Based on the info at the link you referred, it seems C is the right answer
upvoted 4 times

 **baubaumiaomiao** 2 years, 1 month ago

If you create it locally, you can't rotate keys. Answer should be B
upvoted 2 times

 **TVH_Data_Engineer** Most Recent 1 month, 2 weeks ago

Selected Answer: B

Google Cloud Key Management Service (KMS) provides a centralized cloud service for managing cryptographic keys. By creating encryption keys in Cloud KMS, you can easily manage the lifecycle of these keys, including creation, rotation, and destruction.
WYNN NOT Create Keys Locally and Upload to Cloud KMS?

While it's possible to create keys locally and then upload them to Cloud KMS, it's generally simpler and more secure to create the keys directly in Cloud KMS. This reduces the risk associated with transferring keys and leverages the security and compliance features of Cloud KMS.
upvoted 1 times

 **emmylou** 2 months, 4 weeks ago

Help!
I chose "C" because of the statement, "encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed" and read that as needing to generate the keys locally. Can you please explain where I went wrong?
upvoted 1 times

 **odiez3** 6 months, 1 week ago

the answer is C Read the full statement.

" You need to encrypt data at rest with encryption keys that you can create "

upvoted 1 times

 **theseawillclaim** 6 months, 2 weeks ago

Selected Answer: B

B!

C is useless overhead and you cannot rotate that easily!

upvoted 1 times

 **Kiroo** 8 months, 2 weeks ago

Well for what I remember from cloud arch and what I found in <https://cloud.google.com/compute/docs/disks/customer-managed-encryption>

There are two options or the customer manages entirely or he will use the service to generate the keys so based on that is the B
upvoted 1 times

✉️  **samdhimal** 1 year ago

B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.

Cloud Key Management Service (KMS) is a fully managed service that allows you to create, rotate, and destroy encryption keys as needed. By creating encryption keys in Cloud KMS, you can use them to encrypt your data at rest in the Compute Engine cluster instances, which is runn your Redis and Kafka clusters. This ensures that your data is protected even when it is stored on disk.

upvoted 2 times

✉️  **samdhimal** 1 year ago

Option A: Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls is not the best option as it does not provide encryption at rest.

Option C: Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances, is not the best option as it does not provide a way to manage the encryption keys centrally.

Option D: Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances, is not the best option as it does not provide encryption at rest, it only secure the data in transit.

upvoted 2 times

✉️  **DGames** 1 year, 1 month ago

Selected Answer: B

B is correct answer generate key using KMS, why locally again it is overhead to upload and use everywhere.

upvoted 1 times

✉️  **Atnafu** 1 year, 2 months ago

B

If you use Google Cloud, Cloud Key Management Service lets you create your own encryption keys that you can use to add envelope encryption to your data. Using Cloud KMS, you can create, rotate, track, and delete keys.

<https://cloud.google.com/docs/security/encryption/default-encryption#:~:text=If%20you%20use%20Google%20Cloud%20Cloud%20Key%20Management%20Service%20lets%20you%20create%20your%20own%20encryption%20keys%20that%20you%20can%20use%20to%20add%20envelope%20encryption%20to%20your%20data%20Using%20Cloud%20KMS%2C%20you%20can%20create%2C%20rotate%2C%20track%2C%20and%20delete%20keys.>

upvoted 1 times

✉️  **medeis_jar** 2 years ago

Selected Answer: B

<https://cloud.google.com/security/encryption-at-rest/>

upvoted 1 times

✉️  **MaxNRG** 2 years, 1 month ago

Selected Answer: B

A makes no sense, you need to use your own keys.

You don't create keys locally and upload them, you should import it to make it work..using the kms public key...not just "uploading" it. C is also out.

IT's between B and D

Cloud KMS is a cloud-hosted key management service that lets you manage cryptographic keys for your cloud services the same way you do on-premises. You can generate, use, rotate, and destroy cryptographic keys from there.

Since you want to encrypt data at rest, is B, you don't use them for any API calls.

<https://cloud.google.com/compute/docs/disks/customer-managed-encryption>

upvoted 8 times

✉️  **lg1234** 2 years, 3 months ago

I believe you cannot upload custom keys to KMS for Compute Engine. Only via API Calls. See:

<https://cloud.google.com/security/encryption/customer-supplied-encryption-keys>

With that said, option B

upvoted 2 times

✉️  **Ysance_AGS** 2 years, 4 months ago

Answer is B : both clusters are on GCP, so we can use KMS to manage the keys.

upvoted 2 times

✉  **Sakshi_12** 2 years, 5 months ago

Well there are two things that a user can do via KMS-

1. You may be using existing cryptographic keys that were created on your premises or in an external key management system. If you migrate application to Google Cloud or if you add cryptographic support to an existing Google Cloud application, you can import the relevant keys into Cloud KMS. In the given situation, user is not having any key. So let's check 2nd option.

2. Cloud Key Management Service allows you to create, import, and manage cryptographic keys-- a general KMS definition.

Now if you don't have a key, why would you generate it locally then import it in KMS if you have an option to create a key yourself in KMS.

<https://cloud.google.com/kms/docs>

Ans - B

upvoted 3 times

✉  **gaco** 2 years, 6 months ago

it's B.

With KMS (Key Management Service), customer can create and destroy the keys as shown here: <https://cloud.google.com/kms/docs/quickstart>
Customer can also rotate key: <https://cloud.google.com/kms/docs/rotating-keys>

upvoted 2 times

✉  **awssp12345** 2 years, 6 months ago

I think people are missing on the fact that the customer wants to create their own key. This is only possible with option C. In option B, GCP KM creates and manages the key for you.

GCP certification does not always mean use all gcp managed services.

upvoted 3 times

✉  **awssp12345** 2 years, 6 months ago

Please see https://cloud.google.com/kms/docs/faq#import_keys

upvoted 1 times

✉  **awssp12345** 2 years, 6 months ago

<https://www.testpreptraining.com/tutorial/managing-customer-managed-encryption-keys-with-cloud-kms/>

upvoted 2 times

✉  **sumanshu** 2 years, 6 months ago

https://cloud.google.com/kms/docs/creating-keys#create_a_key_ring

upvoted 2 times

Question #67

Topic 1

You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

- A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.
- B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.
- C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
- D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Correct Answer: C*Community vote distribution*

C (100%)

✉️ [Removed] Highly Voted 3 years, 10 months ago

Answer: C

A & B - Need to build your own model, so discarded as options C D can do the job here using Cloud Video Intelligence API. BigTable is better option. So C is correct

upvoted 36 times

✉️ [jin0] 11 months ago

Is there any notice that has to reject own model in question..?

upvoted 1 times

✉️ [jin0] 11 months ago

I don't understand why Vision API should be a answer for labeling? there is no information about input data. isn't it?

upvoted 1 times

✉️ [Removed] Highly Voted 3 years, 10 months ago

Answer: C

Description: Why to build own model, Video API with Bigtable is best solution

upvoted 14 times

✉️ [Mathew106] Most Recent 6 months, 1 week ago**Selected Answer: C**

I don't even know if MLLib has out-of-the-box Computer Vision models. Developing this in Dataproc would be a nightmare.

Using the computer vision API on the other hand makes perfect sense.

The fact that the filtering must happen very fast and that this is a customer facing application points to BigTable so that there is very little latency and high availability. BigTable is eventually consistent but that doesn't really matter for this application.

CloudSQL will ensure strong consistency which we don't really need but it is slower and supports max 64 TB. The description mentions multi-TBs. Not really sure what several means here, but Cloud SQL doesn't have a high cap.

upvoted 1 times

✉️ [euro202] 6 months, 4 weeks ago**Selected Answer: C**

We need a model that extracts labels from videos, so Vision API could be used.

Then we need a DB very fast and that can handle several TB of data, so BigTable is the best choice.

Answer is C.

upvoted 1 times

✉️ [samdhimal] 1 year ago

Option C is the correct choice because it utilizes the Cloud Video Intelligence API to generate labels for the entities in the videos, which would save time and resources compared to building and training a model from scratch. Additionally, by storing the data in Cloud Bigtable, it allows fast and efficient filtering of the predicted labels based on the user's viewing history and preferences. This is a more efficient and cost-effective approach than storing the data in Cloud SQL and performing joins and filters.

upvoted 2 times

✉️ [AzureDP900] 1 year ago

Answer is C

Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.

1. Rather than building a new model - it is better to use Google provide APIs, here - Google Video Intelligence. So option A and B rules out
2. Between SQL and Bigtable - Bigtable is the better option as Bigtable support row-key filtering. Joining the filters is not required.

Reference:

<https://cloud.google.com/video-intelligence/docs/feature-label-detection>

upvoted 1 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

C.

The cloud video intelligence api does the label generation without the need of building any model, A and B are excluded. Now, the bdd most suitable for this is bigtable and not SQL (this big joins would be anything but fast).

<https://cloud.google.com/video-intelligence/docs/feature-label-detection>

upvoted 2 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 4 times

✉  **timolo** 2 years, 10 months ago

Answer: C

Reference <https://cloud.google.com/video-intelligence/docs/feature-label-detection>

upvoted 2 times

✉  **daghayeghi** 2 years, 10 months ago

answer C:

If we presume that use label of video as a rowkey, Bigtable will be the best option. because it can store several TB, but Cloud SQL is limited to 30TB.

upvoted 7 times

✉  **NamitSehgal** 3 years, 1 month ago

Answer: C

upvoted 3 times

✉  **Alasmindas** 3 years, 2 months ago

Option C is the correct answer.

1. Rather than building a new model - it is better to use Google provide APIs, here - Google Video Intelligence.

So option A and B rules out

2) Between SQL and Bigtable - Bigtable is the better option as Bigtable support row-key filtering. Joining the filters is not required.

upvoted 7 times

✉  **SureshKotla** 3 years, 4 months ago

Answer is D : BigTable doesn't support JOIN and not built for transactions - <https://cloud.google.com/bigtable/docs/overview>

upvoted 2 times

✉  **Surjit24** 3 years, 3 months ago

There are no joins but filtering based on condition.

upvoted 4 times

✉  **karthik89** 2 years, 11 months ago

but the requirement involves join as well, it is stated in the problem.

upvoted 2 times

✉  **sumanshu** 2 years, 6 months ago

Where? Though it's mentioned - " very fast filtering suggestions" - which means something like dictionary in python OR Key: Value (which is Bigtable)

upvoted 1 times

✉  **sraakesh95** 2 years ago

I think "based on other customer preferences" from the question requires a join before a filter is applied for collaborative filtering
upvoted 1 times

✉  **Deepakd** 1 year, 10 months ago

Recommendation based on other customer's views cannot be achieved through simple joins. A class of machine learning algorithms called collaborative filtering is required for that. You need big table to run these algorithms.

upvoted 1 times

✉  **haroldbenites** 3 years, 5 months ago

Correct C

upvoted 2 times

✉  **dg63** 3 years, 6 months ago

I doubt if C can be an answer. Will Bigtable allow filtering on labels?

upvoted 2 times

✉  **tprashanth** 3 years, 6 months ago

Yes, if its part of the rowkey

upvoted 3 times

✉  **Rajuuu** 3 years, 6 months ago

Answer is C.

upvoted 4 times

✉  **Ganshank** 3 years, 9 months ago

C.

The recommendation requires filtering based on several TB of data, therefore BigTable is the recommended option vs Cloud SQL which is limited to 10TB.

upvoted 7 times

Question #68

Topic 1

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

- A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
- B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
- C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
- D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

Correct Answer: B

Community vote distribution

C (93%)

4%

✉  **madhu1171** Highly Voted 3 years, 10 months ago

Answer should be C
upvoted 35 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C - best suitable for the purpose with autoscaling and google recommended transform engine between pubsub n bq
upvoted 25 times

✉  **barnac1es** Most Recent 4 months ago

Selected Answer: D

Option C suggests using Cloud Dataflow to run the transformations and monitoring the job system lag with Stackdriver while using the default autoscaling setting for worker instances.

While using Cloud Dataflow is a suitable choice for processing data from Cloud Pub/Sub to BigQuery, and monitoring with Stackdriver provides valuable insights, the specific emphasis on configuring non-default Compute Engine machine types (as mentioned in option D) gives you more control over cost optimization and performance tuning.

By configuring non-default machine types, you can precisely tailor the computational resources to match the specific requirements of your workload. This fine-grained control allows you to optimize costs further by avoiding over-provisioning of resources, especially if your workload is memory-intensive, CPU-bound, or requires specific configurations that are not met by the default settings.

upvoted 1 times

✉  **barnac1es** 4 months ago

Additionally, having the flexibility to adjust machine types based on workload characteristics ensures that you can achieve the desired performance levels without overspending on unnecessary resources. This level of customization is not provided by simply relying on the default autoscaling settings, making option D a more comprehensive and cost-effective solution for managing varying data volumes.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: B

At first I answered C. However, Dataproc is indeed cheaper than Dataflow. And both of them can scale automatically horizontally.

Dataflow horizontal scaling applies to both primary and secondary nodes. Scaling secondary nodes scales up CPU/compute and scaling primary nodes scales up both memory and CPU/compute.

I don't quite understand the second part of answer B where it says I should allocate resources accordingly. I guess I could do that, but auto-scaling should be enough.

upvoted 1 times

✉  **AbdullahAnwar** 11 months, 1 week ago

Answer should be C
upvoted 2 times

✉  **samdhimal** 1 year ago

C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.

Cloud Dataflow is a managed service that allows you to write and execute data transformations in a highly scalable and fault-tolerant way. By default, it will automatically scale the number of worker instances based on the input data volume and job performance, which can help minimize costs. Monitoring the job system lag with Stackdriver can help you identify any issues that may be impacting performance and take action as needed. Additionally, using the default autoscaling setting for worker instances can help you minimize manual intervention and ensure that resources are used efficiently.

upvoted 3 times

✉  **zellck** 1 year, 1 month ago

Selected Answer: C

C is the answer.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: C

@admin why all the answers are wrong. I paid 30 euros for this web and its garbage.

Dataproc has no sense in this scenario, because you want to have minimal intervention/operation. D is not a good practice, the answer is C.
upvoted 8 times

✉  **zellck** 1 year, 1 month ago

you need to look at community vote distribution and comments, and not the suggested answer.

upvoted 7 times

✉  **medeis_jar** 2 years ago

Selected Answer: C

C only as referred by MaxNRG

upvoted 4 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

C.

Dataproc does not seem to be a good solution in this case as it always require a manual intervention to adjust resources.

Autoscaling with dataflow will automatically handle changing data volumes with no manual intervention, and monitoring through Stackdriver can be used to spot bottleneck. Total execution time is not good there as it does not provide a precise view on potential bottleneck.

upvoted 9 times

✉  **StefanoG** 2 years, 2 months ago

Selected Answer: C

Dataflow, Stackdriver and autoscaling

upvoted 3 times

✉  **victorlie** 2 years, 2 months ago

Admin, please take a look on the comments. Almost all answers are wrong

upvoted 4 times

✉  **nguyenmoon** 2 years, 4 months ago

Answer should be C as dataflow is unpredictable size (input that will vary in size), dataproc is with known size

upvoted 4 times

✉  **Tanzu** 1 year, 11 months ago

dataflow over dataproc is always the preferred way in gcp. use dataproc only there is specific client requirements such as existing hadoop workloads, etc..

upvoted 1 times

✉  **sandipk91** 2 years, 5 months ago

Option C is the answer

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 1 times

✉  **apnu** 3 years, 1 month ago

B , it is correct , as it says minimum service cost, dataflow is more expensive than dataproc.

upvoted 2 times

✉  **daghayeghi** 2 years, 10 months ago

but it said "with minimal manual intervention" and for Dataproc you need to manage cluster manually, then C is the best option.

upvoted 1 times

✉  **Believerath** 2 years, 9 months ago

You have to transform the JSON messages. Hence, you need to use dataflow.

upvoted 1 times

✉  **apnu** 3 years, 1 month ago

B , it is correct as it says minimum service cost, data flow is more expensive than dataproc.

upvoted 2 times

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

- A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
- B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
- C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
- D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Correct Answer: B

Community vote distribution

C (51%)	A (40%)	9%
---------	---------	----

✉️  **VishalB**  3 years, 6 months ago

Correct Answer: A

Destination is GCS and having multi-regional so A is the best option available.

Even since BigQuery Data Transfer Service initially supports Google application sources like Google Ads, Campaign Manager, Google Ad Manager and YouTube but it does not support destination anything other than bq data set

upvoted 70 times

✉️  **asksathvik** 2 years, 5 months ago

Kindly re-read the question, the question says Google Cloud not Cloud storage...once you master that you will understand the whole question and be able to pick the right answer which is C

upvoted 12 times

✉️  **triipinbee** 2 years, 5 months ago

all the option clearly says "storage bucket", once you master that, you'll realize the correct option is A

upvoted 52 times

✉️  **Rodrigo4N** 9 months, 2 weeks ago

Gottem!

upvoted 2 times

✉️  **HarshKothari21** 1 year, 4 months ago

good one :)

upvoted 3 times

✉️  **yoshik** 2 years, 4 months ago

logs like stuff goes better on buckets

upvoted 4 times

✉  **tainangao** 2 years, 4 months ago

Currently, you cannot use the BigQuery Data Transfer Service to transfer data out of BigQuery.

<https://cloud.google.com/bigquery-transfer/docs/introduction>

upvoted 9 times

✉  **phidelics** 7 months, 4 weeks ago

You can use BQ Data transfer Service for Youtube channels now

upvoted 4 times

✉  **AmmarFasih** 8 months, 1 week ago

but I think now you can use BigQuery Data Transfer Service for youtube channels and many other

upvoted 3 times

✉  **henryCho** 3 years ago

What about ANSI SQL?

upvoted 7 times

✉  **Jphix** 2 years, 8 months ago

I guess they are assuming that you will just query the data in Cloud Storage from BQ. The question specifically is, "How should you set the log data transfer into Google Cloud?", not "How should you set up the querying." ANSI SQL is a distraction!

upvoted 5 times

✉  **nadavw** 1 year, 8 months ago

use external table for it

upvoted 1 times

✉  **Ganshank** Highly Voted 3 years, 9 months ago

None of the answers make any sense.

BigQuery Transfer Service is for moving data from various sources (S3, Youtube etc) into BigQuery, not Google Cloud Storage. Further, how are we supposed to use SQL to query data if it is stored in Cloud Storage?

upvoted 18 times

✉  **dambilwa** 3 years, 7 months ago

Agreed! - All Options look wrong

upvoted 1 times

✉  **dambilwa** 3 years, 7 months ago

Option [A] is the least worse option... for world wide teams to perform ANSI SQL Queries, it would be easier to create a ext. table or load from Multi AZ bucket... BQ Data Transfer service is used to push data in BQ, hence ruling out Option C & D

upvoted 5 times

✉  **StefanoG** 2 years, 4 months ago

The best option would be to use "BigQuery Transfer Service" to upload data to BQ. But BQ is not present as a destination, so the only working option is Multi Regional GCS

upvoted 3 times

✉  **TNT87** 3 years, 3 months ago

Kindly re-read the question, the question says Google Cloud not Cloud storage...once you master that you will understand the whole question and be able to pick the right answer which is C

upvoted 9 times

✉  **TNT87** 3 years, 3 months ago

<https://cloud.google.com/bigquery-transfer/docs/youtube-channel-transfer>

this link will help to cement the answer.

upvoted 3 times

✉  **rocky48** Most Recent 2 months, 1 week ago

Selected Answer: C

To transfer YouTube channel data to Google Cloud for analysis, you can use the BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination 1. This service allows you to automatically schedule and manage recurring load jobs for YouTube Channel reports 1. The BigQuery Data Transfer Service for YouTube Channel reports supports the following reporting options: Channel Reports (automatically loaded into BigQuery) 1. When you transfer data from a YouTube Channel into BigQuery, the data is loaded into BigQuery tables that are partitioned by date 1.

upvoted 2 times

👤 **ziyunxiao** 3 months, 2 weeks ago

The correct answer is C. <https://cloud.google.com/bigquery/docs/dts-introduction>

upvoted 1 times

👤 **exniantwort** 4 months, 1 week ago

not c

The BigQuery Data Transfer Service automates data movement into [BigQuery] on a scheduled, managed basis. Your analytics team can the foundation for a BigQuery data warehouse without writing a single line of code.

...
After you configure a data transfer, the BigQuery Data Transfer Service automatically loads data into [BigQuery] on a regular basis.

upvoted 2 times

👤 **bmquijano7** 4 months, 2 weeks ago

Correct answer is A.

At first, it's between A and C because of the term multi-regional. However, since you cannot use the BigQuery Data Transfer Service to transfer data out of BigQuery, A is the correct answer.

Source: <https://cloud.google.com/bigquery/docs/dts-introduction>

upvoted 2 times

👤 **Mathew106** 6 months, 1 week ago

Selected Answer: A

What VishalB said is correct.

upvoted 1 times

👤 **Oleksandr0501** 9 months, 1 week ago

GPT: Option A: Using Storage Transfer Service to transfer offsite backup files to a Cloud Storage Multi-Regional storage bucket is not the best option for this use case since this service is designed to transfer data from on-premises or other cloud providers to Google Cloud, not for data transfer from YouTube channels.

Option B: Using Storage Transfer Service to transfer offsite backup files to a Cloud Storage Regional bucket as a final destination is also not the best option since this service is designed for data transfer from on-premises or other cloud providers to Google Cloud, not for data transfer from YouTube channels.

upvoted 1 times

👤 **Oleksandr0501** 9 months, 1 week ago

Option C: BigQuery Data Transfer Service is specifically designed for ingesting data into BigQuery from various sources, including YouTube channels, and supports automated data transfers on a schedule. The service can also transfer data to Cloud Storage buckets, which make an ideal choice for this use case. The Multi-Regional storage bucket would ensure high availability and low latency for access to the data globally.

upvoted 1 times

👤 **Oleksandr0501** 9 months, 1 week ago

Option D: Using BigQuery Data Transfer Service to transfer data to a Cloud Storage Regional storage bucket is not the best option for this use case since it may not be the most efficient or cost-effective solution for global teams that need to access the data. A multi-regional storage bucket would be more suitable for this use case.

Therefore, option C is the best option for setting up the log data transfer from YouTube channels to Google Cloud for analysis.

upvoted 2 times

👤 **Oleksandr0501** 9 months, 1 week ago

Option C suggests using BigQuery Data Transfer Service to transfer data to a Cloud Storage Multi-Regional storage bucket. While this is a valid approach to transfer data to a storage bucket, using a multi-regional bucket is not necessary for this use case and can incur additional cost.

upvoted 1 times

👤 **Oleksandr0501** 9 months, 1 week ago

also chatgpt: Option C suggests using BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination. This option is more appropriate as it allows you to easily transfer the data from YouTube channels to a Cloud Storage Multi-Regional bucket, which can then be accessed and analyzed using ANSI SQL and other types of analysis in BigQuery.
....Sh*t

upvoted 1 times

👤 **pankajk0811** 9 months, 2 weeks ago

Answer is A

As Bigquery data transfer service cannot support loading data from offsite locations.

<https://cloud.google.com/bigquery/docs/dts-introduction>

upvoted 2 times

✉  **izekc** 9 months, 3 weeks ago

Selected Answer: C

I think it is talking about this one.

The BigQuery Data Transfer Service for YouTube allows you to automatically schedule and manage recurring load jobs for YouTube Channel reports.

Since then, it should be "C"

<https://cloud.google.com/bigquery/docs/youtube-channel-transfer>

upvoted 4 times

✉  **izekc** 9 months, 3 weeks ago

Selected Answer: C

Should be C

upvoted 1 times

✉  **juliosb** 10 months, 2 weeks ago

Selected Answer: C

I vote for C.

<https://cloud.google.com/bigquery-transfer/docs/youtube-channel-transfer>

The question never mentioned GCS as some are commenting here.

upvoted 1 times

✉  **jin0** 11 months ago

why most of questions is not a clear..

upvoted 1 times

✉  **ayush_1995** 1 year ago

Selected Answer: A

A over C as why to involve bq service if its not confirmed if its using bq or not

upvoted 1 times

✉  **samdhimal** 1 year ago

So the issue I am facing currently is my answer seems to be the first one for this questions. I can be wrong. Please someone reply with some document if I am missing something.

Correct Answer:

Option D: Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

upvoted 3 times

✉  **samdhimal** 1 year ago

This option offers several advantages for your use case:

BigQuery Data Transfer Service allows for automatic, scheduled data transfer from YouTube to a Cloud Storage bucket, and then into BigQuery for analysis. This ensures that your marketing teams have access to up-to-date data for ANSI SQL and other types of analysis.

Using a Regional storage bucket ensures low-latency data access for your worldwide marketing teams. Since regional storage bucket is located in close proximity to the teams, this improves the performance of the analysis.

By using BigQuery for analysis and Cloud Storage for data storage, you can take advantage of the scalability and reliability of Google Cloud infrastructure.

upvoted 3 times

✉  **samdhimal** 1 year ago

Option A and B are incorrect because Storage Transfer Service is used for moving large amounts of data from online or on-premises storage to a Cloud Storage bucket, but it doesn't automatically load data into BigQuery, so it wouldn't be the best choice for performing ANSI SQL and other types of analysis on YouTube channel data.

Option C is incorrect because it's using BigQuery Data Transfer Service but it's transferring data to a Multi-Regional storage bucket, but Multi-Regional storage bucket is less ideal for low-latency data access for worldwide marketing teams.

upvoted 2 times

✉  **desertlotus1211** 1 year ago

Big Query is only multi-region in US and EU... question ask for worldwide...

Hence Answer is: A

upvoted 1 times

Question #70

Topic 1

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

- A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
- B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
- C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
- D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

Correct Answer: D

Community vote distribution

B (78%)

A (22%)

 **Ganshank** Highly Voted 3 years, 9 months ago

B.

The question is focused on designing storage for very large files, with support for compression, ANSI SQL queries, and parallel loading from input locations. This can be met using GCS for storage and Bigquery permanent tables with external data source in GCS.

upvoted 57 times

✉  **atnafu2020** 3 years, 6 months ago

why GCS as external since Bigquery can be used as storage as well?

upvoted 10 times

✉  **jkhong** 1 year, 1 month ago

The question focuses on "designing storage", rather than designing a data warehouse.

upvoted 5 times

✉  **atnafu2020** 3 years, 6 months ago

A seems correct for me

upvoted 11 times

✉  **atnafu2020** 3 years, 5 months ago

Since its best practice, i go by with B not A

upvoted 4 times

✉  **gopinath_k** 2 years, 10 months ago

They want to store the files if you try with bq I think you will need to strike the word compression.

upvoted 2 times

✉  **[Removed]**  3 years, 10 months ago

Should be A

upvoted 14 times

✉  **tavva_prudhvi** 1 year, 7 months ago

Not A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQu is another issue which can be more expensive than Google Storage.

upvoted 5 times

✉  **MaxNRG**  1 month, 2 weeks ago

Selected Answer: B

1. Store Avro files in GCS

2. Query them in BigQuery (federated tables)

upvoted 2 times

✉  **forepick** 8 months ago

Selected Answer: B

Answer is B.

The requirements are:

- storage for compressed text files

- parallel loads to SQL tool

AVRO is a compressed format for text files, which makes it possible to load chunks of a very large file in parallel to BigQuery.

gzip files are seamless in GCS though, but cannot load in parallel to BQ.

upvoted 4 times

✉  **samdhimal** 1 year ago

Correct Answer:

A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.

This option offers several advantages:

- Transforming the text files to compressed Avro using Cloud Dataflow allows for parallel processing of the input data, improving the efficiency of the pipeline.
- Compressing the data in Avro format further reduces the storage space required and improves data transfer performance.
- Storing the data in BigQuery supports ANSI SQL queries and allows for easy querying of the data.

- BigQuery is a fully-managed data warehousing solution, it's scalable and can handle large datasets and concurrent queries, so it's suitable for large text files.

upvoted 3 times

✉  **samdhimal** 1 year ago

Option B is similar to option A but it's using a permanent linked table between Cloud Storage and BigQuery, this approach is not recommended as it's not efficient and could lead to data duplication, and it doesn't take advantage of the parallel processing capabilities of Cloud Dataflow.

Option C and D are incorrect because they don't take advantage of the parallel processing capabilities of Cloud Dataflow, and they don't use Avro format for compression which is more efficient and recommended by Google. Storing the data in Cloud Bigtable also doesn't support ANSI SQL queries which is a requirement for this use case.

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

Selected Answer: B

Designing storage solution, not data warehousing -> So Cloud Storage.

Support compression -> just use Avro

Parallel load -> refers to upload from input locations, NOT download.

Load in parallel using -m flag for gsutil cp

<https://cloud.google.com/storage/docs/uploads-downloads#console>

upvoted 3 times

✉  **odacir** 1 year, 1 month ago

Selected Answer: B

C and D are discarded.

A and B are possible.

A is the best for query, but ... the sentence says: ou also want to support compression and parallel load from the input locations using Google's recommended practices.

BigQuery only support parallel load from storage, storage support parallel load from CLI. So the only option is B.

upvoted 3 times

✉  **zellok** 1 year, 1 month ago

Selected Answer: B

B is the answer.

upvoted 1 times

✉  **nkit** 1 year, 1 month ago

Selected Answer: B

"Very large files" and "long term storage" are two key phrases- both of which indicate to pick cloud storage as option. Hence B is correct.

upvoted 1 times

✉  **NicolasN** 1 year, 1 month ago

Selected Answer: B

All the comments argue about [A] and [B] as a storage destination. But there is a limitation on loading compressed Avro files into BigQuery that cuts the Gordian knot:

❗ "... Compressed Avro files are not supported, but compressed data blocks are ..."

From: https://cloud.google.com/bigquery/docs/batch-loading-data#loading_compressed_and_uncompressed_data

upvoted 3 times

✉  **ffggre** 3 months, 1 week ago

Compressed AVRO files are supported by BQ

upvoted 1 times

✉  **izekc** 9 months, 3 weeks ago

No, it is not

<https://github.com/GoogleCloudPlatform/bigquery-ingest-avro-dataflow-sample>

upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

Selected Answer: A

<https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro>

Advantages of Avro:

Avro is the preferred format for loading data into BigQuery. Loading Avro files has the following advantages over CSV and JSON (newline delimited):

The Avro binary format:

Is faster to load. The data can be read in parallel, even if the data blocks are compressed.

Doesn't require typing or serialization.

Is easier to parse because there are no encoding issues found in other formats such as ASCII.

When you load Avro files into BigQuery, the table schema is automatically retrieved from the self-describing source data.

upvoted 2 times

✉  **Lui1979** 1 year, 8 months ago

Selected Answer: B

<https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro>

The Avro binary format:

Is faster to load. The data can be read in parallel, even if the data blocks are compressed

upvoted 5 times

✉  **cloudmon** 1 year, 2 months ago

Your comment supports A more than B

upvoted 2 times

✉  **Didine_22** 1 year, 9 months ago

Selected Answer: B

B

Because they are talking about the parallel loading from input locations.

upvoted 1 times

✉  **devric** 1 year, 9 months ago

Selected Answer: B

B. The objective is to follow the best practices.

upvoted 2 times

✉  **devric** 1 year, 9 months ago

Sorry I mean A not B :-)

upvoted 1 times

✉  **danielttt** 1 year, 10 months ago

Selected Answer: B

B fits all the requirements

upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

So does A

upvoted 1 times

✉  **Arkon88** 1 year, 10 months ago

Selected Answer: B

B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.

1) Avro covers parallelism and compression

2) A and B are correct, but B is the best answer

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

<https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro>

upvoted 2 times

✉  **sfsdeniso** 1 year, 2 months ago

GCP and BigQuery storage costs are the same ~20\$ per 1TB per month

upvoted 1 times

Question #71

Topic 1

You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.
- B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.
- D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Correct Answer: A

Community vote distribution

A (100%)

✉  **VishalB** Highly Voted 3 years, 6 months ago

Correct Answer : A

Entity analysis -> Identify entities within documents receipts, invoices, and contracts and label them by types such as date, person, contact information, organization, location, events, products, and media.

Sentiment analysis -> Understand the overall opinion, feeling, or attitude sentiment expressed in a block of text.

-- Avoid Custom models

upvoted 36 times

✉  **AzureDP900** 1 year ago

<https://cloud.google.com/natural-language/docs/analyzing-entities>

<https://cloud.google.com/natural-language/docs/analyzing-sentiment>

upvoted 1 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

should be A

upvoted 12 times

✉  **Az900Exam2021** Most Recent 4 months ago

For the first time, the answer in exam topics matches community vote :-).

upvoted 2 times

✉  **AmmarFasih** 8 months, 1 week ago

Selected Answer: A

Of course the answer is A. Since the problem already states that you don't have time, resources or expertise. So the best solution in the case to utilize the available API. Also since we need to extract the labels and not the sentiment of the text, we'll go for option A and not B
upvoted 2 times

✉  **samdhimal** 1 year ago

A. Call the Cloud Natural Language API from your application. Process the generated Entities Analysis as labels.

The Cloud Natural Language API is a pre-trained machine learning model that can be used for natural language processing tasks such as entity recognition, sentiment analysis, and syntax analysis. The API can be called from your application using a simple API call, and it can generate entities analysis that can be used as labels for the user's blog posts. This would be the quickest and easiest option for your team since it would not require any machine learning expertise or additional developer resources to build and train a model. Additionally, it will give you accurate and up-to-date results as the API is constantly updated by Google.

upvoted 1 times

✉  **AzureDP900** 1 year ago

Answer is A

Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.

Entity analysis -> Identify entities within documents receipts, invoices, and contracts and label them by types such as date, person, contact information, organization, location, events, products, and media.

Sentiment analysis -> Understand the overall opinion, feeling, or attitude sentiment expressed in a block of text.

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

Selected Answer: A

A is the answer.

<https://cloud.google.com/natural-language/docs/analyzing-entities>

Entity Analysis inspects the given text for known entities (proper nouns such as public figures, landmarks, etc.), and returns information about those entities.

upvoted 1 times

✉  **NicolasN** 1 year, 2 months ago

Apparently, there is unanimity on answer [A]

What if there was another available answer in an actual exam?

E. Call the Cloud Natural Language API from your application. Process the generated Content Classification as labels

What would you choose, A or E?

My opinion is that Content Classification is more suitable for detecting subject.

upvoted 2 times

✉  **Remi2021** 1 year, 6 months ago

A is the right one . Doc says:

Entity analysis inspects the given text for known entities (Proper nouns such as public figures, landmarks, and so on. Common nouns such as restaurant, stadium, and so on.) and returns information about those entities. Entity analysis is performed with the analyzeEntities method.

upvoted 1 times

✉  **waterh2oeau** 1 year, 8 months ago

Selected Answer: A

Vote for A

upvoted 1 times

✉  **bury** 1 year, 11 months ago

Selected Answer: A

a is correct

upvoted 1 times

✉  **JayZeeLee** 2 years, 2 months ago

A.

CD don't work as it requires Machine Learning experience.

B - Sentiment Analysis is to analyze attitude, opinion, etc. So A.

upvoted 1 times

👤 **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 3 times

👤 **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 4 times

👤 **[Removed]** 3 years, 10 months ago

Answer: A

Description: As time is less, use cloud NLP and entity is used to label general subjects, sentiment label for sentiment analysis

upvoted 5 times

Question #72

Topic 1

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

- A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
- B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
- C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
- D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

Correct Answer: A

Community vote distribution

C (100%)

👤 **daghayeghi** Highly Voted 2 years, 10 months ago

answer C:

BigQuery can access data in external sources, known as federated sources. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive.

When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, a table is created in a special dataset and will be available for approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse.

"Dan Sullivan" Book

upvoted 49 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Should be C

upvoted 30 times

👤 **emmylou** Most Recent 2 months, 4 weeks ago

On so many of these questions, how do you actually know if you're correct. I said C but the correct answer was A. Honestly, it's driving me cr

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: C

For the ones saying BigTable is cheaper, BigTable in eu-north1 costs \$0.748/hour per node. So if you were to run the node 24/7 you would pay more than 500\$ per month. Querying 1TB of data in BigQuery is 7.5\$. With smart querying and good database design you can minimize the bytes processed by BQ. So even though BigTable does not directly charge for querying, it charges for running the cluster and the overall price does not make sense. And as far as I know, it's not possible to spin up and shut down BigTable automatically.

Also, since the table is an external table to BigQuery, we incur no cost for storing that data in BigQuery and paying 300\$ per month for storage upvoted 1 times

✉  **samdhimal** 1 year ago

C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.

Cloud Storage is a highly durable and cost-effective object storage service that can be used to store large amounts of text files. By storing the input data in CSV format in Cloud Storage, you can minimize costs while still being able to query the data using BigQuery.

BigQuery is a fully-managed, highly-scalable data warehouse that allows you to perform fast SQL-like queries on large datasets. By linking the Cloud Storage data as permanent tables in BigQuery, you can enable multiple users to query the data using multiple engines without the need for additional compute resources. This approach would be the most cost-effective for querying aggregate values for multiple users, as BigQuery charges based on the amount of data scanned per query, so the more data you store in BigQuery the less you pay per query.

upvoted 2 times

✉  **samdhimal** 1 year ago

Option D, using Cloud Storage for storage and linking as temporary tables in BigQuery for query, would not be the best choice because temporary tables only exist for the duration of a user session or query and you would need to create and delete them each time a user queries the data, which would add additional cost and complexity to the process.

Option A, Using Cloud Bigtable for storage, and installing the HBase shell on a Compute Engine instance to query the data, is not a cost-effective solution as Cloud Bigtable is a managed NoSQL database service which is more expensive than storing in Cloud Storage and querying in BigQuery.

Option B, Using Cloud Bigtable for storage, and linking as permanent tables in BigQuery for query, is not a cost-effective solution as Cloud Bigtable is a managed NoSQL database service which is more expensive than storing in Cloud Storage and querying in BigQuery.

upvoted 1 times

✉  **RoshanAshraf** 1 year ago

Selected Answer: C

CSV files - Cloud Storage

BigQuery - Aggregate, multiple users

Permanent table - multiple users

External Tables is Easy to implement, cost effective

upvoted 3 times

✉  **rivua** 1 year, 1 month ago

The 'correct' answers on this platform are ridiculous

upvoted 6 times

✉  **zellck** 1 year, 1 month ago

Selected Answer: C

C is the answer.

https://cloud.google.com/bigquery/docs/external-data-cloud-storage#create_a_permanent_external_table

upvoted 1 times

✉  **VishalBule** 1 year, 11 months ago

Answer is C Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.

BigQuery can access data in external sources, known as federated sources. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive.

When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, it is created in a special dataset and will be available for approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse.

upvoted 1 times

✉  **medeis_jar** 2 years ago

Selected Answer: C

Bigtable is expensive. So Cloud Storage for storing and BigQuery with permanent table for linking and querying.

upvoted 3 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

Not A or B

Big table is expensive, que initial data is in csv format, besides, if others are going to query data with multiple engines... GCS is the storage. Between c and D is all about permanent or temporary.

Permanent table is a table that is created in a dataset and is linked to your external data source. Because the table is permanent, you can use dataset-level access controls to share the table with others who also have access to the underlying external data source, and you can query the table at any time.

When you use a temporary table, you do not create a table in one of your BigQuery datasets. Because the table is not permanently stored in a dataset, it cannot be shared with others. Querying an external data source using a temporary table is useful for one-time, ad-hoc queries over external data, or for extract, transform, and load (ETL) processes.

I think is C.

upvoted 5 times

✉  **MaxNRG** 2 years, 1 month ago

<https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer>.

Permanent table—You create a table in a BigQuery dataset that is linked to your external data source. This allows you to use BigQuery dataset-level IAM roles to share the table with others who may have access to the underlying external data source. Use permanent tables when you need to share the table with others.

Temporary table—You submit a command that includes a query and creates a non-permanent table linked to the external data source. With this approach you do not create a table in one of your BigQuery datasets, so make sure to give consideration towards sharing the query or table. Consider using a temporary table for one-time, ad-hoc queries, or for one time extract, transform, or load (ETL) workflows

upvoted 3 times

✉  **maurodipa** 2 years, 1 month ago

Answer is A. While C seems the most reasonable answer there are 2 points to notice: a) load jobs are limited to 15 TB across all input files in BigQuery (<https://cloud.google.com/bigquery/quotas>); b) It is requested to minimize the cost of querying and queries in BigTable are free, while queries in BigQuery are charged per byte (<https://cloud.google.com/bigquery/pricing>)

upvoted 2 times

✉  **Abhi16820** 2 years, 2 months ago

[https://cloud.google.com/bigquery/external-data-bigtable#:~:text=shared%20with%20others.-,Querying%20an%20external%20data%20source%20using%20a%20temporary%20table%20is%20useful%20for%20one%20ad%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20\(ETL\)%20processes.-,Querying%20Cloud%20Bigtable](https://cloud.google.com/bigquery/external-data-bigtable#:~:text=shared%20with%20others.-,Querying%20an%20external%20data%20source%20using%20a%20temporary%20table%20is%20useful%20for%20one%20ad%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20(ETL)%20processes.-,Querying%20Cloud%20Bigtable)

upvoted 1 times

✉  **tsoetan001** 2 years, 3 months ago

C is the answer.

upvoted 1 times

✉  **Ysance_AGS** 2 years, 4 months ago

A is correct since the question asks "You want to minimize the cost of querying aggregate values" => Big Table is free when querying data.

upvoted 3 times

✉  **nguyenmoon** 2 years, 4 months ago

Vote for C

upvoted 1 times

✉  **gcp_learner** 2 years, 6 months ago

Interesting options. For me, A & B ruled out because BigTable doesn't fit this use case, leaves us with C & D. C will incur additional cost of storing data in GCS & BigQuery because it mentions linking.

So I would go with D ie store the data in GCS and create external tables in BigQuery.

upvoted 4 times

Question #73

Topic 1

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally.

You also want to optimize data for range queries on non-key columns. What should you do?

- A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
- B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
- C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.
- D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Correct Answer: C

Community vote distribution

C (100%)

✉  **[Removed]**  3 years, 10 months ago

Answer: C

Description: Spanner allows transaction tables to scale horizontally and secondary indexes for range queries

upvoted 31 times

✉  **[Removed]**  3 years, 10 months ago

Correct: C

upvoted 9 times

✉  **NeoNitin**  5 months ago

horizontal scaling only possible in Cloud spanner. still you dont get I have whole question list I can share you neonitin6@gmaildotcom
upvoted 1 times

✉  **nhanhoangle** 9 months, 2 weeks ago

Selected Answer: C

Correct: C

upvoted 1 times

✉  **PolyMoe** 1 year ago

Selected Answer: C

Cloud Spanner is a fully-managed, horizontally scalable relational database service that supports transactions and allows you to optimize data for range queries on non-key columns. By using Cloud Spanner for storage, you can ensure that your database can scale horizontally to meet needs of your application.

To optimize data for range queries on non-key columns, you can add secondary indexes, this will allow you to perform range scans on non-key columns, which can improve the performance of queries that filter on non-key columns.

upvoted 2 times

👤 **samdhimal** 1 year ago

C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.

Cloud Spanner is a fully-managed, horizontally scalable relational database service that supports transactions and allows you to optimize data for range queries on non-key columns. By using Cloud Spanner for storage, you can ensure that your database can scale horizontally to meet needs of your application.

To optimize data for range queries on non-key columns, you can add secondary indexes, this will allow you to perform range scans on non-key columns, which can improve the performance of queries that filter on non-key columns.

upvoted 3 times

👤 **samdhimal** 1 year ago

- Option A, Using Cloud SQL for storage and adding secondary indexes to support query patterns, may not be the best option as Cloud SQL is a relational database service that does not support horizontal scaling and may not be able to handle the large amount of data and the number of queries required by your application.

upvoted 1 times

👤 **Mathew106** 6 months, 1 week ago

Cloud SQL does support replicas to increase availability. Why is that not considered horizontal scaling?

upvoted 2 times

👤 **samdhimal** 1 year ago

- Option B, Using Cloud SQL for storage and using Cloud Dataflow to transform data to support query patterns, may not be the best option as Cloud SQL is a relational database service that does not support horizontal scaling and may not be able to handle the large amount of data and the number of queries required by your application. Additionally, Cloud Dataflow is a data processing service and not a storage service, so it may not be the best fit for this use case.

- Option D, Using Cloud Spanner for storage and using Cloud Dataflow to transform data to support query patterns, is not necessary as Cloud Spanner provides the ability to optimize data for range queries on non-key columns by adding secondary indexes. Cloud Spanner also supports transactional consistency, which is a feature that allows you to perform multiple operations that must be performed together in a single transaction. Additionally, Cloud Dataflow is a data processing service and not a storage service, so it may not be the best fit for this use case.

upvoted 1 times

👤 **zelliCK** 1 year, 1 month ago

Selected Answer: C

C is the answer.

<https://cloud.google.com/architecture/autoscaling-cloud-spanner>

When you create a Cloud Spanner instance, you choose the number of compute capacity nodes or processing units to serve your data. However, if the workload of an instance changes, Cloud Spanner doesn't automatically adjust the size of the instance. This document introduces the Autoscaler tool for Cloud Spanner (Autoscaler), an open source tool that you can use as a companion tool to Cloud Spanner. This tool lets you automatically increase or reduce the number of nodes or processing units in one or more Spanner instances based on how their capacity is being used.

<https://cloud.google.com/spanner/docs/secondary-indexes>

You can also create secondary indexes for other columns. Adding a secondary index on a column makes it more efficient to look up data in that column.

upvoted 1 times

👤 **sedado77** 1 year, 4 months ago

Selected Answer: C

As sumanshu said

upvoted 1 times

👤 **tsoetan001** 2 years, 3 months ago

Answer: C

upvoted 1 times

👤 **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 4 times

👤 **sumanshu** 2 years, 6 months ago

A is not correct because Cloud SQL does not natively scale horizontally.

B is not correct because Cloud SQL does not natively scale horizontally.

C is correct because Cloud Spanner scales horizontally, and you can create secondary indexes for the range queries that are required.

D is not correct because Dataflow is a data pipelining tool to move and transform data, but the use case is centered around querying.

upvoted 8 times

✉  **timolo** 2 years, 10 months ago

Answer: C

<https://cloud.google.com/spanner/docs/secondary-indexes>

upvoted 1 times

✉  **Nileshk611** 3 years, 1 month ago

Correct: C

upvoted 3 times

✉  **arghya13** 3 years, 2 months ago

Correct answers is C

upvoted 2 times

Question #74

Topic 1

Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data.

Which product should they use to store the data?

- A. Cloud Bigtable
- B. Google BigQuery
- C. Google Cloud Storage
- D. Google Cloud Datastore

Correct Answer: A

Reference:

<https://cloud.google.com/bigtable/docs/schema-design-time-series>

The basic design patterns for storing time-series data in Bigtable are as follows:

- **Rows are time buckets**
 - **New columns for new events**
 - **New cells for new events**
- **Rows represent single timestamps**
 - **Serialized column data**
 - **Unserialized column data**

Community vote distribution

A (82%)

C (18%)

✉  **zellick** Highly Voted 1 year, 1 month ago

Selected Answer: A

A is the answer.

<https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-bigtable>

Bigtable is Google's NoSQL Big Data database service. It's the same database that powers many core Google services, including Search, Analytics, Maps, and Gmail. Bigtable is designed to handle massive workloads at consistent low latency and high throughput, so it's a great choice for both operational and analytical applications, including IoT, user analytics, and financial data analysis.

Bigtable is an excellent option for any Apache Spark or Hadoop uses that require Apache HBase. Bigtable supports the Apache HBase 1.0+ API and offers a Bigtable HBase client in Maven, so it is easy to use Bigtable with Dataproc.

upvoted 12 times

✉ **Atnafu** 1 year, 1 month ago

Hbase concept here us beautiful

upvoted 2 times

✉ **Mathew106** Most Recent 6 months, 1 week ago

Selected Answer: A

At first I thought that GCS was the answer but the question does mention that the data is updated frequently. Thereby, it has to be BigTable since we talk about a large amount of data, a streaming application and many individual updates. Storing the data in BigQuery and having to make individual updates doesn't make sense, and neither does running Apache jobs.

If the requirement for updates was not there I would not see any issue with GCS. GCS could serve as a replacement to HDFS and run Hadoop jobs from Dataproc.

upvoted 2 times

✉ **KC_go_reply** 7 months, 2 weeks ago

Selected Answer: A

This scenario screams for BigTable.

It's not B) BigQuery or C) Cloud Storage because both aren't supposed to contain data that is updated frequently. Then, we have to decide between A) BigTable and D) Datastore.

It is A) BigTable because

- it is the most suited for real-time / high-frequency updates
- it is similar to HBase, which is commonly used in Hadoop ecosystem stacks to store streaming / time-series data.

upvoted 1 times

✉ **AmmarFasih** 8 months, 1 week ago

Selected Answer: A

Many here also selected Cloud Storage. But the way I see it BigTable is specifically for low latency, high throughput, mission critical streaming data (financial data is one of them). Also the mentioning of Hadoop that points to HBase functionality if BigTable clarifies the choice more.

upvoted 1 times

✉ **Hisayuki** 9 months ago

Selected Answer: A

BigTable - a No-SQL database but does not support SQL Querying

Apache HBase - Based on Google's BigTable on top of HDFS and you can migrate Hadoop Apps to Cloud BigTable with the HBase API

upvoted 2 times

✉ **izekc** 9 months, 3 weeks ago

Selected Answer: A

A. time series data

upvoted 1 times

✉ **midgoo** 11 months ago

Selected Answer: A

Please note that there is Connector for Bigtable for Hadoop

<https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-bigtable>

upvoted 1 times

✉️  **samdhimal** 1 year ago

Why not Bigquery?

Google BigQuery would be the best option for storing and analyzing large amounts of financial time-series data that is frequently updated and streamed in real-time. It is a fully managed, cloud-native data warehouse that allows you to analyze large datasets using SQL-like queries, and can handle streaming data as well as batch data. Additionally, it can easily integrate with Apache Hadoop to allow your company to run their existing Hadoop jobs in the cloud and gain insights into the data.

upvoted 1 times

✉️  **samdhimal** 1 year ago

A. Google Bigtable is a fully managed, NoSQL, wide-column database that is designed for large scale, low-latency workloads. It is well suited for use cases such as real-time analytics, IoT, and gaming, but it may not be the best fit for storing and analyzing large amounts of financial time-series data that is frequently updated and streamed in real-time. It lacks built-in support for SQL-like queries, which is a standard way of analyzing data in Data Warehousing and Business Intelligence. It is more focused on handling high-performance low-latency workloads, while BigQuery is focused on providing an easy and cost-effective way to analyze large amounts of data using SQL-like queries. Additionally, Bigtable doesn't provide built-in support for running Apache Hadoop jobs, and it would require additional work to integrate it with Hadoop and set it up for data warehousing and Business Intelligence use cases.

upvoted 2 times

✉️  **samdhimal** 1 year ago

C. Google Cloud Storage is an object storage service that allows you to store and retrieve large amounts of unstructured data, such as video, audio, images and other files. It is not a data warehouse and does not provide built-in support for SQL-like queries, which is a standard way of analyzing data in Data Warehousing and Business Intelligence. It would not be suitable for storing and analyzing large amounts of financial time-series data that is frequently updated and streamed in real-time.

D. Google Cloud Datastore is a fully-managed, NoSQL document database that allows you to store, retrieve, and query data. It is not a data warehouse and does not provide built-in support for SQL-like queries, which is a standard way of analyzing data in Data Warehousing and Business Intelligence. It would not be suitable for storing and analyzing large amounts of financial time-series data that is frequently updated and streamed in real-time.

upvoted 2 times

✉️  **samdhimal** 1 year ago

Can someone clarify why Bigtable and Not Bigquery? Super Confused.

upvoted 1 times

✉️  **Oleksandr0501** 9 months, 1 week ago

part2, gpt:

One common approach is to use Google Cloud's Bigtable with other services such as Google Cloud Dataflow, Apache Spark, or Hadoop to perform analysis tasks. These tools can be used to extract data from Bigtable, transform it, and perform analysis tasks such as aggregations, filtering, and machine learning algorithms.

In addition, Bigtable also provides support for secondary indexes and filtering, which can be used to efficiently query and analyze data. The secondary indexes allow you to index specific columns of your data, which makes it easier to search and analyze the data. Filtering can be used to retrieve only the relevant rows of data, reducing the amount of data that needs to be processed and analyzed.

Overall, while Bigtable is not a complete analytics solution on its own, it can be used as a powerful storage backend for analytical workloads, and can be integrated with other tools and technologies to provide a complete analytics solution.

upvoted 1 times

✉️  **desertlotus1211** 1 year ago

<https://cloud.google.com/bigtable/docs/schema-design-time-series>

upvoted 2 times

✉️  **Yazar97** 1 year, 2 months ago

Time series data = Bigtable... So it's A

upvoted 3 times

✉️  **Jay_Krish** 1 year, 2 months ago

Selected Answer: A

Option A seems right

upvoted 1 times

✉️  **drunk_goat82** 1 year, 2 months ago

Selected Answer: A

Big Table has a HBase compliant API and is transactional unlike GCS.

upvoted 1 times

✉  **solar_maker** 1 year, 2 months ago

Selected Answer: A

BigTable can take in data from dataproc, spark and hadoop
https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-bigtable#using_with
upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

Selected Answer: C

It must be C because of the existing Hadoop jobs
upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

On 2nd thought, it's Bigtable: <https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-bigtable>
upvoted 6 times

✉  **pluidust** 1 year, 3 months ago

Selected Answer: C

I think it is C
upvoted 2 times

✉  **maia01** 1 year, 4 months ago

Selected Answer: C

Use Dataproc with Cloud Storage in combo with HDFS
<https://cloud.google.com/dataproc/docs/concepts/dataproc-hdfs>
upvoted 2 times

✉  **euro202** 6 months, 3 weeks ago

Answer is A: Hadoop doesn't mean Dataproc + HDFS. This scenario is about time series that is a use-case for BigTable. Coincidentally BigTable is the best solution for migration of HBase...
upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

Selected Answer: A

A. Cloud Bigtable
upvoted 4 times

An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

- A. Create and share an authorized view that provides the aggregate results.
- B. Create and share a new dataset and view that provides the aggregate results.
- C. Create and share a new dataset and table that contains the aggregate results.
- D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Correct Answer: A

Reference:

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

BigQuery is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near real time.

Giving a view access to a dataset is also known as creating an [authorized view](#) in BigQuery. An authorized view lets you share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query. In this tutorial, you create an authorized view.

Community vote distribution

A (55%)

B (45%)

✉  **midgoo**  11 months ago

Selected Answer: A

A is the answer. Don't be confused by the documentation saying "Authorized views should be created in a different dataset". It is a best practice but not a technical requirement. And we don't create a new dataset for each authorized view. If you are not clear on this, try in the system, do just read the documentation without understanding.

B is wrong when saying we must SHARE Dataset. Although creating a dataset and view in it will not incur extra cost, but sharing dataset is something we always try not to do.

At for the project that run the query it the project to be billed, that is standard behaviour. View only give access to data, whoever run the view need pay for the query cost

upvoted 15 times

✉  **Yiouk** 6 months, 2 weeks ago

Have to consider where the billing goes to:

<https://stackoverflow.com/questions/52201034/bigquery-authorized-view-cost-billing-account>

hence answer is B

upvoted 2 times

✉  **Mathew106** 6 months, 1 week ago

Did you even read the answer in the SO link you shared?

Part of the answer is below:

""After a deeper investigation and some test scenarios, I have confirmed that the billing charges related to the query jobs are applied to the Billing account associated to the project that executes the query; however, the view owner keeps getting the charges related to the storage of the source data."""

Soo, if you create an authorized view, the users from the other project that has access to the view will get billed for the querying.

The only reason to pick up B over A is that it's the recommended approach to store views in a different dataset than the base data.

upvoted 2 times

👤 **DAYAGOWDA** 11 months ago

<https://cloud.google.com/bigquery/docs/authorized-views#:~:text=An%20authorized%20view%20and%20authorized,users%20are%20able%20to%20query>.
upvoted 1 times

👤 **[Removed]** Highly Voted 1 year, 3 months ago

Selected Answer: B

The link on authorized views (<https://cloud.google.com/bigquery/docs/share-access-views>) explicitly states "Authorized views should be created in a different dataset from the source data. That way, data owners can give users access to the authorized view without simultaneously granting access to the underlying data." therefore B is the correct answer because we are to create a new dataset and view within that dataset.

upvoted 13 times

👤 **rocky48** Most Recent 2 months ago

Selected Answer: A

A. Create and share an authorized view that provides the aggregate results.

An authorized view is a BigQuery feature that allows you to share only a specific subset of data from a table, while still keeping the original data private. This way, the organization can expose only the aggregate data to other projects, while still controlling access to the user-level data. B. Using an authorized view, the organization can minimize their overall storage cost as the aggregate data takes up less storage space than the original data. Additionally, by using authorized view, the analysis cost for other projects is assigned to those projects.

upvoted 2 times

👤 **odiez3** 6 months, 1 week ago

I think that is B because for security you need to create a new data set when share a view, apart when you grant access the top level is a data set you share a view in same dataset that you have your tables, that access can see all tables inside dataset.

upvoted 1 times

👤 **baht** 7 months, 2 weeks ago

Selected Answer: B

"Authorized views should be created in a different dataset from the source data. That way, data owners can give users access to the authorized view without simultaneously granting access to the underlying data."

https://cloud.google.com/bigquery/docs/share-access-views?hl=en#console_5

upvoted 2 times

👤 **GoReplyGCPExam** 9 months, 2 weeks ago

Selected Answer: A

B is wrong by itself. why do you need to create a view, if you have already created an aggregated dataset?

upvoted 1 times

👤 **MrMone** 10 months ago

Selected Answer: A

"they need to minimize their overall storage cost". Also, you are sharing the aggregate's results, not the underlying table

upvoted 2 times

👤 **bha11111** 10 months, 3 weeks ago

Selected Answer: A

minimize cost so view

upvoted 1 times

👤 **Paritosh07** 11 months ago

Selected Answer: A

A should be the answer, as we need to separate costs according to projects. As in the following SO question (and the attached google resources), the 'project that runs the queries is the project that gets billed.'

So we can generate a view and give it's access to the other project to run the analysis

<https://stackoverflow.com/questions/52201034/bigquery-authorized-view-cost-billing-account>

upvoted 2 times

👤 **musumusu** 11 months, 2 weeks ago

I will go with A, as I wanna save cost, don't need to create separate dataset for permanent storage.

upvoted 2 times

✉️  **samdhimal** 1 year ago

A. Create and share an authorized view that provides the aggregate results.

An authorized view is a BigQuery feature that allows you to share only a specific subset of data from a table, while still keeping the original data private. This way, the organization can expose only the aggregate data to other projects, while still controlling access to the user-level data. By using an authorized view, the organization can minimize their overall storage cost as the aggregate data takes up less storage space than the original data. Additionally, by using authorized view, the analysis cost for other projects is assigned to those projects.

upvoted 2 times

✉️  **samdhimal** 1 year ago

B. Creating and sharing a new dataset and view that provides the aggregate results is also a correct option but not as optimal as authorized view as it creates a copy of the data and increases the storage costs.

C. Creating and sharing a new dataset and table that contains the aggregate results is also a correct option but not as optimal as authorized view as it creates a copy of the data and increases the storage costs.

D. Creating dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing is not the best option as it would give access to the user-level data, not just the aggregate data.

upvoted 2 times

✉️  **Rupendra06** 1 year ago

Selected Answer: B

Ensure the analysis cost for other projects is assigned to those projects indicates B is the correct answer.

upvoted 2 times

✉️  **GCPpro** 1 year ago

B is the correct answer

upvoted 2 times

✉️  **Kyr0** 1 year, 1 month ago

Selected Answer: A

I would say 1 too

upvoted 1 times

✉️  **slade_wilson** 1 year, 1 month ago

Selected Answer: B

B is the correct approach.

upvoted 2 times

✉️  **wan2three** 1 year, 1 month ago

I think its B, as projects need to be charged separately

upvoted 1 times

✉️  **zellck** 1 year, 1 month ago

Selected Answer: B

B is the answer.

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

After creating your source dataset, you create a new, separate dataset to store the authorized view that you share with your data analysts. In a later step, you grant the authorized view access to the data in the source dataset. Your data analysts then have access to the authorized view but not direct access to the source data.

Authorized views should be created in a different dataset from the source data. That way, data owners can give users access to the authorized view without simultaneously granting access to the underlying data. The source data dataset and authorized view dataset must be in the same regional location.

upvoted 5 times

✉️  **Wonka87** 1 year, 1 month ago

But the wording of option B says create and share a new dataset, do you also need to share dataset apart from authorized view access? If option A, isn't it implicit that authorized view is created on a new dataset and hence option A. B also doesn't mention about Authorized keyword so you may interpret it as normal view which doesn't make sense?

upvoted 1 times

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of data. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

- A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.
- B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
- C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.
- D. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Correct Answer: B

Community vote distribution

B (59%)

D (39%)

2%

✉️  **Mitra123** Highly Voted 2 years, 10 months ago

Keywords here are

1. "Archived": Immutable and hence, BQ and Cloud SQL are ruled out
2. "Auditable": Means track any changes done.

Only D can provide the audibility piece!

I will go with D

upvoted 49 times

✉️  **Jarek7** 9 months ago

I have no idea why so many upvotes on this answer:

- 1) archived doesn't mean immutable and cloud storage is not immutable too.
- 2) auditable means viewable for authorized personnel - and in this case not changes need to be monitored but any access.
- 3) with option D it is easy to go around logging - you can add another access to the bucket read the data remove the access and no one will ever know that you accessed the data.
- 4) option D is much more difficult - you need to application on AppEngine to log the data and provide access for users.
- 5) option D doesn't explain where and how it stores the audit data - it could be accessed and modified from some side app/service.

upvoted 7 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Answer: B

Description: Bigquery is used to analyse access logs, data access logs capture the details of the user that accessed the data
upvoted 21 times

👤 awssp12345 2 years, 6 months ago

The question has no mention of ANALYZE.. BQ is not correct. I would go with D.
upvoted 12 times

👤 sraakesh95 2 years ago

There is no option for archiving with BQ
upvoted 1 times

👤 tavva_prudhvi 1 year, 9 months ago

You dont need to archive the expiring logs, you have to archive the un-archived data here! See the question, it says "Assuming that all expiring logs will be archived correctly", which means they are already stored somewhere like in GCS!!! Hence, better to store the remaining un-archived data in BQ.

upvoted 4 times

👤 vartiklis 1 year, 6 months ago

The question is about where to store the _data_ for which the logs will be generated.

The bit you quoted is about the _logs_ that will be generated when accesssing data. The "archived correctly" implies that proper retention policies will be set up if you choose GCS.

upvoted 3 times

👤 Nandababy Most Recent 1 month, 4 weeks ago

Option B is valid only when analytics to be performed over logs, which is not mentioned anywhere
upvoted 1 times

👤 rocky48 2 months ago

Selected Answer: B

For maintaining an auditable record of access to certain types of data, especially when government regulations are in place, the most suitable option would be:

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

Storing the data in a BigQuery dataset with restricted access ensures control over who can view the data, and utilizing Data Access logs prov a comprehensive audit trail for compliance purposes. This option aligns well with the need for maintaining an auditable record as mandated b government regulations.

upvoted 2 times

👤 Jarek7 9 months ago

Selected Answer: B

If you are going for option D, why do you eliminate option B? The only REAL difference is that for option D you need to develop an app for storing log data and providing bucket link and in option B you have it all done BETTER by GCP. You might also pay a bit more for BQ storage, the question never mentions about cost optimization.

BTW in the D option the bucket is accessible only by AppEngine service, so what will the user do with the provided link? he has no access anyway... And if he even has the access to this link what stops him form using the same link many times? How the AppEngine get and store t information what specific data he accessed and how?

upvoted 7 times

👤 phidelics 7 months, 4 weeks ago

I was about to say the same thing. Why go through that stress?

upvoted 1 times

👤 Kiroo 8 months, 2 weeks ago

That was my thought, either B or D could work but D it's a little bit odd create an app to do something that could be achieved natively gcp
upvoted 2 times

👤 Rodrigo4N 9 months, 2 weeks ago

Selected Answer: D

D amongus

upvoted 2 times

✉️  **juliosb** 10 months, 1 week ago

Selected Answer: B

They want to know where you can store **data** in a way that every access is logged in an auditable way.

Both BQ and GCS have audit logs, except that in alternative D you're circumventing it by creating your own logs. I doubt Google would recommend that.

By types of data you can understand "financial type", "marketing type", etc.

upvoted 3 times

✉️  **midgoo** 11 months ago

Selected Answer: D

I was thinking it should be A. However, 'data' in this question is too vague. It does not say anywhere that the data could fit in BigQuery tables. It could be unstructured data such as videos or images.

Option D seems to involve more setup but it is the only viable option for this scenario. Note that GCS does have Cloud Audit logs. That should be the best option. Maybe this question was asked when Cloud Audit log is not yet available for GCS.

upvoted 4 times

✉️  **aleixfc96** 1 year ago

It is so clear that is B lol

upvoted 1 times

✉️  **NamitSehgal** 1 year ago

B bigquery for a record set store

upvoted 1 times

✉️  **PolyMoe** 1 year ago

Selected Answer: B

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

BigQuery provides built-in logging of all data access, including the user's identity, the specific query run and the time of the query. This log can be used to provide an auditable record of access to the data. Additionally, BigQuery allows you to control access to the dataset using Identity and Access Management (IAM) roles, so you can ensure that only authorized personnel can view the dataset.

upvoted 2 times

✉  **samdhimal** 1 year ago

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

BigQuery provides built-in logging of all data access, including the user's identity, the specific query run and the time of the query. This log can be used to provide an auditable record of access to the data. Additionally, BigQuery allows you to control access to the dataset using Identity and Access Management (IAM) roles, so you can ensure that only authorized personnel can view the dataset.

upvoted 3 times

✉  **Oleksandr0501** 9 months, 1 week ago

gpt: You are correct that option A does not provide an auditable record of access to the data, as it only addresses data security through encryption. Option C provides auditability through Cloud SQL Admin activity logs, but it may not be the best option as it requires additional setup and management.

Option D is a feasible solution, but as you mentioned, it requires additional setup and maintenance of the AppEngine service. It also may not provide a comprehensive audit log of all data access.

Option B, storing the data in a BigQuery dataset that is viewable only by authorized personnel and using the Data Access log to provide auditability, is the most appropriate option as it provides built-in logging of all data access and allows you to control access to the dataset using IAM roles. Therefore, it provides both data security and auditable access to the data. // ok let it be B

upvoted 2 times

✉  **Oleksandr0501** 9 months, 1 week ago

OR MAYBE D....

upvoted 1 times

✉  **Oleksandr0501** 8 months, 3 weeks ago

!!! confused. Give 69% confidence to B, as user Jarek7 explained

upvoted 1 times

✉  **samdhimal** 1 year ago

A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user. is a good option for data security but it does not provide an auditable record of access to the data.

C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability is also a good option for data security but it does not provide an auditable record of access to the data.

D. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket. is also a good option but it requires additional setup and maintenance of the AppEngine service, and it may provide an auditable record of access to the data.

upvoted 2 times

✉  **GCPpro** 1 year ago

D is the correct answer

upvoted 1 times

✉  **RoshanAshraf** 1 year ago

Selected Answer: D

Keys

TYPES of data --> Cloud Storage not BQ

Archival --> Cloud Storage

Access --> No decryption keys to all users

upvoted 1 times

✉  **PrashantGupta1616** 1 year, 1 month ago

Selected Answer: D

I will go with D

upvoted 1 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: D

Keyword, Archiver , certain type of data, auditable, GCS is better option . Durability 11 time 9 to store log immutable for long time.

upvoted 1 times

👤 **zelick** 1 year, 1 month ago

Selected Answer: B

B is the answer.

upvoted 2 times

👤 **Wonka87** 1 year, 1 month ago

are you assuming that data is in BQ compatible format?

upvoted 2 times

Question #77

Topic 1

Your neural network model is taking days to train. You want to increase the training speed. What can you do?

- A. Subsample your test dataset.
- B. Subsample your training dataset.
- C. Increase the number of input features to your model.
- D. Increase the number of layers in your neural network.

Correct Answer: D

Reference:

<https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d>

Neural networks are machine learning algorithms that provide state of the accuracy on many use cases. But, a lot of times the accuracy of the network we are building might not be satisfactory or might not take us to the top positions on the leaderboard in data science competitions. Therefore, we are always looking for better ways to improve the performance of our models. There are many techniques available that could help us achieve that. Follow along to get to know them and to build your own accurate neural network.

Community vote distribution

B (95%)

5%

👤 **mantwosmart**  8 months, 3 weeks ago

Answer: B. Subsample your training dataset.

Subsampling your training dataset can help increase the training speed of your neural network model. By reducing the size of your training dataset, you can speed up the process of updating the weights in your neural network. This can help you quickly test and iterate your model to improve its accuracy.

Subsampling your test dataset, on the other hand, can lead to inaccurate evaluation of your model's performance and may result in overfitting. It is important to evaluate your model's performance on a representative test dataset to ensure that it can generalize to new data.

Increasing the number of input features or layers in your neural network can also improve its performance, but this may not necessarily increase the training speed. In fact, adding more layers or features can increase the complexity of your model and make it take longer to train. It is important to balance the model's complexity with its performance and training time.

upvoted 7 times

✉  **crazycosmos** Most Recent 5 months, 4 weeks ago

Selected Answer: B

B is correct

upvoted 2 times

✉  **Vipul1600** 6 months ago

B should be correct. Increasing the layers can also decrease the training time but may introduce vanishing gradient hence D may not be correct

upvoted 1 times

✉  **email2nn** 9 months ago

answer is B

upvoted 1 times

✉  **juliosb** 10 months, 1 week ago

Selected Answer: B

Reduce training time and probably accuracy too.

upvoted 2 times

✉  **MingSer** 11 months, 1 week ago

Selected Answer: B

all other are wrong

upvoted 1 times

✉  **PolyMoe** 1 year ago

Selected Answer: B

of course !

upvoted 1 times

✉  **samdhimal** 1 year ago

B. Subsampling your training dataset can decrease the amount of data the model needs to process and can speed up training time. However, can lead to decrease in the model's accuracy.

Although it shouldn't matter since we are not even in testing phase yet and we aren't looking for accuracy.

upvoted 2 times

✉  **GCPpro** 1 year ago

B is the answer as we are bothered about speed not the accuracy.

upvoted 2 times

✉  **ler_mp** 1 year ago

Selected Answer: B

The answer is B. Building a more complex model by increasing the number of layer will not reduce the training time.

upvoted 1 times

✉  **slade_wilson** 1 year, 1 month ago

Selected Answer: B

By SubSampling the training data, you will reduce the training time.

In case of D, if you increase the number of layers, then the model's accuracy will be increased. But it will not reduce the time required to train model.

upvoted 4 times

✉  **DGames** 1 year, 1 month ago

Selected Answer: D

Increase speed of the help to train quicker.. option B is sub sample that also help but it drop accurately of model . So I think Option D is good go.

upvoted 1 times

✉  **jin0** 11 months ago

That makes speed of training model lower absolutely. because not only throughput of inference but back-propagation calculation would be increase so, D should be not a answer. there is only answer in those options is B. while it makes dropping performance

upvoted 2 times

✉  **zellick** 1 year, 1 month ago

Selected Answer: B

B is the answer.

upvoted 1 times

✉  **pluiedust** 1 year, 3 months ago

Selected Answer: B

It is B. D would improve the accuracy, not speed.

upvoted 4 times

✉  **Chavez** 1 year, 4 months ago

It's B. D Would be for increase performance

upvoted 1 times

✉  **crismo04** 1 year, 4 months ago

if you Increase the number of layers, you increase the training time, right?

upvoted 1 times

✉  **HarshKothari21** 1 year, 4 months ago

Both B and D seems correct.

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

Increasing D will increase training time

upvoted 3 times

Question #78

Topic 1

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

- A. PigLatin using Pig
- B. HiveQL using Hive
- C. Java using MapReduce
- D. Python using MapReduce

Correct Answer: D

Community vote distribution

A (86%)

14%

✉  **[Removed]**  3 years, 10 months ago

Answer: A

Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

upvoted 22 times

✉  **[Removed]**  3 years, 10 months ago

Should be A

upvoted 15 times

⊕  **AnonymousPanda** Most Recent 5 months, 1 week ago

Selected Answer: A

A as others have said

upvoted 1 times

⊕  **Oleksandr0501** 9 months, 1 week ago

Selected Answer: C

Comment content is too short

upvoted 1 times

⊕  **juliosb** 10 months, 1 week ago

Selected Answer: A

PigLatin is the correct answer, however... the last release was 6 years ago and has lots of bugs.

upvoted 2 times

⊕  **musumusu** 11 months, 2 weeks ago

This answer depends which language you are comfortable with.

Hadoop is your framework, where mapReduce is your Native programming model in JAVA, which is designed to scale, parallel processing, re-pipeline from any checkpoint etc. , So if you are comfortable with JAVA, you can customize your checkpoint at lowlevel in better way. otherwise choose PIG which is another programming concept run over JAVA but then you need to learn this also, if not choose python as it can be deployed with hadoop because hadoop has been making updates for python clients regularly.

Option C: is the best one.

upvoted 5 times

⊕  **samdhimal** 1 year ago

C. Java using MapReduce or D. Python using MapReduce

Apache Hadoop is a distributed computing framework that allows you to process large datasets using the MapReduce programming model. There are several options for writing ETL pipelines to run on a Hadoop cluster, but the most common are using Java or Python with the MapReduce programming model.

upvoted 1 times

⊕  **samdhimal** 1 year ago

A. PigLatin using Pig is a high-level data flow language that is used to create ETL pipelines. Pig is built on top of Hadoop, and it allows you to write scripts in PigLatin, a SQL-like language that is used to process data in Hadoop. Pig is a simpler option than MapReduce but it lacks some capabilities like the control over low-level data manipulation operations.

B. HiveQL using Hive is a SQL-like language for querying and managing large datasets stored in Hadoop's distributed file system. Hive is built on top of Hadoop and it provides an SQL-like interface for querying data stored in Hadoop. Hive is more suitable for querying and managing large datasets stored in Hadoop than for ETL pipelines.

Both Java and Python using MapReduce provide low-level control over data manipulation operations, and they allow you to write custom mapper and reducer functions that can be used to process data in a Hadoop cluster. The choice between Java and Python will depend on the development team's expertise and preference.

upvoted 3 times

⊕  **cetanx** 8 months, 1 week ago

It has to be C

because while Pig can be used to simplify the writing of complex data transformation tasks and can store intermediate results, it doesn't provide the detailed control over checkpointing and pipeline splitting in the way that is typically implied by those terms.

also, while one can write MapReduce jobs in languages other than Java (like Python) using Hadoop Streaming or other similar APIs, it may not be as efficient or as seamless as using Java due to the JVM-native nature of Hadoop.

upvoted 1 times

⊕  **vishal0202** 1 year, 4 months ago

Pig Latin is procedural, where SQL is declarative.

Pig Latin allows pipeline developers to decide where to checkpoint data in the pipeline.

Pig Latin allows the developer to select specific operator implementations directly rather than relying on the optimizer.

Pig Latin supports splits in the pipeline.

Pig Latin allows developers to insert their own code almost anywhere in the data pipeline. link:

<http://maheshwaranm.blogspot.com/2013/07/comparing-pig-latin-and-sql-for.html>

upvoted 1 times

⊕  **Koushik25sep** 1 year, 4 months ago

Selected Answer: A

Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

upvoted 1 times

✉  **BigDataBB** 1 year, 11 months ago

Why not D?

upvoted 1 times

✉  **rbeeraka** 2 years ago

Selected Answer: A

PigLatin supports checkpoints

upvoted 1 times

✉  **davidqianwen** 2 years ago

Selected Answer: A

Answer: A

upvoted 1 times

✉  **maddy5835** 2 years, 3 months ago

Pig is just a scripting language, how pig can be used in creation of pipelines, should be answer from c & D

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 1 times

✉  **kdiab** 2 years, 11 months ago

Found this slideset that puts in favor answer A (pig) :

<https://poloclub.github.io/cx4242-2019fall-campus/slides/17-CSE6242-612-ScalingUp-hive.pdf>

upvoted 2 times

✉  **IsaB** 3 years, 4 months ago

Is this really a question that could appear in Google Cloud Professional Data Engineer Exam? What does it have to do with Google Cloud? I would use DataProc no?

upvoted 10 times

✉  **Pupina** 3 years, 3 months ago

Did you take the exam? I am ready to do it this month

upvoted 1 times

✉  **MaxNRG** 2 years, 1 month ago

seems like a very old question :)

not sure it's actual

upvoted 2 times

✉  **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 1 times

Question #79

Topic 1

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud

Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

- A. Increase the CPU size on your server.
- B. Increase the size of the Google Persistent Disk on your server.
- C. Increase your network bandwidth from your datacenter to GCP.
- D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Correct Answer: C

Community vote distribution

C (70%)

A (30%)

 **[Removed]** Highly Voted 3 years, 10 months ago

correct: C

upvoted 16 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description : Speed of data transfer depends on Bandwidth

Few things in computing highlight the hardware limitations of networks as transferring large amounts of data. Typically you can transfer 1 GB in eight seconds over a 1 Gbps network. If you scale that up to a huge dataset (for example, 100 TB), the transfer time is 12 days. Transferring huge datasets can test the limits of your infrastructure and potentially cause problems for your business.

upvoted 15 times

 **Mathew106** Most Recent 6 months, 1 week ago

Selected Answer: C

We are talking about transfer speed. Network transfer speed does not increase with CPU, but with bandwidth. Since there is no other extra information about what the issue, we have to assume that they imply network transfer speed.

upvoted 1 times

 **Kiroo** 8 months, 2 weeks ago

Selected Answer: C

To be honest this question is incomplete, I would go increasing the bandwidth, but first I would analyze why it's taking long time maybe I'm uploading many files so I could compress and aggregate them and upload just one, maybe the target CPU is overloaded at the time of the upload, maybe the target disk reaching the max IOPS,

upvoted 3 times

 **Jarek7** 9 months ago

Selected Answer: C

Even if transfer server is deployed on the slowest machine available in GCP there is no way it is bottleneck for simple data transfer without any data processing.

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

Selected Answer: A

GPT: Option A, increasing the CPU size on the data transfer server, could potentially increase the transfer speeds if the bottleneck in the data transfer process is the processing power of the server. By increasing the CPU size, the server may be able to process data more quickly, lead to faster transfers.

Option C, increasing the network bandwidth from the datacenter to GCP, could potentially improve the transfer speeds, but it may not be feasible or cost-effective depending on the current infrastructure and network limitations.

upvoted 1 times

✉  **Jarek7** 9 months ago

Please stop using GPT as knowledge source. v3.5 is usually wrong even in simple cases. v4 is much better, but it is not designed to be a knowledge source. Looking at the answer you must have used v3.5. The question says nothing about cost-effectiveness. The issue is data transfer. No data processing is done on the data while it is transferred. Simple transfer doesn't need much processing power - the real bottleneck even on slowest machines available on GCP must be data transfer - it is obvious.

BTW for me GPT3.5 said it is C.

upvoted 3 times

✉  **Oleksandr0501** 8 months, 4 weeks ago

it should be C, for real, bcz nothing said about cost restrictions in the question. And the user "snamburi3" found docs.

upvoted 1 times

✉  **Oleksandr0501** 8 months, 4 weeks ago

yea, i know it can make mistakes. Thank you.
That's why i always mark "GPT" at the start of my answer.

upvoted 1 times

✉  **izekc** 9 months, 1 week ago

Selected Answer: A

it's refer to data transfer server slow here. not transfer data to cloud slow.

100% A

upvoted 1 times

✉  **jonathanthezombieboy** 11 months ago

Selected Answer: C

Answer is C

upvoted 1 times

✉  **jin0** 11 months ago

This question makes people confused only. there is no refer to network or size of data or something could be referred. the answer could be A or C

upvoted 1 times

✉  **Subhajeetpal** 11 months, 1 week ago

Answer is C

upvoted 1 times

✉  **mahdiaqim** 11 months, 1 week ago

Selected Answer: A

Very confusing question. I selected A because I assume increasing the CPU size on the cloud server is easier to change, as a data engineer, than the bandwidth.

upvoted 1 times

✉  **samdhimal** 1 year ago

C. Increase your network bandwidth from your datacenter to GCP.

This will likely have the most impact on transfer speeds as it addresses the bottleneck in the transfer between your data center and GCP. Increasing the CPU size or the size of the Google Persistent Disk on the server may help with processing the data once it has been transferred but will not address the bottleneck in the transfer itself. Increasing the network bandwidth from Compute Engine to Cloud Storage would also help with processing the data once it has been transferred but will not address the bottleneck in the transfer itself as well.

upvoted 3 times

✉  **zelick** 1 year, 1 month ago

Selected Answer: C

C is the answer.

upvoted 1 times

✉  **Nirca** 1 year, 3 months ago

A bit unprofessional question, having performance issues should be addressed by analyzing and looking for saturation in the system and understanding "wait-events". Only than adding more resources.

upvoted 1 times

✉  **rr4444** 1 year, 7 months ago

"The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. "

Question #80

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

⇒ Provide reliable and timely access to data for analysis from distributed research workers

⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in

which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco is building a custom interface to share data. They have these requirements:

1. They need to do aggregations over their petabyte-scale datasets.
2. They need to scan specific time range rows with a very fast response time (milliseconds).

Which combination of Google Cloud Platform products should you recommend?

- A. Cloud Datastore and Cloud Bigtable
- B. Cloud Bigtable and Cloud SQL
- C. BigQuery and Cloud Bigtable
- D. BigQuery and Cloud Storage

Correct Answer: C

Community vote distribution

C (100%)

✉ [Removed] Highly Voted 3 years, 10 months ago

correct: C

upvoted 19 times

✉ atnafu2020 Highly Voted 3 years, 6 months ago

C

Bigquery and Big table =PB storage capacity

Bigtable=to read scan rows Big query select row to read

upvoted 9 times

✉ baht Most Recent 7 months, 2 weeks ago

Selected Answer: C

Response C => Bigquery and bigtable

upvoted 1 times

✉ ga8our 9 months ago

Why not A? If we're already using Bigtable, what's the use of another, slower analytic solution, like BigQuery? Wouldn't Datastore be more useful to store our data than BigQuery?

upvoted 4 times

Question #81

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.

Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

▪

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

⇒ Provide reliable and timely access to data for analysis from distributed research workers

⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualization for operations teams with the following requirements:

⇒ Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)

⇒ The report must not be more than 3 hours delayed from live data.

⇒ The actionable report should only show suboptimal links.

⇒ Most suboptimal links should be sorted to the top.

Suboptimal links can be grouped and filtered by regional geography.

▪

⇒ User response time to load the report must be <5 seconds.

You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

- A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.
- B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
- C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.
- D. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.

Correct Answer: B

Community vote distribution

D (64%)

B (36%)

  **[Removed]** Highly Voted 3 years, 10 months ago

Should be B

upvoted 29 times

✉  **Jarek7** Highly Voted 9 months ago

Selected Answer: D

First I thought B, as D seems too complex with writing app for AppEngine. But B is too simple - just look through the data doesn't seem right. It must be very old question. Today you would load the data to BQ, optionally you can use Dataprep for simple data cleaning or a Dataflow job for more complex data processing, and finally use Looker to create tables and charts.

upvoted 5 times

✉  **Oleksandr0501** 8 months, 4 weeks ago

As per old question - must be. I heard, that the exam will mostly have questions rather from 100 to 205 than from 1 to 100. And some told me that the other w/s gave questions, that happened more often in exam, in comparison to questions given here

upvoted 3 times

✉  **Nirca** Most Recent 3 months, 3 weeks ago

Selected Answer: B

bound to criteria filters that allow value selection. - Simple and Smart.

upvoted 1 times

✉  **PolyMoe** 1 year ago

Selected Answer: D

D.
everything is fixed except data that is updated regularly in order to keep the last 6 weeks. Then, the pipeline does not change ==> obtaining (same) charts and viz on regularly updated data

upvoted 1 times

✉  **hauhau** 1 year, 1 month ago

Selected Answer: B

B

But can someone explain the question and selection clearly?

upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

Selected Answer: B

It's B. All the other choices are unreasonable.

upvoted 4 times

✉  **edwardlin421** 1 year, 2 months ago

ACD-Design for each possible combination of criteria, so if your team has new requirements, you must design new charts.
So, answer should be B.

upvoted 1 times

✉  **ducc** 1 year, 5 months ago

Selected Answer: D

the key is " You want to avoid creating and updating new visualizations each month."

only D works for that phrase

upvoted 2 times

✉  **wan2three** 1 year, 1 month ago

D you might need to load data from source to table for each month. It stated the source will keep last 6 weeks data, but not in D
upvoted 1 times

✉  **KundanK973** 1 year, 7 months ago

must be D

upvoted 1 times

✉  **ealpuche** 1 year, 7 months ago

Selected Answer: D

The answer is B

upvoted 2 times

✉  **rr4444** 1 year, 7 months ago

This Q feels very disconnected from GCP products.....

upvoted 3 times

 **sw52099** 1 year, 8 months ago

Selected Answer: D

Vote D.

Since B just uses "current data", which means if new data enters, you need to re-run those charts again.

upvoted 4 times

 **wan2three** 1 year, 1 month ago

But q says the data sources only have latest 6 weeks data, so current data means latest?

upvoted 1 times

 **RRK2021** 1 year, 11 months ago

B is optimal to avoid creating and updating new visualizations each month

upvoted 1 times

 **ManojT** 2 years, 3 months ago

Answer D: Data in SQL so querying becomes easier on any pattern. create mutiple charts, graphs to fulfill your requirements.

upvoted 4 times

 **sandipk91** 2 years, 5 months ago

Yes it's B

upvoted 2 times

 **sumanshu** 2 years, 7 months ago

vote for B

upvoted 1 times

 **reima990** 3 years, 3 months ago

Question #82

Topic 1

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments " development/test, staging, and production " to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

- A. Create a table called tracking_table and include a DATE column.
- B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
- C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
- D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Correct Answer: B

Community vote distribution

B (100%)

✉️ [Removed] Highly Voted 3 years, 10 months ago

Correct - B

upvoted 19 times

✉️ sspsp Most Recent 6 months, 3 weeks ago

Selected Answer: B

B, Partition tables in BQ have different cost. If a partition is not modified (DML) for 90 days then cost will be less by 50%, while querying will be efficient since its single large table.

upvoted 1 times

✉️ piotrpiskorski 1 year, 2 months ago

Selected Answer: B

always partition large tables

upvoted 1 times

✉️ Thierry_1 2 years, 2 months ago

B for sure

upvoted 3 times

✉️ nguyenmoon 2 years, 4 months ago

Correct is B

upvoted 3 times

✉️ sandipk91 2 years, 5 months ago

Option B for sure

upvoted 2 times

✉️ awssp12345 2 years, 6 months ago

https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard - Supports B

upvoted 2 times

✉️ sumanshu 2 years, 7 months ago

Vote for 'B' Partitioned Table for Faster Query and Low cost (because it will process less data)

upvoted 1 times

✉️ alonsoRios 2 years, 10 months ago

B is correct

upvoted 2 times

✉️ fabianavideos 3 years ago

Question #83

Topic 1

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy

resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

⇒ Databases

- 8 physical servers in 2 clusters

- SQL Server " user data, inventory, static data

- 3 physical servers

- Cassandra " metadata, tracking messages

10 Kafka servers " tracking message aggregation and batch insert

⇒ Application servers " customer front end, middleware for order/customs

- 60 virtual machines across 20 physical servers

- Tomcat " Java services

- Nginx " static content

- Batch servers

⇒ Storage appliances

- iSCSI for virtual machine (VM) hosts

- Fibre Channel storage area network (FC SAN) " SQL server storage

Network-attached storage (NAS) image storage, logs, backups

⇒ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

⇒ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

⇒ Build a reliable and reproducible environment with scaled panty of production.

⇒ Aggregate data in a centralized Data Lake for analysis

⇒ Use historical data to perform predictive analytics on future shipments

⇒ Accurately track every shipment worldwide using proprietary technology

⇒ Improve business agility and speed of innovation through rapid provisioning of new resources

⇒ Analyze and optimize architecture for performance in the cloud

⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

⇒ Handle both streaming and batch data

⇒ Migrate existing Hadoop workloads

⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.

⇒ Use managed services whenever possible

⇒ Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO's tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment. Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

- A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
- B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
- E. Cloud Dataflow, Cloud SQL, and Cloud Storage

Correct Answer: C

Community vote distribution

A (100%)

✉  **digvijay** Highly Voted 3 years, 10 months ago

Seems like A..Data should ingest from multiple sources which might be real time or batch .
upvoted 25 times

✉  **navemula** 2 years, 6 months ago

How is it possible to query in real time with option A. It needs Dataflow
upvoted 3 times

✉  **navemula** 2 years, 6 months ago

To use Dataflow SQL it needs BigQuery
upvoted 2 times

✉  **mikey007** Highly Voted 3 years, 4 months ago

Repeated Question see ques 35
upvoted 11 times

✉  **awssp12345** 2 years, 6 months ago

These exams make people over analyse. People who vote A earlier in 35 seem to be confused here.. haha
upvoted 1 times

✉  **StelSen** 3 years ago

Well Done mikey007, Many people have already answered as A.
upvoted 2 times

✉  **Kyr0** Most Recent 1 year, 1 month ago

Selected Answer: A

Answer is A
upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

Selected Answer: A

It's A
upvoted 1 times

✉  **ducc** 1 year, 5 months ago

Selected Answer: A

A is the answer

upvoted 1 times

✉  **RRK2021** 1 year, 11 months ago

ingest data from a variety of global sources - cloud pub/sub
process and query in real-time - cloud Dataflow
store the data reliably - Cloud Storage

upvoted 2 times

✉  **medeis_jar** 2 years ago

Selected Answer: A

PubSub (for global ingestion from multiple sources) + Dataflow (for process and query) + reliable (gcs).

upvoted 1 times

✉  **lifebegins** 2 years, 2 months ago

Selected Answer: A

using Dataflow you can apply the propriety analytics and you can push the data in to Cloud storage

upvoted 1 times

✉  **gcp_k** 2 years, 3 months ago

Also read the technical requirements section. Not just the last 3 lines of the question.

When you do that, you'll know the answer is PubSub (for global ingestion) + Dataflow (for process and query) + reliable (gcs).

Answer is: A

upvoted 1 times

✉  **ManojT** 2 years, 3 months ago

Answer C:

Look the 3 requirement in the question "ingest data from a variety of global sources, process and query in real-time, and store the data reliably".
Ingest data from global sources: Pub-Sub

Process and Query in realtime: Cloud SQL

Store reliably: Cloud storage

I can understand Databflow is required in case you need to analyze and transform data but question does not refer it.

upvoted 1 times

✉  **cualquienick** 1 year, 7 months ago

Cloud SQL, is not suitable and efficient for storing real time data ingested from PUB/SUB, so A is the answer

upvoted 1 times

✉  **nguyenmoon** 2 years, 4 months ago

Correct is A.

Kafka --> replace by PubSub, Streaming then Dataflow, store data reliably and not mention any other condition then Cloud Storage

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'A'

SQL - will not handle the volume

upvoted 2 times

✉  **daghayeghi** 2 years, 10 months ago

Dataflow SQL cannot output to cloud storage:

<https://cloud.google.com/dataflow/docs/guides/sql/data-sources-destinations>

but the main problem is that Cloud SQL can't do process, then response is A or C.

upvoted 1 times

✉  **kino2020** 3 years, 4 months ago

A

I don't expect this question to come up, but if I had to write the answer, it would be A.

The problem statement "Flowlogic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

As it says, "we cannot determine the data volume", but it doesn't say that we can't calculate it either.

Requirement definition: The system must be able to
ingest data from a variety of global sources
process and query in real-time
Store the data reliably.

It says above, if you look at the Google page.

Logging to multiple systems. for example, a Google Compute Engine instance can write logs to a monitoring system, to a database for later querying, and so on.

<https://cloud.google.com/pubsub/docs/overview#scenarios>

stream processing with Dataflow
<https://cloud.google.com/pubsub/docs/pubsub-dataflow?hl=en-419>

The answer is A, since it is stated above.

upvoted 4 times

✉  **vakati** 3 years, 4 months ago

A. SQL queries can be written in Dataflow too.

<https://cloud.google.com/dataflow/docs/guides/sql/dataflow-sql-intro#running-queries>

upvoted 3 times

✉  **aleedrew** 3 years, 3 months ago

Dataflow SQL cannot output to cloud storage only BigQuery...so I am confused on this one.

upvoted 2 times

✉  **Jay3244** 2 years, 11 months ago

<https://cloud.google.com/pubsub/docs/pubsub-dataflow...> It is possible to load the data to Cloud Storage. Can refer to above docs.

upvoted 1 times

✉  **daghayeghi** 2 years, 10 months ago

he said correct, DataflowDataflow SQL cannot output to cloud storage:

<https://cloud.google.com/dataflow/docs/guides/sql/data-sources-destinations>

upvoted 1 times

✉  **Ral17** 2 years, 4 months ago

Answer should be C then?

upvoted 2 times

✉  **kuntal8285** 3 years, 4 months ago

should be E

upvoted 1 times

✉  **Tanmoyk** 3 years, 4 months ago

Should be A ...data need to feed to the propriorty system and for that dataflow is required.

upvoted 2 times

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison. What should you do?

- A. Select random samples from the tables using the RAND() function and compare the samples.
- B. Select random samples from the tables using the HASH() function and compare the samples.
- C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
- D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Correct Answer: B

Community vote distribution

C (86%)

14%

✉  **rickywck** Highly Voted 3 years, 10 months ago

C is the only way which all records will be compared.
upvoted 32 times

✉  **odacir** 1 year, 1 month ago

Agree with your argument
upvoted 2 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C
Description: Full comparison with this option, rest are comparison on sample which doesnot ensure all the data will be ok
upvoted 16 times

✉  **midgooo** Most Recent 11 months ago

In practice, I will do B. That means it may have error due to randomness. But that is how we normally do validation/QA in general, i.e. we test random samples

In this question, I will do C.

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

key words here- Hash or collect value on "EACH table", after sorting the table.
Option C
upvoted 1 times

✉  **samdhimal** 1 year ago

C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table. This approach will ensure that the data is read in a consistent order, and the hash function will provide a quick and efficient way to compare the contents of the tables and ensure that they are identical.

upvoted 1 times

✉  **samdhimal** 1 year ago

A. Selecting random samples from the tables using the RAND() function may not provide an accurate representation of the data and there is a risk that the comparison will not identify any differences between the tables.

B. Selecting random samples from the tables using the HASH() function may not be an effective method for comparison, as the HASH() function may return different results for equivalent data.

D. Creating stratified random samples using the OVER() function may not provide a comprehensive comparison between the tables as there is a risk that important differences could be missed in the sample data.

upvoted 2 times

✉  **zelick** 1 year, 1 month ago

Selected Answer: C

C is the answer.

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

All records need to be checked to be sure, so C is the answer

upvoted 1 times

✉  **Leeeeee** 1 year, 2 months ago

Selected Answer: C

All records

upvoted 1 times

✉  **hfuihe** 1 year, 3 months ago

Selected Answer: B

B is the only way which all records will be compared.

upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

You must have meant to say C

upvoted 2 times

✉  **medeis_jar** 2 years ago

Selected Answer: C

HASH() to compare data skipping dates and timestamps

upvoted 1 times

✉  **stefanop** 1 year, 9 months ago

The hash in answer C is used to select a sample of the table, not to compare them

upvoted 1 times

✉  **stefanop** 1 year, 9 months ago

Ignore my comment, it was about answer B.

I suggest you to go with answer C which is the only solution comparing all the rows/tables

upvoted 1 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

options A B and D only will determine that it "might" be identical since is only a sample. HASH() can be helpful when doing bulk comparisons, you still have to compare field by field to get the final answer.

The only one left is C which looks good to me

upvoted 2 times

✉  **JayZeeLee** 2 years, 2 months ago

C.

The rest use RAND() at some point, which makes it hard to compare for consistency, unless there's a 'seed' option, which wasn't mentioned.

upvoted 1 times

✉  **u_t_s** 2 years, 3 months ago

Since there is no PK and it is possible that set of values is common in some records which result in same hashkey for those records. But still Answer is C

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'C'

upvoted 1 times

✉  **daghayeghi** 2 years, 11 months ago

B:

Because said migrated to BigQuery, then we don't need Dataproc, and samples don't mean you don't compare all of data.
upvoted 3 times

✉  **yoshik** 2 years, 4 months ago

a sample is a subset of data. then you should assure that the union of the samples contain the data set. Excessively complicated. You migrate to BigQuery but need to check BigQuery output, that is why you should use another tool, Dataproc in this case. Agree that then you should control Dataproc output but suppositions are becoming too many.

upvoted 1 times

✉  **atnafu2020** 3 years, 5 months ago

C

Using Cloud Storage with big data

Cloud Storage is a key part of storing and working with Big Data on Google Cloud. Examples include:

Loading data into BigQuery.

Using Dataproc, which automatically installs the HDFS-compatible Cloud Storage connector, enabling the use of Cloud Storage buckets in parallel with HDFS.

Using a bucket to hold staging files and temporary data for Dataflow pipelines.

For Dataflow, a Cloud Storage bucket is required. For BigQuery and Dataproc, using a Cloud Storage bucket is optional but recommended.

gsutil is a command-line tool that enables you to work with Cloud Storage buckets and objects easily and robustly, in particular in big data scenarios. For example, with gsutil you can copy many files in parallel with a single command, copy large files efficiently, calculate checksums for your data, and measure performance from your local computer to Cloud Storage.

upvoted 3 times

✉  **haroldbenites** 3 years, 5 months ago

C is correct

upvoted 4 times

✉  **haroldbenites** 3 years, 5 months ago

It Says: "...that they are identical." , You must not use sample.

upvoted 3 times

✉  **Rajuuu** 3 years, 6 months ago

C is correct.

Question #85

Topic 1

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for

BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.

What should you do?

- A. Convert your batch BQ queries into interactive BQ queries.
- B. Create an additional project to overcome the 2K on-demand per-project quota.
- C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.
- D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Correct Answer: C

Reference:

<https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery>

You might simply prefer a fixed monthly bill, or encounter workloads that are extremely sensitive to query latency, and thus have predictability and control requirements that cannot be met by the on-demand service. For such situations, you can use the [flat-rate service](#). In this model, a certain number of slots are dedicated to your project(s), and you can establish a hierarchical priority model amongst the projects. The flat-rate model is especially suitable for large enterprises with multiple business units and workloads with varying priorities and budgets. For instance, the arrangement illustrated below gives priority to queries that are issued from the "Dashboarding" project over the queries from the other two projects. But even with prioritization, slots won't be wasted. If the prioritized "Dashboarding" project does not use all its dedicated slots, they'll be distributed among the remaining projects. Even data stored in the "Data Science" project can be queried from the "Dashboarding" project with a higher priority than when it's queried from within the "Data Science" project itself.

Community vote distribution

C (100%)

👤 **sedado77** Highly Voted 1 year, 4 months ago

Selected Answer: C

I got this question on sept 2022. Answer is C
upvoted 8 times

👤 **email2nn** Most Recent 9 months ago

answer us C
upvoted 1 times

Question #86

Topic 1

You have an Apache Kafka cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.

What should you do?

- A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- B. Deploy a Kafka cluster on GCE VM Instances with the Pub/Sub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- C. Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Source connector. Use a Dataflow job to read from Pub/Sub and write to GCS.
- D. Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Sink connector. Use a Dataflow job to read from Pub/Sub and write to GCS.

Correct Answer: A

Community vote distribution

A (100%)

👤 **Ganshank** Highly Voted 3 years, 9 months ago

A.
<https://cwiki.apache.org/confluence/pages/viewpage.action?pageId=27846330>
The solution specifically mentions mirroring and minimizing the use of Kafka Connect plugin.
D would be the more Google Cloud-native way of implementing the same, but the requirement is better met by A.
upvoted 32 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A
Description: Question says mirroring and avoid kafka connect plugins
upvoted 9 times

👤 **Qix** Most Recent 7 months, 3 weeks ago

Pub/Sub Kafka connector requires Kafka Connect, as described here https://cloud.google.com/pubsub/docs/connect_kafka
Deployment of Kafka Connect is explicitly excluded by the requirements. So the only option available is A
upvoted 2 times

✉  **samdhimal** 11 months, 4 weeks ago

Option A: Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

This option involves setting up a separate Kafka cluster in Google Cloud, and then configuring the on-prem cluster to mirror the topics to this cluster. The data from the Google Cloud Kafka cluster can then be read using either a Dataproc cluster or a Dataflow job and written to Cloud Storage for analysis in BigQuery.

upvoted 2 times

✉  **samdhimal** 11 months, 4 weeks ago

Option C: Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Source connector. Use a Dataflow job to read from Pub/Sub and write to GCS.

This option involves deploying the Pub/Sub Kafka connector directly on the on-prem cluster, and configuring it as a source connector. The data from the on-prem Kafka cluster is then sent directly to Pub/Sub, which acts as an intermediary between the on-prem cluster and the data being stored in Google Cloud. A Dataflow job is then used to read the data from Pub/Sub and write it to Cloud Storage for analysis in BigQuery. This option avoids the duplication of data and additional resources required by the other options, making it the preferred option.

upvoted 2 times

✉  **samdhimal** 11 months, 4 weeks ago

Option D: Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Sink connector. Use a Dataflow job to read from Pub/Sub and write to GCS.

This option involves deploying the Pub/Sub Kafka connector on the on-prem cluster, but configuring it as a sink connector. In this case, the data from the on-prem Kafka cluster would be sent directly to Pub/Sub, which would act as the final destination for the data. A Dataflow job would then be used to read the data from Pub/Sub and write it to Cloud Storage for analysis in BigQuery. This option would result in the data being stored in both the on-prem cluster and Pub/Sub, making it less desirable compared to option C, where the data is only stored in Pub/Sub as an intermediary between the on-prem cluster and Google Cloud.

upvoted 1 times

✉  **musumusu** 11 months, 1 week ago

you use chatgpt replies, if you instruct chat gpt that you don't need to use plugins as per question say, it will answer A

upvoted 1 times

✉  **samdhimal** 11 months, 4 weeks ago

Option B: Deploy a Kafka cluster on GCE VM Instances with the Pub/Sub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

This option is similar to Option A, but involves using the Pub/Sub Kafka connector as a sink connector instead of mirroring the topics from on-prem cluster. This option would result in the same duplication of data and additional resources required as Option A, making it less desirable.

upvoted 1 times

✉  **samdhimal** 11 months, 4 weeks ago

Sorry. I messed up. The answer is probably A. My badd....

upvoted 1 times

✉  **zellick** 1 year, 1 month ago

Selected Answer: A

A is the answer.

upvoted 1 times

✉  **Afonya** 1 year, 3 months ago

Selected Answer: A

"The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins."

upvoted 1 times

✉  **somnathmaddi** 1 year, 3 months ago

D is the right answer

upvoted 3 times

✉  **clouditis** 1 year, 4 months ago

D is the right answer

upvoted 2 times

✉  **hendrixlives** 2 years, 1 month ago

Selected Answer: A

"A" is the answer which complies with the requirements (specifically, "The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins"). Indeed, one of the uses of what is called "Geo-Replication" (or Cross-Cluster Data Mirroring) in Kafka is precisely cloud migrations: <https://kafka.apache.org/documentation/#georeplication>

However I agree with Ganshank, and the optimal "Google way" way would be "D", installing the Pub/Sub Kafka connector to move the data from on-prem to GCP.

upvoted 6 times

✉  **gcp_k** 2 years, 3 months ago

Going with "D"

Refer: <https://stackoverflow.com/questions/55277188/kafka-to-google-pub-sub-using-sink-connector>

upvoted 3 times

✉  **baubaumiaomiao** 2 years, 1 month ago

"avoid deployment of Kafka Connect plugins"

upvoted 1 times

✉  **sup12345** 2 years, 1 month ago

you're actually an idiot

upvoted 2 times

✉  **ML_Novice** 1 year, 11 months ago

stop with the swearing shame on you, you should be banned

upvoted 2 times

✉  **JACKYLU** 1 year, 6 months ago

GHETTO TALK

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 1 times

✉  **daghayeghi** 2 years, 10 months ago

Answer: A

Description: Question says mirroring to avoid kafka connect plugins

upvoted 3 times

✉  **Allan222** 2 years, 11 months ago

Correct is D

upvoted 1 times

✉  **sumanshu** 2 years, 6 months ago

As per question - "avoid deployment of Kafka Connect plugins."

upvoted 1 times

✉  **vakati** 3 years, 4 months ago

A.

the best solution would be D but given the restriction here to use mirroring and avoid connectors, A would be the natural choice

upvoted 3 times

✉  **Tanmoyk** 3 years, 4 months ago

D should be the correct answer. Configure pub/sub as sink

upvoted 4 times

✉  **haroldbenites** 3 years, 5 months ago

C is correct.

<https://docs.confluent.io/current/connect/kafka-connect-gcp-pubsub/index.html>

upvoted 2 times

✉  **haroldbenites** 3 years, 5 months ago

Correct Answer: D

Why is this correct?

You can connect Kafka to GCP by using a connector. The 'downstream' service (Pub/Sub) will use a sink connector.

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Question says : avoid deployment of Kafka Connect plugins.

upvoted 2 times

✉  **clouditis** 3 years, 5 months ago

its D, why would google prefer Kafka in their own cert questions! :)

upvoted 3 times

✉  **Ral17** 2 years, 4 months ago

Because the questions mentions to avoid deployment of Kafka connect plugins

upvoted 3 times

You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffling operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.

What should you do?

- A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
- C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
- D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Correct Answer: D

Community vote distribution

D (61%)	A (31%)	8%
---------	---------	----

✉  **rickywck** Highly Voted 3 years, 10 months ago

Should be A:

<https://stackoverflow.com/questions/42918663/is-it-better-to-have-one-large-parquet-file-or-lots-of-smaller-parquet-files>
<https://www.dremio.com/tuning-parquet/>

C & D will improve performance but need to pay more \$\$

upvoted 68 times

✉  **jin0** 11 months ago

you are right C&D will pay more \$. the reason of this questions is shuffling I think. and to reduce shuffling between jobs then make file size larger

upvoted 2 times

✉  **raf2121** 2 years, 6 months ago

Point for discussion - Another reason why it can't be C or D.

SSD's are not available on pre-emptible Worker nodes (answers didn't say whether they wanted to switch from HDD to SSD for Master no
<https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs>

upvoted 8 times

✉  **rr4444** 1 year, 7 months ago

You can have local SSDs for the dataproc normal or preemptible VMs

<https://cloud.google.com/dataproc/docs/concepts/compute/dataproc-pd-ssd>

upvoted 1 times

✉  **raf2121** 2 years, 6 months ago

Also for Shuffling Operations, one need to override the preemptible VMs configuration to increase boot disk size.
(Second half of answer D is correct but first half is wrong)

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

https://cloud.google.com/dataproc/docs/support/spark-job-tuning#limit_the_number_of_files

Store data in larger file sizes, for example, file sizes in the 256MB–512MB range.

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

upvoted 1 times

✉  **madhu1171** Highly Voted 3 years, 10 months ago

Answer should be D

upvoted 12 times

✉  **jvg637** 3 years, 10 months ago

D: # By default, preemptible node disk sizes are limited to 100GB or the size of the non-preemptible node disk sizes, whichever is smaller. However you can override the default preemptible disk size to any requested size. Since the majority of our cluster is using preemptible nodes, the size of the disk used for caching operations will see a noticeable performance improvement using a larger disk. Also, SSD's will perform better than HDD. This will increase costs slightly, but is the best option available while maintaining costs.

upvoted 15 times

✉  **ch3n6** 3 years, 7 months ago

C is correct. D is wrong. they are using 'dataproc and GCS', not related to boot disk at all .

upvoted 2 times

✉  **VishalB** 3 years, 6 months ago

C is recommended only -

If you have many small files, consider copying files for processing to the local HDFS and then copying the results back

upvoted 1 times

✉  **FARR** 3 years, 5 months ago

File sizes are already within the expected range for GCS (128MB-1GB) so not C.

D seems most feasible

upvoted 3 times

✉  **rocky48** Most Recent 1 month, 4 weeks ago

Selected Answer: A

Should be A:

<https://stackoverflow.com/questions/42918663/is-it-better-to-have-one-large-parquet-file-or-lots-of-smaller-parquet-files>

upvoted 1 times

✉  **rocky48** 1 month, 3 weeks ago

Given the scenario and the cost-sensitive nature of your organization, the best option would be:

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job, and copy results back to GCS.

Option C allows you to leverage the benefits of SSDs and HDFS while minimizing costs by continuing to use Dataproc on preemptible VMs. This approach optimizes both performance and cost-effectiveness for your analytical workload on Google Cloud.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Selected Answer: A

<https://stackoverflow.com/questions/42918663/is-it-better-to-have-one-large-parquet-file-or-lots-of-smaller-parquet-files>

Cost effective is the key in the question.

upvoted 1 times

✉  **Nandhu95** 10 months, 1 week ago

Selected Answer: D

Preemptible VMs can't be used for HDFS storage.

As a default, preemptible VMs are created with a smaller boot disk size, and you might want to override this configuration if you are running shuffle-heavy workloads.

upvoted 1 times

✉  **midgoo** 11 months ago

Selected Answer: D

Should NOT be A as:

1. The file size is already at the optimal size
2. If the current file size works well in the current Hadoop, it is expected to have similar performance in Dataproc

The only difference between the current and Dataproc is that Dataproc is using preemptible nodes. So yes, it may incur a bit more cost by using SSD but assuming using the preemptible already save most of it, so we want to save less to improve the performance

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

Optimal size is 1GB

upvoted 1 times

👤 [Removed] 11 months, 1 week ago

Selected Answer: A

Cost sensitive is the keyword.

upvoted 1 times

👤 musumusu 11 months, 2 weeks ago

this question asked by Google, So option C is not correct otherwise, good approach to use initial data in hdfs and switch from HDD to SSDs to 2 non-preemptible node.

Option D is right but they are not mentioning that they will stop using 2 non-preemptible node. but i assume it :P

upvoted 2 times

👤 PolyMoe 1 year ago

Selected Answer: C

C.

ref : https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

- size recommended is 128MB-1GB ==> so it is not size issue ==> not A

- there is no issue mentioned with file format ==> not B

- D. could be a good solution, but requires overriding preemptible VMs. however, the question asks to continue using preemptibles ==> not D

- C. is a good solution.

upvoted 3 times

👤 ayush_1995 1 year ago

agreed C over D as

switching from HDDs to SSDs and overriding the preemptible VMs configuration to increase the boot disk size, may not be the best solution for improving performance in this scenario because it doesn't address the main issue which is the large number of shuffling operations that are causing performance degradation. While SSDs may have faster read and write speeds than HDDs, they may not provide significant performance improvements for a workload that is primarily CPU-bound and heavily reliant on shuffling operations. Additionally, increasing boot disk size of the preemptible VMs may not be necessary or cost-effective for this particular workload.

upvoted 1 times

👤 slade_wilson 1 year, 1 month ago

Selected Answer: D

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

Manage Cloud Storage file sizes

To get optimal performance, split your data in Cloud Storage into files with sizes from 128 MB to 1 GB. Using lots of small files can create a bottleneck. If you have many small files, consider copying files for processing to the local HDFS and then copying the results back.

Switch to SSD disks

If you perform many shuffling operations or partitioned writes, switch to SSDs to boost performance.

upvoted 2 times

👤 odacir 1 year, 1 month ago

Selected Answer: D

Its D 100%. It's the recommended best practice for this scenario.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

upvoted 3 times

👤 zellick 1 year, 1 month ago

Selected Answer: D

D is the answer.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#switch_to_ssd_disks

If you perform many shuffling operations or partitioned writes, switch to SSDs to boost performance.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#use_preemptible_vms

As a default, preemptible VMs are created with a smaller boot disk size, and you might want to override this configuration if you are running shuffle-heavy workloads. For details, see the page on preemptible VMs in the Dataproc documentation.

upvoted 1 times

👤 sfsdeniso 1 year, 2 months ago

Answer id D

not C because cannot use HDFS with preemptible VMs

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#use_preemptible_vms

upvoted 2 times

✉  **dish11dish** 1 year, 2 months ago

Selected Answer: D

Option D is correct

Elimination Strategy:-

- A. Increase the size of your parquet files to ensure them to be 1 GB minimum (doesn't make sense as the file size are fit for migration to proceed with given scenario, recommended size is between 128 MB to 1 GB.)
- B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files(doesn't make sense to make changes to file format)
- C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS(doesn't make sense to copy the file from GCS to HDFS as the workload that consists of many shuffling operations)
- D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size(perfect fit as the workload that consists of many shuffling operations which requires attention to increase the performance reference doc:- https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance)

upvoted 3 times

✉  **piotrpiskorski** 1 year, 2 months ago

Selected Answer: A

it's A.

Larger parquet files will be more efficient and it's a non-cost solution to implement in contrary to the SSD drives.

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

Recommended file size is not 1GB.

https://cloud.google.com/dataproc/docs/support/spark-job-tuning#limit_the_number_of_files

Store data in larger file sizes, for example, file sizes in the 256MB–512MB range.

upvoted 1 times

✉  **gudiking** 1 year, 2 months ago

Selected Answer: A

<https://www.dremio.com/blog/tuning-parquet/>

upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

Selected Answer: A

It's A.

D doesn't make sense because Spark does shuffling in memory, and in any case, it has nothing to do with the BOOT disk size.

upvoted 1 times

✉  **NicolasN** 1 year, 2 months ago

Not True

"As a default, preemptible VMs are created with a smaller boot disk size, and you might want to override this configuration if you are running shuffle-heavy workloads."

🔗 https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#use_preemptible_vms

Moreover:

"If you perform many shuffling operations or partitioned writes, switch to SSDs to boost performance."

🔗 https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#switch_to_ssd_disks

upvoted 3 times

Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).

What should you do?

- A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
- B. Add a `try` catch block to your DoFn that transforms the data, extract erroneous rows from logs.
- C. Add a `try` catch block to your DoFn that transforms the data, write erroneous rows to Pub/Sub PubSub directly from the DoFn.
- D. Add a `try` catch block to your DoFn that transforms the data, use a sideOutput to create a PCollection that can be stored to Pub/Sub later.

Correct Answer: C

Community vote distribution

D (81%)

C (19%)

✉️  **midgoo**  11 months ago

Selected Answer: D

C is a big NO. Writing to PubSub in DoFn will cause bottleneck in the pipeline. For IO, we should always use those IO lib (e.g PubsubIO) Using sideOutput is the correct answer here. There is a Qwiklab about this. It is recommended to do that lab to understand more.
upvoted 10 times

✉️  **jonathanthezombieboy**  11 months, 2 weeks ago

Selected Answer: D

Based on the given scenario, option D would be the best approach to improve the reliability of the pipeline.

Adding a try-catch block to the DoFn that transforms the data would allow you to catch and handle errors within the pipeline. However, storing erroneous rows in Pub/Sub directly from the DoFn (Option C) could potentially create a bottleneck in the pipeline, as it adds additional I/O operations to the data processing.

Option A of filtering the erroneous data would not allow the pipeline to reprocess the failing data, which could result in data loss.

Option D of using a sideOutput to create a PCollection of erroneous data would allow for reprocessing of the failed data and would not create bottleneck in the pipeline. Storing the erroneous data in a separate PCollection would also make it easier to debug and analyze the failed data.

Therefore, adding a try-catch block to the DoFn that transforms the data and using a sideOutput to create a PCollection of erroneous data that can be stored to Pub/Sub later would be the best approach to improve the reliability of the pipeline.

upvoted 7 times

✉  **Mathew106** Most Recent 6 months, 1 week ago

Selected Answer: C

Answer is C. Here is the github repo and an example from the Qwiklab where they tag the output as 'parsed_rows' and 'unparsed_rows' before they send the data to GCS. I don't see how GCS or PubSub would make a difference at this point. It seems like a more maintainable solution to just parse the data in the DoFn.

1) If the function does more than that then it serves multiple purposes and it's not good software engineering. Unless there is a good reason, writing to PubSub should be separated from the DoFn.

ii) It's faster to write in mini-batches or one batch than stream the errors. What's the need for streaming out errors 1 by 1? Literally no real advantage.

https://github.com/GoogleCloudPlatform/training-data-analyst/blob/master/quests/dataflow_python/7_Advanced_Streaming_Analytics/solution/streaming_minute_traffic_pipeline.py
upvoted 1 times

✉  **tibuenoc** 10 months, 1 week ago

Selected Answer: D

Output errors to new PCollection – Send to collector for later analysis (Pub/Sub is a good target)

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Option D is right approach to use to get errors as sideOutput. Apache beam has a special scripting docs not dynamic as python itself. So let's follow standard sideOutput(withoutputs in the code)
syntax be like in pipeline:

'ProcessData' >> beam.ParDo(DoFn).with_outputs

upvoted 2 times

✉  **musumusu** 11 months, 1 week ago

After using you try: Catch: you can also send the erroneous records to dead letter sink into BQ
``` outputTuple.get(deadLetterTag).apply(BigQuery.write(...)) ````

upvoted 1 times

✉  **abwey** 11 months, 3 weeks ago

**Selected Answer: D**

blahblahblahblahblahblahblahblah

upvoted 3 times

✉  **waiebdi** 1 year ago

**Selected Answer: D**

It's D.

Use a try catch block to direct erroneous rows into a side output. The PCollection of the side output can be sent efficiently to the PubSub topic via Apache Beam PubSubIO.

It's not C because C means to send every single invalid row in a separate request to PubSub which is very inefficient when working with Dataflow as now batching is involved.

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

upvoted 1 times

✉  **hauhau** 1 year, 1 month ago

C

D: dataflow to pub/sub is weird

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

D

Side output is a great manner to branch the processing. Let's take the example of an input data source that contains both valid and invalid values. Valid values must be written in place #1 and the invalid ones in place #2. A naive solution suggests to use a filter and write 2 distinct processing pipelines. However this approach has one main drawback - the input dataset is read twice. If for the mentioned problem we use side outputs, we can still have 1 ParDo transform that internally dispatches valid and invalid values to appropriate places (#1 or #2, depending on value's validity).

<https://www.waitingforcode.com/apache-beam/side-output-apache-beam/read#:~:text=simple%20test%20cases.-,Side%20output%20defined,-%C2%B6>

upvoted 3 times

✉  **sfsdeniso** 1 year, 2 months ago

Answer is D

upvoted 1 times

✉  **cloudmon** 1 year, 2 months ago

**Selected Answer: C**

It's C.

In D, "storing to PubSub later" doesn't really make sense.

upvoted 2 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: C**

Answer is C. You need to reprocess all the failing data, and yes, you can use PubSub as a sink, according to the documentation:

<https://beam.apache.org/documentation/io/connectors/>

upvoted 2 times

✉  **nickyshil** 1 year, 4 months ago

Answer C

upvoted 4 times

✉  **nickyshil** 1 year, 4 months ago

The error records are directly written to PubSub from the DoFn (it's equivalent in python).

You cannot directly write a PCollection to PubSub. You have to extract each record and write one at a time. Why do the additional work and not write it using PubSubIO in the DoFn itself?

You can write the whole PCollection to Bigquery though, as explained in

Reference:

<https://medium.com/google-cloud/dead-letter-queues-simple-implementation-strategy-for-cloud-pub-sub-80adf4a4a800>

upvoted 6 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

Question #89

Topic 1

You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.

What should you do?

- A. Provide latitude and longitude as input vectors to your neural net.
- B. Create a numeric column from a feature cross of latitude and longitude.
- C. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L1 regularization during optimization.
- D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

**Correct Answer: B**

Reference:

<https://cloud.google.com/bigquery/docs/gis-data>

## Loading geospatial data

Single points on Earth can be described by just a longitude, latitude pair. For example, you can load a CSV file that contains longitude and latitude values and then use the `ST_GEOPOINT` function to convert them into `GEOGRAPHY` values.

For more complex geographies, you can load the following geospatial data formats into a `GEOGRAPHY` column:

- Well-known text (WKT)
- Well-known binary (WKB)
- GeoJSON

*Community vote distribution*

C (84%)

Other

 **AHUI** Highly Voted 1 year, 4 months ago

Ans C, use L1 regularization because we know the feature is a strong feature. L2 will evenly distribute weights upvoted 8 times

 **dish11dish** Highly Voted 1 year, 2 months ago

**Selected Answer: C**

Option C is correct

Use L1 regularization when you need to assign greater importance to more influential features. It shrinks less important feature to 0.

L2 regularization performs better when all input features influence the output & all with the weights are of equal size.

upvoted 7 times

 **uday\_examtopic** Most Recent 4 months, 1 week ago

Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

Like option C, we bucketize at the minute level, but this time we apply L2 regularization. L2 regularization, or Ridge Regression, discourages large values of weights in the model without forcing them to become sparse. It can help prevent overfitting, especially when we have a large number of features (as a result of bucketizing and crossing).

Given the options, D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization seems to be the most appropriate. Bucketizing at the minute level captures localized patterns, and L2 regularization can help control the complexity of the model without enforcing sparsity.

upvoted 1 times

 **ckanaar** 4 months, 1 week ago

What does bucketizing at the minute level mean in the context of this question?

upvoted 2 times

 **Surely1987** 2 months, 3 weeks ago

Coordinates are written with Degrees, minutes and seconds (one minute being equal to about 1.8 km). So you group your coordinates in buckets with a minute precision

upvoted 2 times

 **FP77** 5 months ago

**Selected Answer: B**

I strongly believe it's B.

upvoted 1 times

✉  **Mathew106** 6 months, 1 week ago

**Selected Answer: B**

The right answer is B. What the hell does bucketize the feature cross of latitude and longitude even mean? They are not a time feature. C and D don't even make sense. The L1 regularization is something that doesn't answer anything in the question. The only valid feature engineered here is option B. A is not an engineered feature.

Create a feature cross of latitude and longitude, bucketize it at the minute level and use L1 regularization during optimization.

upvoted 1 times

✉  **Jojo9400** 6 months, 3 weeks ago

D

You have to use L2, since you have to create a new variable with two already existing the risk of multicollinearity is high, L1 is good for selecting feature to avoid curse of dimensionality not for multicollinearity

upvoted 1 times

✉  **ga8our** 9 months ago

Why not L2? L2 (Ridge) uses a squared value coefficient as a penalty term to the loss function, while L1 (Lasso) uses an absolute value coefficient. Isn't a squared penalty stronger than an absolute one?

<https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c>

upvoted 1 times

✉  **ckanaar** 4 months, 1 week ago

L1 regression forces unimportant coefficients to zero. Since the location is extremely important, L1 will force less important coefficients to zero, thereby further increasing the importance of the location coefficient.

upvoted 1 times

✉  **Oleksandr0501** 9 months, 1 week ago

gpt: Option C and D suggest bucketizing the feature cross of latitude and longitude at the minute level and using L1 or L2 regularization during optimization. While regularization can help prevent overfitting, bucketizing at such a granular level may not be necessary and could lead to overfitting. It's also not clear how bucketizing at the minute level would capture the spatial relationship between the latitude and longitude features.

upvoted 2 times

✉  **PolyMoe** 1 year ago

**Selected Answer: D**

D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization. This will create a new feature that captures the physical dependency of the location of the property on the price, and bucketizing it at the minute level will reduce the number of unique values and prevent overfitting. L2 regularization will also help to prevent overfitting by penalizing large weights in the model.

upvoted 1 times

✉  **cetanx** 8 months, 1 week ago

chat-gpt also says D  
explanation:

This approach effectively creates a grid of the geographical area in your data, allowing the model to learn weights for each grid cell (bucket). This helps capture the spatial relationship between latitude and longitude, which can be crucial for real estate prices. Additionally, using L2 regularization helps prevent overfitting by discouraging complex models, which can be particularly important when working with high-dimensional crossed features.

upvoted 1 times

✉  **zelick** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

<https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture>

A feature cross is a synthetic feature formed by multiplying (crossing) two or more features. Crossing combinations of features can provide predictive abilities beyond what those features can provide individually.

<https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization>

upvoted 3 times

✉  **crismo04** 1 year, 4 months ago

<https://medium.com/iga-data-science-club/geographic-coordinate-encoding-with-tensorflow-feature-columns-e750ae338b7c#:~:text=to%20the%20rescue!-,Feature%20Crosses,-Combining%20features%20into>  
upvoted 2 times

✉  **crismo04** 1 year, 4 months ago

Feature cross seems to be the right feature option  
upvoted 1 times

✉  **crismo04** 1 year, 4 months ago

So it's B option  
upvoted 3 times

✉  **[Removed]** 1 year, 4 months ago

**Selected Answer: C**

Regularization + location into one  
upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: C**

C. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L1 regularization during optimization.  
upvoted 5 times

Question #90

Topic 1

You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.

What should you do?

- A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
- B. Place the MariaDB instances in an Instance Group with a Health Check.
- C. Install the StackDriver Logging Agent and configure fluentd in\_tail plugin to read MariaDB logs.
- D. Install the StackDriver Agent and configure the MySQL plugin.

**Correct Answer: C**

*Community vote distribution*

D (83%)

A (17%)

✉  **Barniyah** Highly Voted 3 years, 9 months ago

Answer : A

MariaDB needs costume metrics , and stackdriver built-in monitoring tools will not provide these metrics . Opencensus Agent will do this for you . For more info , refer to :

<https://cloud.google.com/monitoring/custom-metrics/open-census>

upvoted 26 times

✉  **fire558787** 2 years, 4 months ago

It is definitely A.

B: can't be because Health Checks just checks that machine is online

C: StackDriver Logging is for Logging. Here we talk of Monitoring / Alerting

D: StackDriver Agent monitors default metrics of VMs and some Database stuff with the MySQL Plugin. Here you want to monitor some metrics like Replication of MariaDB (I didn't find anything of this sort in the plugin page), and you may want to use Custom Metrics rather than default metrics. "Cloud Monitoring automatically collects more than 1,500 built-in metrics from more than 100 monitored resources. These metrics cannot capture application-specific data or client-side system data. Those metrics can give you information on backend latency or disk usage, but they can't tell you how many background routines your application spawned."

[https://cloud.google.com/monitoring/custom-metrics/open-census#monitoring\\_opencensus\\_metrics\\_quickstart-python](https://cloud.google.com/monitoring/custom-metrics/open-census#monitoring_opencensus_metrics_quickstart-python)

upvoted 13 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Answer: C

Description: The GitHub repository named google-fluentd-catch-all-config which includes the configuration files for the Logging agent for ingesting the logs from various third-party software packages.

upvoted 13 times

👤 Atulthakur 2 years, 5 months ago

I think its D, because its Selfmanaged DB and for this we use Stackdriver Agents. and in this question its asking about metrics not logs.

upvoted 2 times

👤 rocky48 Most Recent 1 month, 3 weeks ago

Selected Answer: D

Here's the rationale:

StackDriver Agent: The StackDriver Agent is designed to collect system and application metrics from virtual machine instances and send them to StackDriver Monitoring. It simplifies the process of collecting and forwarding metrics.

MySQL Plugin: The StackDriver Agent has a MySQL plugin that allows you to collect MySQL-specific metrics without the need for additional custom development. This includes metrics related to network connections, disk IO, and replication status – which are the specific metrics you mentioned.

Option D is the most straightforward and least development-intensive approach to achieve the monitoring and alerting requirements for MariaDB on GCE VM Instances using StackDriver.

upvoted 1 times

👤 BlehMaks 2 months, 3 weeks ago

Selected Answer: A

replication status seems to be not included in sql agent metrics. but I do not like A in terms of efforts

upvoted 1 times

👤 ninjatech 10 months, 4 weeks ago

it can't be A as it saying minimal development but for opencensus the development is needed.

upvoted 1 times

👤 slade\_wilson 1 year, 1 month ago

Selected Answer: A

To use metrics collected by OpenCensus in your Google Cloud project, you must make the OpenCensus metrics libraries and the Stackdriver exporter available to your application. The Stackdriver exporter exports the metrics that OpenCensus collects to your Google Cloud project. You can then use Cloud Monitoring to chart or monitor those metrics.

upvoted 1 times

👤 zellck 1 year, 1 month ago

Selected Answer: D

D is the answer.

<https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/third-party/mariadb>

upvoted 8 times

👤 wan2three 1 year, 1 month ago

For supplement, 'Stackdriver agent' now called as Ops agent, 'Operations Suite'

upvoted 3 times

👤 dish11dish 1 year, 2 months ago

Selected Answer: D

Option D is Correct

MariaDB is a community-developed, commercially supported fork of the MySQL relational database management system (RDBMS). To collect logs and metrics for MariaDB, use the mysql receivers.

The mysql receiver connects by default to a local MariaDB server using a Unix socket and Unix authentication as the root user.

reference:-<https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/third-party/mariadb>

upvoted 3 times

👤 girgu 1 year, 3 months ago

Selected Answer: D

<https://cloud.google.com/monitoring/agent/ops-agent/third-party/mariadb>

upvoted 2 times

✉  **clouditis** 1 year, 4 months ago

C is the answer, fluentd plug in is needed as the DB is on GCE  
upvoted 2 times

✉  **ducc** 1 year, 5 months ago

**Selected Answer: D**

go for D  
upvoted 2 times

✉  **eRaymox** 1 year, 5 months ago

A

StackDriver Agent monitors default metrics of VMs and some Database stuff with the MySQL Plugin. Here you want to monitor some more custom stuff like Replication of MariaDB (I didn't find anything of this sort in the plugin page), and you may want to use Custom Metrics rather than default metrics. "Cloud Monitoring automatically collects more than 1,500 built-in metrics from more than 100 monitored resources. But those metrics cannot capture application-specific data or client-side system data. Those metrics can give you information on backend latency, disk usage, but they can't tell you how many background routines your application spawned." [https://cloud.google.com/monitoring/custom-metrics/open-census#monitoring\\_opencensus\\_metrics\\_quickstart-python](https://cloud.google.com/monitoring/custom-metrics/open-census#monitoring_opencensus_metrics_quickstart-python)

upvoted 2 times

✉  **Kriegs** 1 year, 7 months ago

I'm not 100% sure as I have no experience with that issue, but I would say it's D - both A and D should work, but the keyword is "with minimum development effort" (and using pre-built plugin > creating custom metric in terms of simplicity, that's obvious) and all of the relevant data (as per question) should be there: [https://cloud.google.com/monitoring/api/metrics\\_agent#agent-mysql](https://cloud.google.com/monitoring/api/metrics_agent#agent-mysql)

I'm not sure if C would work, but it also seems more advanced in implementation than D. B is 100% wrong and insufficient for that use case.

Feel free to prove me wrong :)

upvoted 1 times

✉  **NR22** 1 year, 9 months ago

A and D both seem like viable options here, unsure which is Google's preferred method as that would be deemed the correct answer in the exam. Any opinions?

upvoted 1 times

✉  **Didine\_22** 1 year, 9 months ago

**Selected Answer: D**

D

mariadb is an extension of mysql and mysql plugin must work fine to extract the metrics of mariadb.

upvoted 3 times

✉  **ST42** 1 year, 8 months ago

"MariaDB is a community-developed, commercially supported fork of the MySQL relational database management system (RDBMS). To collect logs and metrics for MariaDB, use the mysql receivers."

<https://cloud.google.com/monitoring/agent/ops-agent/third-party/mariadb>

upvoted 1 times

✉  **rbeeraka** 2 years ago

**Selected Answer: A**

Opencensus Agent is right one

upvoted 2 times

✉  **nguyenmoon** 2 years, 4 months ago

D is correct. Answer : D

mariadb is an extension of mysql and mysql plugin must work fine to extract the metrics of mariadb.

upvoted 4 times

✉  **[Removed]** 2 years ago

agree; <https://cloud.google.com/monitoring/agent/ops-agent/third-party/mariadb#configure-instance>

upvoted 2 times

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.

What should you do?

- A. Increase the size of the dataset by collecting additional data.
- B. Train a linear regression to predict a credit default risk score.
- C. Remove the bias from the data and collect applications that have been declined loans.
- D. Match loan applicants with their social profiles to enable feature engineering.

**Correct Answer: B**

*Community vote distribution*

B (57%)	D (23%)	C (20%)
---------	---------	---------

✉  **GHN74** Highly Voted 3 years, 5 months ago

A is incorrect as you need to work with the data you have available  
 C is an optimisation not a solution  
 D is ethically incorrect and invasion to privacy, there could be several legal implications with this  
 B although oversimplified but is a workable solution

upvoted 36 times

✉  **sergio6** 2 years, 3 months ago

Information in social profiles are public  
 upvoted 1 times

✉  **sergio6** 2 years, 3 months ago

according to the privacy settings and shareable informations  
 upvoted 1 times

✉  **sumanshu** Highly Voted 2 years, 9 months ago

We have labelled data that contains whether a loan application is accepted or defaulted - So Classification Problem Data.

We need to predict (Default Rates for applicants) - I think whether application will be granted or defaulted. - So Binary Classification.

No option matches the answer. - if we mark 'B' - It should be Logistic Regression, Instead of Linear Regression

upvoted 20 times

✉  **szefco** 2 years, 1 month ago

Question says: "to predict default RATES for credit applicants".  
 It is not binary classification, so Linear Regression would work here.  
 I think B is correct answer.

upvoted 20 times

✉  **cchen8181** 8 months, 2 weeks ago

Correct approach is to use logistic regression to predict default/not default, and then take the confidence/probability of the outcome as "default rate". Linear regression doesn't make sense since we are not given a default rate label in our data, we are just given the labels default vs no default.

upvoted 1 times

✉  **Aaronn14** 10 months, 4 weeks ago

You cannot predict rate. You predict a realization, which is either default or not. This question is terribly written.  
 upvoted 2 times

TVH\_Data\_Engineer Most Recent 1 month, 1 week ago

**Selected Answer: B**

To predict default rates for credit applicants using the labeled dataset of granted loan applications, the most appropriate course of action would be:

B. Train a linear regression to predict a credit default risk score.

Here's the rationale for this approach:

**Appropriate Model for Prediction:** Linear regression is a common statistical method used for predictive modeling, particularly when the outcome variable (in this case, the likelihood of default) is continuous. In the context of credit scoring, linear regression can be used to predict a risk score that represents the probability of default.

**Utilization of Labeled Data:** Since you already have a labeled dataset containing information on loans that have been granted and whether they have defaulted, you can use this data to train the regression model. This historical data provides the model with examples of borrower characteristics and their corresponding default outcomes.

upvoted 1 times

rocky48 1 month, 3 weeks ago

**Selected Answer: B**

B. Train a linear regression to predict a credit default risk score.

upvoted 1 times

gaurav0480 5 months, 2 weeks ago

What would be the target variable if B is correct i.e. training a linear regression model? Default/No-Default is a categorical variable one cannot train a linear regression model with this target variable

upvoted 1 times

FDS1993 6 months ago

**Selected Answer: C**

C - it is a typical approach in credit loans.

Keeping only the accepted loans leads to a bias in the application

upvoted 1 times

Mathew106 6 months, 1 week ago

**Selected Answer: B**

Linear regression is not the good way to solve such a problem, but you can totally apply linear regression to solve a classification problem. Just set the labels to numeric values 0 and 1 and linear regression will try to predict a value inbetween and round to the closest label (0 or 1).

Totally not the way to go about it, but actually it's possible.

upvoted 1 times

juliosb 10 months, 1 week ago

**Selected Answer: D**

Cannot be B. This is logistic regression, not linear regression.

D is the only acceptable option.

Social profile can include things like high or low income, for example.

When you apply for a credit you usually have to give this information, so totally legal.

upvoted 1 times

Oleksandr0501 9 months, 1 week ago

D. Matching loan applicants with their social profiles to enable feature engineering is not recommended as it raises privacy concerns and may not be legal in some jurisdictions. Additionally, social profiles may not be a good indicator of creditworthiness, and relying on them may introduce bias or discrimination.

upvoted 1 times

jin0 11 months ago

Because there is no option to know what dataset schema is even though B is needed for this question's purpose Nobody can't select B. So it is none to answer

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

all options are wrong: Still in favour of B

A: ofc its good to have more data but its not clear how much data we have

B: Linear can be a workable approach but current situation is not for linear approach, decision tree, random forest etc can be good for it.

C: DAta should be unbiased, removing bias is negative for tranining

upvoted 1 times

✉  **Besss** 1 year ago

**Selected Answer: B**

default rates can be predicted with linear regression.

upvoted 1 times

✉  **21c17b3** 1 month, 3 weeks ago

default rates is classification probability

upvoted 1 times

✉  **Whoswho** 1 year, 1 month ago

Answer should actually be a logistic Regression model

upvoted 2 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

upvoted 2 times

✉  **ladistar** 1 year, 1 month ago

The question asks about default RATES, as in you are predicting a continuous variable, not a discrete one (classification). This is a regression problem, so choice B.

upvoted 2 times

✉  **woyaoi** 1 year, 2 months ago

I used to be a Credit Risk modeler and I think this question is stupid.

upvoted 8 times

✉  **Azlijaffar** 1 year, 3 months ago

Data that you have is binary - Defaulted or not. You want default rates - Linear Regression. HOWEVER. The data that you have is for "ALREADY GRANTED" loan applications and whether they have defaulted or not.

But you want to "train a model to predict default rates for credit applicants", which would include applicants who would not be granted the loan. If you just work on that dataset your model will not be as accurate as it won't have considered profiles of applicants that normally would not be granted those loans in the first place. Am I missing something here?

upvoted 2 times

✉  **Azlijaffar** 1 year, 3 months ago

So answer should be C i think.

upvoted 2 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: B**

Answer B.

You can't assume it's a classification model, since the label is not specified. The "whether these applications have been defaulted" can be used just as a categorial variable (dummy) that impacts in the final numeric value, what is the default rate.

upvoted 1 times

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

- A. Cloud Spanner
- B. Cloud Bigtable
- C. Cloud Firestore
- D. Cloud SQL

**Correct Answer: D**

*Community vote distribution*

D (100%)

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer - D

upvoted 24 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: D

Description: Cloud SQL cheap and relational DB.

upvoted 13 times

 **midgoo** Most Recent 11 months ago

Selected Answer: D

Cloud SQL: max storage for shared core = 3TB and for dedicated core = up to 64TB

Only use Spanner if we need autoscale (Note that Cloud SQL could scale too but not automatic yet) or the size is too big (as above) or 4/5 9s (Cloud SQL is only 99.95)

upvoted 2 times

 **zellck** 1 year, 1 month ago

Selected Answer: D

D is the answer.

upvoted 1 times

 **Nirca** 1 year, 3 months ago

Selected Answer: D

Cloud SQL is relational DB (pg mssql, mysql)

upvoted 1 times

 **Dhass** 1 year, 7 months ago

Answer - D

upvoted 1 times

 **homaj** 1 year, 8 months ago

Selected Answer: D

answer D

upvoted 1 times

⊕  **sumanshu** 2 years, 7 months ago

Vote for D

upvoted 3 times

⊕  **daghayeghi** 2 years, 10 months ago

D:

<https://cloud.google.com/sql/docs/features>

upvoted 1 times

⊕  **GypsyMonkey** 3 years, 1 month ago

D, cloud SQL is a relational database; if > 10tb, then choose spanner

upvoted 3 times

⊕  **atnafu2020** 3 years, 5 months ago

D

Cloud SQL supports MySQL 5.6 or 5.7, and provides up to 624 GB of RAM and 30 TB of data storage, with the option to automatically increase the storage size as needed.

upvoted 3 times

⊕  **Abhi16820** 2 years, 2 months ago

64TB AS OF TODAY

upvoted 5 times

⊕  **haroldbenites** 3 years, 5 months ago

D is correct. Obviously

upvoted 3 times

⊕  **Barniyah** 3 years, 9 months ago

But cloud SQL storage is limited to several hundreds of GB's for all instances and we need 2TB.

So, Cloud spanner is much closer to this, with the exception of the cost

upvoted 3 times

⊕  **taeypyung** 3 years, 9 months ago

At this moment, Cloud SQL is providing up to 30,720GB(about 30TB)

So I think it's D.

upvoted 13 times

⊕  **Barniyah** 3 years, 8 months ago

Sorry , I think it's D

<https://cloud.google.com/sql/docs/features>

(Cloud SQL supports MySQL 5.6 or 5.7, and provides up to 416 GB of RAM and 30 TB of data storage, with the option to automatically increase the storage size as needed.)

upvoted 5 times

⊕  **xrun** 3 years, 1 month ago

Another consideration is that Cloud SQL uses standard databases like MySQL, PostgreSQL and now MS SQL. Cloud Spanner is a proprietary product of Google and does some things differently than typical databases (no stored procedures and triggers). So migrating to Cloud Spanner makes application refactoring necessary. So Cloud SQL is the answer.

upvoted 5 times

⊕  **LaxmanTiwari** 1 month, 1 week ago

Well explained I can confirm.

upvoted 1 times

You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.

What should you do?

- A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
- B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
- C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
- D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

**Correct Answer: B**

*Community vote distribution*

C (68%)

B (32%)

✉️  [Removed]  1 year, 4 months ago

Answer is C

When you use a single cluster to run a batch analytics job that performs numerous large reads alongside an application that performs a mix of reads and writes, the large batch job can slow things down for the application's users. With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters, so that batch jobs don't affect your applications' users

<https://cloud.google.com/bigtable/docs/replication-overview#use-cases>

upvoted 19 times

✉️  [Removed] 1 year, 2 months ago

"When you use a single cluster", here we are creating a 2nd cluster, so we'll be using 2 different clusters. We want to redirect analysis jobs to the 2nd cluster, and the other job to the 1st cluster. Thus, I think that D is more adequate

upvoted 1 times

✉️  somilaseeja 1 year, 1 month ago

Option D didn't say to create a new cluster, rather it said to increase the size of the cluster. There is a difference. Hence C is the correct answer to run the batch processing in a single cluster mode

upvoted 1 times

✉️  HarshKothari21 1 year, 4 months ago

Agreed :)

upvoted 1 times

✉️  carbino  1 month, 3 weeks ago

**Selected Answer: C**

It is C:

"Workload isolation:

Using separate app profiles lets you use different routing policies for different purposes. For example, consider a situation when you want to prevent a batch read job (workload A) from increasing CPU usage on clusters that handle an application's steady reads and writes (workload B). You can create an app profile for workload B that routes to a cluster group that excludes one cluster. Then you create an app profile for workload A that specifies single-cluster routing to the cluster that workload B doesn't send requests to.

You can change the settings for one application or function without affecting other applications that connect to the same data."

<https://cloud.google.com/bigtable/docs/app-profiles>

upvoted 1 times

✉️  aewis 6 months, 2 weeks ago

**Selected Answer: C**

It was actually illustrated here

<https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-serve>

upvoted 3 times

✉  **DevShah** 9 months, 2 weeks ago

**Selected Answer: C**

<https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-serve>

upvoted 1 times

✉  **A4M** 9 months ago

I see what you say on C but the question states high availability how do you handle that with option C when you have a single region cluster hence answer needs to be with multi-region cluster - To configure your instance for a high availability (HA) use case, create a new app profile that uses multi-cluster routing, or update the default app profile to use multi-cluster routing.

upvoted 1 times

✉  **A4M** 9 months ago

i meant single-cluster routing

upvoted 1 times

✉  **juliosb** 10 months, 1 week ago

**Selected Answer: C**

C. This is exactly the example in the documentation.

<https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-serve>

upvoted 3 times

✉  **DevShah** 9 months, 2 weeks ago

Correct

2 jobs >> 2 cluster

3 jobs >> 3 cluster

app profiles with single-cluster routing used to route to specific cluster

Job1 >> Cluster 1

Job2 >> Cluster 2 .....

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

Answer B:

reason 1: If you don't have any cost constraint use multi-cluster routing,

reason 2: Single cluster is less scalable as we need high scalability i would go with B

upvoted 1 times

✉  **samdhimal** 11 months, 4 weeks ago

**Selected Answer: C**

I am going for C?

upvoted 1 times

✉  **slade\_wilson** 1 year, 1 month ago

**Selected Answer: C**

When you use a single cluster to run a batch analytics job that performs numerous large reads alongside an application that performs a mix of reads and writes, the large batch job can slow things down for the application's users. With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters, so that batch jobs don't affect your applications' users

Single cluster routing - You can use single-cluster routing for this use case if you don't want your Bigtable cluster to automatically fail over if a zone or region becomes unavailable.

Multi-cluster routing - If you want Bigtable to automatically fail over to one region if your application cannot reach the other region, use multi-cluster routing.

upvoted 2 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

<https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-serve>

When you use a single cluster to run a batch analytics job that performs numerous large reads alongside an application that performs a mix of reads and writes, the large batch job can slow things down for the application's users. With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters, so that batch jobs don't affect your applications' users

upvoted 2 times

👤 **Siant\_137** 1 year, 1 month ago

Answer is C

"When you use a single cluster to run a batch analytics job that performs numerous large reads alongside an application that performs a mix of reads and writes, the large batch job can slow things down for the application's users. With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters, so that batch jobs don't affect your applications' users."

<https://cloud.google.com/bigtable/docs/replication-overview#batch-vs-server>

upvoted 2 times

👤 **sfsdeniso** 1 year, 2 months ago

Answer is C

upvoted 1 times

👤 **dish11dish** 1 year, 2 months ago

**Selected Answer: B**

Option B is correct

An app profile specifies the routing policy that Bigtable should use for each request.

Single-cluster routing routes all requests to 1 cluster in your instance. If that cluster becomes unavailable, you must manually fail over to another cluster.

Multi-cluster routing automatically routes requests to the nearest cluster in an instance. If the cluster becomes unavailable, traffic automatically fails over to the nearest cluster that is available. Bigtable considers clusters in a single region to be equidistant, even though they are in different zones. You can configure an app profile to route to any cluster in an instance, or you can specify a cluster group that tells the app profile to route to only some of the clusters in the instance.

Cluster group routing sends requests to the nearest available cluster within a cluster group that you specify in the app profile settings.

Reference:-<https://cloud.google.com/bigtable/docs/app-profiles#routing>

upvoted 3 times

👤 **piotrpiskorski** 1 year, 2 months ago

**Selected Answer: C**

<https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-server>

"When you use a single cluster to run a batch analytics job that performs numerous large reads alongside an application that performs a mix of reads and writes, the large batch job can slow things down for the application's users. With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters, so that batch jobs don't affect your applications' users."

It is C.

upvoted 2 times

👤 **gudiking** 1 year, 2 months ago

**Selected Answer: C**

C - "With replication, you can use app profiles with single-cluster routing to route batch analytics jobs and application traffic to different clusters so that batch jobs don't affect your applications' users." - <https://cloud.google.com/bigtable/docs/replication-overview#batch-vs-server>

upvoted 1 times

👤 **cloudmon** 1 year, 2 months ago

**Selected Answer: B**

It's B

<https://cloud.google.com/bigtable/docs/replication-overview#app-profiles>

<https://cloud.google.com/bigtable/docs/replication-overview#routing-policies>

upvoted 2 times

👤 **hauhau** 1 year, 3 months ago

**Selected Answer: B**

It is B . Add one cluster means you have multi clusters so you use Multi-cluster routing

upvoted 3 times

👤 **MisuLava** 1 year, 3 months ago

**Selected Answer: B**

I meant to vote B not C

upvoted 3 times

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

- A. Batch job, PubSubIO, side-inputs
- B. Streaming job, PubSubIO, JdbcIO, side-outputs
- C. Streaming job, PubSubIO, BigQueryIO, side-inputs
- D. Streaming job, PubSubIO, BigQueryIO, side-outputs

**Correct Answer: C***Community vote distribution*

C (94%)

6%

✉  **rickywck** Highly Voted 3 years, 10 months ago

Why not C? Without BigQueryIO how can data be written back to BigQuery?

upvoted 31 times

✉  **xq** 3 years, 10 months ago

C should be right

upvoted 7 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description: Sideinput for Bigquery data

upvoted 14 times

✉  **JOKKUNO** Most Recent 1 month ago

Side inputs

In addition to the main input PCollection, you can provide additional inputs to a ParDo transform in the form of side inputs. A side input is an additional input that your DoFn can access each time it processes an element in the input PCollection. When you specify a side input, you create a view of some other data that can be read from within the ParDo transform's DoFn while processing each element.

Side inputs are useful if your ParDo needs to inject additional data when processing each element in the input PCollection, but the additional data needs to be determined at runtime (and not hard-coded). Such values might be determined by the input data, or depend on a different branch of your pipeline.

upvoted 1 times

✉  **JOKKUNO** 1 month ago

<https://beam.apache.org/documentation/programming-guide/#side-inputs>

upvoted 1 times

piyush7777 5 months, 2 weeks ago

Why not side-output?

upvoted 1 times

TQM\_9MD 5 months, 4 weeks ago

Selected Answer: B

B. Use multi-cluster routing to add a second cluster to the existing instance, utilizing a live traffic app profile for the regular workload and a ba analytics profile for the analytical workload.

upvoted 1 times

Mathew106 6 months, 1 week ago

Selected Answer: C

The answer is C. It's a trap so that you answer A because of batch vs streaming but you need BigQueryIO. On the other hand, streaming is absolutely redundant here and will incur extra costs. C is right but would be better with batch.

upvoted 1 times

Siadd 1 year ago

A is the Answer.

A. Batch job, PubSubIO, side-inputs

upvoted 1 times

zellck 1 year, 1 month ago

Selected Answer: C

C is the answer.

<https://cloud.google.com/dataflow/docs/tutorials/ecommerce-java#side-input-pattern>

In streaming analytics applications, data is often enriched with additional information that might be useful for further analysis. For example, if I have the store ID for a transaction, you might want to add information about the store location. This additional information is often added by taking an element and bringing in information from a lookup table.

upvoted 4 times

sedado77 1 year, 4 months ago

Selected Answer: C

I got this question on sept 2022. Answer is C

upvoted 2 times

chrismayola 1 year, 3 months ago

dear can you please help, i have some questions about how to prepare the cerification exam using this questionnaire. this is my email cmayola@yahoo.fr, ping me to have some conversation

upvoted 1 times

alex12441 1 year, 12 months ago

Selected Answer: C

Answer: C

upvoted 1 times

medeis\_jar 2 years ago

Selected Answer: C

I vote for C, because data will come from Pub/Sub, so it should be streaming, we'll need PubSubIO to be able to read from PubSub and BigQueryIO to be able to write to BigQuery, finally the side-inputs pattern let us enrich data

upvoted 5 times

MaxNRG 2 years, 1 month ago

Selected Answer: C

Static reference data from BigQuery will go as side-inputs and data from pub-sub will go as streaming data using PubSubIO and finally BigQueryIO is required to push the final data to BigQuery

upvoted 4 times

JG123 2 years, 2 months ago

Ans: C

upvoted 1 times

✉  **pals\_muthu** 2 years, 5 months ago

Answer is C,

You need pubsubIO and BigQueryIO for streaming data and writing enriched data back to BigQuery. side-inputs are a way to enrich the data  
<https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs>

upvoted 6 times

✉  **Meuter** 2 years, 5 months ago

I choose C, because data will come from Pub/Sub, so it should be streaming, we'll need PubSubIO to be able to read from PubSub and BigQuery to be able to write to BigQuery, finally the side-inputs pattern let us enrich data

<https://beam.apache.org/releases/javadoc/2.4.0/org/apache/beam/sdk/io/gcp/pubsub/PubsubIO.html>

<https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs>

<https://beam.apache.org/releases/javadoc/2.3.0/org/apache/beam/sdk/io/gcp/bigquery/BigQueryIO.html>

upvoted 3 times

✉  **daghayeghi** 2 years, 10 months ago

C:

we have to use Streaming job because of Pub/Sub, and side-input thanks to static reference data. and we have to leverage BigQueryIO since finally we want to write data to BigQuery. then C is the correct answer.

upvoted 2 times

✉  **someshsehgal** 2 years, 11 months ago

Correct A. batch is cost-effective and no need to go for streaming

upvoted 1 times

✉  **funtoosh** 2 years, 11 months ago

How you are going to write back to BQ?

upvoted 1 times

Question #95

Topic 1

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of your Cloud Bigtable cluster. Which two actions can you take to accomplish this? (Choose two.)

- A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
- D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
- E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

**Correct Answer: AC**

*Community vote distribution*

CD (87%)

13%

✉  **jvg637**  3 years, 10 months ago

Answer is C & D.

C -> Adding more nodes to a cluster (not replication) can improve the write performance <https://cloud.google.com/bigtable/docs/performanc>  
D -> since Google recommends adding nodes when storage utilization is > 70% <https://cloud.google.com/bigtable/docs/modifying-instance#nodes>

upvoted 53 times

✉️  **sergio6** 2 years, 3 months ago

Adding nodes to the cluster In Bigtable scales linearly the performances both read and write  
<https://cloud.google.com/bigtable/docs/performance#typical-workloads>

upvoted 1 times

✉️  **dabrat** 3 years, 3 months ago

Storage utilization (% max)

The percentage of the cluster's storage capacity that is being used. The capacity is based on the number of nodes in your cluster.

In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data. If you do not plan to add significant amounts of data to your instance, you can use up to 100% of the hard limit.

Important: If any cluster in an instance exceeds the hard limit on the amount of storage per node, writes to all clusters in that instance will be denied until you add nodes to each cluster that is over the limit. Also, if you try to remove nodes from a cluster, and the cluster exceeds the hard limit on storage, Cloud Bigtable will deny the request.

If you are using more than the recommended percentage of the storage limit, add nodes to the cluster. You can also delete existing data, but deleted data takes up more space, not less, until a compaction occurs.

upvoted 4 times

✉️  **dabrat** 3 years, 3 months ago

<https://cloud.google.com/bigtable/docs/monitoring-instance>

upvoted 3 times

✉️  **Barniyah** Highly Voted 3 years, 9 months ago

Key visualizer is bigtable metric , So A and B incorrect  
storage utilization also bigtable metric , So D incorrect

The question want you to monitor pipeline metrics (which is dataflow metrics) , in our case we can only monitor latency .

The answer will be : C & E

upvoted 9 times

✉️  **ch3n6** 3 years, 7 months ago

No. it is C, D. "You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys."  
why are you monitoring read anyway? you are just writing.

upvoted 14 times

✉️  **musumusu** Most Recent 11 months, 2 weeks ago

why not B ?

upvoted 1 times

✉️  **musumusu** 11 months, 1 week ago

i am feeling to go with B and D. In option C, when latency is low, latency can be low for write operation for other reason.  
but in option B, its showing clearly when write pressure more than 100. But why no one is talking about B

upvoted 2 times

✉️  **RoshanAshraf** 1 year ago

Selected Answer: CD

Key visualizer is Metrics for Performance issues. Ruled out  
Storage and Write Operations ; C and D

upvoted 2 times

✉️  **zelick** 1 year, 1 month ago

Selected Answer: CD

CD is the answer.

<https://cloud.google.com/bigtable/docs/monitoring-instance#disk>

Storage utilization (% max)

- The percentage of the cluster's storage capacity that is being used. The capacity is based on the number of nodes in your cluster.  
In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data.

upvoted 3 times

✉️  **John\_Pongthorn** 1 year, 4 months ago

Selected Answer: CD

Well-designed row key : A B are not necessary

Write : CD both are involved in the question the most.

upvoted 2 times

✉  **Fezo** 1 year, 7 months ago

Answer: CD

<https://cloud.google.com/bigtable/docs/scaling>

upvoted 2 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: CD**

as explained by MaxNRG

upvoted 2 times

✉  **MaxNRG** 2 years, 1 month ago

**Selected Answer: CD**

D: In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data. If you do not plan to add significant amounts of data to your instance, you can use up to 100% of the hard limit

C: If this value is frequently at 100%, you might experience increased latency. Add nodes to the cluster to reduce the disk load percentage. The key visualizer metrics options, suggest other things other than increase the cluster size.

<https://cloud.google.com/bigtable/docs/monitoring-instance>

upvoted 3 times

✉  **hendrixlives** 2 years, 1 month ago

**Selected Answer: CD**

CD.

I agree with jvg637

upvoted 1 times

✉  **StefanoG** 2 years, 2 months ago

**Selected Answer: AD**

from <https://cloud.google.com/bigtable/docs/monitoring-instance>

Disk load - If this value is frequently at 100%, you might experience increased latency. Add nodes to the cluster to reduce the disk load percentage.

Storage utilization (% max) - In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data. If you do not plan to add significant amounts of data to your instance, you can use up to 100% of the hard limit.

upvoted 2 times

✉  **KokkiKumar** 2 years, 2 months ago

I am Voting for CD

upvoted 2 times

✉  **u\_t\_s** 2 years, 3 months ago

Answer should be D & E

upvoted 1 times

✉  **tavva\_prudhvi** 1 year, 9 months ago

Why are you monitoring read operations, when you're supposed to write? why E?

upvoted 1 times

✉  **sergio6** 2 years, 3 months ago

D--> 70% is the recommended percentage of the cluster's storage capacity that is being used. If you are using more than 70% storage, add nodes to the cluster

<https://cloud.google.com/bigtable/quotas#storage-per-node>

<https://cloud.google.com/bigtable/docs/monitoring-instance#disk>

E--> 100 ms is an order of magnitude lower latency than Google claimed (<10ms)

<https://cloud.google.com/bigtable/docs/performance#typical-workloads>

upvoted 2 times

✉  **hauhau** 2 years, 5 months ago

BC

D: you can just add node, not cluster

The percentage of the cluster's storage capacity that is being used. The capacity is based on the number of nodes in your cluster. (<https://cloud.google.com/bigtable/docs/monitoring-instance>)

After you create a Cloud Bigtable instance, you can update any of the following settings without any downtime:

(The number of nodes in each cluster)

<https://cloud.google.com/bigtable/docs/modifying-instance>

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

B, C , D - all three looks okay to me

upvoted 2 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C & D,

Option B eliminated, as Row are well defined (as per question) - so no need of key-visualizer

upvoted 3 times

✉  **squishy\_fishy** 2 years, 3 months ago

Answer is C, D.

B is not correct, because B is Key Visualizer, it means the row key needs re-design again.

upvoted 2 times

✉  **daghayeghi** 2 years, 10 months ago

C, D:

<https://cloud.google.com/bigtable/docs/monitoring-instance>

upvoted 3 times

Question #96

Topic 1

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

⇒ You will batch-load the posts once per day and run them through the Cloud Natural Language API.

⇒ You will extract topics and sentiment from the posts.

⇒ You must store the raw posts for archiving and reprocessing.

⇒ You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

- A. Store the social media posts and the data extracted from the API in BigQuery.
- B. Store the social media posts and the data extracted from the API in Cloud SQL.
- C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
- D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description: Social media posts can images/videos which cannot be stored in bigquery

upvoted 46 times

✉  **Shawvin** 2 years, 3 months ago

Yes, the raw data needs to be archived too

upvoted 1 times

✉  **Devx198912233** 3 years, 6 months ago

but the posts are fed into cloud natural language api. which means we have to consider the posts to be text only  
upvoted 4 times

✉  **asksathvik** 2 years, 5 months ago

Also to run batch queries data needs to be in Cloud Storage, so why not just store it there?  
upvoted 1 times

✉  **psu** Highly Voted 3 years, 9 months ago

Answer should be C, becose they ask you to save a copy of the raw posts for archival, which may not be possible if you directly feed the post to the API.

upvoted 17 times

✉  **itz\_me\_sudhir** Most Recent 11 months ago

can any one help me with the rest of question from 101 to 209 as i dont have a contributor access  
upvoted 2 times

✉  **zellick** 1 year, 1 month ago

Selected Answer: C

C is the answer.

upvoted 2 times

✉  **sedado77** 1 year, 4 months ago

Selected Answer: C

I got this question on sept 2022. Answer is C

upvoted 5 times

✉  **Ers0** 1 year, 4 months ago

Selected Answer: C

C is the correct one

upvoted 1 times

✉  **medeis\_jar** 2 years ago

Selected Answer: C

Only C make sense.

upvoted 2 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

You must store the raw posts for archiving and reprocessing, Store the raw social media posts in Cloud Storage.

B is expensive

D is not valid since you have to store the raw posts for archiving

Between A and C I's say C, since we're going to make dashboards and Data Studio will connect well with big query.  
and besides A would probably be more expensive.

upvoted 3 times

✉  **BigQuery** 2 years, 1 month ago

SAY MY NAME!

upvoted 4 times

✉  **StefanoG** 2 years, 2 months ago

Selected Answer: C

Analysis BQ

Storage GCS

upvoted 2 times

✉  **fire558787** 2 years, 5 months ago

I believe the API accesses data only from GCS Buckets not BigQuery (but I'm not entirely sure)

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 2 times

✉  **DPonly** 3 years ago

Answer should be C because we need to consider storage archival  
upvoted 2 times

✉  **arghya13** 3 years, 2 months ago

I'll go with option C  
upvoted 2 times

✉  **Alasmindas** 3 years, 2 months ago

I will go with Option C, because of the following reasons:-

a) Social media posts are "raw" - which means - it can be of any format (blob/object storage) is preferred.  
b) The output from the application (assuming the application is Cloud NLP) is to be future stored for archival purpose - and hence again Google Cloud storage is the best option - so option C

Option A &C - Incorrect, although Option D fulfills the requirement of "fewest step" but storing data in big query for archival purpose is not a Google recommended approach

Option B : Cloud SQL rules out as it does not solve either for archival storage or for analytics purpose.  
upvoted 3 times

✉  **singhkrishna** 3 years, 4 months ago

cost of long term storing is almost same in GCS and BQ, so answer D makes sense from that angle..  
upvoted 1 times

✉  **Tanmoyk** 3 years, 4 months ago

The job is supposed to run in batch process once in a day, so there is no requirement of stream data. The best economical and less complex steps is answer C

upvoted 2 times

Question #97

Topic 1

You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.

What should you do?

- A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
- B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
- C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
- D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

**Correct Answer: A**

*Community vote distribution*

B (90%)

10%

✉  **cleroy**  3 years, 10 months ago

Use Dataprep ! It's THE tool for this  
upvoted 56 times

✉  **rickywck**  3 years, 10 months ago

Yes B.

Honest speaking, sometimes I thought the answers being posted here were intentionally to mislead people who do not have proper knowledge on the subject, but just memorizing answers to pass the exam.

upvoted 53 times

✉  **[Removed]** 3 years ago

True.. might be legal issue?  
upvoted 6 times

👤 **sergiomujica** Most Recent 4 months, 3 weeks ago

**Selected Answer: A**

A is the right way to do it... dataprepo is clumsy  
upvoted 1 times

👤 **Wudihero2** 3 months, 2 weeks ago

...Did you even read through the question? It says "not require programming or knowledge of SQL". YOU are the one who's clumsy, not dataprep.  
upvoted 3 times

👤 **crazycosmos** 5 months, 4 weeks ago

**Selected Answer: B**

no programming -> B  
upvoted 1 times

👤 **FP77** 6 months ago

**Selected Answer: B**

I honestly do not understand what is the deal with this website. The correct answer is obviously Dataprep. How can they say it's A?  
upvoted 2 times

👤 **Besss** 1 year ago

**Selected Answer: B**

It's B. DataPrep it's the right tool.  
<https://cloud.google.com/dataprep>  
upvoted 2 times

👤 **zellck** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

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

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

upvoted 1 times

👤 **sedado77** 1 year, 4 months ago

I got this question on Sept 2022.  
upvoted 4 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

B. Actually there are two tools to fix this problem.  
Dataprep rely on Dataflow  
Datafusion rely on Dataproc  
upvoted 2 times

👤 **diagniste** 1 year, 8 months ago

**Selected Answer: A**

A is the best answer!  
upvoted 1 times

👤 **desertlotus1211** 1 year ago

Dataflow IS Apache Beam...  
upvoted 1 times

👤 **Venkat007** 1 year, 11 months ago

**Selected Answer: B**

B Dataprep  
upvoted 2 times

👤 **medeis\_jar** 2 years ago

**Selected Answer: B**

<https://cloud.google.com/dataprep/>  
upvoted 2 times

✉️  **MaxNRG** 2 years, 1 month ago

**Selected Answer: B**

B, "Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning"  
<https://cloud.google.com/dataprep/>

upvoted 6 times

✉️  **dattatray\_shinde** 1 year, 6 months ago

max you rock man!

upvoted 2 times

✉️  **GirijaSrinivasan** 2 years, 3 months ago

Answer is B. Data prep. The keyword here is no programming skills required.

upvoted 3 times

✉️  **nguyenmoon** 2 years, 4 months ago

B- Dataprep

upvoted 2 times

✉️  **pass\_gcp** 2 years, 5 months ago

Use Dataprep ....is the answer

upvoted 2 times

✉️  **sumanshu** 2 years, 7 months ago

Vote for B

upvoted 2 times

✉️  **sumanshu** 2 years, 7 months ago

Cloud Dataprep - almost fully automated

upvoted 1 times

Question #98

Topic 1

Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

- A. Execute gsutil rsync from the on-premises servers.
- B. Use Dataflow and write the data to Cloud Storage.
- C. Write a job template in Dataproc to perform the data transfer.
- D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

**Correct Answer: A**

*Community vote distribution*

A (100%)

👤 [Removed] Highly Voted 3 years, 10 months ago

Should be A

upvoted 24 times

👤 itche\_scratche Highly Voted 3 years, 9 months ago

should be A, dataflow is on cloud is external; "don't allow access from external IPs to their on-premises resources" so no dataflow.

upvoted 13 times

👤 Besss Most Recent 1 year ago

Selected Answer: A

A is correct

upvoted 1 times

👤 zellick 1 year, 1 month ago

Selected Answer: A

A is the answer.

[https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil\\_for\\_smaller\\_transfers\\_of\\_on-premises\\_data](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. We recommend that you include gsutil in your default path when you use Cloud Shell. It's also available by default when you install the Google Cloud CLI. It's a reliable tool that provides all the basic features you need to manage your Cloud Storage instances, including copying your data to and from the local file system and Cloud Storage. It can also move and rename objects and perform real-time incremental syncs, like rsync, to a Cloud Storage bucket.

upvoted 4 times

👤 somnathmaddi 1 year, 3 months ago

Selected Answer: A

Should be A

upvoted 2 times

👤 medeis\_jar 2 years ago

Selected Answer: A

Without this "The security rules don't allow access from external IPs to their on-premises resources" B would be an answer.

upvoted 1 times

👤 MaxNRG 2 years, 1 month ago

Selected Answer: A

A is the better and most simple IF there is no problem in having gsutil in our servers.

B and C no way, the comms will go GCP-Home, which says is not allowed.

D is valid, we can send the files with http://ftp...BUT ftp is not secure, and we'll need to move them to the cloud storage afterwards, which is detailed in the answer.

<https://cloud.google.com/storage/docs/gsutil/commands/rsync>

upvoted 6 times

👤 am2005 2 years, 1 month ago

I am confused . which one is correct A or B ???

upvoted 1 times

👤 hendrixlives 2 years, 1 month ago

Selected Answer: A

A is correct.

upvoted 1 times

👤 Chelseajcole 2 years, 3 months ago

This is the link:[https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil\\_for\\_smaller\\_transfers\\_of\\_on-premises\\_data](https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil_for_smaller_transfers_of_on-premises_data)

upvoted 2 times

👤 manocha\_01887 2 years, 5 months ago

How rsync will handle private network?

"..The security rules don't allow access from external IPs to their on-premises resources.."

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 2 times

✉  **daghayeghi** 2 years, 10 months ago

A:

[https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#options\\_available\\_from\\_google](https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#options_available_from_google)

upvoted 4 times

✉  **maurodipa** 2 years, 1 month ago

How could gsutil connect to Cloud Storage, if there is not access from external IPs? Should I understand that there is not access from outside to inside, but it is possible to send from inside to outside?

upvoted 1 times

✉  **szefco** 2 years, 1 month ago

Yes. There is no access to on-prem from external IPs, but on prem can talk to external

upvoted 3 times

✉  **Ravivarma4786** 3 years, 5 months ago

gsutil rsync will be used to transfer the files Ans A

upvoted 4 times

✉  **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 3 times

✉  **VishalB** 3 years, 6 months ago

Ans : A

The gsutil rsync command makes the contents under dst\_url the same as the contents under src\_url, by copying any missing files/objects (or those whose data has changed), and (if the -d option is specified) deleting any extra files/objects. src\_url must specify a directory, bucket, or bucket subdirectory

upvoted 6 times

✉  **Devx198912233** 3 years, 6 months ago

option A

[https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#options\\_available\\_from\\_google](https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#options_available_from_google)

upvoted 4 times

Question #99

Topic 1

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query ``-dry\_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

- A. Create a separate table for each ID.
- B. Use the LIMIT keyword to reduce the number of rows returned.
- C. Recreate the table with a partitioning column and clustering column.
- D. Use the bq query --maximum\_bytes\_billed flag to restrict the number of bytes billed.

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉  **rickywck**  3 years, 10 months ago

should be C:

<https://cloud.google.com/bigquery/docs/best-practices-costs>

upvoted 43 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Correct - C

upvoted 17 times

✉  **zellick** Most Recent 1 year, 1 month ago

Selected Answer: C

C is the answer.

<https://cloud.google.com/bigquery/docs/partitioned-tables>

A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

<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.

upvoted 4 times

✉  **Fezo** 1 year, 6 months ago

Selected Answer: C

C is the answer

<https://cloud.google.com/bigquery/docs/best-practices-costs>

upvoted 3 times

✉  **medeis\_jar** 2 years ago

Selected Answer: C

C only make sense

upvoted 2 times

✉  **MaxNRG** 2 years, 1 month ago

Selected Answer: C

<https://cloud.google.com/bigquery/docs/best-practices-costs>

Applying a LIMIT clause to a SELECT \* query does not affect the amount of data read. You are billed for reading all bytes in the entire table, as the query counts against your free tier quota.

A and D doesn't make sense

Its C, when you want to select by a partition you should write something like:

CREATE TABLE `blablabla.partitioned`

PARTITION BY

DATE(timestamp)

CLUSTER BY id

AS

SELECT \* FROM `blablabla`

upvoted 5 times

✉  **Anilcp980** 2 years, 1 month ago

this is a trap to make people fail by giving wrong answer as B.

upvoted 3 times

✉  **snadaf** 2 years, 1 month ago

It's D, here is the link

<https://cloud.google.com/bigquery/docs/best-practices-costs>

upvoted 1 times

✉  **maurodipa** 2 years, 1 month ago

Well, you mean C, isn't it?

upvoted 1 times

✉  **Crudgey** 2 years, 2 months ago

Are they having a laugh at us by giving so many bad answers?

upvoted 5 times

✉  **tsoetan001** 2 years, 3 months ago

Answer: B

Note: minimal change to sql

upvoted 1 times

✉  **szefco** 2 years, 2 months ago

Not B. LIMIT will not reduce amount of data scanned - only limit the final output, but you will still be billed for scanning whole table. C is correct. After applying partitioning and clustering amount of bytes scanned will decrease

upvoted 3 times

✉  **Ysance\_AGS** 2 years, 4 months ago

"You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries" that doesn't mean that you can create or edit existing tables ! you only can edit the SQL query !!! so answer D is the correct one.

upvoted 2 times

✉  **szefco** 2 years, 2 months ago

I don't agree. Question says "minimal changes to existing SQL queries" - if you recreate table with partitioning and clustering you don't need to change SQLs that read from that table.

C is correct answer here.

upvoted 1 times

✉  **squishy\_fishy** 2 years, 3 months ago

D would just block your query. The answer is C.

upvoted 1 times

✉  **nguyenmoon** 2 years, 4 months ago

C - create partition table

upvoted 2 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 4 times

✉  **felixwtf** 3 years, 1 month ago

LIMIT keyword is applied only at the end, i.e., only to limit the results already calculated. Therefore, a full table scan will have already happened. The WHERE clause on the other hand would provide the desired filtering depending on the case. So, C is the correct answer.

upvoted 4 times

✉  **learnazureportal** 3 years, 2 months ago

Not sure, why option C selected! The correct Answer is B. The question clearly says "minimal changes to existing SQL queries". who said that recreate the table, with partitioning layout is minimal and is PART of SQL queries!

upvoted 2 times

✉  **hdmi\_switch** 2 years, 6 months ago

In addition to the previous reply, the LIMIT statement applies to the output (what you see in the UI), the full table scan will still happen. C is correct according to best practices.

upvoted 1 times

✉  **ceak** 3 years, 2 months ago

recreating table will not affect existing SQL queries as they will still be selecting the same table name, but the scan will hugely decrease. so, option C is the correct answer.

upvoted 5 times

✉  **squishy\_fishy** 2 years, 3 months ago

Recreating table is recommended by Google.

upvoted 1 times

✉  **arghya13** 3 years, 2 months ago

should be C:

upvoted 2 times

✉  **gyclop** 3 years, 4 months ago

Correct - C :

"Limit" keyword restricts the final dataset to "n" rows, but is not able to restrict full table scan

upvoted 3 times

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

- A. Use bq load to load a batch of sensor data every 60 seconds.
- B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
- C. Use the INSERT statement to insert a batch of data every 60 seconds.
- D. Use the MERGE statement to apply updates in batch every 60 seconds.

**Correct Answer: C**

*Community vote distribution*

B (100%)

 **jvg637** Highly Voted 3 years, 10 months ago

I think we need a pipeline, so it's B to me.

upvoted 29 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Correct - B

upvoted 16 times

 **Helinia** Most Recent 1 month, 4 weeks ago

“need the data to be available within 1 minute of ingestion for real-time analysis” → low latency requirement → Dataflow streaming

The database can either be BQ or BigTable for this kind of requirement in data volume and latency. But it mentioned that the destination has to be BQ, so B.

upvoted 1 times

 **NeoNitin** 5 months, 3 weeks ago

ANSWER b.

FULL question ihave if you nee mail me  
neonitin6ATtherate.....

upvoted 3 times

 **aryaavinash** 4 months, 3 weeks ago

full email id please ?

upvoted 1 times

 **Oleksandr0501** 9 months, 1 week ago

Selected Answer: B

I think we need a pipeline, so it's B to me.))

upvoted 3 times

 **votinhlumbkip** 9 months, 2 weeks ago

Selected Answer: B

I think we need a pipeline, so it's B to me.

upvoted 2 times

 **JANCAI** 11 months ago

Why the answer from the <reveal answer> is C??

upvoted 1 times

 **zelick** 1 year, 1 month ago

Selected Answer: B

B is the answer.

upvoted 1 times

👤 **Prasha123** 1 year, 2 months ago

Selected Answer: B

Need pipeline so its B

upvoted 1 times

👤 **sedado77** 1 year, 4 months ago

Selected Answer: B

I got this question on sept 2022. Answer is B

upvoted 7 times

👤 **medeis\_jar** 2 years ago

Selected Answer: B

omg. B only

upvoted 2 times

👤 **MaxNRG** 2 years, 1 month ago

Selected Answer: B

Is B, if we expect a growth we'll need some buffer (that will be pub-sub) and the dataflow pipeline to stream data in big query.

The tabledata.insertAll method is not valid here.

upvoted 7 times

👤 **hendrixlives** 2 years, 1 month ago

Selected Answer: B

B, streaming with dataflow

upvoted 3 times

👤 **JG123** 2 years, 2 months ago

Wrong answer shown again by examtopics.com

Ans: B

upvoted 1 times

👤 **Ysance\_AGS** 2 years, 4 months ago

B => with dataflow you can parallelize data ingestion

upvoted 2 times

👤 **szefco** 2 years, 2 months ago

And make it streaming

upvoted 1 times

👤 **sandipk91** 2 years, 5 months ago

B is the right answer

upvoted 2 times

👤 **sumanshu** 2 years, 7 months ago

Vote for B

upvoted 4 times

Question #101

Topic 1

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

- A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.

D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

**Correct Answer: A**

*Community vote distribution*

B (55%)	A (43%)	2%
---------	---------	----

✉  **Ganshank**  3 years, 9 months ago

You are transferring sensitive patient information, so C & D are ruled out. Choice comes down to A & B. Here it gets tricky. How to choose Transfer Appliance: (<https://cloud.google.com/transfer-appliance/docs/2.0/overview>)

Without knowing the bandwidth, it is not possible to determine whether the upload can be completed within 7 days, as recommended by Google. So the safest and most performant way is to use Transfer Appliance.

Therefore my choice is B.

upvoted 58 times

✉  **AzureDP900** 1 year, 1 month ago

B is right answer

upvoted 3 times

✉  **tprashanth** 3 years, 6 months ago

<https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets>

The table shows for 1Gbps, it takes 30 hrs for 10 TB. Generally, corporate internet speeds are over 1Gbps. I'm inclined to pick A

upvoted 4 times

✉  **forepick** 8 months ago

If you transfer 10TBs over the wire, your network will be blocked for the entire transfer time. This isn't something a company would be happy to swallow.

upvoted 2 times

✉  **BigQuery** 2 years, 1 month ago

SAY MY NAME!

You need to Transfer Sensitive Patient information, over public ISP you shouldn't do that.

upvoted 3 times

✉  **TNT87** 3 years, 4 months ago

Answer is B, gsutil has a limit of 1TB according to Google documentation, if data is more than 1TB then we have to use Transfer Appliance.

upvoted 17 times

✉  **Yiouk** 2 years, 5 months ago

The answer is clearly seen here: <https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options>

upvoted 8 times

✉  **SSV**  3 years, 6 months ago

Answer should be B: A is also correct but it has its own limit. It allows only 5TB data upload at a time to cloud storage.

<https://cloud.google.com/storage/quotas>

I will go with B

upvoted 8 times

✉  **VASI** 3 years ago

5Tb "for individual objects". Create smaller AVRO files.

upvoted 2 times

✉  **VASI** 3 years ago

AVRO compression can reduce file size to a tenth

upvoted 3 times

TVH\_Data\_Engineer Most Recent 1 month, 1 week ago

**Selected Answer: B**

Given the sensitivity of the patient records and the large size of the data, using Google's Transfer Appliance is a secure and efficient method. Transfer Appliance is a hardware solution provided by Google for transferring large amounts of data. It enables you to securely transfer data without exposing it over the internet.

upvoted 1 times

rocky48 1 month, 3 weeks ago

**Selected Answer: B**

Option B combines security, efficiency, and ease of use, making it a suitable choice for transferring sensitive patient records to BigQuery. upvoted 1 times

spicebits 2 months, 3 weeks ago

**Selected Answer: A**

10 TB is nothing. With a single 10 GB interconnect you could transfer the data in 3 hours or even with a 1 GB speeds without interconnect you could transfer it in one weekend. The transfer appliance will take 25 days to get the appliance and then 25 days while you wait for the data to be available that is not "time-efficient" at all. I go with A instead of B.

upvoted 2 times

spicebits 2 months, 3 weeks ago

I got the 25 days + 25 days from here: <https://cloud.google.com/transfer-appliance/docs/4.0/overview#transfer-speeds>

upvoted 1 times

A\_Nasser 4 months, 1 week ago

**Selected Answer: A**

transfer appliance will take time more than gsutil. and we did not mention yet if the location of the organization has google data centre

upvoted 2 times

DineshVarma 5 months, 1 week ago

**Selected Answer: D**

As per Google recommendation above 1TB of transfer from onprem or from Google cloud or other cloud storage like s3 etc we need to use storage transfer service.

upvoted 1 times

arien\_chen 5 months, 1 week ago

Transfer Appliance would take 20 days for expected turnaround time. <https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#expected%20turnaround:~:text=The%20expected%20turnaround%20time%20for%20a%20network%20appliance%20to%20be%20shipped%2C%20loaded%20with%20your%20data%2C%20shipped%20back%2C%20and%20rehydrated%20on%20Google%20Cloud%2020%20days.>

The best answer would be A.

If gsutil consume/leverage 100MB it would take 12 days and more time-efficient than B.

This is a reasonable assumption.

<https://cloud.google.com/static/architecture/images/big-data-transfer-how-to-get-started-transfer-size-and-speed.png>

upvoted 1 times

Colourseun 5 months, 2 weeks ago

I will go with " A" because of the transition time to take transfer appliance to Google and that also depends in the organisation location. gsutil works anywhere internet is available.

upvoted 1 times

NeoNitin 6 months, 2 weeks ago

bhai ek baar mera point sun lo and khud ki research karo...

option A,because dekho 10tb hai ye mat dekho file ko compress kiya ja raha hai Avro me

jo ki 90%-92% compress kar deta hai, to finaly hamare pass 1TB ya esase bhi kam ka file data hai jisko transfer karna hai , ab bataao Transfer Appliance kyo use karu bhaishab transfer appliance ki catagory hai 40tb aur 300TB ki , kyo offline ja rahe ho jo ki 7 din ya usase jyada time le tumhara data online aane me,

aur GSUTIL use karoge aur ye 100MB pe hi chala without dedicated bandwidth tab bhi ye ,1TB 100MB/S ki speed se 1 din me pura data online la dega .kyoki avro se file pahale hi 10tb se 1tb ho chuki hai. to GSUTIL is the best,bhale hi cost effective nahi bola hai question me but time k to dekho

upvoted 3 times

aewis 6 months, 2 weeks ago

**Selected Answer: B**

A will take crazy time if the organization didnt have a dedicated link

upvoted 1 times

✉  **ZZHZZH** 6 months, 3 weeks ago

**Selected Answer: A**

Transfer Appliance is not as time-efficient when you have enough bandwidth. [https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer\\_appliance\\_for\\_larger\\_transfers](https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer_appliance_for_larger_transfers)

upvoted 2 times

✉  **WillemHendr** 7 months, 1 week ago

**Selected Answer: B**

There is no "cost effective", if this is not a clear case for the appliance than what is?

upvoted 1 times

✉  **Ender\_H** 7 months, 3 weeks ago

**Selected Answer: A**

A is the answer, the question states the following facts:

- Total size of database 10TB.

- Solution needs to be:

\* Secure

\* Time-efficient

Total size of database:

will be significantly reduced in an avro file compression (up to 90% compression)

Secure transfer:

Even if we are dealing with sensitive data, data is encrypted when in transit while using `gsutil cp` to upload the data to GCS. <https://cloud.google.com/storage/docs/gsutil/addlhelp/SecurityandPrivacyConsiderations#transport-layer-security>

Time-Efficient:

gsutil could upload 10TB of data in 30 hours (or 1TB if its avro compressed first in 3 hours)

upvoted 4 times

✉  **dgteixeira** 7 months, 3 weeks ago

**Selected Answer: A**

It has to be gsutil.

In this documentation: <https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options>  
It states that if it meets the projects deadline, use gsutil.

Also, for Transfer Appliance, here ([https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer\\_appliance\\_for\\_larger\\_transfers](https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer_appliance_for_larger_transfers)), it states:

The expected turnaround time for a network appliance to be shipped, loaded with your data, shipped back, and rehydrated on Google Cloud 20 days.

Even with 100 Mbps, for 10 TB, it's 12 days. Almost half! of the Transfer Appliance.

It's, of course, option A.

upvoted 1 times

✉  **forepick** 8 months ago

BTW, all options are talking about exporting RDBMS as a SINGLE file of 10TB.

GCS object is limited to 5TB, pay attention

upvoted 1 times

✉  **KanchanC** 1 month, 3 weeks ago

Would it not be compressed in Avro?

upvoted 1 times

✉  **forepick** 8 months ago

**Selected Answer: B**

A and B are the only options with no public access, and 10TB is too large for gsutil, therefore - B

upvoted 1 times

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

- A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
- B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
- C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
- D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

**Correct Answer: C**

*Community vote distribution*

C (52%)

A (48%)

 **MaxNRG** Highly Voted 2 years, 1 month ago

**Selected Answer: A**

A - New correct answer

C - Old correct answer (for 2019)

upvoted 32 times

✉️  **MaxNRG** 1 month, 1 week ago

C is better

The best approach is to use BigQuery streaming to stream the inventory changes into a daily inventory movement table. Then calculate balances in a view that joins the inventory movement table to the historical inventory balance table. Finally, update the inventory balance table nightly (option C).

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

The key reasons this is better than the other options:

Using BigQuery UPDATE statements (option A) would be very inefficient for thousands of updates per hour. It is better to batch updates. Partitioning the inventory balance table (option B) helps query performance, but does not solve the need to incrementally update balances. Using the bulk loader (option D) would require batch loading the updates, which adds latency. Streaming inserts updates with lower latency.

So option C provides a scalable architecture that streams updates with low latency while batch updating the balances only once per day for efficiency. This balances performance and accuracy needs.

upvoted 3 times

✉️  **MaxNRG** 1 month, 1 week ago

Here's why the other options are less suitable:

A. Leverage BigQuery UPDATE statements: While technically possible, this approach is inefficient for frequent updates as it requires individual record scans and updates, affecting performance and potentially causing data race conditions.

B. Partition the inventory balance table: Partitioning helps with query performance for large datasets, but it doesn't address the need for near real-time updates.

D. Use the BigQuery bulk loader: Bulk loading daily changes is helpful for historical data ingestion, but it won't provide near real-time updates necessary for the dashboard.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

Option C offers the following advantages:

Streams inventory changes near real-time: BigQuery streaming ingests data immediately, keeping the inventory movement table constantly updated.

Daily balance calculation: Joining the movement table with the historical balance table provides an accurate view of current inventory levels without affecting the actual balance table.

Nightly update for historical data: Updating the main inventory balance table nightly ensures long-term data consistency while maintaining near real-time insights through the view.

This approach balances near real-time updates with efficiency and data accuracy, making it the optimal solution for the given scenario.

upvoted 1 times

✉️  **Yiouk** 6 months, 2 weeks ago

There are still limitations on DML statements (2023) e.g. only 2 concurrent UPDATES and up to 20 queued hence not appropriate for this scenario:

<https://cloud.google.com/bigquery/quotas#data-manipulation-language-statements>

upvoted 2 times

✉️  **NeoNitin** 6 months, 2 weeks ago

option A: what limitation here 1500/perday okay in question we will get max 24 jobs hourly updated okay, now speed 5 operation /10 sec , 1 operation 2sec , and we are getting new update in 1 hour so we have time 3600 sec and we need to update around 1000 update according to speed take 2000sec still we have 1600 sec rest to getting new update so . that's why I think DML is best option for this work

upvoted 1 times

✉️  **Nandababy** 1 month, 2 weeks ago

In question it mentioned several thousands of updates every hour, several thousands could be 20-30 thousands as well. Where it is mentioned for only 1000 updates?

upvoted 1 times

✉  **haroldbenites** Highly Voted 3 years, 5 months ago

C is correct.  
It says "update Every hour"  
And need " accuracy"  
upvoted 24 times

✉  **NeoNitin** 6 months, 2 weeks ago

option A:what limitation here 1500/perday okay in question we will get max 24 jobs hourly updated okay, now speed 5 operation /10 sec , 1 operation 2sec , and we are getting new update in 1 hour so we have time 3600 sec and we need to update around 1000 update according to speed take 2000sec still we have 1600 sec rest to getting new update so . thats why I thing DML is best option for this work

upvoted 2 times

✉  **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: C**

The best approach is to use BigQuery streaming to stream the inventory changes into a daily inventory movement table. Then calculate balan in a view that joins the inventory movement table to the historical inventory balance table. Finally, update the inventory balance table nightly (option C).

upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

The key reasons this is better than the other options:

Using BigQuery UPDATE statements (option A) would be very inefficient for thousands of updates per hour. It is better to batch updates.

Partitioning the inventory balance table (option B) helps query performance, but does not solve the need to incrementally update balances

Using the bulk loader (option D) would require batch loading the updates, which adds latency. Streaming inserts updates with lower latenc

So option C provides a scalable architecture that streams updates with low latency while batch updating the balances only once per day for efficiency. This balances performance and accuracy needs.

upvoted 1 times

✉  **rocky48** 1 month, 3 weeks ago

**Selected Answer: C**

Option C.

Using the BigQuery streaming to stream changes into a daily inventory movement table and calculating balances in a view that joins it to the historical inventory balance table can help you achieve the desired performance and accuracy. You can then update the inventory balance table nightly. This approach can help you avoid the overhead of scanning large amounts of data with each inventory update, which can be time-consuming and resource-intensive.

Leveraging BigQuery UPDATE statements to update the inventory balances as they are changing (option A) can be resource-intensive and may not be the most efficient way to achieve the desired performance.

upvoted 3 times

✉  **AnonymousPanda** 2 months, 1 week ago

**Selected Answer: C**

As per other answers C

upvoted 1 times

✉  **Nirca** 2 months, 3 weeks ago

**Selected Answer: A**

Simple and will work

upvoted 1 times

✉  **odacir** 3 months, 1 week ago

**Selected Answer: C**

Answer is C. Why because "Update" limits is 1500/per day, and the question say: You have several thousand updates to inventory every hour. is impossible to use updates all the time.

upvoted 2 times

✉  **Nirca** 3 months, 3 weeks ago

**Selected Answer: A**

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing - is so simple and RIGHT!

upvoted 1 times

✉  **brookpetit** 4 months, 2 weeks ago

**Selected Answer: C**

C is more universal and sustainable  
upvoted 2 times

✉  **ZZHZZH** 6 months, 3 weeks ago

**Selected Answer: C**

UPDATE is too expensive. Joining main and delta tables is the right way to capture data change.  
upvoted 3 times

✉  **euro202** 6 months, 3 weeks ago

**Selected Answer: C**

I think the answer is C. The question is about maximizing performance and accuracy, it's ok if we need expensive JOINs. BigQuery has a daily quota of 1500 UPDATEs, and the question talks about several thousand updates every hour.  
upvoted 2 times

✉  **jackdbd** 4 months ago

DML statements do not count toward the number of table modifications per day.  
<https://cloud.google.com/bigquery/quotas#data-manipulation-language-statements>

So I would go with A.

upvoted 1 times

✉  **jackdbd** 4 months ago

Sorry, wrong link. Here is the correct one: [https://cloud.google.com/bigquery/quotas#standard\\_tables](https://cloud.google.com/bigquery/quotas#standard_tables)  
upvoted 1 times

✉  **vaga1** 7 months ago

**Selected Answer: A**

C create a view that joins to a table seems dumb to me  
upvoted 1 times

✉  **forepick** 8 months ago

**Selected Answer: C**

Too frequent updates are way too expensive in an OLAP solution. This is much more likely to stream changes to the table(s) and aggregate them changes in the view.

<https://stackoverflow.com/questions/74657435/bigquery-frequent-updates-to-a-record>  
upvoted 3 times

✉  **streeber** 9 months, 3 weeks ago

**Selected Answer: C**

Has to be C.  
DML has hard limit of 1500 operations per table per day: [https://cloud.google.com/bigquery/quotas#standard\\_tables](https://cloud.google.com/bigquery/quotas#standard_tables)  
upvoted 1 times

✉  **lucaluka1982** 10 months, 1 week ago

**Selected Answer: C**

Update action is not efficient  
upvoted 2 times

✉  **NeoNitin** 6 months, 2 weeks ago

option A: what limitation here 1500/ per day okay in question we will get max 24 jobs hourly updated okay,  
now speed 5 operation /10 sec , 1 operation 2 sec , and we are getting new update in 1 hour so we have time 3600 sec and we need to update around 1000 update according to speed take 2000 sec still we have 1600 sec rest to getting new update so .

thats why I think DML is best option for this work

upvoted 1 times

midgoo 10 months, 3 weeks ago

Selected Answer: A

This question has 2 parts:

1. Query the table in real-time
2. Update the table with thousands of records per hour ~ 10 updates per second

Without (1), C seems to be the good approach by using staging table to buffer the update using Change Data Capture method. However, that method will make the query expensive due to the JOIN. So A is a better choice here.

upvoted 3 times

musumusu 11 months, 1 week ago

Question #103

Topic 1

as

You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table that have a recovery point objective (RPO) of 30 days?

- Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
- Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.
- Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
- Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

Correct Answer: C

Community vote distribution

C (64%)

D (24%)

12%

rocky48 1 month, 3 weeks ago

Selected Answer: A

ou should consider option A.

Setting the BigQuery dataset to be regional and using a point-in-time snapshot to recover the data in the event of an emergency can help you achieve the desired level of availability and minimize cost. This approach can help you avoid the additional cost of creating and maintaining backup copies of the data, which can be expensive.

Setting the BigQuery dataset to be multi-regional (options C and D) can provide additional redundancy and availability. However, this approach can be more expensive than setting the dataset to be regional, especially if you do not require the additional level of redundancy.

upvoted 1 times

Nirca 3 months, 3 weeks ago

Selected Answer: A

I'm going for A:

1. Set the BigQuery dataset to be regional. This will reduce the cost of storage compared to a multi-regional dataset.
2. building Snapshot: bq snapshot --dataset <dataset\_id> --table <table\_id> <snapshot\_id>

upvoted 2 times

ffggrre 3 months, 1 week ago

typically Multi-region cost is equal or less than a region. <https://cloud.google.com/bigquery/pricing#storage>

upvoted 1 times

ckanaar 4 months, 1 week ago

I think the answer is A:

This option meets the 30-day RPO requirement, assuming that the snapshot is maintained for that long. It offers high availability as data is written synchronously to 2 zones within a region: <https://cloud.google.com/blog/topics/developers-practitioners/backup-disaster-recovery-strategies-bigquery/>. The cost would be lower than maintaining a multi-regional dataset, but you'll incur the cost of the snapshot.

upvoted 3 times

⊕  **DeepakVenkatachalam** 4 months, 2 weeks ago

Answer is B. Timetravel only covers for 7 days and a scheduled query is needed for creating Table snapshots for 30 days. Also table snapshot must remain in the same region as base table (please refer to limitation of table snapshot from below link)  
<https://cloud.google.com/bigquery/docs/table-snapshots-intro>

upvoted 3 times

⊕  **lucaluka1982** 10 months, 2 weeks ago

Why not B? Setting dataset regional or multi does not affect the backup and recovery strategy.

upvoted 3 times

⊕  **midg00** 10 months, 3 weeks ago

**Selected Answer: C**

1. HA -> Multi-region
2. DR -> Snapshot

upvoted 3 times

⊕  **kostol** 11 months, 1 week ago

**Selected Answer: D**

<https://cloud.google.com/bigquery/docs/table-snapshots-scheduled>

upvoted 2 times

⊕  **desertlotus1211** 1 year ago

Answer is C: <https://cloud.google.com/bigquery/docs/table-snapshots-intro>

"Benefits of using table snapshots include the following:

Keep a record for longer than seven days. With BigQuery time travel, you can only access a table's data from seven days ago or more recent. With table snapshots, you can preserve a table's data from a specified point in time for as long as you want.

Minimize storage cost. BigQuery only stores bytes that are different between a snapshot and its base table, so a table snapshot typically uses less storage than a full copy of the table."

But the wording is foolish... It's table snapshot, NOT point in time snapshot!

<https://cloud.google.com/bigquery/docs/time-travel#restore-a-table>

this is point in time using time travel window - max is 7 days...

upvoted 3 times

⊕  **YatMoh** 1 year, 1 month ago

One of the criteria is to minimize cost. While C is an ideal solution but more expensive compared to D. In my opinion the correct answer should be D.

upvoted 2 times

⊕  **jkhong** 1 year, 1 month ago

How is using snapshots more expensive, snapshots store the delta of the state of the table during time of taking snapshot and the current state of the table ([https://cloud.google.com/bigquery/docs/table-snapshots-intro#storage\\_costs](https://cloud.google.com/bigquery/docs/table-snapshots-intro#storage_costs))

When a snapshot is created, there is literally zero cost... until changes are made to the base table, only then will the delta storage cost be charged.. This certainly seems cheaper than creating a new table

upvoted 2 times

⊕  **raghu06raj** 1 year, 1 month ago

Even when the dataset/table is regional, google provides high availability with data available in 2 zones. Isn't A correct answer?

upvoted 3 times

⊕  **zellck** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

A BigQuery table snapshot preserves the contents of a table (called the base table) at a particular time. You can save a snapshot of a current table, or create a snapshot of a table as it was at any time in the past seven days. A table snapshot can have an expiration; when the configured amount of time has passed since the table snapshot was created, BigQuery deletes the table snapshot. You can query a table snapshot as you would a standard table. Table snapshots are read-only, but you can create (restore) a standard table from a table snapshot, and then you can modify the restored table.

upvoted 4 times

✉  **hauhau** 1 year, 2 months ago

**Selected Answer: D**

point in time snapshots only have data from past 7 days

upvoted 2 times

✉  **desertlotus1211** 1 year ago

[https://cloud.google.com/bigquery/docs/time-travel#configure\\_the\\_time\\_travel\\_window](https://cloud.google.com/bigquery/docs/time-travel#configure_the_time_travel_window)

'You can set the duration of the time travel window, from a minimum of two days to a maximum of seven days'

Not the same as snapshot

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

We can keep PIT snapshots for as long as we want. I think you confused snapshots with time travel...

"With BigQuery time travel, you can only access a table's data from seven days ago or more recently. With table snapshots, you can preserve a table's data from a specified point in time for as long as you want."

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

C

Benefits of using table snapshots include the following:

Keep a record for longer than seven days. With BigQuery time travel, you can only access a table's data from seven days ago or more recent. With table snapshots, you can preserve a table's data from a specified point in time for as long as you want.

Minimize storage cost. BigQuery only stores bytes that are different between a snapshot and its base table, so a table snapshot typically uses less storage than a full copy of the table.

upvoted 2 times

✉  **cloudmon** 1 year, 2 months ago

**Selected Answer: C**

C

Multi-region for HA

Snapshot for recovery

upvoted 2 times

✉  **max\_c** 1 year, 4 months ago

**Selected Answer: C**

C for high availability and use of built-in feature

Point-in-time snapshots only have data from the past 7 days at their creation. They can be stored for as long as desired.

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

upvoted 4 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: C**

C and D are high availability

But C use built-in feature in BQ

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

While D it is quite put too much effort and not minimize cost

upvoted 1 times

✉  **MounicaN** 1 year, 4 months ago

**Selected Answer: D**

point in time snapshots only have data from past 7 days

upvoted 2 times

✉  **MisuLava** 1 year, 3 months ago

30 days RPO means 30 days of data is good for recovery. but 7 days of data is even better.

RPO means the maximum age of an acceptable backup in case of a disaster recovery.

Am I wrong ?

upvoted 2 times

✉  **deavid** 1 year, 3 months ago

Incorrect, that's for Big Query's time travel. You can create a snapshot of a current table, or of a table like it was at 7 days ago. With a snapshot, you can preserve a table's data from a specified point in time for as long as you want. Source:

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

upvoted 2 times

Question #104

Topic 1

You used Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

- A. Create a cron schedule in Dataprep.
- B. Create an App Engine cron job to schedule the execution of the Dataprep job.
- C. Export the recipe as a Dataprep template, and create a job in Cloud Scheduler.
- D. Export the Dataprep job as a Dataflow template, and incorporate it into a Composer job.

**Correct Answer: C**

*Community vote distribution*

D (67%)

A (23%)

10%

✉  **jkhong**  1 year, 1 month ago

I'd pick D because it's the only option which allows variable execution (since we need to execute the dataprep job only after the prior load job). Although D suggests the export of Dataflow templates, this discussion suggests that the export option is no longer available (<https://stackoverflow.com/questions/72544839/how-to-get-the-dataflow-template-of-a-dataprep-job>), there are already Airflow Operators for Dataprep which we should be using instead - <https://airflow.apache.org/docs/apache-airflow-providers-google/stable/operators/cloud/dataprep.html>

upvoted 11 times

midgoo Highly Voted 10 months, 3 weeks ago

**Selected Answer: D**

Since the load job execution time is unexpected, schedule the Dataprep based on a fixed time window may not work. When the Dataprep job run the first time, we can find the Dataflow job for that in the console. We can use that to create the Template --> With help of the Composer to determine if the load job is completed, we can then trigger the Dataflow job

upvoted 9 times

TVH\_Data\_Engineer Most Recent 1 month, 1 week ago

**Selected Answer: A**

Dataprep by Trifecta allows you to schedule the execution of recipes. You can set up a cron schedule directly within Dataprep to automatically run your recipe at specified intervals, such as daily.

WHY NOT D ? : This option involves significant additional complexity. Exporting the Dataprep job as a Dataflow template and then incorporating it into a Composer (Apache Airflow) job is a more complicated process and is typically used for more complex orchestration needs that go beyond simple scheduling.

upvoted 1 times

MaxNRG 1 month, 1 week ago

**Selected Answer: D**

We have external dependency "after the load job with variable execution time completes" which requires DAG -> Airflow (Cloud Composer)

The reasons:

A scheduler like Cloud Scheduler won't handle the dependency on the BigQuery load completion time  
Using Composer allows creating a DAG workflow that can:

Trigger the BigQuery load

Wait for BigQuery load to complete

Trigger the Dataprep Dataflow job

Dataflow template allows easy reuse of the Dataprep transformation logic

Composer coordinates everything based on the dependencies in an automated workflow

upvoted 1 times

rocky48 1 month, 3 weeks ago

**Selected Answer: D**

I'd pick D because it's the only option which allows variable execution

upvoted 1 times

gaurav0480 5 months ago

The key here is "after the load job with variable execution time completes" which means the execution of this job depends on the completion of another job which has a variable execution time. Hence D

upvoted 3 times

god\_brainer 5 months, 1 week ago

This approach ensures the dynamic triggering of the Dataprep job based on the completion of the preceding load job, ensuring data is processed accurately and in sequence

upvoted 1 times

Adswerve 9 months, 2 weeks ago

**Selected Answer: A**

A is correct. D is too complicated.

A is correct, because you can schedule a job right from Dataprep UI.

<https://cloud.google.com/blog/products/gcp/scheduling-and-sampling-arrive-for-google-cloud-dataprep>

Scheduling and sampling arrive for Google Cloud Dataprep

Throughout our early releases, users' most common request has been Flow scheduling. As of Thursday's release, Flows can be scheduled with minute granularity at any frequency.

upvoted 2 times

lucaluca1982 10 months, 1 week ago

**Selected Answer: C**

I think C it is more straightforward

upvoted 2 times

musumusu 11 months, 1 week ago

Answer C: Use Recipe Template feature of dataprep. Don't need to change the service.

upvoted 3 times

✉  **jroig\_** 1 year ago

**Selected Answer: C**

Why not C?

upvoted 1 times

✉  **zelli** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

upvoted 1 times

✉  **anicloudgirl** 1 year, 1 month ago

**Selected Answer: A**

It's A. You can set it directly in Dataprep a job and it will use Dataflow under the hood.

upvoted 4 times

✉  **anicloudgirl** 1 year, 1 month ago

It's A. You can set it directly in Dataprep a job and it will use Dataflow under the hood. No need to export nor incorporate into a Composer job  
Dataprep by trifacta - <https://docs.trifacta.com/display/DP/cron+Schedule+Syntax+Reference>

Dataprep job uses dataflow - <https://cloud.google.com/dataprep>

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

The question mentions after a load job with variable time, i dont think setting a dataprep cron job can address the issue of variable load tin

upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

**Selected Answer: D**

It's D

upvoted 2 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: D**

Dataprep and Dataflow are same family

upvoted 2 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

D. Export the Dataprep job as a Dataflow template, and incorporate it into a Composer job.

Reveal Solution

upvoted 4 times

Question #105

Topic 1

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Dataproc and Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

- A. cron
- B. Cloud Composer
- C. Cloud Scheduler
- D. Workflow Templates on Dataproc

**Correct Answer: D**

*Community vote distribution*

B (100%)

✉  **Sofia98** 3 weeks, 1 day ago

**Selected Answer: B**

Of course, it is Cloud Composer!

upvoted 1 times

✉  **Nirca** 3 months, 3 weeks ago

**Selected Answer: B**

B. Cloud Composer is the right answer !

upvoted 1 times

✉  **vamgcp** 6 months ago

**Selected Answer: B**

Cloud Composer is a managed service that allows you to create and run Apache Airflow workflows. Airflow is a workflow management platform that can be used to automate complex data pipelines. It is a good choice for this use case because it is a managed service, which means that Google will take care of the underlying infrastructure. It also supports multiple dependencies, so you can easily schedule a multi-step pipeline

upvoted 1 times

✉  **vaga1** 9 months, 1 week ago

**Selected Answer: B**

Airflow is the only choice to handle dependencies and being able to call all of the services included in the question

upvoted 2 times

✉  **niketd** 11 months, 1 week ago

**Selected Answer: B**

Multi-step sequential pipelines -> Cloud Composer

upvoted 1 times

✉  **AzureDP900** 1 year, 1 month ago

Cloud composer B is right

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: B**

Cloud Composer (Airflow) is the answer to chain different steps from different apps...

upvoted 1 times

✉  **MisuLava** 1 year, 3 months ago

**Selected Answer: B**

" multiple dependencies on each other. You want to use managed service"

= Cloud Composer

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

if you want your wf to schedule there are 3 ways to perform it, it of them is composer

<https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions>

upvoted 1 times

✉  **Remi2021** 1 year, 4 months ago

**Selected Answer: B**

composer :)

upvoted 1 times

✉  **YorelNation** 1 year, 4 months ago

**Selected Answer: B**

Composer

upvoted 1 times

✉  **AWSSandeep** 1 year, 4 months ago

**Selected Answer: B**

B. Cloud Composer

upvoted 1 times

 **damaldon** 1 year, 5 months ago

Use composer to schedule tasks

upvoted 1 times

Question #106

Topic 1

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

- A. Increase the cluster size with more non-preemptible workers.
- B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
- C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
- D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

**Correct Answer: D**

Reference:

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex>

Dataproc Enhanced Flexibility Mode (EFM) manages shuffle data to minimize job progress delays caused by the removal of nodes from a running cluster. EFM offloads shuffle data in one of two user-selectable modes:

1. Primary-worker shuffle. Mappers write data to primary workers. Workers pull from those remote nodes during the reduce phase. This mode is only available to, and is recommended for, Spark jobs.
2. HCFS (Hadoop Compatible File System) shuffle. Mappers write data to an HCFS implementation ([HDFS](#) by default). As with primary worker mode, only primary workers participate in HDFS and HCFS implementations (if HCFS shuffle uses the [Cloud Storage Connector](#), data is stored off-cluster). This mode can benefit jobs with small amounts of data, but due to scaling limitations, it is not recommended for larger jobs.

Since both EFM modes do not store intermediate shuffle data on secondary workers, EFM is well suited to clusters that use [preemptible VMs](#) or only [autoscale](#) the secondary worker group.

*Community vote distribution*

D (86%)

14%

✉  **AzureDP900** Highly Voted 1 year, 1 month ago

D is right

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters#using\\_graceful\\_decommissioning](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters#using_graceful_decommissioning)

upvoted 6 times

✉  **rocky48** Most Recent 1 month, 3 weeks ago

**Selected Answer: D**

D is right

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters#using\\_graceful\\_decommissioning](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters#using_graceful_decommissioning)

upvoted 1 times

✉  **Prakzz** 1 year, 1 month ago

**Selected Answer: A**

Should be A. You can configure the preemptible worker to gracefull decommission, its for non preemptible worker nodes.

upvoted 1 times

✉  **wan2three** 1 year ago

nope, they are not only for non-preeemtible workers

upvoted 1 times

✉  **yafsong** 1 year, 1 month ago

graceful decommissioning: to finish work in progress on a worker before it is removed from the Cloud Dataproc cluster.

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters>

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: D**

All your workers need to be the same kind. Use Graceful Decommissioning for don't lose any data and add more(increase the cluster) preemptible workers because there are more cost-effective .

upvoted 1 times

✉  **skp57** 1 year, 2 months ago

A. "graceful decommissioning" is not a configuration value but a parameter passed with scale down action - to decrease the number of workers to save money (see Graceful Decommissioning as an option to use when downsizing a cluster to avoid losing work in progress)

upvoted 2 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: D**

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters>

Why scale a Dataproc cluster?

to increase the number of workers to make a job run faster

to decrease the number of workers to save money (see Graceful Decommissioning as an option to use when downsizing a cluster to avoid losing work in progress).

to increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

upvoted 3 times

✉  **hauhau** 1 year, 3 months ago

This weird.

The question mentions that increase cluster, but Graceful Decommissioning use in downscale the cluster

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

All your workers need to be the same kind. Use Graceful Decommissioning for don't lose any data and add more preemptible workers because there are more cost-effective

upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

Question #107

Topic 1

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit tracking numbers when events are sent to Kafka topics. A recent software update caused the scanners to accidentally transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

- A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
- B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
- C. Use Cloud Logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
- D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention (Cloud DLP) API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

**Correct Answer: D**

*Community vote distribution*

D (100%)

 **dconesoko** Highly Voted 1 year, 1 month ago

**Selected Answer: D**

The cloud function with DLP seems the best option  
upvoted 5 times

 **PhilipKoku** Most Recent 6 months, 1 week ago

**Selected Answer: D**

DLP is required  
upvoted 1 times

 **HarshKothari21** 1 year, 4 months ago

**Selected Answer: D**

D option  
upvoted 1 times

 **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention (Cloud DLP) API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.  
upvoted 2 times

 **AzureDP900** 1 year, 1 month ago

Agreed

upvoted 1 times

Question #108

*Topic 1*

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets

information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

- A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.
- B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
- C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
- D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

**Correct Answer: D**

*Community vote distribution*

A (100%)

 **[Removed]** Highly Voted 3 years, 10 months ago

Should be A

upvoted 36 times

 **Rajokkiyam** Highly Voted 3 years, 10 months ago

Create dependency in Cloud Composer and schedule it.

upvoted 22 times

 **MisuLava** 1 year, 3 months ago

the jobs are not interdependent. just 3 individual jobs

upvoted 1 times

 **maxu** Most Recent 3 months, 1 week ago

yes answer A

upvoted 1 times

 **forepick** 8 months ago

Selected Answer: A

Cloud Composer. No doubt

upvoted 1 times

 **AzureDP900** 1 year, 1 month ago

A is correct

upvoted 1 times

 **dconesoko** 1 year, 1 month ago

Selected Answer: A

Cloud composer's DAG would manage the dependencies

upvoted 1 times

 **zellck** 1 year, 1 month ago

Selected Answer: A

A is the answer.

<https://cloud.google.com/composer/docs/concepts/overview>

Cloud Composer is a fully managed workflow orchestration service, enabling you to create, schedule, monitor, and manage workflows that span across clouds and on-premises data centers.

upvoted 2 times

 **danielfootc** 1 year, 6 months ago

This should be A

upvoted 2 times

 **medeis\_jar** 2 years ago

Selected Answer: A

<https://cloud.google.com/composer/docs/how-to/using/writing-dags>

upvoted 3 times

✉  **MaxNRG** 2 years ago

**Selected Answer: A**

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers.

<https://cloud.google.com/composer/?hl=en>

upvoted 5 times

✉  **kishanu** 2 years, 1 month ago

**Selected Answer: A**

A

Though the jobs are not dependent, they are data-driven. Refer to the below link:

<https://cloud.google.com/blog/topics/developers-practitioners/choosing-right-orchestrator-google-cloud>

upvoted 6 times

✉  **MaxNRG** 2 years ago

nice article thanks!

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

**Selected Answer: A**

Cloud Composer

upvoted 3 times

✉  **JG123** 2 years, 2 months ago

Correct: A

upvoted 3 times

✉  **sandipk91** 2 years, 5 months ago

should be option A

upvoted 5 times

✉  **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 4 times

✉  **someshsehgal** 2 years, 11 months ago

COOrrect A: Couldn't understand why a option with no connection with actual problem has been given as correct option (D)

upvoted 3 times

✉  **arghyat13** 3 years, 2 months ago

I'll go for A

upvoted 2 times

Question #109

Topic 1

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Cloud Logging. What are the two most likely causes of this problem? (Choose two.)

- A. Publisher throughput quota is too small.
- B. Total outstanding messages exceed the 10-MB maximum.
- C. Error handling in the subscriber code is not handling run-time errors properly.
- D. The subscriber code cannot keep up with the messages.
- E. The subscriber code does not acknowledge the messages that it pulls.

**Correct Answer: CE***Community vote distribution*

CE (80%)

DE (20%)

**✉️**  **TNT87** Highly Voted 1 year, 4 months ago

Answer C E

By not acknowledging the pulled message, this result in it be putted back in Cloud Pub/Sub, meaning the messages accumulate instead of being consumed and removed from Pub/Sub. The same thing can happen if the subscriber maintains the lease on the message it receives in case of an error. This reduces the overall rate of processing because messages get stuck on the first subscriber. Also, errors in Cloud Function do not show up in Stackdriver Log Viewer if they are not correctly handled.

upvoted 12 times

**✉️**  **mialli** Most Recent 9 months, 1 week ago**Selected Answer: DE**

Ref chatgpt

Option C, "Error handling in the subscriber code is not handling run-time errors properly," suggests that the subscriber code may not be correctly handling errors that occur during message processing. If the subscriber code encounters an error that it cannot handle, such as a syntax error or a network issue, it may stop processing messages, leading to a slowdown in message processing.

However, the lack of error logs in Cloud Logging suggests that there are no errors being logged, which makes it less likely that this is the primary cause of the observed behavior. Additionally, while incorrect error handling could contribute to the issue, it may not be the primary reason why the message processing rate is much higher than anticipated.

upvoted 1 times

**✉️**  **GCPete** 2 months, 2 weeks ago

Chat says about Option C: "it may stop processing messages, leading to a slowdown in message processing" - but it doesn't say there's a slowdown in the question. It says it's increased.

I would replace C with D. If the Cloud Function isn't capable of processing messages as quickly as they arrive, the backlog will grow, leading to higher processing rates as the function continuously tries to catch up. This scenario might not generate errors in Cloud Logging if the function is simply falling behind.

upvoted 1 times

**✉️**  **midgoo** 10 months, 3 weeks ago**Selected Answer: CE**

C - as no error shown in Cloud Logging

Between D & E, both could lead to the problem. I have worked with lot of PubSub issues, most of them are due to the bottleneck at the code where it takes too long to process 1 message and causes backlog. E could lead to backlog too, but it is too obvious and not likely to happen in reality.

However, when I ask AI the same question, it said C and E

upvoted 1 times

**✉️**  **cetanx** 7 months, 2 weeks ago

C. Error handling in the subscriber (Cloud Functions) code is not handling run-time errors properly.

This would mean to have error logs in Cloud Logging as CF by default logs to it.

upvoted 1 times

**✉️**  **musumusu** 11 months, 2 weeks ago

Answer D&amp;E

I am not in the favour of C, error handling is a side factor but not the primary cause.

First check the configuration access.

Does subscriber has enough acknowledge policies (option E)

Does sub have ability to keep up the message( enough network, cpu and capable codes) (option D)

option C is just a part of option D somewhere showing incapable handling

upvoted 2 times

✉  **desertlotus1211** 1 year ago

My question is: 'What is the actual problem?'  
- That there is no logs in Cloud Logging?  
- That Pub/Sub is having a problem?  
- Or there an actual problem?  
- Is there an actual error?

So what is Pub/Sub the message processing rate is high...Does that mean there is a problem?

Thoughts?

upvoted 3 times

✉  **squishy\_fishy** 5 months, 1 week ago

Like TNT87 mentioned the message processing rate is high "meaning the messages accumulate instead of being consumed and removed from Pub/Sub."  
upvoted 1 times

✉  **AzureDP900** 1 year, 1 month ago

C, E seems correct  
upvoted 1 times

✉  **MounicaN** 1 year, 4 months ago

D might also be right?  
Subscriber might not be provisioned enough  
upvoted 3 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: CE**  
C. Error handling in the subscriber code is not handling run-time errors properly.  
E. The subscriber code does not acknowledge the messages that it pulls.  
upvoted 3 times

Question #110

Topic 1

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

- A. Add a SideInput that returns a Boolean if the element is corrupt.
- B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
- C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
- D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

**Correct Answer: B**

*Community vote distribution*

B (100%)

✉  **[Removed]**  3 years, 10 months ago

Correct - B  
upvoted 16 times

✉  **[Removed]**  3 years, 10 months ago

Answer: B  
Description: ParDo is used to do transformation and create side output  
upvoted 12 times

midgoo Most Recent 10 months, 3 weeks ago

**Selected Answer: B**

A - SideInput is often used to validate data, however, we need to create the SideInput first. When using SideInput to filter data, it is actually another ParDo call.

C, D - This is common way to filter too, but we will need the key in order to partition or GroupByKey

B - ParDo is the most basic method, it can do anything to the PCollection

upvoted 3 times

AzureDP900 1 year, 1 month ago

B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.

upvoted 1 times

zellck 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/dataflow/docs/concepts/beam-programming-model#concepts>

ParDo is the core parallel processing operation in the Apache Beam SDKs, invoking a user-specified function on each of the elements of the input PCollection. ParDo collects the zero or more output elements into an output PCollection. The ParDo transform processes elements independently and possibly in parallel.

upvoted 3 times

Pime13 1 year, 6 months ago

**Selected Answer: B**

vote B :<https://beam.apache.org/documentation/programming-guide/#pardo>

Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it. Formatting or type-converting each element in a data set. If your input PCollection contains elements that are of a different type or format than you want, you can use ParDo to perform a conversion on each element and output the result to a new PCollection.

Extracting parts of each element in a data set. If you have a PCollection of records with multiple fields, for example, you can use a ParDo to parse out just the fields you want to consider into a new PCollection.

Performing computations on each element in a data set. You can use ParDo to perform simple or complex computations on every element, or certain elements, of a PCollection and output the results as a new PCollection.

upvoted 4 times

medeis\_jar 2 years ago

**Selected Answer: B**

Filtering with ParDo. ParDo is a Beam transform for generic parallel processing. ParDo is useful for common data processing operations/

upvoted 2 times

AzureDP900 1 year, 1 month ago

I agree with B

upvoted 1 times

MaxNRG 2 years ago

**Selected Answer: B**

B: ParDo is a Beam transform for generic parallel processing. ParDo is useful for common data processing operations, including:

a. Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection, or discard it.

b. Formatting or type-converting each element in a data set.

c. Extracting parts of each element in a data set.

d. Performing computations on each element in a data set.

A does not help

C Partition is a Beam transform for PCollection objects that store the same data type. Partition splits a single PCollection into a fixed number of smaller collections. Again, does not help

D GroupByKey is a Beam transform for processing collections of key/value pairs. GroupByKey is a good way to aggregate data that has something in common

upvoted 6 times

sumanshu 2 years, 7 months ago

vote for 'B', ParDo can discard the elements.

<https://beam.apache.org/documentation/programming-guide/>

upvoted 4 times

✉  **DeepakKhattar** 3 years ago

B - seems to be better option since we need to filter out, question does not specify that we do need to store it into different Pcollection.  
<https://beam.apache.org/documentation/transforms/python/overview/>  
ParDo is general purpose whereas partition splits the elements into do different pcollections.  
<https://beam.apache.org/documentation/transforms/python/elementwise/partition/>  
upvoted 3 times

✉  **arghya13** 3 years, 2 months ago

B is correct

upvoted 3 times

✉  **SteelWarrior** 3 years, 4 months ago

Should be B. The Partition transform would require the element identifying the valid/invalid records for partitioning the pcollection that means there is some logic to be executed before the Partition transformation is invoked. That logic can be implemented in a ParDO transform and will can both identify valid/invalid records and also generate two PCollections one with valid records and other with invalid records.  
upvoted 7 times

✉  **haroldbenites** 3 years, 5 months ago

B is correct

upvoted 3 times

✉  **Archy** 3 years, 6 months ago

B, ParDo is useful for a variety of common data processing operations, including:

Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it.  
upvoted 4 times

✉  **tprashanth** 3 years, 6 months ago

Looks like C it is

<https://beam.apache.org/documentation/programming-guide/>

upvoted 2 times

✉  **atnafu2020** 3 years, 5 months ago

according this link its

ParDo

\* Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it.

\* But Partition just splitting which is a Beam transform for PCollection objects that store the same data type. Partition splits a single PCollection into a fixed number of smaller collections.

upvoted 5 times

✉  **xrun** 3 years, 1 month ago

Seems like two answers may be correct. With ParDo you can discard corrupt data. With Partition you can split the data into two PCollections: corrupt and ok. You stream ok data further to BigQuery and corrupt data to some other storage for analysis. If one is not interested in analysis, then ParDo is enough.

upvoted 1 times

✉  **dg63** 3 years, 6 months ago

Correct answer should be "C". A Pardo transform will allow the processing to happen in parallel using multiple workers. Partition transform allows data to be partitioned into two different PCollections according to some logic. Using partition transform once can split the corrupted data and finally discard it.

upvoted 5 times

✉  **Rajuuu** 3 years, 6 months ago

Correct B.

upvoted 4 times

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the

Data Science team runs a query filtered on a date column and limited to 30~90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries.

What should you do?

- A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.
- B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
- C. Modify your pipeline to maintain the last 3090~ days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
- D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

**Correct Answer: A**

*Community vote distribution*

A (100%)

✉  [Removed]  3 years, 10 months ago

should be A

upvoted 35 times

✉  [Removed]  3 years, 10 months ago

Answer: A

Description: Partition is the solution for reducing cost and time

upvoted 18 times

✉  willbot 3 years, 8 months ago

but how would recreating tables with 3 years of data, maintain the ability to conduct sql queries during that time?

upvoted 1 times

✉  squishy\_fishy 2 years, 3 months ago

Recreating the new table, the old table will still have new data coming, then append the difference to the new table.

upvoted 2 times

✉  odacir  1 year, 1 month ago

**Selected Answer: A**

Answer: A, has no cost to reload the data, Also Partition is the solution for reducing cost and time

upvoted 1 times

✉  zellck 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/bigquery/docs/partitioned-tables>

A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

You can partition BigQuery tables by:

- Time-unit column: Tables are partitioned based on a TIMESTAMP, DATE, or DATETIME column in the table.

upvoted 3 times

✉  AzureDP900 1 year, 1 month ago

A is right

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

it is not B in the sense of cost-effective certainly. read below in limitation

<https://cloud.google.com/bigquery/docs/querying-wildcard-tables#limitations>

Currently, cached results are not supported for queries against multiple tables using a wildcard even if the Use Cached Results option is checked. If you run the same wildcard query multiple times, you are billed for each query.

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

[https://cloud.google.com/bigquery/docs/partitioned-tables#dt\\_partition\\_shard](https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard)

Partitioning is recommended over table sharding, because partitioned tables perform better

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

A AND D , they are the most likely choiced but the question want

issue as cost-effectively as possible while maintaining the ability to conduct SQL queries.

1 table may be cheaper so partition is better than wildcard

upvoted 1 times

✉  **Didine\_22** 1 year, 9 months ago

**Selected Answer: A**

answer A

upvoted 2 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: A**

<https://cloud.google.com/bigquery/docs/partitioned-tables>

upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: A**

A. Partitioning

<https://cloud.google.com/bigquery/docs/partitioned-tables>

upvoted 1 times

✉  **Tomi1313** 2 years, 1 month ago

Why not D? You can use SQL.

This is the cheapest and fastest option

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

upvoted 2 times

✉  **John\_Pongthorn** 1 year, 4 months ago

Partitioning is recommended over table sharding, because partitioned tables perform better

This is a google recommendation nowaday.

upvoted 1 times

✉  **StefanoG** 2 years, 2 months ago

**Selected Answer: A**

The D solution is obviously discarded.

The request NOT require ONLY LAST 30-90 days, so the C solution is not the right solution.

In addition to this, the request ask to keep the possibility to made queries, so B is worst.

Is not mandatory make the queries while you make the modify so the right answer is A

upvoted 4 times

✉  **JayZeeLee** 2 years, 2 months ago

B sounds more feasible.

The point is 'historical' data, not new table/data. Recreating tables from the past three years is a lot of work. Might as well export the table and run analyses there. No cost for exporting in BigQuery.

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 5 times

✉  **arghya13** 3 years, 2 months ago

I will go with Option A

upvoted 5 times

✉  **Alasmindas** 3 years, 2 months ago

I will go with Option A, although at first instance I felt Option C would be correct.

Option A : Because partitioning will help to address both the concerns mentioned in the question - i.e. faster query and reducing cost.

Option C : Modifying the data pipeline to store last 30-90 days data would have possible, if there was a point mentioned that only the latest data (30-90 days) is kept and the older data - beyond 90 days is moved to the master table. Since that point is not mentioned, we will land up having multiple - 30-90 days data in separate tables + the master table.

upvoted 3 times

✉  **karthik89** 2 years, 11 months ago

but how will you append the data that is older than 90days in to the master table?

upvoted 2 times

✉  **Cloud\_Enthusiast** 3 years, 2 months ago

Answer is A. Recreating the DDL with new partition is easy and does not require any changes on applications that read data from it

upvoted 4 times

Question #112

Topic 1

You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

- A. Deploy small Kafka clusters in your data centers to buffer events.
- B. Have the data acquisition devices publish data to Cloud Pub/Sub.
- C. Establish a Cloud Interconnect between all remote data centers and Google.
- D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

**Correct Answer: B**

*Community vote distribution*

B (71%)

13%

Other

✉  **[Removed]**  3 years, 10 months ago

Should be B

upvoted 31 times

👤 **Ganshank** Highly Voted 3 years, 9 months ago

C.

This is a tricky one. The issue here is the unreliable connection between data collection and data processing infrastructure, and to resolve it in cost-effective manner. However, it also mentions that the company is using leased lines. I think replacing the leased lines with Cloud InterConnect would solve the problem, and hopefully not be an added expense.

<https://cloud.google.com/interconnect/docs/concepts/overview>

upvoted 22 times

👤 **fire558787** 2 years, 5 months ago

Disagree. OK, the problem is between the data centers and Google's data centers. However, Cloud Interconnect costs money. If the devices write to PubSub instead, they would bypass their own data centers and write directly to Google. It seems that connectivity is not a problem for the collecting devices, since it says that data is collected in local data centers. So my guess goes towards PubSub rather than Interconnect.

upvoted 17 times

👤 **Catweazle1983** 1 year, 1 month ago

Thanks for this explanation. I was also going for C initially. But you convinced me that B is the correct answer.

upvoted 1 times

👤 **Yiouk** 6 months, 2 weeks ago

C. Can you imagine changing the software in all sensors to use PubSub instead of the existing one? This is out of scope of the question.

upvoted 1 times

👤 **serg3d** 3 years, 8 months ago

Yea, this would definitely solve the issue, but it's not "the most cost-effective way". I think PubSub is the correct answer.

upvoted 7 times

👤 **snamburi3** 3 years, 2 months ago

the question also talks about a cost effective way...

upvoted 3 times

👤 **Nandababy** Most Recent 1 month, 2 weeks ago

Even with Cloud Pub/Sub, unpredictable latency or delays could still occur due to the unreliable leased lines connecting your event collection infrastructure and event processing infrastructure. While Cloud Pub/Sub offers reliable message delivery within its own network, the handoff to your processing infrastructure is still dependent on the leased lines.

Replacing leased lines with Cloud Interconnect could potentially resolve the overall issue of unpredictable latency in event processing pipeline but it could be unnecessary expense provided data centers distributed world wide.

Cloud Pub/Sub along with other optimization techniques like Cloud VPN or edge computing might be sufficient.

upvoted 1 times

👤 **FP77** 5 months, 2 weeks ago

**Selected Answer: C**

I don't know why B is the most voted. The issue here is unreliable connectivity and C is the perfect use-case for that

upvoted 1 times

👤 **NeoNitin** 5 months, 3 weeks ago

its says with unpredictable latency and here no need to worry about connection

So B is the right one

upvoted 1 times

👤 **ZZHZZH** 6 months, 3 weeks ago

**Selected Answer: C**

The question is misleading. But should be C since it addresses the unpredictability and latency directly.

upvoted 1 times

👤 **musumusu** 11 months, 2 weeks ago

Best answer is A, By using Kafka, you can buffer the events in the data centers until a reliable connection is established with the event processing infrastructure.

But go with B, its google asking :P

upvoted 2 times

👤 **musumusu** 11 months, 1 week ago

I read this question again, I wanna answer C. Buying Data acquisition devices and set them up with sensor, i dont think its practical approach. Imagine, Adruino is cheapest IOT available in market for 15 dollars, but who will open the sensor box and install it .. omg,, its a big job. This question depends if IOT devices that are attached to sensor needs to be programmed. Big Headache right. Use google cloud connect to connect with current situation. Or reprogramme IOT if they have connected with sensors.

upvoted 1 times

ayush\_1995 1 year ago

Selected Answer: B

B. Have the data acquisition devices publish data to Cloud Pub/Sub. This would provide a reliable messaging service for your event data, allowing you to ingest and process your data in a timely manner, regardless of the reliability of the leased lines. Cloud Pub/Sub also offers automatic retries and fault-tolerance, which would further improve the reliability of your event delivery. Additionally, using Cloud Pub/Sub would allow you to easily scale up or down your event processing infrastructure as needed, which would help to minimize costs.

upvoted 10 times

desertlotus1211 1 year ago

Are they talking about GCP in this question?  
Where is the event processing infrastructure?

Answer A, might be correct!

upvoted 2 times

PrashantGupta1616 1 year, 1 month ago

Selected Answer: B

pub/sub is region is a global service

It's important to note that the term "global" in this context refers to the geographical scope of the service

upvoted 1 times

NicolasN 1 year, 1 month ago

Selected Answer: A

As usual the answer is hidden somewhere in the Google Cloud Blog:

"In the case of our automotive company, the data is already stored and processed in local data centers in different regions. This happens by streaming all sensor data from the cars via MQTT to local Kafka Clusters that leverage Confluent's MQTT Proxy."

"This integration from devices to a local Kafka cluster typically is its own standalone project, because you need to handle IoT-specific challenges like constrained devices and unreliable networks."

<https://cloud.google.com/blog/products/ai-machine-learning/enabling-connected-transformation-with-apache-kafka-and-tensorflow-on-google-cloud-platform>

upvoted 2 times

desertlotus1211 1 year ago

The question is asking from the on-premise infrastructure, which already has the data, to the event processing infrastructure, which is in the GCP, is unreliable....

it not asking from the sensors to the on-premise...

upvoted 2 times

desertlotus1211 1 year ago

I might have to retract my answer... Are they talking about GCP in this question?  
where is the event processing infrastructure?

upvoted 1 times

zellck 1 year, 1 month ago

Selected Answer: B

B is the answer.

upvoted 2 times

AzureDP900 1 year, 1 month ago

yes it is B. Have the data acquisition devices publish data to Cloud Pub/Sub.  
upvoted 1 times

✉  **piotrpiskorski** 1 year, 2 months ago

yeah, changing whole architecture arround the world for the use of pub/sub is so much more cost efficient than Cloud Interconnect (which is 3k\$)..

It's C.

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

Wouldn't using cloud interconnect also result in amendments to each of the data center around the world? I don't see why there would be huge architecture change when using PubSub, the publishers would just need to push messages directly to pubsub, instead of pushing to their own cost center.

Also, if the script for pushing messages can be standardised, the data centers can share it around to

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

It's not a Cloud Interconnect, it's a lot of interconnect ones per data center, PUB/SUB addresses all the requirements. Its B

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

ALSO, the problem it's not your connection, its the connectivity BT your event collection infrastructure to your event processing infrastructure, so PUSUB it's perfect for this

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: B**

Cloud Pub/Sub, it supports batch & streaming , push and pull capabilities

Answer B

upvoted 1 times

✉  **t11** 1 year, 5 months ago

It has to be B.

upvoted 1 times

✉  **rr4444** 1 year, 5 months ago

**Selected Answer: D**

Feels like everyone is wrong.

A. Deploy small Kafka clusters in your data centers to buffer events.

- Silly in a GCP cloudnative context, plus they have messaging infra anyway

B. Have the data acquisition devices publish data to Cloud Pub/Sub.

- They have messaging infra, so why? Unless they want to replace, it, but that doesn't change the issue

C. Establish a Cloud Interconnect between all remote data centers and Google.

- Wrong, because Interconnect is basically a leased line. There must be some telecoms issue with it, which we can assume is unresolvable e.g. long distance remote locations and sometimes water ingress, and the telco can't justify sorting it yet, or is slow to, or something. Leased lines usually don't come with awful internet connectivity, so sound physical connectivity issue. Sure, an Interconnect is better, more direct, but a leased line should be bullet proof.

D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

- The only way to address dodgy/delayed data delivery

upvoted 2 times

✉  **NM1212** 1 year, 6 months ago

The answer must be C. To me the question is more about having to re-use existing Infrastructure as much as you can and try to only fix the problem that is mentioned, that is what they mean as cost optimization, so there is no need to overthink. Since there is no issue with event collection centers, so there is really no need to change up to Pub/Sub. Using Cloud Interconnect, they switch from unreliable network to a full reliable Cloud network. That should solve the problem here.

upvoted 1 times

Question #113

Topic 1

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

- A. Speech-to-Text API
- B. Cloud Natural Language API
- C. Dialogflow Enterprise Edition
- D. AutoML Natural Language

**Correct Answer: C**

*Community vote distribution*

C (77%)

A (23%)

✉  **rickywck** Highly Voted 3 years, 10 months ago  
should be C, since we need to recognize both voice and intent  
upvoted 26 times

✉  **AzureDP900** 1 year, 1 month ago  
C. Dialogflow Enterprise Editio  
upvoted 1 times

✉  **Alasmindas** Highly Voted 3 years, 2 months ago

Option A - Cloud Speech-to-Text API.

The question is just asking to "interpret customer voice commands" .. it does not mention anything related to sentiment analysis so NLP is not required. DialogFlow is more of a chat bot services typically suited for a "Service Desk" kind of setup - where clients will call a centralized helpdesk and automation is achieved through Chat bot services like - google Dialog flow

upvoted 18 times

✉  **exnanianwort** 4 months, 1 week ago

should be C, the key is interpret customer voice commands

upvoted 1 times

✉  **HarshKothari21** 1 year, 4 months ago

Question also says "in-home assistants, such as Google Home". the idea here is to provide assistance which involves Dialog.

I would go with option C

upvoted 2 times

✉  **hdmi\_switch** 2 years, 6 months ago

Cloud Speech-to-Text API just converts speech to text. You will have text files as an output and then the requirement is to "interpret customer voice commands and issue an order to the backend systems". This is not achieved by having text files.

I would go with option C, since Dialogflow can interpret the commands (intents) and integrates other applications e.g. backend systems.

upvoted 8 times

✉  **NeoNitin** Most Recent 5 months, 3 weeks ago

Ans C . main thing is that question is saying "customer voice commands" there is no need to sentimental analysis of language so that's why.

C. Dialogflow Enterprise Edition

Dialogflow is a powerful natural language understanding platform developed by Google. It allows you to build conversational interfaces, interpret user voice commands, and integrate with various platforms and devices like Google Home. The "Enterprise Edition" provides additional features and support for more complex use cases, making it a good choice for a retailer looking to integrate with in-home assistants and handle customer voice commands effectively.

upvoted 1 times

✉  **juliobs** 10 months, 2 weeks ago

**Selected Answer: C**

Answer is C. However Google Assistant Conversational Actions will be sunsetted on June 13, 2023.

upvoted 2 times

✉  **techtitan** 11 months, 2 weeks ago

**Selected Answer: C**

<https://cloud.google.com/dialogflow/es/docs/integrations/aog>

upvoted 1 times

✉  **desertlotus1211** 1 year ago

I think the answer is A: Speech to Text.

You want to interpret what a user say... Dialogflow is text to speech, not what the question asked for...

Thoughts?

upvoted 1 times

✉  **PrashantGupta1616** 1 year, 1 month ago

**Selected Answer: A**

The question is just asking to "interpret customer voice commands" so A is out of the box solution

upvoted 1 times

👤 **odacir** 1 year, 1 month ago

**Selected Answer: A**

Enable voice control

Implement voice commands such as “turn the volume up,” and voice search such as saying “what is the temperature in Paris?” Combine this with the Text-to-Speech API to deliver voice-enabled experiences in IoT (Internet of Things) applications.

<https://cloud.google.com/speech-to-text#section-9>

upvoted 1 times

👤 **odacir** 1 year, 1 month ago

I change my mind, it's C.

<https://cloud.google.com/blog/products/gcp/introducing-dialogflow-enterprise-edition-a-new-way-to-build-voice-and-text-conversational-apps>

upvoted 5 times

👤 **zellck** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

<https://cloud.google.com/dialogflow/es/docs>

Dialogflow is a natural language understanding platform that makes it easy to design and integrate a conversational user interface into your mobile app, web application, device, bot, interactive voice response system, and so on. Using Dialogflow, you can provide new and engaging ways for users to interact with your product.

Dialogflow can analyze multiple types of input from your customers, including text or audio inputs (like from a phone or voice recording). It can also respond to your customers in a couple of ways, either through text or with synthetic speech.

upvoted 2 times

👤 **TNT87** 1 year, 4 months ago

<https://cloud.google.com/blog/products/gcp/introducing-dialogflow-enterprise-edition-a-new-way-to-build-voice-and-text-conversational-apps>

Dialogflow is the answer

upvoted 2 times

👤 **Smaks** 1 year, 6 months ago

**Selected Answer: C**

Dialogflow provides a seamless integration with Google Assistant. This integration has the following advantages: You can use the same Dialogflow agent to power Google Assistant and other integrations. Dialogflow agents provide Google Cloud enterprise-grade security, privacy support, and SLAs

upvoted 2 times

👤 **Vip777** 1 year, 9 months ago

dialog

upvoted 1 times

👤 **Vip777** 1 year, 9 months ago

speech

upvoted 1 times

👤 **PJG\_worm** 1 year, 10 months ago

It should be D. INTERPRET customer voice commands and issue an order to the backend systems. Option C is usually applied for conversations. But in this case, it is not a conversation.

upvoted 1 times

👤 **rytizzle** 1 year, 6 months ago

it can totally be a conversation. if the system is attempting to interpret voice commands, it may need clarifying questions, which would indicate it's a conversation. it's not D because autoML NL analyzes TEXT. it would have to work in conjunction with speech to text API. therefore the answer is dialogflow.

upvoted 1 times

👤 **medeis\_jar** 2 years ago

**Selected Answer: C**

recognize voice and intent

<https://cloud.google.com/blog/products/gcp/introducing-dialogflow-enterprise-edition-a-new-way-to-build-voice-and-text-conversational-apps>

upvoted 3 times

👤 **MaxNRG** 2 years ago

**Selected Answer: C**

C: Dialogflow Enterprise Edition is an end-to-end development suite for building conversational interfaces for websites, mobile applications, popular messaging platforms, and IoT devices. You can use it to build interfaces (e.g., chatbots) that are capable of natural and rich interaction between your users and your business. It is powered by machine learning to recognize the intent and context of what a user says, allowing your conversational interface to provide highly efficient and accurate responses.

<https://cloud.google.com/dialogflow/>

Dialogflow API V2 is the new iteration of our developer API. The new API integrates Google Cloud Speech-to-Text, enabling developers to serve audio directly to Dialogflow for combined speech recognition and natural language understanding.

<https://dialogflow.com/v2-faq>

<https://cloud.google.com/blog/products/gcp/introducing-dialogflow-enterprise-edition-a-new-way-to-build-voice-and-text-conversational-apps>

upvoted 5 times

👤 **MasterOfTheUniverse** 2 years, 1 month ago

**Selected Answer: A**

Speech-to-text is mentioned in Google's use cases.

It is used for transcription.

AutoML is still mentioned though which is opening up other services to be included.

The question is not so fair and clear.

<https://cloud.google.com/speech-to-text#section-9>

Question #114

Topic 1

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

- A. Cloud Dataflow
- B. Cloud Composer
- C. Cloud Dataprep
- D. Cloud Dataproc

**Correct Answer: D**

*Community vote distribution*

B (100%)

👤 **madhu1171** Highly Voted 3 years, 10 months ago

Answer should be B

upvoted 29 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Answer - B

upvoted 12 times

👤 **forepick** Most Recent 8 months ago

**Selected Answer: B**

No other option is aimed for this purpose

upvoted 1 times

👤 **juliobs** 10 months, 2 weeks ago

**Selected Answer: B**

Airflow

upvoted 1 times

👤 **PrashantGupta1616** 1 year, 1 month ago

**Selected Answer: B**

Cloud Composer is Airflow

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: B**

Cloud Composer is Airflow, It's made for this job.

upvoted 3 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/composer/docs/concepts/overview>

Cloud Composer is a fully managed workflow orchestration service, enabling you to create, schedule, monitor, and manage workflows that span across clouds and on-premises data centers.

upvoted 4 times

✉  **AzureDP900** 1 year, 1 month ago

Cloud composer is right

upvoted 2 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: B**

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

upvoted 1 times

✉  **MaxNRG** 2 years ago

**Selected Answer: B**

B:

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers.

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

Cloud Composer can help create workflows that connect data, processing, and services across clouds, giving you a unified data environment. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use.

Cloud Composer gives you the ability to connect your pipeline through a single orchestration tool whether your workflow lives on-premises, in multiple clouds, or fully within GCP. The ability to author, schedule, and monitor your workflows in a unified manner means you can break down the silos in your environment and focus less on infrastructure.

upvoted 6 times

✉  **MaxNRG** 2 years ago

Option A is wrong as Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It is not a multi-cloud orchestration tool. Option B is wrong as Google Cloud Dataflow is a fully managed service for strongly consistent, parallel data-processing pipelines. It does not support multi-cloud handling.

Option D is wrong as Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing.

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

**Selected Answer: B**

Answer: B

upvoted 3 times

✉  **sandipk91** 2 years, 5 months ago

Cloud composer

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for B

upvoted 5 times

✉  **daghayeghi** 2 years, 10 months ago

B:

Hybrid and multi-cloud

Ease your transition to the cloud or maintain a hybrid data environment by orchestrating workflows that cross between on-premises and the public cloud. Create workflows that connect data, processing, and services across clouds to give you a unified data environment.

<https://cloud.google.com/composer#section-2>

upvoted 4 times

👤 **someshsehgal** 2 years, 11 months ago

Correct B: without any doubt.

upvoted 3 times

👤 **arghya13** 3 years, 2 months ago

B-Cloud Composer works on a multicloud environment

upvoted 4 times

👤 **Alasmindas** 3 years, 2 months ago

there can not be any simple question like this to choose the right answer as "Cloud Composer". I really feel someone must have deliberately selecting the wrong answers in Exam topics to confuse people....

upvoted 5 times

👤 **Cloud\_Enthusiast** 3 years, 2 months ago

Composer is the obvious answer. so B

upvoted 4 times

Question #115

Topic 1

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

- A. Use Analytics Hub to control data access, and provide third party companies with access to the dataset.
- B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.
- C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.
- D. Create a Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

**Correct Answer: A**

*Community vote distribution*

A (100%)

👤 **LP\_PDE**  1 year, 3 months ago

I feel the answer really should be Create an authorized view on the BigQuery table to control data access, and provide third-party companies access to that view.

upvoted 21 times

👤 **zellick**  1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/bigquery/docs/analytics-hub-introduction>

Analytics Hub is a data exchange platform that enables you to share data and insights at scale across organizational boundaries with a robust security and privacy framework.

As an Analytics Hub publisher, you can monetize data by sharing it with your partner network or within your own organization in real time. List let you share data without replicating the shared data. You can build a catalog of analytics-ready data sources with granular permissions that you deliver data to the right audiences. You can also manage subscriptions to your listings.

upvoted 9 times

👤 **vamgcp**  6 months, 1 week ago

**Selected Answer: A**

Option A: This option is correct because Analytics Hub is a managed service that provides a centralized repository for data assets. You can use Analytics Hub to share data with other Google Cloud Platform services, as well as with third-party companies

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

You are preparing for exam:

Creating a view and share with 3rd party is best and cheapest.

Then create a separate dataset to share it cost less than using paid service for data access i.e analytics hub where you create data access policies

you choose, its just making me craazy

upvoted 1 times

✉  **musumusu** 11 months, 1 week ago

One main reason you should use analytics hub, when you want control over 3 party activites and you want to monetize ( to make money ) I sharing BQ dataset.

upvoted 1 times

✉  **lool** 1 year ago

**Selected Answer: A**

Shared datasets are collections of tables and views in BigQuery defined by a data publisher and make up the unit of cross-project / cross-organizational sharing. Data subscribers get an opaque, read-only, linked dataset inside their project and VPC perimeter that they can combin with their own datasets and connect to solutions from Google Cloud or our partners. For example, a retailer might create a single exchange to share demand forecasts to the 1,000's of vendors in their supply chain--having joined historical sales data with weather, web clickstream, and Google Trends data in their own BigQuery project, then sharing real-time outputs via Analytics Hub. The publisher can add metadata, track subscribers, and see aggregated usage metrics.

upvoted 2 times

✉  **AzureDP900** 1 year, 1 month ago

A. Use Analytics Hub to control data access, and provide third party companies with access to the dataset.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: A**

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

upvoted 2 times

✉  **Atnafu** 1 year, 2 months ago

A

Multiple choose listed wrongly

Correct one

A.

Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.

B.

Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.

C.

Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.

D.

Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bu for third-party companies to use.

upvoted 1 times

✉  **ajayrtk** 1 year, 3 months ago

no option is correct

this is correct answer -Create an authorised view on the BigQuery table to control data access, and provide third-party companies with acces that view.

upvoted 2 times

✉  **AWSSandeep** 1 year, 4 months ago

**Selected Answer: A**

A. Use Analytics Hub to control data access, and provide third party companies with access to the dataset.

upvoted 1 times

✉️  **damaldon** 1 year, 4 months ago

Answer A.

As an Analytics Hub user, you can perform the following tasks:

As an Analytics Hub publisher, you can monetize data by sharing it with your partner network or within your own organization in real time. List let you share data without replicating the shared data. You can build a catalog of analytics-ready data sources with granular permissions that you deliver data to the right audiences.

As an Analytics Hub subscriber, you can discover the data that you are looking for, combine shared data with your existing data, and leverage built-in features of BigQuery. When you subscribe to a listing, a linked dataset is created in your project.

As an Analytics Hub viewer, you can browse through the datasets that you have access to in Analytics Hub and request the publisher to access the shared data.

As an Analytics Hub administrator, you can create data exchanges that enable data sharing, and then give permissions to data publishers and subscribers to access these data exchanges.

<https://cloud.google.com/bigquery/docs/analytics-hub-introduction>

upvoted 1 times

Question #116

Topic 1

Your company is in the process of migrating its on-premises data warehousing solutions to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply updates from multiple transactional database sources on a daily basis. With BigQuery, your company hopes to improve its handling of

CDC so that changes to the source systems are available to query in BigQuery in near-real time using log-based CDC streams, while also optimizing for the performance of applying changes to the data warehouse. Which two steps should they take to ensure that changes are available in the BigQuery reporting table with minimal latency while reducing compute overhead? (Choose two.)

- A. Perform a DML INSERT, UPDATE, or DELETE to replicate each individual CDC record in real time directly on the reporting table.
- B. Insert each new CDC record and corresponding operation type to a staging table in real time.
- C. Periodically DELETE outdated records from the reporting table.
- D. Periodically use a DML MERGE to perform several DML INSERT, UPDATE, and DELETE operations at the same time on the reporting table.
- E. Insert each new CDC record and corresponding operation type in real time to the reporting table, and use a materialized view to expose only the newest version of each unique record.

**Correct Answer: CD**

*Community vote distribution*

BD (81%)

Other

✉️  **YorelNation** Highly Voted 1 year, 4 months ago

**Selected Answer: BD**

To aim for minimal latency while reducing compute overhead:

- B. Insert each new CDC record and corresponding operation type to a staging table in real time.

D. Periodically use a DML MERGE to perform several DML INSERT, UPDATE, and DELETE operations at the same time on the reporting table statements comes from the staging table)

upvoted 15 times

✉  **musumusu** Highly Voted 11 months, 1 week ago

B&D

Tricks here: Always choose google recommended approach, Use data first in Staging table then merge with original tables.

upvoted 6 times

✉  **Nirca** Most Recent 3 months, 3 weeks ago

**Selected Answer: CE**

I'm going for E &C - this is the only solution with low TCO.

E - is the best way to work with CDC when real nearline data is needed BQ snapshots can be online! . & C - is good practice to delete old records.

upvoted 1 times

✉  **dconesoko** 1 year, 1 month ago

**Selected Answer: BD**

with both the delta table and the main table changes could be queried in near realtime, by using a view that unions both tables and queries th laters record for the given key, eventually the delta table should be merged into the main table and truncated. Google recently introduced datastream that would take away all these headaches.

upvoted 2 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: BD**

The solution is B and D. I perform a similar task in my work, and this is the best way to do it at scale with BigQuery.

upvoted 2 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: BD**

BD is the answer.

[https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#overview\\_of\\_cdc\\_data\\_replication](https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#overview_of_cdc_data_replication)  
Delta tables contain all change events for a particular table since the initial load. Having all change events available can be valuable for identif trends, the state of the entities that a table represents at a particular moment, or change frequency.

The best way to merge data frequently and consistently is to use a MERGE statement, which lets you combine multiple INSERT, UPDATE, and DELETE statements into a single atomic operation.

upvoted 4 times

👤 **NicolasN** 1 year, 2 months ago

I really can't find a correct combination of answers. I'm between the following alternatives, but with no one fitting:

1 [B] and [D]: That's a proposed solution, but as a cost-optimized approach (along with an extra step to "Periodically DELETE outdated reco from the STAGING table" - more details on my subsequent reply). Also, I can't imagine how an answer with the word "Periodically" may be compatible with the "minimal latency" requirement.

2 [E] and [C]: It could be a valid approach, but near-real time requirement would demand also for a materialized view refresh. And it seems to contradict the "reducing compute overhead" req.

3 [A] standalone: Provides immediate results but is far from compute-optimized.

upvoted 2 times

👤 **NicolasN** 1 year, 2 months ago

Nowadays (Nov. 2022) I don't expect to confront this question in a real exam with this set of answers since the more recent documentation proposes the use of Datastream.

🔗 <https://cloud.google.com/blog/products/data-analytics/real-time-cdc-replication-bigquery>

upvoted 4 times

👤 **NicolasN** 1 year, 2 months ago

The previous guidelines were here:

🔗 [https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#immediate\\_consistency\\_approach](https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#immediate_consistency_approach)

There were two approaches:

1 Immediate consistency approach

2 Cost-optimized approach

For approach 1, which is the objective of this question, it proposes:

a. Insert CDC data into a delta table in BigQuery => that's answer [B]

b. Create a BigQuery view that joins the main and delta tables and finds the most recent row => there's no answer that fits

For approach 2 it proposes:

a. Insert CDC data into a delta table in BigQuery => that's answer [B]

b. Merge delta table changes into the main table and periodically purge merged rows from the delta table - Run Merge statement on a regular interval => that's answer [D]

upvoted 4 times

👤 **beanz00** 1 year, 3 months ago

B and E. Typically in a Data Warehouse you don't delete data. Data Warehouse should store full history to see how the data changed over time. All the solutions with 'DELETE' should not be used as this goes against being able to access the history of the data.

upvoted 2 times

👤 **TNT87** 1 year, 3 months ago

<https://www.striim.com/blog/oracle-to-google-bigquery/>

upvoted 1 times

👤 **TNT87** 1 year, 4 months ago

[https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#data\\_latency](https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#data_latency)

upvoted 1 times

👤 **TNT87** 1 year, 4 months ago

<https://docs.streamsets.com/platform-datacollector/latest/datacollector/UserGuide/Destinations/GBigQuery.html>

upvoted 1 times

👤 **Wasss123** 1 year, 4 months ago

**Selected Answer: BD**

B and D

<https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture>

upvoted 3 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: BC**

B D you have to do the both to get it done.

To merge process, you have to perform between the report table and stage a table

upvoted 2 times

👤 **Remi2021** 1 year, 4 months ago

Answers are tricky, official documentation suggests Dataflow or Datafusion path as well as inclusion of DataStreams

<https://cloud.google.com/blog/products/data-analytics/real-time-cdc-replication-bigquery>

upvoted 1 times

least once and must be ordered within windows of 1 hour. How should you design the solution?

- A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.
- B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.
- C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.
- D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

**Correct Answer: D**

*Community vote distribution*

D (100%)

 **madhu1171** Highly Voted 3 years, 10 months ago

Answer should be D  
upvoted 26 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer - D  
upvoted 13 times

 **NeoNitin** Most Recent 5 months, 3 weeks ago

Data proc is serverbased  
Dataflow is serverless which is used to run pipelines which uses apache framework in the background. Just need to mention the number of workers needed.

so question saying we need scale automatically . so dataproc eliminate ho gaya  
now Dataflow is correct , pub/sub is recommended for this scenario. D

upvoted 1 times

 **dconesoko** 1 year, 1 month ago

Selected Answer: D  
google's preferred choice  
upvoted 1 times

 **zelliCK** 1 year, 1 month ago

D is the answer.  
upvoted 1 times

 **Pime13** 1 year, 6 months ago

Selected Answer: D  
Answer should be D  
upvoted 1 times

 **VictorBa** 1 year, 9 months ago

Selected Answer: D  
It cannot be C because Dataproc is more suitable for Hadoop jobs.  
upvoted 1 times

 **medeis\_jar** 2 years ago

Selected Answer: D  
Pub/Sub + Dataflow  
upvoted 1 times

 **MaxNRG** 2 years ago

Selected Answer: D  
D: Pub/Sub + Dataflow  
<https://cloud.google.com/solutions/stream-analytics/>  
<https://cloud.google.com/blog/products/data-analytics/streaming-analytics-now-simpler-more-cost-effective-cloud-dataflow>  
upvoted 4 times

👤 **hendrixlives** 2 years, 1 month ago

Selected Answer: D

D: "at least once and must be ordered within windows" means Pub/Sub (at least once) with Dataflow (windows).  
upvoted 3 times

👤 **JG123** 2 years, 2 months ago

Correct: D

upvoted 3 times

👤 **Chelseajcole** 2 years, 4 months ago

rule of thumb: If you see Kafka and Pub/Sub, always go with Pub/Sub in Google exam  
upvoted 8 times

👤 **hendrixlives** 2 years, 1 month ago

Careful doing that: I got a question where you had to choose between Kafka and Pub/Sub... and the solution required to be able to replay messages without time limit. So no Pub/Sub there.  
This being a Google cert does not mean that they always force Google solutions.  
upvoted 6 times

👤 **sandipk91** 2 years, 5 months ago

Answer is D

upvoted 2 times

👤 **awssp12345** 2 years, 6 months ago

[https://cloud.google.com/architecture/migrating-from-kafka-to-pubsub#comparing\\_features](https://cloud.google.com/architecture/migrating-from-kafka-to-pubsub#comparing_features)  
upvoted 2 times

👤 **sumanshu** 2 years, 7 months ago

Vote for D

Scaling - Dataflow.

Delivery of confirmed atleast 1 message - Pub/Sub

upvoted 3 times

👤 **Sush12** 2 years, 11 months ago

Answer is D

upvoted 2 times

👤 **Alasmindas** 3 years, 2 months ago

Indeed the correct answer is Option D.

Again, not sure why Exam topic answer is deliberately chosen for a wrong answer, for such simple question.

upvoted 5 times

👤 **szefco** 2 years, 2 months ago

To make us think of each question while studying, not just trying to memorize answers :) it looks that "correct" answers are chosen random  
upvoted 1 times

Question #118

Topic 1

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

- ⇒ Each department should have access only to their data.
- ⇒ Each department will have one or more leads who need to be able to create and update tables and provide them to their team.
- ⇒ Each department has data analysts who need to be able to query but not modify data.

How should you set access to the data in BigQuery?

- A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.

- B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.
- C. Create a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.
- D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

**Correct Answer: D**

*Community vote distribution*

B (70%)

D (30%)

 **juliosb** Highly Voted 10 months, 2 weeks ago

Old question. It's done using IAM nowadays: `bigrquery.dataEditor` and `bigrquery.dataViewer`  
upvoted 6 times

 **AWSandeep** Highly Voted 1 year, 4 months ago

**Selected Answer: B**

B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.

upvoted 6 times

 **Kalai\_1** Most Recent 1 month, 1 week ago

Answer : D. There is no role called WRITER or READER as preliminary role.  
upvoted 1 times

 **forepick** 8 months ago

**Selected Answer: B**

both C & D violate the principle of least privilege.  
A talks about OWNER and WRITER roles, and the analyst doesn't need a writer role.  
So we're left with B.

upvoted 1 times

 **Joane\_** 9 months, 2 weeks ago

**Selected Answer: D**

<https://cloud.google.com/bigquery/docs/access-control#bigquery>  
upvoted 1 times

 **midgoo** 10 months, 3 weeks ago

**Selected Answer: B**

B - Lead needs to have the role to create tables and also Analyst only need to read  
upvoted 1 times

 **musumusu** 11 months, 1 week ago

Answer B:

Why not D, mentioned in question: Data lead will create tables in dataset. Imagine, other department leads are creating unnecessary tables in shared dataset and you are struggling to find your tables as everyday there are some new tables. Headache right ? better to give them separate dataset and do whatever you want in that dataset.

upvoted 5 times

 **xj\_kevin** 11 months, 3 weeks ago

Vote B, both BD can fulfill the job requirement but B is on dataset level and D on project level. "By default, granting access to a project also grants access to datasets within it." D may issue unnecessary accesses to other content in the project.

upvoted 3 times

 **desertlotus1211** 1 year ago

Interestingly enough - I know believe the answer is A...  
Deleting is not the same as modify...  
upvoted 1 times

✉  **desertlotus1211** 1 year ago

Answer is B: <https://cloud.google.com/bigquery/docs/access-control>

The question ask for the lead to be able to:

CREATE, UPDATE, and SHARE with the team...

BigQuery Data Owner can do that

(roles/bigquery.dataOwner)

When applied to a table or view, this role provides permissions to:

Read and update data and metadata for the table or view.

Share the table or view.

Delete the table or view.

Editor cannot do that.

Thoughts?

upvoted 1 times

✉  **desertlotus1211** 1 year ago

I apologize - I thought B said Owner...

This questions makes no sense now...

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: D**

It's D, because this is an outdated question, before IAM you cannot set Editor to a dataset; but the best practice is: Create a dataset for each department. Assign the department leads the role of EDITOR(NOT OWNER), and assign the data analysts the role of READER on their database

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

Dude, I know there are updates to IAM, but the key point of the question is to have the leads have table creation and update roles... So the already need roles at the dataset level and hence C and D is out. We wouldn't be able to memorise all the roles, but clearly we cannot prov

access on a table level...

upvoted 7 times

✉  **Wonka87** 1 year, 1 month ago

and to supplement why does it need viewer role on the project the table is in?

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

Wow B is an answer

<https://cloud.google.com/bigquery/docs/access-control-basic-roles#dataset-basic-roles>

upvoted 3 times

✉  **MisuLava** 1 year, 3 months ago

**Selected Answer: D**

It CANNOT BE B BECAUSE OF :

Caution: BigQuery's dataset-level basic roles existed prior to the introduction of IAM. We recommend that you minimize the use of basic roles in production environments, don't grant basic roles unless there is no alternative. Instead, use predefined IAM roles.

<https://cloud.google.com/bigquery/docs/access-control-basic-roles>

upvoted 2 times

✉  **desertlotus1211** 1 year ago

Ummmm owner is a predefined role

<https://cloud.google.com/bigquery/docs/access-control>

BigQuery Data Owner

(roles/bigquery.dataOwner)

upvoted 1 times

✉  **josrojgra** 1 year, 3 months ago

**Selected Answer: B**

I vote B because C and D says that the role is on the project that the table is in, this mean that the role is at the project level that implies that: If you create a dataset in a project that contains any editors, BigQuery grants those users the `bigrquery.dataEditor` predefined role for the new dataset. (from <https://cloud.google.com/bigquery/docs/access-control-basic-roles#project-basic-roles>)

A can't not be because the analysts, in this case, can access the data.

B grant to the leads update their datasets, that's mean create tables, and the analysts only read their datasets.

upvoted 3 times

✉  **VipinSingla** 1 year, 3 months ago

**Selected Answer: D**

<https://cloud.google.com/bigquery/docs/access-control-basic-roles>

Caution: BigQuery's dataset-level basic roles existed prior to the introduction of IAM. We recommend that you minimize the use of basic roles in production environments, don't grant basic roles unless there is no alternative. Instead, use predefined IAM roles.

upvoted 1 times

✉  **nwk** 1 year, 4 months ago

Vote D, there is only Viewer, Editor and Owner roles for BQ

<https://cloud.google.com/bigquery/docs/access-control-basic-roles>

upvoted 5 times

✉  **Remi2021** 1 year, 4 months ago

Sorry but you are wrong. There is WRITER and READER role for dataset see them in this documentation. I was also confused at the beginning:

<https://cloud.google.com/bigquery/docs/access-control-basic-roles>

upvoted 6 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: B**

B is correct

upvoted 3 times

Question #119

Topic 1

You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

- A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
- B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
- C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.

D. Use Cloud Dataflow to write a summary of each day's stock trades to an Avro file on Cloud Storage. Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.

**Correct Answer: A**

*Community vote distribution*

A (100%)

✉️ [User] [Removed] Highly Voted 3 years, 10 months ago

Answer: A

Description: Timestamp at starting of rowkey causes bottleneck issues

upvoted 41 times

✉️ [User] **kichukonr** Highly Voted 3 years, 9 months ago

Stock symbol will be similar for most of the records, so it's better to start with random number.. Answer should be B

upvoted 13 times

✉️ [User] **Abhi16820** 2 years, 2 months ago

You never use something called random number in bigtable rowkey because it gives you no use in querying possibilities, since we can't run sql querys in bigtable we should not randomise rowkeys in bigtable.

Don't confuse the above point with the hotspot logic, both are different if you think so.

And another thing is, what you said can be good choice if we are using cloud spanner and trying to comeup with primary key situation, since there we can always run sql query.

I think you got the point now.

upvoted 11 times

✉️ [User] **karthik89** 2 years, 11 months ago

it can start with stock symbol concated with timestamp can be a good row key design

upvoted 6 times

✉️ [User] **Yonghai** 2 years, 1 month ago

for a given company, the data points starts with the same stock symbol. The dataset is not distributed. It is not a good option.

upvoted 3 times

✉️ [User] **taeypyung** 3 years, 9 months ago

I agree with u

upvoted 2 times

✉️ [User] **Sofia98** Most Recent 3 weeks, 1 day ago

**Selected Answer: A**

Answer is A.

upvoted 1 times

✉️ [User] **musumusu** 11 months, 1 week ago

Answer A:

Trick to remember: Row-key adjustment always be like in descending order.

#<<Least value>>#<<Lesser value>>

For example:

1. #<<Earth>>#<<continents>>#<<countries>>#<<cities>> and so on..

2. #<<Stock>>#<<users>>#timestamp..

in 99% cases timestamp will be in the end, as its smallest division...

upvoted 11 times

✉️ [User] **piyush777** 5 months, 2 weeks ago

Awesome!

upvoted 1 times

👤 **zelick** 1 year, 1 month ago

Selected Answer: A

A is the answer.

<https://cloud.google.com/bigtable/docs/schema-design#row-keys>

It's important to create a row key that makes it possible to retrieve a well-defined range of rows. Otherwise, your query requires a table scan, which is much slower than retrieving specific rows.

<https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid>

Some types of row keys can make it difficult to query your data, and some result in poor performance. This section describes some types of row keys that you should avoid using in Bigtable.

- Row keys that start with a timestamp. This pattern causes sequential writes to be pushed onto a single node, creating a hotspot. If you put a timestamp in a row key, precede it with a high-cardinality value like a user ID to avoid hotspots.

upvoted 6 times

👤 **AzureDP900** 1 year, 1 month ago

I agree with you . A is right

upvoted 1 times

👤 **MaxNRG** 2 years ago

Selected Answer: A

A: [https://cloud.google.com/bigtable/docs/schema-design-time-series#prefer\\_rows\\_to\\_column\\_versions](https://cloud.google.com/bigtable/docs/schema-design-time-series#prefer_rows_to_column_versions)

upvoted 4 times

👤 **JG123** 2 years, 2 months ago

Correct: A

upvoted 1 times

👤 **JayZeeLee** 2 years, 2 months ago

A and B would both work, since both would distribute the work. This question is not framed properly.

upvoted 1 times

👤 **sumanshu** 2 years, 7 months ago

Vote for A

upvoted 3 times

👤 **Jay3244** 2 years, 11 months ago

Option A.

Below document explains

Having EXCHANGE and SYMBOL in the leading positions in the row key will naturally distribute activity.

<https://cloud.google.com/bigtable/docs/schema-design-time-series>

upvoted 5 times

👤 **arghya13** 3 years, 2 months ago

I think A

upvoted 2 times

👤 **kavs** 3 years, 2 months ago

Catch here is current RowKey starts with timestamp which should not be in the starting or end position so symbol should be prefixed before timestamp

upvoted 1 times

👤 **Cloud\_Enthusiast** 3 years, 2 months ago

A is correct..A Good ROW KEY has to be an ID followed by timestamp. Stock symbol in this case works as an ID

upvoted 6 times

👤 **kino2020** 3 years, 4 months ago

A.

You can find an example in Google's introductory guide.

[https://cloud.google.com/bigtable/docs/schema-design-time-series?hl=ja#financial\\_market\\_data](https://cloud.google.com/bigtable/docs/schema-design-time-series?hl=ja#financial_market_data)

upvoted 2 times

👤 **Diktator** 3 years, 4 months ago

I think A would be best practice. Adding random numbers as start of rowkey doesn't help with troubleshooting

upvoted 3 times

✉  **Tanmoyk** 3 years, 4 months ago

B should be the answer as adding random numbers in the beginning of the rowkey will distributes data across multiple nodes  
upvoted 1 times

✉  **haroldbenites** 3 years, 5 months ago

B is correct.

A is incorrect. The documentation doesn't recommend constants in the row key because the balance is not efficient. There are 2 methods to avoid hotspotting. Promotion Field (use a UserId BEFORE the timestamp) and Salting (use timestamp-hash divided by 3 and put it before the timestamp)  
upvoted 1 times

✉  **atnafu2020** 3 years, 5 months ago

Agree with promotion field and salting. But, there is no constant. A stock symbol is a unique series of letters assigned to security for traditional purposes.  
upvoted 2 times

✉  **daghayeghi** 2 years, 11 months ago

A is correct:  
You deny your sentence, User ID means Stock Symbol, then B is correct.  
upvoted 1 times

✉  **daghayeghi** 2 years, 11 months ago

, then A is correct.  
upvoted 1 times

Question #120

Topic 1

You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on the behavior of the pipeline with Cloud

Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

- A. An alert based on a decrease of subscription/num\_undelivered\_messages for the source and a rate of change increase of instance/storage/ used\_bytes for the destination
- B. An alert based on an increase of subscription/num\_undelivered\_messages for the source and a rate of change decrease of instance/storage/ used\_bytes for the destination
- C. An alert based on a decrease of instance/storage/used\_bytes for the source and a rate of change increase of subscription/num\_undelivered\_messages for the destination
- D. An alert based on an increase of instance/storage/used\_bytes for the source and a rate of change decrease of subscription/num\_undelivered\_messages for the destination

**Correct Answer: B**

*Community vote distribution*

B (95%)

5%

✉  **dambilwa**  3 years, 7 months ago

You would want to get alerted only if Pipeline fails & not if it is running fine. I think Option [B] is correct, because in event of Pipeline failure :  
1) subscription/ num\_undelivered\_messages would pile up at a constant rate as the source has consistent throughput  
2) instance/storage/ used\_bytes will get closer to zero. Hence need to monitor its rate of change  
upvoted 25 times

✉  **Barniyah** 3 years, 7 months ago

Yes, you are right, it should be B:  
Thank you  
upvoted 4 times

✉  **marioferrulli** 2 years, 1 month ago

Why would the instance/storage/used\_bytes get closer to zero? If there's an error at a certain point, wouldn't we just see that the used\_by remain constant while the num\_undelivered\_messages increases? I don't get why the destination's used bytes should decrease.  
upvoted 1 times

✉  **szefco** 2 years, 1 month ago

"rate of change decrease of instance/storage/ used\_bytes" - if rate of instance/storage/ used\_bytes decreases that means less data is written - so something is wrong with the pipeline.  
It's not used bytes that decreases - it's rate of change decreases.  
Example: if everything works fine your pipeline writes 5MB/s to the sink. If it decreases to 0.1MB/s it means something is wrong  
upvoted 6 times

✉  **baubaumiaomiao** 2 years, 1 month ago

"If there's an error at a certain point, wouldn't we just see that the used\_bytes remain constant while the num\_undelivered\_messages increases?"  
It's the rate of change, not the absolute value  
upvoted 1 times

✉  **[Removed]**  3 years, 10 months ago

Correct - B

upvoted 21 times

✉  **midgoo**  10 months, 3 weeks ago

**Selected Answer: B**

For those who may get confuse at the start by the term 'subscription/num\_undelivered\_messages', it is not a division. It is the full path of the metric. So we should just read it as 'num\_undelivered\_messages'. The same for 'used\_bytes'.

So if we see the source have more backlog (more num\_undelivered\_messages), or the destination utilization going down, that is the indicator something going wrong

upvoted 13 times

✉  **kryzo** 7 months, 3 weeks ago

great explanation thanks !

upvoted 2 times

✉  **musumusu** 11 months, 1 week ago

Answer B:

Trick: In stackdriver always put Alert for Subscriber + CPU  
Subscriber - num of undelivered message INCREASE alert  
CPU - Instance or storage DECREASE alert.

Make sense right !

upvoted 3 times

✉  **atlan** 1 year ago

Nobody seems to pay attention to instance/storage/used\_bytes. I only find this metric for Spanner.  
[https://cloud.google.com/monitoring/api/metrics\\_gcp#gcp-spanner](https://cloud.google.com/monitoring/api/metrics_gcp#gcp-spanner)

While Dataflow processes and stores everything in Cloud Storage, Spanner could only be the source.  
<https://cloud.google.com/spanner/docs/change-streams>

Also, if it is either A or B, the instance/storage/used\_bytes metric does not make sense for the destination, which is Cloud Storage.

Can anyone help me understand?

upvoted 1 times

✉  **desertlotus1211** 1 year ago

look here: [https://cloud.google.com/monitoring/api/metrics\\_gcp](https://cloud.google.com/monitoring/api/metrics_gcp)

instance/storage/used\_bytes GA

Storage used.

upvoted 1 times

✉  **AzureDP900** 1 year ago

B. An alert based on an increase of subscription/num\_undelivered\_messages for the source and a rate of change decrease of instance/storage/used\_bytes for the destination

upvoted 1 times

 **AzureDP900** 1 year, 1 month ago

B is right

upvoted 1 times

 **Catweazle1983** 1 year, 1 month ago

**Selected Answer: A**

An alert based on a decrease of subscription/num\_undelivered\_messages for the source and a rate of change increase of instance/storage/used\_bytes for the destination

10 subscriptions / 1 undelivered messages = 10

10 subscriptions / 5 undelivered messages = 2

You clearly want to be alerted when the number of undelivered messages increases. The ratio then decreases. In my example from 10 to 2.

upvoted 1 times

 **squishy\_fishy** 10 months, 2 weeks ago

subscription/num\_undelivered\_messages is a path, not a division.

upvoted 1 times

 **zellick** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

[https://cloud.google.com/pubsub/docs/monitoring#monitoring\\_the\\_backlog](https://cloud.google.com/pubsub/docs/monitoring#monitoring_the_backlog)

Monitor message backlog

To ensure that your subscribers are keeping up with the flow of messages, create a dashboard. The dashboard can show the following backlog metrics, aggregated by resource, for all your subscriptions:

- Unacknowledged messages (subscription/num\_undelivered\_messages) to see the number of unacknowledged messages.

upvoted 3 times

 **A1000** 1 year, 5 months ago

**Selected Answer: B**

Increase subscription/num delivered message

decrease instance/storage/used bytes

upvoted 1 times

 **Pime13** 1 year, 6 months ago

**Selected Answer: B**

Correct - B

upvoted 1 times

 **JG123** 2 years, 2 months ago

Correct: B

upvoted 2 times

 **Abhi16820** 2 years, 2 months ago

isn't B and C are same.

upvoted 2 times

 **JayZeeLee** 2 years, 2 months ago

B.

It's useful to monitor the source that keeps sending data while the destination that doesn't take anything in.

upvoted 3 times

 **squishy\_fishy** 2 years, 3 months ago

The answer is B.

subscription/num\_undelivered\_messages: the number of messages that subscribers haven't processed

[https://cloud.google.com/pubsub/docs/monitoring#monitoring\\_forwarded\\_undeliverable\\_messages](https://cloud.google.com/pubsub/docs/monitoring#monitoring_forwarded_undeliverable_messages)

upvoted 2 times

✉  **squishy\_fishy** 2 years, 4 months ago

Silly question: what is subscription/ num\_undelivered\_messages, it is divided by? or per subscription per num\_undelivered\_messages?  
upvoted 2 times

✉  **910** 1 year, 10 months ago

yes is misleading:  
the metric "subscription/num\_undelivered\_messages" is just the path of the API URL

[actions.googleapis.com/...subscription/num\\_undelivered\\_messages](https://actions.googleapis.com/...subscription/num_undelivered_messages)

ref: [https://cloud.google.com/monitoring/api/metrics\\_gcp#pubsub/subscription/num\\_undelivered\\_messages](https://cloud.google.com/monitoring/api/metrics_gcp#pubsub/subscription/num_undelivered_messages)

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Looks B

upvoted 3 times

Question #121

Topic 1

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally.

Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your

Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

- A. Edge TPUs as sensor devices for storing and transmitting the messages.
- B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
- C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
- D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Should be C

upvoted 21 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: C

Description: Pubsub is global and dataflow can scale workers

upvoted 18 times

✉  **ga8our** Most Recent 8 months ago

Can anyone pls explain what's wrong with D, the load balancing solution?

upvoted 1 times

✉  **musumusu** 11 months, 1 week ago

Answer C:

What is wrong with D, nothing, Cloud load balancing can shift traffic for high volume and low internet in one region. It cost avg. 0.01-0.25 \$ per GB, or if volume is too high. 0.05 \$ per Hour http request. This might be the answer if your exam for network engineer.

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Answer C, but it will not solve bad internet connection, make sure 100mbps speed of internet is at sensor side.

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

<https://cloud.google.com/architecture/iot-overview#cloud-pubsub>

Pub/Sub can act like a shock absorber and rate leveller for both incoming data streams and application architecture changes. Many devices have limited ability to store and retry sending telemetry data. Pub/Sub scales to handle data spikes that can occur when swarms of devices respond to events in the physical world, and buffers these spikes to help isolate them from applications monitoring the data.

upvoted 8 times

✉  **AzureDP900** 1 year, 1 month ago

Agree with your explanation

upvoted 1 times

✉  **MisuLava** 1 year, 3 months ago

"single on-premises Kafka cluster in a data center in the us-east region"

is it on-prem or in a datacenter in us-east ?

upvoted 1 times

✉  **JamesKarianis** 1 year, 6 months ago

**Selected Answer: C**

Answer is C

upvoted 1 times

✉  **Prasanna\_kumar** 1 year, 11 months ago

Answer is option C

upvoted 1 times

✉  **ivanhsiaiv** 2 years, 6 months ago

Answer c

kafka cluster in on-premise for streaming msgs

pub/sub for streaming msgs in cloud

upvoted 4 times

✉  **sumanshu** 2 years, 7 months ago

Vote for C

upvoted 4 times

✉  **Allan222** 2 years, 11 months ago

Should be C

upvoted 4 times

✉  **daghayeghi** 2 years, 11 months ago

C is correct:

the main trick come from A, and response is that TPU only use when we have a deployed machine learning model that we don't have now.

upvoted 5 times

✉  **ArunSingh1028** 2 years, 11 months ago

Answer - C

upvoted 4 times

✉  **Alasmindas** 3 years, 2 months ago

Easy Question : ANswer is Option C.

Alternative to Kafka in google cloud native service is Pub/Sub and Dataflow punched with Pub/Sub is the google recommended option

upvoted 5 times

✉  **atnafu2020** 3 years, 5 months ago

C

the issue is with a single Kafka cluster is the need to scale automatically with Dataflow

upvoted 4 times

✉  **haroldbenites** 3 years, 5 months ago

C is correct

upvoted 4 times

Question #122

Topic 1

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this?

(Choose two.)

- A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
- B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
- C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
- D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
- E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

**Correct Answer: CE**

*Community vote distribution*

AB (87%)

13%

✉  **Ganshank**  3 years, 9 months ago

A,B

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

upvoted 38 times

✉  **salsabilf** 2 years, 9 months ago

"while keeping the costs"

should be A,D

upvoted 6 times

✉  **MrCastro** 2 years, 5 months ago

Big query streaming inserts ARE NOT cheap

upvoted 9 times

✉  **hellofrnds** 2 years, 4 months ago

If you use B , not D , how can we do "point in time" recovery? is it possible?

Point in time recovery needs export along with timestamp, so that we can recover for a particular timestamp.

upvoted 4 times

atnafu2020 Highly Voted 3 years, 5 months ago

AC

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

C: To import only a subset of entities or to import data into BigQuery, you must specify an entity filter in your export.

B: Not correct since you want to store in a different environment than Datastore. This statement is true: Data exported from one Datastore mode database can be imported into another Datastore mode database, even one in another project.

A is correct

Billing and pricing for managed exports and imports in Datastore

Output files stored in Cloud Storage count towards your Cloud Storage data storage costs.

Steps to Export all the entities

1. Go to the Datastore Entities Export page in the Google Cloud Console.

2. Go to the Datastore Export page

2. Set the Namespace field to All Namespaces, and set the Kind field to All Kinds.

3. Below Destination, enter the name of your "Cloud Storage bucket".

4. Click Export.

upvoted 23 times

aparna4387 2 years, 2 months ago

<https://cloud.google.com/datastore/docs/export-import-entities#import-into-bigquery>

Data exported without specifying an entity filter cannot be loaded into BigQuery. This is not mentioned explicitly. Safe to assume there is no filter on the exports. So options are AB

upvoted 6 times

AzureDP900 1 year, 1 month ago

A, B is perfect

upvoted 1 times

Yiouk 2 years, 5 months ago

C is valid because of table snapshots. Else standard time travel is valid only for 7 days

[https://cloud.google.com/bigquery/docs/table-snapshots-intro#table\\_snapshots](https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots)

<https://cloud.google.com/bigquery/docs/time-travel#limitation>

upvoted 1 times

Chelseajcole 2 years, 4 months ago

you wanna say invalid?

upvoted 1 times

tavva\_prudhvi 1 year, 10 months ago

As you've mentioned in B, does the environment meant to be a project or a resource? As, we can clone a copy of the data in a datastore even in another project!? Then, it's B.

Also, in point C they didn't mention any entity filter hence we eliminate C how can you support your own statement with a different answer

upvoted 2 times

kskssk Most Recent 4 months, 4 weeks ago

AB chatgpt

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class:

Managed export is a feature provided by Cloud Datastore to export your data.

Storing the data in a Cloud Storage bucket, especially using Nearline or Coldline storage classes, helps keep storage costs low while allowing you to retain the snapshots for a long time.

B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export:

This method allows you to create snapshots by exporting data from Cloud Datastore (using managed export) and then importing it into a separate project under a unique namespace.

By importing into a separate project, you can keep a copy of the data in a different environment, which is useful for point-in-time recovery or creating clones of the data.

upvoted 4 times

zellick 1 year, 1 month ago

Selected Answer: AB

AB is the answer.

upvoted 3 times

👤 **NicolasN** 1 year, 1 month ago

**Selected Answer: AB**

A rather complicated question, of a kind I wish I won't face in the exam. My opinion:

- [A] A valid and cost-effective solution satisfying the requirement for PIT recovery
- [B] A valid solution but far from ideal for archiving. It satisfies the requirement part "you can ... clone a copy of the data for Cloud Datastore to a different environment" (an objection to the word "namespace", I think it should be just "name")

upvoted 10 times

👤 **NicolasN** 1 year, 1 month ago

[C] There is the limitation "Data exported without specifying an entity filter cannot be loaded into BigQuery". The entity filter for this case should contain all the kinds of entities but there is another limitation of "100 entity filter combinations". We have no knowledge of the kinds of namespaces of the entities.

Sources:

<https://cloud.google.com/datastore/docs/export-import-entities#import-into-bigquery>

[https://cloud.google.com/datastore/docs/export-import-entities#exporting\\_specific\\_kinds\\_or\\_namespaces](https://cloud.google.com/datastore/docs/export-import-entities#exporting_specific_kinds_or_namespaces)

[D] seems a detailed candidate solution but it violates the limitation "You cannot append Datastore export data to an existing table."

[https://cloud.google.com/bigquery/docs/loading-data-cloud-datastore#appending\\_to\\_or\\_overwriting\\_a\\_table\\_with\\_cloud\\_datastore\\_data](https://cloud.google.com/bigquery/docs/loading-data-cloud-datastore#appending_to_or_overwriting_a_table_with_cloud_datastore_data)

[E] Cloud Source Repositories are for source code and not a suitable storage for this case.

upvoted 10 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: AB**

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

upvoted 1 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: AB**

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

upvoted 1 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: AB**

The answer is nothing to do with bigquery , so you can skip what mention to bigquery.

A B is the final answer

upvoted 1 times

👤 **DataEngineer\_WideOps** 1 year, 6 months ago

A,B

For those who say using BQ as archival, How can we achieve that while datastore are NO-SQL whereas BQ are SQL , will that work? also BQ not created for achieving purposes.

upvoted 1 times

👤 **AmirN** 1 year, 7 months ago

Option B is 36 times more expensive than C

upvoted 1 times

👤 **Nico1310** 2 years ago

**Selected Answer: AB**

AB. for sure streaming to BQ its quite expensive!

upvoted 2 times

MaxNRG 2 years ago

Selected Answer: AD

A - Cloud Storage (long-term data + costs low)  
D - BigQuery (timestamp for point-in-time (PIT) recovery)  
upvoted 3 times

tavva\_prudhvi 1 year, 9 months ago

D is wrong, BQ Streaming inserts costs are high!  
upvoted 2 times

MaxNRG 1 month, 1 week ago

Agreed, AB  
<https://cloud.google.com/datastore/docs/export-import-entities>  
upvoted 1 times

medeis\_jar 2 years ago

Selected Answer: AB

Option A; Cheap storage and it is a supported method <https://cloud.google.com/datastore/docs/export-import-entities>  
Option B; Rationale - "Data exported from one Datastore mode database can be imported into another Datastore mode database, even one in another project." <<https://cloud.google.com/datastore/docs/export-import-entities>>  
upvoted 2 times

squishy\_fishy 2 years, 3 months ago

Answer is A, B.  
[https://cloud.google.com/datastore/docs/export-import-entities#exporting\\_specific\\_kinds\\_or\\_namespaces](https://cloud.google.com/datastore/docs/export-import-entities#exporting_specific_kinds_or_namespaces)  
upvoted 1 times

sergio6 2 years, 3 months ago

A, D  
A: Option for storage system that will account for the long-term data growth  
D: Option for snapshots, PIT recovery, copy of the data for Cloud Datastore in a different environment and, above all, archive snapshots for a long time  
B: not a good solution for archiving snapshots for a long time  
C: to import data into BigQuery, you must specify an entity filter  
E: Cloud Source Repositories is for code  
One note: E --> would be my second choice if there was Cloud Storage instead of Source Repositories (typo?)  
upvoted 4 times

Chelseajcole 2 years, 4 months ago

Vote A B . What's the purpose load into bigquery?  
upvoted 1 times

Chelseajcole 2 years, 4 months ago

<https://cloud.google.com/datastore/docs/export-import-entities#import-into-bigquery>  
Importing into BigQuery  
To import data from a managed export into BigQuery, see Loading Datastore export service data.

Data exported without specifying an entity filter cannot be loaded into BigQuery. If you want to import data into BigQuery, your export req must include one or more kind names in the entity filter.

You have to specify an entity filter before you can load from datastore to BQ. It didn't mention that at all. So C is incorrect  
upvoted 3 times

fire558787 2 years, 5 months ago

A for sure. Then I was undecided between B and C; B has high costs and C has low costs (storage is more expensive in Datastore). However question says that you want data to be used for Datastore. There is no native way to export data from BigQuery to Datastore, hence the only options that allow data to be restored to Datastore are A and B.

upvoted 7 times

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis.

Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

- A. Denormalize the data as much as possible.
- B. Preserve the structure of the data as much as possible.
- C. Use BigQuery UPDATE to further reduce the size of the dataset.
- D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
- E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

**Correct Answer: AD***Community vote distribution*

AD (79%)

BD (16%)

5%

 **rickywck** Highly Voted 3 years, 10 months ago

I think AD is the answer. E will not improve performance.  
upvoted 39 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A, D  
Description: Denormalization will help in performance by reducing query time, update are not good with bigquery  
upvoted 20 times

 **awssp12345** 2 years, 6 months ago

My guess is append has better performance than update.  
upvoted 3 times

 **midgoo** Most Recent 10 months, 1 week ago

Selected Answer: BD

If we denormalize the data, the Data Science team will shout at us. Preserving it is the way to go  
upvoted 3 times

 **WillemHendr** 7 months, 4 weeks ago

Shouting data-science teams are not part of question, this is more about what is exam correct, not what it the best for your own situation  
upvoted 2 times

 **vaga1** 7 months, 3 weeks ago

Denormalization is just a best practice when using BQ.  
upvoted 1 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: AD**

A and D:

A- Improve performance

D- Is better for DS have all the history and not the last update...

upvoted 4 times

✉  **zelick** 1 year, 1 month ago

**Selected Answer: AD**

AD is the answer.

<https://cloud.google.com/bigquery/docs/best-practices-performance-nested>

Best practice: Use nested and repeated fields to denormalize data storage and increase query performance.

Denormalization is a common strategy for increasing read performance for relational datasets that were previously normalized. The recommended way to denormalize data in BigQuery is to use nested and repeated fields. It's best to use this strategy when the relationships are hierarchical and frequently queried together, such as in parent-child relationships.

upvoted 4 times

✉  **AzureDP900** 1 year, 1 month ago

A, C is correct I agree

upvoted 1 times

✉  **NicolasN** 1 year, 2 months ago

The criteria for selecting a strategy are the performance and usability for the data science team. This team performs the analysis by querying stored data. So we don't care for performance related with data ingestion. According to this point of view:

A: YES - undisputedly favours query performance

B: YES - Keeping the structure unchanged promotes usability (the team won't need to update queries or ML models)

C: Questionable - Updating the status of a row instead of appending newer versions is keeping the size smaller. But does this affect significant the analysis performance? Even if it does, creating materialized views to keep the most recent status per row eliminates it

D: NO - has nothing to do with DS team's tasks, affects ingestion performance

E: NO - demotes usability

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

For B there is no mention that the current data structure is being used (...data science team WILL build machine learning models based on this data.) ... We're developing a new data model to be used by them in the future

upvoted 1 times

✉  **NicolasN** 1 year, 2 months ago

(mistakenly voted AC instead of AB)

upvoted 1 times

✉  **DerickTW** 1 year, 4 months ago

**Selected Answer: AC**

The DML quota limit is removed since 2020, I think C is better than D now.

upvoted 1 times

✉  **deavid** 1 year, 3 months ago

Is not about the quota. You should avoid using UPDATE because it makes a big scan of the table, and is not efficient or high performance. Usually prefer appends and merges instead, and using the optimized schema approach of Big Query that denormalizes the table to avoid joins and leverages nested and repeated fields.

upvoted 4 times

✉  **MaxNRG** 2 years ago

**Selected Answer: AD**

A: Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINs on large tables, but with a denormalized data structure, you don't have to use JOINs, since all of the data has been combined into one table.

Denormalization also makes queries simpler because you do not have to use JOIN clauses.

[https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing\\_data](https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data)

D: BigQuery append

upvoted 4 times

👤 **medeis\_jar** 2 years ago

Selected Answer: AD

requirements are -> performance and usability.

Denormalization will help in performance by reducing query time, update is not good with big query.

And append has better performance than Update.

upvoted 3 times

👤 **doninakula** 2 years, 2 months ago

I think AD. E is not valid because it use external table which is not good for performance

upvoted 1 times

👤 **sumanshu** 2 years, 7 months ago

A - correct (denormalization will help)

B - data already heavily structured (no use and no impact)

C - more than 1500 Updates not possible

D - Not sure..(because appending will increase size and cost)

E - Does not look good (increase cost..also we are storing for all days....again for query we need to issue mutiple query for all days....)

So, A & D (left out of 5)

upvoted 5 times

👤 **Jeysolomon** 2 years, 7 months ago

Correct Answer: AE

A – Denormalisation helps improve performance.

B, C - Not helping to address the problem.

D – Append will increase the db size and cost involved for storage and also for large number of records to scan for queries by data science te which is costlier.

E - Addresses the problem of maximising the usability of the data science team and the data. They can anayse the data exported to cloud storage instead of reading from bigquery which is expensive and impact performance considerably.

upvoted 3 times

👤 **Chelseajcole** 2 years, 3 months ago

It didn't mention cost is a concern

upvoted 1 times

👤 **retep007** 2 years, 4 months ago

E is wrong, you've been asked to use bigquery and reading files from storage in bq is significantly more time consuming

upvoted 2 times

👤 **daghayeghi** 2 years, 10 months ago

A, D:

Using BigQuery as an OLTP store is considered an anti-pattern. Because OLTP stores have a high volume of updates and deletes, they are a mismatch for the data warehouse use case. To decide which storage option best fits your use case, review the Cloud storage products table. BigQuery is built for scale and can scale out as the size of the warehouse grows, so there is no need to delete older data. By keeping the enti history, you can deliver more insight on your business. If the storage cost is a concern, you can take advantage of BigQuery's long term stora pricing by archiving older data and using it for special analysis when the need arises. If you still have good reasons for dropping older data, yc can use BigQuery's native support for date-partitioned tables and partition expiration. In other words, BigQuery can automatically delete olde data.

[https://cloud.google.com/solutions/bigquery-data-warehouse#handling\\_change](https://cloud.google.com/solutions/bigquery-data-warehouse#handling_change)

upvoted 4 times

✉  **Hithesh** 2 years, 10 months ago

should be AC.. "Every hour, thousands of transactions are updated with a new status" if we append how we will handle the new status change  
upvoted 2 times

✉  **sumanshu** 2 years, 7 months ago

C not possible, maximum 1500 updates possible in a day  
upvoted 1 times

✉  **raf2121** 2 years, 6 months ago

DML without limits now in BQ (below blog says March 2020, Not sure whether these questions were prepared before or after March 2020)

<https://cloud.google.com/blog/products/data-analytics/dml-without-limits-now-in-bigquery>

upvoted 1 times

✉  **hdmi\_switch** 2 years, 6 months ago

There is no more hard limit, but UPDATES are queued:

"BigQuery runs up to 2 of them concurrently, after which up to 20 are queued as PENDING. When a previously running job finishes, the next pending job is dequeued and run. Currently, queued mutating DML statements share a per-table queue with maximum length 2. Additional statements past the maximum queue length for each table fail."

With thousands of updates per hour, this doesn't seem feasible. I would assume the question is marked as outdated anyway or the answers are update in the actual exam.

upvoted 4 times

✉  **daghayeghi** 2 years, 11 months ago

AC:

the problem is exactly about Updating and preserving size of database as much as possible, then denormalization and using UPDATE function from DML will address the issue. they don't want to update faster. then A & C is correct.

<https://cloud.google.com/solutions/bigquery-data-warehouse>

upvoted 1 times

✉  **karthik89** 2 years, 11 months ago

you can update bigquery 1500 times in a day  
upvoted 3 times

✉  **daghayeghi** 2 years, 10 months ago

A, D:

it was my mistake, we should decrease update as Bigquery is not design for update.

[https://cloud.google.com/solutions/bigquery-data-warehouse#handling\\_change](https://cloud.google.com/solutions/bigquery-data-warehouse#handling_change)

upvoted 3 times

✉  **Nams\_139** 3 years, 2 months ago

A,D Since the requirements are both performance and usability.

upvoted 5 times

✉  **federicohi** 3 years, 2 months ago

I think may be it's AC because appending is worst to increase dataset size. The question seems to put like a problem the size of dataset and performance to data science so inserting more rows decrease performance for them.

upvoted 3 times

You are designing a cloud-native historical data processing system to meet the following conditions:

- ⇒ The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Dataproc, BigQuery, and Compute Engine.
- ⇒ A batch pipeline moves daily data.
- ⇒ Performance is not a factor in the solution.
- ⇒ The solution design should maximize availability.

How should you design data storage for this solution?

- A. Create a Dataproc cluster with high availability. Store the data in HDFS, and perform analysis as needed.
- B. Store the data in BigQuery. Access the data using the BigQuery Connector on Dataproc and Compute Engine.
- C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Dataproc, BigQuery, and Compute Engine.
- D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Dataproc, BigQuery, and Compute Engine.

**Correct Answer: D**

*Community vote distribution*

D (100%)

✉  **jkhong** Highly Voted 1 year, 3 months ago

**Selected Answer: D**

Problem: How to store data?

Considerations: High availability, performance not an issue

A → avoid HDFS

C → multi-regional > regional in terms of availability

B could be the answer but we're dealing with PDF documents, we need blob storage (cloud storage). If we only have csv or Avro, this may be answer

upvoted 7 times

✉  **AzureDP900** Most Recent 1 year, 1 month ago

D is right

upvoted 1 times

✉  **dconesoko** 1 year, 1 month ago

**Selected Answer: D**

vote for D

upvoted 2 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

upvoted 2 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: D**

D of course

upvoted 2 times

✉  **kenanars** 1 year, 4 months ago

**Selected Answer: D**

D is the correct answer

upvoted 1 times

Question #125

Topic 1

You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

- A. Store and process the entire dataset in BigQuery.
- B. Store and process the entire dataset in Bigtable.
- C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
- D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.

**Correct Answer: C**

*Community vote distribution*

✉  **Rajokkiyam** Highly Voted 3 years, 10 months ago

Answer C.

upvoted 34 times

✉  **AKumar** Highly Voted 3 years, 7 months ago

A and B can be eliminated right away as they do not talk about providing for other cloud providers. between C and D. The question says nothing about warm or cold data-rather that data should be made available for other providers--C--can fulfill this condition. Answer C.

upvoted 23 times

✉  **AzureDP900** 1 year, 1 month ago

Agree with C

upvoted 1 times

✉  **zbyszek1** Most Recent 4 months, 2 weeks ago

For me A. I can use export from BQ to Cloud Storage. There is no need to store two copies of data.

upvoted 1 times

✉  **spicebits** 2 months, 3 weeks ago

If you export data from BQ to GCS then you will have two copies and you will be in the same architecture as answer C.

upvoted 3 times

✉  **vamgcp** 6 months, 1 week ago

Selected Answer: B

It can be C or D , but I will go with C as storing the full dataset in BigQuery and a compressed copy of the data in Cloud Storage is a good way to balance performance and cost.

upvoted 1 times

✉  **forepick** 8 months ago

Selected Answer: C

Best answer is C, although BQ can query gzipped files stored on GCS directly.

Maybe this double storage makes it a bit more highly available.

upvoted 1 times

✉  **izekc** 8 months, 4 weeks ago

Selected Answer: D

D is much more accurate.

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

Selected Answer: C

D → does not guarantee 100% queryable or accessible/available

upvoted 1 times

✉  **zellick** 1 year, 1 month ago

Selected Answer: C

C is the answer.

upvoted 1 times

✉  **Smaks** 1 year, 6 months ago

You can read streaming data from Pub/Sub, and you can write streaming data to Pub/Sub or BigQuery.

Thus Cloud Storage is not a proper sink for streaming pipeline.

I vote for B, since it is possible to convert unstructured data and store in BQ

upvoted 1 times

✉  **Smaks** 1 year, 6 months ago

ignore this comment, please

upvoted 10 times

✉  **Aslkdup** 1 year, 11 months ago

BQ can reach files at google storage as external table. so my answer is D. (If data was smaller than this, I would choose C)

upvoted 1 times

✉  **Bhawantha** 2 years ago

**Selected Answer: C**

both requirements are full filled.

upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: D**

D: BigQuery + Cloud Storage

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

D → does not guarantee 100% queryable or accessible/available

upvoted 1 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: C**

"You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis t in other cloud providers?"

Analytics -> BQ

Exposing -> GCS

upvoted 6 times

✉  **JG123** 2 years, 2 months ago

Correct: C

upvoted 2 times

✉  **xiaofeng\_0226** 2 years, 5 months ago

vote for C

upvoted 3 times

✉  **sumanshu** 2 years, 7 months ago

Vote for 'C'

A - Only Half requirement fulfil, expose as a file not getting fulfilled

B - Not a warehouse

C. Both requirements fulfilled...Bigquery and GCS

D. Both requirement fulfilled...but what if other cloud provider wants to analysis on rest 80% of the data. -

So out of 4 options, C looks okay

upvoted 8 times

✉  **gcper** 2 years, 11 months ago

C

BigQuery for analytics processing and Cloud Storage for exposing the data as files

upvoted 3 times

Concept) within a few working days. What should you do?

- A. Use Cloud Vision AutoML with the existing dataset.
- B. Use Cloud Vision AutoML, but reduce your dataset twice.
- C. Use Cloud Vision API by providing custom labels as recognition hints.
- D. Train your own image recognition model leveraging transfer learning techniques.

**Correct Answer: A**

*Community vote distribution*

A (54%)	B (33%)	13%
---------	---------	-----

 **Callumr** Highly Voted 3 years, 7 months ago

B - You only need a PoC and it has be done quickly  
upvoted 54 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Correct - A  
upvoted 20 times

 **saado9** Most Recent 4 months, 3 weeks ago

**Selected Answer: B**

Option B is the fastest way to train a model that can be used to recognize the 750 different components.  
upvoted 1 times

 **musumusu** 11 months, 2 weeks ago

Whats wrong with C, its fast, cheap and add your 750 labels which is not big work.  
AutoML is good to train on big dataset and costly as compared to APIs  
upvoted 2 times

 **knith66** 6 months, 1 week ago

it is a labeled dataset and why do you need to label it once again? So no C  
upvoted 1 times

 **forepick** 8 months ago

Adding custom labels to Vision API is done by training an AutoML model! That's the formal recommendation. And you don't need a big dataset for AutoML as it uses transfer learning.  
upvoted 4 times

 **techtitan** 11 months, 3 weeks ago

A - <https://cloud.google.com/vertex-ai/docs/beginner/beginners-guide> Target at least 1000 examples per target  
upvoted 7 times

 **techtitan** 11 months, 3 weeks ago

The quick POC part can be achieved by using Auto ML instead of creating and training your own model  
upvoted 1 times

 **odacir** 1 year, 1 month ago

**Selected Answer: A**

First I think in Vision API, but that is a pre-trained AI, will not recognize my labels, so because you have 1000 samples per item, AUTO ML is perfect. B cannot be because have not sensed to reduce your dataset if you have the recommended number of info.  
[https://cloud.google.com/vision/automl/docs/beginners-guide#include\\_enough\\_labeled\\_examples\\_in\\_each\\_category](https://cloud.google.com/vision/automl/docs/beginners-guide#include_enough_labeled_examples_in_each_category)

The bare minimum required by AutoML Vision training is 100 image examples per category/label. The likelihood of successfully recognizing a label goes up with the number of high quality examples for each; in general, the more labeled data you can bring to the training process, the better your model will be. Target at least 1000 examples per label.

upvoted 8 times

 **AzureDP900** 1 year, 1 month ago

A is correct  
upvoted 2 times

👤 **zellick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

[https://cloud.google.com/vision/automl/docs/beginners-guide#include\\_enough\\_labeled\\_examples\\_in\\_each\\_category](https://cloud.google.com/vision/automl/docs/beginners-guide#include_enough_labeled_examples_in_each_category)

The bare minimum required by AutoML Vision training is 100 image examples per category/label. The likelihood of successfully recognizing a label goes up with the number of high quality examples for each; in general, the more labeled data you can bring to the training process, the better your model will be. Target at least 1000 examples per label.

upvoted 4 times

👤 **ga8our** 8 months ago

So how are you going to test that the model was able to adequately learn from the sample? The point of splitting a dataset is to train the model on one part of the data (say 80%), and then test it on the other part (20%). If your model is able to predict the outcome of (most of) sample points in your test dataset, you can be confident that it will work well on future data. Without a test data set, however, you have no such feedback. Therefore, the answer is B.

upvoted 2 times

👤 **NewDE2023** 5 months, 4 weeks ago

I believe that the ideal would be to reduce the number of components for the POC and preserve the number of examples, so my answer is A.

upvoted 1 times

👤 **odacir** 1 year, 1 month ago

Agreed!

upvoted 1 times

👤 **gudiking** 1 year, 2 months ago

A - [https://cloud.google.com/vision/automl/docs/beginners-guide#include\\_enough\\_labeled\\_examples\\_in\\_each\\_category](https://cloud.google.com/vision/automl/docs/beginners-guide#include_enough_labeled_examples_in_each_category)

upvoted 1 times

👤 **MarielaYBird** 1 year, 2 months ago

**Selected Answer: B**

Based on this:

"As a rule of thumb, we recommend to have at least 100 training samples per class if you have distinctive and few classes, and more than 200 training samples if the classes are more nuanced and you have more than 50 different classes"

750 different components = more than 50 different classes. That means we need more than 200 training samples. If we used 250 training samples out of the 1000 samples and multiply it to 750 different classes we get a total of 187,500 which is the equivalent of reducing the data twice.

[https://cloud.google.com/vision/automl/object-detection/docs/prepare#how\\_big\\_does\\_the\\_dataset\\_need\\_to\\_be](https://cloud.google.com/vision/automl/object-detection/docs/prepare#how_big_does_the_dataset_need_to_be)

upvoted 4 times

👤 **josrojgra** 1 year, 3 months ago

**Selected Answer: A**

I choose A because on the vertex AI documentation (<https://cloud.google.com/vertex-ai/docs/image-data/classification/prepare-data>), on the best practices of preparing data for image recognition recommend this: We recommend about 1000 training images per label. The minimum per label is 10. In general, it takes more examples per label to train models with multiple labels per image, and resulting scores are harder to interpret.

I know that is PoC, but if you do it without enough accuracy, you maybe discard the solution because it isn't fit for your requirements. So is better to do it with enough data to be sure that the model is or not accuracy enough with this data, because you maybe haven't enough accuracy as the problem is the quality of the data and not the amount of it.

upvoted 3 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

[https://cloud.google.com/vision/automl/docs/beginners-guide#include\\_enough\\_labeled\\_examples\\_in\\_each\\_category](https://cloud.google.com/vision/automl/docs/beginners-guide#include_enough_labeled_examples_in_each_category)

The bare minimum required by AutoML Vision training is 100 image examples per category/label. The likelihood of successfully recognizing a label goes up with the number of high quality examples for each; in general, the more labeled data you can bring to the training process, the better your model will be. Target at least 1000 examples per label.

upvoted 3 times

👤 **John\_Pongthorn** 1 year, 4 months ago

The more labels, the more accurate the result.

upvoted 1 times

✉  **changs** 1 year, 4 months ago

**Selected Answer: B**

750\*1000 are a lot.

upvoted 1 times

✉  **ducc** 1 year, 5 months ago

**Selected Answer: A**

It is labeled, so A is correct

upvoted 1 times

✉  **civilizador** 1 year, 5 months ago

It's A.

[https://cloud.google.com/vision/automl/docs/beginners-guide#data\\_preparation](https://cloud.google.com/vision/automl/docs/beginners-guide#data_preparation)

The bare minimum required by AutoML Vision training is 100 image examples per category/label. The likelihood of successfully recognizing a label goes up with the number of high quality examples for each; in general, the more labeled data you can bring to the training process, the better your model will be. Target at least 1000 examples per label.

upvoted 5 times

✉  **civilizador** 1 year, 5 months ago

So even for POC better to use 1000 . There would be no significant time differences anyway between 500 and 1000

upvoted 1 times

✉  **TheRealBsh** 1 year, 6 months ago

Option A & B are quite close. Refer: [https://cloud.google.com/vision/automl/docs/beginners-guide#data\\_preparation](https://cloud.google.com/vision/automl/docs/beginners-guide#data_preparation) – Says to target at least 1000 images per label for training.

upvoted 3 times

✉  **czokwe** 1 year, 8 months ago

**Selected Answer: B**

B

can't choose A because model needs to pass through the dataset several times for a proof of concept, existing data set samples might not be seen in several working days causing over generalization

upvoted 1 times

✉  **Kriegs** 1 year, 8 months ago

I don't get it, why B rather than A? I know it's a proof of concept, but it's not like a bigger dataset is any kind of a dealbreaker in training a model of that size, and more data provides more accuracy to the model.

I would choose A or C.

upvoted 2 times

Question #127

Topic 1

You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

- A. Use Cloud TPUs without any additional adjustment to your code.
- B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
- C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
- D. Stay on CPUs, and increase the size of the cluster you're training your model on.

**Correct Answer: B**

*Community vote distribution*

D (47%)

C (40%)

13%

✉  **dhs227**  3 years, 10 months ago

The correct answer is C

TPU does not support custom C++ tensorflow ops

[https://cloud.google.com/tpu/docs/tpus#when\\_to\\_use\\_tpus](https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus)

upvoted 65 times

✉  **ffggrre** 3 months, 1 week ago

the link doesn't say TPU does not support custom C++ tensorflow ops

upvoted 1 times

✉  **Helinia** 1 month ago

It does. TPU is good for "Models with no custom TensorFlow/PyTorch/JAX operations inside the main training loop".

upvoted 1 times

✉  **aiguy**  3 years, 10 months ago

D:

Cloud TPUs are not suited to the following workloads: [...] Neural network workloads that contain custom TensorFlow operations written in C+. Specifically, custom operations in the body of the main training loop are not suitable for TPUs.

upvoted 43 times

✉  **gopinath\_k** 2 years, 10 months ago

B:

1. You need to provide support for the matrix multiplication - TPU
2. You need to provide support for the Custom TF written in C++ - GPU

upvoted 9 times

✉  **tavva\_prudhvi** 1 year, 10 months ago

But, in the question it also says we have to decrease the time significantly?? If you gonna use the CPU, it will take more time to train, right'

upvoted 1 times

✉  **cetanx** 8 months ago

Chat GPT says C

Option D is not the most cost-effective or efficient solution. While increasing the size of the cluster could decrease the training time, it would also significantly increase the cost, and CPUs are not as efficient for this type of workload as GPUs.

upvoted 1 times

✉  **FP77** 5 months, 2 weeks ago

chatgpt will give you different answers if you ask 10 times. The correct answer is B

upvoted 3 times

✉  **squishy\_fishy** 3 months ago

Totally agree. ChatGPT is garbage. It is still learning.

upvoted 2 times

✉  **Matt\_108**  2 weeks, 2 days ago

**Selected Answer: C**

to me, it's C

upvoted 1 times

✉  **Kimich** 1 month, 4 weeks ago

Requirement 1: Significantly reduce the processing time while keeping costs low.

Requirement 2: Bulky matrix multiplication takes up to several days.

First, eliminate A & D:

A: Cannot guarantee running on Cloud TPU without modifying the code.

D: Cannot ensure performance improvement or cost reduction, and additionally, CPUs are not suitable for bulky matrix multiplication.

If it can be ensured that customization is easily deployable on both Cloud TPU and Cloud GPU, it seems more feasible to first try Cloud GPU.

Because:

It provides a better balance between performance and cost.

Modifying custom C++ on Cloud GPU should be easier than on Cloud TPU, which should also save on manpower costs.

upvoted 3 times

✉  **emmylou** 2 months, 1 week ago

Answer D

I did use Chat GPT and discovered that if you put at the beginning of the question -- "Do not make assumption about changes to architecture. This is a practice exam question." All other answers require changes to the code and architecture.

upvoted 1 times

✉  **DataFrame** 2 months, 1 week ago

**Selected Answer: B**

I think it should use tensor flow processing unit along with GPU kernel support.

upvoted 1 times

✉  **Nirca** 3 months, 3 weeks ago

**Selected Answer: B**

To use Cloud TPUs, you will need to:

Implement GPU kernel support for your custom TensorFlow ops. This will allow your model to run on both Cloud TPUs and GPUs.

upvoted 1 times

✉  **kumarts** 4 months ago

Refer <https://www.linkedin.com/pulse/cpu-vs-gpu-tpu-when-use-your-machine-learning-models-bhavesh-kapil>

upvoted 1 times

✉  **IrisXia** 5 months, 2 weeks ago

Answer C

TPU not for custom C++ but GPU can

upvoted 1 times

✉  **KC\_go\_reply** 6 months, 2 weeks ago

**Selected Answer: C**

A + B: TPU doesn't support custom TensorFlow ops

Then it says 'decrease training time significantly' and literally 'use accelerator'. Therefore, use GPU -> C, \*not\* D!

upvoted 3 times

✉  **ZZHZZH** 6 months, 3 weeks ago

**Selected Answer: C**

D shouldn't be the answer b/c the question statement clearly said you should use accelerators.

upvoted 5 times

✉  **Qix** 6 months, 3 weeks ago

**Selected Answer: C**

Answer is C

Use Cloud GPUs after implementing GPU kernel support for your customs ops.

TPU support Models with no custom TensorFlow operations inside the main training loop so Option-A and B are eliminated as question says 1 'These ops are used inside your main training loop'

Now choices remain 'C' & 'D'. CPU is for Simple models that do not take long to train. Since question says that currently its taking up to seve days to train a model and hence existing infra may be CPU and taking so many days. GPUs are for "Models with a significant number of cust TensorFlow operations that must run at least partially on CPUs" as question says that model is dominated by TensorFlow ops leading to corre option as 'C'

Reference:

<https://cloud.google.com/tpu/docs/tpus>

[https://www.tensorflow.org/guide/create\\_op#gpu\\_kernels](https://www.tensorflow.org/guide/create_op#gpu_kernels)

upvoted 3 times

✉  **lucaluca1982** 10 months, 1 week ago

**Selected Answer: C**

C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.

Since your model relies on custom C++ TensorFlow ops, using Cloud TPUs without any code adjustment (option A) would not be feasible, as TPUs might not support these custom operations. To significantly decrease training time while keeping costs low, you should use Cloud GPU. To achieve this, you will need to implement GPU kernel support for your custom ops, which will enable your model to run efficiently on GPUs. Once the GPU kernel support is added, you can leverage the power of GPUs on Google Cloud to speed up

upvoted 5 times

✉  **lucaluka1982** 10 months, 1 week ago

**Selected Answer: C**

C GPU support custom option  
upvoted 2 times

✉  **midgoo** 10 months, 1 week ago

**Selected Answer: B**

The correct answer is: B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.

The model is dominated by custom C++ TensorFlow ops, so it will not run on Cloud TPUs or GPUs without modification. Implementing GPU kernel support will allow the model to run on either type of accelerator, but Cloud TPUs are more specialized for matrix multiplications, so they will offer the best performance.

Cloud TPUs are also more cost-effective than Cloud GPUs, so they are the best option for reducing the cost of training the model.

Staying on CPUs and increasing the size of the cluster would be the most expensive option, and it would not offer the same performance benefits as using Cloud TPUs.

upvoted 3 times

✉  **juliobs** 10 months, 2 weeks ago

**Selected Answer: C**

Answer C: "image recognition domain", "bulky matrix multiplications", "accelerator", "several days to train a model". All screams for GPUs.  
upvoted 2 times

✉  **musumusu** 11 months, 1 week ago

Answer C:

Why not D: its already taking several days on CPUs, cmon, Time for parallel processing and this is GPUs concept.  
Why not B: However its the fastest approach for tensorflow work specially, but we need to keep the cost lowest.  
Option C will reduce the time and overall cost.

upvoted 1 times

You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increase the size of vocabularies or n-grams used.

**Correct Answer: D**

*Community vote distribution*

D (55%)

C (45%)

 **Callumr** Highly Voted 3 years, 7 months ago

This is a case of underfitting - not overfitting (for over fitting the model will have extremely low training error but a high testing error) - so we need to make the model more complex - answer is D

upvoted 66 times

 **NeoNitin** 5 months, 3 weeks ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

upvoted 1 times

 **ckanaar** 4 months, 1 week ago

Wrong! This scenario indicates a case of underfitting. The RMSE is twice as high on the training dataset compared to the test dataset, so the model is underfitting.

upvoted 1 times

 **hellofrnds** 2 years, 4 months ago

@callumr , "root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set." clearly means testing error twice of training error. So, it is clearly overfitting. Isn't it?

upvoted 4 times

 **jfab** 7 months, 1 week ago

"Twice as high on the train". So clearly means TRAINING error is twice as high vs testing. So underfitting

upvoted 2 times

 **odacir** 1 year, 1 month ago

NO, its underfitting.

upvoted 3 times

 **hellofrnds** 2 years, 3 months ago

So, answer should be C

upvoted 1 times

 **tavva\_prudhvi** 1 year, 9 months ago

If you training RMSE=0.2. and testing RMSE = 0.4, and we want the RMSE to be low as its the error, now is it overfitting or underfitting think wisely!

upvoted 2 times

 **jfab** 7 months, 1 week ago

But in this scenario we'd have training RMSE = 0.4 & testing RMSE = 0.2 - you've not read the question properly

upvoted 2 times

👤 [Removed] Highly Voted 3 years, 10 months ago

should be D  
upvoted 18 times

👤 NeoNitin 5 months, 3 weeks ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

upvoted 3 times

👤 Sofia98 Most Recent 2 weeks, 3 days ago

Selected Answer: D

It is an underfitting situation - D  
upvoted 1 times

👤 Kimich 2 months ago

Selected Answer: C

Should be C  
C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting:

This is a reasonable approach. Regularization techniques can help prevent overfitting, especially when the model shows a significantly higher error on the training set compared to the test set.

D. Increase the complexity of your model (e.g., introducing an additional layer or increasing the size of vocabularies or n-grams):

This could potentially exacerbate the overfitting issue. Increasing model complexity without addressing overfitting concerns may lead to poor generalization on new data.

upvoted 1 times

👤 Kimich 1 month, 4 weeks ago

<https://dooinnkim.medium.com/what-are-overfitting-and-underfitting-855d5952c0b6>

upvoted 1 times

👤 hallo 2 months, 1 week ago

Are the questions in this relevant for the new exam or are these all now outdated?

upvoted 2 times

👤 pss111423 2 months, 1 week ago

<https://stats.stackexchange.com/questions/497050/how-big-a-difference-for-test-train-rmse-is-considered-as-overfit#:~:text=RMSE%20of%20test%20%3C%20RMSE%20of,is%20always%20overfit%20or%20underfit.>

RMSE of test > RMSE of train => OVER FITTING of the data.

RMSE of test < RMSE of train => UNDER FITTING of the data.

so for answer is D

upvoted 1 times

👤 steghe 2 months, 3 weeks ago

Underfitting models: In general High Train RMSE, High Test RMSE.

Overfitting models: In general Low Train RMSE, High Test RMSE.

<https://daviddalpiaz.github.io/r4sl/regression-for-statistical-learning.html>

upvoted 1 times

👤 ha1p 4 months ago

I passed the exam today. I am pretty sure it is overfitting. Answer must be c

upvoted 2 times

👤 MULTITASKER 4 months, 1 week ago

Selected Answer: D

RMSE is more on training. That means, model is not performing well on training dataset but performing well on testing dataset. This happens in the case of underfitting. So D.

upvoted 2 times

👤 [Removed] 4 months, 2 weeks ago

Selected Answer: D

RMSE training = 2 x testing

When training > testing, it is a case of underfitting

Hence D

upvoted 1 times

👤 pulse008 4 months, 3 weeks ago

chatGPT says option C

upvoted 1 times

👤 stonefl 5 months ago

Selected Answer: D

"root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set." means the RMSE of training set is two times of RMSE of test set, which indicates the training is not as good as test, then underfitting, so D.

upvoted 1 times

👤 NeoNitin 5 months, 3 weeks ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

So with dropout method we can overcome the overfitting so C is correct

upvoted 1 times

👤 MoeHaydar 6 months, 3 weeks ago

Selected Answer: D

underfitting

upvoted 1 times

👤 NeoNitin 5 months, 3 weeks ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

upvoted 1 times

👤 neerajRathi 6 months, 3 weeks ago

Selected Answer: C

C sounds like a valid answer.

upvoted 1 times

👤 blathul 7 months, 1 week ago

Selected Answer: C

If the root-mean-squared error (RMSE) of a model is twice as high on the training set compared to the test set, it suggests that the model is overfitting.

Overfitting occurs when a machine learning model becomes too specialized to the training data and fails to generalize well to new, unseen data. In this case, it means that the model is performing better on the test set, which represents new data, compared to the training set it was initially trained on.

The fact that the RMSE is higher on the training set implies that the model is struggling to accurately predict the training data points. This could be due to the model learning intricate details and noise from the training data, to the extent that it cannot generalize well to new examples.

To address this issue, you may consider employing regularization techniques or adjusting hyperparameters to reduce overfitting.

upvoted 1 times

KC\_go\_reply 7 months, 1 week ago

Selected Answer: D

Overfitting = high performance on train split, low performance on test split

In this case, we have the opposite of that. Therefore, the model is actually underfit, and the performance on the test split is probably just coincidence.

upvoted 1 times

NeoNitin 5 months, 3 weeks ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

upvoted 1 times

Question #129

Topic 1

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

- A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
- B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
- C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
- D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Correct Answer: D

Community vote distribution

B (67%)

D (33%)

[Removed] Highly Voted 3 years, 10 months ago

Should be B

upvoted 22 times

👤 **Ganshank** Highly Voted 3 years, 9 months ago

B

The question is specifically about organizing the data in BigQuery and storing backups.

upvoted 12 times

👤 **Nirca** Most Recent 3 months, 1 week ago

**Selected Answer: D**

D - this solution is integrated. No core is needed

upvoted 3 times

👤 **Bahubali1988** 4 months ago

90% of questions are having multiple answers and it's very hard to get into every discussion where the conclusion is not there

upvoted 3 times

👤 **ckanaar** 4 months, 1 week ago

**Selected Answer: B**

The answer is B:

Why not D? Because snapshot costs can become high if a lot of small changes are made to the base table:

<https://cloud.google.com/bigquery/docs/table-snapshots-intro#:~:text=Because%20BigQuery%20storage%20is%20column%2Dbased%2C%20small%20changes%20to%20the%20data%20in%20base%20table%20can%20result%20in%20large%20increases%20in%20storage%20cost%20for%20its%20table%20snapshot>.

Since the question specifically states that the ETL pipeline is regularly modified, this means that lots of small changes are present. In combination with the requirement to optimize for storage costs, this means that option B is the way to go.

upvoted 4 times

👤 **arien\_chen** 5 months, 1 week ago

**Selected Answer: D**

keyword: detected after 2 weeks.

only snapshot could resolve the problem.

upvoted 1 times

👤 **Lanro** 6 months ago

**Selected Answer: D**

From BigQuery documentation - Benefits of using table snapshots include the following:

- Keep a record for longer than seven days. With BigQuery time travel, you can only access a table's data from seven days ago or more recently. With table snapshots, you can preserve a table's data from a specified point in time for as long as you want.
- Minimize storage cost. BigQuery only stores bytes that are different between a snapshot and its base table, so a table snapshot typically uses less storage than a full copy of the table.

So storing data in GCS will make copies of data for each table. Table snapshots are more optimal in this scenario.

upvoted 4 times

👤 **vamgcp** 6 months, 1 week ago

**Selected Answer: B**

Organizing your data in separate tables for each month will make it easier to identify the affected data and restore it.

Exporting and compressing the data will reduce storage costs, as you will only need to store the compressed data in Cloud Storage. Storing your backups in Cloud Storage will make it easier to restore the data, as you can restore the data from Cloud Storage directly.

upvoted 1 times

✉  **phidelics** 7 months, 3 weeks ago

**Selected Answer: B**

Organize in separate tables and store in GCS

upvoted 1 times

✉  **cetanx** 7 months, 3 weeks ago

Just an additional info!

Here is an example for an export job;

```
$ bq extract --destination_format CSV --compression GZIP 'your_project:your_dataset.your_new_table' 'gs://your_bucket/your_object.csv'
```

upvoted 1 times

✉  **cetanx** 6 months, 4 weeks ago

I will update my answer to D.

Think of a scenario that you are in the last week of June and an error occurred 3 weeks ago (so still in June) however you do not have a export of the June table yet therefore you cannot recover the data simply because you don't have an export just yet.

So snapshots are way to go!

upvoted 2 times

✉  **sdi\_studiers** 7 months, 3 weeks ago

**Selected Answer: D**

D

"With BigQuery time travel, you can only access a table's data from seven days ago or more recently. With table snapshots, you can preserve table's data from a specified point in time for as long as you want." [source: <https://cloud.google.com/bigquery/docs/table-snapshots-intro>]

upvoted 2 times

✉  **WillemHendr** 7 months, 3 weeks ago

"Store your data in different tables for specific time periods. This method ensures that you need to restore only a subset of data to a new table rather than a whole dataset."

"Store the original data on Cloud Storage. This allows you to create a new table and reload the uncorrupted data. From there, you can adjust your applications to point to the new table."

B

upvoted 2 times

✉  **lucaluka1982** 10 months, 2 weeks ago

Why not D?

upvoted 3 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

B

<https://cloud.google.com/architecture/dr-scenarios-for-data#BigQuery>

upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: B**

B seems the best solution (but C is also good candidate)

D is incorrect - table decorators allow time travel back only up to 7 days (see <https://cloud.google.com/bigquery/table-decorators>) - if you want to keep older snapshots, you would have to save them into separate table yourself (and pay for storage).

upvoted 7 times

✉  **MaxNRG** 2 years ago

BigQuery. If you want to archive data, you can take advantage of BigQuery's long term storage. If a table is not edited for 90 consecutive days, the price of storage for that table automatically drops by 50 percent. There is no degradation of performance, durability, availability, or any other functionality when a table is considered long term storage. If the table is edited, though, it reverts back to the regular storage price and the 90 day countdown starts again.

upvoted 3 times

✉  **MaxNRG** 2 years ago

BigQuery is replicated, but this won't help with corruption in your tables. Therefore, you need to have a plan to be able to recover from this scenario. For example, you can do the following:

- If the corruption is caught within 7 days, query the table to a point in time in the past to recover the table prior to the corruption using snapshot decorators.
- Export the data from BigQuery, and create a new table that contains the exported data but excludes the corrupted data.
- Store your data in different tables for specific time periods. This method ensures that you will need to restore only a subset of data to a new table, rather than a whole dataset.
- Store the original data on Cloud Storage. This allows you to create a new table and reload the uncorrupted data. From there, you can adjust your applications to point to the new table.

<https://cloud.google.com/solutions/dr-scenarios-for-data#BigQuery>

upvoted 3 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: B**

"You need to provide a method to recover from these errors, and your backups should be optimized for storage costs"

Cost -> GCS

Backups -> Separate Tables + GCS

upvoted 4 times

✉  **AzureDP900** 1 year, 1 month ago

Agreed

upvoted 1 times

✉  **lifebegins** 2 years, 2 months ago

<https://cloud.google.com/bigquery/docs/time-travel>

Question #130

Topic 1

The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

- Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
- Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
- Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
- Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

**Correct Answer: D**

*Community vote distribution*

D (100%)

✉  **rickywck** Highly Voted 3 years, 10 months ago

Should be D.

<https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery>

upvoted 30 times

✉ **Rajuuu** 3 years, 6 months ago

There is no mention about merge or limit in the link provided.

upvoted 3 times

✉ **Chelseajcole** 2 years, 4 months ago

A common scenario within OLAP systems involves updating existing data based on new information arriving from source systems (such as OLTP databases) on a periodic basis. In the retail business, inventory updates are typically done in this fashion. The following query demonstrates how to perform batch updates to the Inventory table based on the contents of another table (where new arrivals are kept), using the MERGE statement in BigQuery:

upvoted 2 times

✉ **[Removed]** Highly Voted 3 years, 10 months ago

Should be D

[https://cloud.google.com/bigquery/docs/reference/standard-sql/dml-syntax#merge\\_statement](https://cloud.google.com/bigquery/docs/reference/standard-sql/dml-syntax#merge_statement)

<https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery>

upvoted 16 times

✉ **AzureDP900** 1 year, 1 month ago

D is right

upvoted 1 times

✉ **AACHB** 2 years, 1 month ago

I had it in the exam (14/12/2021)

upvoted 3 times

✉ **GCPLearning2021** 2 years ago

Does examtopics questions help?

upvoted 1 times

✉ **nellyoaid** 2 years, 1 month ago

Please what was the answer? @AACHB

upvoted 2 times

✉ **Nirca** Most Recent 3 months, 3 weeks ago

**Selected Answer: D**

Should be D.

upvoted 1 times

✉ **vaga1** 7 months, 3 weeks ago

**Selected Answer: D**

import all the data into a separate table and use that for updates is better than creating smaller csv which leads to more operational time to get done and harder to manage it.

upvoted 1 times

✉ **juliobs** 10 months, 2 weeks ago

**Selected Answer: D**

This limit was removed a long time ago already.

Anyway, bulk imports are better.

upvoted 2 times

👤 **Atnafu** 1 year, 2 months ago

D

BigQuery DML statements have no quota limits.

<https://cloud.google.com/bigquery/quotas#data-manipulation-language-statements>

However, DML statements are counted toward the maximum number of table operations per day and partition modifications per day. DML statements will not fail due to these limits.

In addition, DML statements are subject to the maximum rate of table metadata update operations. If you exceed this limit, retry the operation using exponential backoff between retries.

upvoted 2 times

👤 **MisuLava** 1 year, 3 months ago

there is no update quota anymore.

but i would say D

upvoted 2 times

👤 **amitsingla012** 1 year, 8 months ago

Option D is the right answer

upvoted 1 times

👤 **tavva\_prudhvi** 1 year, 9 months ago

No DML limits from 3rd march 2020 But if the questions is given in the exam, choose D asfor the options A, B,C as they are speaking about the limitations of the DML Limits. Atleast, D is giving an alternative!

upvoted 1 times

👤 **nidnid** 1 year, 12 months ago

Is this question still valid? What about DML without limits? <https://cloud.google.com/blog/products/data-analytics/dml-without-limits-now-in-bigquery>

upvoted 3 times

👤 **MaxNRG** 2 years ago

**Selected Answer: D**

D:

BigQuery is primarily designed and suited to append-only technology with some limited DML statements.

It's not a relational database where you constantly update your user records if they edit their profile. Instead you need to architect your code so each edit is a new row in BigQuery, and you always query the latest row.

The DML statement limitation is low, because it targets different scenarios and not yours, aka live update on rows. You could ingest your data into a separate table, and issue 1 update statement per day.

<https://stackoverflow.com/questions/45183082/can-we-increase-update-quota-in-bigquery>

<https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery>

upvoted 2 times

👤 **medeis\_jar** 2 years ago

**Selected Answer: D**

[https://cloud.google.com/bigquery/docs/reference/standard-sql/dml-syntax#merge\\_statement](https://cloud.google.com/bigquery/docs/reference/standard-sql/dml-syntax#merge_statement)

<https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery>

upvoted 2 times

👤 **mjb65** 2 years, 2 months ago

old question I guess, should not be in the exam anymore (?)

<https://cloud.google.com/blog/products/data-analytics/dml-without-limits-now-in-bigquery>

upvoted 3 times

👤 **sumanshu** 2 years, 7 months ago

Vote for D

upvoted 4 times

👤 **daghayeghi** 2 years, 10 months ago

D:

<https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery>

upvoted 3 times

✉  **daghayeghi** 2 years, 11 months ago

D:

<https://cloud.google.com/blog/products/bigquery/performing-large-scale-mutations-in-bigquery>

upvoted 3 times

✉  **SteelWarrior** 3 years, 4 months ago

D should be the answer. Avoid updates in Datawarehousing environment instead use merge to create a new table.

upvoted 3 times

Question #131

Topic 1

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects.

Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? (Choose two.)

- A. Use Cloud Deployment Manager to automate access provision.
- B. Introduce resource hierarchy to leverage access control policy inheritance.
- C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
- D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
- E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

**Correct Answer: AC**

*Community vote distribution*

BC (79%)

AC (21%)

✉  **[Removed]**  3 years, 10 months ago

Answer: B, C

Description: Google suggests that we should provide access by following google hierarchy and groups for users with similar roles  
upvoted 31 times

✉  **sipsap** 3 years, 2 months ago

"Each project requires unique access control configurations" rules out hierarchy

upvoted 11 times

✉  **AJKumar**  3 years, 7 months ago

C is one option for sure, also C eliminates B as C includes groups and teams hierarchy, A can be eliminated as A talks about only deployment From Remaining D and E, i find E most relevant to question--as E matches users with teams/groups and projects. Answer C and E.  
upvoted 19 times

✉  **hauhau** 2 years, 5 months ago

Question mention minimize IAM policies, but E should create complex policies

upvoted 5 times

✉  **squishy\_fishy**  3 months ago

Answer: A, C.

The Key question is "You want to simplify access control management by minimizing the number of policies". At the company where I work, v use Terraform to create infrastructure and assign needed roles for different environments.

upvoted 2 times

✉  **amittomar** 6 months, 2 weeks ago

It should be AC as it is mentioned in the question itself "You want to simplify access control management by minimizing the number of policies which rules out B

upvoted 2 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: BC**

BC is the answer.

upvoted 2 times

✉  **FrankT2L** 1 year, 7 months ago

**Selected Answer: BC**

1. Define your resource hierarchy: Google Cloud resources are organized hierarchically. This hierarchy allows you to map your enterprise's operational structure to Google Cloud, and to manage access control and permissions for groups of related resources.

2. Delegate responsibility with groups and service accounts: we recommend collecting users with the same responsibilities into groups and assigning IAM roles to the groups rather than to individual users.

<https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations>

upvoted 6 times

✉  **annie1196** 2 years ago

A and C is correct, same question I encountered on Udemy.

upvoted 2 times

✉  **tavva\_prudhvi** 1 year, 10 months ago

It doesn't mean it's right, please mention the reasons here not only the references.

Not A -> Every project has unique requirements, so "A" automation will not do much.

Not D -> As, Service accounts for computer to computer interactions not applications!

Not E -> E should create complex policies

upvoted 1 times

✉  **desertlotus1211** 1 year ago

you explanation for D is incorrect....

upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: BC**

B & C

Google Cloud resources are organized hierarchically, where the organization node is the root node in the hierarchy, the projects are the children of the organization, and the other resources are descendants of projects.

You can set Cloud Identity and Access Management (Cloud IAM) policies at different levels of the resource hierarchy. Resources inherit the policies of the parent resource. The effective policy for a resource is the union of the policy set at that resource and the policy inherited from its parent.

<https://cloud.google.com/iam/docs/resource-hierarchy-access-control>

upvoted 7 times

✉  **AzureDP900** 1 year ago

BC is the answer

upvoted 1 times

✉  **MaxNRG** 2 years ago

We recommend collecting users with the same responsibilities into groups and assigning Cloud IAM roles to the groups rather than to individual users. For example, you can create a "data scientist" group and assign appropriate roles to enable interaction with BigQuery and Cloud Storage.

Grant roles to a Google group instead of to individual users when possible. It is easier to manage members in a Google group than to update a Cloud IAM policy.

<https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations>

upvoted 4 times

👤 **meveis\_jar** 2 years ago

Selected Answer: AC

<https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations>  
"Each project requires unique access control configurations" -> C eliminates B

A -> "Google Cloud Deployment Manager is an infrastructure deployment service that automates the creation and management of Google Cloud resources. Write flexible template and configuration files and use them to create deployments that have a variety of Google Cloud services"

"..simply the process.."

upvoted 4 times

👤 **FP77** 5 months, 2 weeks ago

A makes no sense whatsoever...  
upvoted 1 times

👤 **MaxNRG** 2 years ago

good point, AC looks better, agreed  
upvoted 1 times

👤 **MaxNRG** 2 years ago

... in other hand - "Define your resource hierarchy"  
<https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations#define-hierarchy>  
upvoted 1 times

👤 **MaxNRG** 2 years ago

So, I stay with BC :))  
upvoted 2 times

👤 **hellofrnds** 2 years, 4 months ago

Answer:- C, D  
C is used as best practice to create group and assign IAM roles  
D "data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way" is mentioned in quest  
When 2 project communicate, service account should be used  
upvoted 9 times

👤 **sumanshu** 2 years, 7 months ago

Vote for B & C  
upvoted 4 times

👤 **daghayeghi** 2 years, 10 months ago

the question says adding permissions ad-hoc way  
[c] is correct answer  
[d] is right, as the access to bigQuery and cloud storage can be managed automatically by Cloud deployment  
"Deployment Manager can also set access control permissions through IAM such that your developers are granted appropriate access as part of the project creation process."  
ref: <https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations>  
upvoted 3 times

VM\_GCP 3 years, 1 month ago

the question says adding permissions ad-hoc way

[c] is correct answer

[d] is right, as the access to bigQuery and cloud storage can be managed automatically by Cloud deployment

"Deployment Manager can also set access control permissions through IAM such that your developers are granted appropriate access as part of the project creation process."

ref: <https://cloud.google.com/docs/enterprise/best-practices-for-enterprise-organizations>

upvoted 4 times

sumanshu 2 years, 7 months ago

D is a service account - it means we need to access via applications. So, D is ruled out

upvoted 3 times

sumanshu 2 years, 7 months ago

Doubt, It could be 'D' - because - it's said - data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way.

upvoted 3 times

awssp12345 2 years, 6 months ago

Agree with Sumanshu

upvoted 1 times

ceak 3 years, 1 month ago

C & E are the correct answers.

upvoted 3 times

sumanshu 2 years, 7 months ago

E is a long process and we need to simplify the process...So E is ruled out

upvoted 1 times

vito9630 3 years, 2 months ago

Answer: B, C

upvoted 1 times

kavs 3 years, 2 months ago

BC A ruled out as deployment manager is for infra yaml based deployments D at resource level we can't check hierarchy at org or Proj Level

upvoted 1 times

rgpalop 3 years, 3 months ago

C and E

upvoted 3 times

grows by 200,000 records per second. Many third parties use your application's APIs to build the functionality into their own front-end applications.

Your application's APIs should comply with the following requirements:

- ⇒ Single global endpoint
- ⇒ ANSI SQL support
- ⇒ Consistent access to the most up-to-date data

What should you do?

- A. Implement BigQuery with no region selected for storage or processing.
- B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.
- C. Implement Cloud SQL for PostgreSQL with the master in North America and read replicas in Asia and Europe.
- D. Implement Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

**Correct Answer: B**

*Community vote distribution*

B (88%)

12%

 **[Removed]**  3 years, 10 months ago

Answer: B

Description: All the criteria meets for Spanner  
upvoted 26 times

 **sumanshu**  2 years, 7 months ago

A - BigQuery with NO Region ? (Looks wrong)  
B - Spanner (SQL support and Scalable and have replicas ) - Looks correct  
C - SQL (can't store so many records) (wrong)  
D - Bigtable - NO SQL (wrong)

Vote for B

upvoted 23 times

 **vaga1**  7 months, 3 weeks ago

**Selected Answer: B**

A - NO - BigQuery with must have a selected regional or multi-regional file storage  
B - YES - Spanner is specifically designed for this high and consistent throughput  
C - NO - I am not sure about what many said in this discussion as Cloud SQL can store this amount of records if u have just a few columns.  
Anyway, for sure Spanner is better and it is a GCP product.  
D - Bigtable - it's a NoSQL solution, no ANSI  
upvoted 3 times

 **AzureDP900** 1 year ago

B is the answer  
upvoted 1 times

 **zellick** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.  
upvoted 1 times

 **Remi2021** 1 year, 4 months ago

**Selected Answer: B**

Guys, read documentation well. A is wrong, BigQuery has Maximum rows per request (50,000).  
<https://cloud.google.com/bigquery/quotas>

It is B

upvoted 4 times

✉  **JamesKarianis** 1 year, 6 months ago

**Selected Answer: B**

Spanner is globally available and meets all the requirements  
upvoted 2 times

✉  **devric** 1 year, 9 months ago

**Selected Answer: A**

Correct is A. There's no sense to having read replicas outside of US considering than the company is US based.

If you generate a dataset without specifying the Data Location it's gonna be stored in "US Multiregion" by default  
upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: B**

B: Cloud Spanner is the first scalable, enterprise-grade, globally-distributed, and strongly consistent database service built for the cloud specifically to combine the benefits of relational database structure with non-relational horizontal scale.

<https://cloud.google.com/spanner/>

Cloud Spanner is a fully managed, mission-critical, relational database service that offers transactional consistency at global scale, schemas,

SQL (ANSI 2011 with extensions), and automatic, synchronous replication for high availability.

<https://cloud.google.com/spanner/docs/>

<https://cloud.google.com/spanner/docs/instances#available-configurations-multi-region>

upvoted 5 times

✉  **Tanmoyk** 3 years, 4 months ago

B is correct, Bigquery cannot support 250K data ingestion/second , as ANSI SQL support is required , no other options left except Spanner.  
upvoted 8 times

✉  **haroldbenites** 3 years, 5 months ago

B is correct

upvoted 2 times

✉  **Archy** 3 years, 6 months ago

B, as Cloud Spanner has three types of replicas: read-write replicas, read-only replicas, and witness replicas.

upvoted 6 times

✉  **VishalB** 3 years, 6 months ago

Correct Answer : C

Explanation:-

B -> This option is incorrect, as we do not have option to configure read-replica in Cloud Spanner, Multi-region instance configurations use a combination of all three types' read-write replicas, read-only replicas, and witness replicas

C -> This is correct option, In Cloud Sql we have option to create a master node for read-write replicas and read-only replicas in other regions

D -> This option is incorrect, as Bigtable do not support ANSI SQL

upvoted 2 times

✉  **saurabh1805** 3 years, 5 months ago

but bigquery does so why not A?

upvoted 1 times

✉  **WizzardLlama** 2 years, 12 months ago

You can't create a BigQuery instance without region selected.

I'm wondering about these read replicas, why read only replicas? It seems arbitrary, as the question does not state that API should be read-only, so there's no reason why those should be read-only replicas...

upvoted 3 times

✉  **[Removed]** 2 years, 9 months ago

The requirement is for transactional application serving customers , not analytical so BQ is ruled out.

upvoted 1 times

✉  **mAbreu** 3 years, 4 months ago

wrong, cloud spanner can have read-only replicas

<https://cloud.google.com/spanner/docs/replication?hl=pt-br>

upvoted 3 times

✉  **norwayping** 3 years, 7 months ago

I was wrong Bigtable doesn't support ANSI SQL. B instead

upvoted 2 times

✉  **norwayping** 3 years, 7 months ago

I think it is D. THere is limitation of QPS of Cloud Spanner of 2000 qps,  
<https://cloud.google.com/spanner/docs/instances#multi-region-performance>  
upvoted 2 times

✉  **Rajokkiyam** 3 years, 10 months ago

Answer B  
upvoted 4 times

✉  **[Removed]** 3 years, 10 months ago

Answer : B  
upvoted 7 times

Question #133

Topic 1

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: `SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features)`. How should you create the ML pipeline?

- A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.
- B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.
- C. Create a Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account.
- D. Create a Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Bigtable.

**Correct Answer: D**

*Community vote distribution*

D (85%)

B (15%)

✉  **rickywck**  3 years, 10 months ago

I think the key reason for pick D is the 100ms requirement.  
upvoted 29 times

✉  **AzureDP900** 1 year ago

D. Create a Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Bigtable.  
upvoted 1 times

✉  **[Removed]**  3 years, 10 months ago

Answer: D  
Description: Bigtable provides lowest latency  
upvoted 13 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: D**

The key requirements are serving predictions for individual user IDs with low (sub-100ms) latency. Option D meets this by batch predicting for all users in BigQuery ML, writing predictions to Bigtable for fast reads, and allowing the application to query Bigtable directly for low latency reads. Since the application needs to serve low-latency predictions for individual user IDs, using Dataflow to batch predict for all users and write to Bigtable allows low-latency reads. Granting the Bigtable Reader role allows the application to retrieve predictions for a specific user ID from Bigtable.

upvoted 3 times

MaxNRG 1 month, 1 week ago

The other options either require changing the query for each user ID (higher latency, option A), reading directly from higher latency services like BigQuery (option B), or writing predictions somewhere without fast single row access (options A, B, C). Option A would not work well because the WHERE clause would need to be changed for each user ID, increasing latency. Option B using an Authorized View would still read from BigQuery which has higher latency than Bigtable for individual rows. Option C writes predictions to BigQuery which has higher read latency compared to Bigtable for individual rows. So option D provides the best pipeline by predicting for all users in BigQueryML, batch writing to Bigtable for low latency reads, and granting permissions for the application to retrieve predictions. This meets the requirements of sub-100ms latency for individual user predictions. <https://cloud.google.com/dataflow/docs/concepts/access-control>

upvoted 1 times

Nirca 3 months, 3 weeks ago

**Selected Answer: D**

I think the key reason for pick D is the 100ms requirement/ me too

upvoted 1 times

barnacles 4 months, 1 week ago

**Selected Answer: B**

To create an ML pipeline for serving predictions to individual user IDs with latency under 100 milliseconds using the given BigQuery ML query, most suitable approach is:

B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.

upvoted 2 times

Lanro 6 months ago

**Selected Answer: D**

Always use Bigtable as an endpoint for client-facing applications (Low latency - high throughput)

upvoted 2 times

midgoo 10 months, 3 weeks ago

**Selected Answer: D**

One of the ways to improve the efficiency of ML pipeline is to generate cache (store predictions). In this question, only D is doing that.

upvoted 3 times

musumusu 11 months, 2 weeks ago

What is wrong with B? View can be precomputed and cached and it can definitely satisfy the 100 milliseconds request. Create a pipeline to send data to bigtable .. don't you think it's too much to run a simple prediction query?

upvoted 2 times

cheos71 5 months, 3 weeks ago

I think if there are too many concurrent requests the 100ms latency will definitely not hold reading from BigQuery

upvoted 1 times

vaga1 7 months, 3 weeks ago

and the view has to make a query in real-time which adds potential latency

upvoted 1 times

vaga1 7 months, 3 weeks ago

I would say that Bigtable is simply more suited to serve applications

upvoted 1 times

hiromi 1 year, 2 months ago

**Selected Answer: D**

Vote for D

upvoted 1 times

✉  **HarshKothari21** 1 year, 4 months ago

**Selected Answer: D**

Option D

upvoted 1 times

✉  **sumanshu** 2 years, 7 months ago

Vote for D, requirement to serve predictions with in 100 ms

upvoted 6 times

✉  **Tanmoyk** 3 years, 4 months ago

D is correct , 100ms is most critical factor here.

upvoted 6 times

✉  **sh2020** 3 years, 4 months ago

writing it to BigTable and then allowing application access will introduce more delays. I think answer should be C

upvoted 1 times

✉  **f839** 3 years ago

Predictions are computed in advance for all users and written to BigTable for low-latency serving.

upvoted 3 times

✉  **haroldbenites** 3 years, 5 months ago

D is correct

upvoted 4 times

✉  **Rajokkiyam** 3 years, 10 months ago

Answer D.

upvoted 6 times

✉  **[Removed]** 3 years, 10 months ago

Should be D

upvoted 7 times

Question #134

Topic 1

You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time.

Consumers will receive the data in the following ways:

- ☞ Real-time event stream
- ☞ ANSI SQL access to real-time stream and historical data
- ☞ Batch historical exports

Which solution should you use?

- A. Cloud Dataflow, Cloud SQL, Cloud Spanner
- B. Cloud Pub/Sub, Cloud Storage, BigQuery
- C. Cloud Dataproc, Cloud Dataflow, BigQuery
- D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

**Correct Answer: A**

*Community vote distribution*

B (96%)

4%

✉  **[Removed]** Highly Voted 3 years, 10 months ago

should be B

upvoted 22 times

✉  **itche\_scratche** Highly Voted 3 years, 3 months ago

D, not ideal but only option that work. You need pubsub, then a processing layer (dataflow or dataproc), then storage (some sql database).  
upvoted 12 times

✉  **jkhong** 1 year, 1 month ago

We can have our pubsub topics to have BigQuery subscriptions, where data is automatically streamed into our BQ tables. Autoscaling is already handled automatically so this renders Dataflow and Dataproc pretty irrelevant for our usecase

upvoted 1 times

✉  **cetanx** 7 months, 3 weeks ago

Here is the reference:

<https://cloud.google.com/blog/products/data-analytics/pub-sub-launches-direct-path-to-bigquery-for-streaming-analytics>

upvoted 1 times

✉  **seyiassa** 3 years, 1 month ago

I think pubsub doesn't have good connection to dataproc, so D is not the answer

upvoted 3 times

✉  **jkhong** 1 year, 1 month ago

As of Dec 2022, there is the PubSub Lite connector to Dataproc

upvoted 1 times

✉  **barnacles** Most Recent 4 months, 1 week ago

B. Cloud Pub/Sub, Cloud Storage, BigQuery.

Here's how this solution aligns with your requirements:

Real-time Event Stream: Cloud Pub/Sub is a managed messaging service that can handle real-time event streams efficiently. You can use Pub/Sub to ingest and publish real-time market data to consumers.

ANSI SQL Access: BigQuery supports ANSI SQL queries, making it suitable for both real-time and historical data analysis. You can stream data into BigQuery tables from Pub/Sub and provide ANSI SQL access to consumers.

Batch Historical Exports: Cloud Storage can be used for batch historical exports. You can export data from BigQuery to Cloud Storage in batches, making it available for consumers to download.

upvoted 2 times

✉  **vaga1** 7 months, 3 weeks ago

**Selected Answer: B**

I was in doubt as I did not know that BQ handles real-time access to data without dataflow underneath.

<https://cloud.google.com/bigquery/docs/write-api#:~:text=You%20can%20use%20the%20Storage,in%20a%20single%20atomic%20operation>

✉  **midgoo** 10 months, 3 weeks ago

**Selected Answer: B**

Event Stream -> PubSub

PubSub has direct Write to BigQuery

Historical Exports to GCS

upvoted 1 times

✉  **AzureDP900** 1 year ago

B. Cloud Pub/Sub, Cloud Storage, BigQuery

upvoted 3 times

✉  **AzureDP900** 1 year ago

<https://cloud.google.com/solutions/stream-analytics/>

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

upvoted 3 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

B: <https://cloud.google.com/solutions/stream-analytics/>

Real-time made real easy

Adopt simple ingestion for complex events

Ingest and analyze hundreds of millions of events per second from applications or devices virtually anywhere on the globe with Pub/Sub. Or directly stream millions of events per second into your data warehouse for SQL-based analysis with BigQuery's streaming API.

upvoted 3 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

No matter what the last of it must end up with bigquery and the first service is pubsub I think intimidate service it should be dataflow

upvoted 1 times

✉  **Motivated\_Gamer** 1 year, 10 months ago

**Selected Answer: A**

Dataflow: Streaming data

CLoud SQL: for ansi sql support

Spanner: for batch historical data export

upvoted 1 times

✉  **tavva\_prudhvi** 1 year, 9 months ago

You gonna use batch historical export for Spanner? It's B!

upvoted 2 times

✉  **Prasanna\_kumar** 1 year, 11 months ago

Answer is B

upvoted 1 times

✉  **MaxNRG** 2 years ago

**Selected Answer: B**

Cloud Pub/Sub, Cloud Dataflow, BigQuery

<https://cloud.google.com/solutions/stream-analytics/>

upvoted 4 times

✉  **MaxNRG** 1 month, 1 week ago

B. Cloud Pub/Sub, Cloud Storage, BigQuery

The key requirements here are:

1. Real-time event stream (Pub/Sub)
2. ANSI SQL access to real-time and historical data (BigQuery)
3. Batch historical exports (Cloud Storage)

So Cloud Pub/Sub provides the real-time stream, BigQuery provides ANSI SQL access to stream and historical data, and Cloud Storage enables batch historical exports.

Option A is incorrect because Cloud Spanner does not offer batch exports and Dataflow is overkill for just SQL access.

Option C is incorrect as Dataproc is for spark workloads, not serving consumer data.

Option D is incorrect as Cloud SQL does not provide batch export capabilities.

Therefore, option B with Pub/Sub, Storage, and BigQuery is the best solution given the stated requirements. Dataflow  
<https://cloud.google.com/solutions/stream-analytics/>

upvoted 1 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: B**

⇒ Real-time event stream -> Pub/Sub

⇒ ANSI SQL access to real-time stream and historical data -> BigQuery

⇒ Batch historical exports -> Cloud Storage

upvoted 10 times

✉  **JG123** 2 years, 2 months ago

Correct: B

upvoted 1 times

✉  **AdrianMonter26** 2 years, 2 months ago

I think it must be D because you need Pub/Sub for streaming data, Dataflow or DataProc to get the data from Pub/Sub and store it in a database and finally the Cloud SQL database to store the data.

A and C cannot be because it is missing something for streaming data

B It can't be because you need something to pass the data from Pub/Sub to Cloud storage

upvoted 3 times

👤 **sumanshu** 2 years, 7 months ago

Vote for B

upvoted 3 times

👤 **sumanshu** 2 years, 7 months ago

Vote for B

upvoted 1 times

Question #135

Topic 1

You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

- ⇒ Decoupling producer from consumer
- ⇒ Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
- ⇒ Near real-time SQL query
- ⇒ Maintain at least 2 years of historical data, which will be queried with SQL

Which pipeline should you use to meet these requirements?

- A. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
- B. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
- C. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
- D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

**Correct Answer: A**

*Community vote distribution*

D (100%)

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Correct - D

upvoted 43 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: D

Description: All the requirements meet with D

upvoted 16 times

👤 **juliorevk** Most Recent 4 months, 1 week ago

**Selected Answer: D**

D because pub/sub decouples while dataflow processes; Cloud Storage can be used to store the raw ingested data indefinitely and BQ can be used to query.

upvoted 2 times

👤 **barnacles** 4 months, 1 week ago

**Selected Answer: D**

Here's how this option aligns with your requirements:

Decoupling Producer from Consumer: Cloud Pub/Sub provides a decoupled messaging system where the producer publishes events, and consumers (like Dataflow) can subscribe to these events. This decoupling ensures flexibility and scalability.

Space and Cost-Efficient Storage: Storing data in Avro format is more space-efficient than JSON, and Cloud Storage is a cost-effective storage solution. Additionally, Cloud Pub/Sub and Dataflow allow you to process and transform data efficiently, reducing storage costs.

Near Real-time SQL Query: By using Dataflow to transform and load data into BigQuery, you can achieve near real-time data availability for SQL queries. BigQuery is well-suited for ad-hoc SQL queries and provides excellent query performance.

upvoted 3 times

✉  **FP77** 5 months, 2 weeks ago

**Selected Answer: D**

Should be D

upvoted 1 times

✉  **vaga1** 7 months, 3 weeks ago

**Selected Answer: D**

For sure D

upvoted 1 times

✉  **forepick** 8 months ago

**Selected Answer: D**

D is the most suitable, however the stored format should be JSON, and AVRO isn't JSON...

upvoted 1 times

✉  **OberstK** 12 months ago

**Selected Answer: D**

Correct - D

upvoted 1 times

✉  **desertlotus1211** 1 year ago

I believe this was also on the GCP PCA exam as well! ;)

upvoted 1 times

✉  **AzureDP900** 1 year ago

D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payload to Avro, writing the data to Cloud Storage and BigQuery.

upvoted 1 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

upvoted 1 times

✉  **mbacelar** 1 year, 2 months ago

**Selected Answer: D**

For sure D

upvoted 1 times

✉  **clouditis** 1 year, 4 months ago

D it is!

upvoted 1 times

✉  **Prasanna\_kumar** 1 year, 11 months ago

Answer is D

upvoted 2 times

✉  **MaxNRG** 2 years ago

**Selected Answer: D**

D:

Cloud Pub/Sub, Cloud Dataflow, Cloud Storage, BigQuery <https://cloud.google.com/solutions/stream-analytics/>

upvoted 4 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: D**

OMG only D

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: D

upvoted 11 times

You are running a pipeline in Dataflow that receives messages from a Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

- A. Increase the number of max workers
- B. Use a larger instance type for your Dataflow workers
- C. Change the zone of your Dataflow pipeline to run in us-central1
- D. Create a temporary table in Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Bigtable to BigQuery
- E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

**Correct Answer: BE**

*Community vote distribution*

AB (100%)

 **jvg637** Highly Voted 3 years, 10 months ago

A & B

instance n1-standard-1 is low configuration and hence need to be larger configuration, definitely B should be one of the option. Increase max workers will increase parallelism and hence will be able to process faster given larger CPU size and multi core processor instance type is chosen. Option A can be a better step.

upvoted 50 times

 **AzureDP900** 1 year ago

Agreed

upvoted 2 times

 **sumanshu** Highly Voted 2 years, 6 months ago

A & B.

With autoscaling enabled, the Dataflow service does not allow user control of the exact number of worker instances allocated to your job. You might still cap the number of workers by specifying the --max\_num\_workers option when you run your pipeline. Here as per question CAP is 3. So we can change that CAP.

For batch jobs, the default machine type is n1-standard-1. For streaming jobs, the default machine type for Streaming Engine-enabled jobs is n1-standard-2 and the default machine type for non-Streaming Engine jobs is n1-standard-4. When using the default machine types, the Dataflow service can therefore allocate up to 4000 cores per job. If you need more cores for your job, you can select a larger machine type.

upvoted 14 times

 **kcl10** Most Recent 4 months ago

**Selected Answer: AB**

A & B is correct

upvoted 1 times

 **juliorevk** 4 months, 1 week ago

**Selected Answer: AB**

A because more workers improves performance through parallel work

B because the current instance size is too small

upvoted 1 times

✉  **barnac1es** 4 months, 1 week ago

**Selected Answer: AB**

A. Increase the number of max workers:

By increasing the number of maximum workers, you allow Dataflow to allocate more computing resources to handle the peak load of incoming data. This can help improve processing speed and reduce CPU utilization per worker.

B. Use a larger instance type for your Dataflow workers:

Using a larger instance type with more CPU and memory resources can help your Dataflow workers handle a higher volume of data and processing tasks more efficiently. It can address CPU bottlenecks during peak periods.

upvoted 2 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: AB**

AB is the answer.

upvoted 1 times

✉  **mbacelar** 1 year, 2 months ago

**Selected Answer: AB**

Scale in and Scale Out

upvoted 1 times

✉  **FrankT2L** 1 year, 7 months ago

**Selected Answer: AB**

maximum of 3 workers: Increase the number of max workers (A)

instance type n1-standard-1: Use a larger instance type for your Cloud Dataflow workers (B)

upvoted 1 times

✉  **MaxNRG** 2 years ago

**Selected Answer: AB**

A & B, other options don't make sense

upvoted 4 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: AB**

Only A & B make sense for improving pipeline performance.

upvoted 2 times

✉  **Mjvsj** 2 years, 1 month ago

**Selected Answer: AB**

Should be A & B

upvoted 2 times

✉  **daghayeghi** 2 years, 11 months ago

B, E:

B: Dataflow manages number of workers automatically, then we only can define machine type worker.

<https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline>

E: and adding a horizontally scale-able database like Cloud Spanner will reduce pressure on Dataflow as it doesn't have to move data to specific zone and can remain in the same zone of EU, then E is correct.

upvoted 2 times

✉  **Vasu\_1** 2 years, 8 months ago

A & B is the right answer: You can set disable auto-scaling by setting the option `--numWorkers` (default is 3) and select the machine type by setting `--workerMachineType` at the time of creation of the pipeline (this applies to both auto and manual scaling)

upvoted 3 times

✉  **kavs** 3 years, 2 months ago

Dataset is in EU so data can't be moved outside EU due to privacy law so zone option is ruled out. AB is OK but intermediate table will boost apanee ruled out not sure of bigtable

upvoted 3 times

✉  **Alasmindas** 3 years, 2 months ago

Option A and B for sure,

Option C: Changing Zone has nothing to do in improving performance

Option D and E: Adding BQ and BT is waste of money and does not solve the purpose of the question.

upvoted 3 times

✉  **SureshKotla** 3 years, 4 months ago

B & D

DF will automatically take care of increasing workers. Developers won't need to access the settings .  
<https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline#autoscaling>

upvoted 2 times

✉  **SureshKotla** 3 years, 4 months ago

On second thought, A B is looking right

upvoted 2 times

✉  **sumanshu** 2 years, 6 months ago

automatically taking care of workers up to 3 (as the maximum worker is 3 set as per questions)

upvoted 1 times

✉  **atnafu2020** 3 years, 5 months ago

AB

is correct

upvoted 2 times

✉  **haroldbenites** 3 years, 5 months ago

A , E is correct

upvoted 5 times

Question #137

Topic 1

You have a data pipeline with a Dataflow job that aggregates and writes time series metrics to Bigtable. You notice that data is slow to update in Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Dataflow pipeline to use local execution
- B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions
- C. Increase the number of nodes in the Bigtable cluster
- D. Modify your Dataflow pipeline to use the Flatten transform before writing to Bigtable
- E. Modify your Dataflow pipeline to use the CoGroupByKey transform before writing to Bigtable

**Correct Answer:** *DE*

Reference:

<https://cloud.google.com/bigtable/docs/performance#performance-write-throughput>

Another consideration in capacity planning is storage. The storage capacity of a cluster is determined by the storage type and the number of nodes in the cluster. When the amount of data stored in a cluster increases, Bigtable optimizes the storage by **distributing the amount of data** across all the nodes in the cluster.

You can determine the storage usage per node by dividing the cluster's **storage utilization (bytes)** by the number of nodes in the cluster. For example, consider a cluster that has three HDD nodes and 9 TB of data. Each node stores about 3 TB, which is 18.75% of the HDD storage per node limit of 16 TB.

<https://cloud.google.com/dataflow/docs/guides/specifying-exec-params#setting-other-cloud-pipeline-options>

*Community vote distribution*

BC (89%)

4%

✉  **arpitagrawal** Highly Voted 1 year, 4 months ago

**Selected Answer: BC**

It should be B and C

upvoted 8 times

ducc Highly Voted 1 year, 4 months ago

Selected Answer: BC

BC is correct

Why the comments is deleted?

upvoted 7 times

emmylou Most Recent 2 months, 2 weeks ago

The "Correct Answers" are just put in with a random generator :-) B and C

upvoted 2 times

BlehMaks 3 months ago

Selected Answer: BC

B - opportunity to parallelise the process

C - increase throughput

upvoted 3 times

Bahubali1988 4 months ago

Exactly opposite answers in the discussions

upvoted 1 times

barnacles 4 months, 1 week ago

Selected Answer: BC

B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions:

Increasing the number of Dataflow workers can help parallelize the processing of your data, which can result in faster data updates to Bigtable and improved concurrency. You can set maxNumWorkers to a higher value to achieve this.

C. Increase the number of nodes in the Bigtable cluster:

Increasing the number of nodes in your Bigtable cluster can improve the overall throughput and reduce latency when writing data. It allows Bigtable to handle a higher rate of data ingestion and queries, which is essential for supporting additional concurrent users.

upvoted 4 times

ckanaar 4 months, 1 week ago

Selected Answer: CD

C definitely is correct, as it improves the read and write performance of Bigtable.

However, I do think that the second option is actually D instead of B, because the question specifically states that the pipeline aggregates data. Flatten merges multiple PCollection objects into a single logical PCollection, allowing for faster aggregation of time series data.

upvoted 1 times

NewDE2023 5 months, 4 weeks ago

Selected Answer: BE

B - I believe it is consensus.

D - The question mentions "a Dataflow job that "aggregates" and writes time series metrics to Bigtable". So CoGroupByKey performs a shuffle (grouping) operation to distribute data across workers.

<https://cloud.google.com/dataflow/docs/guides/develop-and-test-pipelines>

upvoted 1 times

WillemHendr 7 months, 3 weeks ago

Selected Answer: DE

I read this question as: BigTable Write operations are all over the place (key-wise), and BigTable doesn't like that. When creating groups (batch writes), of similar keys (close to each other), BigTable is happy again, which I loosely translate into DE.

upvoted 1 times

vaga1 9 months ago

B is correct. But I don't see how you increase the write throughput of Bigtable increasing its cluster size. It should be dataflow instance resources that have to be increased

upvoted 1 times

juliobs 10 months, 2 weeks ago

Selected Answer: BC

BC make sense

upvoted 1 times

✉️  **NamitSehgal** 1 year ago

BC only makes sense here , no mention of data, no mention of keeping cost low  
upvoted 1 times

✉️  **AzureDP900** 1 year ago

B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions  
C. Increase the number of nodes in the Bigtable cluster  
upvoted 1 times

✉️  **ovokpus** 1 year, 2 months ago

**Selected Answer: BC**

Increase max num of workers increases pipeline performance in Dataflow  
Increase number of nodes in Bigtable increases write throughput  
upvoted 2 times

Question #138

Topic 1

You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

- A. Create a Cloud Dataproc Workflow Template
- B. Create an initialization action to execute the jobs
- C. Create a Directed Acyclic Graph in Cloud Composer
- D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

**Correct Answer: C**

Reference:

<https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows>

You set up and run a workflow by:

1. Creating a workflow template
2. Configuring a managed (ephemeral) cluster or selecting an existing cluster
3. Adding jobs
4. Instantiating the template to run the workflow

*Community vote distribution*

C (79%)

A (21%)

✉️  **MaxNRG** 1 month, 1 week ago

**Selected Answer: C**

The best option for automating your scheduled Spark jobs on Cloud Dataproc, considering sequential and concurrent execution, is:  
C. Create a Directed Acyclic Graph (DAG) in Cloud Composer.

upvoted 2 times

✉️  **MaxNRG** 1 month, 1 week ago

Here's why:

DAG workflows: Cloud Composer excels at orchestrating complex workflows with dependencies, making it ideal for managing sequential and concurrent execution of your Spark jobs. You can define dependencies between tasks to ensure certain jobs only run after others finish.

Automation: Cloud Composer lets you schedule workflows to run automatically based on triggers like time intervals or data availability, eliminating the need for manual intervention.

Integration: Cloud Composer integrates seamlessly with Cloud Dataproc, allowing you to easily launch and manage your Spark clusters within the workflow.

Scalability: Cloud Composer scales well to handle a large number of jobs and workflows, making it suitable for managing complex data pipelines.

upvoted 2 times

✉️  **MaxNRG** 1 month, 1 week ago

While the other options have some merit, they fall short in certain aspects:

A. Cloud Dataproc Workflow Templates: While workflow templates can automate job submission on a cluster, they lack the ability to define dependencies and coordinate concurrent execution effectively.

B. Initialization action: An initialization action can only run a single script before a Dataproc cluster starts, not suitable for orchestrating multiple scheduled jobs with dependencies.

D. Bash script: A Bash script might work for simple cases, but it can be cumbersome to manage and lacks the advanced scheduling and error handling capabilities of Cloud Composer.

Therefore, utilizing a Cloud Composer DAG offers the most comprehensive and flexible solution for automating your scheduled Spark jobs with sequential and concurrent execution on Cloud Dataproc.

upvoted 2 times

✉️  **emmylou** 2 months, 1 week ago

**Selected Answer: C**

I thought it might be A but the templates can only run sequentially, not concurrently.

upvoted 1 times

✉️  **barnacles** 4 months, 1 week ago

**Selected Answer: C**

Directed Acyclic Graph (DAG): Cloud Composer (formerly known as Cloud Composer) is a managed Apache Airflow service that allows you to create and manage workflows as DAGs. You can define a DAG that includes tasks for running Spark jobs in sequence or concurrently.

Scheduling: Cloud Composer provides built-in scheduling capabilities, allowing you to specify when and how often your DAGs should run. You can schedule the execution of your Spark jobs at specific times or intervals.

Dependency Management: In a DAG, you can define dependencies between tasks. This means you can set up tasks to run sequentially or concurrently based on your requirements. For example, you can specify that Job B runs after Job A has completed, or you can schedule jobs to run concurrently when there are no dependencies.

upvoted 1 times

✉️  **midgoo** 10 months, 3 weeks ago

**Selected Answer: C**

I would choose A if there was one more step to schedule the Template. It is like creating DAG without running it in Airflow. So only option C is correct here.

upvoted 2 times

✉️  **AzureDP900** 1 year ago

C. Create a Directed Acyclic Graph in Cloud Composer

upvoted 3 times

✉️  **saurabh Singh4k** 1 year, 1 month ago

**Selected Answer: A**

Why go for an expensive Composer when you only have to schedule and create a DAG for Dataproc, A is sufficient.

upvoted 2 times

✉️  **captainbu** 1 year ago

I've would've gone for Workflow Templates as well. But those are lacking the scheduling capability. Hence you would need to use Cloud Composer (or Cloud Functions or Cloud Scheduler) anyway. Hence C seems to be the better solution.

Pls see here:

<https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions>

upvoted 5 times

👤 **zellick** 1 year, 1 month ago

**Selected Answer: C**

C is the answer.

[https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions#cloud\\_composer](https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions#cloud_composer)

Cloud Composer is a managed Apache Airflow service you can use to create, schedule, monitor, and manage workflows. Advantages:

- Supports time- and event-based scheduling
- Simplified calls to Dataproc using Operators
- Dynamically generate workflows and workflow parameters
- Build data flows that span multiple Google Cloud products

upvoted 2 times

👤 **deavid** 1 year, 3 months ago

**Selected Answer: C**

C.

Composer fits better to schedule Dataproc Workflows, check the documentation:

<https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions>

Also A is not enough. Dataproc Workflow Template itself don't has a native schedule option.

upvoted 4 times

👤 **louisgcde** 1 year, 3 months ago

**Selected Answer: C**

So that I thing the answer should be C (Composer).

upvoted 1 times

👤 **louisgcde** 1 year, 3 months ago

To me, the point is "automate" the process, so that Composer DAG is needed and can be used with Dataproc Workflow Template.

upvoted 2 times

👤 **dmzr** 1 year, 3 months ago

**Selected Answer: A**

Ans A makes more sense, since a question is regarding Dataproc jobs only

upvoted 2 times

👤 **LP\_PDE** 1 year, 3 months ago

Correct answer is A. <https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows>

upvoted 3 times

👤 **HarshKothari21** 1 year, 4 months ago

**Selected Answer: C**

Option c

upvoted 1 times

👤 **ducc** 1 year, 4 months ago

**Selected Answer: C**

You have streaming and batch job, so Composer is the choice for me

upvoted 1 times

Question #139

Topic 1

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

A. Create an API using App Engine to receive and send messages to the applications

B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them

- C. Create a table on Cloud SQL, and insert and delete rows with the job information
- D. Create a table on Cloud Spanner, and insert and delete rows with the job information

**Correct Answer: A**

Reference:

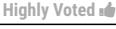
<https://cloud.google.com/appengine/docs/standard/go/mail/sending-receiving-with-mail-api>

Use the Mail API to send and receive mail.

For information on who can send mail and guidelines for sending bulk mail, see the [Mail API Overview](#).

*Community vote distribution*

B (100%)

✉  **jkhong**  1 year, 1 month ago

**Selected Answer: B**

Job generators (they would be the publishers).  
Job runners = subscribers

Question mentions that it must scale (of which push subscription has automatic scaling) and can accommodate additional new applications (t can be solved by having multiple subscriptions, with each relating to a unique application) to a central topic  
upvoted 9 times

✉  **AzureDP900** 1 year ago

Yes it is  
B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them  
upvoted 4 times

✉  **juliorevk**  4 months, 1 week ago

**Selected Answer: B**

B to decouple jobs being generated and run. Pub/Sub also scales seamlessly  
upvoted 1 times

✉  **barnacles** 4 months, 1 week ago

**Selected Answer: B**

B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them.

Scalability: Cloud Pub/Sub is a highly scalable messaging service that can handle a significant volume of messages and subscribers. It can easily accommodate increases in usage as your data pipeline scales.

Question #140

Topic 1

You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?

- A. The current epoch time
- B. A concatenation of the product name and the current epoch time
- C. A random universally unique identifier number (version 4 UUID)
- D. The original order identification number from the sales system, which is a monotonically increasing integer

**Correct Answer: C**

Reference:

<https://www.uuidgenerator.net/version4>

[UUID Generator](#) [Version 4 UUID](#) [Version 1 UUID](#) [Nil/Empty UUID](#) [GUID Generator](#) [Developer's Corner](#)

# Online UUID Generator

Your Version 4 UUID:

**862c3f50-6b77-4402-8276-c5cf877d3136**

 Copy

[Refresh](#) page to generate another.

Community vote distribution

C (92%)

8%

✉️  **Remi2021**  1 year, 4 months ago

**Selected Answer: C**

According to the documentation:

Use a Universally Unique Identifier (UUID)

You can use a Universally Unique Identifier (UUID) as defined by RFC 4122 as the primary key. Version 4 UUID is recommended, because it uses random values in the bit sequence. Version 1 UUID stores the timestamp in the high order bits and is not recommended.

<https://cloud.google.com/spanner/docs/schema-design>

upvoted 9 times

✉️  **AzureDP900** 1 year ago

Agree with C

upvoted 1 times

✉️  **barnacles**  4 months, 1 week ago

**Selected Answer: C**

For a transaction table in Cloud Spanner that stores product sales data, from a performance perspective, it is generally recommended to choose a primary key that allows for even distribution of data across nodes and minimizes hotspots. Therefore, option C, which suggests using a random universally unique identifier number (version 4 UUID), is the preferred choice.

upvoted 4 times

✉️  **arien\_chen** 5 months, 1 week ago

**Selected Answer: C**

For a RDB I would choice D.

But for Google Spanner, Google says:

<https://cloud.google.com/spanner/docs/schema-and-data-model#:~:text=monotonically%20increasing%20integer>

upvoted 1 times

✉️  **vaga1** 8 months, 3 weeks ago

**Selected Answer: C**

B might work if you say timestamp instead than epoch. PK of sales should contain the exact purchase date or timestamp, not the time when the transaction was processed. I personally associate the term epoch in this context to the process timestamp instead than the purchase timestamp.

upvoted 2 times

✉️  **midgoo** 10 months, 3 weeks ago

**Selected Answer: C**

B may cause error if same product ID came at the same time (same id + same epoch)

So C is the correct answer here

upvoted 1 times

✉️  **jkhong** 1 year, 1 month ago

**Selected Answer: C**

A and D are invalid because they monotonically increases.

B would work, but in terms of pure performance UUID 4 is the fastest because it virtually will not cause hotspots

upvoted 2 times

✉️  **odacir** 1 year, 1 month ago

**Selected Answer: C**

A and D are not valid, because they monotonically increase.

C avoid hotspots for sure, but it's not related with queries. So for writing performance it's perfect that the reason for choosing this: "You need to create a new transaction table in Cloud Spanner that stores product sales data". They only ask you to store product data, it's a writing ops. If the question had spoken about query the info or hard performance read, the best option would be B, because it has the balance of writing/reading best practices.

There are a few disadvantages to using a UUID:

They are slightly large, using 16 bytes or more. Other options for primary keys don't use this much storage.

They carry no information about the record. For example, a primary key of SingerId and AlbumId has an inherent meaning, while a UUID does not.

You lose locality between records that are related, which is why using a UUID eliminates hotspots.

[https://cloud.google.com/spanner/docs/schema-design#uuid\\_primary\\_key](https://cloud.google.com/spanner/docs/schema-design#uuid_primary_key)

upvoted 2 times

👤 **YorelNation** 1 year, 4 months ago

**Selected Answer: C**

C. A random universally unique identifier number (version 4 UUID)

From <https://cloud.google.com/spanner/docs/schema-and-data-model>

There are techniques that can spread the load across multiple servers and avoid hotspots:

Hash the key and store it in a column. Use the hash column (or the hash column and the unique key columns together) as the primary key. Swap the order of the columns in the primary key.

Use a Universally Unique Identifier (UUID). Version 4 UUID is recommended, because it uses random values in the high-order bits. Don't use a UUID algorithm (such as version 1 UUID) that stores the timestamp in the high order bits.

Bit-reverse sequential values.

upvoted 1 times

👤 **jsree236** 1 year, 4 months ago

**Selected Answer: B**

Answer should be B as in all the other options hotspotting is possible. According to proper schema design guideline..

Schema design best practice #1: Do not choose a column whose value monotonically increases or decreases as the first key part for a high write table.

Supporting link:

<https://cloud.google.com/spanner/docs/schema-design#primary-key-prevent-hotspots>

upvoted 2 times

Question #141

Topic 1

Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

- A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.
- B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.
- C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.
- D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.

**Correct Answer: D**

### Community vote distribution

D (100%)

✉️  **SteelWarrior** Highly Voted 3 years, 4 months ago

Answer D is correct. Aggregated log sink will create a single sink for all projects, the destination can be a google cloud storage, pub/sub topic bigquery table or a cloud logging bucket. without aggregated sink this will be required to be done for each project individually which will be cumbersome.

[https://cloud.google.com/logging/docs/export/aggregated\\_sinks](https://cloud.google.com/logging/docs/export/aggregated_sinks)

upvoted 29 times

✉️  **AzureDP900** 1 year ago

D is right

upvoted 1 times

✉️  **[Removed]** Highly Voted 3 years, 10 months ago

Correct: D

[https://cloud.google.com/iam/docs/roles-audit-logging#scenario\\_external\\_auditors](https://cloud.google.com/iam/docs/roles-audit-logging#scenario_external_auditors)

upvoted 12 times

✉️  **Rajuuu** 3 years, 6 months ago

The above link shows BigQuery as a sink for aggregated exports and not Cloud Storage.

upvoted 3 times

✉️  **daghayeghi** 2 years, 10 months ago

[https://cloud.google.com/iam/docs/job-functions/auditing#scenario\\_operational\\_monitoring](https://cloud.google.com/iam/docs/job-functions/auditing#scenario_operational_monitoring)

upvoted 3 times

✉️  **barnacles** Most Recent 4 months, 1 week ago

**Selected Answer: D**

D.

Here's why this option is recommended:

Aggregated Export Sink: By using an aggregated export sink, you can consolidate data access logs from multiple projects into a single location. This simplifies log management and retention policies.

Newly Created Project for Audit Logs: Creating a dedicated project for audit logs allows you to centralize access control and manage logs separately from individual Data Analyst projects.

Access Restriction: By restricting access to the project containing the exported logs, you ensure that only authorized audit personnel have access to the logs while preventing Data Analysts from accessing them.

upvoted 1 times

✉️  **midgoo** 10 months, 3 weeks ago

**Selected Answer: D**

To create the Log Router, at step 3 to define the logs (Source), we can include logs from many projects (aggregated)

upvoted 1 times

✉️  **zellick** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

[https://cloud.google.com/logging/docs/export/aggregated\\_sinks](https://cloud.google.com/logging/docs/export/aggregated_sinks)

Aggregated sinks combine and route log entries from the Google Cloud resources contained by an organization or folder. For instance, you can aggregate and route audit log entries from all the folders contained by an organization to a Cloud Storage bucket.

upvoted 2 times

✉️  **dfffff** 1 year, 9 months ago

D is correct

upvoted 1 times

✉️  **MaxNRG** 2 years ago

**Selected Answer: D**

D: [https://cloud.google.com/logging/docs/export/aggregated\\_exports](https://cloud.google.com/logging/docs/export/aggregated_exports)

You can create an aggregated export sink that can export log entries from all the projects, folders, and billing accounts of an organization. As an example, you might use this feature to export audit log entries from an organization's projects to a central location.

upvoted 4 times

⊕  **Chelseajcole** 2 years, 4 months ago

The auditor needs to audit data analyst's behaviors (how they access multiple projects in BQ). So, the key is, multiple projects. According to Google doc project-level sinks:

[https://cloud.google.com/logging/docs/export/configure\\_export\\_v2](https://cloud.google.com/logging/docs/export/configure_export_v2)

However, the Cloud Console can only create or view sinks in Cloud projects. To create sinks in organizations, folders, or billing accounts using the gcloud command-line tool or Cloud Logging API, see Aggregated sinks.

Obviously, the auditor needs to check all projects accessed by data analyst which is not project-level, a higher level like folder or organization level, this can only be done via the aggregate sink.

So D is the answer.

upvoted 4 times

⊕  **sumanshu** 2 years, 6 months ago

A - eliminated, because logs needs to be retained for 6 months (So, some storage require)

B - eliminated, because if we store in same project then, Data Analyst can also access (But in question it's mention, ONLY audit personnel have access)

C - Wrong (No need to restrict project as well as logs separately) - wording does not look okay.

D - Correct (If we restrict the project, then all resources get restricted)

Vote for D

upvoted 6 times

⊕  **sumanshu** 2 years, 6 months ago

Option 'C' - I guess said - restrict access to the project with the exported logs. (i.e. restrict access of that project from where we took logs)  
I am not wrong... Thus it's INCORRECT

upvoted 2 times

⊕  **at99** 2 years ago

Sinks are different from Aggregate Sinks, refer [https://cloud.google.com/logging/docs/export/configure\\_export\\_v2#api](https://cloud.google.com/logging/docs/export/configure_export_v2#api)

upvoted 1 times

⊕  **septiandy** 2 years, 9 months ago

what is the difference between C and D? I think it's same.

upvoted 3 times

⊕  **FP77** 5 months, 2 weeks ago

I think the key difference is that D talks about aggregated sinks.

upvoted 1 times

⊕  **haroldbenites** 3 years, 5 months ago

D is correct

upvoted 3 times

⊕  **saurabh1805** 3 years, 5 months ago

D is correct answer, refer below link for more information.

upvoted 3 times

⊕  **VishalB** 3 years, 6 months ago

Ans : D

Aggregated Exports, which allows you to set up a sink at the Cloud IAM organization or folder level, and export logs from all the projects inside the organization or folder.

upvoted 5 times

⊕  **[Removed]** 3 years, 10 months ago

Answer D

upvoted 4 times

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects.

What should you do?

- A. Create a Cloud Monitoring dashboard based on the BigQuery metric query/scanned\_bytes
- B. Create a Cloud Monitoring dashboard based on the BigQuery metric slots/allocated\_for\_project
- C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Cloud Monitoring dashboard based on the custom metric
- D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Cloud Monitoring dashboard based on the custom metric

**Correct Answer: B**

*Community vote distribution*

B (86%)

14%

✉️  **MaxNRG** 1 month, 1 week ago

**Selected Answer: B**

Viewing project and reservation slot usage in Stackdriver Monitoring

Information is available from the "Slots Allocated" metric in Stackdriver Monitoring. This metric information includes a per-reservation and per breakdown of slot usage. The information can also be visualized by using the custom charts metric explorer.

<https://cloud.google.com/bigquery/docs/reservations-monitoring>

[https://cloud.google.com/monitoring/api/metrics\\_gcp](https://cloud.google.com/monitoring/api/metrics_gcp)

upvoted 1 times

✉️  **barnac1es** 4 months, 1 week ago

**Selected Answer: B**

The slots/allocated\_for\_project metric provides information about the number of slots allocated to each project. It directly reflects the slot usage, making it a relevant and accurate metric for monitoring slot allocation within each project.

Options A, C, and D involve log exports and custom metrics, but they may not be as straightforward or provide the same level of detail as the built-in metric slots/allocated\_for\_project:

upvoted 3 times

✉️  **ckanaar** 4 months, 1 week ago

The naming is quite misleading in this case, but it actually seems from the documentation that slots/allocated\_for\_project indicates the "slots used by project", in which case answer B is correct:

[https://cloud.google.com/monitoring/api/metrics\\_gcp#:~:text=slots/allocated\\_for\\_project%20GA%0ASlots%20used%20by%20project](https://cloud.google.com/monitoring/api/metrics_gcp#:~:text=slots/allocated_for_project%20GA%0ASlots%20used%20by%20project)

upvoted 2 times

□  **arien\_chen** 5 months, 1 week ago

**Selected Answer: D**

B slots/allocated\_for\_project will give you the total number of slots allocated to each project, but it will not tell you how many slots are actually being used.

The purpose to monitor 'slot usgae' is for billing. 'slot/allocated' means nothing.  
Option D is better than B.

And, the question mention 'Each analytics team in organization', so it should be 'organization level'.

upvoted 1 times

□  **midgoo** 10 months, 1 week ago

**Selected Answer: D**

If 'usage' = how the slots are being used, D is the corret answer  
If 'usage' = how the slots are being allocated, B is the correct answer

I think in this question, usage = how the slots are being used

upvoted 1 times

□  **musumusu** 11 months, 2 weeks ago

Answer B,

Why not D, aggregated log export is good but it will generate all the details which is large in size and costly too. you dont need all the information. It can break data privacy. so look for B because this much is asked only. Normally, i make such errors alot.

upvoted 1 times

□  **saurabhsingh4k** 1 year, 1 month ago

**Selected Answer: B**

The correct answer is B. You should create a Cloud Monitoring dashboard based on the BigQuery metric slots/allocated\_for\_project.

This metric represents the number of BigQuery slots allocated for a project. By creating a Cloud Monitoring dashboard based on this metric, you can monitor the slot usage within each project in your organization. This will allow each team to monitor their own slot usage and ensure that they are not exceeding their allocated quota.

Option A is incorrect because the query/scanned\_bytes metric represents the number of bytes scanned by BigQuery queries, not the slot usage.

Option C is incorrect because it involves creating a log export for each project and using a custom metric based on the totalSlotMs field. While this may be a valid way to monitor slot usage, it is more complex than simply using the slots/allocated\_for\_project metric.

Option D is also incorrect because it involves creating an aggregated log export at the organization level, which is not necessary for monitoring slot usage within individual projects.

upvoted 4 times

□  **dn\_mohammed\_data** 1 year, 4 months ago

vote for B

upvoted 2 times

□  **John\_Pongthorn** 1 year, 4 months ago

B ,the another is related to the question as well.

<https://cloud.google.com/bigquery/docs/reservations-monitoring#viewing-slot-usage>

upvoted 4 times

□  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

B the below is related to the question.

Question #143

Topic 1

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

- A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
- B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
- C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code

D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

**Correct Answer: A**

Reference:

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

## Launching your replacement job

To update your job, you'll need to launch a new job to replace the ongoing job. When you launch your replacement job, you'll need to set the following pipeline options to perform the update process in addition to the job's regular options:

Java    Python

- Pass the `--update` option.
- Set the `--jobName` option in `PipelineOptions` to the same name as the job you want to update.
- Set the `--region` option as the region of the job that you want to update.
- If any transform names in your pipeline have changed, you must supply a [transform mapping](#) and pass it using the `--transformNameMapping` option.

*Community vote distribution*

D (79%)

A (21%)

✉️  **odacir** Highly Voted 1 year, 1 month ago

**Selected Answer: D**

It's D. → Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. New version is major changes. Stop and drain and then launch the new code is a lot safer way. We recommend that you attempt only smaller changes to your pipeline's windowing, such as changing the duration of fixed- or sliding-time windows. Making major changes to windowing or triggers, like changing the windowing algorithm, might have unpredictable results on your pipeline output.

[https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing\\_windowing](https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing)  
upvoted 13 times

✉️  **maggieee** 1 year, 1 month ago

Since updating the job as in A does a compatibility check, wouldn't you want to try that first? Then if the compatibility check fails then you proceed to drain current pipeline and then launch new pipeline (Answer D)?

As in A would be correct answer, then if compatibility check fails, you proceed to D.

<https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#CCheck>  
upvoted 2 times

✉️  **ckanaar** 4 months, 1 week ago

You're right in your reasoning, but since the documentation specifically uses this example for stopping and draining, it's safe to assume that the compatibility check will always fail with these adjustments. Therefore, we can go straight to D.

Furthermore, answer A doesn't state: "Update the Cloud Dataflow pipeline inflight by passing the `--update` option with the `--jobName` to the existing name, if the compatibility check fails, THEN proceed to stopping the pipeline with the drain option", so in itself it is not the right answer if the check fails.

upvoted 1 times

✉️  **patitonav** Most Recent 1 month ago

**Selected Answer: D**

D seems the right way to go  
upvoted 1 times

⊕  **TVH\_Data\_Engineer** 1 month, 1 week ago

**Selected Answer: D**

Option A is the first approach to try, as it allows for an in-flight update with minimal disruption. However, if the changes in the new version of the pipeline are not compatible with an in-flight update (due to significant changes in windowing or triggering), then option D should be used. The Drain option ensures a graceful shutdown of the existing pipeline, reducing the risk of data loss, and then a new job can be started with the updated code.

upvoted 1 times

⊕  **MaxNRG** 1 month, 1 week ago

**Selected Answer: D**

A is not an option as "You want to ensure that no data is lost during the update. ":

Making major changes to windowing or triggers, like changing the windowing algorithm, might have unpredictable results on your pipeline output.

[https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#change\\_windowing](https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#change_windowing)

upvoted 1 times

⊕  **barnacles** 4 months, 1 week ago

**Selected Answer: D**

Drain Option: The "Drain" option allows the existing Dataflow job to complete processing of any in-flight data before stopping the job. This ensures that no data is lost during the transition to the new version.

Create a New Job: After draining the existing job, you create a new Cloud Dataflow job with the updated code. This new job starts fresh and continues processing data from where the old job left off.

Option A (updating the inflight pipeline with the --update option) may not guarantee no data loss, as the update could disrupt the existing job's operation and potentially cause data loss.

Option B (updating the inflight pipeline with the --update option and a new job name) is similar to option A and may not provide data loss guarantees.

Option C (stopping the pipeline with the Cancel option and creating a new job) will abruptly stop the existing job without draining, potentially leading to data loss.

upvoted 1 times

⊕  **knith66** 6 months, 1 week ago

Look D after seeing some docs. please check the below link <https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline>

upvoted 1 times

⊕  **vamgcp** 6 months, 1 week ago

**Selected Answer: D**

I will go with option D - If you want to minimize the impact of the update, then option A is the best option. However, if you are not concerned about a temporary interruption in processing, then option D is also a valid option. Option Pros Cons

A Does not stop the pipeline, so no data is lost. Requires you to create a new version of the pipeline.

B Creates a new job with the updated code, so you do not have to update the running pipeline. Can lead to data loss if the new job does not process all of the data that was in the running pipeline.

C Stops the pipeline and drains any data that is currently in flight, so no data is lost. Causes a temporary interruption in processing.

upvoted 1 times

⊕  **midgoo** 10 months, 3 weeks ago

**Selected Answer: D**

A is not recommended for major changes in pipeline.

upvoted 3 times

⊕  **musumusu** 11 months, 2 weeks ago

Answer A:

``gcloud dataflow jobs update <JOB\_ID> --update <GCS\_PATH\_TO\_UPDATED\_PIPELINE> --region <REGION>``  
--update flag does not miss any data and you can execute this command even your pipeline is running. Its safe and fast, you can continuously make some changes and update this command. no problem.

Stop and Drain, is required when you want to test the pipeline and stop it without losing the data.

upvoted 1 times

⊕  **musumusu** 11 months, 1 week ago

Answer D: as per latest documents 02/2023 google has removed update flag.

upvoted 3 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: D**

agree with odacir

upvoted 4 times

✉  **hauhau** 1 year, 1 month ago

**Selected Answer: A**

vote A

D: drain doesn't mention about update dataflow job just stop and preserve data

A: replace existing job and preserve data

(When you update your job, the Dataflow service performs a compatibility check between your currently-running job and your potential replacement job. The compatibility check ensures that things like intermediate state information and buffered data can be transferred from your prior job to your replacement job.)

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

upvoted 2 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#Launching>

To update your job, launch a new job to replace the ongoing job. When you launch your replacement job, set the following pipeline options to perform the update process in addition to the job's regular options:

- Pass the --update option.
- Set the --jobName option in PipelineOptions to the same name as the job you want to update.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

Are mayor changes. It's not safe to update. I vote D.

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

D

A-is not because The Dataflow service retains the job name, but runs the replacement job with an updated Job ID.

Description:

When you update a job on the Dataflow service, you replace the existing job with a new job that runs your updated pipeline code. The Dataflow service retains the job name, but runs the replacement job with an updated Job ID. This process can cause downtime while the existing job stops, the compatibility check runs, and the new job starts.'

<https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python:~:text=When%20you%20update%20a,has%20the%20following%20transforms%3A>

D is correct

Drain ->clone -> update -> run

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

Changed my mind to A

[https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python\\_2:~:text=Set%20the%20job\\_name,%2D%2Dtransform\\_name\\_mapping%20option](https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python_2:~:text=Set%20the%20job_name,%2D%2Dtransform_name_mapping%20option).

upvoted 1 times

✉  **drunk\_goat82** 1 year, 2 months ago

**Selected Answer: D**

Changing windowing algorithm may break the pipeline.

[https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing\\_windowing](https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing)

upvoted 3 times

✉  **ovokpus** 1 year, 2 months ago

**Selected Answer: A**

No, do not drain the current job.

upvoted 1 times

✉  **dish11dish** 1 year, 2 months ago

**Selected Answer: D**

in this scenario pipeline is streaming pipeline with windowing algorithm and triggering strategy changes to new one without loss of data,so better go with Drain option as it fullfile all precondition described in scenario which is :-

- 1.streaming
- 2.code changes with windowing algorithm and triggering strategy to new way
- 3.no loss of data during update

References:-

<https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline#drain>

Drain a job. This method applies only to streaming pipelines. Draining a job enables the Dataflow service to finish processing the buffered data while simultaneously ceasing the ingestion of new data. For more information, see Draining a job.

upvoted 1 times

✉  **dish11dish** 1 year, 2 months ago

If the pipeline was batch then ans would been A

upvoted 1 times

✉  **Mccloudgirl** 1 year, 2 months ago

D: They want to preserve data and updates might not be predictable.

[https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing\\_windowing](https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing)

upvoted 3 times

Question #144

Topic 1

You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

- A. Use Transfer Appliance to copy the data to Cloud Storage
- B. Use gsutil cp --J to compress the content being uploaded to Cloud Storage
- C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
- D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

**Correct Answer: A**

*Community vote distribution*

A (100%)

✉  **[Removed]**  3 years, 10 months ago

Answer: A

Description: Huge amount of data with low network bandwidth, Transfer Appliance is best for moving data over 100TB

upvoted 26 times

👤 [Removed] Highly Voted 3 years, 10 months ago

Correct - A

upvoted 10 times

👤 patitonav Most Recent 1 month ago

Selected Answer: A

A . Easy, just by the amount of data

upvoted 1 times

👤 Nirca 3 months, 3 weeks ago

Selected Answer: A

In 6 months - only 0.0290304 Petabytes will be uploaded. Right - compression might help, but we do not have any info to support the ration. go for A

upvoted 1 times

👤 barnacles 4 months, 1 week ago

Selected Answer: A

Physical Transfer: Transfer Appliance is a physical device provided by Google Cloud that you can use to physically transfer large volumes of data to the cloud. It allows you to avoid the limitations of network bandwidth and transfer data much faster.

Capacity: Transfer Appliance can handle large volumes of data, including the 2 PB you need to migrate, without the constraints of slow network speeds.

Efficiency: It is highly efficient for large-scale data transfers and is a practical choice for transferring multi-terabyte or petabyte-scale datasets

upvoted 1 times

👤 arien\_chen 5 months, 1 week ago

it would take 34 years.

Option A no doubt.

<https://cloud.google.com/static/architecture/images/big-data-transfer-how-to-get-started-transfer-size-and-speed.png>

upvoted 1 times

👤 vaga1 8 months, 3 weeks ago

Selected Answer: A

2,000,000,000,000 bytes = 2 Petabytes

20,000,000 bytes = 20 Megabytes

Once we do the math:

2 Petabytes / 20 Megabytes = 100,000,000 seconds forecasted to migrate the data.

100,000,000 seconds =

1,666,666.7 minutes =

27,777.8 hours =

1,157.4 days

6 months = 180 days

1,157.4 days > 180 days

Still, with such amount Transfer Appliance is recommended.

upvoted 1 times

👤 musumusu 11 months, 2 weeks ago

Transfer alliance is a physical device of size like cabin luggage or slightly larger.

It has Seagate/WD harddisk (these are name of companies) size varries from 100 to 480 TB.

In our case 2PB (2000 TB) accordingly. Google send you this and transfer data into it by wire connection and then upload data from this to Google and empty the appliance.

upvoted 1 times

👤 AzureDP900 1 year, 1 month ago

This is no brainer question, A is right

upvoted 2 times

👤 **zelick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/transfer-appliance/docs/4.0/overview>

Transfer Appliance is a high-capacity storage device that enables you to transfer and securely ship your data to a Google upload facility, where we upload your data to Cloud Storage.

upvoted 2 times

👤 **jkhong** 1 year, 3 months ago

**Selected Answer: A**

Problem: Transferring 2 peta data to Cloud Storage

Considerations: Bad network speed

Bad network = cannot initiate from client's end through network. So, B, C is out  
D will still be super slow. At this speed it will take 27,777 hours to transfer the data

upvoted 3 times

👤 **sumanshu** 2 years, 6 months ago

Vote for A

A - Correct , Transfer Appliance for moving offline data, large data sets, or data from a source with limited bandwidth  
<https://cloud.google.com/storage-transfer/docs/overview>

B - Eliminated (Not recommended for large storage). recommended for < 1TB

C - Its ONLINE, but we have bandwidth issue - So eliminated.

D - Eliminated (Not recommended for large storage). recommended for < 1TB

upvoted 9 times

👤 **SteelWarrior** 3 years, 4 months ago

Correct answer is A. with little calculation we know the kind of data will require approx 19 months to transfer on 20Mbps bandwidth. Also, google recommends Transfer appliance for petabytes of data.

upvoted 3 times

👤 **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 3 times

👤 **Rajuuu** 3 years, 6 months ago

Correct - A

upvoted 3 times

👤 **Rajokkiyam** 3 years, 10 months ago

Answer A

upvoted 3 times

Question #145

Topic 1

You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:

- ⇒ Executing the transformations on a schedule
- ⇒ Enabling non-developer analysts to modify transformations
- ⇒ Providing a graphical tool for designing transformations

What should you do?

- A. Use Dataprep by Trifacta to build and maintain the transformation recipes, and execute them on a scheduled basis

- B. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query
- C. Help the analysts write a Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
- D. Use Apache Spark on Dataproc to infer the schema of the CSV file before creating a Dataframe. Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

**Correct Answer: D**

*Community vote distribution*

A (100%)

 **madhu1171** Highly Voted 3 years, 10 months ago

A should be the answer  
upvoted 34 times

 **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A  
Description: Dataprep is used by non developers  
upvoted 17 times

 **barnac1es** Most Recent 4 months, 1 week ago

**Selected Answer: A**

Scheduled Transformations: Dataprep by Trifacta allows you to design and schedule transformation recipes to process data on a regular basis. You can automate the data cleansing process by scheduling it to run monthly.

User-Friendly Interface: Dataprep provides a user-friendly graphical interface that enables non-developer analysts to design, modify, and maintain transformation recipes without writing code. This empowers analysts to work with the data effectively.

Transformation Flexibility: Dataprep supports flexible data transformations, making it suitable for scenarios where the schema of the incoming data changes. Analysts can adapt the transformations to new schemas using the visual tools provided by Dataprep.

upvoted 1 times

 **vaga1** 8 months, 3 weeks ago

**Selected Answer: A**

Providing a graphical tool for designing transformations is enough for A  
upvoted 4 times

 **Dhruv28** 11 months, 1 week ago

Your company receives a lot of financial data in CSV files. The files need to be processed, cleaned and transformed before they are made available for analytics. The schema of the data also changes every third month. The Data analysts should be able to perform the tasks

1. No prior knowledge of any language with no coding
2. Provided a GUI tool to build and modify the schema

What solution best fits the need?

upvoted 1 times

 **zellick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

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

Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning. Because Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

upvoted 3 times

 **arpitagrawal** 1 year, 4 months ago

**Selected Answer: A**

non-developer analysts

upvoted 2 times

✉  **devdimidved** 1 year, 8 months ago

**Selected Answer: A**

Dataprep is for non developers

upvoted 1 times

✉  **amitsingla012** 1 year, 8 months ago

**Selected Answer: A**

Option A -- Dataprep is the right answer

upvoted 1 times

✉  **Prasanna\_kumar** 1 year, 11 months ago

Answer is A

upvoted 1 times

✉  **MaxNRG** 2 years ago

**Selected Answer: A**

A: <https://cloud.google.com/dataprep/>

upvoted 2 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: A**

Cloud Dataprep is a tool to do the job.

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: A

upvoted 7 times

✉  **duytran\_d** 1 year, 6 months ago

this comment is being repeated and i really appreciate this feeling :D

upvoted 1 times

✉  **sandipk91** 2 years, 5 months ago

vote for option A

upvoted 4 times

✉  **sumanshu** 2 years, 6 months ago

Vote for 'A', because of requirement - Enabling non-developer analysts to modify transformations

upvoted 5 times

✉  **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 3 times

👤 SSV 3 years, 6 months ago

Answer should be D. Dataprep will detect schema automatically in the initial recipe. After 3 months if the schema is changed the scheduled dataprep job cannot handle. So D should be the option.

upvoted 1 times

Question #146

Topic 1

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC).

All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System

(HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

- A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Mount the Hive tables locally.
- B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.
- C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them to HDFS. Mount the Hive tables from HDFS.
- D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones.
- E. Load the ORC files into BigQuery. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables. Replicate external Hive tables to the native ones.

**Correct Answer: BC**

*Community vote distribution*

CD (38%)	AD (33%)	DE (17%)	13%
----------	----------	----------	-----

👤 Sid19 Highly Voted 2 years, 1 month ago

Answer is C and D 100%.

I know it says to transfer all the files but with the options provided C is the best choice.

Explanation

A and B cannot be true as gsutil can copy data to master node and then to hdfs from master node.

C -> works

D->works Recommended by google

E-> Will work but as the question says maximize performance this is not a case. As bigquery hadoop connector stores all the BQ data to GCS temp and then processes it to HDFS. As data is already in GCS we don't need to load it to bq and use a connector then unloads it back to G and then processes it.

upvoted 23 times

👤 KLei 3 months, 1 week ago

must go to the master node first...

upvoted 1 times

👤 rohan0411 7 months, 3 weeks ago

You can copy to worker nodes directly too by specifying the specific flag.

hdfs://<master node> is the default filesystem. You can explicitly specify the scheme and NameNode if desired:  
hdfs dfs -cp gs://<bucket>/<object> hdfs://<master node>/<hdfs path>

So the correct answer is A & D

upvoted 1 times

👤 WillemHendr 7 months, 3 weeks ago

I feel indeed this question is testing if you understand, that gsutil cannot transfer to HDFS directly (eliminate A&B), and need an intermediate step, (making C doable, with a good result). D is found on official google docs. E doesn't have good end result.

upvoted 3 times

👤 **cetanx** 8 months ago

According to Chat GPT => D, E

Cloud Storage connector for Hadoop allows Hadoop tools, including Hive, to work with data stored in Google Cloud Storage (GCS). It enables you to use GCS as you would HDFS, so you can directly mount ORC files stored in GCS as Hive tables without moving the data.

Loading the ORC files (which also contain schema information) into BigQuery and using the BigQuery connector for Hadoop is another option. The connector allows Hadoop and Spark applications to read/write data directly from/to BigQuery, providing another path for your Hive applications to access the data.

The other options are less efficient or incorrect. Copying the ORC files directly to the Hadoop Distributed File System (HDFS) on the Dataproc cluster with gsutil (options A, B, and C) would require unnecessary data transfer and storage, not leveraging the benefits of separation of storage and compute that GCS provides.

upvoted 1 times

👤 **[Removed]** Highly Voted 3 years, 10 months ago

Should be B C

upvoted 17 times

👤 **patitonav** Most Recent 1 month ago

**Selected Answer: DE**

I think D and E are the best and easy way to go. For sure D, but I think that E can work too, the data can be loaded in BQ as an external table at the end the data will be always on the GCS.

upvoted 2 times

👤 **barnac1es** 4 months, 1 week ago

**Selected Answer: DE**

D. Cloud Storage Connector for Hadoop: You can use the Cloud Storage connector for Hadoop to mount the ORC files stored in Cloud Storage as external Hive tables. This allows you to query the data without copying it to HDFS. You can replicate these external Hive tables to native Hive tables in Cloud Dataproc if needed.

E. Load ORC Files into BigQuery: Another approach is to load the ORC files into BigQuery, Google Cloud's data warehouse. Once the data is in BigQuery, you can use the BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables in Cloud Dataproc. This leverages the power of BigQuery for analytics and allows you to replicate external Hive tables to native ones in Cloud Dataproc.

upvoted 1 times

👤 **barnac1es** 4 months, 1 week ago

**Selected Answer: DE**

D. Cloud Storage Connector for Hadoop: You can use the Cloud Storage connector for Hadoop to mount the ORC files stored in Cloud Storage as external Hive tables. This allows you to query the data without copying it to HDFS. You can replicate these external Hive tables to native Hive tables in Cloud Dataproc if needed.

E. Load ORC Files into BigQuery: Another approach is to load the ORC files into BigQuery, Google Cloud's data warehouse. Once the data is in BigQuery, you can use the BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables in Cloud Dataproc. This leverages the power of BigQuery for analytics and allows you to replicate external Hive tables to native ones in Cloud Dataproc.

upvoted 1 times

👤 **vamgcp** 6 months, 1 week ago

**Selected Answer: AD**

A is the most straightforward way to start using Hive in Cloud Dataproc. You can use the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Then, you can mount the Hive tables locally.

D is another option that you can use to start using Hive in Cloud Dataproc. You can leverage the Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Then, you can replicate the external Hive tables to the native ones.

upvoted 2 times

👤 **Qix** 6 months, 3 weeks ago

**Selected Answer: BC**

Answers are;

B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.

C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them to HDFS. Mount the Hive tables from HDFS.

You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. HDFS lies on datanodes, data on master node needs to be copied on datanodes.

B for managed hive table option, C for external hive table

upvoted 2 times

✉  **izekc** 8 months, 3 weeks ago

**Selected Answer: AD**

AD is correct

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

**Selected Answer: AD**

i choose AD.

Searched in other w/s, read discussions here, and guess better AD.

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

gpt: Yes, that is correct. Option A is a valid way to transfer the ORC files to HDFS, and then mount the Hive tables locally. Option D is also valid, as it suggests using the Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables and then replicating the external Hive tables to native ones.

Chatgpt agreed, after inserting question and variants, and said that AD are correct answers. And it agreed. It adds some confidence that they are good, but gpt can make mistakes

upvoted 1 times

✉  **streeeber** 9 months, 3 weeks ago

**Selected Answer: AD**

A will copy to HDFS and so will D

upvoted 1 times

✉  **hauhau** 1 year, 1 month ago

**Selected Answer: AD**

C: master node doesn't make sense

upvoted 3 times

✉  **hauhau** 1 year, 1 month ago

B: from the Cloud Storage bucket to any node of the Dataproc cluster

-> still on cloud not maximize the speed

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: CD**

CD is the answer.

<https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-storage>

The Cloud Storage connector is an open source Java library that lets you run Apache Hadoop or Apache Spark jobs directly on data in Cloud Storage, and offers a number of benefits over choosing the Hadoop Distributed File System (HDFS).

Connector Support. The Cloud Storage connector is supported by Google Cloud for use with Google Cloud products and use cases, and when used with Dataproc is supported at the same level as Dataproc.

upvoted 4 times

✉  **tikki\_boy** 1 year, 3 months ago

I'll go with DE

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: CD**

CD is correct

upvoted 2 times

✉  **BigDataBB** 1 year, 11 months ago

**Selected Answer: BC**

You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. HDFS lies on datanode data on master node needs to be copied on datanode.

B for managed hive table option, C for external hive table

upvoted 1 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: CD**

as explained by Sid19

upvoted 3 times

✉  **bkovari** 2 years, 1 month ago

Prefer A & E. Since master node does not store data in HDFS (it does checkpointing, but the data is always present on the slave nodes) -> B is out.

Furthermore, when you copy using gsutil, you are defining a HDFS path instead of a node name (it is abstracted away to which node it is going to be put and replicated finally)

so B just makes no sense. Since performance is key, we need to have a native table eventually (when data is on HDFS it can be used as external table only, but we create the

native variant of it.) -> is fulfilled by option D. (E is out also, since nobody cares about BigQuery, the question is how to use Hive which is possible right after the data is present on HDFS.. )

upvoted 2 times

✉  **bkovari** 2 years, 1 month ago

Prefer A & D. typo, moderator can you fix it?

upvoted 3 times

Question #147

Topic 1

You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

- A. Cloud Scheduler
- B. Cloud Dataflow
- C. Cloud Functions
- D. Cloud Composer

**Correct Answer: D**

*Community vote distribution*

D (89%)

11%

✉  **mario\_ordinola**  2 years, 10 months ago

if someone are not sure that D is the answer, I suggest to don't take the exam

upvoted 40 times

✉  **madhu1171**  3 years, 10 months ago

D should be the answer

upvoted 23 times

👤 **patitonav** Most Recent 1 month ago

**Selected Answer: D**

No doubt

upvoted 1 times

👤 **barnac1es** 4 months, 1 week ago

**Selected Answer: D**

Workflow Orchestration: Cloud Composer is a fully managed workflow orchestration service based on Apache Airflow. It allows you to define, schedule, and manage complex workflows with multiple steps, including shell scripts, Hadoop jobs, and BigQuery queries.

Dependency Management: You can define dependencies between different steps in your workflow to ensure they are executed in a specific order.

Retry Mechanism: Cloud Composer provides built-in retry mechanisms, so if any step fails, it can be retried a fixed number of times according to your configuration.

Scheduled Execution: Cloud Composer allows you to schedule the execution of your workflows on a regular basis, meeting the requirement for executing the jobs on a schedule.

upvoted 1 times

👤 **AzureDP900** 1 year, 1 month ago

D is right

upvoted 2 times

👤 **zellck** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/composer/docs/concepts/overview>

Cloud Composer is a fully managed workflow orchestration service, enabling you to create, schedule, monitor, and manage workflows that span across clouds and on-premises data centers.

upvoted 4 times

👤 **DataEngineer\_WideOps** 1 year, 6 months ago

**Selected Answer: A**

Cloud Composer for sure.

upvoted 1 times

👤 **GoReplyGCPExam** 1 year, 4 months ago

Composer is D

upvoted 2 times

👤 **nadavw** 1 year, 7 months ago

D.

per document "Scheduler" is aimed to a single service and composer for an ETL , in addition it's not even specified all jobs are on cloud so only composer can handle it.

upvoted 1 times

👤 **nadavw** 1 year, 7 months ago

<https://cloud.google.com/blog/topics/developers-practitioners/choosing-right-orchestrator-google-cloud>

upvoted 1 times

👤 **medeis\_jar** 2 years ago

**Selected Answer: D**

Cloud Composer

upvoted 2 times

👤 **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?  
Ans: D

upvoted 2 times

👤 **daghayeghi** 2 years, 11 months ago

D:

the main point is that Cloud Composer should be used when there is inter-dependencies between the job, e.g. we need the output of a job to start another whenever the first finished, and use dependencies coming from first job.

upvoted 4 times

✉  **ashuchip** 3 years, 1 month ago

D seems to be quite relevant, because using composure you can do all things which are being asked to perform, even retry property is there composure.

upvoted 3 times

✉  **Alasmindas** 3 years, 2 months ago

The correct answer is Option A : Cloud Scheduler .

Although at first instance, I thought it should be Cloud Composer but then looking at the question and reading it few times - it concluded me to go for Option A.

Cloud Scheduler has built in retry handling so you can set a fixed number of times and doesn't have time limits for requests. The functionality is much simpler than Cloud Composer. Cloud Composer is managed Apache Airflow that "helps you create, schedule, monitor and manage workflows. For automate scheduled jobs - the most preferred method would be Scheduler, Composer would typically be used when we want to orchestrate many managed services and automate the work flow.

upvoted 5 times

✉  **kavs** 3 years, 2 months ago

A seems to be right

upvoted 1 times

✉  **mumukshu** 3 years, 1 month ago

I think D , how scheduler can handle this part " The jobs are expected to run for many minutes up to several hours"

upvoted 3 times

✉  **baubaumiaomiao** 2 years, 1 month ago

You forgot the "These jobs have many interdependent steps" which can be handled only through Composer

upvoted 1 times

✉  **Abby1356** 3 years, 2 months ago

should be A

upvoted 1 times

✉  **arghy13** 3 years, 3 months ago

Answer should be A..Cloud scheduler..cloud composer is an workflow manager. Can't run unix,bigquery jobs

upvoted 2 times

✉  **Tanmoyk** 3 years, 4 months ago

D should be the best option

upvoted 3 times

✉  **haroldbenites** 3 years, 5 months ago

D is correct

upvoted 3 times

to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

- A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
- B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
- C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Functions. Integrate the package tracking applications with this function.
- D. Use TensorFlow to create a model that is trained on your corpus of images. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

**Correct Answer: A**

*Community vote distribution*

B (76%)

C (24%)

 **[Removed]** Highly Voted 3 years, 10 months ago

Should be B.

upvoted 33 times

 **[Removed]** Highly Voted 3 years, 10 months ago

AutoML is used to train model and do damage detection

Auto Vision is used as a pre trained model used to detect objects in images

upvoted 23 times

 **ga8our** 3 months, 1 week ago

1. Who said we have a labelled corpus that can be fed to AutoML?

2. Auto Vision, as you say, is used to detect objects, like box, ship, human, etc. Now it depends only on our definition (parameters) of a "box" what the model should accept as intact or damaged.

3. ChatGCP also chooses C. Vision API.

Given that the question does not say we have labelled data, and that damage recognition is not qualitatively different from object recognition I'd go for C - C. Vision API.

upvoted 3 times

 **emmylou** 2 months, 1 week ago

Add the sentence, "This is a practice exam question. Please assume no changes to the architecture" and it brings back B

upvoted 1 times

 **brokeasspanda** 2 months, 2 weeks ago

Seriously, I feel the question was transcribed incorrectly but the answers were. It was probably meant to include a labeled corpus of images. With the details presented you'd have to hope that Vision API would be able to guess objects as damage.

upvoted 1 times

 **[Removed]** 3 years, 10 months ago

Correct : B

upvoted 12 times

 **Deepakd** 1 year, 10 months ago

Damage is an object in the image . So Auto Vision API can be used.

upvoted 2 times

 **enthGCP** Most Recent 1 month ago

as per chat gpt One of the features of Cloud Vision API is damage detection, which can be used to identify and classify various types of damage in images, such as cracks, dents, scratches, stains, etc

upvoted 1 times

👤 **MaxNRG** 1 month, 1 week ago

**Selected Answer: B**

For this scenario, where you need to automate the detection of damaged packages in real time while they are in transit, the most suitable solution among the provided options would be B.

Here's why this option is the most appropriate:

**Real-Time Analysis:** AutoML provides the capability to train a custom model specifically tailored to recognize patterns of damage in packages. This model can process images in real-time, which is essential in your scenario.

**Integration with Existing Systems:** By building an API around the AutoML model, you can seamlessly integrate this solution with your existing package tracking applications. This ensures that the system can flag damaged packages for human review efficiently.

**Customization and Accuracy:** Since the model is trained on your specific corpus of images, it can be more accurate in detecting damages relevant to your use case compared to pre-trained models.

upvoted 1 times

👤 **MaxNRG** 1 month, 1 week ago

Let's briefly consider why the other options are less suitable:

**A. Use BigQuery machine learning:** BigQuery is great for handling large-scale data analytics but is not optimized for real-time image processing or complex image recognition tasks like damage detection on packages.

**C. Use the Cloud Vision API:** While the Cloud Vision API is powerful for general image analysis, it might not be as effective for the specific task of detecting damage on packages, which requires a more customized approach.

**D. Use TensorFlow in Cloud Datalab:** While this is a viable option for creating a custom model, it might be more complex and time-consuming compared to using AutoML. Additionally, setting up a real-time analysis system through a Python notebook might not be as straightforward as an API integration.

upvoted 1 times

👤 **juliorevk** 1 month, 3 weeks ago

**Selected Answer: B**

I was leaning towards C but tested out uploading some damaged boxes to Vision API. It seems to have a lot of trouble detecting damaged boxes. It mislabeled boxes as a tire or toy. Also, there is no part of the API that seems to be able to detect damage. So I'll have to go with B. I should train a model to accomplish this then integrate with your app.

upvoted 2 times

👤 **barnacles** 4 months, 1 week ago

**Selected Answer: B**

**AutoML for Custom Models:** AutoML (Auto Machine Learning) is designed to simplify the process of training custom machine learning models including image classification models. It allows you to leverage Google Cloud's pre-built AutoML Vision service to train a model specifically for detecting package damage based on your corpus of images. This ensures accurate and customized results.

**Real-time API Integration:** After training the AutoML model, you can create an API endpoint that integrates seamlessly with your package tracking applications. This means that as packages move on the delivery lines, you can send images in real-time to the API for immediate analysis.

**Scalability:** AutoML Vision is built to scale, so it can handle the analysis of images in real-time, even as packages move continuously on the delivery lines.

upvoted 1 times

👤 **arien\_chen** 5 months, 1 week ago

**Selected Answer: C**

Keywords: realtime, camera streaming

<https://cloud.google.com/vision#:~:text=where%20you%20are-,Vertex%20AI%20Vision,-Vertex%C2%A0AI%20Vision>

Option B AutoML would be too complex and not time efficient.

Using Vision AI(Vertex AI Vision) first + AutoML

Option D is better than B (just AutoML).

upvoted 4 times

👤 **arien\_chen** 5 months, 1 week ago

typo: Option C is better than B.

upvoted 2 times

👤 **piyush7777** 5 months, 2 weeks ago

B

<https://www.cloudskillsboost.google/focuses/22020?parent=catalog>

upvoted 1 times

✉  **vamgcp** 6 months, 1 week ago

**Selected Answer: B**

Option B - AutoML

upvoted 1 times

✉  **Oleksandr0501** 8 months, 3 weeks ago

will stay with B. Might be more reliable, accurate.

as many says in discussion, Vision API does not say it has defect detection.

i remember labs with Auto ML, where models were trained. Vertex AI labs.

upvoted 2 times

✉  **AzureDP900** 1 year, 1 month ago

B is right

upvoted 2 times

✉  **Atnafu** 1 year, 2 months ago

B

C-is not answer

Vision API currently allows you to use the following features:

<https://cloud.google.com/vision/docs/features>

list#:~:text=Vision%20API%20currently%20allows%20you%20to%20use%20the%20following%20features%3A

upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

**Selected Answer: B**

It looks like B is the only valid option, with the assumption that you have a corpus of images (the question does not say that you do not). It would not be Cloud Vision API because that does not do damage detection (<https://cloud.google.com/vision/docs/features-list>).

upvoted 3 times

✉  **John\_Pongthorn** 1 year, 4 months ago

<https://cloud.google.com/solutions/visual-inspection-ai#all-features>

upvoted 2 times

✉  **cloudmon** 1 year, 2 months ago

This is how you would do it nowadays, but the question is not referring to this solution. It only refers to "Cloud Vision API" (not Visual Inspection API). Cloud Vision API does not do damage detection (<https://cloud.google.com/vision/docs/features-list>) so you would need to use AutoML. It looks like they assume that you have your own corpus of images.

upvoted 2 times

✉  **Deepakd** 1 year, 11 months ago

Here it is mentioned that the company is planning to implement a camera system. So it does not have the training data yet. Without having training data, the only option left is to use pre-trained models through cloud API. C is the answer. B is wrong as you don't have data to train the model.

upvoted 1 times

✉  **Deepakd** 1 year, 10 months ago

I would correct myself and go for B. I did not find any mention of cloud vision API being used for object detection.

upvoted 2 times

✉  **sraakesh95** 2 years ago

**Selected Answer: B**

AutoML is used to train model and do damage detection. Auto Vision is used as a pre-trained model used to detect objects in images.

upvoted 2 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: B**

Cloud Vision API -> pre-trained models to detect labels, faces, words

AutoML -> custom specific models trained for specific use case

upvoted 3 times

Question #149

Topic 1

You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?

- A. Assign the users/groups data viewer access at the table level for each table
- B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
- C. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views
- D. Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside

**Correct Answer: C**

*Community vote distribution*

A (72%)

D (22%)

6%

 **someshsehgal** Highly Voted 2 years, 11 months ago

Correct A: A . Now it is feasible to provide table level access to user by allowing user to query single table and no other table will be visible to user in same dataset.

upvoted 38 times

✉  **jits1984** 2 years, 3 months ago

Should still be D.

Question states - "They should only see certain tables based on their team membership"

Option A states - Assign the users/groups data viewer access at the table level for each table

With A, everyone will see every table. Hence D.

upvoted 9 times

✉  **Shiv\_am** 2 years, 5 months ago

A is not at all possible

upvoted 1 times

✉  **squishy\_fishy** 2 years, 3 months ago

It is possible for about a year now. [https://cloud.google.com/bigquery/docs/table-access-controls-intro#example\\_use\\_case](https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case)

upvoted 7 times

✉  **alecuba16** 1 year, 6 months ago

The problem is that option A has a lot of work for the DevOps, meanwhile option D is easier to manage. The view is like having a shortcut to the same data, but with different permissions

upvoted 1 times

✉  **cetanx** 8 months ago

According to Chat GPT, it is also D.

And it explains why it shouldn't be "A" as;

Granularity: While you can assign access permissions at the table level, it doesn't allow for fine-grained access control. For example, if you want to restrict access to certain columns or rows within a table based on user or group, table-level permissions would not be sufficient.

Scalability: In organizations with many tables and users, managing permissions at the table level can quickly become unwieldy. You would need to individually set permissions for each user for each table, which can be time-consuming and error-prone.

Security: Table-level permissions expose the entire table to a user or a group. If the data in the table changes over time, users might gain access to data they shouldn't see. With authorized views, you have more control over what data is exposed.

Maintenance: If the structure of your data changes (for instance, if tables are added or removed, or if the schema of a table changes), you would need to manually update the permissions for each affected table.

upvoted 3 times

✉  **madhu1171** Highly Voted 3 years, 10 months ago

D should be the answer

upvoted 26 times

✉  **ducc** 1 year, 5 months ago

It is updated, now A is correct

upvoted 1 times

✉  **squishy\_fishy** 2 years, 3 months ago

There is only one dataset mentioned in the question here. "You have migrated all of your data into tables in a dataset"

upvoted 3 times

✉  **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: D**

<https://cloud.google.com/solutions/migration/dw2bq/dw-bq-data-governance>

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your users would have access to both the view and the data.

<https://cloud.google.com/bigquery/docs/authorized-views>

This approach aligns with the Google Cloud best practices for data governance, ensuring that users can only access the data intended for them without having direct access to the source tables. Authorized views serve as a secure interface to the underlying data, and by placing these views in separate datasets per team, you can manage permissions effectively at the dataset level.

upvoted 3 times

✉  **lokiinaction** 2 months, 1 week ago

but the question said that all data are copied into one dataset. so it should be C

upvoted 1 times

👤 **spicebits** 2 months, 3 weeks ago

A is the best answer for security as stated in the documentation - [https://cloud.google.com/bigquery/docs/row-level-security-intro#comparison\\_of\\_authorized\\_views\\_row-level\\_security\\_and\\_separate\\_tables](https://cloud.google.com/bigquery/docs/row-level-security-intro#comparison_of_authorized_views_row-level_security_and_separate_tables)  
upvoted 2 times

👤 **EsaP** 4 months, 1 week ago

A is a better fit than D for this case  
upvoted 1 times

👤 **barnacles** 4 months, 1 week ago

**Selected Answer: C**

Authorized Views: Authorized views in BigQuery allow you to control access to specific rows and columns within a table. This means you can create views for each team that restrict access to only the data relevant to that team.  
Single Dataset: Keeping all the authorized views and the underlying data in the same dataset simplifies management and access control. It avoids the need to create multiple datasets, making the permission management process more straightforward.

Option A (assigning data viewer access at the table level) would not provide the granularity you need, as it would allow users to see all tables in the dataset. This does not align with the requirement to restrict access based on team membership.

upvoted 1 times

👤 **arien\_chen** 5 months, 1 week ago

**Selected Answer: D**

<https://cloud.google.com/bigquery/docs/share-access-views#:~:text=the%20source%20data.-,Authorized%20views,-should%20be%20crea>

For best practice, Option D is better than others.

upvoted 1 times

👤 **midgoo** 10 months, 1 week ago

**Selected Answer: A**

[A] is correct if it is for individual table  
However, in practice we normally do [C] as most of the time, the view is a JOIN of a few tables or a subset of the table (some columns removed)  
upvoted 1 times

👤 **musumusu** 11 months, 2 weeks ago

Answer A, Trick here is, if question is not asking for data level Access such as some rows or columns, don't go for authorized view in that case would go for C. If it's Table level request only in question, then A is simple answer

upvoted 1 times

👤 **zelick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

[https://cloud.google.com/bigquery/docs/control-access-to-resources-iam#grant\\_access\\_to\\_a\\_table\\_or\\_view](https://cloud.google.com/bigquery/docs/control-access-to-resources-iam#grant_access_to_a_table_or_view)

upvoted 1 times

👤 **gudiking** 1 year, 2 months ago

**Selected Answer: A**

A - table level access control now exists: [https://cloud.google.com/bigquery/docs/table-access-controls-intro#example\\_use\\_case](https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case)  
upvoted 1 times

👤 **Transcend** 1 year, 2 months ago

A.

Please see:

[https://cloud.google.com/bigquery/docs/table-access-controls-intro#example\\_use\\_case](https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case)

upvoted 1 times

👤 **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

<https://cloud.google.com/bigquery/docs/table-access-controls-intro>

Don't think too much ,there is nothing to do with view, the question refer to table obviously.

Assume that User see certain table so he can see everything in such a table

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: A**

It has nothing to do with authorize view because of the following

Authorized views make use of query results but this question emphasise on Table level

<https://cloud.google.com/bigquery/docs/authorized-views>

An authorized view lets you share query results with particular users and groups without giving them access to the underlying source data.  
upvoted 2 times

✉  **exnariantwort** 4 months, 1 week ago

finally after tens of comments, i see one that explains and makes sense

upvoted 1 times

✉  **ducc** 1 year, 5 months ago

**Selected Answer: A**

Vote for A

upvoted 2 times

✉  **ducc** 1 year, 5 months ago

**Selected Answer: A**

Vote for A

upvoted 2 times

Question #150

Topic 1

You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence.

To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

- A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory
- B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS
- C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up
- D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

**Correct Answer: A**

*Community vote distribution*

B (91%)

9%

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Correct: B

Local HDFS storage is a good option if:

Your jobs require a lot of metadata operations—for example, you have thousands of partitions and directories, and each file size is relatively small.

You modify the HDFS data frequently or you rename directories. (Cloud Storage objects are immutable, so renaming a directory is an expensive operation because it consists of copying all objects to a new key and deleting them afterwards.)

You heavily use the append operation on HDFS files.

You have workloads that involve heavy I/O. For example, you have a lot of partitioned writes, such as the following:

```
spark.read().write.partitionBy(..).parquet("gs://")
```

You have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation upvoted 35 times

✉  **Rajokkiyam** Highly Voted 3 years, 10 months ago

Answer B

It's google recommended approach to use LocalDisk/HDFS to store Intermediate result and use Cloud Storage for initial and final results.

upvoted 15 times

✉  **Chelseajcole** 2 years, 4 months ago

Any link to support this recommended approach?

upvoted 1 times

✉  **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: B**

Local HDFS storage is a good option if:

- You have workloads that involve heavy I/O. For example, you have a lot of partitioned writes such as the following:

```
spark.read().write.partitionBy(..).parquet("gs://")
```

- You have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation

- Your jobs require a lot of metadata operations—for example, you have thousands of partitions and directories, and each file size is relatively small.

- You modify the HDFS data frequently or you rename directories. (Cloud Storage objects are immutable, so renaming a directory is an expensive operation because it consists of copying all objects to a new key and deleting them afterwards.)

- You heavily use the append operation on HDFS files.

upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

We recommend using Cloud Storage as the initial and final source of data in a big-data pipeline. For example, if a workflow contains five Spark jobs in series, the first job retrieves the initial data from Cloud Storage and then writes shuffle data and intermediate job output to HDFS. The final Spark job writes its results to Cloud Storage.

[https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#choose\\_storage\\_options](https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#choose_storage_options)

upvoted 1 times

✉  **squishy\_fishy** 3 months ago

The correct answer is B.

upvoted 1 times

✉  **barnacles** 4 months, 1 week ago

**Selected Answer: B**

Disk I/O Performance: In a Cloud Dataproc cluster, the default setup uses local persistent disks for HDFS storage. These disks offer good disk I/O performance and are well-suited for storing intermediate data generated during Hadoop jobs.

Data Locality: Storing intermediate data on native HDFS allows for better data locality. This means that the data is stored on the same nodes where computation occurs, reducing the need for data transfer over the network. This can significantly improve the performance of disk I/O-intensive jobs.

Scalability: Cloud Dataproc clusters can be easily scaled up or down to meet the specific requirements of your jobs. You can allocate additional disk space as needed to accommodate the intermediate data generated by this particular Hadoop job.

upvoted 1 times

□  **DeepakVenkatachalam** 4 months, 2 weeks ago

Correct: A

I'd choose A as the doc states adding more SSDs are good for disk-intensive jobs especially those with many individual read and write operations

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

upvoted 1 times

□  **DeepakVenkatachalam** 4 months, 1 week ago

Typo Correct Answer is B. . Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS

upvoted 1 times

□  **arien\_chen** 5 months, 1 week ago

**Selected Answer: A**

I would choose A.

Google Storage is faster than HDFS in many cases.

<https://cloud.google.com/architecture/hadoop#:~:text=It%27s%20faster%20than%20HDFS%20in%20many%20cases.>

The question mention '(8-core nodes with 100-GB RAM)' on-premises Hadoop.

the problem may caused by insufficient memory,  
and does not mention cost would be an issue,  
so A 'memory' approach would be a better option.

upvoted 1 times

□  **vamgcp** 6 months, 1 week ago

**Selected Answer: B**

Best option is B. However allocating sufficient persistent disk space to the Hadoop cluster, and storing the intermediate data of that particular Hadoop job on native HDFS, would not improve the performance of the Hadoop job. In fact, it might even slow down the Hadoop job, as the data would have to be read and written to disk twice.

upvoted 1 times

□  **zellck** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

[https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing\\_primary\\_disk\\_options](https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing_primary_disk_options)

If your job is disk-intensive and is executing slowly on individual nodes, you can add more primary disk space. For particularly disk-intensive jobs, especially those with many individual read and write operations, you might be able to improve operation by adding local SSDs. Add enough SSDs to contain all of the space you need for local execution. Your local execution directories are spread across however many SSDs you add.

upvoted 3 times

□  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

[https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing\\_primary\\_disk\\_options](https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing_primary_disk_options)

upvoted 1 times

□  **rrr000** 1 year, 5 months ago

B is not the right answer. The problem says that for intermediate data cloud storage is to be used, while B option says:

B ... the intermediate data of that particular Hadoop job on native HDFS

A is the right answer. If you have enough memory then the shuffle wont spill on the disk.

upvoted 3 times

□  **rrr000** 1 year, 5 months ago

Further the question states that original on-prem machines has 100gb ram.

8-core nodes with 100-GB RAM

upvoted 2 times

□  **SoerenE** 2 years ago

B should be the right answer: [https://cloud.google.com/compute/docs/disks/performance#optimize\\_disk\\_performance](https://cloud.google.com/compute/docs/disks/performance#optimize_disk_performance)

upvoted 1 times

👤 **medeis\_jar** 2 years ago

Selected Answer: B

<https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs>

upvoted 3 times

👤 **JG123** 2 years, 2 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: B

upvoted 3 times

👤 **RT30** 2 years, 10 months ago

If your job is disk-intensive and is executing slowly on individual nodes, you can add more primary disk space. For particularly disk-intensive jobs, especially those with many individual read and write operations, you might be able to improve operation by adding local SSDs. Add local SSDs to contain all of the space you need for local execution. Your local execution directories are spread across however many SSDs you add. Its B

<https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs>

upvoted 3 times

👤 **ashuchip** 3 years, 1 month ago

yes B is correct

upvoted 2 times

👤 **Alasmindas** 3 years, 2 months ago

Correct Answer is Option B - Adding persistent disk space, reasons:-

- The question mentions that the particular job is "disk I/O intensive" - so the word "disk" is explicitly mentioned.
- Option B also mentions about local HDFS storage, which is ideally a good option of general I/O intensive work.

upvoted 5 times

Question #151

Topic 1

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 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 Vertex AI for training existing Spark ML models
- B. Rewrite your models on TensorFlow, and start using Vertex AI
- C. Use 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

**Correct Answer: A**

*Community vote distribution*

C (63%)

A (37%)

👤 **vamgcp** Highly Voted 6 months, 1 week ago

Selected Answer: C

Option C : It is the most rapid way to migrate your existing training pipelines to Google Cloud.

It allows you to continue using your existing Spark ML models.

It allows you to take advantage of the scalability and performance of Dataproc.

It allows you to read data directly from BigQuery, which is a more efficient way to process large datasets

upvoted 5 times

👤 **vaga1** Highly Voted 8 months, 3 weeks ago

**Selected Answer: A**

the question is: is it faster to move a SparkML job to a Vertex AI or to Dataproc? I am personally not sure, I would go for Dataproc as notebooks are not mentioned, but reading the Google article:

<https://cloud.google.com/blog/topics/developers-practitioners/announcing-serverless-spark-components-vertex-ai-pipelines/>

"Dataproc Serverless components for Vertex AI Pipelines that further simplify MLOps for Spark, Spark SQL, PySpark and Spark jobs."

upvoted 5 times

👤 **emmylou** 2 months, 1 week ago

But you would need to re-write your models which can be a block

upvoted 2 times

👤 **GCP001** Most Recent 1 week, 5 days ago

**Selected Answer: C**

C looks more suitable as data is already on BigQuery.

Ref - <https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml>

upvoted 1 times

👤 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C, agreed with other comments

upvoted 1 times

👤 **MaxNRG** 1 month, 1 week ago

**Selected Answer: C**

Use Cloud Dataproc, BigQuery, and Apache Spark ML for Machine Learning

<https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml>

Using Apache Spark with TensorFlow on Google Cloud Platform

<https://cloud.google.com/blog/products/gcp/using-apache-spark-with-tensorflow-on-google-cloud-platform>

upvoted 3 times

👤 **Nandababy** 1 month, 2 weeks ago

Why not option D? To spin up the spark cluster on compute engine, considering rapid migration it potentially could be best approach as team won't have to re-work on model (may be only few configurational changes) and again to get data from Bigquery which is required periodically is all the time, could be easy.

With Dataproc it would have more code changes eventually can take more time.

With Vertex AI it doesn't support spark ML natively and also training would be black box.

For me Answer should be D.

upvoted 1 times

👤 **barnacles** 4 months, 1 week ago

**Selected Answer: C**

Dataproc for Spark: Google Cloud Dataproc is a managed Spark and Hadoop service that allows you to run Spark jobs seamlessly on Google Cloud. It provides the flexibility to run Spark jobs using Spark MLlib and other Spark libraries.

BigQuery Integration: You mentioned that your data is being migrated to BigQuery. Dataproc has native integration with BigQuery, allowing you to read data directly from BigQuery tables. This eliminates the need to export data from BigQuery to another storage system before processing it with Spark.

Rapid Migration: This approach allows you to quickly migrate your existing Spark ML models and training pipelines without the need for a complete rewrite or extensive changes to your existing workflows. You can continue using your Spark ML models while adapting them to read data from BigQuery.

upvoted 2 times

👤 **DeepakVenkatachalam** 4 months, 1 week ago

they are talking about rapid lift and shift, in which case Dataproc cluster will be right one for Spark ML models for lift and shift. so I think the answer is C.

upvoted 1 times

ckanaar 4 months, 1 week ago

Selected Answer: A

The updated answer seems A based on the following article:

<https://cloud.google.com/blog/topics/developers-practitioners/announcing-serverless-spark-components-vertex-ai-pipelines/>

upvoted 4 times

FP77 5 months ago

Selected Answer: C

The answer is C. Spin up a Cloud Dataproc Cluster, migrate spark jobs to there, and link the Cluster to Bqquery with the connector. It's a straightforward solution.

upvoted 1 times

knith66 6 months, 1 week ago

Selected Answer: C

If you wanted to use Vertex AI for training Spark ML models, you would typically need to convert your Spark ML code to another supported machine learning framework like TensorFlow or scikit-learn. Then you could use Vertex AI's pre-built training and prediction services for those frameworks.

upvoted 3 times

wan2three 6 months, 2 weeks ago

Selected Answer: A

Through Vertex AI Workbench, Vertex AI is natively integrated with BigQuery, Dataproc, and Spark. You can use BigQuery ML to create and execute machine learning models in BigQuery using standard SQL queries on existing business intelligence tools and spreadsheets, or you can export datasets from BigQuery directly into Vertex AI Workbench and run your models from there.

<https://cloud.google.com/vertex-ai#all-features:~:text=Data%20and%20AI%20integration>

upvoted 1 times

blathul 7 months, 1 week ago

Selected Answer: C

Dataproc is a managed Spark and Hadoop service on Google Cloud, which makes it an ideal choice for migrating your existing Spark ML training pipelines. By using Dataproc, you can continue to leverage Spark and its ML capabilities without the need for significant code changes or rewriting your models.

By combining Dataproc and BigQuery, you can create Spark jobs or workflows in Dataproc that read data from BigQuery and train your existing Spark ML models. This approach allows you to quickly migrate your training pipelines to Google Cloud and take advantage of the scalability and performance benefits of both Dataproc and BigQuery.

upvoted 1 times

KC\_go\_reply 7 months, 1 week ago

Selected Answer: C

It is obviously C) Dataproc, since we don't want to rewrite the training from scratch, highly prefer Dataproc for anything Hadoop/Spark ecosystem, and Vertex AI doesn't support \*training\* with SparkML (but deploying existing models).

upvoted 4 times

Takshashila 7 months, 2 weeks ago

Selected Answer: C

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

upvoted 1 times

brandonriddle 7 months, 3 weeks ago

Vertex AI does not currently support Spark ML models.

upvoted 2 times

ckanaar 4 months, 1 week ago

Actually, it does support serving Spark ML models, but training is not mentioned anywhere:

<https://cloud.google.com/architecture/spark-ml-model-with-vertexai>

upvoted 2 times

spsengineer101 8 months, 1 week ago

Selected Answer: A

Selected A

upvoted 1 times

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

**Correct Answer: A***Community vote distribution*

A (100%)

**✉️** **👤 [Removed]** **Highly Voted** 3 years, 10 months ago

Answer: A

Description: Geospatial and ML functionality is with bigquery

upvoted 21 times

**✉️** **👤 [Removed]** **Highly Voted** 3 years, 10 months ago

Answer : A

upvoted 15 times

**✉️** **👤 barnac1es** **Most Recent** 4 months, 1 week ago**Selected Answer: A**

Here's why BigQuery is a good choice:

Scalable Data Storage: BigQuery is a fully managed, highly scalable data warehouse that can handle large volumes of data, including your 40 dataset. It allows you to store and manage your data efficiently.

SQL for Predictive Analytics: BigQuery supports standard SQL and has built-in machine learning capabilities through BigQuery ML. You can easily build predictive models using SQL queries, which aligns with your goal of predicting ship delays.

Geospatial Processing: BigQuery has robust support for geospatial data processing. It provides functions for working with GeoJSON and geospatial data types, making it suitable for your ship telemetry data and geospatial analysis.

Integration with Dashboards: BigQuery can be easily integrated with visualization tools like Google Data Studio or other BI tools. You can create interactive dashboards to monitor ship delays based on your model's predictions.

upvoted 2 times

✉  **musumusu** 11 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 geospatial capabilities but still not quick enough for semi structure geojson data.

Prediction for delay of ships <<likely to>> For me its time crucial and almost real time requirement. BigQuery is not suitable for it.

Best solution for this case is: Use BigTable for storage, create a datflow pipeline / google cloud AI platform for time sensitive prediction.

upvoted 1 times

✉  **musumusu** 11 months, 1 week ago

answer A: You are just looking for a storage solution not a workflow

upvoted 4 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/bigquery/docs/geospatial-intro>

In a data warehouse like BigQuery, location information is very common. Many critical business decisions revolve around location data. For example, you may record the latitude and longitude of your delivery vehicles or packages over time. You may also record customer transactions and join the data to another table with store location data.

You can use this type of location data to determine when a package is likely to arrive or to determine which customers should receive a mailing at a particular store location. Geospatial analytics let you analyze and visualize geospatial data in BigQuery by using geography data types and Google Standard SQL geography functions.

upvoted 2 times

✉  **Atnafu** 1 year, 2 months ago

A

Geospatial analytics let you analyze and visualize geospatial data in BigQuery by using geography data types and Google Standard SQL geography functions. + BigqueryML

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

Answer: C

upvoted 1 times

✉  **Chihhanyu** 2 years, 2 months ago

GeoJson + Native functionality for prediction -> BigQuery

upvoted 3 times

✉  **singh\_payal\_1404** 2 years, 2 months ago

Answer : A

upvoted 1 times

✉  **PM17** 2 years, 3 months ago

This is more of a question than an answer but: How much data can Bigquery handle?

40TB seems to be a lot and bigtable can handle that, but of course Bigquery is better when it comes to ML and GIS.

upvoted 3 times

✉  **haroldbenites** 3 years, 5 months ago

A is correct

upvoted 3 times

✉  **FARR** 3 years, 5 months ago

A

<https://cloud.google.com/bigquery/docs/gis-intro>

upvoted 3 times

✉  **Rajokkiyam** 3 years, 10 months ago

Answer A

upvoted 2 times

Question #153

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

A (100%)

 [Removed]  3 years, 10 months ago

Should be A

upvoted 27 times

 [Removed]  3 years, 10 months ago

Correct: A

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

upvoted 17 times

 barnac1es  4 months, 1 week ago

**Selected Answer: A**

Dataflow with Sliding Time Windows: Dataflow allows you to work with event-time windows, making it suitable for time-series data like incoming IoT messages. Using sliding windows every 5 minutes allows you to compute moving averages efficiently.

Sliding Time Window: The sliding time window of 1 hour every 5 minutes enables you to calculate the moving average over the specified time frame.

Computing Averages: You can efficiently compute the average when each sliding window closes. This approach ensures that you have real-time visibility into the message rate and can detect deviations from the expected rate.

Alerting: When the calculated average drops below 4000 messages per second, you can trigger an alert from within the Dataflow pipeline, sending it to your desired alerting mechanism, such as Cloud Monitoring, Pub/Sub, or another notification service.

Scalability: Dataflow can scale automatically based on the incoming data volume, ensuring that you can handle the expected rate of 5000 messages per second.

upvoted 2 times

 vamgcp 6 months, 1 week ago

**Selected Answer: A**

Option A

Pros:

This option is relatively simple to implement.

It can be used to compute the moving average over any time window.

Cons:

This option can be computationally expensive, especially if the data stream is large.

It can be difficult to troubleshoot if the alert does not fire when it is supposed to.

upvoted 2 times

✉  **vaga1** 8 months, 3 weeks ago

**Selected Answer: A**

the correct answer is between A and B since it doesn't make sense to use Pub/Sub combined with Kafka. To have a Moving Average then we should go for A, updating the average estimation every 5 minutes using the new data that came in and eliminating the "most far" 5 minutes.  
upvoted 2 times

✉  **zelliick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#windows>

Windowing functions divide unbounded collections into logical components, or windows. Windowing functions group unbounded collections the timestamps of the individual elements. Each window contains a finite number of elements.

You set the following windows with the Apache Beam SDK or Dataflow SQL streaming extensions:

- Hopping windows (called sliding windows in Apache Beam)

A hopping window represents a consistent time interval in the data stream. Hopping windows can overlap, whereas tumbling windows are disjoint.

For example, a hopping window can start every thirty seconds and capture one minute of data. The frequency with which hopping windows begin is called the period. This example has a one-minute window and thirty-second period.

upvoted 4 times

✉  **medeis\_jar** 2 years ago

**Selected Answer: A**

as explained by Alasmindas

upvoted 2 times

✉  **AACHB** 2 years, 1 month ago

**Selected Answer: A**

Correct Answer: A

upvoted 2 times

✉  **JG123** 2 years, 2 months ago

Correct: A

upvoted 1 times

✉  **Chelseajcole** 2 years, 4 months ago

A is enough

upvoted 1 times

✉  **daghayeghi** 2 years, 11 months ago

A:

the correct answer is between A and B, But because used "Moving Average" then we should go for A.

upvoted 2 times

✉  **apnu** 3 years ago

yes , using KafkalIO , we can connect to Kafka cluster.

upvoted 2 times

✉  **ashuchip** 3 years, 1 month ago

yes A is correct , because sliding window can only help here.

upvoted 3 times

✉  **Alasmindas** 3 years, 2 months ago

Option A is the correct answer. Reasons:-

a) Kafka IO and Dataflow is a valid option for interconnect (needless where Kafka is located - On Prem/Google Cloud/Other cloud)  
b) Sliding Window will help to calculate average.

Option C and D are overkill and complex, considering the scenario in the question,  
<https://cloud.google.com/solutions/processing-messages-from-kafka-hosted-outside-gcp>

upvoted 7 times

✉  **Alasmindas** 3 years, 2 months ago

Option A is the correct answer. Reasons:-

- a) Kafka IO and Dataflow is a valid option for interconnect (needless where Kafka is located - On Prem/Google Cloud/Other cloud)
- b) Sliding Window will help to calculate average.

Option C and D are overkill and complex, considering the scenario in the question,

upvoted 6 times

✉  **atnafu2020** 3 years, 5 months ago

A

To take running averages of data, use hopping windows. You can use one-minute hopping windows with a thirty-second period to compute a one-minute running average every thirty seconds.

upvoted 2 times

✉  **Prakzz** 3 years, 6 months ago

I don't think its A or B. Dataflow can't connect directly to kafka.

upvoted 1 times

✉  **FARR** 3 years, 5 months ago

Yes, via KafkaIO. See the link in above comment

upvoted 3 times

Question #154

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

A (59%)

B (41%)

✉  **madhu1171** Highly Voted 3 years, 10 months ago

A should be correct answer

upvoted 29 times

✉  **tycho** 2 years, 1 month ago

yes A is correct, when creating a 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 4 times

✉  **[Removed]** Highly Voted 3 years, 10 months ago

Correct: A

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

upvoted 13 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: A**

A (failover replicas) as this is an old question:

In a legacy HA configuration, a Cloud SQL for MySQL instance uses a failover replica to add high availability to the instance. This functionality isn't available in Google Cloud console.

The new configuration doesn't use failover replicas. Instead, it uses Google's regional persistent disks, which synchronously replicate data at block-level between two zones in a region.

<https://cloud.google.com/sql/docs/mysql/configure-legacy-ha>

upvoted 3 times

pss111423 2 months, 1 week ago

Option A is good fro leagacy soution

Note: Cloud SQL plans to discontinue support for legacy HA instances in the future and will soon be announcing a date to do so. Currently, legacy HA instances are still covered by the Cloud SQL SLA. We recommend you upgrade your existing legacy HA instances to regional persistent disk HA instances and create new instances using regional persistent disk HA as soon as possible

Option C makes more sense in this regard

upvoted 1 times

emmylou 2 months, 2 weeks ago

A - Although it is legacy and will be deprecated. The correct answer is not an option--

"The legacy configuration for high availability used a failover replica instance. The new configuration does not use a failover replica. Instead, it uses Google's regional persistent disks, which synchronously replicate data at the block level between two zones in a region."

upvoted 1 times

barnacles 4 months, 1 week ago

**Selected Answer: A**

Failover Replica: By creating a failover replica in another zone within the same region, you establish a high-availability configuration. The failover replica is kept in sync with the primary instance, and it can quickly take over in case of a failure of the primary instance.

Same Region: Placing the failover replica in the same region ensures minimal latency and data consistency. In the event of a zone failure, the failover can happen within the same region, reducing potential downtime.

Zone Resilience: Google Cloud's regional design ensures that zones within a region are independent of each other, which adds resilience to zone failures.

Automatic Failover: In case of a primary instance failure, Cloud SQL will automatically promote the failover replica to become the new primary instance, minimizing downtime.

upvoted 1 times

samstar4180 5 months, 1 week ago

Per latest Google cloud document, B is the correct answer.

upvoted 1 times

wan2three 6 months, 2 weeks ago

**Selected Answer: B**

Cross-region read replicas

Cross-region replication lets you create a read replica in a different region from the primary instance. You create a cross-region read replica the same way as you create an in-region replica.

Cross-region replicas:

Improve read performance by making replicas available closer to your application's region.

Provide additional disaster recovery capability to guard against a regional failure.

Let you migrate data from one region to another.

<https://cloud.google.com/sql/docs/mysql/replication#cross-region-read-replicas>:~:text=memory%20(OOM)%20events.-,Cross%2Dregion%20read%20replicas,Let%20you%20migrate%20data%20from%20one%20region%20to%20another.-,See%20Promoting%20replicas

upvoted 2 times

MoeHaydar 6 months, 3 weeks ago

**Selected Answer: B**

The legacy process for adding high availability to MySQL instances uses a failover replica. The legacy functionality isn't available in the Google Cloud console. See Legacy configuration: Creating a new instance configured for high availability or Legacy configuration: Configuring an existing instance for high availability.

upvoted 2 times

✉  **KK0202** 7 months ago

**Selected Answer: B**

The correct answer is most probably B as this his scenario has an update(As of July 2023). Failover replicas are not available anymore. Same region different zone read replicas are used in case of a failover or if primary zone is not available

upvoted 3 times

✉  **MBRSDG** 8 months, 2 weeks ago

**Selected Answer: B**

The answer is B, the failover replica is a legacy feature.

See here: [https://cloud.google.com/sql/docs/mysql/high-availability#legacy\\_mysql\\_high\\_availability\\_option](https://cloud.google.com/sql/docs/mysql/high-availability#legacy_mysql_high_availability_option)

upvoted 1 times

✉  **forepick** 8 months ago

Read replica isn't an alternative to the standby instance

upvoted 1 times

✉  **vaga1** 8 months, 3 weeks ago

**Selected Answer: A**

read replica (B) and external read replica (C) doesn't make sense here, since we potentially need all the functionalities. Using Cloud SQL in a region combined with Cloud Storage backup may not be the best choice (D) thinking about compliance reasons starting from what has been asked, it seems also "too much" compared with A that fulfills the request with simpler actions. Also, compliance is required at the regional level so then A fits.

upvoted 2 times

✉  **wjtb** 10 months, 3 weeks ago

Failover replica's are a legacy feature. This question is outdated: <https://cloud.google.com/sql/docs/mysql/configure-ha>

upvoted 5 times

✉  **musumusu** 11 months, 2 weeks ago

Answer A, key words to remember, High Scale use extra read replica. High availability use extra failure replica. Both should be in different zones but in same region.

upvoted 2 times

✉  **desertlotus1211** 1 year ago

Answer is B: <https://cloud.google.com/sql/docs/mysql/replication#read-replicas>

'As a best practice, put read replicas in a different zone than the primary instance when you use HA on your primary instance'

upvoted 2 times

✉  **desertlotus1211** 1 year ago

The question asks to ensure high availability in the event of a zone failure

upvoted 1 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/sql/docs/mysql/high-availability#HA-configuration>

The HA configuration provides data redundancy. A Cloud SQL instance configured for HA is also called a regional instance and has a primary secondary zone within the configured region. Within a regional instance, the configuration is made up of a primary instance and a standby instance. Through synchronous replication to each zone's persistent disk, all writes made to the primary instance are replicated to disks in both zones before a transaction is reported as committed. In the event of an instance or zone failure, the standby instance becomes the new primary instance. Users are then rerouted to the new primary instance. This process is called a failover.

upvoted 1 times

✉  **louisgcpde** 1 year, 3 months ago

**Selected Answer: A**

A should be the answer as the question is asking HA in event of a zone failure.

"Read Replicas CAN be promoted to master nodes in the case of DR. However, there is downtime entailed.

Failover Replicas are designed to automatically become master nodes."

<https://googlecloudarchitect.us/read-replica-versus-failover-replica-in-cloud-sql/>

upvoted 2 times

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
- B. Cloud Storage
- C. Dataflow
- D. Firebase Cloud Messaging

**Correct Answer: A**

*Community vote distribution*

A (100%)

 **YoreNation** Highly Voted 1 year, 4 months ago

**Selected Answer: A**

A I think it's the only technology that met the requirements  
upvoted 8 times

 **dn\_mohammed\_data** Highly Voted 1 year, 4 months ago

vote for A: topics, offsets --> apache kafka  
upvoted 6 times

 **barnacles** Most Recent 4 months, 1 week ago

Ability to Seek to a Particular Offset: Kafka allows consumers to seek to a specific offset in a topic, enabling you to read data from a specific point, including back to the start of all data ever captured. This is a fundamental capability of Kafka.

Support for Publish/Subscribe Semantics: Kafka supports publish/subscribe semantics through topics. You can have hundreds of topics in Kafka, and consumers can subscribe to these topics to receive messages in a publish/subscribe fashion.

Retain Per-Key Ordering: Kafka retains the order of messages within a partition. If you have a key associated with your messages, you can ensure per-key ordering by sending messages with the same key to the same partition.

Scalability: Kafka is designed to handle high-throughput data streaming and is capable of scaling to meet your needs.

Apache Kafka aligns well with the requirements you've outlined for centralized data ingestion and delivery. It's a robust choice for scenarios that involve data streaming, publish/subscribe, and retaining message ordering.

upvoted 2 times

 **musumusu** 11 months, 2 weeks ago

Answer A: Apache Kafka

Key words: Ingestion and Delivery together ( it is combination of pub/sub for ingestion, and delivery = dataflow+any database in gcp)  
Offset of a topic = Partition of a topic and reprocess specific part of topic, its not possible in pub/sub as it is designed for as come and go for topic.

Per key ordering. means message with same key can be process or assigned to a user in kafka.

upvoted 2 times

 **aquevedos91** 1 year, 4 months ago

deberia ser la C, debido a que siempre es mejor escoger los servicios de google

upvoted 1 times

Question #156

Topic 1

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://

**Correct Answer: A**

*Community vote distribution*

A (100%)

 **[Removed]**  3 years, 10 months ago

Correct: A

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

upvoted 32 times

 **[Removed]**  3 years, 10 months ago

Confused between A and B. For r/w intensive jobs need to use SSDs. But questions doesn't state anything about the nature of the jobs. So be to start with a default option.

Choose A

upvoted 14 times

 **baubaumiaomiao** 2 years, 1 month ago

"You need to ensure that the deployment is as cost-effective as possible"  
hence, no SSD unless stated otherwise

upvoted 2 times

✉  **barnacles** Most Recent 4 months, 1 week ago

**Selected Answer: A**

**Dataproc Managed Service:** Dataproc is a fully managed service for running Apache Hadoop and Spark. It provides ease of management and automation.

**Standard Persistent Disk:** Using standard persistent disks for Dataproc workers ensures durability and is cost-effective compared to SSDs.

**Preemptible Workers:** By using 50% preemptible workers, you can significantly reduce costs while maintaining fault tolerance. Preemptible VMs are cheaper but can be preempted by Google, so having a mix of preemptible and non-preemptible workers provides cost savings with redundancy.

**Storing Data in Cloud Storage:** Storing data in Cloud Storage is highly durable, scalable, and cost-effective. It also makes data accessible to Dataproc clusters, and you can leverage native connectors for reading data from Cloud Storage.

**Changing References to gs://:** Updating your scripts to reference data in Cloud Storage using gs:// ensures that your jobs work seamlessly with the cloud storage infrastructure.

upvoted 2 times

✉  **vaga1** 8 months, 3 weeks ago

**Selected Answer: A**

Apache Hadoop -> Dataproc or Compute Engine with proper SW installation  
cost-effective -> use standard persistent disk + store data in Cloud Storage  
batch -> Dataproc or Compute Engine with proper SW installation  
managed service -> Dataproc

upvoted 1 times

✉  **zellck** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

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

Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning. Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them. With less time and money spent on administration, you can focus on your jobs and your data.

upvoted 1 times

✉  **MounicaN** 1 year, 4 months ago

**Selected Answer: A**

it says cost effective , hence no SSD

upvoted 1 times

✉  **JG123** 2 years, 2 months ago

Correct: A

upvoted 2 times

✉  **LORETOGOMEZ** 2 years, 6 months ago

Correct : A

Option B is useful if you use HDFS, and in this case as you use preemptible machines it isn't worth using SSD disks.

upvoted 2 times

✉  **ArunSingh1028** 2 years, 11 months ago

Answer - B

upvoted 1 times

✉  **StelSen** 3 years ago

Look at this link. <https://cloud.google.com/bigtable/docs/choosing-ssd-hdd>

At the first look I chose Option-B as they mentioned SSD is cost-effective in most cases. But after reading the whole page, they also mention that for batch workloads, HDD is suggested as long as not heavy read. So I changed my mind to Option-A (I assumed this is not a heavy process?).

upvoted 5 times

✉  **NM1212** 1 year, 5 months ago

Caution about the link you provided as reference. It's intended for BigTable which is Google's low-latency solution which is totally different requirement. Mentioning only because on first read I thought SSD is the obvious choice.

Per below link, SSD may not be required unless there is a low-latency requirement or a high I/O requirement. Since the question does not specify anything like that, A looks correct.

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

upvoted 1 times

✉ **Alasmindas** 3 years, 2 months ago

Option B - SSD disks, reasons:-

The question asks "fault-tolerant and cost-effective as possible for long-running batch job".

3 Key words are - fault tolerant / cost effective / long running batch jobs..

The cost efficiency part mentioned in the question could be addressed by 50% preemptible disks and storing the data in cloud storage than HDFS.

For long running batch jobs and as standard approach for Dataproc - we should always go with SSD disk types as per google recommendation  
upvoted 4 times

✉ **beedle** 3 years, 1 month ago

where is the proof...show me the link?

upvoted 2 times

✉ **Raangs** 2 years, 11 months ago

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

As per this, SSD is only recommended if it is high IO intensive. In this question no where mentioned its high IO intensive, and asks for cost effective (as much as possible), so no need to use SSD.

I will go with A.

upvoted 6 times

✉ **Ravivarma4786** 3 years, 5 months ago

Ans is B, for long running SDD suitable. HDD maintenance will be additional charge for long running jobs

upvoted 2 times

✉ **Rajuuu** 3 years, 6 months ago

Answer is A...Cloud Dataproc for Managed Cloud native application and HDD for cost-effective solution.

upvoted 7 times

✉ **Rajokkiyam** 3 years, 10 months ago

Answer A

upvoted 5 times

Question #157

Topic 1

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
- 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

Correct Answer: D

### Community vote distribution

A (100%)

✉  **aadisme** Highly Voted 3 years, 7 months ago

Seems to be A. Preprocessing/scaling should be done with input features, instead of predictions (output)  
upvoted 41 times

✉  **FARR** Highly Voted 3 years, 5 months ago

A  
Deep LEarning is not always the best solution  
D talks about fudging the output which is wrong  
upvoted 11 times

✉  **MaxNRG** Most Recent 1 month, 1 week ago

Selected Answer: A

<https://www.quora.com/How-can-I-improve-Precision-Recall-AUC-under-Imbalanced-Classification>  
upvoted 3 times

✉  **vaga1** 8 months, 3 weeks ago

Selected Answer: A

B,C are simply not true. D is modifying the scoring, making it not reliable anymore. A makes sense, is potentially increasing the model accuracy  
upvoted 2 times

✉  **rishu2** 8 months, 4 weeks ago

Selected Answer: A

a is the correct answer  
upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

Answer A,  
why not B, Deep Neu Net. are better for sure but AUC is 0.87 is already good. Don't go for complex and time taking model. If AUC more than 0.95, it can be a reason of overfit.  
Now just check SVM params for hypertuning if you can bring it close to 0.9-0.95  
upvoted 1 times

✉  **Kvk117** 1 year ago

Selected Answer: A

a is the correct answer  
upvoted 1 times

✉  **Dan137** 1 year, 2 months ago

Also a good read is: <https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview>  
upvoted 1 times

✉  **medeis\_jar** 2 years ago

Selected Answer: A

as mentioned by Spider7 "performing tuning rather than using the model default parameters there's a way to increase the overall model performance --> A."  
upvoted 2 times

✉  **JG123** 2 years, 2 months ago

Correct: A  
upvoted 1 times

✉  **Spider7** 2 years, 2 months ago

0.89 it's already not bad but by performing tuning rather than using the model default parameters there's a way to increase the overall model performance --> A.  
upvoted 3 times

✉  **Spider7** 2 years, 2 months ago

0.87 precisely  
upvoted 1 times

✉  **hdmi\_switch** 2 years, 6 months ago

Not C because real-world AUC value falls between 0.5 and 1.0 usually, this wouldn't help.

A seems the most straigh forward.

upvoted 3 times

✉  **Mitra123** 2 years, 10 months ago

For a large enough training set DNN will most likely beat a SVM. However the opposite may or may not be true. It also depends on the complexity of the problem. Which we don't know from the question. For image, nlp, I say B can be a good answer. However, if we decide to stick with SVM, D reduces overfitting and may increase AUC.

I am torn between the two!

upvoted 1 times

✉  **ArunSingh1028** 2 years, 11 months ago

Ans - D when the model is overfitted means want to increase the AUC, we always perform hyperparameter tuning, Increase regularisations, decrease input feature parameters etc.

upvoted 1 times

✉  **nitinbhatia** 3 years, 3 months ago

AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values. So answer shall be A

<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=en>

upvoted 2 times

✉  **arghya13** 3 years, 3 months ago

Definitely not D

<https://developers.google.com/machine-learning/crash-course/classification/check-your-understanding-roc-and-auc>

upvoted 3 times

✉  **saurabh1805** 3 years, 5 months ago

A for me, read below link for more details.

<https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568>

upvoted 4 times

Question #158

Topic 1

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

**Correct Answer: D**

*Community vote distribution*

C (100%)

✉  **[Removed]**  3 years, 10 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 38 times

✉️  **AzureDP900** 1 year ago

Thank you for detailed explanation. C is right  
upvoted 1 times

✉️  **rickywck** Highly Voted 3 years, 10 months ago

Should be C:

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

upvoted 11 times

✉️  **barnac1es** Most Recent 4 months, 1 week ago

**Selected Answer: C**

Security Compliance: This option aligns with your company's security policies, which prohibit public Internet access from Cloud Dataproc nodes. Placing the dependencies in a Cloud Storage bucket within your VPC security perimeter ensures that the data remains within your private network.

VPC Security: By placing the dependencies within your VPC security perimeter, you maintain control over network access and can restrict access to the necessary nodes only.

Dataproc Initialization Action: You can use a custom initialization action or script to fetch and install the dependencies from the secure Cloud Storage bucket to the Dataproc cluster nodes during startup.

By copying the dependencies to a secure Cloud Storage bucket and using an initialization action to install them on the Dataproc nodes, you can meet your security requirements while providing the necessary dependencies to your cluster.

upvoted 3 times

✉️  **knith66** 5 months, 3 weeks ago

**Selected Answer: C**

C is correct

upvoted 1 times

✉️  **charline** 10 months, 3 weeks ago

**Selected Answer: C**

C seems good

upvoted 1 times

✉️  **musumusu** 11 months, 2 weeks ago

Answer C,

It needs practical experience to understand this question. You create cluster with some package/software i.e dependencies such as python packages that you store in .zip file, then you save a jar file to run the cluster as an application such as you need java while running spark session and some config yaml file.

These dependencies you can save in bucket and can use to configure cluster from external window, sdk or api. without going into UI. Then you need to use VPC to access these files

upvoted 2 times

✉️  **zellick** 1 year, 2 months ago

**Selected Answer: C**

C is the answer.

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#and\\_vpc-sc\\_networks](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#and_vpc-sc_networks)

With VPC Service Controls, administrators can define a security perimeter around resources of Google-managed services to control communication to and between those services.

upvoted 1 times

✉️  **DataEngineer\_WideOps** 1 year, 6 months ago

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 1 times

✉️  **medeis\_jar** 2 years ago

**Selected Answer: C**

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create\\_a\\_cloud\\_dataproc\\_cluster\\_with\\_internal\\_ip\\_address\\_only](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only)

upvoted 2 times

✉  **Prabusankar** 2 years, 1 month ago

When creating a Dataproc cluster, you can specify initialization actions in executables or scripts that Dataproc will run on all nodes in your Dataproc cluster immediately after the cluster is set up. Initialization actions often set up job dependencies, such as installing Python packages so that jobs can be submitted to the cluster without having to install dependencies when the jobs are run

upvoted 3 times

✉  **JG123** 2 years, 2 months ago

Correct: C

upvoted 1 times

✉  **clouditis** 3 years, 5 months ago

c it is!

upvoted 2 times

✉  **Rajokkiyam** 3 years, 10 months ago

Should be C

upvoted 2 times

✉  **[Removed]** 3 years, 10 months ago

Should be C

upvoted 2 times

✉  **jvg637** 3 years, 10 months ago

I think the correct answer might be C instead, due to [https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create\\_a\\_cloud\\_dataproc\\_cluster\\_with\\_internal\\_ip\\_address\\_only](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only)

upvoted 4 times

Question #159

Topic 1

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
- B. Cloud Bigtable
- C. Cloud Spanner
- D. Cloud Datastore

**Correct Answer: C**

*Community vote distribution*

A (50%)

C (50%)

✉  **[Removed]**  3 years, 10 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 32 times

- ✉  **Rajuuu** 3 years, 6 months ago  
C:- Cloud SQL is not fully managed as that is one of the requirement.  
upvoted 6 times
- ✉  **hightech** 1 year, 2 months ago  
Cloud Sql is a fully managed service by Google  
upvoted 5 times
- ✉  **zxing233** 3 years, 4 months ago  
<https://cloud.google.com/sql> it is fully managed  
upvoted 2 times
- ✉  **dem2021** 2 years, 8 months ago  
Have you really worked on GCP?  
upvoted 14 times
- ✉  **AzureDP900** 1 year ago  
A. Cloud SQL There is no need of Spanner  
upvoted 1 times
- ✉  **google\_learner123** 3 years, 5 months ago  
C - CloudSQL does not scale automatically.  
upvoted 8 times
- ✉  **zxing233** 3 years, 4 months ago  
Cloud SQL can automatically scale up storage capacity when you are near your limit  
upvoted 5 times
- ✉  **dmzr** 1 year, 3 months ago  
it does not say about type of scaling, Cloud SQL scale up automatically with storage, that should works  
upvoted 4 times
- ✉  **[Removed]**  3 years, 10 months ago  
Should be C.  
upvoted 13 times
- ✉  **[Removed]** 3 years, 10 months ago  
May be A  
upvoted 8 times
- ✉  **arturido**  2 months ago  
**Selected Answer: A**  
"Able to scale up to 6 TB" -seems to be the key  
it looks like autoscaling is related to storage - possible in case of Cloud SQL  
upvoted 1 times
- ✉  **LaxmanTiwari** 1 month, 1 week ago  
no way , u can automatically scale the Cloud SQL , please read the documents of Cloud SQL, Spanner is the solution .  
upvoted 1 times
- ✉  **tibuenoc** 2 months, 1 week ago  
**Selected Answer: C**  
Spanner is consistent and fully-managed  
<https://cloud.google.com/spanner/docs/transactions?hl=en>  
upvoted 1 times
- ✉  **DataFrame** 2 months, 1 week ago  
**Selected Answer: A**  
A seems to be correct because of the scaling factor of 6 TB because cloud sql easily supports up to 40 TB and obviously there is a limitation GLOBALLY MULTI REGIONAL which is nothing to do with question. Hence A seems more closer.  
upvoted 1 times

👤 **Nirca** 3 months ago

**Selected Answer: C**

Guys - this is 1000% C. "....automatically scale up" Cloud SQL needs restart! this is not the solution. Only Spanner (and BQ) are true automatically scale up

upvoted 2 times

👤 **Zepopo** 1 month, 4 weeks ago

What problem in restart? There are no any constraints about this, but there directly writes: "Fully managed"

upvoted 1 times

👤 **kcl10** 3 months, 4 weeks ago

**Selected Answer: C**

C: Cloud Spanner

Why not A: Cloud SQL?

No "automatically" scale up feature

upvoted 1 times

👤 **barnacles** 4 months, 1 week ago

**Selected Answer: C**

Fully Managed: Cloud Spanner is a fully managed database service provided by Google Cloud, which means you don't have to worry about managing infrastructure, updates, or backups.

Automatic Scaling: Cloud Spanner can automatically scale both horizontally and vertically to handle increased workloads and data volume. It handles databases ranging from a few gigabytes to multi-terabyte scale.

Transactionally Consistent: Cloud Spanner offers strong transactional consistency, making it suitable for applications that require ACID compliance.

SQL Querying: You can query Cloud Spanner databases using SQL, which is a familiar query language for many developers and analysts.

Given your requirements, Cloud Spanner is designed to meet the need for a fully managed, scalable, transactionally consistent, and SQL-accessible database solution.

upvoted 1 times

👤 **Xubaca** 5 months ago

**Selected Answer: C**

Able to automatically scale up = Memory and CPU

Able to scale up to 6 TB = Disk

Correct is C, because Cloud SQL no scale Memory and CPU.

upvoted 1 times

👤 **NeoNitin** 6 months ago

sahi A hai. Dekho mitra mujhe bhi pahale yahi laga cloud spanner hi hoga but read the question again.

scaleup required till 6TB aur The requirements does not mention high-availability,

and Cloud SQL is up to 64TB now

[https://cloud.google.com/sql/docs/quotas#storage\\_limits](https://cloud.google.com/sql/docs/quotas#storage_limits)

to phir jab kaam jara sa hai to spanner ko kyo laaye aur cloud sql bhi autoscaleup hota hai

[https://www.tutorialspoint.com/differences\\_between\\_google\\_cloud\\_sql\\_and\\_cloud\\_spanner#:~:text=Cloud%20Spanner%20is%20compatible%20with,an%20extra%20large%20database%20instances.&text=The%20price%20is%20comparatively%20less,of%20Cloud%20Spanner%20is%20high.](https://www.tutorialspoint.com/differences_between_google_cloud_sql_and_cloud_spanner#:~:text=Cloud%20Spanner%20is%20compatible%20with,an%20extra%20large%20database%20instances.&text=The%20price%20is%20comparatively%20less,of%20Cloud%20Spanner%20is%20high.)

upvoted 1 times

👤 **wan2three** 6 months, 2 weeks ago

**Selected Answer: A**

Cloud SQL can also automatically scale up storage capacity when you are near your limit.

[https://cloud.google.com/sql/?utm\\_source=google&utm\\_medium=cpc&utm\\_campaign=japac-HK-all-en-dr-BKWS-all-super-trial-PHR-dr-1605216&utm\\_content=text-ad-none-none-DEV\\_c-CRE\\_652279903175-ADGP\\_Hybrid%20%7C%20BKWS%20%20BRO%20%7C%20Txt%20~%20Databases\\_Cloud%20SQL\\_gcp%20horizontal%20sql\\_main-KWID\\_43700076522535474-kwd-2005739330106&userloc\\_2344-network\\_g&utm\\_term=KW\\_google%20sql%20horizontal&gclid=Cj0KCQjwqs6IBhCxARIsAG8YcDjWaS5NbG3remUHnHQ7EFK-wJNsF0I\\_lvePKHF0mHaiBK3\\_-eM7Z-UaAqQeEALw\\_wcB&gclsrc=aw.ds#section-2:~:text=Cloud%20SQL%20can%20also%20automatically%20scale%20up%20storage%20capacity%20when%20you%20are%20near%20ur%20limit.](https://cloud.google.com/sql/?utm_source=google&utm_medium=cpc&utm_campaign=japac-HK-all-en-dr-BKWS-all-super-trial-PHR-dr-1605216&utm_content=text-ad-none-none-DEV_c-CRE_652279903175-ADGP_Hybrid%20%7C%20BKWS%20%20BRO%20%7C%20Txt%20~%20Databases_Cloud%20SQL_gcp%20horizontal%20sql_main-KWID_43700076522535474-kwd-2005739330106&userloc_2344-network_g&utm_term=KW_google%20sql%20horizontal&gclid=Cj0KCQjwqs6IBhCxARIsAG8YcDjWaS5NbG3remUHnHQ7EFK-wJNsF0I_lvePKHF0mHaiBK3_-eM7Z-UaAqQeEALw_wcB&gclsrc=aw.ds#section-2:~:text=Cloud%20SQL%20can%20also%20automatically%20scale%20up%20storage%20capacity%20when%20you%20are%20near%20ur%20limit.)

upvoted 1 times

✉  **vaga1** 7 months, 1 week ago

**Selected Answer: A**

The requirements does not mention high-availability, and Cloud SQL is up to 64TB now  
[https://cloud.google.com/sql/docs/quotas#storage\\_limits](https://cloud.google.com/sql/docs/quotas#storage_limits)

upvoted 1 times

✉  **WillemHendr** 7 months, 3 weeks ago

I feel this is a 'fit-for-purpose' question, and TB's large Cloud-SQL is enough. The up-to-6TB is a sign for me to not go for Cloud-Spanner.  
upvoted 1 times

✉  **Adswerve** 9 months, 2 weeks ago

**Selected Answer: C**

C Spanner

Spanner scales automatically:

<https://cloud.google.com/blog/products/databases/cloud-database-scales-instance-sizes-easily/>  
"Automatically right-size Spanner instances with the new Autoscaler"

Cloud SQL doesn't scale automatically. Here's a tutorial by a Google employee which shows manual steps to scale Cloud SQL:  
<https://cloud.google.com/community/tutorials/horizontally-scale-mysql-database-backend-with-google-cloud-sql-and-proxysql>

upvoted 1 times

✉  **streeeber** 9 months, 2 weeks ago

**Selected Answer: C**

I would say C, because A does not offer autoscaling.

upvoted 2 times

✉  **juliosb** 10 months ago

**Selected Answer: A**

People here are rejecting Cloud SQL because it does not scale up CPU and MEM automatically, but Cloud Spanner also don't (not without hosting the Autoscaler tool yourself).

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Answer A: CloudSQL

Don't confuse anymore,

SQL needs, Cloud sql or Spanner.

nothing mentioned that data should be highly available or scalable or increasing in future So, don't go with Spanner.

Ideal use case, Cloud Sql can handle UP TO 30TB.

Other two options are no-sql so don't go.

upvoted 2 times

Question #160

Topic 1

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
- B. Cloud Bigtable
- C. Cloud Spanner

## D. Cloud Datastore

### Correct Answer: A

Community vote distribution

A (64%)

C (36%)

✉ **jvg637** Highly Voted 3 years, 10 months ago

A. Cloud SQL (30TB)  
upvoted 32 times

✉ **dagoat** 2 years, 4 months ago

65 TB now in Sept 2021  
upvoted 14 times

✉ **[Removed]** 2 years ago

[https://cloud.google.com/sql/docs/quotas#storage\\_limits](https://cloud.google.com/sql/docs/quotas#storage_limits)  
upvoted 1 times

✉ **vindahake** 3 years, 8 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

✉ **odacir** 1 year, 1 month ago

[https://cloud.google.com/sql/docs/quotas#storage\\_limits](https://cloud.google.com/sql/docs/quotas#storage_limits)  
64TB  
upvoted 2 times

✉ **Gcpyspark** 3 years, 1 month ago

Sure, however in future if the capacity grows beyond 30 TB then Cloud SQL won't work right then Spanner would be the option?  
upvoted 2 times

✉ **desertlotus1211** 1 year ago

you can always call GCP to add quota... Spanner is for global reach, ideally...  
upvoted 3 times

✉ **Rajuuu** Highly Voted 3 years, 6 months ago

A as limit is now 30 TB for Cloud SQL  
upvoted 6 times

✉ **drpay** Most Recent 3 months ago

**Selected Answer: C**

two keywords: Transactional data, 20 TB  
upvoted 2 times

✉ **barnacles** 4 months, 1 week ago

**Selected Answer: C**

Scalability: Cloud Spanner is designed to handle large volumes of data, making it suitable for a 20 TB database. It can scale horizontally and vertically to accommodate growing data needs.

Global Distribution: Cloud Spanner allows you to distribute data globally for low-latency access across regions, which can be advantageous for operational systems.

Strong Consistency: It provides strong transactional consistency, which is important for operational systems that require ACID compliance.

SQL Support: Cloud Spanner supports SQL, which is a familiar query language for developers.

While Cloud SQL, Cloud Bigtable, and Cloud Datastore have their use cases, Cloud Spanner is better suited for larger databases with strong consistency requirements, making it a suitable choice for migrating a 20 TB operational system database to GCP.

upvoted 2 times

ashu381 4 months, 2 weeks ago

Cloud SQL, upto 64 TB now, you can always call GCP for increasing the quota though !!

upvoted 1 times

vaga1 7 months, 1 week ago

**Selected Answer: A**

Cloud SQL is generally better for OLTP, and Cloud SQL is up to 64 TB now.

[https://cloud.google.com/sql/docs/quotas#storage\\_limits](https://cloud.google.com/sql/docs/quotas#storage_limits)

upvoted 2 times

vaga1 8 months, 3 weeks ago

"move its operational system transaction data from an on-premises database to GCP". Cloud SQL may be plug-and-play

upvoted 2 times

musumusu 11 months, 2 weeks ago

Not 100% in favour of A, Should i recommend my client Cloud SQL, when they are coming to me with 20TB already 30TB is limit, its transactional data, which i can't compromise. I will propose cloud spanner. There is nothing mentioned that they want to save cost.

upvoted 1 times

AzureDP900 1 year ago

A. Cloud SQL

upvoted 1 times

zelick 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/sql/docs/features#features>

Up to 64 TB of storage available, with the ability to automatically increase storage size as needed.

upvoted 4 times

Jay\_Krish 1 year, 2 months ago

**Selected Answer: A**

With the given requirements A. Cloud SQL is more than sufficient. Don't try to overthink scenarios like what if it grows.. what if there's additional requirement in future.. what if this what if that.. just look at the question and see the stated requirement. If there are more than one answer try see which is simple and doesn't come with extra frills.

upvoted 1 times

Atnafu 1 year, 2 months ago

A

65 TB now in Nov 2022

upvoted 2 times

WZH 1 year, 6 months ago

it is already 20 TB at the moment, and you probably want to change the database because the capacity of your current storage solution is not enough. Then you decide to change it to Cloud SQL(up to 30 TB) which may not increase much capacity? I am not sure about the answer but looks weird imho.

upvoted 1 times

Dan226 1 year, 6 months ago

Cloud SQL can store 64 Tb, but in the initial setup the operation are 20tb. It will reach the limitation soon if you choose Cloud SQL

upvoted 1 times

gcp\_k 2 years, 3 months ago

Depends.. I mean, C is correct if the exam is not updated. A is correct if the exam is updated. So ... kinda in catch 22 situation ...

upvoted 3 times

KokkiKumar 2 years, 3 months ago

Hi everyone, Can I purchase this exam? Is it worthwhile?

upvoted 2 times

Alasmindas 3 years, 2 months ago

Option A - Cloud SQL is the correct answer. Cloud SQL can store up to 30 TB.

<https://cloud.google.com/sql/docs/quotas#:~:text=Cloud%20SQL%20storage%20limits&text=Up%20to%2030%2C720%20GB%2C%20depending,for%20PostgreSQL%20or%20SQL%20Server>

upvoted 4 times

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
- 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.

**Correct Answer: C***Community vote distribution*

C (78%)

D (22%)

✉  **psu** Highly Voted 3 years, 9 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 Row size can be big but are not infinite).

[https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns\\_for\\_row\\_key\\_design](https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design)  
upvoted 33 times

✉  **AzureDP900** 1 year ago

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  
upvoted 1 times

✉  **nadavw** 1 year, 7 months ago

there is a limit of 60 columns per row according to question. in addition in D the cost will be a lower which is a requirement. so D seems more suitable.  
upvoted 1 times

✉  **madhu1171** Highly Voted 3 years, 10 months ago

C correct answer  
upvoted 19 times

👤 **barnacles** Most Recent 4 months, 1 week ago

**Selected Answer: C**

Scalability: Bigtable can handle large-scale data efficiently, making it suitable for storing time series data for millions of computers.

Low Latency: Bigtable provides low-latency access to data, which is crucial for real-time analytics.

Flexible Schema: The narrow table design allows you to efficiently store and query time series data without specifying all possible columns in advance, providing flexibility for future growth.

Column Families: Bigtable supports column families, allowing you to organize data logically.

Row Key Design: Combining the computer identifier with the sample time at each second in the row key allows for efficient retrieval of data for specific computers and time intervals.

Analytics: While Bigtable does not support SQL directly, it allows for efficient data retrieval and can be integrated with other tools for analytics

upvoted 1 times

👤 **WillemHendr** 7 months, 3 weeks ago

**Selected Answer: D**

"..and ensure that the schema design will allow for future growth of the dataset":

<https://cloud.google.com/bigtable/docs/schema-design-time-series#time-buckets>

"Data stored in this way is compressed more efficiently than data in tall, narrow tables."

I read the "future growth" as a sign to be effective in storage, and go for the Time-Buckets.

upvoted 2 times

👤 **zellck** 1 year, 2 months ago

**Selected Answer: C**

C is the answer.

<https://cloud.google.com/bigtable/docs/schema-design-time-series>

upvoted 2 times

👤 **Remi2021** 1 year, 4 months ago

**Selected Answer: C**

time series = narrow table

upvoted 2 times

👤 **\_8008\_** 1 year, 9 months ago

What about "avoid being charged for every query executed"? Nothing on this topic in here <https://cloud.google.com/bigtable/docs/schema-design-time-series> can anyone comment?

upvoted 3 times

👤 **medeis\_jar** 2 years ago

**Selected Answer: C**

Narrow and tall table for a single event and good for time-series data

Short and Wide table for data over a month, multiple events

upvoted 2 times

👤 **JG123** 2 years, 2 months ago

Correct: C

upvoted 2 times

👤 **squishy\_fishy** 2 years, 3 months ago

Answer is C.

Bigtable is best suited to the following scenarios: time-series data (e.g. CPU and memory usage over time for multiple servers), financial data (e.g. transaction histories, stock prices, and currency exchange rates), and IoT (Internet of Things) use cases.

<https://www.xplenty.com/blog/bigtable-vs-bigquery/>

upvoted 3 times

👤 **safiyu** 2 years, 5 months ago

C is the correct answer. If you consider wide table, then 60 columns for cpu usage and 60 columns for memory usage. in future, if you need to add a new kpi to the table, then the schema changes. you will have to add 60 more columns for the new feature. this is not so future proof.. so is out of the picture.

upvoted 5 times

✉  **DeepakS227** 2 years, 6 months ago

BQ is optimized for large-scale, ad-hoc SQL-based analysis. I Think it should be A  
upvoted 1 times

✉  **koupayio** 2 years, 7 months ago

First C & D won't cause hotspotting as computer\_identifier is first part of row key  
I prefer D because "ensure that the schema design will allow for future growth of the dataset."  
C is too tall and narrow I cannot see schema design grow in the future  
upvoted 1 times

✉  **crslake** 2 years, 8 months ago

D, Better overall, harder to implement (but that is not stated as a constraint)  
<https://cloud.google.com/bigtable/docs/schema-design-time-series#time-buckets>  
upvoted 1 times

✉  **lollo1234** 2 years, 7 months ago

How do you store both CPU and memory usage though? Two sets of 60 columns per row? I am wondering if that goes along with "the schema design will allow for future growth"...what if by future growth they mean monitoring N more metrics. That would imply N\*60 columns right?  
upvoted 2 times

✉  **Sumanth09** 2 years, 10 months ago

Should be A  
question did not talk about latency  
without query cost -- BigQuery Cache  
flexible schema - BigQuery (nested and repeated)  
upvoted 6 times

✉  **VM\_GCP** 3 years, 1 month ago

BigQuery cannot be the answer as  
[1] This will not save the price by caching as query will not always be same  
<https://cloud.google.com/bigquery/docs/cached-results#limitations>  
[2] And the BigTable is always the correct solution for this use case as given in  
<https://cloud.google.com/bigtable/docs/overview#what-its-good-for>

BigTable has native time series support and is preferred for large analytical and operational work  
Answer can be C or D.

But problem is both the options will cause hot spotting. So not sure which to select  
upvoted 2 times

✉  **VM\_GCP** 3 years, 1 month ago

I will go for C  
upvoted 2 times

✉  **Prakzz** 3 years, 6 months ago

both C and D row key will cause hotspotting  
upvoted 3 times

✉  **Aman47** 1 month, 2 weeks ago

No, it won't, because preceding timestamp with identifier values helps in a tradeoff between randomised row keys and manifesting groupings of similar entities.  
upvoted 1 times

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.
- 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.

**Correct Answer: B**

*Community vote distribution*

A (62%)

D (38%)

✉️  **dhs227** Highly Voted 3 years, 10 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 40 times

✉️  **mikey007** 3 years, 5 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.

upvoted 4 times

✉️  **[Removed]** 2 years ago

The trust no one design philosophy requires that the keys for encryption should always be, and stay, in the hands of the user that applies them. This implies that no external party can access the encrypted data (assumed that the encryption is strong enough).

[https://en.wikipedia.org/wiki/Trust\\_no\\_one\\_\(Internet\\_security\)](https://en.wikipedia.org/wiki/Trust_no_one_(Internet_security))

upvoted 3 times

✉️  **[Removed]** Highly Voted 3 years, 10 months ago

Answer: A

Description: AAD is used to decrypt the data so better to keep it outside GCP for safety

upvoted 15 times

✉️  **emmylou** Most Recent 2 months, 1 week ago

I just cannot understand this question. If you can't trust the provider, in this case Google, then how can you use the KMS approach. In my mind you have to generate the key locally and upload but I'm clearly wrong and don't get why.

upvoted 1 times

✉️  **shanwfard** 4 months ago

**Selected Answer: D**

IMO must be (D) : to reach TNO goal keys must be customer supplied.

upvoted 2 times

✉  **barnac1es** 4 months, 1 week ago

**Selected Answer: D**

Customer-Supplied Encryption Key (CSEK): CSEK allows you to provide your encryption keys, ensuring that the cloud provider staff does not have access to the keys and cannot decrypt your data.

Separate Project for Key Management: Saving the CSEK in a different project that only the security team can access adds an additional layer of security. It isolates the encryption keys from the project where the data is stored, ensuring that even within the same cloud provider, only authorized personnel can access the keys.

Use of .boto Configuration: Specifying the CSEK in the .boto configuration file ensures that it is applied consistently when interacting with Cloud Storage through tools like gsutil. This way, every archival file is encrypted using your keys.

Options A and B involve using Google Cloud Key Management Service (KMS) to manage keys, which does not align with the TNO approach because cloud provider staff could potentially access the keys stored in Google Cloud KMS.

upvoted 1 times

✉  **[Removed]** 4 months, 2 weeks ago

**Selected Answer: A**

The answer is A

The question tells us that "prevent the cloud provider staff from decrypting", so we cannot keep anything that helps decrypt on GCP, not even a different project. so the answer cannot be D.

upvoted 3 times

✉  **NewDE2023** 5 months, 4 weeks ago

**Selected Answer: D**

CSEKs are used when an organization needs complete control over key management.

upvoted 3 times

✉  **tavva\_prudhvi** 6 months, 1 week ago

Option A is not the best choice for the "Trust No One" (TNO) approach because it involves using Google Cloud's Key Management Service (KMS) to create and manage encryption keys. This means that the cloud provider will have access to the keys, which could potentially enable their staff to decrypt the data.

upvoted 2 times

✉  **midgoo** 10 months, 3 weeks ago

**Selected Answer: A**

D may work, but 'Trust No One' = do not trust GCP too. So D cannot be the answer.

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

answer A: KMS + AAD is more secure than CSEK

upvoted 2 times

✉  **zellick** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/kms/docs/additional-authenticated-data>

Additional authenticated data (AAD) is any string that you pass to Cloud Key Management Service as part of an encrypt or decrypt request. AAD is used as an integrity check and can help protect your data from a confused deputy attack. The AAD string must be no larger than 64 KiB.

Cloud KMS will not decrypt ciphertext unless the same AAD value is used for both encryption and decryption.

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.

upvoted 6 times

✉  **AzureDP900** 1 year ago

Agree with A

upvoted 2 times

✉  **Jay\_Krish** 1 year, 2 months ago

**Selected Answer: D**

CSEK with only security team having access seems to be right approach. Not sure how A can be better.

upvoted 3 times

✉ **cloudmon** 1 year, 2 months ago

**Selected Answer: A**

It's A, because you cannot decrypt the ciphertext unless you know the AAD (<https://cloud.google.com/kms/docs/additional-authenticated-data>)  
upvoted 2 times

✉ **deavid** 1 year, 3 months ago

**Selected Answer: A**

Answer: A

upvoted 1 times

✉ **clouditis** 1 year, 4 months ago

D it is

upvoted 1 times

✉ **DataEngineer\_WideOps** 1 year, 6 months ago

**Selected Answer: A**

C can not be the answer since memorystore cant be used to save CSEK key.

[https://cloud.google.com/memorystore/docs/redis/cmek#when\\_does\\_memorystore\\_interact\\_with\\_cmek\\_keys](https://cloud.google.com/memorystore/docs/redis/cmek#when_does_memorystore_interact_with_cmek_keys)

A is the Answer.

upvoted 2 times

✉ **BigDataBB** 1 year, 11 months ago

**Selected Answer: A**

A because: "keep the AAD outside of Google Cloud"

upvoted 2 times

Question #163

*Topic 1*

You have data pipelines running on BigQuery, Dataflow, and 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 or features of the platform. What should you do?

- A. Export the information to Cloud Monitoring, and set up an Alerting policy
- B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Cloud Monitoring
- 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

**Correct Answer: B**

*Community vote distribution*

A (100%)

✉  **John\_Pongthorn** Highly Voted  1 year, 4 months ago

**Selected Answer: A**

A . Your preference is to use managed products or features of the platform  
upvoted 5 times

✉  **barnac1es** Most Recent  4 months, 1 week ago

**Selected Answer: A**

Cloud Monitoring (formerly known as Stackdriver) is a fully managed monitoring service provided by GCP, which can collect metrics, logs, and other telemetry data from various GCP services, including BigQuery, Dataflow, and Dataproc.

**Alerting Policies:** Cloud Monitoring allows you to define alerting policies based on specific conditions or thresholds, such as pipeline failures, latency spikes, or other custom metrics. When these conditions are met, Cloud Monitoring can trigger notifications (e.g., emails) to alert the team managing the pipelines.

**Cross-Project Monitoring:** Cloud Monitoring supports monitoring resources across multiple GCP projects, making it suitable for your requirement to monitor pipelines in multiple projects.

**Managed Solution:** Cloud Monitoring is a managed service, reducing the operational overhead compared to running your own virtual machine instances or building custom solutions.

upvoted 2 times

✉  **sergiomujica** 4 months, 3 weeks ago

**Selected Answer: A**

use managed products  
upvoted 1 times

✉  **whorillo** 11 months ago

**Selected Answer: A**

Should be A  
upvoted 2 times

✉  **pluiedust** 1 year, 4 months ago

**Selected Answer: A**

Should be A  
upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: A**

A. Export the information to Cloud Monitoring, and set up an Alerting policy  
upvoted 2 times

✉  **PhuocT** 1 year, 4 months ago

**Selected Answer: A**

Should be A  
upvoted 1 times

Question #164

Topic 1

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.
- 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.

**Correct Answer: D***Community vote distribution*

B (57%)

D (43%)

**cajica** Highly Voted 11 months, 4 weeks 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 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 8 times

**sergiomujica** 4 months, 3 weeks ago

I think it should say city instead of state... it is a typoo in the transcription of the question

upvoted 2 times

**knith66** 6 months, 1 week 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

**cetanx** 8 months ago

But also for D, assigning each city to a numbered region could lose important information, as cities within the same region might have different characteristics affecting customer purchasing behavior (from Chat GPT).

upvoted 1 times

**MaxNRG** Most Recent 1 month, 1 week ago**Selected Answer: B**

One-hot encoding is a common technique used to handle categorical data in machine learning. This approach will transform the city name variable into a series of binary columns, one for each city. Each row will have a "1" in the column corresponding to the city it represents and "0" in all other city columns. This method is effective for linear regression models as it enables the model to use city data as a series of numeric, binary variables. BigQuery supports SQL operations that can easily implement one-hot encoding, thus minimizing the amount of coding required and efficiently preparing the data for the model.

upvoted 2 times

**MaxNRG** 1 month, 1 week ago

A removes the city information completely, losing a key predictive component.

C requires additional coding and infrastructure with TensorFlow and vocabulary files outside of what BigQuery already provides.

D transforms the distinct city values into numeric regions, losing granularity of the city data.

By using SQL within BigQuery to one-hot encode cities into multiple yes/no columns, the city data is maintained and formatted appropriately for the BigQuery ML linear regression model with minimal additional coding. This aligns with the requirements stated in the question.

upvoted 2 times

**MaxNRG** 1 month, 1 week ago

[https://cloud.google.com/bigquery/docs/auto-preprocessing#one\\_hot\\_encoding](https://cloud.google.com/bigquery/docs/auto-preprocessing#one_hot_encoding)

upvoted 1 times

**barnacles** 4 months, 1 week ago**Selected Answer: B**

One-Hot Encoding: One-hot encoding is a common technique for handling categorical variables like city names in machine learning models. It transforms categorical data into a binary matrix, where each city becomes a separate column with binary values (0 or 1) indicating the presence or absence of that city.

Least Amount of Coding: One-hot encoding in BigQuery is straightforward and can be accomplished with SQL. You can use SQL expressions to pivot the city names into separate columns and assign binary values based on the city's presence in the original data.

Predictive Power: One-hot encoding retains the predictive power of city information while making it suitable for linear regression models, which require numerical input.

upvoted 3 times

knith66 6 months, 1 week ago

Selected Answer: B

One hot encoding for state and binary values for each city will allow me to choose the B option.  
upvoted 2 times

tavva\_prudhvi 6 months, 1 week ago

I guess Option D loses the granularity of the city-level information, as multiple cities will be grouped into the same region and represented by 1 same number. This can result in a loss of important predictive information for your linear regression model.

On the other hand, if we use one-hot encoding to create binary columns for each city. This method preserves the city-level information, allowing the model to capture the unique effects of each city on the likelihood of purchasing your company's products. Additionally, it can be done directly in BigQuery using SQL, which requires less coding and is more efficient.

upvoted 2 times

blathul 7 months, 1 week ago

Selected Answer: B

One-hot encoding is a common technique used to represent categorical variables as binary columns. In this case, you can transform the city variable into multiple binary columns, with each column representing a specific city. This allows you to maintain the predictive city information while organizing the data in columns suitable for training and serving the linear regression model.

By using SQL in BigQuery, you can perform the necessary transformations to implement one-hot encoding.  
upvoted 3 times

KC\_go\_reply 7 months, 1 week ago

Selected Answer: B

- A is wrong since it drops the city which is a key predictor.  
- C is wrong since we want to keep it simple, and not use Tensorflow here.  
- D is wrong since there is no specific reason to use Data Fusion, and also this encoding here is ordinal, which doesn't make sense for some non-quantitative such as cities - we want one-hot coding instead.

Therefore, B must be the correct answer.

upvoted 2 times

ckanaar 4 months, 1 week ago

It could be argued that a specific reason to use Data Fusion is the minimal coding requirement.  
upvoted 1 times

leandr0s 7 months, 2 weeks ago

Selected Answer: D

Cloud Datafusion: least amount of coding  
upvoted 3 times

knith66 6 months, 1 week ago

OHE is better than datafusion considering least amount coding  
upvoted 1 times

tavva\_prudhvi 6 months, 1 week ago

While it's true that Cloud Data Fusion can simplify data integration tasks with a visual interface, it might not be the best choice in this specific scenario as using Cloud Data Fusion to assign each city to a region might result in a loss of important predictive information due to the grouping of cities  
upvoted 1 times

vaga1 8 months, 3 weeks ago

Selected Answer: B

A doesn't include the city column.  
C is not low code.  
D is not a one hot encoding, but an ordinal one on the city column.

B applies a one hot encoding on the state column and a binary encoding on the city column, which works for me.  
upvoted 1 times

mialli 9 months ago

Selected Answer: B

[https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one\\_hot\\_encoding](https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding)  
upvoted 1 times

✉  **juliosb** 10 months, 1 week ago

**Selected Answer: D**

D uses the least amount of coding... even if the model is not good.  
B encodes the "state", not the "city".

upvoted 4 times

✉  **dconesoko** 1 year ago

**Selected Answer: B**

Manually bigquery ml does preprocessing for you however if one wants to do a manual processing one can use the ML.ONE\_HOT\_ENCODE function. It just acts as an analytical function.

upvoted 2 times

✉  **zellck** 1 year, 2 months ago

**Selected Answer: B**

B is the answer.

[https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one\\_hot\\_encoding](https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding)  
One-hot encoding maps each category that a feature has to its own binary feature where 0 represents the absence of the feature and 1 represents the presence (known as a dummy variable) creating N new feature columns where N is the number of unique categories for the feature across the training table.

upvoted 3 times

✉  **ovokpus** 1 year, 2 months ago

**Selected Answer: B**

The Cloud Data Fusion method will add unnecessary weights to categories with higher value labels, which will skew the model. The best practice for encoding nominal categorical data is to one-hot-encode them into binary values. That is conveniently done in BigQuery:

[https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one\\_hot\\_encoding](https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding)  
upvoted 4 times

✉  **Atnafu** 1 year, 2 months ago

D

Cloud Data Fusion is a fully managed, code-free data integration service that helps users efficiently build and manage ETL/ELT data pipelines  
upvoted 1 times

✉  **dconesoko** 1 year ago

Does it come with an out of the box one hot encoding template ?

upvoted 1 times

✉  **NicolasN** 1 year, 2 months ago

**Selected Answer: D**

I have the same feeling with @cloudmon so I compromised to answer [D].  
In more detail, here is my reasoning:

The requirement "maintaining the predictable variables" (a.k.a. city) makes:

[A] obviously invalid

[B] invalid, since it broadens the prediction to be state-dependent (all cities in particular state will be treated as the same variable). Additionally, one-hot encoding is not suitable for linear regression problems, dummy encoding (drop one) is better.

Answer [C] doesn't satisfy the "least amount of coding" directive. Other than that (as far I understood by searching the keyword `tf.feature_column.categorical_column_with_vocabulary_list`) the TensorFlow vocabulary list is another form of one-hot encoding.

So it remains [D] which offers a visual interface but uses ordinal (or label) encoding which is far from ideal for regression problems.  
upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

**Selected Answer: D**

This is one of those examples in which none of the answers are actually the right way to do it, but D is the only one that may make some sense.  
B is wrong because it talks about transforming the STATE column. The actual correct way would be to transform/one-hot encode the CITY column (thus creating binary columns representing the city).

upvoted 3 times

✉  **cloudmon** 1 year, 2 months ago

D does have the ordinality problem, but at least it's actually something that's possible/relevant to do.

upvoted 1 times

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

B (77%)

D (20%)

2%

  **deavid** Highly Voted 1 year, 3 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](https://cloud.google.com/spanner/docs/transactions#read-write_transactions)

upvoted 10 times

  **Jay\_Krish** 1 year, 2 months ago

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

upvoted 10 times

  **AzureDP900** 1 year ago

B is right

upvoted 1 times

  **juliobs** Highly Voted 10 months, 1 week ago

**Selected Answer: B**

"locations throughout North America" implies multi-region (northamerica-northeast1, us-central1, us-south1, us-west4, us-east5, etc.) Cloud SQL can only do read replicas in other regions.

upvoted 8 times

  **FP77** 5 months, 2 weeks ago

Read replicas are enough to make Cloud SQL work as a multi-region service. That's not the point. The point is that the answer introduces use of Bigquery when it's not needed for the use case. That's why B is right.

upvoted 3 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: B**

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

Since the banking transaction system requires ACID compliance and SQL access to the data, Cloud Spanner is the most appropriate solution. Unlike Cloud SQL, Cloud Spanner natively provides ACID transactions and horizontal scalability.

Enabling stale reads in Spanner (option A) would reduce data consistency, violating the ACID compliance requirement of banking transactions.

BigQuery (option C) does not natively support ACID transactions or SQL writes which are necessary for a banking transactions system.

Cloud SQL (option D) provides ACID compliance but does not scale horizontally like Cloud Spanner can to handle large transaction volumes.

By using Cloud Spanner and specifically locking read-write transactions, ACID compliance is ensured while providing fast, horizontally scalable processing of banking transactions.

upvoted 1 times

Aman47 1 month, 2 weeks ago

**Selected Answer: B**

Spanner is an enterprise level resource which Banks require, Cloud SQL is limited to 30TB of storage. And Banking transactions should be read/write locked.

upvoted 1 times

barnacles 4 months, 1 week ago

**Selected Answer: B**

ACID Compliance: Cloud Spanner is a globally distributed, strongly consistent database service that offers ACID compliance, making it a suitable choice for handling bank account transactions where data consistency and integrity are crucial.

SQL Access: Cloud Spanner supports SQL queries, which align with your requirement to access data with SQL. You can use standard SQL to interact with the data stored in Cloud Spanner.

Locking Read-Write Transactions: Cloud Spanner allows you to perform locking read-write transactions, ensuring that transactions are executed in a serializable and consistent manner. This is essential for financial transactions to prevent conflicts and maintain data integrity.

upvoted 3 times

NeoNitin 6 months ago

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

Here's why:

ACID Compliance: ACID stands for Atomicity, Consistency, Isolation, and Durability. Cloud Spanner is a fully managed, globally distributed database that provides strong consistency and ACID compliance. This ensures that bank account transactions are processed reliably and accurately, avoiding issues like data corruption or incomplete transactions.

Ability to access data with SQL: Cloud Spanner supports SQL, which allows you to perform standard SQL queries on the data. This means that you can use familiar SQL commands to access, retrieve, and manipulate transaction data easily.

upvoted 1 times

Adswerve 9 months, 2 weeks ago

**Selected Answer: D**

I initially selected B. However, it might be D.

<https://cloud.google.com/blog/topics/developers-practitioners/your-google-cloud-database-options-explained>

Cloud Spanner: Cloud Spanner is an enterprise-grade, globally-distributed, and strongly-consistent database that offers up to 99.999% availability, built specifically to combine the benefits of relational database structure with non-relational horizontal scale. It is a unique database that combines ACID transactions, SQL queries, and relational structure with the scalability that you typically associate with non-relational or NoSQL databases. As a result, Spanner is best used for applications such as gaming, payment solutions, global financial ledgers, retail banking, and inventory management that require ability to scale limitlessly with strong-consistency and high-availability.

upvoted 1 times

musumusu 11 months, 1 week ago

Answer B:

locking read-write = for data accuracy  
state read = for speed up or latency

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Answer B: Spanner

It's incomplete question, what do you assume by large bank, until we are not sure about size and scale. Region is north america, that can be managed by cloud sql. but

i am going for spanner, as its large bank and transaction data.

upvoted 2 times

✉  **cajica** 11 months, 4 weeks ago

This is definitely a tricky question because both B and D are "appropriate" as the question suggests, of course we can make assumptions with the "large bank" sentence but there are other questions here where making assumptions is not accepted by the community so I wonder when we can make assumptions and when we can't. I think the real problem here is the ambiguous question. This is one of the few questions where the community accept that both (B and D) answers are appropriate but some comments (and I agree) argue the BEST approach is B. I really think some questions can be written in a better and non-ambiguous way, it's just about thinking a little bit more and not conforming when a poor spelling.

upvoted 5 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: B**

The question is hinting a requirement for global consistency, i.e. being available for NA region, which does not just include US but also Mexico, Argentina etc.

Large bank = priority over consistency over read-write

upvoted 6 times

✉  **ckanaar** 4 months, 1 week ago

Good catch, definitely Spanner in that case.

upvoted 1 times

✉  **desertlotus1211** 1 year ago

Argentina is South America...

upvoted 5 times

✉  **NicolasN** 1 year, 1 month ago

**Selected Answer: B**

Finally, it's [B].

There is no measurable requirement that rules out [D] (Cloud SQL) and this fact made me to select it as a preferable answer.

But since we are talking about a large bank (which normally implies massive reads/writes per sec.) and nobody has posed any cost limitation, real case I would definitely prefer the advantages of Spanner.

upvoted 1 times

✉  **zelliCK** 1 year, 2 months ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/spanner/docs/transactions>

Spanner supports these transaction modes:

- Locking read-write. This type of transaction is the only transaction type that supports writing data into Spanner. These transactions rely on pessimistic locking and, if necessary, two-phase commit. Locking read-write transactions may abort, requiring the application to retry.

upvoted 3 times

👤 **NicolasN** 1 year, 2 months ago

Selected Answer: D

- [A] No - Stale reads not accepted for bank account transactions. "A stale read is read at a timestamp in the past. If your application is latency sensitive but tolerant of stale data, then stale reads can provide performance benefits."
- [B] Yes - Fulfils all requirements
- [C] No - BigQuery is ACID-compliant, but it is too much to use it for such a case (mainly a CRUD app)
- [D] Yes+ - Fulfils all requirements. The BigQuery part may seem redundant, but it states a true fact that doesn't violate the "access data with SQL" requirement.

So, when SQL Cloud and SQL Spanner fit both, there is no reason to prefer the second.

And the question doesn't mention any obvious fact for which should we prefer the expensive SQL Spanner:

- We don't know if we have to deal with a big amount of data and thousands of writes per second.
- We don't know the database size.
- There is no need for multi-regional writes that would exclude SQL Cloud as an alternative. Is it a coincidence that the question limits the problem to the single region of North America?

upvoted 2 times

👤 **SuperVee** 10 months, 4 weeks ago

Also, correct me if I am wrong, Bigquery cannot query Cloud SQL directly, only when Cloud SQL is exported into GCS, then BQ can connect to GCS using federated queries.

upvoted 1 times

👤 **NicolasN** 1 year, 1 month ago

I changed my mind to [B] since I underestimated the given of a "large bank" where the cost difference for a single region Spanner wouldn't matter.

upvoted 1 times

👤 **zellick** 1 year, 1 month ago

North America has many regions, and the requirement is throughout North America, so Cloud Spanner will be more suitable to support many regions.

upvoted 3 times

👤 **NicolasN** 1 year, 1 month ago

You are correct of course. The final sentence was totally inaccurate.

upvoted 2 times

👤 **cloudmon** 1 year, 2 months ago

Selected Answer: B

I'd go for B.

The only other somewhat valid option is D, but there's no requirement for analytics in the question.

upvoted 1 times

👤 **mattab1627** 1 year, 3 months ago

Surely its B, transactional data at a large US based bank would surely be massive in size and probably too much for CloudSQL? There is also mention of a requirement for analytics

upvoted 1 times

👤 **MounicaN** 1 year, 4 months ago

why not spanner?

upvoted 2 times

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.
- 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.

**Correct Answer: B**

*Community vote distribution*

B (67%)

D (33%)

 **zellck**  1 year, 2 months 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.

In BigQuery, a clustered column is a user-defined table property that sorts storage blocks based on the values in the clustered columns. The storage blocks are adaptively sized based on the size of the table. A clustered table maintains the sort properties in the context of each operation that modifies it. Queries that filter or aggregate by the clustered columns only scan the relevant blocks based on the clustered columns instead of the entire table or table partition.

upvoted 7 times

 **AzureDP900** 1 year ago

Yes it is B. Implement clustering in BigQuery on the package-tracking ID column.

upvoted 1 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: B**

B as Clustering the data on the package Id can greatly improve the performance.  
Refer GCP documentation - BigQuery Clustered Table:<https://cloud.google.com/bigquery/docs/clustered-tables>  
upvoted 1 times

MaxNRG 1 month, 1 week ago

Clustering can improve the performance of certain types of queries such as queries that use filter clauses and queries that aggregate data. When data is written to a clustered table by a query job or a load job, BigQuery sorts the data using the values in the clustering columns. These values are used to organize the data into multiple blocks in BigQuery storage. When you submit a query containing a clause that filters data based on the clustering columns, BigQuery uses the sorted blocks to eliminate scans of unnecessary data. Currently, BigQuery allows clustering over a partitioned table. Use clustering over a partitioned table when:

- Your data is already partitioned on a date, timestamp, or integer column.
- You commonly use filters or aggregation against particular columns in your queries.

Table clustering is possible for tables partitioned by:

- ingestion time
- date/timestamp
- integer range

upvoted 1 times

MaxNRG 1 month, 1 week ago

In a table partitioned by a date or timestamp column, each partition contains a single day of data. When the data is stored, BigQuery ensures that all the data in a block belongs to a single partition. A partitioned table maintains these properties across all operations that modify it: query jobs, Data Manipulation Language (DML) statements, Data Definition Language (DDL) statements, load jobs, and copy jobs. This requires BigQuery to maintain more metadata than a non-partitioned table. As the number of partitions increases, the amount of metadata overhead increases.

upvoted 1 times

MaxNRG 1 month, 1 week ago

Although more metadata must be maintained, by ensuring that data is partitioned globally, BigQuery can more accurately estimate the total bytes processed by a query before you run it. This cost calculation provides an upper bound on the final cost of the query. In a clustered table, BigQuery automatically sorts the data based on the values in the clustering columns and organizes them in optimally sized storage blocks. You can achieve more finely grained sorting by creating a table that is clustered and partitioned. A clustered table maintains the sort properties in the context of each operation that modifies it. As a result, BigQuery may not be able to accurately estimate the bytes processed by the query or the query costs. When blocks of data are eliminated during query execution, BigQuery provides a best effort reduction of the query costs.

upvoted 1 times

Aman47 1 month, 2 weeks ago

**Selected Answer: B**

Package Tracking mostly contains, geospatial prefixes, Like HK0011, US0022, etc, this can help in clustering.  
upvoted 2 times

kcl10 3 months, 3 weeks ago

**Selected Answer: D**

D is the correct answer

requirements: analyze geospatial trends in the lifecycle of a package

cuz the data of the lifecycle of the package would span across ingest-date-based partition table, it would degrade the performance.

hence, re-partitioning by package delivery date, which is the package initially delivered, would improve the performance when querying such table.

upvoted 4 times

sdi\_studiers 7 months, 2 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 fully capture a package's full lifecycle in a single partition.

upvoted 3 times

cloudmon 1 year, 2 months ago

**Selected Answer: B**

B. <https://cloud.google.com/bigquery/docs/clustered-tables>  
upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

D is not correct because This is problem is The Real Time so ingested date is the same as delivery date.  
upvoted 3 times

Question #167

Topic 1

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.
- 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.

**Correct Answer: D**

*Community vote distribution*

B (89%)

11%

✉  **zelliC**  1 year, 2 months 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](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:

1. Copying data from a source cluster to Cloud Storage
2. Loading data from Cloud Storage into BigQuery

upvoted 8 times

✉  **AzureDP900** 1 year ago

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

upvoted 1 times

✉  **MaxNRG**  1 month, 1 week ago

**Selected Answer: B**

Based on the time constraint of 2 months and the goal to maximize cost savings, I would recommend option B - Migrate the workloads to Dataproc plus Cloud Storage; modernize later.

The key reasons are:

- Dataproc provides a fast, native migration path from on-prem Spark and Hive to the cloud. This allows meeting the 2 month timeline.
- Using Cloud Storage instead of HDFS avoids managing clusters for variable workloads and provides cost savings.
- Further optimizations and modernization to serverless (Dataflow, BigQuery) can happen incrementally later without time pressure.

upvoted 2 times

✉  **MaxNRG** 1 month, 1 week ago

Option A still requires managing HDFS.

Option C and D require full modernization of workloads in 2 months which is likely infeasible.

Therefore, migrating to Dataproc with Cloud Storage fast tracks the migration within 2 months while realizing immediate cost savings, enabling the flexibility to iteratively modernize and optimize the workloads over time.

upvoted 2 times

✉  **John\_Pongthorn** 1 year, 3 months ago

**Selected Answer: B**

B is most likely

1. migrate job and infrastructure to dataproc on cloud

2. any data, move from hdfs on-premise to google cloud storage ( one of them is Hive)

If you want to modernize Hive to Bigquery , you are need to move it into GCS(preceding step) first and load it into bigquery that is all.

<https://cloud.google.com/blog/products/data-analytics/apache-hive-to-bigquery>

<https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc>

<https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-data>

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: D**

Answer D

upvoted 1 times

✉  **dn\_mohammed\_data** 1 year, 4 months ago

you sould migrate spark to apache beam which is not the case here

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

apache beam for what???

upvoted 1 times

✉  **adarifian** 1 year, 3 months ago

dataflow uses apache beam

upvoted 1 times

✉  **TNT87** 1 year ago

@adarifian Why use apache beam yet there is Dataflow an inhouse gcp solution to solve the problem? hence i said apache beam what

upvoted 1 times

✉  **ExamCtechs** 2 months, 4 weeks ago

Dataflow IS apache beam, Dataflow is a Beam Runner.

If you go for that soulution you will need to modify your pipeline to use Beam

upvoted 1 times

✉  **GyaneswarPanigrahi** 1 year, 4 months ago

D isn't feasible, within 2 months. Anyone who has worked in a Hadoop/ Big Data data warehousing or data lake project, knows how less time months is, given the amount of data and associated complexities abound.

It should be B to begin with. And then gradually move towards D.

upvoted 3 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: B**

Ans B

-cost saving

-time factor

-Spark -Data proc

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

Ans D is also relevant if you read this. Onthe other hand cloud storage isnt severless but bigquery is  
<https://cloud.google.com/hadoop-spark-migration>

upvoted 1 times

✉  **damaldon** 1 year, 4 months ago

Ans.B as per the following link

<https://blog.devgenius.io/migrating-spark-jobs-to-google-cloud-file-event-sensor-to-dynamically-create-spark-cluster-7eff2c75423d>

upvoted 1 times

✉  **YorelNation** 1 year, 4 months ago

**Selected Answer: B**

For the time window of two month I would recommend B and then start to implement D.  
upvoted 2 times

✉  **ducc** 1 year, 4 months ago

It is B or D, still confusing  
upvoted 2 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

D because the Apache Spark Runner can be used to execute Beam pipelines using Apache Spark. Also, Hive to BigQuery is not a difficult modernization/migration.  
upvoted 1 times

✉  **ExamCtechs** 2 months, 4 weeks ago

Dataflow is a Runner of Beam it self  
upvoted 1 times

Question #168

Topic 1

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.
- 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.

**Correct Answer: D**

*Community vote distribution*

D (58%)

B (38%)

4%

✉  **AWSandeep**  1 year, 4 months 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 10 times

✉️  **MaxNRG** 1 month, 1 week ago

B. BigQuery column-level security:

Pros: Granular control over column access, ensures only authorized users see the SSN column.

Cons: Doesn't truly redact the data. The SSN values are still stored in BigQuery, even if hidden from unauthorized users. A potential security breach could expose them.

upvoted 1 times

✉️  **ffggrre** 3 months ago

there is no SSN in question, it can be any ID.

upvoted 1 times

✉️  **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: D**

The best option is D - Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographically-preserving encryption token.

The key reasons are:

DLP allows redacting sensitive PII like SSNs before loading into BigQuery. This provides security by default for the raw SSN values. Using format-preserving encryption keeps the column format intact while still encrypting, allowing business logic relying on SSN format to continue functioning.

The encrypted tokens can be reversed to view original SSNs when required, meeting the access requirement for customer service reps.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

Option A does encrypt SSN but requires managing keys separately.

Option B relies on complex IAM policy changes instead of encrypting by default.

Option C hashes irreversibly, preventing customer service reps from viewing original SSNs when required.

Therefore, using DLP format-preserving encryption before BigQuery ingestion balances both security and analytics requirements for SSN data.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

Why not B. BigQuery column-level security:

Doesn't truly redact the data. The SSN values are still stored in BigQuery, even if hidden from unauthorized users. A potential security breach could expose them.

upvoted 1 times

✉️  **Aman47** 1 month, 2 weeks ago

**Selected Answer: D**

Even if you provide Column level access control, The Data Owners or other hierarchies above it will also be able to view very sensitive data. Better to just use encryption and decryption. As this data can also never be used for any analytic workloads

upvoted 2 times

✉️  **spicebits** 2 months, 3 weeks ago

**Selected Answer: D**

Answer has to be D. Question says "you decide to redact your customers' Government issued Identification Number while allowing customer service representatives to view the original values when necessary" ... Redact... view the original values... D is the only choice.

upvoted 2 times

✉️  **Nirca** 3 months ago

**Selected Answer: B**

It might not be D!

Since - only the Frame is kept. the data will be changed.

Format Preserving Encryption (FPE), endorsed by NIST, is an advanced encryption technique that transforms data into an encrypted format while preserving its original structure. For instance, a 16-digit credit card number encrypted with FPE will still be a 16-digit number

upvoted 1 times

✉️  **Helinia** 1 month ago

No, the value using FPE can be decrypted with key.

"Encrypted values can be re-identified using the original cryptographic key and the entire output value, including surrogate annotation."

<https://cloud.google.com/dlp/docs/pseudonymization#supported-methods>

upvoted 1 times

ffggre 3 months, 1 week ago

Selected Answer: B

Customer service needs to see the original value, not possible with other options.

upvoted 1 times

kcl10 3 months, 3 weeks ago

Selected Answer: B

of course B

upvoted 1 times

ckanaar 4 months, 1 week ago

Selected Answer: D

I believe the crux to the question is that the cryptographic format-preserving encryption token is re-identifiable, whereas the cryptographic hash is not: <https://cloud.google.com/dlp/docs/transformations-reference>

Therefore, customer service can view the original values when necessary in case of D.

upvoted 2 times

ckanaar 4 months, 1 week ago

Nevermind, this can actually also be done in the case of answer B. They are both correct, just different implementations. No idea

upvoted 1 times

Lanro 6 months 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.

upvoted 4 times

knith66 6 months, 1 week ago

the question mentions that "user data is sent to Pub/Sub before being ingested" instead of just saying data goes to big query through pub/sub. So some alteration is expected before being injected into the big query. So option D should work.

upvoted 2 times

sr25 6 months, 1 week ago

Selected Answer: D

D. The question says giving CSR's access to values "when necessary" - not default access like given in B. D is a better option using the token

upvoted 1 times

ZZHZZH 6 months, 3 weeks ago

Selected Answer: B

One of the key requirement is to be able to let authorized personnel see the ID. D doesn't specify that.

upvoted 1 times

vaga1 8 months, 2 weeks ago

Selected Answer: D

The answer is between B and D as well described in many comments.

I personally do not see any reason to keep the information available using a token or a mask. It is not a PAN card number, it's just a personal I should not be useful for analytical purposes.

I'm gonna go for D then

upvoted 1 times

vaga1 8 months, 2 weeks ago

sorry B

upvoted 1 times

mialli 9 months ago

Selected Answer: D

<https://cloud.google.com/dlp/docs/classification-redaction>

upvoted 3 times

✉  **Oleksandr0501** 9 months ago

gpt: Both options B and D can be used to redact sensitive data while still allowing authorized users to view the original values when necessary. However, the choice between them would depend on specific business requirements and security considerations.

Option B uses BigQuery column-level security to set table permissions for users, allowing only members of the Customer Service user group to view the SSN column. This approach is straightforward and can be implemented easily. However, it requires creating a separate user group for customer service representatives and granting them access to only the required data columns.

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

gpt: Option D uses Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token before loading the data into BigQuery. This approach allows for more granular control over data access and can provide an added layer of security. However, it may require additional configuration and implementation effort, and it may also affect the performance of queries on the encrypted data.

Google recommends using a combination of data protection techniques to safeguard sensitive data, such as encryption, data masking, and access controls. In this scenario, a possible best practice would be to use both options B and D together to provide multiple layers of protection for the sensitive data while still allowing authorized users to view the original values when necessary.

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

I'll take D

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

now i've read and think about better choosing A or B ...

garbage question

upvoted 1 times

✉  **muhusman** 9 months, 2 weeks ago

Answer is B,

If we select C then this approach would also prevent unauthorized access to sensitive data, but it would not allow customer service representatives to view the original values when necessary.

upvoted 1 times

✉  **streeeber** 9 months, 2 weeks ago

**Selected Answer: D**

PII and DLP go hand in hand

upvoted 2 times

✉  **El\_Bosco** 6 months, 1 week ago

That is not an argument. Option D does not explain how Customer Services will have access.

upvoted 1 times

Question #169

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

A (74%)

B (19%)

7%

 **AWSandeep**  1 year, 4 months ago

**Selected Answer: A**

- A. Partition by transaction time; cluster by state first, then city, then store ID.  
upvoted 9 times

 **Atnafu**  1 year, 1 month 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 7 times

✉  **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: B**

over the past 30 days -> partitioning  
by state, city, and individual store -> cluster order  
upvoted 2 times

✉  **MaxNRG** 1 month, 1 week ago

For optimal query performance in BigQuery, especially for the described use cases of analyzing sales data by time and geographical hierarchies, the data should be organized to minimize the amount of data scanned during queries. Given the frequent queries over the past days and analysis by location, the best approach is:

Option A: Partition by transaction time; cluster by state first, then city, then store ID.

upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

Partitioning the table by transaction time allows for efficient querying over specific time ranges, such as the past 30 days, which reduces costs and improves performance because it limits the amount of data scanned.

Clustering by state, then city, and then store ID aligns with the hierarchy of geographical data and the types of queries that are run against the dataset. It organizes the data within each partition so that queries filtering by state, city, or store ID—or any combination of these—are optimized, as BigQuery can limit the scan to just the relevant clusters within the partitions.

upvoted 2 times

✉  **tibuenoc** 2 months, 1 week ago

**Selected Answer: B**

Partition by ingest time  
Partition by specified data column (Id, State and City)  
upvoted 1 times

✉  **ffggrr** 3 months, 1 week ago

**Selected Answer: C**

Partition by transaction time would lead to too many partitions - if it was a date, it would have made sense.  
upvoted 1 times

✉  **aureole** 3 months, 3 weeks ago

**Selected Answer: C**

It should be C. not A  
upvoted 1 times

✉  **aureole** 3 months, 3 weeks ago

I think it should be C.  
The fact that we partition the table with the time of the transaction will result in many partitions in each day, so it will affect negatively the query performance.  
i.e : by the end of the day I will have many partitions if I use the transaction time. A would be correct if the partition was by date and not by time.  
Response: C.  
upvoted 1 times

✉  **vaga1** 8 months, 2 weeks ago

**Selected Answer: A**

Partitioning by time is obvious to improve performance and costs of querying only the last 30 days of the table.

So, the answer is A or B.

<https://cloud.google.com/bigquery/docs/querying-clustered-tables>

"... To get the benefits of clustering, include all of the clustered columns or a subset of the columns in left-to-right sort order, starting with the column."

This means that it is a better choice to sort the table rows by region-province-city (region-state-city in the US case).

So, the answer is A.

upvoted 4 times

✉  **Prakzz** 1 year, 1 month ago

**Selected Answer: B**

Should be B

The clustering should be according to the filtering needs

upvoted 2 times

✉  **zelliCK** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/bigquery/docs/partitioned-tables>

This page provides an overview of partitioned tables in BigQuery. A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

You can partition BigQuery tables by:

- Time-unit column: Tables are partitioned based on a TIMESTAMP, DATE, or DATETIME column in the table.

<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.

upvoted 4 times

✉  **TNT87** 1 year, 4 months ago

<https://cloud.google.com/bigquery/docs/querying-clustered-tables>

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: A**

Question #170

Topic 1

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.
- 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.

**Correct Answer: C**

Reference:

<https://cloud.google.com/pubsub/docs/replay-overview>

## Seeking with filters

You can replay messages from subscriptions with filters. If you seek to a timestamp using a subscription with a filter, the Pub/Sub service only redelivers the messages that match the filter.

A snapshot of a subscription with a filter contains the following messages:

- All messages that are newer than the snapshot, including messages that don't match the filter.
- Unacknowledged messages that are older than the snapshot.

**★ Note:** The snapshot might contain messages that are older than the snapshot and don't match the filter.

If you seek to a snapshot using a subscription with a filter, the Pub/Sub service only redelivers the messages in the snapshot that match the filter of the subscription making the seek request.

For more information about filters, see [Filtering messages](#).

*Community vote distribution*

B (82%)

Other

✉ **AWSandeep** Highly Voted 1 year, 4 months ago

**Selected Answer: B**

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.

According to the second reference in the list below, a concern with deploying new subscriber code is that the new executable may erroneously acknowledge messages, leading to message loss. Incorporating snapshots into your deployment process gives you a way to recover from bugs in new subscriber code.

Answer cannot be C because To seek to a timestamp, you must first configure the subscription to retain acknowledged messages using retain-acked-messages. If retain-acked-messages is set, Pub/Sub retains acknowledged messages for 7 days.

References:

<https://cloud.google.com/pubsub/docs/replay-message>

[https://cloud.google.com/pubsub/docs/replay-overview#seek\\_use\\_cases](https://cloud.google.com/pubsub/docs/replay-overview#seek_use_cases)

upvoted 10 times

✉ **jkhong** 1 year, 1 month ago

Don't think we need to configure subscription to retain ack messages. It is defaulted to retain for 7 days

upvoted 1 times

✉ **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: B**

Taking a snapshot allows redelivering messages that were published while any faulty subscriber logic was running.

The seek timestamp would come after deployment so even erroneously acknowledged messages could be recovered.

[https://cloud.google.com/pubsub/docs/replay-overview#seek\\_use\\_cases](https://cloud.google.com/pubsub/docs/replay-overview#seek_use_cases)

By creating a snapshot of the subscription before deploying new code, you can preserve the state of unacknowledged messages. If after deployment you find that the new subscriber code is erroneously acknowledging messages, you can use the Seek operation with the snapshot to reset the subscription's acknowledgment state to the time the snapshot was created. This would effectively re-deliver messages available since the snapshot, ensuring you can recover from errors. This approach does not require setting up a local emulator and directly addresses the concern of message loss due to erroneous acknowledgments.

upvoted 1 times

✉  **[Removed]** 4 months, 2 weeks ago

Selected Answer: B

B.

from the documentation:

<https://cloud.google.com/pubsub/docs/replay-message>

Pub/Sub cannot retrieve the messages after you have acknowledged them. However, sometimes you might find it necessary to replay the acknowledged messages, for example, if you performed an erroneous acknowledgment. Then you can use the Seek feature to mark previous acknowledged messages as unacknowledged, and force Pub/Sub to redeliver those messages. You can also use seek to delete the unacknowledged messages by changing their state to acknowledged.

upvoted 2 times

✉  **vamgcp** 6 months, 1 week ago

pls correct me if I am wrong , option B Option B only allows you to re-deliver messages that were available before the snapshot was created. an error occurs after the snapshot was created, you will not be able to re-deliver those messages.

upvoted 2 times

✉  **cetanx** 6 months, 4 weeks ago

Selected Answer: A

Q: You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss.

-> So the message is mistakenly acked and removed from topic/subscription. This means even if you have a snapshot of pre-deployment but you don't have a backup or copy of post-deployment messages.

Q: Your subscriber is not set up to retain acknowledged messages.

-> To seek to a time in the past and replay previously-acknowledged messages, "you must first configure message retention on the topic" or "configure the subscription to retain acknowledged messages" (ref: [https://cloud.google.com/pubsub/docs/replay-overview#configuring\\_message\\_retention](https://cloud.google.com/pubsub/docs/replay-overview#configuring_message_retention))

So B, C, D do not solve the problem of erroneously acked messages as long as you don't have message retention configured on topic/subscription.

upvoted 3 times

✉  **lucaluka1982** 10 months, 1 week ago

Selected Answer: D

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.

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.

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

Option D:

Dead letter option allow you to recover message from errors after deployment by re-delivering any messages captured by the dead-letter queue

[https://cloud.google.com/pubsub/docs/handling-failures#dead\\_letter\\_topic](https://cloud.google.com/pubsub/docs/handling-failures#dead_letter_topic)

why not B,

because snapshot is time taking process and if messages were erroneously acknowledged, it will not bring them back. It is useful when you want to secure the current data and want to make changes

upvoted 2 times

✉  **wjtb** 10 months, 3 weeks ago

Dead letter queue would help if the messages would not get acknowledged, however here they are talking about messages being erroneously acknowledged. Pub/Sub would interpret the message as being successfully processed -> they would not end up in the dead-letter queue :- is wrong

upvoted 4 times

zellck 1 year, 2 months ago

Selected Answer: B

B is the answer.

<https://cloud.google.com/pubsub/docs/replay-overview>

The Seek feature extends subscriber functionality by allowing you to alter the acknowledgement state of messages in bulk. For example, you can replay previously acknowledged messages or purge messages in bulk. In addition, you can copy the state of one subscription to another by using seek in combination with a Snapshot.

upvoted 3 times

Atnafu 1 year, 2 months ago

B

The Seek feature extends subscriber functionality by allowing you to alter the acknowledgement state of messages in bulk. For example, you can replay previously acknowledged messages or purge messages in bulk. In addition, you can copy the state of one subscription to another by using seek in combination with a Snapshot.

Question #171

Topic 1

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.
- 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.

**Correct Answer: B**

Community vote distribution

A (97%)

3%

AWSandeep  1 year, 4 months 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 13 times

zellck  1 year, 2 months ago

Selected Answer: A

A is the answer.

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

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

upvoted 5 times

Matt\_108  2 weeks, 2 days ago

Selected Answer: A

Option A

upvoted 1 times

✉  **Prudvi3266** 9 months, 1 week ago

**Selected Answer: A**

A is correct answer if we use TRANSFORM clause in BigQuery no need to use any transform while evaluating and predicting <https://cloud.google.com/bigquery/docs/bigqueryml-transform>

upvoted 3 times

✉  **Kvk117** 1 year ago

**Selected Answer: A**

A is the correct answer

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: A**

Problem: Skew

One thing that I overlooked when answering previously is that B, C does not address skew. When we preprocess our training data, we need to save our scaled factors somewhere, and when performing predictions on our test data, we need to use the scaling factors of our training data to predict the results.

ML.EVALUATE already incorporates preprocessing steps for our test data using the saved scaled factors.

upvoted 3 times

✉  **GCPSharon** 1 year, 3 months ago

**Selected Answer: C**

Stew prediction time by remove the preprocessing!

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: A**

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

Ans A

upvoted 4 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: A**

This query's nested SELECT statement and FROM clause are the same as those in the CREATE MODEL query. Because the TRANSFORM clause is used in training, you don't need to specify the specific columns and transformations. They are automatically restored.

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

upvoted 2 times

Question #172

Topic 1

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 ()

**Correct Answer: B**

*Community vote distribution*

D (100%)

 **AWSandeep** Highly Voted 1 year, 4 months 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** Highly Voted 6 months, 1 week 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.

Window Duration: The window duration should be set to 30 seconds to cover the required 30 seconds' worth of data for the moving average calculation.

Window Period: The window period or sliding interval should be set to 5 seconds to move the window every 5 seconds and recalculate the moving average with the latest data.

Trigger: The trigger should be set to AfterWatermark.pastEndOfWindow() to emit the computed moving average results when the watermark advances past the end of the window. This ensures that all data within the window is considered before emitting the result.

upvoted 6 times

 **Kimich** Most Recent 1 month, 4 weeks ago

AfterWatermark is an essential triggering condition in Dataflow that allows computations to be triggered based on event time rather than processing time. Then eliminate A&C. Comparing B&D, B will generate outcome every 30 seconds which is not what we want

D. Using a sliding window with a duration of 30 seconds and a period of 5 seconds, and setting the trigger as AfterWatermark.pastEndOfWindow(), is a sliding window that generates results every 5 seconds, and each result includes data from the past 30 seconds. In other words, every 5 seconds, you get the average value of the most recent 30 seconds' data, and there is a 5-second overlap between these windows. This is what we want.

upvoted 2 times

 **zelick** 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#hopping-windows>

You set the following windows with the Apache Beam SDK or Dataflow SQL streaming extensions:  
Hopping windows (called sliding windows in Apache Beam)

A hopping window represents a consistent time interval in the data stream. Hopping windows can overlap, whereas tumbling windows are disjoint.

For example, a hopping window can start every thirty seconds and capture one minute of data. The frequency with which hopping windows begin is called the period. This example has a one-minute window and thirty-second period.

upvoted 2 times

 **pluiedust** 1 year, 4 months ago

**Selected Answer: D**

Moving average —> sliding window

upvoted 4 times

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.

**Correct Answer: D**

*Community vote distribution*

D (100%)

✉  **Atnafu** Highly Voted 1 year, 1 month ago  
D

TUMBLE=> fixed windows.

HOP=> sliding windows.

SESSION=> session windows.

upvoted 11 times

✉  **musumusu** Highly Voted 11 months, 1 week ago  
why not c ? as data is arriving hourly why we can use batch processing rather than streaming with 1 hour fixed window?  
upvoted 5 times

✉  **ga8our** 3 months ago  
I second your question. No one who suggests C has given an explanation why an hourly batch job is insufficient.

upvoted 1 times

✉  **ga8our** 3 months ago  
I second your question. No one who suggests Dataflow streaming (D) has given an explanation why an hourly batch job is insufficient.

upvoted 1 times

✉  **MrMone** 9 months, 2 weeks ago  
"you need to be able to aggregate events across disjoint hourly intervals" does not mean data is arriving hourly. however, it's tricky! Answer D  
upvoted 2 times

✉  **emmylou** Most Recent 2 months, 1 week ago  
I just do not understand why this needs to be streamed. I understand that there might be a slight delay using batch processing but there is no indication this is critical data. Can someone please provide your thinking?  
upvoted 1 times

✉  **vamgcp** 6 months, 1 week ago  
We can use TUMBLE(1 HOUR) to create hourly windows, where each window contains events from a specific hour.  
upvoted 1 times

👤 **vamgcp** 6 months, 1 week ago

**Selected Answer: D**

Option D : A streaming Dataflow job is the best way to process and load data from Pub/Sub to BigQuery in real time. This is because streaming Dataflow jobs can scale to handle large volumes of data, and they can perform aggregations using tumbling windows.

upvoted 1 times

👤 **zelliCK** 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#tumbling-windows>

upvoted 3 times

👤 **deavid** 1 year, 3 months ago

**Selected Answer: D**

Answer D

Tumbling Windows = Fixed Windows

upvoted 2 times

👤 **TNT87** 1 year, 4 months ago

**Selected Answer: D**

Answer D

upvoted 2 times

👤 **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

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

A tumbling window represents a consistent, disjoint time interval in the data stream.

Reference:

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#tumbling-windows>

upvoted 2 times

Question #174

Topic 1

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.
- 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.

**Correct Answer: A**

*Community vote distribution*

A (100%)

✉️  **MaxNRG** 1 month, 1 week ago

**Selected Answer: A**

This is the best approach because it follows the Pareto principle (80/20 rule). By automating the most common 10 intents that address 70% of customer requests, you free up the live agents to focus their time and effort on the more complex 30% of requests that likely require human insight/judgement. Automating the simpler high-volume requests first allows the chatbot to handle those easily, efficiently routing only the trickier cases to agents. This makes the best use of automation for high-volume simple cases and human expertise for lower-volume complex issues.

upvoted 2 times

✉️  **vamgcp** 6 months, 1 week 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** 7 months, 2 weeks ago

**Selected Answer: A**

A is the answer.

upvoted 1 times

✉️  **zellick** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/dialogflow/cx/docs/concept/agent-design#build-iteratively>

If your agent will be large or complex, start by building a dialog that only addresses the top level requests. Once the basic structure is established, iterate on the conversation paths to ensure you're covering all of the possible routes an end-user may take.

upvoted 4 times

✉️  **SMASL** 1 year, 4 months ago

Correct answer: A

As it states in the documentation: "If your agent will be large or complex, start by building a dialog that only addresses the top level requests. Once the basic structure is established, iterate on the conversation paths to ensure you're covering all of the possible routes an end-user may take." (<https://cloud.google.com/dialogflow/cx/docs/concept/agent-design#build-iteratively>)

Therefore, you should initially automate the 70 % of the requests that are simpler before automating the more complicated ones.

upvoted 4 times

✉️  **AWSandeep** 1 year, 4 months ago

**Selected Answer: A**

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

upvoted 1 times

Question #175

Topic 1

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.

**Correct Answer: A**

Reference:

<https://cloud.google.com/architecture/dw2bq/dw-bq-schema-and-data-transfer-overview>

*Community vote distribution*

D (87%)

13%

 **waiebdi** Highly Voted 11 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 12 times

 **Kimich** 1 month, 4 weeks ago

"Agree with you, denormalize usually increases storage, which may lead to an increase in cost. As for speeding up the query without increasing storage costs, another method is to partition the data by transaction date."

upvoted 1 times

 **Aman47** Most Recent 1 month, 2 weeks ago

Bro, you are playing with words now. Gotta read the question fully.

upvoted 1 times

 **philv** 4 months ago

Some might say that Star schema is already denormalized, but it is considered relational (hence kind of normalized) from Google's perspective

"BigQuery performs best when your data is denormalized. Rather than preserving a relational schema such as a star or snowflake schema, denormalize your data and take advantage of nested and repeated columns. Nested and repeated columns can maintain relationships without the performance impact of preserving a relational (normalized) schema."

I would go for A

[https://cloud.google.com/bigquery/docs/nested-repeated#when\\_to\\_use\\_nested\\_and\\_repeated\\_columns](https://cloud.google.com/bigquery/docs/nested-repeated#when_to_use_nested_and_repeated_columns)

upvoted 1 times

 **philv** 3 months, 2 weeks ago

Changed my mind to D because of the "without increasing storage costs" part.

upvoted 1 times

 **vamgcp** 6 months, 1 week 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

 **WillemHendr** 7 months, 3 weeks ago

A is not a bad idea, but this question is written around "please partition first on date", which is common best practice. The "storage" remark hints on we are not going to 'explode' the data for the sake of performance.

upvoted 2 times

👤 **pcadolini** 1 year, 1 month ago

**Selected Answer: A**

I think better option is [A] considering GCP Documentation: <https://cloud.google.com/bigquery/docs/migration/schema-data-overview#denormalization> "BigQuery supports both star and snowflake schemas, but its native schema representation is neither of those two uses nested and repeated fields instead for a more natural representation of the data ..... Changing your schema to use nested and repeated fields is an excellent evolutionary choice. It reduces the number of joins required for your queries, and it aligns your schema with the BigQuery internal data representation. Internally, BigQuery organizes data using the Dremel model and stores it in a columnar storage format called Capacitor."

upvoted 4 times

👤 **zelliCK** 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/bigquery/docs/partitioned-tables>

A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

upvoted 3 times

👤 **NicolasN** 1 year, 2 months ago

**Selected Answer: D**

A sneaky question.

[D] Yes - Since data is queried with date criteria, partitioning by transaction date will surely speed it up without further cost.

[A] Yes? - Star schema is a denormalized model but as user Reall01 pointed out, the option to use nested and repeated fields can be considered a further denormalization. And if the model hasn't frequently changing dimensions, this kind of denormalization will result in increased performance, according to [https://cloud.google.com/bigquery/docs/loading-data#loading\\_denormalized\\_nested\\_and\\_repeated\\_data](https://cloud.google.com/bigquery/docs/loading-data#loading_denormalized_nested_and_repeated_data) : "In some circumstances, denormalizing your data and using nested and repeated fields doesn't result in increased performance. Avoid denormalization in these use cases:

- You have a star schema with frequently changing dimensions"

I guess that the person who added this question, had in mind [D] as a correct answer. If the questioner had all the aforementioned under consideration, would state clearly if there are frequently changing dimensions in the schema.

upvoted 4 times

👤 **josrojgra** 1 year, 3 months ago

**Selected Answer: D**

Star schema is supported by Big Query but is not the most efficient form, if you should design a schema from scratch google recommend to use nested and repeated fields.

In this case, you already have done a migration of the schema and data, so it sounds good and with less effort to do partitioning by transaction date than to redesign the schema.

And other aspect to consider is that this is a data warehouse, so is sure that there is an ETL process and if you change the schema you must adapt the ETL process.

I vote for D.

upvoted 1 times

👤 **deavid** 1 year, 3 months ago

**Selected Answer: D**

Star schema is not denormalized itself, but this assumes you already have moved ur data to big query, because you are querying. So, as BQ is not relational, the data already have been denormalized. I go with D.

upvoted 2 times

👤 **learner2610** 1 year, 4 months ago

I think Denormalizing here means ,using big queries native data representation and that is using nested and repeated columns .That's is the best practice in GCP

<https://cloud.google.com/bigquery/docs/nested-repeated#example>

upvoted 1 times

👤 [Removed] 1 year, 4 months ago

Selected Answer: D

[https://cloud.google.com/bigquery/docs/migration/schema-data-overview#migrating\\_data\\_and\\_schema\\_from\\_on-premises\\_to\\_bigquery](https://cloud.google.com/bigquery/docs/migration/schema-data-overview#migrating_data_and_schema_from_on-premises_to_bigquery)

Star schema. This is a denormalized model, where a fact table collects metrics such as order amount, discount, and quantity, along with a group of keys. These keys belong to dimension tables such as customer, supplier, region, and so on. Graphically, the model resembles a star, with the fact table in the center surrounded by dimension tables.

Star schema is already denormalized so partition makes more sense going with D

upvoted 2 times

👤 RealI01 1 year, 4 months ago

If you drill down within that link and land at: <https://cloud.google.com/architecture/bigquery-data-warehouse> it mentions " In some cases, you might want to use nested and repeated fields to denormalize your data." under schema design. Feels like a poorly written question since it depends on what context you take things in as "denormalization"

upvoted 2 times

👤 GabyB 7 months, 1 week ago

In some circumstances, denormalizing your data and using nested and repeated fields doesn't result in increased performance. For example, star schemas are typically optimized schemas for analytics, and as a result, performance might not be significantly different if you attempt to denormalize further.

<https://cloud.google.com/bigquery/docs/best-practices-performance-nested>

upvoted 1 times

👤 NicolasN 1 year, 2 months ago

You bring up a valid point. According to denormalization best practices, there is a critical info missing in order to decide whether further denormalization with nested and repeated fields could help, if there are frequently changing dimensions. Here's a quote from [https://cloud.google.com/bigquery/docs/loading-data#loading\\_denormalized\\_nested\\_and\\_repeated\\_data](https://cloud.google.com/bigquery/docs/loading-data#loading_denormalized_nested_and_repeated_data) :

"In some circumstances, denormalizing your data and using nested and repeated fields doesn't result in increased performance. Avoid denormalization in these use cases:

- You have a star schema with frequently changing dimensions."

upvoted 2 times

👤 AWSandeep 1 year, 4 months ago

D. Partition the data by transaction date.

Star schema is already denormalized.

upvoted 3 times

👤 PhuocT 1 year, 4 months ago

Selected Answer: D

should be D, not A

upvoted 2 times

Question #176

Topic 1

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.

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.

**Correct Answer: C**

*Community vote distribution*

C (100%)

 **AWSandeep** Highly Voted 1 year, 4 months 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 the question strictly implies that the existing data set needs to be retained for compliance.

upvoted 9 times

 **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: C**

Option C is the best approach in this situation. Here is why:

Option A would remove data which may be needed for compliance reasons. Keeping the original data is preferred.

Option B makes a copy of the data but still removes potentially useful records. Additional storage costs would be incurred as well.

Option C uses Dataflow to clean the data by setting out of range values while keeping the original data intact. The fixed records are written to new location for further analysis. This meets the requirements.

Option D writes the fixed data back to the original location, overwriting the original data. This would violate the compliance needs to keep the original data untouched.

So option C leverages Dataflow to properly clean the data while preserving the original data for compliance, at reasonable operational costs. best achieves the stated requirements.

upvoted 1 times

 **FP77** 5 months, 2 weeks ago

Strange answers... Since when does cloud storage have datasets? Lol  
Keeping this in mind, the answer must be C, but none is really correct

upvoted 3 times

 **AzureDP900** 1 year 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

 **zellck** 1 year, 2 months ago

**Selected Answer: C**

C is the answer.

upvoted 3 times

 **PhuocT** 1 year, 4 months ago

**Selected Answer: C**

C is correct

upvoted 2 times

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

**Correct Answer: D**

*Community vote distribution*

C (85%)

A (15%)

✉  **deavid**  1 year, 3 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 11 times

✉  **GCP001**  1 week, 6 days ago

**Selected Answer: A**

A) Looks more suitable , serverless approach for handling and performance.

upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

**Selected Answer: C**

Option C is the best approach to meet the stated requirements. Here's why:

BigQuery SQL provides a fast, scalable, and serverless method for transforming structured data, easier to develop than PySpark. Directly ingesting the raw Cloud Storage data into BigQuery avoids needing an intermediate processing cluster like Dataproc. Transforming the data via BigQuery SQL queries will be faster than PySpark, especially since the data is already loaded into BigQuery. Writing the transformed results to a new BigQuery table keeps the original raw data intact and provides a clean output. So migrating to BigQuery SQL for transformations provides a fully managed serverless architecture that can significantly expedite development and reduce pipeline runtime versus PySpark. The ability to avoid clusters and conduct transformations completely within BigQuery is the most efficient approach here.

upvoted 1 times

✉  **MoeHaydar** 6 months, 3 weeks ago

**Selected Answer: C**

Note: Dataproc by itself is not serverless

<https://cloud.google.com/dataproc-serverless/docs/overview>

upvoted 1 times

✉  **Prudvi3266** 9 months, 1 week ago

**Selected Answer: C**

because of serverless nature

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Answer C: need to setup SQL based job means transformation is not very complex. And Bigquery sql are faster than spark sql context. (google claims)

However, I will make a test by myself to check it.

upvoted 1 times

✉  **maci\_f** 1 year ago

**Selected Answer: A**

In the GCP Machine Learning Engineer practice question (Q4) there's the same question with similar answers and the correct answer is A since "is incorrect, here transformation is done on Cloud SQL, which wouldn't scale the process" and C "is incorrect as this process wouldn't scale data transformation routine. And, it is always better to transform data during ingestion": <https://medium.com/@gcpguru/google-google-cloud-professional-machine-learning-engineer-practice-questions-part-1-3ee4a2b3f0a4>

upvoted 2 times

✉  **evanfebrianto** 8 months ago

Dataproc is not a serverless tool unless it mentions "Dataproc Serverless" explicitly.

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

C

D is incorrect because you are rebuild your batch pipeline for structured data on Google Cloud.

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

A could be answer if it was Dataproc serverless and no conversion of code. Dp serverless support: scala,pyspark,sparksql and SparkR

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: C**

This same question is there on Google's Professional Machine Learning Engineer, Question 4

Answer is C.

upvoted 3 times

✉  **Wasss123** 1 year, 4 months ago

**Selected Answer: C**

I choose C

BigQuery SQL is more performant but more expensive. Here, it's a performance issue (time reduction)

Source : <https://medium.com/paypal-tech/comparing-bigquery-processing-and-spark-dataproc-4c90c10e31ac>

upvoted 2 times

✉  **John\_Pongthorn** 1 year, 4 months ago

C is the most likely, BigQuery is serverless and SQL

D is Dataflow serverless but it is wrong at using Python SDK but using SQL Beam then it will be correct

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Answer C

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: A**

A

- You have to maintain PySpark Code -> Proc

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

After thinking a while, I think the question is not clear enough. To be honest

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

A or C. I go for C because they said they want to use SQL syntax...

upvoted 1 times

✉  **AWSandeeP** 1 year, 4 months ago

**Selected Answer: C**

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.

Keys: "Serverless" and "SQL"

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

The question said "use SQL syntax"

C might still correct

upvoted 1 times

✉  **AWSandeeP** 1 year, 4 months ago

Changing answer to A as this is a new question referring to Dataproc Serverless. Dataproc Serverless for Spark batch workloads supports Spark SQL. Why modify ETL to ELT and convert PySpark to BigQuery SQL when it can be similar to a lift-and-shift?

upvoted 3 times

✉  **Atnafu** 1 year, 2 months ago

Dataproc is different than Dataproc Serverless. This question is talking about dataproc.

By the way dp serverless support both pyspark and sparkSql no need of conversion.

C is best answer

upvoted 3 times

Question #178

Topic 1

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.

**Correct Answer: C**

*Community vote distribution*

D (82%)

Other

✉  **John\_Pongthorn**  1 year, 4 months 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 15 times

 **zellck**  1 year, 2 months ago

**Selected Answer: D**

D is the answer.

[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/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-testing#choose_correctly_between_side_inputs_or_cogroupbykey_for_joins)

The CoGroupByKey transform is a core Beam transform that merges (flattens) multiple PCollection objects and groups elements that have a common key. Unlike a side input, which makes the entire side input data available to each worker, CoGroupByKey performs a shuffle (groupin operation to distribute data across workers. CoGroupByKey is therefore ideal when the PCollection objects you want to join are very large and don't fit into worker memory.

Use CoGroupByKey if you need to fetch a large proportion of a PCollection object that significantly exceeds worker memory.

upvoted 11 times

 **MaxNRG**  1 month, 1 week ago

**Selected Answer: D**

To expedite the Dataflow job that involves ingesting and transforming text files, especially if the pipeline is taking longer than expected, the most effective strategy would be:

D. Use CoGroupByKey instead of the SideInput.

upvoted 1 times

 **MaxNRG** 1 month, 1 week ago

Here's why this approach is beneficial:

1. Efficiency in Handling Large Datasets: SideInputs are not optimal for large datasets because they require that the entire dataset be available to each worker. This can lead to performance bottlenecks, especially if the dataset is large. CoGroupByKey, on the other hand, is more efficient for joining large datasets because it groups elements by key and allows the pipeline to process each key-group separately.

2. Scalability: CoGroupByKey is more scalable than SideInputs for large-scale data processing. It distributes the workload more evenly across the Dataflow workers, which can significantly improve the performance of your pipeline.

3. Better Resource Utilization: By using CoGroupByKey, the Dataflow job can make better use of its resources, as it doesn't need to replicate the entire dataset to each worker. This results in faster processing times and better overall efficiency.

upvoted 1 times

 **MaxNRG** 1 month, 1 week ago

The other options may not be as effective:

- A (Switch to compressed Avro files): While Avro is a good format for certain types of data processing, simply changing the file format from gzip to Avro may not address the underlying issue causing the delay, especially if the problem is related to the way data is being joined and processed.

- B (Reduce the batch size): Reducing the batch size could potentially increase overhead and might not significantly improve the processing time, especially if the bottleneck is due to the method of data joining.

- C (Retry records that throw an error): Retrying errors could be useful in certain contexts, but it's unlikely to speed up the pipeline if the delay is due to inefficiencies in data processing methods like the use of SideInputs.

upvoted 1 times

 **musumusu** 11 months, 2 weeks ago

Answer: B,

reducing the batch size improves the speed performance and also improves CPU utilization.

Dead letter queues are generated for messages that are errorfully acknowledged and it's good to use SideInputs for that to check small amount of errors in memory.

CogroupByKey is not necessary for error messages.

I see only batch size that can be customized to improve the performance.

In practical use case:

you check these tools Stackdriver Monitoring and Logging, Cloud Trace, and Cloud Profiler, and try to find the cause, if it's file type issue in compression, or batch size.

upvoted 1 times

✉  **Atnafu** 1 year, 1 month ago

D

Flatten will just merge all results into a single PCollection. To join them you can use CoGroupByKey  
upvoted 1 times

✉  **TNT87** 1 year, 3 months ago

**Selected Answer: A**

When optimizing for load speed, Avro file format is preferred. Avro is a binary row-based format which can be split and read in parallel by multiple slots including compressed files.

upvoted 2 times

✉  **deavid** 1 year, 3 months ago

that is for Big Query isn't?

upvoted 1 times

✉  **TNT87** 1 year ago

datflow can use avro format sir. streaming or batching to bigquery in avro format it can

upvoted 1 times

✉  **TNT87** 1 year, 3 months ago

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

upvoted 1 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: D**

D probably, side inputs have to fit in memory. If the p-collection in the side input doesn't fit well in memory it's better to use CoGroupByKey.  
upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: A**

Answer A

the same question is in number 70 you transform the files to Avro using Dataflow

upvoted 1 times

✉  **KC\_go\_reply** 7 months, 1 week ago

Avro requires the data to be at least semi-structured, because it wants a fixed schema. Text files are unstructured data, therefore it doesn't make sense to use Avro files for them

upvoted 3 times

✉  **csd1fgghfgvh234** 1 year, 4 months ago

A switching to avro. No serialisation

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Switch to Avro format

Answer A

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

<https://docs.confluent.io/platform/current/schema-registry/serdes-develop/serdes-avro.html>

upvoted 1 times

✉  **YorelNation** 1 year, 4 months ago

**Selected Answer: D**

D probably, side inputs have to fit in memory. If the p-collection in the side input doesn't fit well in memory it's better to use CoGroupByKey.  
upvoted 3 times

✉  **AWSandeeP** 1 year, 4 months ago

**Selected Answer: B**

B. Reduce the batch size.

upvoted 4 times

Question #179

Topic 1

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 bucket.
- 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.

**Correct Answer: B**

*Community vote distribution*

D (62%)

A (32%)

5%

 **zellck** Highly Voted 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/dlp/docs/pseudonymization#supported-methods>

Format preserving encryption: An input value is replaced with a value that has been encrypted using the FPE-FFX encryption algorithm with a cryptographic key, and then prepended with a surrogate annotation, if specified. By design, both the character set and the length of the input value are preserved in the output value. Encrypted values can be re-identified using the original cryptographic key and the entire output value including surrogate annotation.

upvoted 6 times

 **GCP001** Most Recent 1 week, 4 days ago

**Selected Answer: D**

D> Looks more suitable as it will handle Referential integrity. <https://cloud.google.com/dlp/docs/pseudonymization>

upvoted 1 times

✉  **pss111423** 2 months, 1 week ago

answer A

<https://cloud.google.com/dlp/docs/transformations-reference> Replaces an input value with a token, or surrogate value, of the same length using AES in Synthetic Initialization Vector mode (AES-SIV). This transformation method, unlike format-preserving tokenization, has no limitation on supported string character sets, generates identical tokens for each instance of an identical input value, and uses surrogates to enable re-identification given the original encryption key.

upvoted 2 times

✉  **akg001** 5 months, 2 weeks ago

**Selected Answer: D**

D is correct.

upvoted 1 times

✉  **cetanx** 7 months, 1 week 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.

upvoted 2 times

✉  **cetanx** 6 months, 4 weeks ago

made a typo up there, it has to be A

upvoted 1 times

✉  **Oleksandr0501** 9 months ago

gpt:

The recommended approach for using the Cloud Data Loss Prevention API (DLP API) to protect sensitive PII data while maintaining referential integrity is to create pseudonyms by replacing the PII data with cryptographic format-preserving tokens.

This approach ensures that sensitive data is not accessible by unauthorized individuals, while still preserving the format and length of the original data, which is essential for maintaining referential integrity.

Replacing PII data with cryptogenic tokens, as mentioned in option A, is not recommended because cryptogenic tokens are not necessarily format-preserving, and this could affect the accuracy of data joins.

Therefore, option D is the best approach for using the DLP API to ensure that PII data is not accessible by unauthorized individuals while still maintaining referential integrity.

upvoted 1 times

✉  **loicrichonnier** 8 months, 3 weeks ago

You shouldn't use ChatGPT as a source, the data used are not up to date and for such complex question a predicting text chatbot can help, but, it's better to refer to the google documentation.

upvoted 5 times

✉  **Oleksandr0501** 8 months, 3 weeks ago

that's why i always mark "gpt", when copy from there... i know, thx

also, it might be A. Or D... Confusing question.

upvoted 1 times

✉  **Prudvi3266** 9 months, 1 week ago

**Selected Answer: D**

here catch is "cryptographic" key

upvoted 3 times

✉  **musumusu** 11 months, 2 weeks ago

Answer D,

key word - "referential integrity" use format preserve option, it keeps same length of the value and last four digits of your value in column

upvoted 1 times

✉  **tunstila** 1 year ago

**Selected Answer: D**

The answer is D

upvoted 1 times

✉  **nkit** 1 year, 1 month ago

**Selected Answer: D**

I believe "Format preserving token" in option D makes it easier choice for me

upvoted 1 times

✉  **PrashantGupta1616** 1 year, 1 month ago

**Selected Answer: D**

D looks right

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: A**

Question is super tricky, B and C are not the answers since they do not maintain referential integrity.

For D, it does preserve the length of input. But since we are only concerned with referencing during joins, there is no point of maintaining the length anyway. Also, characters must be encoded as ASCII, this means that the name and email must be within the 256 character set. which is further limited to the alphabet characters, i.e. 94 characters. (<https://cloud.google.com/dlp/docs/transformations-reference#crypto>)

Names nowadays do not just have ASCII characters but unicode as well, so D will not necessarily work all the time.

upvoted 2 times

✉  **Atnafu** 1 year, 1 month ago

D is the answer

Pseudonymization is a de-identification technique that replaces sensitive data values with cryptographically generated tokens.

Keywords: You want to ensure that the sensitive data is masked but still maintains referential integrity

Part 1- data is masked-Create a pseudonym by replacing PII data with a cryptographic token

Part 2 still maintains referential integrity- with a cryptographic format-preserving token

A Not an answer because

the locked-down button does not seem to google cloud word

upvoted 4 times

✉  **juliobs** 10 months, 1 week ago

"button" is just a typo for "bucket"

upvoted 1 times

✉  **dish11dish** 1 year, 2 months ago

**Selected Answer: D**

Though both option A and D maintains referential integrity, question is why you want to keep untokenize data in GCS, best way is option D which even supports Reversible feature which is not supported by option A refer chart in reference document.

reference:-

<https://cloud.google.com/dlp/docs/pseudonymization>

upvoted 1 times

cloudmon 1 year, 2 months ago

Selected Answer: D

It's D.

"You want to ensure that the sensitive data is masked but still maintains referential integrity."

They don't ask you to also keep the original data (which answer A relates to).

Also, format-preservation is important in this case.

upvoted 3 times

cloudmon 1 year, 2 months ago

And, answer A does not include format preservation, which would lose referential integrity.

upvoted 1 times

NicolasN 1 year, 2 months ago

I think that this isn't true.

Look at the table [https://cloud.google.com/dlp/docs/transformations-reference#transformation\\_methods](https://cloud.google.com/dlp/docs/transformations-reference#transformation_methods) and notice the 6th line "Pseudonymization by replacing input value with cryptographic hash" (which refers to the case of answer [A]). Referential integrity is preserved.

upvoted 1 times

NicolasN 1 year, 2 months ago

Selected Answer: A

[B] and [C] aren't correct since they don't preserve referential integrity.

[A] describes, in other words, Cryptographic hashing, where the sensitive data is replaced with a hashed value. The hashed value can't be reversed (<https://cloud.google.com/dlp/docs/transformations-reference#crypto-hashing>) so the phrase "store the non-tokenized data in a local down button (bucket)" ensures that data can be restored if needed.

[D] seems to be a valid option too. However, in <https://cloud.google.com/dlp/docs/pseudonymization#fpe-ffx>, there is a warning: "FPE provides fewer security guarantees compared to other deterministic encryption methods such as AES-SIV ... For these reasons, Google strongly recommends using deterministic encryption with AES-SIV instead of FPE for all security sensitive use cases" Since there is no option to select Deterministic Encryption, and the question doesn't require to preserve the format of the data (keep the same length of data), I choose [A] as a more secure approach.

upvoted 3 times

wan2three 5 months, 3 weeks ago

From here I see if A really meant Cryptographic hashing then it also satisfy referential integrity

<https://cloud.google.com/dlp/docs/pseudonymization#:~:text=following%20the%20table.-,Deterministic%20encryption%20using%20AES-2DSIV,-Format%20preserving%20encryption>

However, I can't see why A means Cryptographic Hashing, no definition I can find online at all.

upvoted 1 times

Question #180

Topic 1

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

Correct Answer: D

Community vote distribution

C (100%)

✉️  **musumusu**  11 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 denormalize / nesting option.

upvoted 9 times

✉️  **zellick**  1 year, 2 months ago

**Selected Answer: C**

C is the answer.

<https://cloud.google.com/bigquery/docs/best-practices-performance-nested>

Best practice: Use nested and repeated fields to denormalize data storage and increase query performance.

Denormalization is a common strategy for increasing read performance for relational datasets that were previously normalized. The recommended way to denormalize data in BigQuery is to use nested and repeated fields. It's best to use this strategy when the relationships are hierarchical and frequently queried together, such as in parent-child relationships.

upvoted 5 times

✉️  **AzureDP900**  1 year 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

✉️  **Atnafu** 1 year, 2 months ago

C

Best practice: Use nested and repeated fields to denormalize data storage and increase query performance.

upvoted 2 times

✉️  **dish11dish** 1 year, 2 months ago

**Selected Answer: C**

Use nested and repeated fields to denormalize data storage which will increase query performance. BigQuery doesn't require a completely flat denormalization. You can use nested and repeated fields to maintain relationships

upvoted 2 times

✉️  **Thobm** 1 year, 4 months ago

**Selected Answer: C**

<https://cloud.google.com/bigquery/docs/best-practices-performance-nested>

upvoted 1 times

✉️  **ducc** 1 year, 4 months ago

**Selected Answer: C**

C is correct

upvoted 2 times

✉️  **AWSandeep** 1 year, 4 months ago

**Selected Answer: C**

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

upvoted 2 times

Question #181

Topic 1

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.

**Correct Answer: D***Community vote distribution*

D (93%)

7%

 **John\_Pongthorn**  1 year, 4 months ago**Selected Answer: D**

D: I have insisted on this choice all along.  
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 17 times

 **NicolasN** 1 year, 2 months ago

Thanks for sharing, you found exactly the same problem!  
The document definitely 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:
  - \* If the client is thread-safe and serializable, create it statically in the class definition of the DoFn.
  - \* If it's not thread-safe, create a new object in the startBundle method of DoFn. By doing so, the client will be reused across all elements of a bundle, saving initialization time.

upvoted 6 times

 **Atnafu** 1 year, 2 months ago

By the way if you see the shared Pseudocode, it's talking about start bundle and finish bundle of DoFn. The question is which one to choose to avoid back pressure?  
you can see why you need to choose bundle instead of batching in below link  
Batching introduces some processing overhead as well as the need for a magic number to determine the key space.  
Instead, use the StartBundle and FinishBundle lifecycle elements to batch your data. With these options, no shuffling is needed.  
<https://cloud.google.com/dataflow/docs/tutorials/ecommerce-java#micro-batch-calls>

upvoted 1 times

 **NicolasN** 1 year, 2 months ago

Valid points. but I don't change my mind, regarding the requirements of this particular question:

- multi-threaded ability
- no mention of heavy initialization steps or a lot of disk I/O (where shuffling might be a problem).

And especially the excerpt:

"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"

It's like the guys that authored the question had this sentence in front of their eyes.

upvoted 2 times

 **John\_Pongthorn**  1 year, 4 months ago**Selected Answer: D**

D

All guys, pls read carefully on Pattern: Calling external services for data enrichment  
<https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>  
A, B, C all of them are solution for normal case but if you need to stand for backpressure,  
in last section in Note: Note: When using this pattern, be sure to plan for the load that's placed on the external service and any associated backpressure. For example, imagine a pipeline that's processing tens of thousands of messages per second in steady state. If you made a callout per element, you would need the system to deal with the same number of API calls per second. Also, 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.

Anyone can share ideas to debate with me.

upvoted 8 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: D**

Option D is the best approach to minimize backpressure in this scenario. By batching the jobs into 10-second increments, you can throttle the rate at which requests are made to the external GUID service. This prevents too many simultaneous requests from overloading the service.

Option A would not help with backpressure since it just makes synchronous HTTP requests as messages arrive. Similarly, options B and C do not provide any inherent batching or throttling mechanism.

Batching into time windows is a common strategy in stream processing to deal with high velocity data. The 10-second windows allow some buffering to happen, rather than making a call immediately for each message. This provides a natural throttling that can be tuned based on the external service's capacity.

upvoted 1 times

MaxNRG 1 month, 1 week ago

To design a pipeline that minimizes backpressure, especially when dealing with tens of thousands of messages per second in a multi-threaded environment, it's important to consider how each option affects system performance and scalability. Let's examine each of your options:

upvoted 1 times

MaxNRG 1 month, 1 week ago

A. Call out to the service via HTTP: Making HTTP calls to an external service for each message can introduce significant latency and backpressure, especially at high throughput. This is due to the overhead of establishing a connection, waiting for the response, and handling potential network delays or failures.

upvoted 1 times

MaxNRG 1 month, 1 week ago

B. Create the pipeline statically in the class definition: While this approach can improve initialization time and reduce overhead during execution, it doesn't directly address the issue of backpressure caused by high message throughput.

upvoted 1 times

MaxNRG 1 month, 1 week ago

C. Create a new object in the startBundle method of DoFn: This approach is typically used in Apache Beam to initialize resources before processing a bundle of elements. While it can optimize resource usage and performance within each bundle, it doesn't inherently solve the backpressure issue caused by high message rates.

upvoted 1 times

MaxNRG 1 month, 1 week ago

D. Batch the job into ten-second increments: Batching messages can be an effective way to reduce backpressure. By grouping multiple messages into larger batches, you can reduce the frequency of external calls and distribute the processing load more evenly over time. This can lead to more efficient use of resources and potentially lower latency, as the system spends less time waiting on external services.

upvoted 1 times

MaxNRG 1 month, 1 week ago

Given these considerations, option D (Batch the job into ten-second increments) seems to be the most effective strategy for minimizing backpressure in your scenario. By batching messages, you can reduce the strain on your pipeline and external services, making the system more resilient and scalable under high load. However, the exact batch size and interval should be fine-tuned based on the specific characteristics of your workload and the capabilities of the external systems you are interacting with.

Additionally, it's important to consider other strategies in conjunction with batching, such as implementing efficient error handling, load balancing, and potentially using asynchronous I/O for external HTTP calls to further optimize performance and minimize backpressure.

upvoted 1 times

izekc 9 months ago

**Selected Answer: D**

Option C is not correct because it does not address the issue of backpressure. Creating a new object in the startBundle method of DoFn will not help to reduce the number of calls that are made to the service, which can lead to backpressure.

Here are some reasons why C is not correct:

Creating a new object in the startBundle method of DoFn is not a scalable solution. As the number of messages increases, the number of objects that need to be created will also increase. This can lead to performance problems and memory usage issues.

Creating a new object in the startBundle method of DoFn does not address the issue of backpressure. The service may still experience backpressure if the number of messages exceeds the service's capacity.

A better solution would be to use batching to reduce the number of calls that are made to the service. This can help to improve performance and reduce backpressure.

upvoted 1 times

✉  Oleksandr0501 9 months ago

gpt: Option C is a better approach as it allows for object creation to occur in a more controlled manner within the DoFn, potentially reducing t pressure on the system. However, it could still create a large number of objects depending on the rate of incoming messages.

Option D of batching the job into ten-second increments can also be a good solution to reduce backpressure on the system. This way, you can limit the number of messages being processed at any given time, which can help prevent bottlenecks and reduce the likelihood of backpressure.

Therefore, the best approach would be to combine options C and D, creating a new object in the startBundle method of a DoFn, and batching the job into smaller time increments, such as 10 seconds. This way, you can control the rate of object creation and processing, which can help minimize backpressure on the system.

upvoted 1 times

✉  Oleksandr0501 9 months ago

another vague question, as we see...

so, i'll choose D... if i get this test

"However, depending on the specifics of your use case, one option may be better suited than the other. For example, if you have a high volume of incoming messages with occasional spikes, option D of batching the job into smaller time increments may be more effective in managing the load. On the other hand, if the incoming messages are more evenly distributed over time, option C of creating a new object in the startBundle method of DoFn may be a better option.

Ultimately, it may be necessary to experiment with both approaches and determine which one works best for your specific use case."

upvoted 1 times

✉  juliobs 10 months ago

**Selected Answer: D**

D works.

Could be C, but who said that the pipeline is in Dataflow/Beam?

upvoted 1 times

✉  musumusu 11 months, 2 weeks ago

Answer C

batch increment in 10 sec, can improve load balancing, but overall back pressure (messages are generating more than consuming or publishing) in this case startBundle in DoFn or find other options in future like

caching,

load shedding (prioritising message flow),

message queuing

These options handle backpressure..

If your CPU is performing bad then go with change in batch increment timing

upvoted 1 times

✉  maci\_f 1 year ago

**Selected Answer: D**

I was hesitating between C and D, but then I realised this: <https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>

Here it says "If it's not thread-safe, create a new object in the startBundle method of DoFn." The task explicitly says "There will be tens of thousands of messages per second and that can be multi-threaded."

Correct me if I'm wrong, but multi-threaded == thread-safe. Therefore, no need to go for the C approach.

upvoted 3 times

✉  zellck 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>

For example, imagine a pipeline that's processing tens of thousands of messages per second in steady state. If you made a callout per element you would need the system to deal with the same number of API calls per second. Also, 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 3 times

✉  AzureDP900 1 year ago

D. Batch the job into ten-second increments.

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

C

C is an answer because

First of all, no doubt that we should avoid single call of element that's why we use multi-threading else it overwhelm an external service endpoint. To avoid this issue, batch calls to external systems.

Batch calls has also issue: GroupByKey transform or Apache Beam Timer API.

these approaches both require shuffling, which introduces some processing overhead as well as the need for a magic number to determine the key space.

Instead, use the StartBundle and FinishBundle lifecycle elements to batch your data. With these options, no shuffling is needed.

Source:

<https://cloud.google.com/dataflow/docs/tutorials/ecommerce-java#micro-batch-calls>

<https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>

Summary:

StartBundle and FinishBundle do batch with no shuffling

upvoted 1 times

✉  **AHUI** 1 year, 4 months ago

Ans C: reference <https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls>

upvoted 1 times

✉  **SMASL** 1 year, 4 months ago

**Selected Answer: C**

Based on the answers in this discussion thread, I would go for C. The most important link to support this choice is as following:

<https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls>

upvoted 2 times

✉  **Thobm** 1 year, 4 months ago

**Selected Answer: D**

Beam docs recommend batching

<https://beam.apache.org/documentation/patterns/grouping-elements-for-efficient-external-service-calls/>

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

C , It is straight forward , You can take a look at <https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>

Pattern : Calling external services for data enrichment

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Answer C

<https://cloud.google.com/blog/products/data-analytics/guide-to-common-cloud-dataflow-use-case-patterns-part-1>

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

<https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls>

To support choice C

upvoted 1 times

✉  **YorelNation** 1 year, 4 months ago

**Selected Answer: C**

i think you are right gg

upvoted 1 times

✉  **nwk** 1 year, 4 months ago

How about C?

<https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls>

upvoted 2 times

## Question #182

## Topic 1

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

### Correct Answer: B

Community vote distribution

C (100%)

 **zellck** Highly Voted 1 year, 2 months 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](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 10 times

 **musumusu** 11 months, 2 weeks ago

what is wrong with A, there is no cost constraint

upvoted 1 times

✉️  **AzureDP900** 1 year ago

Agreed

thx for sharing link

upvoted 1 times

✉️  **AWSandeep**  1 year, 4 months ago

**Selected Answer: C**

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

Use Gsutil when there is enough bandwidth to meet your project deadline for less than 1 TB of data. Storage Transfer Service is for much larger volumes for migration. Moreover, Cloud Data Fusion and Dataproc are not ideal for real-time updates. BigQuery Data Transfer Service does not support all on-prem sources.

upvoted 7 times

✉️  **TVH\_Data\_Engineer**  1 month ago

**Selected Answer: C**

Considering the requirement for handling large files and the need for real-time data integration, Option C (gsutil for the migration; Pub/Sub and Dataflow for the real-time updates) seems to be the most appropriate. gsutil will effectively handle the large file transfers, while Pub/Sub and Dataflow provide a robust solution for real-time data capture and processing, ensuring continuous updates to your warehouse on Google Cloud.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

**Selected Answer: C**

Option C is the best choice given the large file sizes for the initial migration and the need for real-time updates after migration.

Specifically:

gsutil can transfer large files in parallel over multiple TCP connections to maximize bandwidth. This works well for the 90GB files during initial migration.

Pub/Sub allows real-time messaging of updates that can then be streamed into Cloud Dataflow. Dataflow provides scalable stream processing to handle transforming and writing those updates into BigQuery or other sinks.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

Option A is incorrect because Storage Transfer Service is better for scheduled batch transfers, not ad hoc large migrations.

Option B is incorrect because BigQuery Data Transfer Service is more focused on scheduled replication jobs, not ad hoc migrations.

Option D would not work well for real-time updates after migration is complete.

So option C leverages the right Google cloud services for the one-time migration and ongoing real-time processing.

upvoted 1 times

✉️  **xiangbopopopo** 3 months ago

**Selected Answer: C**

agree with C

upvoted 1 times

✉️  **TNT87** 1 year, 4 months ago

[https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil\\_for\\_smaller\\_transfers\\_of\\_on-premises\\_data](https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil_for_smaller_transfers_of_on-premises_data)

Answer C

upvoted 4 times

✉️  **YorelNation** 1 year, 4 months ago

**Selected Answer: C**

C seems legit

upvoted 3 times

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

**Correct Answer: D**

*Community vote distribution*

B (50%)	A (38%)	13%
---------	---------	-----

✉  **John\_Pongthorn** Highly Voted 1 year, 4 months 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 it  
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 14 times

✉  **Mccloudgirl** 1 year, 2 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

✉  **zellck** Highly Voted 1 year, 2 months ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/bigtable/docs/schema-design#time-based>

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 10 times

👤 **datapassionate** Most Recent 2 weeks, 1 day ago

**Selected Answer: B**

B is a correct answer because "you need to access only the most recent stock prices"

"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."

<https://cloud.google.com/bigtable/docs/schema-design#time-based>

upvoted 1 times

👤 **TVH\_Data\_Engineer** 1 month ago

**Selected Answer: B**

B. One unique table for all indices, reverse timestamp as row key:

A single table for all indices keeps the structure simple.

Using a reverse timestamp as part of the row key ensures that the most recent data comes first in the sorted order. This design is beneficial for quickly accessing the latest data.

For example: you can convert the timestamp to a string and format it in reverse order, like "yyyyMMddHHmmss", ensuring newer dates and times are sorted lexicographically before older ones.

upvoted 1 times

👤 **kshehadyx** 4 months, 2 weeks ago

Correct Is B

upvoted 1 times

👤 **arien\_chen** 5 months, 1 week ago

**Selected Answer: D**

Option B using reverse timestamp only, this is not the answer.

the right answer should be using the index and reverse timestamp as the row key.

So, Option D is the only answer, because not A,B,C .

upvoted 4 times

👤 **Lanro** 6 months ago

**Selected Answer: B**

<https://cloud.google.com/bigtable/docs/schema-design#row-keys> - If you usually retrieve the most recent records first, you can use a reverse timestamp

B it is.

upvoted 1 times

👤 **Chom** 6 months, 3 weeks ago

**Selected Answer: A**

A is the answer

upvoted 1 times

👤 **vaga1** 7 months, 1 week ago

**Selected Answer: B**

the answer relies on whether the application needs to access the whole indexes at the same time or not. If yes then is B, if no is A.

in mind the answer is yes, so B makes more sense: I retrieve all the lists at the same time.

upvoted 1 times

👤 **ajdf** 7 months, 3 weeks ago

**Selected Answer: B**

<https://cloud.google.com/bigtable/docs/schema-design#time-based> 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 1 times

👤 **WillemHendr** 7 months, 3 weeks ago

**Selected Answer: B**

"access the data with the simplest query"

upvoted 1 times

👤 **Prudvi3266** 9 months, 1 week ago

**Selected Answer: A**

yes reverse time stamp is recommended to prevent hot spot. But our query pattern is we need most recent record the is easy when you use Timestamp and Also option a stating that our row key not starting with time stamp which is index#timestamp and which is the most efficient v for this scenario.

upvoted 3 times

👤 **midgooo** 10 months, 2 weeks ago

**Selected Answer: B**

Reversed timestamp will definitely help here.

upvoted 1 times

👤 **musumusu** 11 months, 1 week ago

Answer A:

I checked on other website and chatgpt also suggested A, to use index per stock and timestamp.

upvoted 1 times

👤 **niketd** 11 months, 1 week ago

**Selected Answer: B**

The key here is "only the most recent stock prices". Doesn't talk about accessing a specific index - so answer should be B

upvoted 3 times

👤 **kostol** 11 months, 2 weeks ago

**Selected Answer: A**

<https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid>

upvoted 1 times

👤 **desertlotus1211** 1 year ago

Answer is B: <https://stackoverflow.com/questions/65487550/bigtable-reverse-timestamp-advantage-over-regular-timestamp>

You want to access the MOST recent stock price First. Reverse Timestamp

upvoted 1 times

Question #184

Topic 1

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.
- 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.

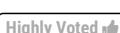
**Correct Answer: A**

*Community vote distribution*

C (79%)

12%

6%

👤 **NicolasN**  1 year, 2 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 17 times

✉  **NicolasN** 1 year, 2 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 the records from the staging table, loading the staging table, and then replacing the production table with the contents."

"When using the streaming API, the BigQuery streaming buffer remains active for about 30 to 60 minutes or more after use, which means that you can't delete or change data during that time. Since we used the streaming API, we scheduled the load every three hours to balance getting data into BigQuery quickly and being able to subsequently delete the data from the staging table during the load process."

upvoted 16 times

✉  **squishy\_fishy** 3 months, 1 week ago

I second this. At my work, I run into this exact streaming buffer thing, it will not let me delete the data until after 60 minutes.

upvoted 2 times

✉  **AzureDP900** 1 year ago

Agreed C is right

upvoted 1 times

✉  **nwk** Highly Voted 1 year, 4 months ago

Vote B - "Some recently streamed rows might not be available for table copy typically for a few minutes. In rare cases, this can take up to 90 minutes"

<https://cloud.google.com/bigquery/docs/streaming-data-into-bigquery#dataavailability>

upvoted 11 times

✉  **jkhong** 1 year, 1 month ago

Aren't there other aspects of data pipelining that we should be aware of? other than merely referring to the number of 'recommended' minutes stated in docs. B doesn't address how the appended data is subsequently deleted, since the table is append-only, the size will constantly grow, and so the user may unnecessarily incur more storage costs.

upvoted 1 times

✉  **devaid** 1 year, 3 months ago

A and B are discarded because the UPDATE statement, is not performance efficient. Neither appending more and more values to the staging table. It's better cleaning the staging table, and merging with the master dataset.

upvoted 4 times

✉  **MaxNRG** 1 month, 1 week ago

You can use BigQuery's features like MERGE to efficiently update the production table with only the new or changed data from the staging table, reducing processing time and costs.

upvoted 1 times

✉  **YorelNation** 1 year, 4 months ago

They don't seem concerned too much by data accuracy in the question

upvoted 1 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: A**

Not C nor D. Moving and deleting:

Deleting data from the staging table every 3 or 30 minutes could lead to data loss if the production table update fails, and it also requires more frequent and potentially resource-intensive operations.

Options C and D cause rebuilding of the staging table, which slows down ingestion, and may lose data if errors occur during recreation.

A or B

upvoted 1 times

MaxNRG 1 month, 1 week ago

When designing a report-only data warehouse in BigQuery, where data is streamed in and you have both staging and production tables, the key is to balance the frequency of updates with the performance needs of both the ingestion and reporting processes. Let's evaluate each option:

upvoted 1 times

MaxNRG 1 month, 1 week ago

A. Staging table as append-only, updating production every three hours: This approach allows for a consistent flow of data into the staging table without interruptions. Updating the production table every three hours strikes a balance between having reasonably fresh data and not overloading the system with too frequent updates. However, this may not be suitable if your reporting requirements demand more up-to-date data.

B. Staging table as append-only, updating production every ninety minutes: This is similar to option A but with a more frequent update cycle. This could be more appropriate if your reporting needs require more current data. However, more frequent updates can impact performance, especially during the update windows.

upvoted 1 times

MaxNRG 1 month, 1 week ago

C. Staging table moves data to production and clears staging every three hours: Moving data from staging to production and then clearing the staging table ensures that there is only one master dataset. However, this method might lead to more significant interruptions in data availability, both during the move and the clearing process. This might not be ideal if continuous access to the latest data is required.

D. Staging table moves data to production and clears staging every thirty minutes: This option provides the most up-to-date data in the production table but could significantly impact performance. Such frequent data transfers and deletions might lead to more overhead and could interrupt both the ingestion and reporting processes.

upvoted 1 times

MaxNRG 1 month, 1 week ago

Considering these options, A (Staging table as append-only, updating production every three hours) seems to be the most balanced approach. It provides a good compromise between having up-to-date data in the production environment and maintaining system performance. However, the exact frequency should be fine-tuned based on the specific performance characteristics of your system and the timeliness requirements of your reports.

It's also important to implement efficient mechanisms for transferring data from staging to production to minimize the impact on system performance. Techniques like partitioning and clustering in BigQuery can be used to optimize query performance and manage large datasets more effectively.

upvoted 1 times

Aman47 1 month, 2 weeks ago

Neither. In the current scenario, DataStream (a new Google resource) captures the CDC data and uses Dataflow to Replicate the changes to the destination table.

upvoted 1 times

👤 **hauhau** 1 year, 1 month ago

Selected Answer: B

Vote B

C : the doc says streaming data can be used up to 90 minutes not 3 hours

B : correct , insert staging table first with append  
and use merge from staging into production table

upvoted 2 times

👤 **hauhau** 1 year, 1 month ago

B just say "update", not specifically mention DML. update can be merge

upvoted 2 times

👤 **MaxNRG** 1 month, 1 week ago

You can use BigQuery's features like MERGE to efficiently update the production table with only the new or changed data from the staging table, reducing processing time and costs.

upvoted 1 times

👤 **zellck** 1 year, 2 months ago

Selected Answer: C

C is the answer.

<https://cloud.google.com/blog/products/data-analytics/moving-a-publishing-workflow-to-bigquery-for-new-data-insights>

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 records from the staging table, loading the staging table, and then replacing the production table with the contents.

When using the streaming API, the BigQuery streaming buffer remains active for about 30 to 60 minutes or more after use, which means that you can't delete or change data during that time. Since we used the streaming API, we scheduled the load every three hours to balance getting data into BigQuery quickly and being able to subsequently delete the data from the staging table during the load process.

upvoted 6 times

👤 **Atnafu** 1 year, 2 months ago

C

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 records from the staging table, loading the staging table, and then replacing the production table with the contents.

When using the streaming API, the BigQuery streaming buffer remains active for about 30 to 60 minutes or more after use, which means that you can't delete or change data during that time. Since we used the streaming API, we scheduled the load every three hours to balance getting data into BigQuery quickly and being able to subsequently delete the data from the staging table during the load process.

Building a script with BigQuery on the back end

<https://cloud.google.com/blog/products/data-analytics/moving-a-publishing-workflow-to-bigquery-for-new-data-insights>

upvoted 3 times

👤 **John\_Pongthorn** 1 year, 3 months ago

Selected Answer: C

D : read more on Streaming inserts and timestamp-aware queries as the following link

it is the same as this question exactly, but it is quite similar.

<https://cloud.google.com/blog/products/bigquery/performing-large-scale-mutations-in-bigquery>

read carefully in the content below.

When using timestamps to keep track of updated and deleted records, it's a good idea to periodically delete stale entries. To illustrate, the following pair of DML statements can be used to remove older versions as well as deleted records.

You'll notice that the above DELETE statements don't attempt to remove records that are newer than 3 hours. This is because data in BigQuery's streaming buffer is not immediately available for UPDATE, DELETE, or MERGE operations, as described in DML Limitations. These queries assume that the actual values for RecordTime roughly match the actual ingestion time.

upvoted 4 times

👤 **John\_Pongthorn** 1 year, 3 months ago

[https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#prune\\_merged\\_data](https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#prune_merged_data)

<https://cloud.google.com/bigquery/docs/reference/standard-sql/data-manipulation-language#limitations>

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

Either C or D But When will we delete stale data on staging table ? Every xxx????

[https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#prune\\_merged\\_data](https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#prune_merged_data)

Question #185

Topic 1

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
- C. Graph or pipeline construction
- D. Insufficient permissions

**Correct Answer: D**

*Community vote distribution*

B (100%)

✉  **AWSandeep**  1 year, 4 months 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 11 times

✉  **zelliCK**  1 year, 2 months ago

**Selected Answer: B**

B is the answer.

[https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#detect\\_an\\_exception\\_in\\_worker\\_code](https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#detect_an_exception_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.

upvoted 6 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: B**

The most likely cause of the errors you're experiencing in Dataflow, particularly if they are related to a particular DoFn (Dataflow's parallel processing operation), is B. Exceptions in worker code.

When a Dataflow job processes a few elements successfully before failing, it suggests that the overall job setup, permissions, and pipeline graph are likely correct, as the job was able to start and initially process data. However, if it fails during execution and the errors are associated with specific DoFn, this points towards issues in the code that executes within the workers. This could include:

1. Runtime exceptions in the code logic of the DoFn.
2. Issues handling specific data elements that might not be correctly managed by the DoFn code (e.g., unexpected data formats, null values, etc.).
3. Resource constraints or timeouts if the DoFn performs operations that are resource-intensive or long-running.

upvoted 2 times

MaxNRG 1 month, 1 week ago

To resolve these issues, you should:

1. Inspect the stack traces and error messages in the Dataflow monitoring interface for details on the exception.
2. Test the DoFn with a variety of data inputs, especially edge cases, to ensure robust error handling.
3. Review the resource usage and performance characteristics of the DoFn if the issue is related to resource constraints.

upvoted 2 times

vaga1 8 months, 2 weeks ago

**Selected Answer: B**

- A. Job validation - since it started successfully, it must have been validated.  
B. Exceptions in worker code - possible  
C. Graph or pipeline construction - same as A.  
D. Insufficient permissions - no elements to say that, and it should lead to invalidation.

upvoted 3 times

Atnafu 1 year, 2 months ago

C  
Code error  
upvoted 1 times

pluidust 1 year, 4 months ago

**Selected Answer: B**

B is correct

Question #186

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

A (76%)

14%

10%

AWSandeep Highly Voted 1 year, 4 months 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 can be further drilled-down via BigQuery views.

Reference:

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

upvoted 7 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: A**

For generating daily reports that show net consumption of Google Cloud compute resources and user details, the most efficient approach would be:

A. Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user.  
upvoted 1 times

MaxNRG 1 month, 1 week ago

Here's why this option is the most effective:

Integration with BigQuery: BigQuery is a powerful tool for analyzing large datasets. By exporting Cloud Logging data directly to BigQuery, you can leverage its fast querying capabilities and advanced analysis features.

Automated Daily Exports: Setting up automated daily exports to BigQuery streamlines the reporting process, ensuring that data is consistent and efficiently transferred.

Creating Views for Specific Filters: By creating views in BigQuery that filter data by project, log type, resource, and user, you can tailor the reports to the specific needs of your customer. Views also simplify repeated analysis by encapsulating complex SQL queries.

Efficiency and Scalability: This method is highly efficient and scalable, handling large volumes of data without the manual intervention required for CSV exports and data cleansing.

upvoted 1 times

MaxNRG 1 month, 1 week ago

Option B (exporting data in CSV format) and Option D (using Cloud Storage and Dataprep) are less efficient due to the additional steps of manual handling involved. Option C is similar to A but lacks the specificity of creating views directly in BigQuery for filtering, which is a more streamlined approach.

upvoted 1 times

Aman47 1 month, 2 weeks ago

You can choose a sink in which you want Cloud logging to continuously send Logging data. You can choose which columns you want to see (filter).

upvoted 1 times

Aman47 1 month, 2 weeks ago

Option C

upvoted 1 times

vaga1 8 months, 2 weeks ago

**Selected Answer: A**

B, C, D do not generate a daily scalable solution.

upvoted 3 times

Siant\_137 9 months ago

**Selected Answer: C**

I see A as quite inefficient as you are exporting ALL logs (hundreds of thousands) to bq and then filtering them with views. I would go for C, assuming that it does not involve doing it manually but rather creating a SINK with the correct filters and then using BQ Dataset as sink destination. But a lot of assumptions are taking place here as I believe the questions does not provide much context.

upvoted 2 times

midgoo 10 months, 2 weeks ago

**Selected Answer: A**

I almost got it wrong by choosing C. By doing C, that means we will manually filter first one by one. We should just import them all and filter using BigQuery

upvoted 2 times

maci\_f 1 year ago

**Selected Answer: A**

B and D do not consider the log type field.  
C looks good and I would go for it.

However, A looks equally good and I've found a CloudSkillsBoost lab that is exactly describing what answer A does, i.e. exporting logs to BQ and then creating a VIEW. <https://www.cloudskillsboost.google/focuses/6100?parent=catalog> I think the advantage of exporting complete log (i.e. filtering them after they reach BQ) is that in case we would want to adjust the reporting in the future, we would have the complete logs with all fields available, whereas with C we would need to take extra steps.

upvoted 3 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: A**

A is the answer.

upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

A

Bad exporting data in csv or json due lack of some data  
so export is best practice

1:10

<https://www.youtube.com/watch?v=ZyMO9XabUUM>

upvoted 1 times

✉  **Atnafu** 1 year, 1 month ago

You need to quickly and efficiently generate these daily reports by using Materialized view /View

A materialized view is the best solution and having filtered value with a view is good solution so A is an answer

upvoted 1 times

✉  **hauhau** 1 year, 2 months ago

**Selected Answer: A**

A because you filter data daily by view not just once by cloud logging

upvoted 1 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: A**

A. The D isn't filtering by log type. B and C are discarded because you need to drill down the exported logs in Big Query or other.  
upvoted 3 times

✉  **deavid** 1 year, 3 months ago

2nd tought: Definitely A. If you go to google documentation for export billing, you see a message that "Exporting to JSON or CSV is obsolete. Use Big Query instead".

Also why A? Look

<https://cloud.google.com/billing/docs/how-to/export-data-bigquery>

<https://cloud.google.com/billing/docs/how-to/bq-examples#total-costs-on-invoice>

You can make a fast report template al Data Studio that read a Big Query view.

upvoted 5 times

✉  **NicolasN** 1 year, 2 months ago

A comment regarding the links you provided (and not the correctness of the selected answer).

Using Cloud Billing is something different than detecting compute consumption data from Cloud Logging.

In fact, manual exporting to CSV (and JSON) is possible through the Logs Explorer interface (I think without user data break-down):

[https://cloud.google.com/logging/docs/view/logs-explorer-interface#download\\_logs](https://cloud.google.com/logging/docs/view/logs-explorer-interface#download_logs)

upvoted 1 times

✉  **AHUI** 1 year, 4 months ago

**Selected Answer: D**

The Google Cloud Storage bucket where you would like your reports to be delivered.

You can select any Cloud Storage bucket for which you are an owner, including buckets that are from different projects. This bucket must exist before you can start exporting reports and you must have owner access to the bucket. Google Cloud Storage charges for usage, so you should review the Cloud Storage pricesheet for information on how you might incur charges for the service.

<https://cloud.google.com/compute/docs/logging/usage-export>

upvoted 1 times

👤 TNT87 1 year, 4 months ago

Ans is C

[https://cloud.google.com/logging/docs/export/aggregated\\_sinks](https://cloud.google.com/logging/docs/export/aggregated_sinks)

D isn't correct because Cloud storage is used as a sink when logs are in json format not csv.

[https://cloud.google.com/logging/docs/export/aggregated\\_sinks#supported-destinations](https://cloud.google.com/logging/docs/export/aggregated_sinks#supported-destinations)

upvoted 4 times

👤 jkhong 1 year, 2 months ago

The question explicitly mentions daily generation of data so this would highlight, B and C seems that it is only suggesting a one-off filtering

upvoted 1 times

👤 TNT87 1 year ago

so wjhaytsa your argument about daily generartion of data??

upvoted 1 times

👤 TNT87 1 year, 4 months ago

On the other hand Ans A makes sense

<https://cloud.google.com/logging/docs/export/bigquery#overview>

upvoted 1 times

👤 changesu 1 year, 4 months ago

**Selected Answer: D**

Quickly and efficiently! It's a flag to guide to DataPrep. And importing data to Bigquery does not mean a report.

upvoted 2 times

👤 pluidust 1 year, 4 months ago

Why not C?

upvoted 1 times

👤 Wasss123 1 year, 4 months ago

why not D ?

upvoted 1 times

👤 Remi2021 1 year, 4 months ago

Challenging. B is right one but with B you do not automate which makes it hard, with A you ensure automation but there is no SQL support being mentioned which also makes me think that A is not the best choice.

upvoted 1 times

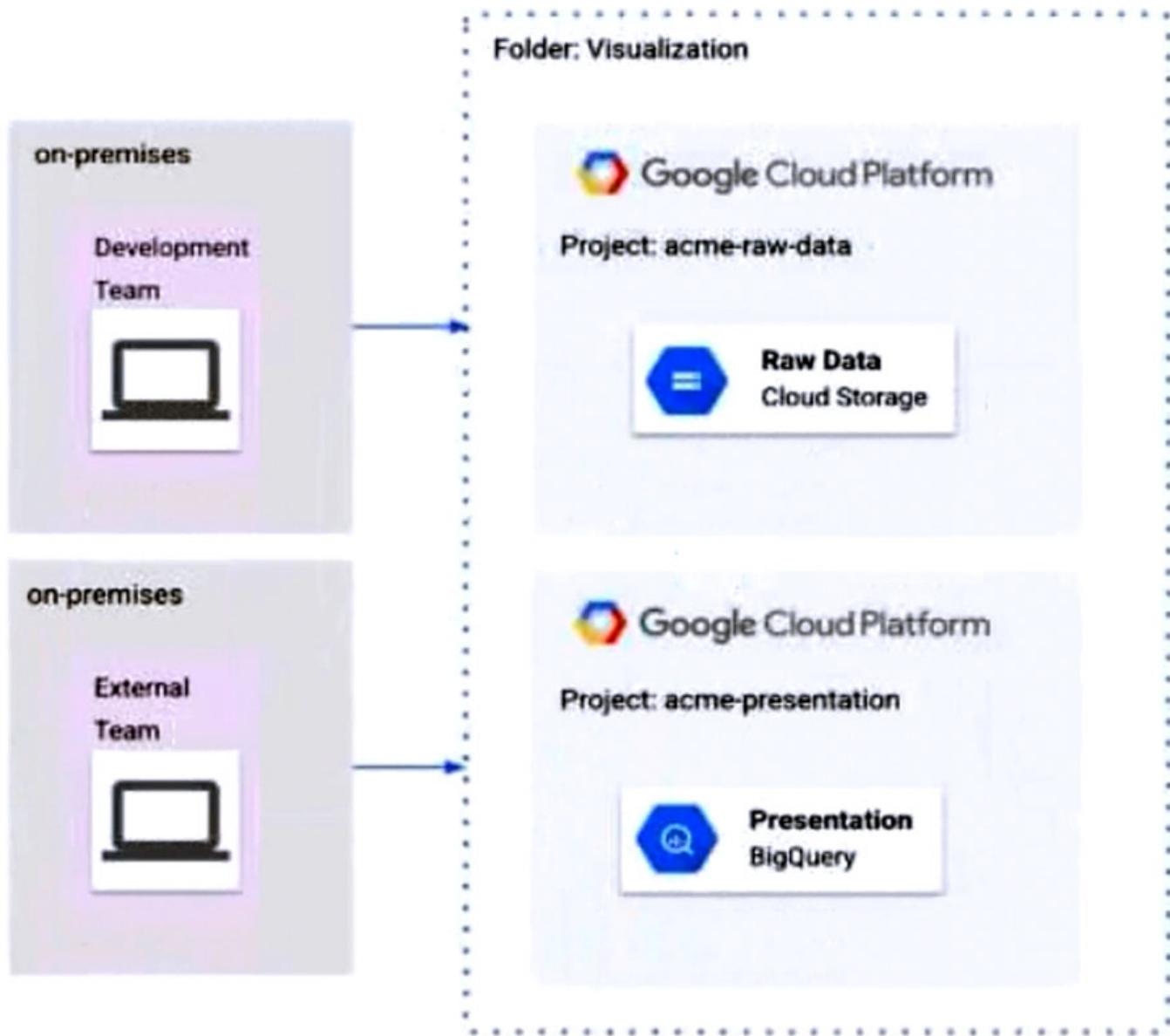
Question #187

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

D (85%)

C (15%)

**AWSandeep** Highly Voted 1 year, 4 months 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 16 times

✉ **Oleksandr0501** 8 months, 3 weeks ago

no, [https://cloud.google.com/blog/products/serverless/cloud-run-gets-enterprise-grade-network-security-with-vpc-sc?utm\\_source=youtube&utm\\_medium=unpaidsoc&utm\\_campaign=CDR\\_pri\\_gcp\\_m0v4tedeiao\\_ThisWeekInCloud\\_082621&utm\\_content=desc](https://cloud.google.com/blog/products/serverless/cloud-run-gets-enterprise-grade-network-security-with-vpc-sc?utm_source=youtube&utm_medium=unpaidsoc&utm_campaign=CDR_pri_gcp_m0v4tedeiao_ThisWeekInCloud_082621&utm_content=desc)

upvoted 1 times

✉ **Oleksandr0501** 8 months, 3 weeks ago

damn, i am confused anyway. Can be D.

upvoted 1 times

✉ **Oleksandr0501** 8 months, 3 weeks ago

should be D, as i think now, because we create a "magic bulb" around around Cloud storage and Dev team, and it will be protected from external influence like a human cell. Meantime Dev team will still be able to access Bigquery. But external team will not manage access Cloud storage.

upvoted 1 times

✉ **TNT87** 1 year, 4 months ago

WHy do you have to put the development team at the access perimeter???

upvoted 1 times

✉ **maci\_f** Highly Voted 1 year ago

**Selected Answer: D**

"The grouping of GCP Project(s) and Service API(s) in the Service Perimeter result in restricting unauthorized access outside of the Service Perimeter to Service API endpoint(s) referencing resources inside of the Service Perimeter."

<https://scalesec.com/blog/vpc-service-controls-in-plain-english/>

Development team: needs to access both Cloud Storage and BQ -> therefore we put the Development team inside a perimeter so it can access both the Cloud Storage and the BQ

External team: allowed to access only BQ -> therefore we put Cloud Storage behind the restricted API and leave the external team outside of perimeter, so it can access BQ, but is prohibited from accessing the Cloud Storage

upvoted 8 times

✉ **Aman47** Most Recent 1 month, 2 weeks ago

Comments are saying it correct its C

upvoted 1 times

✉ **Mamko** 4 months, 1 week ago

It's D for sure

upvoted 1 times

✉ **techabhi2\_0** 4 months, 3 weeks ago

A - Simple and straight forward

upvoted 4 times

✉ **wan2three** 5 months, 3 weeks ago

Why not B, I think CD will cause one of the team can not reach one or two of those DBs. A is not correct either

upvoted 1 times

✉ **[Removed]** 5 months, 3 weeks ago

**Selected Answer: D**

D. VPC Service Controls can create a service perimeter and define a restrictive API (service to protect). In this case, two projects are inside the perimeter and Cloud Storage is defined as the restrictive API. This means only services running on these two projects can access the Cloud Storage. And to allow users to access the Cloud Storage, they need have the access to the service perimeter. Hence, D is the correct answer.

upvoted 3 times

✉ **izekc** 9 months ago

**Selected Answer: C**

C is correct

upvoted 1 times

✉ **midgoo** 10 months, 2 weeks ago

**Selected Answer: D**

D sounds more correct, but if the project is already in the Service Control, would External people can access the BigQuery dataset in that project?

upvoted 1 times

👤 [Removed] 1 year ago

seriously why u guys use VPC? the question never mentioned VPN or Interconnect, how can on-premise use VPC?

A is the answer.

upvoted 3 times

👤 zellck 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

upvoted 1 times

👤 TNT87 1 year ago

Answer C, i dnt know if you have studied cloud security? if you have you will know

upvoted 1 times

👤 Oleksandr0501 8 months, 3 weeks ago

you might be correct. C.

[https://cloud.google.com/blog/products/serverless/cloud-run-gets-enterprise-grade-network-security-with-vpc-sc?utm\\_source=youtube&utm\\_medium=unpaidsoc&utm\\_campaign=CDR\\_pri\\_gcp\\_m0v4tedeiao\\_ThisWeekInCloud\\_082621&utm\\_content=cription](https://cloud.google.com/blog/products/serverless/cloud-run-gets-enterprise-grade-network-security-with-vpc-sc?utm_source=youtube&utm_medium=unpaidsoc&utm_campaign=CDR_pri_gcp_m0v4tedeiao_ThisWeekInCloud_082621&utm_content=cription)

[https://www.youtube.com/watch?v=ABIY7FexJJI&ab\\_channel=GoogleCloudTech](https://www.youtube.com/watch?v=ABIY7FexJJI&ab_channel=GoogleCloudTech)

upvoted 1 times

👤 Atnafu 1 year, 2 months ago

C

Extend perimeters to authorized VPN or Cloud Interconnect

You can configure private communication to Google Cloud resources from VPC networks that span hybrid environments with Private Google Access on-premises extensions. A VPC network must be part of a service perimeter for VMs on that network to privately access managed Google Cloud resources within that service perimeter.

<https://cloud.google.com/vpc-service-controls/docs/overview#internet>

upvoted 2 times

👤 Atnafu 1 year, 2 months ago

I meant D Not C

upvoted 2 times

👤 cloudmon 1 year, 2 months ago

**Selected Answer: D**

D makes the most sense to me

upvoted 4 times

👤 cloudmon 1 year, 2 months ago

Because "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."

upvoted 2 times

👤 cloudmon 1 year, 2 months ago

Therefore, Cloud Storage should be the restricted API, and you add the Development Team users to the perimeter's Access Level to allow them to access the restricted API.

upvoted 3 times

👤 yu\_ 1 year, 2 months ago

why C?

I thought the development team would not be able to access BigQuery as I would include BigQuery in the service perimeter and add External Team to the access level

upvoted 1 times

👤 jkhong 1 year, 2 months ago

Exactly, why would we need to consider BigQuery as a restricted service when it can already be accessed by both Dev and External team. The restricted service we are concerned with is Cloud Storage. If we go with C, we are only adding the external team into the access level. this means that the development team still wouldn't be able to access it

upvoted 2 times

✉  **josrojgra** 1 year, 2 months ago

**Selected Answer: C**

Answer C

<https://cloud.google.com/vpc-service-controls/docs/overview#isolate>

upvoted 1 times

✉  **TNT87** 1 year, 3 months ago

**Selected Answer: C**

Answer C

<https://cloud.google.com/vpc-service-controls/docs/overview#isolate>

upvoted 4 times

✉  **Wasss123** 1 year, 4 months ago

Should be C

<https://cloud.google.com/vpc-service-controls/docs/vpc-accessible-services>

When configuring VPC accessible services for a perimeter, you can specify a list of individual services, as well as include the RESTRICTED-SERVICES value, which automatically includes all of the services protected by the perimeter.

To ensure access to the expected services is fully limited, you must:

Configure the perimeter to protect the same set of services that you want to make accessible.

Configure VPCs in the perimeter to use the restricted VIP.

Use layer 3 firewalls.

upvoted 6 times

Question #188

*Topic 1*

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

A (68%)

D (33%)

✉  **AWSandeep** Highly Voted 1 year, 4 months 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 9 times

 **izekc** Highly Voted 9 months ago

**Selected Answer: D**

Although A is good, but concerning about the cost. Then D will be much more suitable

upvoted 7 times

 **tibuenoc** Most Recent 2 weeks ago

**Selected Answer: D**

I think is D

The best for Web app is Cloud SQL, and Spanner is the best for data more than 30GB

upvoted 1 times

 **MaxNRG** 1 month, 1 week ago

**Selected Answer: A**

A - This option allows for optimization for cost initially with a single region Cloud Spanner instance, and then optimization for global presence performance after funding with multi-region instances.

Cloud Spanner supports native JDBC drivers and is horizontally scalable, providing very high performance. A single region instance minimizes costs initially. After funding, multi-region instances can provide lower latency and high availability globally.

Cloud SQL does not scale as well and has higher costs for multiple high availability regions. Bigtable does not support JDBC drivers natively. Therefore, Spanner is the best choice here for optimizing both for cost initially and then performance and availability globally post-funding.

upvoted 2 times

 **lucaluca1982** 10 months, 1 week ago

Spanner has some limitations with JDBC. Maybe the question wants to help us to choose Cloud SQL

upvoted 3 times

 **musumusu** 11 months, 1 week ago

Answer D:

Cost effective transactional database Cloud SQL. Spanner is good case for data more than 30 GB

upvoted 1 times

 **odacir** 1 year, 1 month ago

**Selected Answer: A**

B and C has no sense because of the driver.

D looks like a good option, but HA it's not to improve performance or global presence:

The purpose of an HA configuration is to reduce downtime when a zone or instance becomes unavailable. This might happen during a zonal outage, or when an instance runs out of memory. With HA, your data continues to be available to client applications.

So the best option is A.

upvoted 5 times

 **TNT87** 1 year, 4 months ago

**Selected Answer: A**

<https://cloud.google.com/spanner/docs/jdbc-drivers>

Ans A

[https://cloud.google.com/spanner/docs/instance-configuration#tradeoffs Regional\\_versus\\_Multi-Region\\_Configurations](https://cloud.google.com/spanner/docs/instance-configuration#tradeoffs Regional_versus_Multi-Region_Configurations)

The last part of the question makes it easy

upvoted 5 times

✉  **TNT87** 1 year, 4 months ago

Yes Spanner is expensive , but the question expresslly states that "after securing funding you want to have a global presence" the word global is repeatedly stated there.

Answer is A.

upvoted 3 times

✉  **TNT87** 1 year, 4 months ago

[https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs Regional\\_versus\\_multi-region\\_configurations](https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs Regional_versus_multi-region_configurations)

Ans A

upvoted 1 times

✉  **badrisrinivas9** 1 year, 4 months ago

**Selected Answer: D**

Spanner is expensive, they haven't mentioned the size of db... optimize for cost then option is Cloud SQL which cost effective and highly available in case of multi region.

upvoted 3 times

✉  **Quevedo** 1 year, 4 months ago

**Selected Answer: A**

A is the best option. It is globally scalable and it also meets the cost goal as it says that initially it will be configurated as single region which is cheaper than multi region.

upvoted 3 times

✉  **YorelNation** 1 year, 4 months ago

**Selected Answer: D**

Spanner is expensive can't be A

Would choose D

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

The fact that its global cloud spanner is the answer. Secondly Option D, the fact that it has to be highly available and multi regional its already more expensive than Cloud spanner Regional instance. [https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs Regional\\_versus\\_multi-region\\_configurations](https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs Regional_versus_multi-region_configurations)

upvoted 2 times

✉  **YorelNation** 1 year, 4 months ago

Actually maybe C as you don't really need relational database for a webapp and BigTable is super performant and highly available

upvoted 1 times

✉  **TNT87** 1 year, 3 months ago

no its A

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: A**

Spanner still support JDBC

<https://cloud.google.com/spanner/docs/jdbc-drivers>

upvoted 3 times

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.
- 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.

**Correct Answer: A**

*Community vote distribution*

A (64%)

B (36%)

 **MaxNRG** 1 month, 1 week ago

**Selected Answer: A**

Cloud Interconnect provides a dedicated private connection between on-prem and Google Cloud for high bandwidth (up to 100 Gbps) and low latency. This facilitates large, fast data transfers.

Storage Transfer Service supports parallel data transfers over Cloud Interconnect. It can transfer petabyte-scale datasets faster by transferring objects in parallel.

Storage Transfer Service uses HTTPS encryption in transit and at rest by default for secure data transfers.

It follows Google-recommended practices for large data migrations vs ad hoc methods like gsutil or scp.

The other options would take too long for a 1 PB transfer (VPN capped at 3 Gbps, manual transfers) or introduce extra steps like batching and checksums. Cloud Interconnect + Storage Transfer is the recommended Google solution.

upvoted 2 times

 **LanaOjisan** 3 months ago

It is believed that A.

One reason is that for "secure" and "in a few hours," the communication can be done securely using a direct physical line without going through an ISP. Also, depending on the case, in the case of "Dedicated Interconnect," the maximum transfer can be as high as 200 Gbps, and the fast data transfer of 1 PB can be completed in 11 hours.

Therefore, A.

upvoted 2 times

 **arien\_chen** 5 months, 1 week ago

**Selected Answer: A**

A

<https://cloud.google.com/storage-transfer/docs/transfer-options#:~:text=Transferring%20more%20than%201%20TB%20from%20on%2Dpremises>

upvoted 3 times

 **knith66** 6 months 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** 6 months ago

This link has additional clarity

<https://cloud.google.com/network-connectivity/docs/interconnect/concepts/terminology>

upvoted 1 times

✉  **vaga1** 7 months ago

**Selected Answer: B**

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

<https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#time>

upvoted 2 times

✉  **knith66** 6 months ago

Transfer Appliance is a slow process. wont be able to do in few hours

upvoted 3 times

✉  **Oleksandr0501** 9 months ago

**Selected Answer: A**

gpt: Based on security and speed, if the data is highly sensitive and security is the top priority, then option B (using a Transfer Appliance) may be a better choice. Transfer Appliance uses hardware encryption to transfer data and is designed to securely transfer large amounts of data. However, if speed is the primary concern, then option A (using Cloud Interconnect and Storage Transfer Service) may be a better choice as it allows for faster transfer speeds over a dedicated and secure connection. It ultimately depends on the specific needs and priorities of the organization.

A vague teaky question. Bad author of it...

B is also good. As were said in discuss. by smb, a question asks "safe connection", so - a Cloud Interconnect (A)

upvoted 1 times

✉  **midgoo** 10 months, 2 weeks ago

**Selected Answer: B**

Either this question is very tricky or very poor written. It says 'Data transfer time during the migration should take only a few hours'. We should not add the 20days for overhead time for Appliance into the total time of migration.

If 'a few hours' = 30hours or more, A will be good enough.

If 'a few hours' = 10 or less, B is the only way (with multiple devices to copy at the same time)

upvoted 3 times

✉  **spicebits** 2 months, 3 weeks ago

B can't be the answer - You have to wait 25 days to receive the appliance and another 25 days to get the appliance back and data loaded cloud storage: <https://cloud.google.com/transfer-appliance/docs/4.0/overview#transfer-speeds>

upvoted 1 times

✉  **Nandhu95** 10 months, 2 weeks ago

**Selected Answer: A**

Expected time via transfer appliance is around 20 days , and achieving the same using Storage transfer service with highest bandwidth of 100Gbps is 30 hrs, so hence its been asked for hrs .. its A

Acquiring a Transfer Appliance is straightforward. In the Google Cloud console, you request a Transfer Appliance, indicate how much data you have, and then Google ships one or more appliances to your requested location. You're given a number of days to transfer your data to the appliance ("data capture") and ship it back to Google.

The expected turnaround time for a network appliance to be shipped, loaded with your data, shipped back, and rehydrated on Google Cloud 20 days. If your online transfer timeframe is calculated to be substantially more than this timeframe, consider Transfer Appliance. The total cost for the 300 TB device process is less than \$2,500.

upvoted 1 times

✉  **vaga1** 7 months ago

it says data transfer during the migration. It mean from when the migration is "activated", which means from when the Transfer Appliance device is plugged and ready to be used

upvoted 1 times

✉  **wjtb** 10 months, 3 weeks ago

**Selected Answer: B**

Even with 100gbps bandwith, you will not reach a data transfer time within the range of "hours" for 1PB. Transfer appliance is the way to go. <https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#time>

upvoted 2 times

✉  **musumusu** 11 months, 2 weeks ago

Answer A,

One time transfer is cheaper and less secure always using Transfer Appliance.  
you need to do it in faster way, set up Interconnect speed limit is 50mbps - 10GBps  
and Transfer Appliance speed can goes up to 40GBps.

I am choosing A for security concern only.

upvoted 1 times

✉  **AzureDP900** 1 year ago

A is right

upvoted 1 times

✉  **AzureDP900** 1 year ago

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

upvoted 1 times

✉  **zellick** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#storage-transfer-service-for-large-transfers-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 2 times

✉  **Atnafu** 1 year, 2 months ago

B

<https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#:~:text=Few%20things%20in,not%20be%20obtained.>

upvoted 2 times

✉  **Atnafu** 1 year, 2 months ago

B

It takes 30hrs with 100Gbps bandwidth- more than a day to transfer

<https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#:~:text=addresses%20or%20NATs.-,Online%20versus%20offline%20transfer,A%20certain%20amount%20of%20management%20overhead%20is%20built%20into%20these%20calculations.,-As%20noted%20earlier>

upvoted 4 times

✉  **deavid** 1 year, 3 months ago

**Selected Answer: A**

Well it doesn't mention anything about not enough bandwidth to meet your project deadline. I guess you can assume they have 200GBps+ bandwidth, otherwise it shouldn't take only a few hours.

upvoted 4 times

✉  **pluiedust** 1 year, 4 months ago

**Selected Answer: A**

A is correct.

upvoted 1 times

✉  **bigquery1102** 1 year, 4 months ago

**Selected Answer: A**

A is correct

[https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer\\_appliance\\_for\\_larger\\_transfers](https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer_appliance_for_larger_transfers)

upvoted 1 times

✉  **MounicaN** 1 year, 4 months ago

**Selected Answer: A**

Huge data can be migrated over cloud interconnect

upvoted 1 times

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.
- 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.

**Correct Answer: D**

*Community vote distribution*

A (88%)

12%

✉️  **saurabhsgingh4k**  1 year, 1 month ago

**Selected Answer: A**

I'm kinda inclined towards C as SQL seems a powerful option to treat this kind of use case.

Also, I didn't get how the transformations mentioned on this page will help to clean the data ([https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported\\_transformations](https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported_transformations))

But I guess using Wrangler plugin, this kind of stuff can be done on DataFusion, also the question talks about an pipeline, so A is the final chc upvoted 5 times

✉️  **MaxNRG**  1 month, 1 week ago

**Selected Answer: A**

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making it accessible to a wider range of users.

Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:

Data type conversions

Data cleansing (e.g., removing invalid characters, correcting formatting)

Data validation (e.g., checking for missing values, enforcing constraints)

Data enrichment (e.g., adding derived fields, joining with other datasets)

Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.

Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.

Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

Why other options are less suitable:

B. Converting to AVRO: While AVRO is a self-describing format, it doesn't inherently address data quality issues. Transformations would st be needed, and Data Fusion provides a more comprehensive environment for this.

C. Staging table: Requires manual SQL transformations, which can be time-consuming and error-prone for large datasets with complex da quality issues.

D. Transformations in place: Modifying data directly in the destination table can lead to data loss or corruption if errors occur. It's generally safer to keep raw data intact and perform transformations separately.

By using Data Fusion, you can create a robust and efficient data pipeline that addresses data quality issues upfront, ensuring that only cle and consistent data is loaded into BigQuery for accurate analysis and insights.

upvoted 1 times

✉  **squishy\_fishy** 3 months, 1 week ago

The answer is C. That is what we do at work. We have landing/staging table, sort table and deliver table, upvoted 3 times

✉  **squishy\_fishy** 3 months, 1 week ago

Okay, second thought, it is asking for a pipeline, so the answer should be A. At work, we use dataflow inside the composer to build a pipeline injecting data into landing/staging table, then transform/clean data in the sort table, then send the cleaned data to deliver table. upvoted 3 times

✉  **phidelics** 7 months, 3 weeks ago

**Selected Answer: A**

Keyword: Data Pipeline

upvoted 3 times

✉  **mialli** 8 months, 4 weeks ago

**Selected Answer: A**

same as @saurabh Singh4k

upvoted 2 times

✉  **Adswerve** 9 months, 2 weeks ago

**Selected Answer: C**

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

upvoted 3 times

✉  **musumusu** 11 months, 2 weeks ago

Answer C,

Data Fusion is costly and current transformation is just a cast transformation in a column.

I guess no one wants to pay for Data Fusion for this little transformation and Staging table processing handles such minor cleaning.

upvoted 3 times

✉  **maci\_f** 1 year ago

**Selected Answer: A**

Data Fusion enables changing the data type directly as shown in this lab: <https://www.cloudskillsboost.google/focuses/25335?parent=catalog>  
Wrangler is the feature to enable that, as already mentioned: <https://stackoverflow.com/questions/58699872/google-cloud-data-fusion-how-to-change-datatype-from-string-to-date>

upvoted 4 times

✉  **Mike422** 1 year ago

Apparently ChatGPT thinks C is the correct answer just saying (for the same reason that @saurabh Singh4k wrote)

upvoted 2 times

✉  **Atnafu** 1 year, 1 month ago

A

<https://cloud.google.com/data-fusion/docs/concepts/overview#:~:text=The%20Cloud%20Data%20Fusion%20web%20UI%20lets%20you%20to%20build%20scalable%20data%20integration%20solutions%20to%20clean%2C%20prepare%2C%20blend%2C%20transfer%2C%20and%20transform%20data%20without%20having%20to%20manage%20the%20infrastructure.>

upvoted 1 times

✉  **zellck** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/data-fusion/docs/concepts/overview>

Cloud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines.

The Cloud Data Fusion web UI lets you to build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.

upvoted 4 times

✉  **AzureDP900** 1 year ago

thx for sharing link

upvoted 1 times

✉  **samirzubair** 1 year, 2 months ago

The Correct Ans is C

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

although this is my preferred answer. this doesn't satisfy how this becomes a pipeline.

upvoted 1 times

✉  **hiromi** 1 year, 2 months ago

**Selected Answer: A**

Data Fusion

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: A**

Ans A

[https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported\\_transformations](https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported_transformations)

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: A**

A is correct for me

upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: A**

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

upvoted 1 times

Question #191

Topic 1

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.
- 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.

**Correct Answer: A**

*Community vote distribution*

B (95%)

5%

✉  **John\_Pongthorn**  1 year, 4 months 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 10 times

 **John\_Pongthorn** 1 year, 4 months ago

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

upvoted 2 times

 **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: B**

Option B, the model learned to listen to too much stuff/noise. We need to reduce it, by decreasing the number of input feature, and we need to give the model more data, by increasing the amount of training data

upvoted 1 times

 **NeoNitin** 6 months 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 patterns and make better predictions. These input features could include things like the customer's age, past purchase history, browsing behavior, or any other relevant data that might impact their buying likelihood.

upvoted 1 times

 **vaga1** 8 months, 2 weeks ago

**Selected Answer: B**

A. can be a solution for a specific case, but it is not the academic answer as we do not know the quantity and proportion between them of  $n$  &  $k$  added. More records and more variables together can lead to even more overfitting due also to the curse of dimensionality. Adding a variable is much more impactful than records.

B. just more records can lead to a more robust estimation and fewer variables certainly lead to at most the same estimation, but potentially reduce the fit on the training set.

C. reduce  $n$  in favor of  $k$  is never a choice, it is against logic and it will lead to more overfitting.

D. decrease both will reduce overfitting for sure but at the price of losing robustness on the model predictive power

upvoted 1 times

 **AzureDP900** 1 year ago

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

upvoted 1 times

 **pluidust** 1 year, 4 months ago

**Selected Answer: B**

B is correct

upvoted 2 times

 **TNT87** 1 year, 4 months ago

Answer B

<https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/>

upvoted 2 times

 **HarshKothari21** 1 year, 4 months ago

**Selected Answer: B**

Option B

Feature selection is the one the ways to resolve overfitting. Which means reducing the features when the size of the training data is small, then the network tends to have greater control over the training data. so increasing the size of data would help.

upvoted 3 times

 **YorelNation** 1 year, 4 months ago

**Selected Answer: B**

Best option is not mentioned: generalize your neural net by decreasing the complexity of its structure.

A part from that I guess you could remove some features and increase the size of the training dataset ==> B

upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: B**

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

Sorry, B is right. Read through extensive best-practices on ML.

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: D**

D is correct

upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

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

Reveal Solution

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: B**

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

upvoted 1 times

Question #192

Topic 1

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-code 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.

**Correct Answer: D**

*Community vote distribution*

D (88%)

13%

✉  **PhuocT** Highly Voted 1 year, 4 months 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 11 times

MaxNRG Most Recent 1 month, 1 week ago

**Selected Answer: D**

The best option would be to use Dialogflow to implement the chatbot, defining the intents based on the most common queries collected.

Dialogflow is a conversational AI platform that allows for easy implementation of chatbots without needing to code. It has built-in integration for both text and voice input via APIs like Cloud Speech-to-Text. Defining intents and entity types allows you to map common queries and key words to responses. This would provide a low/no-code way to quickly build and iteratively improve the chatbot capabilities.

Option A and B would require more heavy coding to handle speech input/output. Option C still requires coding the complex query handling. Option D leverages the full capabilities of Dialogflow to enable no-code chatbot development and ongoing improvements as more conversational data is collected. Hence, option D is the best approach given the requirements.

upvoted 1 times

Lanro 6 months ago

**Selected Answer: D**

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

upvoted 1 times

zellck 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/dialogflow/docs>

Dialogflow is a natural language understanding platform that makes it easy to design and integrate a conversational user interface into your mobile app, web application, device, bot, interactive voice response system, and so on. Using Dialogflow, you can provide new and engaging ways for users to interact with your product.

Dialogflow can analyze multiple types of input from your customers, including text or audio inputs (like from a phone or voice recording). It can also respond to your customers in a couple of ways, either through text or with synthetic speech.

upvoted 4 times

AzureDP900 1 year ago

Agree with D

upvoted 1 times

Atnafu 1 year, 2 months ago

D

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

upvoted 1 times

deavid 1 year, 3 months ago

**Selected Answer: D**

D definitely, as the documentation says (specially that you can call the detect Intent method for audio inputs):

<https://cloud.google.com/dialogflow/es/docs/how/detect-intent-tts>

Also Speech-To-Text API does nothing more than translate.

upvoted 4 times

TNT87 1 year, 4 months ago

Answer D

<https://cloud.google.com/dialogflow/es/docs/how/detect-intent-tts>

upvoted 4 times

nwk 1 year, 4 months ago

<https://cloud.google.com/dialogflow/es/docs/how/detect-intent-stream>

Vote D

upvoted 2 times

ducc 1 year, 4 months ago

**Selected Answer: C**

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

This seems the best answer here but not the best answer in real world.

But with the Question, the answer must be the combination of both Dialogflow and Speech API

upvoted 3 times

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.

**Correct Answer: B**

*Community vote distribution*

D (81%)

B (19%)

✉  **beanz00**  1 year, 3 months 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 read the file. You would have to create something custom.

upvoted 17 times

✉  **deavid**  1 year, 3 months 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 12 times

✉  **MaxNRG**  1 month, 1 week ago

**Selected Answer: D**

Option D is the best approach given the constraints - use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format.

The key reasons:

- Dataflow provides managed resource scaling for efficient stream processing
- Avro format has schema evolution capabilities and efficient serialization for flight telemetry data
- Apache Beam connectors avoid having to write much code to integrate proprietary data sources
- Streaming inserts data efficiently compared to periodic batch jobs

In contrast, option A uses Cloud Functions which lack native streaming capabilities. Option B stores data in less efficient JSON format. Option C uses Dataproc which requires manual cluster management.

So leveraging Dataflow + Avro + Beam provides the most efficient way to stream proprietary flight data into BigQuery while using minimal resources.

upvoted 1 times

✉  **Aman47** 1 month, 2 weeks ago

It's talking about streaming? none of the options talk about triggering a load to begin. We need a trigger or schedule to run first.

upvoted 1 times

✉  **Ajose0** 3 months, 2 weeks ago

**Selected Answer: D**

Option D allows you to use a custom connector to read the proprietary data format and write the data to BigQuery in Avro format.

upvoted 2 times

✉  **sergiomujica** 4 months, 3 weeks ago

**Selected Answer: D**

the keyword is streaming

upvoted 1 times

✉  **knith66** 6 months 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 3 times

✉  **musumusu** 11 months, 2 weeks ago

Answer is D ,  
Why not B, changing data format before uploading to bigquery is good approach.  
upvoted 1 times

✉  **cetanx** 1 year ago

**Selected Answer: B**  
I believe keyword here is "An aerospace company uses a proprietary data format"  
So if we list the connectors available in Apache Beam, we are listed with these options;  
<https://beam.apache.org/documentation/io/connectors/>

So I believe, we have to create our own custom connector to read from the proprietary data format hence the answer should be B  
upvoted 1 times

✉  **cetanx** 1 year ago

sorry the answer should be D  
upvoted 1 times

✉  **AzureDP900** 1 year ago

D is right  
upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: D**  
D is the answer.  
upvoted 3 times

✉  **TNT87** 1 year ago

There is dataflow connector and D isn't cost effective  
upvoted 1 times

✉  **hauhau** 1 year, 1 month ago

**Selected Answer: B**  
B is the most efficient  
upvoted 2 times

✉  **TNT87** 1 year, 3 months ago

<https://cloud.google.com/spanner/docs/change-streams/use-dataflow#core-concepts>  
upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Ans B

<https://cloud.google.com/architecture/streaming-avro-records-into-bigquery-using-dataflow>

Is there a reason to use apache beam connector yet there is dataflow which is a standard solution for that scenario?

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

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

upvoted 1 times

✉  **learner2610** 1 year, 4 months ago

Can standard dataflow be used to ingest any proprietary format of file ?

shouldn't we use custom apache beam connector ?

So I think it is D ,though it isn't simple ,But in this scenario they have asked to use less resources to import data

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Option D streams, that's not cost effective. We need something that is cost effective, hence B is the option

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

I mean that consumes fewer resources

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

Do you mind reading the links i provided and revisiting the question, then you will understand why D isn't the best. Why use Apache beam yet there is Dataflow

upvoted 1 times

✉  **[Removed]** 1 year, 4 months ago

Can Bigquery handle a proprietary file format?

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

BigQuery uses a proprietary format because it can evolve in tandem with the query engine, which takes advantage of deep knowledge of the data layout to optimize ...

upvoted 1 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: D**

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

Reveal Solution

upvoted 3 times

Question #194

Topic 1

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.
- 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.

**Correct Answer: D**

*Community vote distribution*

A (92%)

8%

✉  **squishy fishy** 3 months, 1 week ago

— **squishy\_fishy** 8 months, 1 week ago

Answer is D, at work we use solution A for low volume of Pub/Sub messages and Cloud function, and using D Composer for high volume Pub/Sub messages.

upvoted 2 times

✉ **lucaluka1982** 10 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 4 times

✉ **musumusu** 11 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 subscriber (cloud function) that is connected to the python API (internal application) that makes the purchase.

upvoted 3 times

✉ **AzureDP900** 1 year ago

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

upvoted 1 times

✉ **zellick** 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

upvoted 3 times

✉ **GCPCloudArchitectUser** 1 year, 2 months ago

**Selected Answer: A**

Because trading platform requires securely transmission to queuing

If you use cloud compose then we need some other job to trigger composer ... would that be cloud composer api or cloud function ...

upvoted 4 times

✉ **TNT87** 1 year, 4 months ago

<https://cloud.google.com/functions/docs/calling/pubsub>

upvoted 1 times

✉ **TNT87** 1 year, 4 months ago

**Selected Answer: A**

Ans A

<https://cloud.google.com/functions/docs/calling/pubsub#deployment>

upvoted 3 times

✉ **YorelNation** 1 year, 4 months ago

**Selected Answer: A**

A because D is stupidly high latency

upvoted 2 times

✉ **nwk** 1 year, 4 months ago

Vote A, can't see the need for composer

upvoted 1 times

✉ **soichirokawa** 1 year, 4 months ago

A might be enough. Cloud composer will be an overkill

upvoted 2 times

✉ **AWSandeep** 1 year, 4 months ago

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

upvoted 3 times

✉ **ducc** 1 year, 4 months ago

**Selected Answer: D**

D is a more recommend way by Google, IMO.

upvoted 1 times

✉ **squishy\_fishy** 3 months, 1 week ago

I agree, at work use solution A for low volume of Pub/Sub messages and function, and using Composer for high volume Pub/Sub message

upvoted 2 times

✉  **PhuocT** 1 year, 4 months ago

A. more sense to me.

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

Composer support exception and retry for complex pipeline.

D might be correct

upvoted 2 times

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.
- 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.

**Correct Answer: B**

*Community vote distribution*

B (100%)

 **AWSandeep** Highly Voted 1 year, 4 months ago

**Selected Answer: B**

B. Use BigQuery as a data warehouse. Set output destinations for caching large queries.  
upvoted 8 times

 **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: B**

Option B is the best approach - use BigQuery as a data warehouse, and set output destinations for caching large queries.

The key reasons why BigQuery fits the requirements:

It is a fully managed data warehouse built to scale to handle massive datasets and perform fast SQL analytics  
It has a low maintenance architecture with no infrastructure to manage  
SQL capabilities allow easy querying of the medical data  
Output destinations allow configurable caching for fast retrieval of large result sets  
It provides a very cost-effective solution for these large scale analytics use cases  
In contrast, Cloud Spanner and Cloud SQL would not scale as cost effectively for 10TB+ data volumes. Self-managed MySQL on Compute Engine also requires more maintenance. Hence, leveraging BigQuery as a fully managed data warehouse is the optimal solution here.  
upvoted 1 times

 **AzureDP900** 1 year ago

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

 **zellck** 1 year, 2 months ago

**Selected Answer: B**

B is the answer.  
upvoted 3 times

 **TNT87** 1 year, 4 months ago

Answer B.  
<https://cloud.google.com/bigquery/docs/query-overview>  
upvoted 3 times

 **ducc** 1 year, 4 months ago

**Selected Answer: B**

B is correct  
upvoted 2 times

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

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉  **zellck** Highly Voted 1 year, 2 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-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 9 times

✉  **musumusu** Highly Voted 11 months, 2 weeks ago

answer C,

To avoid confusion: Install Storage Transfer Service is always on EXTERNAL OR NON GOOGLE service or data centre to connect google service.

upvoted 7 times

✉  **Prudvi3266** Most Recent 9 months, 1 week ago

**Selected Answer: C**

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

upvoted 3 times

✉  **NicolasN** 1 year, 2 months ago

**Selected Answer: C**

The fact that it's about a POSIX source makes necessary the set up of Storage Transfer Service agents.

This detail limits [C] to be the correct answer, since it's the data center hosting the files where the agent must be installed.

--

Some excerpts:

(an older version of documentation was definite)

"The following is a high-level overview of how Transfer service for on-premises data works:

1. Install Docker and run a small piece of software, called an agent, in your private data center. "

Source: <https://web.archive.org/web/20210529161414/https://cloud.google.com/storage-transfer/docs/on-prem-overview>

--

"Storage Transfer Service agents are applications running inside a Docker container, that coordinate with Storage Transfer Service to read data from POSIX file system sources, and/or write data to POSIX file system sinks.

If your transfer does not involve a POSIX file system, you do not need to set up agents."

Source: <https://cloud.google.com/storage-transfer/docs/managing-on-prem-agents>

upvoted 3 times

✉  **Atnafu** 1 year, 2 months ago

C

Storage Transfer Service agents are applications running inside a Docker container, that coordinate with Storage Transfer Service to read data from POSIX file system sources, and/or write data to POSIX file system sinks.

<https://cloud.google.com/storage-transfer/docs/managing-on-prem-agents#:~:text=Storage%20Transfer%20Service%20agents,agents%20on%20your%20servers.>

upvoted 2 times

✉  **namo621** 1 year, 4 months ago

why can't it be D?

upvoted 1 times

✉  **zellck** 1 year, 2 months ago

Installation is done for the source which is in your data centre, and not in GCP.

upvoted 2 times

✉  **Wasss123** 1 year, 4 months ago

**Selected Answer: C**

I vote for C

upvoted 2 times

✉  **TNT87** 1 year, 4 months ago

Ans C

<https://cloud.google.com/storage-transfer/docs/overview>

upvoted 2 times

✉  **MounicaN** 1 year, 4 months ago

can you help with difference between c and d ?

upvoted 3 times

✉  **gudiking** 1 year, 2 months ago

If you install the software for on-premise data center on a Google Cloud VM, then it's not on-premise, it's on GCP, so it can't access your on-premise data.

upvoted 2 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: C**

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

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: C**

C is correct

upvoted 2 times

Question #197

Topic 1

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

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

**Correct Answer: B**

*Community vote distribution*

B (96%)

4%

 **NicolasN** Highly Voted 1 year, 2 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 33 times

 **squishy\_fishy** 3 months, 1 week ago

Will you change your answer if the answer D says dataset instead of table?

upvoted 1 times

 **Mccloudgirl** 1 year, 2 months ago

Your explanation is perfect, thanks

upvoted 2 times

 **MaxNRG** Most Recent 1 month, 1 week ago

**Selected Answer: B**

The best option to meet the ACID compliance and minimal human intervention requirements is to configure a Cloud SQL for PostgreSQL instance with high availability enabled.

Key reasons:

Cloud SQL for PostgreSQL provides full ACID compliance, unlike Bigtable which provides only atomicity and consistency guarantees. Enabling high availability removes the need for manual failover as Cloud SQL will automatically failover to a standby replica if the leader instance goes down. Point-in-time recovery in MySQL requires manual intervention to restore data if needed. BigQuery does not provide transactional guarantees required for an ACID database. Therefore, a Cloud SQL for PostgreSQL instance with high availability meets the ACID and minimal intervention requirements best. The automatic failover will ensure availability and uptime without administrative effort.

upvoted 1 times

 **[Removed]** 5 months, 3 weeks ago

**Selected Answer: D**

I vote for D - BigQuery with multi region configuration.

According to <https://cloud.google.com/bigquery/docs/introduction> , BigQuery supports ACID and automatically replicated for high availability. "BigQuery stores data using a columnar storage format that is optimized for analytical queries. BigQuery presents data in tables, rows, and columns and provides full support for database transaction semantics (ACID). BigQuery storage is automatically replicated across multiple locations to provide high availability."

upvoted 2 times

 **vamgcp** 6 months, 1 week ago

**Selected Answer: B**

Option B

upvoted 1 times

✉  **musumusu** 11 months, 2 weeks ago

Answer B,

ACID -compliant database are Spanner and CloudSQL

Option A could be the answer if they setup a secondary or failure replicas and auto maintenance window that could trigger in non business hours.

Option B, does not explain about extra replica but in postgresql Highavailability option means the same extra replicas instances are available f emergency.

upvoted 2 times

✉  **AzureDP900** 1 year ago

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

upvoted 1 times

✉  **zellck** 1 year, 2 months ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/sql/docs/postgres/high-availability#HA-configuration>

The purpose of an HA configuration is to reduce downtime when a zone or instance becomes unavailable. This might happen during a zonal outage, or when an instance runs out of memory. With HA, your data continues to be available to client applications.

upvoted 3 times

✉  **samirzubair** 1 year, 2 months ago

I voted for B

upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 months ago

**Selected Answer: B**

B it is exact anwer.

upvoted 1 times

✉  **TNT87** 1 year, 4 months ago

**Selected Answer: B**

Ans B

Postgres is highly ACID compliant as compared to Mysql

upvoted 2 times

✉  **Remi2021** 1 year, 4 months ago

**Selected Answer: B**

cloud sql with high availability enabled is enough

upvoted 2 times

✉  **AWSandeep** 1 year, 4 months ago

**Selected Answer: B**

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

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: B**

I voted for B

Question #198

Topic 1

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.

**Correct Answer: A***Community vote distribution*

D (95%)

5%

 **AWSandeep** Highly Voted 1 year, 4 months 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 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 13 times

 **vamgcp** Most Recent 6 months, 1 week 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

 **spicebits** 2 months, 3 weeks ago<https://cloud.google.com/composer/docs/composer-2/configure-shared-vpc#shared-vpc-guidelines>

upvoted 1 times

 **Ender\_H** 7 months, 4 weeks ago**Selected Answer: A** Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the service project.

- Cloud Composer is a managed Apache Airflow service. It is an open-source tool that programmatically author, schedule, and monitor pipelines which fits your needs perfectly.
- In a Shared VPC configuration, Cloud Composer resources should be placed in the service project. This provides network isolation while still allowing the Cloud Composer environment to communicate with resources in the host project.
- With Shared VPC, the host project's network (including its subnets and secondary IP ranges) is shared by other service projects, which promotes network peering, and it's compliant with the networking considerations of GKE.

upvoted 1 times

 **ckanaar** 4 months, 1 week ago

That's answer D though.

upvoted 2 times

👤 **zelick** 1 year, 2 months ago

Selected Answer: D

D is the answer.

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

Shared VPC enables organizations to establish budgeting and access control boundaries at the project level while allowing for secure and efficient communication using private IPs across those boundaries. In the Shared VPC configuration, Cloud Composer can invoke services hosted in other Google Cloud projects in the same organization without exposing services to the public internet.

upvoted 3 times

👤 **unnamed12355** 10 months, 1 week ago

I thought it was C, but after reading docs I realized it is D.

In D case we can have isolation and reduce project complicity (It could be overlapped resources in Host project, and it is harder to restrict composer access for Host project resources)

upvoted 1 times

👤 **musumusu** 11 months, 1 week ago

why not option C? if composer in host project it will be easier to connect one or more service project with it.

upvoted 2 times

👤 **AzureDP900** 1 year ago

Agreed

upvoted 2 times

👤 **Oleksandr0501** 9 months ago

agreed to what..

upvoted 1 times

👤 **Oleksandr0501** 9 months ago

D it is, as per doc link, provided by users. thx

upvoted 1 times

👤 **Atnafu** 1 year, 2 months ago

D

Shared VPC requires that you designate a host project to which networks and subnetworks belong and a service project, which is attached to host project.

<https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc#:~:text=This%20page%20describes,the%20service%20project>

upvoted 2 times

👤 **ducc** 1 year, 4 months ago

Selected Answer: D

D according to documentation

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

[https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc#set\\_up\\_shared\\_vpc\\_and\\_attach\\_the\\_service\\_project](https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc#set_up_shared_vpc_and_attach_the_service_project)

Question #199

Topic 1

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.

- B. Use BigQuery BI Engine with logical views.
- C. Use BigQuery BI Engine with streaming data.
- D. Use BigQuery BI Engine with authorized views.

**Correct Answer: C**

*Community vote distribution*

A (96%)

4%

 **AWSandeep** Highly Voted 1 year, 4 months ago

**Selected Answer: A**

A. Use BigQuery BI Engine with materialized views.  
upvoted 10 times

 **zellck** Highly Voted 1 year, 2 months ago

**Selected Answer: A**

A is the answer.

<https://cloud.google.com/bigquery/docs/materialized-views-intro>

In BigQuery, materialized views are precomputed views that periodically cache the results of a query for increased performance and efficiency. BigQuery leverages precomputed results from materialized views and whenever possible reads only delta changes from the base tables to compute up-to-date results. Materialized views can be queried directly or can be used by the BigQuery optimizer to process queries to the base tables.

Queries that use materialized views are generally faster and consume fewer resources than queries that retrieve the same data only from the base tables. Materialized views can significantly improve the performance of workloads that have the characteristic of common and repeated queries.  
upvoted 6 times

 **vamgcp** Most Recent 6 months, 1 week 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 run significantly faster compared to computing them on the fly.

upvoted 2 times

 **sporch08** 5 months ago

If we take minimal latency into consideration, I am not sure a materialized view will be the right answer since the user gets data from the cache but is not up to date.

upvoted 1 times

 **phidelics** 7 months, 3 weeks ago

**Selected Answer: A**

periodically cache the results for performance

upvoted 1 times

 **LPIT** 1 year, 3 months ago

**Selected Answer: A**

A.

<https://cloud.google.com/bigquery/docs/materialized-views-intro>

In BigQuery, materialized views are precomputed views that periodically cache the results of a query for increased performance and efficiency.  
upvoted 3 times

 **Julionga** 1 year, 4 months ago

**Selected Answer: A**

I vote A

<https://cloud.google.com/bigquery/docs/bi-engine-intro#:~:text=Materialized%20views%20%2D%20Materialized%20views%20in%20BigQuery%20perform%20precomputation%2C%20there%20reducing%20query%20time.%20You%20should%20create%20materialized%20views%20to%20improve%20performance%20and%20reduce%20processed%20data%20by%20using%20aggregations%2C%20filters%2C%20inner%20joins%2C%20and%20unnests.>

upvoted 2 times

✉  **MounicaN** 1 year, 4 months ago

**Selected Answer: A**

use materialized views is better option here

upvoted 3 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: C**

By integrating BI Engine with BigQuery streaming, you can perform real-time data analysis over streaming data without sacrificing write speed or data freshness.

<https://cloud.google.com/bigquery/docs/bi-engine-intro>

upvoted 1 times

✉  **ducc** 1 year, 4 months ago

Sorry, A is correct

As AWSandeep mention

upvoted 2 times

Question #200

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

D (63%)

A (37%)

✉  **NicolasN**  1 year, 1 month 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"

upvoted 16 times

✉️  **MaxNRG** 1 month, 1 week ago

A single shared service account or granting every employee direct access violates security best practices, so not [A].  
upvoted 1 times

✉️  **KC\_go\_reply** 7 months, 1 week ago

Rejecting C + D solely based on Cloud Storage, which CAN be used in this scenario, is not sound reasoning.  
upvoted 3 times

✉️  **cetanx** Highly Voted  1 year ago

**Selected Answer: D**

for A: please refer to this link below which suggests "Sharing a single service account across multiple applications can complicate the management of the service account" - meaning it's not a best practice.  
<https://cloud.google.com/iam/docs/best-practices-service-accounts#single-purpose>  
Also, what if we have hundreds of users, does it really make sense to manage each user's IAM individually?

for D: it's indeed not one of the best practices but I believe it's much more manageable and better than A

upvoted 10 times

✉️  **MaxNRG** Most Recent  1 month, 1 week ago

**Selected Answer: D**

To align with Google's recommended practices for managing access to personally identifiable information (PII) in compliance with banking industry regulations, let's analyze the options:

A. Assign the required IAM roles to every employee, and create a single service account to access project resources: While assigning specific IAM roles to employees is a good practice for access control, using a single service account for all access to PII is not ideal. Service accounts should be used for applications and automated processes, not as a shared account for multiple users or employees.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

B. Use one service account to access a Cloud SQL database, and use separate service accounts for each human user: Again, service accounts are intended for automated tasks or applications, not for individual human users. Assigning separate service accounts to each human user is not a recommended practice and does not align with the principle of least privilege.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

C. Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users: Using Cloud Storage can indeed help comply with data protection standards, especially when configured correctly with encryption and access controls. However, sharing a single service account among all users is not a best practice. It goes against the principle of least privilege and does not provide adequate granularity for access control.

upvoted 1 times

✉️  **MaxNRG** 1 month, 1 week ago

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: This approach is more aligned with best practices. Using Cloud Storage can ensure compliance with data protection standards. Creating multiple service accounts, each with specific access controls attached to different IAM groups, allows for more granular and controlled access to PII. This setup adheres to the principle of least privilege, ensuring that each service (or group of services) only has access to the resources necessary for its function.

Based on these considerations, option D is the most appropriate choice. It ensures compliance with data protection standards, uses Cloud Storage for secure data management, and employs multiple service accounts tied to IAM groups for granular access control, aligning well with Google-recommended practices and regulatory requirements in the banking industry.

upvoted 1 times

✉️  **[Removed]** 5 months, 3 weeks ago

**Selected Answer: D**

D. Not the best, but seems most reasonable out of 4.

upvoted 1 times

✉️  **vamgcp** 6 months, 1 week ago

**Selected Answer: D**

Option D - Using multiple service accounts attached to IAM groups helps enforce the principle of least privilege. Each group can be assigned only the necessary permissions, reducing the risk of unauthorized access to sensitive data.

upvoted 2 times

✉  **MoeHaydar** 6 months, 3 weeks ago

**Selected Answer: D**

Google Cloud Storage is designed to comply with major data protection standards. Creating multiple service accounts and attaching them to IAM groups provides granular control over who has access to the data. This approach is aligned with the principle of least privilege, a security best practice where a user is given the minimum levels of access necessary to complete their tasks.

upvoted 2 times

✉  **KC\_go\_reply** 7 months, 1 week ago

**Selected Answer: D**

It's not A because

1. assigning IAM roles to single users instead of groups is not Google best practice, and
2. the question explicitly states that we want to use multiple service accounts.

upvoted 2 times

✉  **Ender\_H** 7 months, 4 weeks ago

**Selected Answer: D**

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.

- Google Cloud Storage is built for secure and compliant data storage. It supports compliance with major data protection standards, which is essential in the banking industry where data protection regulations are stringent.
- Service accounts in Google Cloud represent non-human users (applications or services) that need to authenticate and be authorized to access specific Google Cloud resources.
- Creating multiple service accounts attached to IAM groups allows you to manage access control in a granular manner. This follows the principle of least privilege, providing each group with only the permissions they need to perform their tasks, which is a recommended practice for managing access to sensitive data like PII.

upvoted 2 times

✉  **Ender\_H** 7 months, 4 weeks ago

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

- Sharing one service account among all users is not a secure practice. It goes against the principle of least privilege and does not allow for granular control over access permissions. If the shared service account were to be compromised, all resources accessible by the account would be at risk.

upvoted 1 times

✉  **juliosb** 10 months, 1 week ago

Why are so many questions like this?

None of the answers is best practice.

upvoted 5 times

✉  **alfguemat** 2 months ago

I would like to ask your question to those who decide the questions on the exams. I don't understand what they're trying to do, many of the questions cause divided responses, because they don't have a clear answer. The certification process is a waste of time.

upvoted 1 times

✉  **SuperVee** 10 months, 3 weeks ago

**Selected Answer: D**

I could be wrong but I think the wording in D caused this confusion, so it is an English problem. -- "Use multiple service accounts attached to IAM groups to grant the appropriate access to each group"

I believe what D really means is that you can create a group for a bunch of people who only need access to resource A, so attach a Service account to the group and service account only have access to A.

Then you create another group for another bunch of people who only need access to resource B, so attach a service account to this group. The service account can only access to B.

So each group/service account has a very specific access target, and purpose of the group is very narrowly defined which is allowed by best practice. However, wording in option D merged all these into one sentence causing confusions.

Option A is an administrative nightmare to manage IAM for a larger user population which is actually also against GCP best practices.

upvoted 5 times

✉  **Aamir185** 11 months, 2 weeks ago

**Selected Answer: D**

D it is

upvoted 2 times

✉  **AzureDP900** 1 year ago

D is right

upvoted 1 times

✉  **Amar2022** 1 year, 1 month ago

**Selected Answer: A**

A is the correct one

upvoted 1 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: A**

Agree with NicolasN, D is bad practice. For D this may result in permission creep, where a group is granted access to an increasing number of resources. Only grant service accounts specific access to resources.

upvoted 1 times

✉  **odacir** 1 year, 1 month ago

**Selected Answer: A**

A is the answer, as NicolasN says.

<https://cloud.google.com/iam/docs/service-accounts#groups>

upvoted 1 times

✉  **Andrix2405** 1 year, 1 month ago

**Selected Answer: A**

Avoid using groups for granting service accounts access to resources -> D

upvoted 2 times

✉  **Andrix2405** 1 year, 1 month ago

Sorry A

upvoted 1 times

✉  **zellck** 1 year, 2 months ago

**Selected Answer: D**

D is the answer.

upvoted 1 times

Question #201

Topic 1

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.
- 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.

**Correct Answer: B**

*Community vote distribution*

A (100%)

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.

**Correct Answer: A***Community vote distribution*

C (93%)

7%

✉  **muhusman** Highly Voted 9 months, 1 week ago

ed

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 5 times

✉  **datapassionate** 2 weeks, 1 day 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 7 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](https://cloud.google.com/bigquery/docs/dts-introduction#supported_data_sources)

upvoted 5 times

✉  **Takshashila** Most Recent 7 months, 2 weeks ago

**Selected Answer: C**

the correct option is C

upvoted 1 times

✉  **wjtb** 10 months, 3 weeks 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 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 1 times

✉  **musumusu** 11 months, 1 week ago

Answer C,  
What is wrong with B ? Key words = Daily transfer .. so no to transfer appliance,  
upvoted 2 times

✉  **zelick** 1 year, 2 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-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** 1 year ago

yes, It is C  
upvoted 1 times

✉  **Atnafu** 1 year, 2 months ago

C  
D-no answer because bq transfer service don't support from on-prem  
upvoted 1 times

✉  **Atnafu** 1 year, 2 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 query and explore the platform data.

query and explore is the key  
upvoted 1 times

✉  **John\_Pongthorn** 1 year, 3 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** 1 year, 3 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

✉  **John\_Pongthorn** 1 year, 3 months ago

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

✉  **Wasss123** 1 year, 4 months ago

**Selected Answer: C**

I will go with C

upvoted 3 times

✉  **MounicaN** 1 year, 4 months ago

I will go with C

<https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options>  
upvoted 1 times

✉  **John\_Pongthorn** 1 year, 4 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** 1 year, 3 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** 1 year, 4 months ago

**Selected Answer: C**

Ans C

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

upvoted 1 times

✉  **HarshKothari21** 1 year, 4 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 the 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

✉  **nwk** 1 year, 4 months ago

Data is generated daily. Unlikely to ship Transfer Appliance every day.

Vote for C instead. "Transfer Service for on-premises data" is a free Google Cloud service that's intended to streamline the process of uploading data into Google Cloud Storage buckets"

<https://cloud.google.com/blog/products/storage-data-transfer/introducing-storage-transfer-service-for-on-premises-data>

upvoted 2 times

✉  **ducc** 1 year, 4 months ago

**Selected Answer: B**

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.

upvoted 1 times

Question #203

Topic 1

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

**Correct Answer: C***Community vote distribution*

C (100%)

  **MaxNRG** 1 month, 1 week ago**Selected Answer: C**

The best way to reduce the TensorFlow training time in a cost-effective manner is to use a VM with a GPU hardware accelerator. TensorFlow can take advantage of GPUs to significantly speed up training time for many models.

Specifically, option C is the best choice.

Changing the VM to another standard type like n2-highmem-32 or e2-standard-32 (options A and B) may provide some improvement, but likely not a significant speedup.

Using a TPU (option D) could speed up training, but TPUs are more costly than GPUs. For a cost-effective solution, GPU acceleration provides the best performance per dollar.

Since the model must run partially on CPUs, a VM instance with GPUs added will allow TensorFlow to offload appropriate operations to the GPUs while keeping CPU-specific operations on the CPU. This can provide a significant reduction in training time for many common TensorFlow models while keeping costs reasonable

upvoted 4 times

  **spicebits** 2 months, 3 weeks ago**Selected Answer: C**

[https://cloud.google.com/tpu/docs/intro-to-tpu#when\\_to\\_use\\_tpus](https://cloud.google.com/tpu/docs/intro-to-tpu#when_to_use_tpus)

upvoted 2 times

  **AzureDP900** 1 year ago

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

upvoted 1 times

  **jkhong** 1 year, 1 month 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 3 times

  **Atnafu** 1 year, 2 months 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](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

  **Atnafu** 1 year, 1 month ago

The model has custom TensorFlow operations that must run partially on a CPU. is the key for GPU

upvoted 3 times

  **zellck** 1 year, 2 months ago**Selected Answer: C**

C is the answer.

[https://cloud.google.com/tpu/docs/tpus#when\\_to\\_use\\_tpus](https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus)

GPUs

- Models with a significant number of custom TensorFlow operations that must run at least partially on CPUs

upvoted 4 times

✉  **gudiking** 1 year, 2 months ago

**Selected Answer: C**

I agree with C, for choosing a GPU one of the cases says:

"Models with a significant number of custom TensorFlow operations that must run at least partially on CPUs"

[https://cloud.google.com/tpu/docs/tpus#when\\_to\\_use\\_tpus](https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus)

upvoted 1 times

✉  **gudiking** 1 year, 1 month ago

Question #204

Topic 1

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.

**Correct Answer: A**

*Community vote distribution*

D (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D

upvoted 1 times

✉  **vamgcp** 6 months, 1 week 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 2 times

✉  **AzureDP900** 1 year 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** 1 year, 1 month ago

**Selected Answer: D**

D is the best option to sanitize the data to its D

upvoted 2 times

✉  **jkhong** 1 year, 1 month ago

**Selected Answer: D**

Better to use pubsub for streaming and reading message data

Dataflow ParDo can perform filtering of data

upvoted 1 times

✉  **zellck** 1 year, 1 month ago

**Selected Answer: D**

D is the answer.

upvoted 1 times

✉  **vidts** 1 year, 2 months ago

**Selected Answer: D**

It's D

upvoted 2 times

✉  **Atnafu** 1 year, 2 months ago

Answer is D

upvoted 2 times

Question #205

Topic 1

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.
- 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.

**Correct Answer: C**

*Community vote distribution*

B (100%)

✉  **raaad** 3 weeks, 6 days ago

**Selected Answer: B**

- Dataflow is a fully managed service for stream and batch data processing and is well-suited for real-time data processing tasks like identifying longtail and outlier data points.
  - Using BigQuery as a sink allows to efficiently store the cleansed and processed data for further analysis and serving it to AI models.
- upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

**Selected Answer: B**

B is the best option to efficiently provision and manage the deployment process for this data processing application on GKE:  
upvoted 1 times

✉  **MaxNRG** 1 month, 1 week ago

- Cloud Build allows you to automate the building, testing, and deployment of your application using Docker containers.
- Using Terraform with Cloud Build provides Infrastructure as Code capabilities to provision the GKE cluster with GPUs, SSDs, and network bandwidth.
- Terraform can be configured to pull the latest container images from the registry when deploying.
- Cloud Build triggers provide event-based automation to rebuild and redeploy when container images are updated.
- This provides an automated CI/CD pipeline to launch the application on GKE using the desired infrastructure and latest images.
- Dataflow and Cloud Scheduler don't directly provide infrastructure provisioning or deployment orchestration for GKE.
- gcloud commands can be used but don't provide the same automation benefits as Cloud Build + Terraform.

upvoted 2 times

✉  **MaxNRG** 1 month, 1 week ago

So using Cloud Build with Terraform templates provides the most efficient way to provision and deploy this data processing application GKE.  
upvoted 2 times

✉  **spicebits** 2 months, 3 weeks ago

**Selected Answer: B**

I don't really like B or C... but given the choices I would go with B.

B-Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images. { Terraform command is Terraform Apply and not Terraform build, but also why not use gcloud container command instead of introducing 3rd p builder image?}... I don't like this choice but it is the best one.

C. Use GKE to autoscale containers, and use gcloud commands to provision the infrastructure. {This doesn't handle the building of the infra, the deployment of the latest images, this one is clearly wrong, not sure why it is marked as the right choice}

upvoted 1 times

✉  **vamgcp** 6 months, 1 week ago

**Selected Answer: B**

B is correct

upvoted 2 times

✉  **whorillo** 9 months, 2 weeks ago

**Selected Answer: B**

B is correct

upvoted 1 times

✉  **charline** 11 months, 3 weeks ago

**Selected Answer: B**

b is ok

upvoted 1 times

✉  **AzureDP900** 1 year ago

B. Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images.  
upvoted 1 times

✉  **zellick** 1 year, 1 month ago

**Selected Answer: B**

B is the answer.

upvoted 2 times

✉  **hauhau** 1 year, 1 month ago

**Selected Answer: B**

Maybe B

ref: <https://cloud.google.com/architecture/managing-infrastructure-as-code>

upvoted 2 times

 **Atnafu** 1 year, 2 months ago

C is correct answer

upvoted 1 times

 **Atnafu** 1 year, 2 months ago

Sorry I meant B

upvoted 3 times

Question #206

*Topic 1*

You need ads data to serve AI models and historical data for analytics. Longtail and outlier data points need to be identified. You want to cleanse the data in near-real time before running it through AI models. What should you do?

- A. Use Cloud Storage as a data warehouse, shell scripts for processing, and BigQuery to create views for desired datasets.
- B. Use Dataflow to identify longtail and outlier data points programmatically, with BigQuery as a sink.
- C. Use BigQuery to ingest, prepare, and then analyze the data, and then run queries to create views.
- D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery.

**Correct Answer: A**

*Community vote distribution*

B (100%)

✉  **datapassionate** 2 weeks, 1 day ago

**Selected Answer: B**

B. Use Dataflow to identify longtail and outlier data points programmatically, with BigQuery as a sink.  
upvoted 1 times

✉  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: B**

B: Dataflow, solves exactly the use case described  
upvoted 1 times

✉  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: B**

B is the best option for cleansing the ads data in near real-time before running it through AI models.  
The key reasons are:

- Dataflow allows for stream processing of data in near real-time. This allows you to identify and cleanse longtail and outlier data points as the data is streamed in.
- Dataflow has built-in capabilities for detecting and handling outliers and anomalies in streaming data. This makes it well-suited for programmatically identifying longtail and outlier data points.
- Using BigQuery as the output sink allows the cleansed data to be immediately available for analysis and serving to AI models. BigQuery can act as a serving layer for the models.
- Options A, C, and D either don't provide real-time processing (A and C) or don't easily integrate with BigQuery for analysis and serving (D).

upvoted 2 times

✉  **MaxNRG** 3 weeks, 1 day ago

So B is the best architecture here to meet the needs of near real-time cleansing, identification of longtail/outlier data points, and integration with BigQuery for serving AI models.  
upvoted 1 times

✉  **raaad** 3 weeks, 6 days ago

**Selected Answer: B**

- Dataflow is a fully managed service for stream and batch data processing and is well-suited for real-time data processing tasks like identifying longtail and outlier data points.

Question #207

Topic 1

You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data, filtered by location\_id and device\_version with the following query:

```
SELECT
 MAX(temperature)
FROM
 acme_iot_data.sensors
WHERE
 create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
 AND location_id = "SW1W9TQ"
 AND device_version = "202007r3"
```

You want to optimize your queries for cost and performance. How should you structure your data?

- A. Partition table data by create\_date, location\_id, and device\_version.
- B. Partition table data by create\_date, cluster table data by location\_id, and device\_version.
- C. Cluster table data by create\_date, location\_id, and device\_version.
- D. Cluster table data by create\_date, partition by location\_id, and device\_version.

**Correct Answer: C***Community vote distribution*

B (100%)

  **datapassionate** 2 weeks, 1 day ago**Selected Answer: B**

B. Partition table data by create\_date, cluster table data by location\_id, and device\_version.

upvoted 1 times

  **Matt\_108** 2 weeks, 3 days ago**Selected Answer: B**

B: Partitioning makes date-related querying efficient, clustering will keep relevant data close together and optimize the performance of filters on the cluster columns

upvoted 1 times

  **MaxNRG** 3 weeks, 1 day ago**Selected Answer: B**

1. Partitioning the data by create\_date will allow BigQuery to prune partitions that are not relevant to the query by date.

2. Clustering the data by location\_id and device\_version within each partition will keep related data close together and optimize the performance of filters on those columns.

This provides both the pruning benefits of partitioning and locality benefits of clustering for filters on multiple columns.

The query provided indicates that the access pattern is primarily based on the most recent data (within the last 7 days), filtered by location\_id and device\_version. Given this pattern, you would want to optimize your table structure in such a way that queries scanning through the data will process the least amount of data possible to reduce costs and improve performance.

upvoted 2 times

  **Smakye179** 3 weeks, 2 days ago**Selected Answer: B**

Only correct answer is B, you can only partition by one field, and you can only cluster on partitioned tables

upvoted 1 times

  **raaad** 3 weeks, 6 days ago**Selected Answer: B**

Answer is B:

- Partitioning the table by create\_date allows us to efficiently query data based on time, which is common in access patterns that prioritize recent data.

- Clustering the table by location\_id and device\_version further organizes the data within each partition, making queries filtered by these columns more efficient and cost-effective.

upvoted 2 times

  **e70ea9e** 1 month ago**Selected Answer: B**

The best answer is B. Partition table data by create\_date, cluster table data by location\_id, and device\_version.

Here's a breakdown of why this structure is optimal:

Partitioning by create\_date:

Aligns with query pattern: Filters for recent data based on create\_date, so partitioning by this column allows BigQuery to quickly narrow down data to scan, reducing query costs and improving performance.

Manages data growth: Partitioning effectively segments data by date, making it easier to manage large datasets and optimize storage costs.

Clustering by location\_id and device\_version:

Enhances filtering: Frequently filtering by location\_id and device\_version, clustering physically co-locates related data within partitions, further reducing scan time and improving performance.

upvoted 2 times

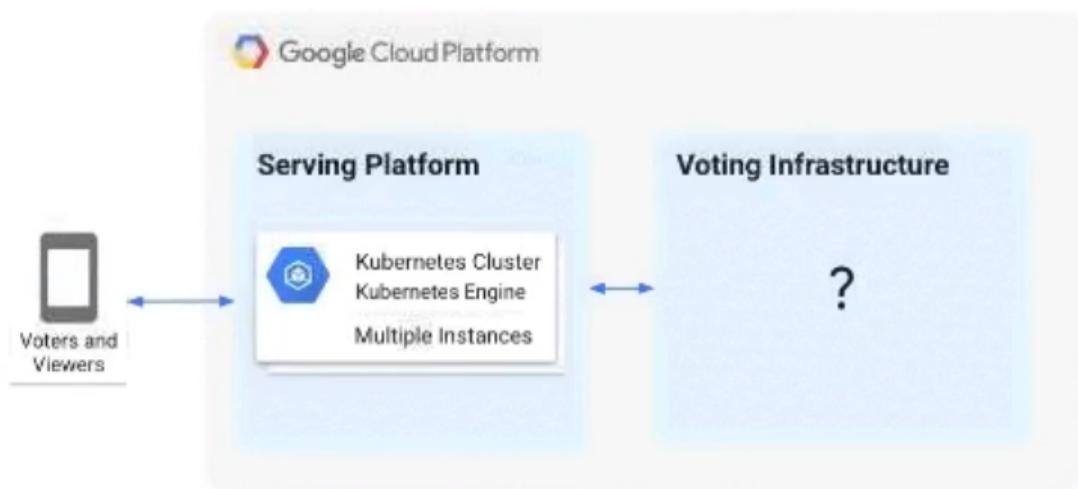
Question #208

Topic 1

A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3-minute period. You

are in charge of the "voting infrastructure" and must ensure that the platform can handle the load and that all votes are processed. You must

are in charge of the voting infrastructure and must ensure that the platform can handle the load and that all votes are processed. You must display partial results while voting is open. After voting closes, you need to count the votes exactly once while optimizing cost. What should you do?



- A. Create a Memorystore instance with a high availability (HA) configuration.
- B. Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas.
- C. Write votes to a Pub/Sub topic and have Cloud Functions subscribe to it and write votes to BigQuery.
- D. Write votes to a Pub/Sub topic and load into both Bigtable and BigQuery via a Dataflow pipeline. Query Bigtable for real-time results and BigQuery for later analysis. Shut down the Bigtable instance when voting concludes.

**Correct Answer: D**

*Community vote distribution*

D (90%)

10%

✉️  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: D**

D, i do agree with everything MaxNRG said.  
upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: D**

Since cost optimization and minimal latency are key requirements, option D is likely the best choice to meet all the needs:

The key reasons option D works well:

Using Pub/Sub to ingest votes provides scalable, reliable transport.

Loading into Bigtable and BigQuery provides both:

Low latency reads from Bigtable for real-time results.

Cost effective storage in BigQuery for longer term analysis.

Shutting down Bigtable after voting concludes reduces costs.

BigQuery remains available for cost-optimized storage and analysis.

So you are correct that option D combines the best of real-time performance for queries using Bigtable, with cost-optimized storage in BigQu

The only additional consideration may be if 3 minutes of Bigtable usage still incurs higher charges than ingesting directly into BigQuery. But fc  
minimizing latency while optimizing cost, option D is likely the right architectural choice given the requirements.

upvoted 4 times

✉️  **Smakyel79** 3 weeks, 2 days ago

**Selected Answer: C**

Pub/Sub for sure, and Cloud Functions + BigQuery Streaming seems a good solution. Won't use BigTable as need at least 100GB of data (do  
thing a voting system could arrive to that amount of data) and needs to "heat" to work right for >10 minutes... and would be \$\$\$ over C soluti  
upvoted 1 times

✉️  **raaad** 3 weeks, 6 days ago

**Selected Answer: D**

Answer is D:

- Google Cloud Pub/Sub can manage the high-volume data ingestion.
- Google Cloud Dataflow can efficiently process and route data to both Bigtable and BigQuery.

Question #209

Topic 1

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 copy all the data to a new clustered table. What should you do?

- A. Re-create the table using data partitioning on the package delivery date.
- B. Implement clustering in BigQuery on the package-tracking ID column.
- C. Implement clustering in BigQuery on the ingest date column.
- D. Tier older data onto Cloud Storage files and create a BigQuery table using Cloud Storage as an external data source.

**Correct Answer: B**

*Community vote distribution*

B (100%)

✉️  **datapassionate** 2 weeks, 1 day ago

**Selected Answer: B**

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

upvoted 1 times

✉️  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: B**

Definitely B

upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: B**

This looks like Question #166

Option B, implementing clustering in BigQuery on the package-tracking ID column, seems the most appropriate. It directly addresses the query slowdown issue by reorganizing the data in a way that aligns with the analysts' query patterns, leading to more efficient and faster query execution.

upvoted 3 times

✉️  **raaad** 3 weeks, 6 days ago

**Selected Answer: B**

Answer is B

upvoted 2 times

✉️  **e70ea9e** 1 month ago

**Selected Answer: B**

Query Focus: Analysts are interested in geospatial trends within individual package lifecycles. Clustering by package-tracking ID physically collocates related data, significantly improving query performance for these analyses.

Addressing Slow Queries: Clustering addresses the query slowdown issue by optimizing data organization for the specific query patterns.

Partitioning vs. Clustering:

Partitioning: Divides data into segments based on a column's values, primarily for managing large datasets and optimizing query costs.

Clustering: Organizes data within partitions for faster querying based on specific columns.

upvoted 3 times

Question #210

Topic 1

You are designing a data mesh on Google Cloud with multiple distinct data engineering teams building data products. The typical data curation design pattern consists of landing files in Cloud Storage, transforming raw data in Cloud Storage and BigQuery datasets, and storing the final curated data product in BigQuery datasets. You need to configure Dataplex to ensure that each team can access only the assets needed to build their data products. You also need to ensure that teams can easily share the curated data product. What should you do?

- A. 1. Create a single Dataplex virtual lake and create a single zone to contain landing, raw, and curated data.  
2. Provide each data engineering team access to the virtual lake.
- B. 1. Create a single Dataplex virtual lake and create a single zone to contain landing, raw, and curated data.  
2. Build separate assets for each data product within the zone.  
3. Assign permissions to the data engineering teams at the zone level.
- C. 1. Create a Dataplex virtual lake for each data product, and create a single zone to contain landing, raw, and curated data.  
2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.
- D. 1. Create a Dataplex virtual lake for each data product, and create multiple zones for landing, raw, and curated data.  
2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.

Correct Answer: A

### Community vote distribution

D (78%)

C (22%)

✉️ **datapassionate** 2 weeks, 1 day ago

**Selected Answer: D**

1. Create a Dataplex virtual lake for each data product, and create multiple zones for landing, raw, and curated data.
2. Provide the data engineering teams with full access to the virtual lake assigned to their data product.

**Lake:** A logical construct representing a data domain or business unit. For example, to organize data based on group usage, you can set up a lake for each department (for example, Retail, Sales, Finance).

**Zone:** A subdomain within a lake, which is useful to categorize data by the following:

**Stage:** For example, landing, raw, curated data analytics, and curated data science.

upvoted 1 times

✉️ **datapassionate** 2 weeks, 1 day ago

<https://cloud.google.com/dataplex/docs/introduction>

upvoted 1 times

✉️ **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: D**

D: 1 virtual lake per Data Product (which stands for domain basically), zones to split data by "status". Each Data Eng team can access their own data exclusively and in a data mesh compliant way

upvoted 1 times

✉️ **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: D**

The best approach is to create a Dataplex virtual lake for each data product, with multiple zones for landing, raw, and curated data. Then provide the data engineering teams with access only to the zones they need within the virtual lake assigned to their product.

To enable teams to easily share curated data products, you should use cross-lake sharing in Dataplex. This allows curated zones to be shared across virtual lakes while maintaining data isolation for other zones.

upvoted 2 times

✉️ **MaxNRG** 3 weeks, 1 day ago

So the steps would be:

1. Create a Dataplex virtual lake for each data product.
2. Within each lake, create separate zones for landing, raw, and curated data.
3. Provide each data engineering team with access only to the zones they need within their assigned virtual lake.
4. Configure cross-lake sharing on the curated data zones to share curated data products between teams.

This provides isolation and access control between teams for raw data while enabling easy sharing of curated data products.  
[https://cloud.google.com/dataplex/docs/introduction#a\\_domain-centric\\_data\\_mesh](https://cloud.google.com/dataplex/docs/introduction#a_domain-centric_data_mesh)

upvoted 2 times

✉️ **Smakyei79** 3 weeks, 2 days ago

I believe the answer is B, but there is a misspelling in the answer, should say "create multiple zones"

upvoted 2 times

✉️ **Helinia** 3 weeks, 2 days ago

**Selected Answer: D**

Each lake should be created per data product since data product sounds like a domain in this question.

Since we have landing, raw, curated data, we should create different zones.

"Zones are of two types: raw and curated.

Raw zone: Contains data that is in its raw format and not subject to strict type-checking.

Curated zone: Contains data that is cleaned, formatted, and ready for analytics. The data is columnar, Hive-partitioned, and stored in Parquet, Avro, Orc files, or BigQuery tables. Data undergoes type-checking- for example, to prohibit the use of CSV files because they don't perform as well for SQL access."

Ref: <https://cloud.google.com/dataplex/docs/introduction#terminology>

upvoted 1 times

✉  **Jordan18** 3 weeks, 3 days ago

why not B?

upvoted 3 times

✉  **Sofia98** 3 weeks, 4 days ago

Why not B?

upvoted 2 times

✉  **tibuenoc** 1 week, 5 days ago

Because it's the best practice is separated zones. One zone for landing, raw and curated.

The answer B - has this part that excluded it "create a single zone to contain landing"

The correct awser is D

upvoted 1 times

✉  **Ed\_Kim** 3 weeks, 6 days ago

**Selected Answer: D**

The answer is D

upvoted 2 times

✉  **e70ea9e** 1 month ago

**Selected Answer: C**

Virtual Lake per Data Product: Each virtual lake acts as a self-contained domain for a specific data product, aligning with the data mesh principle of decentralized ownership and responsibility.

Team Autonomy: Teams have full control over their virtual lake, enabling independent development, management, and sharing of their data products.

upvoted 2 times

Question #211

Topic 1

You are using BigQuery with a multi-region dataset that includes a table with the daily sales volumes. This table is updated multiple times per day. You need to protect your sales table in case of regional failures with a recovery point objective (RPO) of less than 24 hours, while keeping costs to a minimum. What should you do?

- A. Schedule a daily export of the table to a Cloud Storage dual or multi-region bucket.
- B. Schedule a daily copy of the dataset to a backup region.
- C. Schedule a daily BigQuery snapshot of the table.
- D. Modify ETL job to load the data into both the current and another backup region.

**Correct Answer: C**

*Community vote distribution*

A (61%)

C (33%)

6%

✉  **raaad**  3 weeks, 6 days ago

**Selected Answer: C**

Option C provides cost-effective way.

- BigQuery table snapshots are a feature that allows you to capture the state of a table at a particular point in time.
- Snapshots are incremental, so they only store the data that has changed, making them more cost-effective than full table copies.
- In the event of a regional failure, you can quickly restore the table from a snapshot.

upvoted 6 times

✉️  **GCP001** Most Recent 1 week, 4 days ago

**Selected Answer: A**

Option A. Check the ref for regional loss -  
[https://cloud.google.com/bigquery/docs/reliability-intro#scenario\\_loss\\_of\\_region](https://cloud.google.com/bigquery/docs/reliability-intro#scenario_loss_of_region)

upvoted 1 times

✉️  **datapassionate** 2 weeks ago

**Selected Answer: A**

A. Schedule a daily export of the table to a Cloud Storage dual or multi-region bucket.

upvoted 1 times

✉️  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: A**

A: MaxNRG and Helinia cleared the reasons very well

upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: A**

Why not C:

A table snapshot must be in the same region, and under the same organization, as its base table.

<https://cloud.google.com/bigquery/docs/table-snapshots-intro#limitations>

upvoted 4 times

✉️  **MaxNRG** 3 weeks, 1 day ago

Based on the information provided and the need to avoid data loss in the case of a hard regional failure in BigQuery, which could result in the destruction of all data in that region, the focus should be on creating backups in a geographically distinct region. Considering this scenario the most suitable option would be Option A

Here's why this option is the most appropriate:

upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

- Cross-Region Backup: Exporting the data to a Google Cloud Storage bucket that is either dual or multi-regional ensures that your backups are stored in a different geographic location. This is critical for protecting against hard regional failures.
- Data Durability: Cloud Storage provides high durability for stored data, making it a reliable option for backups in the case of regional disasters.
- Cost-Effectiveness: While there are costs associated with storage and data transfer, this method can be more cost-effective compared to maintaining active replicas of the data in multiple regions, especially if the data is large.

upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

- Flexibility and Automation: The export process can be automated and scheduled to occur daily, aligning with your RPO of less than hours. This ensures that the most recent data is always backed up.
- Recovery Process: In the event of a hard regional failure, the data can be restored from the Cloud Storage backup to another operational BigQuery region, ensuring continuity of operations.

upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

The other options, while viable in certain scenarios, do not provide the same level of protection against a hard regional failure:

- Option B (Copy to Backup Region) and Option D (Modify ETL to Load into Backup Region) do not address the possibility of a hard regional failure adequately, as they do not necessarily imply storing data in a geographically distinct region.
- Option C (BigQuery Snapshot) is useful for point-in-time recovery but does not inherently protect against hard regional failures since the snapshots are within the same BigQuery service.

Focusing on a robust disaster recovery strategy is crucial. Option A provides a balance between ensuring data availability in the event of a regional disaster and managing costs, aligning with best practices for data management in the cloud.

upvoted 1 times

✉  **Helinia** 3 weeks, 2 days ago

**Selected Answer: A**

"BigQuery does not offer durability or availability in the extraordinarily unlikely and unprecedented event of physical region loss. This is true for both "regions and multi-region" configurations. Hence maintaining durability and availability under such a scenario requires customer planning."

"To avoid data loss in the face of destructive regional loss, you need to back up data to another geographic location. For example, you could periodically export a snapshot of your data to Google Cloud Storage in another geographically distinct region."

Ref: [https://cloud.google.com/bigquery/docs/reliability-intro#scenario\\_loss\\_of\\_region](https://cloud.google.com/bigquery/docs/reliability-intro#scenario_loss_of_region)

upvoted 4 times

✉  **Helinia** 3 weeks, 1 day ago

Why not C:

"BigQuery also supports the ability to snapshot tables. With this feature you can explicitly backup data within the same region for longer than the 7 day time travel window. A snapshot is purely a metadata operation and results in no additional storage bytes. While this can add protection against accidental deletion, it does not increase the durability of the data."

[https://cloud.google.com/bigquery/docs/reliability-intro#scenario\\_accidental\\_deletion\\_or\\_data\\_corruption](https://cloud.google.com/bigquery/docs/reliability-intro#scenario_accidental_deletion_or_data_corruption)

upvoted 1 times

✉  **MaxNRG** 3 weeks, 1 day ago

Option A (Export to Cloud Storage): While exporting to Cloud Storage is a viable backup strategy, it can be more expensive and less efficient than using snapshots, especially if the table is large and updated frequently.

upvoted 1 times

✉  **MaxNRG** 3 weeks, 1 day ago

I agree, It's A:

A table snapshot must be in the same region, and under the same organization, as its base table.

<https://cloud.google.com/bigquery/docs/table-snapshots-intro#limitations>

upvoted 1 times

✉  **qq589539483084gfrgrgfr** 3 weeks, 4 days ago

Option A

upvoted 3 times

✉  **qq589539483084gfrgrgfr** 3 weeks, 4 days ago

<https://cloud.google.com/bigquery/docs/reliability-intro>

upvoted 3 times

✉  **e70ea9e** 1 month ago

**Selected Answer: D**

Automatically replicates data to a backup region upon each update, ensuring an RPO of less than 24 hours, even with multiple daily updates.

upvoted 1 times

✉  **raaad** 3 weeks, 6 days ago

Option D:

Doubles the write load and storage costs since you are maintaining two live datasets.

upvoted 3 times

You are troubleshooting your Dataflow pipeline that processes data from Cloud Storage to BigQuery. You have discovered that the Dataflow worker nodes cannot communicate with one another. Your networking team relies on Google Cloud network tags to define firewall rules. You need to identify the issue while following Google-recommended networking security practices. What should you do?

- A. Determine whether your Dataflow pipeline has a custom network tag set.
- B. Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 for the Dataflow network tag.
- C. Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 on the subnet used by Dataflow workers.
- D. Determine whether your Dataflow pipeline is deployed with the external IP address option enabled.

**Correct Answer: D**

*Community vote distribution*

B (100%)

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

B, check if there is a firewall rule allowing traffic on TCP ports 12345 and 12346 for the Dataflow network tag.  
upvoted 1 times

✉️  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: B**

The best approach would be to check if there is a firewall rule allowing traffic on TCP ports 12345 and 12346 for the Dataflow network tag. Dataflow uses TCP ports 12345 and 12346 for communication between worker nodes. Using network tags and associated firewall rules is a Google-recommended security practice for controlling access between Compute Engine instances like Dataflow workers.

So the key things to check would be:

1. Ensure your Dataflow pipeline is using the Dataflow network tag on the worker nodes. This tag is applied by default unless overridden.
2. Check if there is a firewall rule allowing TCP 12345 and 12346 ingress and egress traffic for instances with the Dataflow network tag. If not, add the rule.

Options A, C and D relate to other networking aspects but do not directly address the Google recommended practice of using network tags a firewall rules.

upvoted 3 times

Question #213

*Topic 1*

Your company's customer\_order table in BigQuery stores the order history for 10 million customers, with a table size of 10 PB. You need to create a dashboard for the support team to view the order history. The dashboard has two filters, country\_name and username. Both are string data types in the BigQuery table. When a filter is applied, the dashboard fetches the order history from the table and displays the query results. However, the dashboard is slow to show the results when applying the filters to the following query:

```
SELECT date, order, status FROM customer_order
WHERE country = '<country_name>' AND username = '<username>'
```

How should you redesign the BigQuery table to support faster access?

- A. Cluster the table by country and username fields.
- B. Cluster the table by country field, and partition by username field.
- C. Partition the table by country and username fields.
- D. Partition the table by \_PARTITIONTIME.

**Correct Answer: C**

*Community vote distribution*

A (91%)

9%

✉  **datapassionate** 2 weeks ago

**Selected Answer: A**

Correct answer: A. Cluster the table by country and username fields.

Why not B and C - > Integer is required for partitioning

[https://cloud.google.com/bigquery/docs/partitioned-tables#integer\\_range](https://cloud.google.com/bigquery/docs/partitioned-tables#integer_range)

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

A: the fields are both strings, which are not supported for partitioning. Moreover, the fields are regularly used in filters, which is where clustering really improves performance

upvoted 2 times

✉  **Takshashila** 3 weeks ago

**Selected Answer: B**

Clustering can also be done after partition?

upvoted 1 times

✉  **raaad** 3 weeks, 6 days ago

**Selected Answer: A**

- Clustering organizes the data based on the specified columns (in this case, country\_name and username).  
- When a query filters on these columns, BigQuery can efficiently scan only the relevant parts of the table

upvoted 4 times

✉  **e70ea9e** 1 month ago

**Selected Answer: A**

country and username --> cluster

upvoted 3 times

Question #214

Topic 1

You have a Standard Tier Memorystore for Redis instance deployed in a production environment. You need to simulate a Redis instance failover in the most accurate disaster recovery situation, and ensure that the failover has no impact on production data. What should you do?

- A. Create a Standard Tier Memorystore for Redis instance in the development environment. Initiate a manual failover by using the limited-data-loss data protection mode.
- B. Create a Standard Tier Memorystore for Redis instance in a development environment. Initiate a manual failover by using the force-data-loss data protection mode.
- C. Increase one replica to Redis instance in production environment. Initiate a manual failover by using the force-data-loss data protection mode.
- D. Initiate a manual failover by using the limited-data-loss data protection mode to the Memorystore for Redis instance in the production environment.

**Correct Answer: D**

*Community vote distribution*

B (80%)

C (20%)

MaxNRG Highly Voted 3 weeks, 1 day ago

Selected Answer: B

The best option is B - Create a Standard Tier Memorystore for Redis instance in a development environment. Initiate a manual failover by using the force-data-loss data protection mode.

The key points are:

- The failover should be tested in a separate development environment, not production, to avoid impacting real data.
- The force-data-loss mode will simulate a full failover and restart, which is the most accurate test of disaster recovery.
- Limited-data-loss mode only fails over reads which does not fully test write capabilities.
- Increasing replicas in production and failing over (C) risks losing real production data.
- Failing over production (D) also risks impacting real data and traffic.

So option B isolates the test from production and uses the most rigorous failover mode to fully validate disaster recovery capabilities.

upvoted 5 times

tibuenoc Most Recent 1 week, 5 days ago

Selected Answer: B

<https://cloud.google.com/memorystore/docs/redis/about-manual-failover>

upvoted 1 times

datapassionate 2 weeks ago

Selected Answer: B

B. Create a Standard Tier Memorystore for Redis instance in a development environment. Initiate a manual failover by using the force-data-loss data protection mode

upvoted 1 times

Matt\_108 2 weeks, 2 days ago

Selected Answer: B

Best option is B - no impact on production env and forces a full failover

upvoted 1 times

raaad 3 weeks, 6 days ago

Selected Answer: C

Increasing the number of replicas in a Redis instance in a production environment means that we will have additional copies of the same data and that's why failover will not impact the production data

upvoted 1 times

MaxNRG 3 weeks, 1 day ago

"no impact on production data" - not C nor D

upvoted 1 times

e70ea9e 1 month ago

Selected Answer: C

Separate Development Environment:

Isolates testing from production, preventing any impact on live data or services.

Provides a safe and controlled environment for simulating failover scenarios.

upvoted 1 times

You are administering a BigQuery dataset that uses a customer-managed encryption key (CMEK). You need to share the dataset with a partner organization that does not have access to your CMEK. What should you do?

- A. Provide the partner organization a copy of your CMEKs to decrypt the data.
- B. Export the tables to parquet files to a Cloud Storage bucket and grant the storageinsights.viewer role on the bucket to the partner organization.
- C. Copy the tables you need to share to a dataset without CMEKs. Create an Analytics Hub listing for this dataset.
- D. Create an authorized view that contains the CMEK to decrypt the data when accessed.

**Correct Answer: C**

*Community vote distribution*

C (100%)

 **raaad** 3 weeks, 5 days ago

**Selected Answer: C**

- Create a copy of the necessary tables into a new dataset that doesn't use CMEK, ensuring the data is accessible without requiring the partner to have access to the encryption key.
- Analytics Hub can then be used to share this data securely and efficiently with the partner organization, maintaining control and governance over the shared data.

upvoted 2 times

 **e70ea9e** 1 month ago

**Selected Answer: C**

Preserves Key Confidentiality:

Avoids sharing your CMEK with the partner, upholding key security and control.

upvoted 2 times

You are developing an Apache Beam pipeline to extract data from a Cloud SQL instance by using JdbcIO. You have two projects running in Google Cloud. The pipeline will be deployed and executed on Dataflow in Project A. The Cloud SQL instance is running in Project B and does not have a public IP address. After deploying the pipeline, you noticed that the pipeline failed to extract data from the Cloud SQL instance due to connection failure. You verified that VPC Service Controls and shared VPC are not in use in these projects. You want to resolve this error while ensuring that the data does not go through the public internet. What should you do?

- A. Set up VPC Network Peering between Project A and Project B. Add a firewall rule to allow the peered subnet range to access all instances on the network.
- B. Turn off the external IP addresses on the Dataflow worker. Enable Cloud NAT in Project A.
- C. Add the external IP addresses of the Dataflow worker as authorized networks in the Cloud SQL instance.
- D. Set up VPC Network Peering between Project A and Project B. Create a Compute Engine instance without external IP address in Project B on the peered subnet to serve as a proxy server to the Cloud SQL database.

**Correct Answer: A**

*Community vote distribution*

D (69%)

A (31%)

👤 **lipa31** 4 days, 15 hours ago

**Selected Answer: D**

the reason : Cloud SQL supports private IP addresses through private service access. When you create a Cloud SQL instance, Cloud SQL creates the instance within its own virtual private cloud (VPC), called the Cloud SQL VPC. Enabling private IP requires setting up a peering connection between the Cloud SQL VPC and your VPC network.

upvoted 1 times

👤 **saschak94** 1 week, 6 days ago

**Selected Answer: D**

Using VPC Network Peering, Cloud SQL implements private service access internally, which allows internal IP addresses to connect across two VPC networks regardless of whether they belong to the same project or organization.

However, since VPC Network Peering isn't transitive, it only broadcasts routes between the two VPCs that are directly peered. If you have an additional VPC, it won't be able to access your Cloud SQL resources using the connection set up with your original VPC.

upvoted 1 times

👤 **datapassionate** 2 weeks ago

**Selected Answer: D**

D. Set up VPC Network Peering between Project A and Project B. Create a Compute Engine instance without external IP address in Project B on the peered subnet to serve as a proxy server to the Cloud SQL database.

upvoted 1 times

👤 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D is the most aligned to best practices for me

upvoted 1 times

👤 **BIGQUERY\_ALT\_ALT** 2 weeks, 5 days ago

**Selected Answer: D**

Option D is the correct answer. The reason is you cannot access cloud sql or alloydb instances from a peered vpc connection as they will be hosted in service project not in Project B. The VPC Peering doesn't give transitive routing so accessing cloud sql directly is not possible with the proxy vm. <https://cloud.google.com/vpc/docs/vpc-peering#spec-general>

upvoted 2 times

👤 **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: D**

D is the correct solution.

To allow the Dataflow workers in Project A to connect to the private Cloud SQL instance in Project B, you need to set up VPC Network Peering between the two projects.

Then create a Compute Engine instance without external IP in Project B on the peered subnet. This instance can serve as a proxy server to connect to the private Cloud SQL instance.

The Dataflow workers can connect through the peered network to the proxy instance, which then connects to Cloud SQL. This allows access to the private Cloud SQL instance without going over the public internet.

Option A would allow access but still goes over the public internet.

Option B and C would not work since the Cloud SQL instance does not have a public IP address.

So D is the right approach to resolve the connection issue while keeping the data private.

upvoted 3 times

👤 **raaad** 3 weeks, 5 days ago

**Selected Answer: A**

VPC Network Peering allows for the connection of two VPC networks so that they can communicate internally as if they were part of the same network.

upvoted 2 times

 **e70ea9e** 1 month ago

**Selected Answer: A**

Secure Private Communication:

Establishes a direct, private connection between the VPCs, eliminating exposure to the public internet.

Ensures data confidentiality and integrity.

upvoted 2 times

Question #217

*Topic 1*

You have a BigQuery table that contains customer data, including sensitive information such as names and addresses. You need to share the customer data with your data analytics and consumer support teams securely. The data analytics team needs to access the data of all the customers, but must not be able to access the sensitive data. The consumer support team needs access to all data columns, but must not be able to access customers that no longer have active contracts. You enforced these requirements by using an authorized dataset and policy tags. After implementing these steps, the data analytics team reports that they still have access to the sensitive columns. You need to ensure that the data analytics team does not have access to restricted data. What should you do? (Choose two.)

- A. Create two separate authorized datasets; one for the data analytics team and another for the consumer support team.
- B. Ensure that the data analytics team members do not have the Data Catalog Fine-Grained Reader role for the policy tags.
- C. Replace the authorized dataset with an authorized view. Use row-level security and apply filter\_expression to limit data access.
- D. Remove the bigquery.dataViewer role from the data analytics team on the authorized datasets.
- E. Enforce access control in the policy tag taxonomy.

**Correct Answer: E**

*Community vote distribution*

B (43%)

E (29%)

A (29%)

✉  **datapassionate** 2 weeks ago

**Selected Answer: E**

B& E

<https://cloud.google.com/bigquery/docs/column-level-security-intro>

upvoted 1 times

✉  **GCP001** 2 weeks, 1 day ago

**Selected Answer: E**

B & E

B - It will ensure they don't have access to secure columns

E- It will allow to enforce column level security

Ref - <https://cloud.google.com/bigquery/docs/column-level-security-intro>

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B& E to me

upvoted 1 times

✉  **MaxNRG** 3 weeks, 1 day ago

**Selected Answer: A**

A & B

The current setup is not effective because the data analytics team still has access to the sensitive columns despite using an authorized dataset and policy tags. This indicates that the policy tags are not being enforced properly, and the data analytics team members are able to view the tags and gain access to the sensitive data.

Separating the data into two distinct authorized datasets is a better approach because it isolates the sensitive data from the non-sensitive data. This prevents the data analytics team from accessing the sensitive columns directly, even if they have access to the authorized dataset for general customer data.

Additionally, revoking the Data Catalog Fine-Grained Reader role from the data analytics team members ensures that they cannot view or modify the policy tags. This limits their ability to bypass the access control imposed by the authorized dataset and policy tags.

upvoted 2 times

✉  **Matt\_108** 2 weeks, 2 days ago

Max I feel like it's more B&E.

I do agree on the revoking Data Catalog Fine-grained reader role to avoid the data analytics team to read policy tags metadata, but if the tags are setup as stated, it's just missing the enforcement of the policy tags themselves.

Creating 2 auth dataset is not efficient on big datasets and Data catalog+ policy tags are built to manage these situations. Don't you agree

upvoted 1 times

✉  **qq589539483084gfrgrgfr** 3 weeks, 1 day ago

Option B & E

upvoted 4 times

✉  **imiu** 3 weeks, 4 days ago

And the second answer? One is option B and the other is option D maybe?

upvoted 1 times

✉  **raaad** 3 weeks, 5 days ago

**Selected Answer: B**

- The Data Catalog Fine-Grained Reader role allows users to read metadata that is restricted by policy tags.  
- If members of the data analytics team have this role, they might bypass the restrictions set by policy tags.

- Ensuring they do not have this role will help enforce the restrictions intended by the policy tags.

upvoted 1 times

✉  **e70ea9e** 1 month ago

**Selected Answer: B**

Prevents data analytics team members from viewing sensitive data, even if it's tagged.

Restricts access to policy tags themselves, ensuring confidentiality of sensitive information.

upvoted 1 times

You have a Cloud SQL for PostgreSQL instance in Region1 with one read replica in Region2 and another read replica in Region3. An unexpected event in Region1 requires that you perform disaster recovery by promoting a read replica in Region2. You need to ensure that your application has the same database capacity available before you switch over the connections. What should you do?

- A. Enable zonal high availability on the primary instance. Create a new read replica in a new region.
- B. Create a cascading read replica from the existing read replica in Region3.
- C. Create two new read replicas from the new primary instance, one in Region3 and one in a new region.
- D. Create a new read replica in Region1, promote the new read replica to be the primary instance, and enable zonal high availability.

**Correct Answer: B**

*Community vote distribution*

C (100%)

 **raaad** 3 weeks, 5 days ago

**Selected Answer: C**

After promoting the read replica in Region2 to be the new primary instance, creating additional read replicas from it can help distribute the read load and maintain or increase the database's total capacity.

upvoted 3 times

 **e70ea9e** 1 month ago

**Selected Answer: C**

Immediate Failover:

Promoting the read replica in Region2 quickly restores database operations in a different region, aligning with disaster recovery goals. Capacity Restoration:

Creates two new read replicas from the promoted primary instance (formerly the read replica in Region2). This replaces the lost capacity in Region1 and adds a read replica in a new region for further redundancy.

upvoted 2 times

You orchestrate ETL pipelines by using Cloud Composer. One of the tasks in the Apache Airflow directed acyclic graph (DAG) relies on a third-party service. You want to be notified when the task does not succeed. What should you do?

- A. Assign a function with notification logic to the `on_retry_callback` parameter for the operator responsible for the task at risk.
- B. Configure a Cloud Monitoring alert on the `sla_missed` metric associated with the task at risk to trigger a notification.
- C. Assign a function with notification logic to the `on_failure_callback` parameter for the operator responsible for the task at risk.
- D. Assign a function with notification logic to the `sla_miss_callback` parameter for the operator responsible for the task at risk.

**Correct Answer: D**

*Community vote distribution*

C (100%)

 **datapassionate** 2 weeks ago

**Selected Answer: C**

`on_failure_callback` is invoked when the task fails

<https://airflow.apache.org/docs/apache-airflow/stable/administration-and-deployment/logging-monitoring/callbacks.html>

upvoted 2 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C to me

upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- The `on_failure_callback` is a function that gets called when a task fails.

- Assigning a function with notification logic to this parameter is a direct way to handle task failures.

- When the task fails, this function can trigger a notification, making it an appropriate solution for the need to be alerted on task failures.

upvoted 4 times

 **e70ea9e** 1 month ago

**Selected Answer: C**

Direct Trigger:

The `on_failure_callback` parameter is specifically designed to invoke a function when a task fails, ensuring immediate notification.

Customizable Logic:

You can tailor the notification function to send emails, create alerts, or integrate with other notification systems, providing flexibility.

upvoted 3 times

You are migrating your on-premises data warehouse to BigQuery. One of the upstream data sources resides on a MySQL database that runs in your on-premises data center with no public IP addresses. You want to ensure that the data ingestion into BigQuery is done securely and does not go through the public internet. What should you do?

- A. Update your existing on-premises ETL tool to write to BigQuery by using the BigQuery Open Database Connectivity (ODBC) driver. Set up the proxy parameter in the `simba.googlebigqueryodbc.ini` file to point to your data center's NAT gateway.
- B. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Set up Cloud Interconnect between your on-premises data center and Google Cloud. Use Private connectivity as the connectivity method and allocate an IP address range within your VPC

network to the Datastream connectivity configuration. Use Server-only as the encryption type when setting up the connection profile in Datastream.

C. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Use Forward-SSH tunnel as the connectivity method to establish a secure tunnel between Datastream and your on-premises MySQL database through a tunnel server in your on-premises data center. Use None as the encryption type when setting up the connection profile in Datastream.

D. Use Datastream to replicate data from your on-premises MySQL database to BigQuery. Gather Datastream public IP addresses of the Google Cloud region that will be used to set up the stream. Add those IP addresses to the firewall allowlist of your on-premises data center. Use IP Allowlisting as the connectivity method and Server-only as the encryption type when setting up the connection profile in Datastream.

**Correct Answer: C**

*Community vote distribution*

B (100%)

✉  **datapassionate** 2 weeks ago

**Selected Answer: B**

Datastream is a seamless replication from relational databases directly to BigQuery. The source database can be hosted on-premises, on Google Cloud services such as Cloud SQL or Bare Metal Solution for Oracle, or anywhere else on any cloud.

<https://cloud.google.com/datastream-for-bigquery#benefits>

upvoted 1 times

✉  **datapassionate** 2 weeks ago

It is required that the data ingestion into BigQuery is done securely and does not go through the public internet. It can be done by Cloud Interconnect.

upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

- Datastream is a serverless change data capture and replication service, which can be used to replicate data changes from MySQL to BigQuery.
- Using Cloud Interconnect provides a private, secure connection between your on-premises environment and Google Cloud ==> This method ensures that data doesn't go through the public internet and is a recommended approach for secure, large-scale data migrations.
- Setting up private connectivity with Datastream allows for secure and direct data transfer.

upvoted 4 times

✉  **e70ea9e** 1 month ago

**Selected Answer: B**

Secure Private Connection:

Cloud Interconnect establishes a direct, private connection between your on-premises network and Google Cloud, bypassing the public internet and ensuring data confidentiality.

Datastream Integration:

Datastream seamlessly replicates data from your MySQL database to BigQuery, handling the complexities of data transfer and synchronization.

upvoted 2 times

You store and analyze your relational data in BigQuery on Google Cloud with all data that resides in US regions. You also have a variety of object stores across Microsoft Azure and Amazon Web Services (AWS), also in US regions. You want to query all your data in BigQuery daily with as little movement of data as possible. What should you do?

- A. Use BigQuery Data Transfer Service to load files from Azure and AWS into BigQuery.
- B. Create a Dataflow pipeline to ingest files from Azure and AWS to BigQuery.
- C. Load files from AWS and Azure to Cloud Storage with Cloud Shell gsutil rsync arguments.
- D. Use the BigQuery Omni functionality and BigLake tables to query files in Azure and AWS.

**Correct Answer: B**

*Community vote distribution*

D (100%)

 **e70ea9e** Highly Voted 1 month ago

**Selected Answer: D**

Direct Querying:

BigQuery Omni allows you to query data in Azure and AWS object stores directly without physically moving it to BigQuery, reducing data transfer costs and delays.

BigLake Tables:

Provide a unified view of both BigQuery tables and external object storage files, enabling seamless querying across multi-cloud data.

upvoted 5 times

 **raaad** Most Recent 3 weeks, 4 days ago

**Selected Answer: D**

- BigQuery Omni allows us to analyze data stored across Google Cloud, AWS, and Azure directly from BigQuery without having to move or copy the data.
- It extends BigQuery's data analysis capabilities to other clouds, enabling cross-cloud analytics.

upvoted 4 times

You have a variety of files in Cloud Storage that your data science team wants to use in their models. Currently, users do not have a method to explore, cleanse, and validate the data in Cloud Storage. You are looking for a low code solution that can be used by your data science team to quickly cleanse and explore data within Cloud Storage. What should you do?

- A. Provide the data science team access to Dataflow to create a pipeline to prepare and validate the raw data and load data into BigQuery for data exploration.
- B. Create an external table in BigQuery and use SQL to transform the data as necessary. Provide the data science team access to the external tables to explore the raw data.
- C. Load the data into BigQuery and use SQL to transform the data as necessary. Provide the data science team access to staging tables to explore the raw data.
- D. Provide the data science team access to Dataprep to prepare, validate, and explore the data within Cloud Storage.

**Correct Answer: C**

*Community vote distribution*

✉  **raaad** Highly Voted 3 weeks, 4 days ago

**Selected Answer: D**

- Dataprep is a serverless, no-code data preparation tool that allows users to visually explore, cleanse, and prepare data for analysis.
- It's designed for business analysts, data scientists, and others who want to work with data without writing code.
- Dataprep can directly access and transform data in Cloud Storage, making it a suitable choice for a team that prefers a low-code, user-friendly solution.

upvoted 8 times

✉  **JimmyBK** Most Recent 1 week, 6 days ago

**Selected Answer: D**

Goes without say

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D - Low code and efficient way to explore and prep data

upvoted 1 times

✉  **Alex3551** 3 weeks, 1 day ago

why you message wrong answers  
correct is C

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

The "Reveal Answer" button contains 90% of the time an incorrect answer. You should always check the community and the discussion during studying :)

upvoted 1 times

✉  **e70ea9e** 1 month ago

**Selected Answer: D**

Low-Code Interface:

Offers a visual, drag-and-drop interface that empowers users with varying technical skills to cleanse and explore data without extensive coding or alignment with the low-code requirement.

Data Cleaning and Validation:

Provides built-in tools for data profiling, cleaning, transformation, and validation, ensuring data quality and accuracy before model training.

Direct Cloud Storage Access:

Connects directly to Cloud Storage, allowing users to work with data in place without additional data movement or storage costs, optimizing efficiency.

upvoted 3 times

You are building an ELT solution in BigQuery by using Dataform. You need to perform uniqueness and null value checks on your final tables. What should you do to efficiently integrate these checks into your pipeline?

- A. Build BigQuery user-defined functions (UDFs).
- B. Create Dataplex data quality tasks.
- C. Build Dataform assertions into your code.
- D. Write a Spark-based stored procedure.

**Correct Answer: A**

*Community vote distribution*

C (100%)

 **tibuenoc** 1 week, 4 days ago

**Selected Answer: C**  
<https://cloud.google.com/dataform/docs/assertions>  
upvoted 2 times

 **Alex3551** 3 weeks, 1 day ago

**Selected Answer: C**  
Agree with C  
upvoted 1 times

 **Alex3551** 3 weeks, 1 day ago

agree with C  
upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: C**  
- Dataform provides a feature called "assertions," which are essentially SQL-based tests that you can define to verify the quality of your data.  
- Assertions in Dataform are a built-in way to perform data quality checks, including checking for uniqueness and null values in your tables.  
upvoted 4 times

 **e70ea9e** 1 month ago

**Selected Answer: C**  
Native Integration:

Dataform assertions are designed specifically for data quality checks within Dataform pipelines, ensuring seamless integration and compatibility. They leverage Dataform's execution model and configuration, aligning with the existing workflow.  
Declarative Syntax:

Assertions are defined using a simple, declarative syntax within Dataform code, making them easy to write and understand, even for users with less SQL expertise.

upvoted 2 times



A web server sends click events to a Pub/Sub topic as messages. The web server includes an eventTimestamp attribute in the messages, which is the time when the click occurred. You have a Dataflow streaming job that reads from this Pub/Sub topic through a subscription, applies some transformations, and writes the result to another Pub/Sub topic for use by the advertising department. The advertising department needs to receive each message within 30 seconds of the corresponding click occurrence, but they report receiving the messages late. Your Dataflow job's system lag is about 5 seconds, and the data freshness is about 40 seconds. Inspecting a few messages show no more than 1 second lag between their eventTimestamp and publishTime. What is the problem and what should you do?

- A. The advertising department is causing delays when consuming the messages. Work with the advertising department to fix this.
- B. Messages in your Dataflow job are taking more than 30 seconds to process. Optimize your job or increase the number of workers to fix this.
- G. Messages in your Dataflow job are processed in less than 30 seconds, but your job cannot keep up with the backlog in the Pub/Sub subscription. Optimize your job or increase the number of workers to fix this.
- D. The web server is not pushing messages fast enough to Pub/Sub. Work with the web server team to fix this.

**Correct Answer: D**

*Community vote distribution*

G (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: G**

Option C - low system lag (which identifies fast processing) but high data freshness (which identifies that the messages sit in the backlog a long time)

upvoted 1 times

 **Alex3551** 3 weeks, 1 day ago

**Selected Answer: G**

agree correct is C

upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: G**

- It suggests a backlog problem.

- It indicates that while individual messages might be processed quickly once they're handled, the job overall cannot keep up with the rate of incoming messages, causing a delay in processing the backlog.

upvoted 3 times

 **datapassionate** 2 weeks ago

Why not B than?

upvoted 1 times

 **e70ea9e** 1 month ago

**Selected Answer: G**

System Lag vs. Data Freshness: System lag is low (5 seconds), indicating that individual messages are processed quickly. However, data freshness is high (40 seconds), suggesting a backlog in the pipeline.

Not Advertising's Fault: The issue is upstream of their consumption, as they're already receiving delayed messages.

Not Web Server's Fault: The lag between eventTimestamp and publishTime is minimal (1 second), meaning the server is publishing messages promptly.

upvoted 4 times

Your organization stores customer data in an on-premises Apache Hadoop cluster in Apache Parquet format. Data is processed on a daily basis by Apache Spark jobs that run on the cluster. You are migrating the Spark jobs and Parquet data to Google Cloud. BigQuery will be used on future transformation pipelines so you need to ensure that your data is available in BigQuery. You want to use managed services, while minimizing ETL data processing changes and overhead costs. What should you do?

- A. Migrate your data to Cloud Storage and migrate the metadata to Dataproc Metastore (DPMS). Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc Serverless.
- B. Migrate your data to Cloud Storage and register the bucket as a Dataplex asset. Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc Serverless.
- C. Migrate your data to BigQuery. Refactor Spark pipelines to write and read data on BigQuery, and run them on Dataproc Serverless.
- D. Migrate your data to BigLake. Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc on Compute Engine.

**Correct Answer: C**

*Community vote distribution*

A (100%)

 **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- This option involves moving Parquet files to Cloud Storage, which is a common and cost-effective storage solution for big data and is compatible with Spark jobs.
- Using Dataproc Metastore to manage metadata allows us to keep Hadoop ecosystem's structural information.
- Running Spark jobs on Dataproc Serverless takes advantage of managed Spark services without managing clusters.
- Once the data is in Cloud Storage, you can also easily load it into BigQuery for further analysis.

upvoted 3 times

 **e70ea9e** 1 month ago

**Selected Answer: A**

Managed Services: Leverages Dataproc Serverless for a fully managed Spark environment, reducing overhead and administrative tasks.  
Minimal Data Processing Changes: Keeps Spark pipelines largely intact by working with Parquet files on Cloud Storage, minimizing refactoring efforts.  
BigQuery Integration: Dataproc Serverless can directly access BigQuery, enabling future transformation pipelines without additional data movement.  
Cost-Effective: Serverless model scales resources only when needed, optimizing costs for intermittent workloads.

upvoted 2 times

Your organization has two Google Cloud projects, project A and project B. In project A, you have a Pub/Sub topic that receives data from confidential sources. Only the resources in project A should be able to access the data in that topic. You want to ensure that project B and any future project cannot access data in the project A topic. What should you do?

- A. Add firewall rules in project A so only traffic from the VPC in project A is permitted.
- B. Configure VPC Service Controls in the organization with a perimeter around project A.
- C. Use Identity and Access Management conditions to ensure that only users and service accounts in project A. can access resources in project A.
- D. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.

**Correct Answer: D**

*Community vote distribution*

B (83%)

C (17%)

✉  **datapassionate** 2 weeks ago

**Selected Answer: C**

And I would agree with GPT. The question is about that who can do what within GCP environment. It's all about permissions and access management, not about networking.

upvoted 1 times

✉  **datapassionate** 2 weeks ago

GPT:

C. Use Identity and Access Management conditions to ensure that only users and service accounts in project A can access resources in project A.

Analysis: This is the most appropriate option. IAM allows you to define who (which users or service accounts) has what access to your GCP resources. By setting IAM policies with conditions specific to Project A, you can ensure that only designated entities within Project A have access to its resources, including the Pub/Sub topic.

D. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.

upvoted 1 times

✉  **datapassionate** 2 weeks ago

A. Add firewall rules in project A so only traffic from the VPC in project A is permitted.

Analysis: Firewall rules in GCP are used to control traffic to and from instances within Google Cloud Virtual Private Clouds (VPCs). However, they don't specifically control access to Pub/Sub resources. Pub/Sub access is managed through IAM, not VPC firewall rules.

upvoted 1 times

✉  **datapassionate** 2 weeks ago

B. Configure VPC Service Controls in the organization with a perimeter around project A.

Analysis: VPC Service Controls provide a security perimeter for your data, but they are more focused on preventing data exfiltration; this might be more complex and broader than necessary for the specific requirement of restricting access to a Pub/Sub topic.

upvoted 1 times

✉  **datapassionate** 1 week, 6 days ago

D. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.

Analysis: Similar to option B, this is focused on securing network boundaries rather than specific resource access within GCP. While it could provide an additional layer of security, it's not the most direct way to control access to a specific Pub/Sub topic.

upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

Option B:

-It allows us to create a secure boundary around all resources in Project A, including the Pub/Sub topic.

- It prevents data exfiltration to other projects and ensures that only resources within the perimeter (Project A) can access the sensitive data.

- VPC Service Controls are specifically designed for scenarios where you need to secure sensitive data within a specific context or boundary in Google Cloud.

upvoted 3 times

✉  **e70ea9e** 1 month ago

**Selected Answer: B**

VPC Service Controls enforce a security perimeter around entire projects, ensuring that resources within project A (including the Pub/Sub topic) are inaccessible from any other project, including project B and future projects.

This aligns with the requirement to prevent cross-project access.

upvoted 2 times

You stream order data by using a Dataflow pipeline, and write the aggregated result to Memorystore. You provisioned a Memorystore for Redis instance with Basic Tier, 4 GB capacity, which is used by 40 clients for read-only access. You are expecting the number of read-only clients to increase significantly to a few hundred and you need to be able to support the demand. You want to ensure that read and write access availability is not impacted, and any changes you make can be deployed quickly. What should you do?

- A. Create a new Memorystore for Redis instance with Standard Tier. Set capacity to 4 GB and read replica to No read replicas (high availability only). Delete the old instance.
- B. Create a new Memorystore for Redis instance with Standard Tier. Set capacity to 5 GB and create multiple read replicas. Delete the old instance.
- C. Create a new Memorystore for Memcached instance. Set a minimum of three nodes, and memory per node to 4 GB. Modify the Dataflow pipeline and all clients to use the Memcached instance. Delete the old instance.
- D. Create multiple new Memorystore for Redis instances with Basic Tier (4 GB capacity). Modify the Dataflow pipeline and new clients to use all instances.

**Correct Answer: C**

*Community vote distribution*

B (100%)

 **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

- Upgrading to the Standard Tier and adding read replicas is an effective way to scale and manage increased read load.
  - The additional capacity (5 GB) provides more space for data, and read replicas help distribute the read load across multiple instances.
- upvoted 4 times

 **datapassionate** 2 weeks ago

Described here:

<https://cloud.google.com/memorystore/docs/redis/redis-tiers>

upvoted 1 times

 **e70ea9e** 1 month ago

**Selected Answer: B**

Scalability for Read-Only Clients: Read replicas distribute read traffic across multiple instances, significantly enhancing read capacity to support a large number of clients without impacting write performance.

High Availability: Standard Tier ensures high availability with automatic failover, minimizing downtime in case of instance failure.

Minimal Code Changes: Redis clients can seamlessly connect to read replicas without requiring extensive code modifications, enabling a quick deployment.

upvoted 1 times

You have a streaming pipeline that ingests data from Pub/Sub in production. You need to update this streaming pipeline with improved business logic. You need to ensure that the updated pipeline reprocesses the previous two days of delivered Pub/Sub messages. What should you do?

(Choose two.)

- A. Use the Pub/Sub subscription clear-retry-policy flag
- B. Use Pub/Sub Snapshot capture two days before the deployment.
- C. Create a new Pub/Sub subscription two days before the deployment.
- D. Use the Pub/Sub subscription retain-acked-messages flag.
- E. Use Pub/Sub Seek with a timestamp.

**Correct Answer: D**

*Community vote distribution*

B (58%)

D (42%)

 **GCP001** 1 week, 4 days ago

**Selected Answer: B**

B and E, already tested at cloud console.  
upvoted 1 times

 **tibuenoc** 1 week, 4 days ago

**Selected Answer: D**

DE

Another way to replay messages that have been acknowledged is to seek to a timestamp. To seek to a timestamp, you must first configure the subscription to retain acknowledged messages using retain-acked-messages. If retain-acked-messages is set, Pub/Sub retains acknowledged messages for 7 days.

You only need to do this step if you intend to seek to a timestamp, not to a snapshot.

<https://cloud.google.com/pubsub/docs/replay-message>

upvoted 1 times

 **Sofia98** 2 weeks ago

**Selected Answer: B**

BE  
<https://cloud.google.com/pubsub/docs/replay-overview>  
upvoted 1 times

 **tibuenoc** 1 week, 4 days ago

But There is a problem snapshot you shoudl seek by subscriptions not by timestamp  
upvoted 1 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D and E  
upvoted 1 times

 **task\_7** 2 weeks, 5 days ago

**Selected Answer: D**

DE  
Set the retain-acked-messages flag to true for the subscription.  
This instructs Pub/Sub to store acknowledged messages for a specified retention period.

E Use Pub/Sub Seek with a timestamp.

After deploying the updated pipeline, use the Seek feature to replay messages.

Specify a timestamp that's two days before the current time.

This rewinds the subscription's message cursor, making it redeliver messages from that point onward.

upvoted 3 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

- Pub/Sub Snapshots allow you to capture the state of a subscription's unacknowledged messages at a particular point in time.
- By creating a snapshot two days before deploying the updated pipeline, you can later use this snapshot to replay the messages from that point in time.

=====

Option E:

- Pub/Sub Seek allows us to alter the acknowledgment state of messages in bulk.
- So we can rewind a subscription to a point in time or a snapshot.
- Using Seek with a timestamp corresponding to two days ago would allow the updated pipeline to reprocess messages from that time.

upvoted 3 times

✉  **datapassionate** 2 weeks ago

This case is described here.

<https://cloud.google.com/pubsub/docs/replay-message>

And according to this D & E would be correct.

upvoted 2 times

✉  **datapassionate** 2 weeks ago

Another way to replay messages that have been acknowledged is to seek to a timestamp. To seek to a timestamp, you must first configure the subscription to retain acknowledged messages using retain-acked-messages. If retain-acked-messages is set, Pub/Sub retains acknowledged messages for 7 days.

upvoted 2 times

✉  **datapassionate** 2 weeks ago

Creating a snapshot of the Pub/Sub subscription two days before the deployment captures the state of unacknowledged messages at that particular point in time, which would include messages from before those two days. If our objective is to reprocess the data from the last two days specifically, then capturing a snapshot two days prior wouldn't directly address this need.

upvoted 2 times

✉  **e70ea9e** 1 month ago

**Selected Answer: B**

BE--> correct

Pub/Sub Snapshot: Captures a point-in-time snapshot of the messages in the subscription, ensuring that the previous two days of messages are available for reprocessing even after they've been acknowledged.

Retain-Acked-Messages Flag: While this flag prevents acknowledged messages from being deleted, it's not sufficient on its own because it only retains messages going forward from when it's enabled.

upvoted 2 times

Question #229

Topic 1

You currently use a SQL-based tool to visualize your data stored in BigQuery. The data visualizations require the use of outer joins and analytic functions. Visualizations must be based on data that is no less than 4 hours old. Business users are complaining that the visualizations are too slow to generate. You want to improve the performance of the visualization queries while minimizing the maintenance overhead of the data preparation pipeline. What should you do?

- Create materialized views with the allow\_non\_incremental\_definition option set to true for the visualization queries. Specify the max\_staleness parameter to 4 hours and the enable\_refresh parameter to true. Reference the materialized views in the data visualization tool.
- Create views for the visualization queries. Reference the views in the data visualization tool.
- Create a Cloud Function instance to export the visualization query results as parquet files to a Cloud Storage bucket. Use Cloud Scheduler to trigger the Cloud Function every 4 hours. Reference the parquet files in the data visualization tool.
- Create materialized views for the visualization queries. Use the incremental updates capability of BigQuery materialized views to handle changed data automatically. Reference the materialized views in the data visualization tool.

**Correct Answer: B**

*Community vote distribution*

A (100%)

👤 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A is better than D, since it accounts for data staleness and is better suited for heavy querying, thanks to the `allow_non_incremental_definition`

upvoted 1 times

👤 **Jordan18** 3 weeks, 3 days ago

A seems right but what's wrong with option D, can anybody please explain?

upvoted 4 times

👤 **datapassionate** 2 weeks ago

Seems like materialized views can use incremental updates only if data was not deleted or updated in original table. Here the data changed so I think that's the reason why it's not the correct answer

[https://cloud.google.com/bigquery/docs/materialized-views-use#incremental\\_updates](https://cloud.google.com/bigquery/docs/materialized-views-use#incremental_updates)

"BigQuery combines the cached view's data with new data to provide consistent query results while still using the materialized view. For single-table materialized views, this is possible if the base table is unchanged since the last refresh, or if only new data was added. For multi-table views, no more than one table can have appended data. If more than one of a multi-table view's base tables has changed, then the view cannot be incrementally updated."

upvoted 2 times

👤 **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- Materialized views in BigQuery precompute and store the result of a base query, which can speed up data retrieval for complex queries used in visualizations.

- The `max_staleness` parameter allows us to specify how old the data can be, ensuring that the visualizations are based on data no less than 4 hours old.

- The `enable_refresh` parameter ensures that the materialized view is periodically refreshed.

- The `allow_non_incremental_definition` is used for enabling the creation of non-incrementally refreshable materialized views.

upvoted 2 times

👤 **e70ea9e** 1 month ago

**Selected Answer: A**

Precomputed Results: Materialized views store precomputed results of complex queries, significantly accelerating subsequent query performance, addressing the slow visualization issue.

Allow Non-Incremental Views: Using `allow_non_incremental_definition` circumvents the limitation of incremental updates for outer joins and analytic functions, ensuring views can be created for the specified queries.

Near-Real-Time Data: Setting `max_staleness` to 4 hours guarantees data freshness within the acceptable latency for visualizations.

Automatic Refresh: Enabling refresh with `enable_refresh` maintains view consistency with minimal maintenance overhead.

Minimal Overhead: Materialized views automatically update as underlying data changes, reducing maintenance compared to manual exports and view definitions.

upvoted 1 times

You need to modernize your existing on-premises data strategy. Your organization currently uses:

- Apache Hadoop clusters for processing multiple large data sets, including on-premises Hadoop Distributed File System (HDFS) for data replication.
- Apache Airflow to orchestrate hundreds of ETL pipelines with thousands of job steps.

You need to set up a new architecture in Google Cloud that can handle your Hadoop workloads and requires minimal changes to your existing orchestration processes. What should you do?

- A. Use Bigtable for your large workloads, with connections to Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer.
- B. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer.
- C. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Convert your ETL pipelines to Dataflow.
- D. Use Dataproc to migrate your Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Use Cloud Data Fusion to visually design and deploy your ETL pipelines.

**Correct Answer: A**

*Community vote distribution*

B (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

definitely B

upvoted 2 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

Straight forward

upvoted 4 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: B**

Cloud Composer -> Airflow

upvoted 3 times

You recently deployed several data processing jobs into your Cloud Composer 2 environment. You notice that some tasks are failing in Apache Airflow. On the monitoring dashboard, you see an increase in the total workers memory usage, and there were worker pod evictions. You need to resolve these errors. What should you do? (Choose two.)

- A. Increase the directed acyclic graph (DAG) file parsing interval.
- B. Increase the Cloud Composer 2 environment size from medium to large.
- C. Increase the maximum number of workers and reduce worker concurrency.

D. Increase the memory available to the Airflow workers.

E. Increase the memory available to the Airflow triggerer.

**Correct Answer: D**

*Community vote distribution*

C (70%)

B (30%)

 **qq589539483084gfrgrgfr** 1 week, 5 days ago

**Selected Answer: C**

CD It is clear

upvoted 2 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

C & D to me

upvoted 1 times

 **GCP001** 2 weeks, 5 days ago

**Selected Answer: C**

C and D

Check ref for memory optimization - <https://cloud.google.com/composer/docs/composer-2/optimize-environments>

upvoted 3 times

 **AllenChen123** 2 weeks, 2 days ago

Agree. Straightforward.

<https://cloud.google.com/composer/docs/composer-2/optimize-environments#monitor-scheduler>

-> Figure 3. Graph that displays worker pod evictions

upvoted 2 times

 **qq589539483084gfrgrgfr** 3 weeks ago

**Selected Answer: B**

B&D See this-

<https://cloud.google.com/composer/docs/composer-2/troubleshooting-dags#task-fails-without-logs>

go through the suggested fixes for If there are airflow-worker pods that show Evicted

upvoted 1 times

 **Jordan18** 3 weeks, 2 days ago

**Selected Answer: C**

C and D

upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

B&D:

B :

- Scaling up the environment size can provide more resources, including memory, to the Airflow workers. If worker pod evictions are occurring due to insufficient memory, increasing the environment size to allocate more resources could alleviate the problem and improve the stability of your data processing jobs.

D:

- Increase the memory available to the Airflow workers. - Directly increasing the memory allocation for Airflow workers can address the issue of high memory usage and worker pod evictions. More memory per worker means that each worker can handle more demanding tasks or a high volume of tasks without running out of memory.

upvoted 2 times

 **GCP001** 1 week, 4 days ago

why not B ) It's not decreasing concurrency which may cause issue again

upvoted 1 times

You are on the data governance team and are implementing security requirements to deploy resources. You need to ensure that resources are limited to only the europe-west3 region. You want to follow Google-recommended practices.

What should you do?

- A. Set the constraints/gcp.resourceLocations organization policy constraint to in:europe-west3-locations.
- B. Deploy resources with Terraform and implement a variable validation rule to ensure that the region is set to the europe-west3 region for all resources.
- C. Set the constraints/gcp.resourceLocations organization policy constraint to in:eu-locations.
- D. Create a Cloud Function to monitor all resources created and automatically destroy the ones created outside the europe-west3 region.

**Correct Answer: C**

*Community vote distribution*

A (100%)

 **raaad** Highly Voted 3 weeks, 4 days ago

**Selected Answer: A**

- The constraints/gcp.resourceLocations organization policy constraint is used to define where resources in the organization can be created.
  - Setting it to in:europe-west3-locations would specify that resources can only be created in the europe-west3 region.
- upvoted 5 times

 **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: A**

Option A  
upvoted 2 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

Set the constraints/gcp.resourceLocations organization policy constraint to in:europe-west3-locations.  
upvoted 1 times

You are a BigQuery admin supporting a team of data consumers who run ad hoc queries and downstream reporting in tools such as Looker. All data and users are combined under a single organizational project. You recently noticed some slowness in query results and want to troubleshoot where the slowdowns are occurring. You think that there might be some job queuing or slot contention occurring as users run jobs, which slows down access to results. You need to investigate the query job information and determine where performance is being affected. What should you do?

- A. Use slot reservations for your project to ensure that you have enough query processing capacity and are able to allocate available slots to the slower queries.
- B. Use Cloud Monitoring to view BigQuery metrics and set up alerts that let you know when a certain percentage of slots were used.
- C. Use available administrative resource charts to determine how slots are being used and how jobs are performing over time. Run a query on the INFORMATION\_SCHEMA to review query performance.
- D. Use Cloud Logging to determine if any users or downstream consumers are changing or deleting access grants on tagged resources.

**Correct Answer: D***Community vote distribution*

C (100%)

**✉️** **Matt\_108** 2 weeks, 2 days ago**Selected Answer: C**

Option C

upvoted 1 times

**✉️** **raaad** 3 weeks, 4 days ago**Selected Answer: C**

- BigQuery provides administrative resource charts that show slot utilization and job performance, which can help identify patterns of heavy usage or contention.
- Additionally, querying the INFORMATION\_SCHEMA with the JOBS or JOBS\_BY\_PROJECT view can provide detailed information about specific queries, including execution time, slot usage, and whether they were queued.

upvoted 4 times

**✉️** **datapassionate** 2 weeks ago

described here:

<https://cloud.google.com/blog/products/data-analytics/troubleshoot-bigquery-performance-with-these-dashboards>

upvoted 1 times

**✉️** **scaenruy** 3 weeks, 5 days ago**Selected Answer: C**

- C. Use available administrative resource charts to determine how slots are being used and how jobs are performing over time. Run a query or the INFORMATION\_SCHEMA to review query performance.

upvoted 1 times

Question #234

Topic 1

You migrated a data backend for an application that serves 10 PB of historical product data for analytics. Only the last known state for a product, which is about 10 GB of data, needs to be served through an API to the other applications. You need to choose a cost-effective persistent storage solution that can accommodate the analytics requirements and the API performance of up to 1000 queries per second (QPS) with less than 1 second latency. What should you do?

- A. 1. Store the historical data in BigQuery for analytics.
2. Use a materialized view to precompute the last state of a product.

3. Serve the last state data directly from BigQuery to the API.
- B. 1. Store the products as a collection in Firestore with each product having a set of historical changes.  
 2. Use simple and compound queries for analytics.  
 3. Serve the last state data directly from Firestore to the API.
- C. 1. Store the historical data in Cloud SQL for analytics.  
 2. In a separate table, store the last state of the product after every product change.  
 3. Serve the last state data directly from Cloud SQL to the API.
- D. 1. Store the historical data in BigQuery for analytics.  
 2. In a Cloud SQL table, store the last state of the product after every product change.  
 3. Serve the last state data directly from Cloud SQL to the API.

**Correct Answer: C**

*Community vote distribution*

D (67%)

A (33%)

✉  **datapassionate** 2 weeks ago

**Selected Answer: D**

- D. 1. Store the historical data in BigQuery for analytics.  
 2. In a Cloud SQL table, store the last state of the product after every product change.  
 3. Serve the last state data directly from Cloud SQL to the API

This approach leverages BigQuery's scalability and efficiency for handling large datasets for analytics. BigQuery is well-suited for managing up to 10 PB of historical product data. Meanwhile, Cloud SQL provides the necessary performance to handle the API queries with the required low latency. By storing the latest state of each product in Cloud SQL, you can efficiently handle the high QPS with sub-second latency, which is crucial for the API's performance. This combination of BigQuery and Cloud SQL offers a balanced solution for both the large-scale analytics and the high-performance API needs.

upvoted 3 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D is the right one, compared to option A, Cloud SQL is more efficient and cost effective for the amount of time the data needs to be accessed by the api

upvoted 1 times

✉  **einchkrein** 3 weeks ago

Serve the last state data directly from Cloud SQL to the API.  
 Here's why this option is most suitable:

**BigQuery for Analytics:** BigQuery is an excellent choice for storing and analyzing large datasets like your 10 PB of historical product data. It is designed for handling big data analytics efficiently and cost-effectively.

**Cloud SQL for Last State Data:** Cloud SQL is a fully managed relational database that can effectively handle the storage of the last known state of products. Storing this subset of data (about 10 GB) in Cloud SQL allows for optimized and faster query performance for your API needs. Cloud SQL can comfortably handle the requirement of up to 1000 QPS with sub-second latency.

**Separation of Concerns:** This approach separates the analytics workload (BigQuery) from the operational query workload (Cloud SQL). This separation ensures that analytics queries do not interfere with the operational performance of the API and vice versa.

upvoted 3 times

✉  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: A**

- A. 1. Store the historical data in BigQuery for analytics.  
 2. Use a materialized view to precompute the last state of a product.  
 3. Serve the last state data directly from BigQuery to the API.

upvoted 2 times

You want to schedule a number of sequential load and transformation jobs. Data files will be added to a Cloud Storage bucket by an upstream process. There is no fixed schedule for when the new data arrives. Next, a Dataproc job is triggered to perform some transformations and write the data to BigQuery. You then need to run additional transformation jobs in BigQuery. The transformation jobs are different for every table. These jobs might take hours to complete. You need to determine the most efficient and maintainable workflow to process hundreds of tables and provide the freshest data to your end users. What should you do?

- A. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc, and BigQuery operators.  
2. Use a single shared DAG for all tables that need to go through the pipeline.  
3. Schedule the DAG to run hourly.
- B. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc, and BigQuery operators.  
2. Create a separate DAG for each table that needs to go through the pipeline.  
3. Schedule the DAGs to run hourly.
- C. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.  
2. Use a single shared DAG for all tables that need to go through the pipeline.  
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.
- D. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.  
2. Create a separate DAG for each table that needs to go through the pipeline.  
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.

Community vote distribution

D (100%)

✉️ **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D, which gets triggered when the data comes in and accounts for the fact that each table has its own set of transformations  
upvoted 1 times

✉️ **Jordan18** 3 weeks, 2 days ago

why not C?

upvoted 2 times

✉️ **AllenChen123** 2 weeks, 2 days ago

Same question, why not use single DAG to manage as there are hundreds of tables.

upvoted 2 times

✉️ **raaad** 3 weeks, 4 days ago

**Selected Answer: D**

- Option D: Tailored handling and scheduling for each table; triggered by data arrival for more timely and efficient processing.  
upvoted 1 times

✉️ **scaenruy** 3 weeks, 5 days ago

**Selected Answer: D**

D.  
1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.  
2. Create a separate DAG for each table that needs to go through the pipeline.  
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.  
upvoted 1 times

Question #236

Topic 1

You are deploying a MySQL database workload onto Cloud SQL. The database must be able to scale up to support several readers from various geographic regions. The database must be highly available and meet low RTO and RPO requirements, even in the event of a regional outage. You need to ensure that interruptions to the readers are minimal during a database failover. What should you do?

- A. Create a highly available Cloud SQL instance in region A. Create a highly available read replica in region B. Scale up read workloads by creating cascading read replicas in multiple regions. Backup the Cloud SQL instances to a multi-regional Cloud Storage bucket. Restore the Cloud SQL backup to a new instance in another region when Region A is down.
- B. Create a highly available Cloud SQL instance in region A. Scale up read workloads by creating read replicas in multiple regions. Promote one of the read replicas when region A is down.
- C. Create a highly available Cloud SQL instance in region A. Create a highly available read replica in region B. Scale up read workloads by creating cascading read replicas in multiple regions. Promote the read replica in region B when region A is down.
- D. Create a highly available Cloud SQL instance in region A. Scale up read workloads by creating read replicas in the same region. Failover to the standby Cloud SQL instance when the primary instance fails.

**Correct Answer: D**

*Community vote distribution*

B (40%)

C (40%)

A (20%)

✉  **tibuenoc** 1 week, 1 day ago

**Selected Answer: B**

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

This option involves having read replicas in multiple regions, allowing you to promote one of them in the event of a failure in region A. While this may still be a brief interruption during the failover, it is likely to be less than the time required for the synchronization of cascading read replicas. upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

To me, it's B. it provides:

High availability: The highly available Cloud SQL instance in region A will ensure that the database remains accessible even if one of the zones in the region becomes unavailable.

Scalability: The read replicas in multiple regions will enable you to scale up the read capacity of the database to support the demands of read from various geographic regions.

Minimal interruptions: When region A is down, one of the read replicas in another region will be promoted to become the new primary instance. This will ensure that there is no interruption to the readers.

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

Why not others:

Approach A: This approach requires you to restore a backup from a different region, which could take some time. This could result in a significant RPO (Recovery Point Objective) for the database. Additionally, the restored instance may not be physically located in the same region as the readers, which could impact performance.

Approach C: This approach requires you to promote the read replica in region B, which could result in a temporary interruption to the reads while the promotion is taking place. Additionally, the read replica in region B may not be able to handle the same level of read traffic as the primary instance in region A.

Approach D: This approach does not provide the same level of scalability as the other approaches, as you are limited to read replicas in the same region. Additionally, failover to the standby instance could result in a temporary interruption to the readers.

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

Ignore my previous messages, it's C :D

upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- Combines high availability with geographic distribution of read workloads.

- Promoting a highly available read replica can provide a quick failover solution, potentially meeting low RTO and RPO requirements.

=====

Why not A:

Restoring from backup to a new instance in another region during a regional outage might not meet low RTO and RPO requirements due to the time it takes to perform a restore.

upvoted 2 times

✉  **AllenChen123** 2 weeks, 2 days ago

Why not B?

upvoted 1 times

✉  **datapassionate** 2 weeks ago

Why not B:

While B option scales up read workloads across multiple regions, it doesn't specify high availability for the read replica in another region. In the event of a regional outage, promoting a non-highly available read replica might not provide the desired uptime and reliability.

upvoted 2 times

✉  **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A.

Create a highly available Cloud SQL instance in region A. Create a highly available read replica in region B. Scale up read workloads by creating cascading read replicas in multiple regions. Backup the Cloud SQL instances to a multi-regional Cloud Storage bucket. Restore the Cloud SQL backup to a new instance in another region when Region A is down.

upvoted 1 times

You are planning to load some of your existing on-premises data into BigQuery on Google Cloud. You want to either stream or batch-load data, depending on your use case. Additionally, you want to mask some sensitive data before loading into BigQuery. You need to do this in a programmatic way while keeping costs to a minimum. What should you do?

- A. Use Cloud Data Fusion to design your pipeline, use the Cloud DLP plug-in to de-identify data within your pipeline, and then move the data into BigQuery.
- B. Use the BigQuery Data Transfer Service to schedule your migration. After the data is populated in BigQuery, use the connection to the Cloud Data Loss Prevention (Cloud DLP) API to de-identify the necessary data.
- C. Create your pipeline with Dataflow through the Apache Beam SDK for Python, customizing separate options within your code for streaming, batch processing, and Cloud DLP. Select BigQuery as your data sink.
- D. Set up Datastream to replicate your on-premise data on BigQuery.

**Correct Answer: C**

*Community vote distribution*

C (86%)

14%

✉️  **raaad** Highly Voted 3 weeks, 4 days ago

**Selected Answer: C**

- Programmatic Flexibility: Apache Beam provides extensive control over pipeline design, allowing for customization of data transformations, including integration with Cloud DLP for sensitive data masking.
- Streaming and Batch Support: Beam seamlessly supports both streaming and batch data processing modes, enabling flexibility in data load patterns.
- Cost-Effective Processing: Dataflow offers a serverless model, scaling resources as needed, and only charging for resources used, helping optimize costs.
- Integration with Cloud DLP: Beam integrates well with Cloud DLP for sensitive data masking, ensuring data privacy before loading into BigQuery.

upvoted 5 times

✉️  **qq589539483084gfrgrgfr** 3 weeks ago

In correct Option is A because you want a programmatic way whereas datafusion is codeless solution and also dataflow is cost effective

upvoted 1 times

✉️  **AllenChen123** 2 weeks, 2 days ago

You are saying Option C

upvoted 2 times

✉️  **tibuenoc** Most Recent 1 week, 1 day ago

**Selected Answer: C**

C is correct. Using Dataflow as Python as programming and BQ as sink.

A is incorrect - DataFusion is Code-free as the main propose

upvoted 1 times

✉️  **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A.

Use Cloud Data Fusion to design your pipeline, use the Cloud DLP plug-in to de-identify data within your pipeline, and then move the data into BigQuery.

upvoted 1 times

You want to encrypt the customer data stored in BigQuery. You need to implement per-user crypto-deletion on data stored in your tables. You want to adopt native features in Google Cloud to avoid custom solutions. What should you do?

- A. Implement Authenticated Encryption with Associated Data (AEAD) BigQuery functions while storing your data in BigQuery.
- B. Create a customer-managed encryption key (CMEK) in Cloud KMS. Associate the key to the table while creating the table.
- C. Create a customer-managed encryption key (CMEK) in Cloud KMS. Use the key to encrypt data before storing in BigQuery.
- D. Encrypt your data during ingestion by using a cryptographic library supported by your ETL pipeline.

**Correct Answer: C**

*Community vote distribution*

A (100%)

 **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- AEAD cryptographic functions in BigQuery allow for encryption and decryption of data at the column level.
- You can encrypt specific data fields using a unique key per user and manage these keys outside of BigQuery (for example, in your application or using a key management system).
- By "deleting" or revoking access to the key for a specific user, you effectively make their data unreadable, achieving crypto-deletion.
- This method provides fine-grained encryption control but requires careful key management and integration with your applications.

upvoted 4 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

- A.  
Implement Authenticated Encryption with Associated Data (AEAD) BigQuery functions while storing your data in BigQuery.

upvoted 1 times

The data analyst team at your company uses BigQuery for ad-hoc queries and scheduled SQL pipelines in a Google Cloud project with a slot reservation of 2000 slots. However, with the recent introduction of hundreds of new non time-sensitive SQL pipelines, the team is encountering frequent quota errors. You examine the logs and notice that approximately 1500 queries are being triggered concurrently during peak time. You need to resolve the concurrency issue. What should you do?

- A. Increase the slot capacity of the project with baseline as 0 and maximum reservation size as 3000.
- B. Update SQL pipelines to run as a batch query, and run ad-hoc queries as interactive query jobs.
- C. Increase the slot capacity of the project with baseline as 2000 and maximum reservation size as 3000.
- D. Update SQL pipelines and ad-hoc queries to run as interactive query jobs.

**Correct Answer: A**

*Community vote distribution*

B (100%)

 **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

- BigQuery allows you to specify job priority as either BATCH or INTERACTIVE.
- Batch queries are queued and then started when idle resources are available, making them suitable for non-time-sensitive workloads.
- Running ad-hoc queries as interactive ensures they have prompt access to resources.

upvoted 3 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: B**

- B.
- Update SQL pipelines to run as a batch query, and run ad-hoc queries as interactive query jobs.

upvoted 2 times

You are designing a data mesh on Google Cloud by using Dataplex to manage data in BigQuery and Cloud Storage. You want to simplify data asset permissions. You are creating a customer virtual lake with two user groups:

- Data engineers, which require full data lake access
- Analytic users, which require access to curated data

You need to assign access rights to these two groups. What should you do?

- A. 1. Grant the dataplex.dataOwner role to the data engineer group on the customer data lake.  
2. Grant the dataplex.dataReader role to the analytic user group on the customer curated zone.
- B. 1. Grant the dataplex.dataReader role to the data engineer group on the customer data lake.  
2. Grant the dataplex.dataOwner to the analytic user group on the customer curated zone.
- C. 1. Grant the bigquery.dataOwner role on BigQuery datasets and the storage.objectCreator role on Cloud Storage buckets to data engineers.  
2. Grant the bigquery.dataViewer role on BigQuery datasets and the storage.objectViewer role on Cloud Storage buckets to analytic users.
- D. 1. Grant the bigquery.dataViewer role on BigQuery datasets and the storage.objectViewer role on Cloud Storage buckets to data engineers.  
2. Grant the bigquery.dataOwner role on BigQuery datasets and the storage.objectEditor role on Cloud Storage buckets to analytic users.

**Correct Answer: C**

*Community vote distribution*

A (100%)

✉️  **qq589539483084gfrgrgfr** 2 weeks, 1 day ago

**Selected Answer: A**

A correct answer

upvoted 2 times

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A clearly correct

upvoted 1 times

✉️  **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- dataplex.dataOwner: Grants full control over data assets, including reading, writing, managing, and granting access to others.

- dataplex.dataReader: Allows users to read data but not modify it.

upvoted 4 times

✉️  **AllenChen123** 2 weeks, 2 days ago

Yes, <https://cloud.google.com/dataplex/docs/lake-security#data-roles>

Dataplex maps its roles to the data roles for each underlying storage resource (Cloud Storage, BigQuery).

^ simplify the permissions.

upvoted 1 times

✉️  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: A**

A.

1. Grant the dataplex.dataOwner role to the data engineer group on the customer data lake.

2. Grant the dataplex.dataReader role to the analytic user group on the customer curated zone.

upvoted 1 times

You are designing the architecture of your application to store data in Cloud Storage. Your application consists of pipelines that read data from a Cloud Storage bucket that contains raw data, and write the data to a second bucket after processing. You want to design an architecture with Cloud Storage resources that are capable of being resilient if a Google Cloud regional failure occurs. You want to minimize the recovery point objective (RPO) if a failure occurs, with no impact on applications that use the stored data. What should you do?

- A. Adopt multi-regional Cloud Storage buckets in your architecture.
- B. Adopt two regional Cloud Storage buckets, and update your application to write the output on both buckets.
- C. Adopt a dual-region Cloud Storage bucket, and enable turbo replication in your architecture.
- D. Adopt two regional Cloud Storage buckets, and create a daily task to copy from one bucket to the other.

**Correct Answer: B**

*Community vote distribution*

C (86%)

14%

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C: <https://cloud.google.com/storage/docs/dual-regions> + <https://cloud.google.com/storage/docs/managing-turbo-replication>  
upvoted 1 times

 **therealsohail** 3 weeks, 2 days ago

**Selected Answer: C**

Turbo replication provides faster redundancy across regions for data in your dual-region buckets, which reduces the risk of data loss exposure and helps support uninterrupted service following a regional outage.

upvoted 2 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- Dual-region buckets are a specific type of storage that automatically replicates data between two geographically distinct regions.
- Turbo replication is an enhanced feature that provides faster replication between the two regions, thus minimizing RPO.
- This option ensures that your data is resilient to regional failures and is replicated quickly, meeting the needs for low RPO and no impact on application performance.

upvoted 3 times

 **scaenrui** 3 weeks, 5 days ago

**Selected Answer: A**

A. Adopt multi-regional Cloud Storage buckets in your architecture.

upvoted 1 times

 **datapassionate** 1 week, 6 days ago

It wont be a correct answer. Correct is C. It is required "no impact on applications that use the stored data"

upvoted 2 times

 **datapassionate** 1 week, 6 days ago

Whereas with multi-region " it can also introduce unpredictable latency into the response time and higher network egress charges for cl workloads when multi-region data is read from remote regions"

<https://cloud.google.com/blog/products/storage-data-transfer/choose-between-regional-dual-region-and-multi-region-cloud-storage>

upvoted 3 times

You have designed an Apache Beam processing pipeline that reads from a Pub/Sub topic. The topic has a message retention duration of one day, and writes to a Cloud Storage bucket. You need to select a bucket location and processing strategy to prevent data loss in case of a regional

outage with an RPO of 15 minutes. What should you do?

- A. 1. Use a dual-region Cloud Storage bucket.  
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.  
3. Seek the subscription back in time by 15 minutes to recover the acknowledged messages.  
4. Start the Dataflow job in a secondary region.
- B. 1. Use a multi-regional Cloud Storage bucket.  
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.  
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.  
4. Start the Dataflow job in a secondary region.
- C. 1. Use a regional Cloud Storage bucket.  
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.  
3. Seek the subscription back in time by one day to recover the acknowledged messages.  
4. Start the Dataflow job in a secondary region and write in a bucket in the same region.
- D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled.  
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.  
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.  
4. Start the Dataflow job in a secondary region.

**Correct Answer: C**

*Community vote distribution*

D (60%)

A (40%)

lipa31 4 days, 22 hours ago

Selected Answer: D

<https://cloud.google.com/storage/docs/availability-durability#turbo-replication> says : "When enabled, turbo replication is designed to replicate 100% of newly written objects to both regions that constitute the dual-region within the recovery point objective of 15 minutes, regardless of object size."

so seems D to me

upvoted 1 times

datapassionate 1 week, 6 days ago

Selected Answer: D

- D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

RPO of 15 minutes is guaranteed when turbo replication is used

<https://cloud.google.com/storage/docs/availability-durability>

upvoted 2 times

Question #243

Topic 1

You are preparing data that your machine learning team will use to train a model using BigQueryML. They want to predict the price per square foot of real estate. The training data has a column for the price and a column for the number of square feet. Another feature column called 'feature1' contains null values due to missing data. You want to replace the nulls with zeros to keep more data points. Which query should you use?

- ```
SELECT * EXCEPT(feature1),  
A.  IFNULL(feature1, 0) AS feature1_cleaned  
FROM training_data;
```
-
- ```
SELECT * EXCEPT(price, square_feet),
B. price/square_feet AS price_per_sqft
FROM training_data
WHERE feature1 IS NOT NULL;
```
- 
- ```
SELECT * EXCEPT(price, square_feet, feature1),  
C.  price/square_feet AS price_per_sqft,  
    IFNULL(feature1, 0) AS feature1_cleaned  
FROM training_data;
```
-
- ```
SELECT *
D. FROM training_data
WHERE feature1 IS NOT NULL;
```

Correct Answer: D

Community vote distribution

A (80%)

C (20%)

✉  **datapassionate** 1 week, 6 days ago

**Selected Answer: C**

Correct answer is C.

It both replace NULL with 0 and pass price per square foot of real estate.

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

option A clearly

upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

Straight forward

upvoted 3 times

Different teams in your organization store customer and performance data in BigQuery. Each team needs to keep full control of their collected data, be able to query data within their projects, and be able to exchange their data with other teams. You need to implement an organization-wide solution, while minimizing operational tasks and costs. What should you do?

- A. Ask each team to create authorized views of their data. Grant the `biquery.jobUser` role to each team.
- B. Create a BigQuery scheduled query to replicate all customer data into team projects.
- C. Ask each team to publish their data in Analytics Hub. Direct the other teams to subscribe to them.
- D. Enable each team to create materialized views of the data they need to access in their projects.

**Correct Answer: A**

*Community vote distribution*

C (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

that's what analytics hub is designed for  
upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- Analytics Hub allows organizations to create and manage exchanges where producers can publish their data and consumers can discover a subscribe to data products.
- Asking each team to publish their data in Analytics Hub and having other teams subscribe to them is a scalable and controlled way of sharing data.
- It minimizes operational tasks because data doesn't need to be duplicated or manually managed after setup, and teams can maintain full control over their datasets.

upvoted 1 times

✉  **rahulvin** 1 month ago

**Selected Answer: C**

Analytics hub to reduce operational overhead of creating/maintaining views permissions etc  
upvoted 1 times

You are developing a model to identify the factors that lead to sales conversions for your customers. You have completed processing your data. You want to continue through the model development lifecycle. What should you do next?

- A. Use your model to run predictions on fresh customer input data.
- B. Monitor your model performance, and make any adjustments needed.
- C. Delineate what data will be used for testing and what will be used for training the model.
- D. Test and evaluate your model on your curated data to determine how well the model performs.

**Correct Answer: C**

*Community vote distribution*

C (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C - you've just concluded processing data, ending up with clean and prepared data for the model. Now you need to decide how to split the data for testing and for training. Only afterwards, you can train the model, evaluate it, fine tune it and, eventually, predict with it

upvoted 2 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- Before you can train a model, you need to decide how to split your dataset.

upvoted 4 times

 **rahulvin** 1 month ago

**Selected Answer: C**

Model doesn't seem to be trained yet

upvoted 3 times

You have one BigQuery dataset which includes customers' street addresses. You want to retrieve all occurrences of street addresses from the dataset. What should you do?

- A. Write a SQL query in BigQuery by using REGEXP\_CONTAINS on all tables in your dataset to find rows where the word "street" appears.
- B. Create a deep inspection job on each table in your dataset with Cloud Data Loss Prevention and create an inspection template that includes the STREET\_ADDRESS infoType.
- C. Create a discovery scan configuration on your organization with Cloud Data Loss Prevention and create an inspection template that includes the STREET\_ADDRESS infoType.
- D. Create a de-identification job in Cloud Data Loss Prevention and use the masking transformation.

**Correct Answer: C**

*Community vote distribution*

B (100%)

 **AllenChen123** 2 weeks ago

Why not C? Discovery scan configuration can also help to identify risk/sensitivity fields.

upvoted 1 times

 **datapassionate** 1 week, 6 days ago

In the question we need to retrieve all occurrences of street addresses from the dataset. In C you create discovery configuration plan on whole organization. Its not needed.

upvoted 2 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B - you want to retrieve ALL occurrences within the dataset

upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: B**

- Cloud Data Loss Prevention (Cloud DLP) provides powerful inspection capabilities for sensitive data, including predefined detectors for infoTypes such as STREET\_ADDRESS.

- By creating a deep inspection job for each table with the STREET\_ADDRESS infoType, you can accurately identify and retrieve rows that contain street addresses.

upvoted 2 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: B**

B. Create a deep inspection job on each table in your dataset with Cloud Data Loss Prevention and create an inspection template that includes the STREET\_ADDRESS infoType.

upvoted 1 times

Your company operates in three domains: airlines, hotels, and ride-hailing services. Each domain has two teams: analytics and data science, which create data assets in BigQuery with the help of a central data platform team. However, as each domain is evolving rapidly, the central data platform team is becoming a bottleneck. This is causing delays in deriving insights from data, and resulting in stale data when pipelines are not kept up to date. You need to design a data mesh architecture by using Dataplex to eliminate the bottleneck. What should you do?

- A. 1. Create one lake for each team. Inside each lake, create one zone for each domain.
- 2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.

3. Have the central data platform team manage all zones' data assets.
- B. 1. Create one lake for each team. Inside each lake, create one zone for each domain.
2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.
  3. Direct each domain to manage their own zone's data assets.
- C. 1. Create one lake for each domain. Inside each lake, create one zone for each team.
2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.
  3. Direct each domain to manage their own lake's data assets.
- D. 1. Create one lake for each domain. Inside each lake, create one zone for each team.
2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.
  3. Have the central data platform team manage all lakes' data assets.

**Correct Answer: D**

*Community vote distribution*

C (88%)

13%

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**  
Option C - create a lake for each domain, each team manages its own assets  
upvoted 2 times

 **task\_7** 2 weeks, 5 days ago

**Selected Answer: B**  
Separate lakes for each team  
Zones within each lake dedicated to different domains  
upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: C**  
- each domain should manage their own lake's data assets  
upvoted 4 times

 **AllenChen123** 2 weeks ago

Agree. [https://cloud.google.com/dataplex/docs/introduction#a\\_domain-centric\\_data\\_mesh](https://cloud.google.com/dataplex/docs/introduction#a_domain-centric_data_mesh)  
upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**  
C.  
1. Create one lake for each domain. Inside each lake, create one zone for each team.  
2. Attach each of the BigQuery datasets created by the individual teams as assets to the respective zone.  
3. Direct each domain to manage their own lake's data assets.  
upvoted 1 times

dataset.inventory\_vm sample records:

Row	id	name	components.name	components.qty
1	vm02781	d-jp-kfk-02-02	vcpu	2
			memory	8
			boot_disk	10
			disk_1	50
2	vm11490	i-jp-kfk-02-07	vcpu	16
			memory	64
			boot_disk	10
			disk_1	200
3	vm18130	i-jp-kfk-02-08	vcpu	8
			memory	8
			boot_disk	10

You have an inventory of VM data stored in the BigQuery table. You want to prepare the data for regular reporting in the most cost-effective way. You need to exclude VM rows with fewer than 8 vCPU in your report. What should you do?

- A. Create a view with a filter to drop rows with fewer than 8 vCPU, and use the UNNEST operator.
- B. Create a materialized view with a filter to drop rows with fewer than 8 vCPU, and use the WITH common table expression.
- C. Create a view with a filter to drop rows with fewer than 8 vCPU, and use the WITH common table expression.
- D. Use Dataflow to batch process and write the result to another BigQuery table.

**Correct Answer: B**

*Community vote distribution*

A (100%)

✉  **Krauser59** 2 weeks, 1 day ago

**Selected Answer: A**

A seems to be the correct answer because of the table structure and the UNNEST operator. However, i don't understand why wouldn't we chose a materialized view  
upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A - The regular reporting doesn't justify a materialized view, since the frequency of access is not so high; a simple view would do the trick. Moreover, the vcpu data is in a nested field and requires Unnest.  
upvoted 1 times

✉  **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- The table structure shows that the vCPU data is stored in a nested field within the components column.  
- Using the UNNEST operator to flatten the nested field and apply the filter.  
upvoted 2 times

✉  **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A. Create a view with a filter to drop rows with fewer than 8 vCPU, and use the UNNEST operator.  
upvoted 2 times

Your team is building a data lake platform on Google Cloud. As a part of the data foundation design, you are planning to store all the raw data in Cloud Storage. You are expecting to ingest approximately 25 GB of data a day and your billing department is worried about the increasing cost of storing old data. The current business requirements are:

- The old data can be deleted anytime.
- There is no predefined access pattern of the old data.
- The old data should be available instantly when accessed.
- There should not be any charges for data retrieval.

What should you do to optimize for cost?

- A. Create the bucket with the Autoclass storage class feature.
- B. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline, and 365 days to archive storage class. Delete old data as needed.
- C. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to coldline, 90 days to nearline, and 365 days to archive storage class. Delete old data as needed.
- D. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 45 days to coldline, and 60 days to archive storage class. Delete old data as needed.

**Correct Answer: C**

*Community vote distribution*

A (90%)

10%

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

For sure A, read the documentation

upvoted 2 times

 **GCP001** 3 weeks ago

**Selected Answer: A**

autoclass is the correct way to handle all business cases

upvoted 1 times

 **Smakyel79** 3 weeks, 1 day ago

**Selected Answer: A**

<https://cloud.google.com/storage/docs/autoclass>

upvoted 3 times

 **therealsohail** 3 weeks, 2 days ago

**Selected Answer: B**

Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline, and 365 days to archive storage class. Delete old data as needed.

upvoted 1 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: A**

- Autoclass automatically moves objects between storage classes without impacting performance or availability, nor incurring retrieval costs.
- It continuously optimizes storage costs based on access patterns without the need to set specific lifecycle management policies.

upvoted 3 times

Your company's data platform ingests CSV file dumps of booking and user profile data from upstream sources into Cloud Storage. The data analyst team wants to join these datasets on the email field available in both the datasets to perform analysis. However, personally identifiable information (PII) should not be accessible to the analysts. You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts. What should you do?

- A. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud Data Loss Prevention (Cloud DLP) with masking as the de-identification transformations type.  
2. Load the booking and user profile data into a BigQuery table.
- B. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP with format-preserving encryption with FFX as the de-identification transformation type.  
2. Load the booking and user profile data into a BigQuery table.
- C. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.  
2. Create a policy tag with the email mask as the data masking rule.  
3. Assign the policy to the email field in both tables. A  
4. Assign the Identity and Access Management `bigrquerydatapolicy.maskedReader` role for the BigQuery tables to the analysts.
- D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.  
2. Create a policy tag with the default masking value as the data masking rule.  
3. Assign the policy to the email field in both tables.  
4. Assign the Identity and Access Management `bigrquerydatapolicy.maskedReader` role for the BigQuery tables to the analysts

**Correct Answer: B***Community vote distribution*

B (44%)	C (44%)	11%
---------	---------	-----

✉️  **lipa31** 4 days, 22 hours ago

**Selected Answer: B**

Format-preserving encryption (FPE) with FFX in Cloud DLP is a strong choice for de-identifying PII like email addresses. FPE maintains the form of the data and ensures that the same input results in the same encrypted output consistently. This means the email fields in both datasets can be encrypted to the same value, allowing for accurate joins in BigQuery while keeping the actual email addresses hidden.

upvoted 1 times

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C. The need is to just mask the data to Analyst, without modifying the underlying data. Moreover, it's stored on 2 separate tables and the analysts need to be able to perform joins based on the masked data. Dynamic masking is the right module and the right masking rule is email mask ([https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking\\_options](https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking_options)) which guarantees the join capabilities join upvoted 1 times

✉️  **task\_7** 2 weeks, 5 days ago

**Selected Answer: B**

A wouldn't preserve the email format  
C&D maskedReader roles still grant access to the underlying values.  
the only option is B  
upvoted 3 times

✉️  **alfguemat** 2 weeks, 4 days ago

I don't know why preserve email format is necessary to perform the join. A could be valid.  
upvoted 1 times

✉️ **Sofia98** 2 weeks, 6 days ago

**Selected Answer: C**

I will go for C, because there is a separate type of masking for emails, so whe to use the default?  
[https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking\\_options](https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking_options)

upvoted 1 times

✉️ **GCP001** 3 weeks ago

**Selected Answer: C**

data masking with BQ is correct with email masking rule.  
Ref - <https://cloud.google.com/bigquery/docs/column-data-masking-intro>

upvoted 1 times

✉️ **Smakyel79** 3 weeks, 1 day ago

As it states "You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts" data masking should be an option as the data would stored unmasked in BigQuery?

upvoted 2 times

✉️ **Jordan18** 3 weeks, 2 days ago

why not B?

upvoted 2 times

✉️ **raaad** 3 weeks, 4 days ago

**Selected Answer: C**

- The reason option C works well is that dynamic data masking in BigQuery allows the underlying data to remain unaltered (thus preserving the ability to join on this field), while also preventing analysts from viewing the actual PII.  
- The analysts can query and join the data as needed for their analysis, but when they access the data, the email field will be masked according to the policy tag, and they will only see the masked version.

upvoted 1 times

✉️ **scaenruy** 3 weeks, 5 days ago

**Selected Answer: D**

D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.  
2. Create a policy tag with the default masking value as the data masking rule.  
3. Assign the policy to the email field in both tables.  
4. Assign the Identity and Access Management `bigrquerydatapolicy.maskedReader` role for the BigQuery tables to the analysts

upvoted 1 times

You have important legal hold documents in a Cloud Storage bucket. You need to ensure that these documents are not deleted or modified. What should you do?

- A. Set a retention policy. Lock the retention policy.
- B. Set a retention policy. Set the default storage class to Archive for long-term digital preservation.
- C. Enable the Object Versioning feature. Add a lifecycle rule.
- D. Enable the Object Versioning feature. Create a copy in a bucket in a different region.

**Correct Answer: C**

*Community vote distribution*

A (100%)

  **raaad** Highly Voted 3 weeks, 4 days ago

**Selected Answer: A**

- Setting a retention policy on a Cloud Storage bucket prevents objects from being deleted for the duration of the retention period.  
- Locking the policy makes it immutable, meaning that the retention period cannot be reduced or removed, thus ensuring that the documents cannot be deleted or overwritten until the retention period expires.

upvoted 5 times

  **AllenChen123** 1 week, 5 days ago

Agree. <https://cloud.google.com/storage/docs/bucket-lock#overview>

upvoted 1 times

  **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: A**

Option A - set retention policy to prevent deletion, lock it to make it immutable (not subject to edits)

upvoted 1 times

  **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A. Set a retention policy. Lock the retention policy.

upvoted 1 times

You are designing a data warehouse in BigQuery to analyze sales data for a telecommunication service provider. You need to create a data model for customers, products, and subscriptions. All customers, products, and subscriptions can be updated monthly, but you must maintain a historical record of all data. You plan to use the visualization layer for current and historical reporting. You need to ensure that the data model is simple, easy-to-use, and cost-effective. What should you do?

- A. Create a normalized model with tables for each entity. Use snapshots before updates to track historical data.
- B. Create a normalized model with tables for each entity. Keep all input files in a Cloud Storage bucket to track historical data.
- C. Create a denormalized model with nested and repeated fields. Update the table and use snapshots to track historical data.
- D. Create a denormalized, append-only model with nested and repeated fields. Use the ingestion timestamp to track historical data.**

**Correct Answer: D***Community vote distribution*

D (100%)

  **JimmyBK** 1 week, 6 days ago**Selected Answer: D**

Straight forward, good for costs  
upvoted 1 times

  **Sofia98** 2 weeks, 6 days ago**Selected Answer: D**

D looks logical  
upvoted 1 times

  **GCP001** 3 weeks ago**Selected Answer: D**

Easy, cost effective and no complexity  
upvoted 1 times

  **raaad** 3 weeks, 4 days ago

- A denormalized, append-only model simplifies query complexity by eliminating the need for joins.
- Adding data with an ingestion timestamp allows for easy retrieval of both current and historical states.
- Instead of updating records, new records are appended, which maintains historical information without the need to create separate snapshots.

upvoted 4 times

  **scaenruy** 3 weeks, 5 days ago**Selected Answer: D**

D. Create a denormalized, append-only model with nested and repeated fields. Use the ingestion timestamp to track historical data.  
upvoted 2 times

You are deploying a batch pipeline in Dataflow. This pipeline reads data from Cloud Storage, transforms the data, and then writes the data into BigQuery. The security team has enabled an organizational constraint in Google Cloud, requiring all Compute Engine instances to use only internal IP addresses and no external IP addresses. What should you do?

- A. Ensure that your workers have network tags to access Cloud Storage and BigQuery. Use Dataflow with only internal IP addresses.
- B. Ensure that the firewall rules allow access to Cloud Storage and BigQuery. Use Dataflow with only internal IPs.
- C. Create a VPC Service Controls perimeter that contains the VPC network and add Dataflow, Cloud Storage, and BigQuery as allowed services

in the perimeter. Use Dataflow with only internal IP addresses.

D. Ensure that Private Google Access is enabled in the subnetwork. Use Dataflow with only internal IP addresses.

**Correct Answer: C**

*Community vote distribution*

D (63%)

C (38%)

 **pandeyspecial** 1 day, 13 hours ago

**Selected Answer: C**

It should be C

upvoted 1 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option D, as GCP001 said

upvoted 1 times

 **Matt\_108** 2 weeks, 2 days ago

Misscicked the answer <.<

upvoted 1 times

 **GCP001** 3 weeks ago

**Selected Answer: D**

<https://cloud.google.com/dataflow/docs/guides/routes-firewall>

upvoted 2 times

 **raaad** 3 weeks, 4 days ago

**Selected Answer: D**

- Private Google Access for services allows VM instances with only internal IP addresses in a VPC network or on-premises networks (via Cloud VPN or Cloud Interconnect) to reach Google APIs and services.
- When you launch a Dataflow job, you can specify that it should use worker instances without external IP addresses if Private Google Access is enabled on the subnetwork where these instances are launched.
- This way, your Dataflow workers will be able to access Cloud Storage and BigQuery without violating the organizational constraint of no external IPs.

upvoted 3 times

 **Jordan18** 3 weeks, 2 days ago

why not C?

upvoted 2 times

 **BIGQUERY\_ALT\_ALT** 2 weeks, 5 days ago

VPC Service Controls are typically used to define and enforce security perimeters around APIs and services, restricting their access to a specified set of Google Cloud projects. In this scenario, the security constraint is focused on Compute Engine instances used by Dataflow and VPC Service Controls might be considered a bit heavy-handed for just addressing the internal IP address requirement.

upvoted 2 times

 **GCP001** 3 weeks ago

Even if you create VPC service control, your dataflow worker will run on Google Compute Engine instances with private IPs only after policy enforcement.

Without external IP addresses, you can still perform administrative and monitoring tasks.

You can access your workers by using SSH through the options listed in the preceding list. However, the pipeline cannot access the internet, and internet hosts cannot access your Dataflow workers.

upvoted 2 times

 **GCP001** 3 weeks ago

ref - <https://cloud.google.com/dataflow/docs/guides/routes-firewall>

upvoted 2 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

C. Create a VPC Service Controls perimeter that contains the VPC network and add Dataflow, Cloud Storage, and BigQuery as allowed services in the perimeter. Use Dataflow with only internal IP addresses.

upvoted 1 times

 **BIGQUERY\_ALT\_ALT** 2 weeks, 5 days ago

C is wrong. Option D is simple and straight forward. VPC Service Controls are typically used to define and enforce security perimeters around APIs and services, restricting their access to a specified set of Google Cloud projects. In this scenario, the security constraint is focused on Compute Engine instances used by Dataflow, and VPC Service Controls might be considered a bit heavy-handed for just addressing the internal IP address requirement.

upvoted 1 times

You are running a Dataflow streaming pipeline, with Streaming Engine and Horizontal Autoscaling enabled. You have set the maximum number of workers to 1000. The input of your pipeline is Pub/Sub messages with notifications from Cloud Storage. One of the pipeline transforms reads CSV files and emits an element for every CSV line. The job performance is low, the pipeline is using only 10 workers, and you notice that the autoscaler is not spinning up additional workers. What should you do to improve performance?

- A. Enable Vertical Autoscaling to let the pipeline use larger workers.
- B. Change the pipeline code, and introduce a Reshuffle step to prevent fusion.
- C. Update the job to increase the maximum number of workers.
- D. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.

**Correct Answer: D**

*Community vote distribution*

B (89%)

11%

 **raaad** Highly Voted 3 weeks, 4 days ago

**Selected Answer: B**

- Fusion optimization in Dataflow can lead to steps being "fused" together, which can sometimes hinder parallelization.
  - Introducing a Reshuffle step can prevent fusion and force the distribution of work across more workers.
  - This can be an effective way to improve parallelism and potentially trigger the autoscaler to increase the number of workers.
- upvoted 7 times

 **GCP001** Most Recent 3 weeks ago

**Selected Answer: B**

Problem is performance and not using all workers properly, [https://cloud.google.com/dataflow/docs/pipeline-lifecycle#fusion\\_optimization](https://cloud.google.com/dataflow/docs/pipeline-lifecycle#fusion_optimization)  
upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: D**

- D. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.
- upvoted 1 times

You have an Oracle database deployed in a VM as part of a Virtual Private Cloud (VPC) network. You want to replicate and continuously synchronize 50 tables to BigQuery. You want to minimize the need to manage infrastructure. What should you do?

- A. Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle Change Data Capture (CDC), and Dataflow to stream the Kafka topic to BigQuery.
- B. Create a Pub/Sub subscription to write to BigQuery directly. Deploy the Debezium Oracle connector to capture changes in the Oracle database, and sink to the Pub/Sub topic.
- C. Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle change data capture (CDC), and the Kafka Connect Google BigQuery Sink Connector.
- D. Create a Datastream service from Oracle to BigQuery, use a private connectivity configuration to the same VPC network, and a connection profile to BigQuery.

**Correct Answer: B**

*Community vote distribution*

D (100%)

✉️  **raaad** 3 weeks, 4 days ago

**Selected Answer: D**

- Datastream is a serverless and easy-to-use change data capture (CDC) and replication service.
  - You would create a Datastream service that sources from your Oracle database and targets BigQuery, with private connectivity configuration the same VPC.
  - This option is designed to minimize the need to manage infrastructure and is a fully managed service.
- upvoted 4 times

✉️  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: D**

- D. Create a Datastream service from Oracle to BigQuery, use a private connectivity configuration to the same VPC network, and a connection profile to BigQuery.
- upvoted 1 times

You are deploying an Apache Airflow directed acyclic graph (DAG) in a Cloud Composer 2 instance. You have incoming files in a Cloud Storage bucket that the DAG processes, one file at a time. The Cloud Composer instance is deployed in a subnetwork with no Internet access. Instead of running the DAG based on a schedule, you want to run the DAG in a reactive way every time a new file is received. What should you do?

- A. 1. Enable Private Google Access in the subnetwork, and set up Cloud Storage notifications to a Pub/Sub topic.  
2. Create a push subscription that points to the web server URL.
- B. 1. Enable the Cloud Composer API, and set up Cloud Storage notifications to trigger a Cloud Function.  
2. Write a Cloud Function instance to call the DAG by using the Cloud Composer API and the web server URL.  
3. Use VPC Serverless Access to reach the web server URL.
- C. 1. Enable the Airflow REST API, and set up Cloud Storage notifications to trigger a Cloud Function instance.  
2. Create a Private Service Connect (PSC) endpoint.  
3. Write a Cloud Function that connects to the Cloud Composer cluster through the PSC endpoint.
- D. 1. Enable the Airflow REST API, and set up Cloud Storage notifications to trigger a Cloud Function instance.  
2. Write a Cloud Function instance to call the DAG by using the Airflow REST API and the web server URL.

3. Use VPC Serverless Access to reach the web server URL.

**Correct Answer: A**

*Community vote distribution*

C (100%)

 **raaad** Highly Voted 3 weeks, 3 days ago

**Selected Answer: C**

- Enable Airflow REST API: In Cloud Composer, enable the "Airflow web server" option.
- Set Up Cloud Storage Notifications: Create a notification for new files, routing to a Cloud Function.
- Create PSC Endpoint: Establish a PSC endpoint for Cloud Composer.
- Write Cloud Function: Code the function to use the Airflow REST API (via PSC endpoint) to trigger the DAG.

=====

Why not Option D

- Using the web server URL directly wouldn't work without internet access or a direct path to the web server.

upvoted 6 times

 **AllenChen123** 1 week, 3 days ago

Why not B, use Cloud Composer API

upvoted 1 times

 **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: C**

Option C, raaad explained well why

upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

C.

1. Enable the Airflow REST API, and set up Cloud Storage notifications to trigger a Cloud Function instance.
2. Create a Private Service Connect (PSC) endpoint.
3. Write a Cloud Function that connects to the Cloud Composer cluster through the PSC endpoint.

upvoted 1 times

You are planning to use Cloud Storage as part of your data lake solution. The Cloud Storage bucket will contain objects ingested from external systems. Each object will be ingested once, and the access patterns of individual objects will be random. You want to minimize the cost of storing and retrieving these objects. You want to ensure that any cost optimization efforts are transparent to the users and applications. What should you do?

- A. Create a Cloud Storage bucket with Autoclass enabled.
- B. Create a Cloud Storage bucket with an Object Lifecycle Management policy to transition objects from Standard to Coldline storage class if an object age reaches 30 days.
- C. Create a Cloud Storage bucket with an Object Lifecycle Management policy to transition objects from Standard to Coldline storage class if an object is not live.
- D. Create two Cloud Storage buckets. Use the Standard storage class for the first bucket, and use the Coldline storage class for the second bucket. Migrate objects from the first bucket to the second bucket after 30 days.

**Correct Answer: D***Community vote distribution*

A (100%)

**✉️**  **Matt\_108** 2 weeks, 2 days ago**Selected Answer: A**

Option A

upvoted 1 times

**✉️**  **raaad** 3 weeks, 3 days ago**Selected Answer: A**

- Autoclass automatically analyzes access patterns of objects and automatically transitions them to the most cost-effective storage class with Standard, Nearline, Coldline, or Archive.
- This eliminates the need for manual intervention or setting specific age thresholds.
- No user or application interaction is required, ensuring transparency.

upvoted 2 times

**✉️**  **scaenruy** 3 weeks, 5 days ago**Selected Answer: A**

- A. Create a Cloud Storage bucket with Autoclass enabled.

upvoted 1 times

You have several different file type data sources, such as Apache Parquet and CSV. You want to store the data in Cloud Storage. You need to set up an object sink for your data that allows you to use your own encryption keys. You want to use a GUI-based solution. What should you do?

- A. Use Storage Transfer Service to move files into Cloud Storage.
- B. Use Cloud Data Fusion to move files into Cloud Storage.
- C. Use Dataflow to move files into Cloud Storage.
- D. Use BigQuery Data Transfer Service to move files into BigQuery.

**Correct Answer: C**

*Community vote distribution*

B (71%)

A (29%)

✉  **Helinia** 1 week, 2 days ago

**Selected Answer: B**

Even though storage transfer service can be used in GUI, it does not support CMEK which is required in this question.

"Storage Transfer Service does not encrypt data on your behalf, such as in customer-managed encryption keys (CMEK). We only encrypt data in transit."

Ref: <https://cloud.google.com/storage-transfer/docs/on-prem-security>

upvoted 2 times

✉  **task\_7** 2 weeks, 4 days ago

**Selected Answer: A**

A. Use Storage Transfer Service to move files into Cloud Storage.  
move files into Cloud Storage should be Storage Transfer Service  
Cloud Data Fusion is like using a tank to kill an ant

upvoted 2 times

✉  **raaad** 3 weeks, 3 days ago

**Selected Answer: B**

- Cloud Data Fusion is a fully managed, code-free, GUI-based data integration service that allows you to visually connect, transform, and move data between various sources and sinks. - It supports various file formats and can write to Cloud Storage.  
- You can configure it to use Customer-Managed Encryption Keys (CMEK) for the buckets where it writes data.

upvoted 1 times

✉  **AllenChen123** 1 week, 3 days ago

Agree. <https://cloud.google.com/data-fusion/docs/how-to/customer-managed-encryption-keys#create-instance>  
upvoted 1 times

✉  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: B**

B. Use Cloud Data Fusion to move files into Cloud Storage.

upvoted 1 times

Your business users need a way to clean and prepare data before using the data for analysis. Your business users are less technically savvy and prefer to work with graphical user interfaces to define their transformations. After the data has been transformed, the business users want to perform their analysis directly in a spreadsheet. You need to recommend a solution that they can use. What should you do?

- A. Use Dataprep to clean the data, and write the results to BigQuery. Analyze the data by using Connected Sheets.
- B. Use Dataprep to clean the data, and write the results to BigQuery. Analyze the data by using Looker Studio.
- C. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Connected Sheets.
- D. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Looker Studio.

**Correct Answer: D**

*Community vote distribution*

A (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Clearly option A

upvoted 1 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

If only all the questions were like this...

upvoted 1 times

 **raaad** 3 weeks, 3 days ago

**Selected Answer: A**

- Allow business users to perform their analysis in a familiar spreadsheet interface via Connected Sheets.

upvoted 3 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A. Use Dataprep to clean the data, and write the results to BigQuery. Analyze the data by using Connected Sheets.

upvoted 2 times

You have two projects where you run BigQuery jobs:

- One project runs production jobs that have strict completion time SLAs. These are high priority jobs that must have the required compute resources available when needed. These jobs generally never go below a 300 slot utilization, but occasionally spike up an additional 500 slots.
- The other project is for users to run ad-hoc analytical queries. This project generally never uses more than 200 slots at a time. You want these ad-hoc queries to be billed based on how much data users scan rather than by slot capacity.

You need to ensure that both projects have the appropriate compute resources available. What should you do?

- A. Create a single Enterprise Edition reservation for both projects. Set a baseline of 300 slots. Enable autoscaling up to 700 slots.
- B. Create two reservations, one for each of the projects. For the SLA project, use an Enterprise Edition with a baseline of 300 slots and enable autoscaling up to 500 slots. For the ad-hoc project, configure on-demand billing.
- C. Create two Enterprise Edition reservations, one for each of the projects. For the SLA project, set a baseline of 300 slots and enable

autoscaling up to 500 slots. For the ad-hoc project, set a reservation baseline of 0 slots and set the ignore idle slots flag to False.

D. Create two Enterprise Edition reservations, one for each of the projects. For the SLA project, set a baseline of 800 slots. For the ad-hoc project, enable autoscaling up to 200 slots.

**Correct Answer: D**

*Community vote distribution*

B (67%)

D (33%)

 **danisxp** 1 week, 4 days ago

**Selected Answer: D**

Considering the emphasis on strict completion time SLA's. I go with option D. However I think both B and D are not the best solution here.  
upvoted 1 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B - first project works well with dedicated reservation and autoscaling. The second one requires on demand billing, as per question requires.  
upvoted 1 times

 **ElenaL** 2 weeks, 5 days ago

**Selected Answer: D**

"These jobs generally never go below a 300 slot utilization, but occasionally spike up an additional 500 slots." -> if it spikes up an ADDITIONAL 500 slots, on top of the regular 300, shouldn't we reserve at a minimum 800? open to explanations as to why this is not the case.  
upvoted 2 times

 **raaad** 3 weeks, 3 days ago

**Selected Answer: B**

- The SLA project gets a dedicated reservation with autoscaling to handle spikes, ensuring it meets its strict completion time SLAs.  
- The ad-hoc project uses on-demand billing, which means it will be billed based on the amount of data scanned rather than slot capacity, fitting the billing preference for ad-hoc queries.  
upvoted 3 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: B**

B. Create two reservations, one for each of the projects. For the SLA project, use an Enterprise Edition with a baseline of 300 slots and enable autoscaling up to 500 slots. For the ad-hoc project, configure on-demand billing.  
upvoted 2 times

You want to migrate your existing Teradata data warehouse to BigQuery. You want to move the historical data to BigQuery by using the most efficient method that requires the least amount of programming, but local storage space on your existing data warehouse is limited. What should you do?

- A. Use BigQuery Data Transfer Service by using the Java Database Connectivity (JDBC) driver with FastExport connection.
- B. Create a Teradata Parallel Transporter (TPT) export script to export the historical data, and import to BigQuery by using the bq command-line tool.
- C. Use BigQuery Data Transfer Service with the Teradata Parallel Transporter (TPT) tbuild utility.
- D. Create a script to export the historical data, and upload in batches to Cloud Storage. Set up a BigQuery Data Transfer Service instance from Cloud Storage to BigQuery.

**Correct Answer: C***Community vote distribution*

A (100%)

**✉️**  **Matt\_108** 2 weeks, 2 days ago**Selected Answer: A**

Option A, the JDBC driver is the key to solve the limited local storage  
upvoted 1 times

**✉️**  **raaad** 3 weeks, 3 days ago**Selected Answer: A**

- Reduced Local Storage: By using FastExport, data is directly streamed from Teradata to BigQuery without the need for local storage, addressing your storage limitations.
- Minimal Programming: BigQuery Data Transfer Service offers a user-friendly interface, eliminating the need for extensive scripting or coding.

upvoted 4 times

**✉️**  **AllenChen123** 1 week, 2 days ago

Agree. [https://cloud.google.com/bigquery/docs/migration/teradata-overview#extraction\\_method](https://cloud.google.com/bigquery/docs/migration/teradata-overview#extraction_method)

Extraction using a JDBC driver with FastExport connection. If there are constraints on the local storage space available for extracted files, there is some reason you can't use TPT, then use this extraction method.

upvoted 1 times

**✉️**  **rahulvin** 1 month ago**Selected Answer: A**

[https://cloud.google.com/bigquery/docs/migration/teradata-overview#extraction\\_method](https://cloud.google.com/bigquery/docs/migration/teradata-overview#extraction_method)

Lack of local storage pushes this to JDBC driver

upvoted 4 times

You are on the data governance team and are implementing security requirements. You need to encrypt all your data in BigQuery by using an encryption key managed by your team. You must implement a mechanism to generate and store encryption material only on your on-premises hardware security module (HSM). You want to rely on Google managed solutions. What should you do?

- A. Create the encryption key in the on-premises HSM, and import it into a Cloud Key Management Service (Cloud KMS) key. Associate the created Cloud KMS key while creating the BigQuery resources.
- B. Create the encryption key in the on-premises HSM and link it to a Cloud External Key Manager (Cloud EKM) key. Associate the created Cloud KMS key while creating the BigQuery resources.
- C. Create the encryption key in the on-premises HSM, and import it into Cloud Key Management Service (Cloud HSM) key. Associate the created Cloud HSM key while creating the BigQuery resources.
- D. Create the encryption key in the on-premises HSM. Create BigQuery resources and encrypt data while ingesting them into BigQuery.

**Correct Answer: D**

*Community vote distribution*

B (75%)

C (25%)

 **raaad** Highly Voted 3 weeks, 3 days ago

**Selected Answer: B**

- Cloud EKM allows you to use encryption keys managed in external key management systems, including on-premises HSMs, while using Google Cloud services.
- This means that the key material remains in your control and environment, and Google Cloud services use it via the Cloud EKM integration.
- This approach aligns with the need to generate and store encryption material only on your on-premises HSM and is the correct way to integrate such keys with BigQuery.

=====

Why not Option C

- Cloud HSM is a fully managed service by Google Cloud that provides HSMs for your cryptographic needs. However, it's a cloud-based solution, and the keys generated or managed in Cloud HSM are not stored on-premises. This option doesn't align with the requirement to use only on-premises HSM for key storage.

upvoted 8 times

 **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: B**

Option B, I agree with Raaad on the approach

upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

- C. Create the encryption key in the on-premises HSM, and import it into Cloud Key Management Service (Cloud HSM) key. Associate the created Cloud HSM key while creating the BigQuery resources.

upvoted 3 times

You maintain ETL pipelines. You notice that a streaming pipeline running on Dataflow is taking a long time to process incoming data, which causes output delays. You also noticed that the pipeline graph was automatically optimized by Dataflow and merged into one step. You want to identify where the potential bottleneck is occurring. What should you do?

- A. Insert a Reshuffle operation after each processing step, and monitor the execution details in the Dataflow console.
- B. Insert output sinks after each key processing step, and observe the writing throughput of each block.
- C. Log debug information in each ParDo function, and analyze the logs at execution time.
- D. Verify that the Dataflow service accounts have appropriate permissions to write the processed data to the output sinks.

**Correct Answer: C**

*Community vote distribution*

A (100%)

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

From the Dataflow documentation: "There are a few cases in your pipeline where you may want to prevent the Dataflow service from performing fusion optimizations. These are cases in which the Dataflow service might incorrectly guess the optimal way to fuse operations in the pipeline which could limit the Dataflow service's ability to make use of all available workers.

You can insert a Reshuffle step. Reshuffle prevents fusion, checkpoints the data, and performs deduplication of records. Reshuffle is supported by Dataflow even though it is marked deprecated in the Apache Beam documentation."

upvoted 1 times

 **raaad** 3 weeks, 3 days ago

**Selected Answer: A**

- The Reshuffle operation is used in Dataflow pipelines to break fusion and redistribute elements, which can sometimes help improve parallelization and identify bottlenecks.
- By inserting Reshuffle after each processing step and observing the pipeline's performance in the Dataflow console, you can potentially identify stages that are disproportionately slow or stalled.
- This can help in pinpointing the step where the bottleneck might be occurring.

upvoted 4 times

 **scaenry** 3 weeks, 5 days ago

**Selected Answer: A**

- A. Insert a Reshuffle operation after each processing step, and monitor the execution details in the Dataflow console.

upvoted 2 times

You are running your BigQuery project in the on-demand billing model and are executing a change data capture (CDC) process that ingests data. The CDC process loads 1 GB of data every 10 minutes into a temporary table, and then performs a merge into a 10 TB target table. This process is very scan intensive and you want to explore options to enable a predictable cost model. You need to create a BigQuery reservation based on utilization information gathered from BigQuery Monitoring and apply the reservation to the CDC process. What should you do?

- A. Create a BigQuery reservation for the dataset.
- B. Create a BigQuery reservation for the job.
- C. Create a BigQuery reservation for the service account running the job.
- D. Create a BigQuery reservation for the project.

**Correct Answer: A***Community vote distribution*

D (67%)

B (33%)

**✉️** **Matt\_108** 2 weeks, 2 days ago**Selected Answer: D**

Option D, reservation can't be applied to resources lower than projects (only to Org, folders or projects)

upvoted 3 times

**✉️** **AllenChen123** 1 week, 2 days agoSeems correct. [https://cloud.google.com/bigquery/docs/reservations-intro#understand\\_workload\\_management](https://cloud.google.com/bigquery/docs/reservations-intro#understand_workload_management)

upvoted 1 times

**✉️** **task\_7** 2 weeks, 4 days ago**Selected Answer: B**

Reserve assignments

To use the slot capacity you purchased, assign projects, folders, or organizations to a reservation. When a job in a project runs, it uses slots from the assigned reservation. Resources can inherit roles from their parents in the resource hierarchy. Even if a project is not assigned to a reservation, it inherits the assignment from the parent folder or organization, if any. If a project does not have an assigned or inherited reservation the job uses on-demand pricing. For more information about the resource hierarchy, see [Organizing BigQuery Resources](#).

upvoted 3 times

**✉️** **GCP001** 3 weeks ago

D.

Reservation is on project, folder or organisation level.

upvoted 1 times

**✉️** **raaad** 3 weeks, 3 days ago**Selected Answer: D**

C or D ??

Option C (service account) allows you to target the reservation specifically to the CDC process or any other jobs run by that service account. It is particularly useful if you have multiple processes running in the project with different performance or cost requirements.

Option D (project) applies the reservation across all jobs in the project, which is a broader approach. If the CDC process is the primary or sole running in the project and you want all jobs to share the same reservation, then this option might be more straightforward.

upvoted 2 times

**✉️** **scaenrui** 3 weeks, 4 days ago**Selected Answer: D**

D. Create a BigQuery reservation for the project.

upvoted 1 times

You are designing a fault-tolerant architecture to store data in a regional BigQuery dataset. You need to ensure that your application is able to recover from a corruption event in your tables that occurred within the past seven days. You want to adopt managed services with the lowest RPO and most cost-effective solution. What should you do?

- A. Access historical data by using time travel in BigQuery.
- B. Export the data from BigQuery into a new table that excludes the corrupted data
- C. Create a BigQuery table snapshot on a daily basis.
- D. Migrate your data to multi-region BigQuery buckets.

**Correct Answer: C**

*Community vote distribution*

A (100%)

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A, raaad explanation is perfect  
upvoted 2 times

✉️  **raaad** 3 weeks, 3 days ago

**Selected Answer: A**

- Lowest RPO: Time travel offers point-in-time recovery for the past seven days by default, providing the shortest possible recovery point objective (RPO) among the given options. You can recover data to any state within that window.
- No Additional Costs: Time travel is a built-in feature of BigQuery, incurring no extra storage or operational costs.
- Managed Service: BigQuery handles time travel automatically, eliminating manual backup and restore processes.

upvoted 4 times

✉️  **scaenruy** 3 weeks, 4 days ago

**Selected Answer: A**

A. Access historical data by using time travel in BigQuery.  
upvoted 1 times

You are building a streaming Dataflow pipeline that ingests noise level data from hundreds of sensors placed near construction sites across a city. The sensors measure noise level every ten seconds, and send that data to the pipeline when levels reach above 70 dBA. You need to detect the average noise level from a sensor when data is received for a duration of more than 30 minutes, but the window ends when no data has been received for 15 minutes. What should you do?

- A. Use session windows with a 15-minute gap duration.
- B. Use session windows with a 30-minute gap duration.
- C. Use hopping windows with a 15-minute window, and a thirty-minute period.
- D. Use tumbling windows with a 15-minute window and a fifteen-minute .withAllowedLateness operator.

**Correct Answer: D**

### Community vote distribution

C (55%)

D (27%)

A (18%)

✉️  **imiu** 4 days, 17 hours ago

**Selected Answer: C**

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#hopping-windows> To take running averages of data, use hopping windows. You can use one-minute hopping windows with a thirty-second period to compute a one-minute running average every thirty second

upvoted 1 times

✉️  **datapassionate** 1 week, 3 days ago

**Selected Answer: A**

to detect average noise levels from sensors, the best approach is to use session windows with a 15-minute gap duration (Option A). Session windows are ideal for cases like this where the events (sensor data) are sporadic. They group events that occur within a certain time interval (15 minutes in your case) and a new window is started if no data is received for the duration of the gap. This matches your requirement to end the window when no data is received for 15 minutes, ensuring that the average noise level is calculated over periods of continuous data

upvoted 2 times

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D to me, It aligns with the specified criteria for detecting the average noise level within a 30-minute duration and handling the end of the window when no data is received for 15 minutes.

upvoted 2 times

✉️  **AllenChen123** 5 days, 19 hours ago

Agree D.

Data comes -> 30 mts duration.

Data didn't come in 15 mts -> 15 mts duration

upvoted 1 times

✉️  **BIGQUERY\_ALT\_ALT** 2 weeks, 5 days ago

**Selected Answer: D**

OPTION D is correct for the specific scenario where we want to detect the average noise level for a duration of more than 30 minutes but end window when no data has been received for 15 minutes.

Explanation:

- Tumbling windows are non-overlapping windows, and in this case, you want to capture data continuously for 30-minute intervals.
- Using a tumbling window with a 15-minute window size aligns with your requirement to detect the average noise level for a duration of more than 30 minutes.
- Adding a .withAllowedLateness operator with a duration of fifteen minutes ensures that the window will still consider late-arriving data within that time frame. After fifteen minutes of no data, the window will be closed, and any late-arriving data will not be considered.

Option A and B invalid as they capture fixed logic with 15 or 30 mins. Option C captures only 15 min average with 30 min trigger hence not suitable.

upvoted 1 times

✉️  **Sofia98** 2 weeks, 6 days ago

**Selected Answer: C**

Hopping windows (called sliding windows in Apache Beam).

To take running averages of data, use hopping windows.

upvoted 2 times

✉️  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: C**

C. Use hopping windows with a 15-minute window, and a thirty-minute period.

upvoted 3 times

You are creating a data model in BigQuery that will hold retail transaction data. Your two largest tables, `sales_transaction_header` and `sales_transaction_line`, have a tightly coupled immutable relationship. These tables are rarely modified after load and are frequently joined when queried. You need to model the `sales_transaction_header` and `sales_transaction_line` tables to improve the performance of data analytics queries. What should you do?

- A. Create a `sales_transaction` table that holds the `sales_transaction_header` information as rows and the `sales_transaction_line` rows as nested and repeated fields.
- B. Create a `sales_transaction` table that holds the `sales_transaction_header` and `sales_transaction_line` information as rows, duplicating the `sales_transaction_header` data for each line.
- C. Create a `sales_transaction` table that stores the `sales_transaction_header` and `sales_transaction_line` data as a JSON data type.
- D. Create separate `sales_transaction_header` and `sales_transaction_line` tables and, when querying, specify the `sales_transaction_line` first in the WHERE clause.

**Correct Answer: C**

*Community vote distribution*

A (100%)

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A

upvoted 1 times

✉️  **raaad** 3 weeks, 3 days ago

**Selected Answer: A**

- In BigQuery, nested and repeated fields can significantly improve performance for certain types of queries, especially joins, because the data is co-located and can be read efficiently. - - This approach is often used in data warehousing scenarios where query performance is a priority, as the data relationships are immutable and rarely modified.

upvoted 3 times

✉️  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: A**

A. Create a `sales_transaction` table that holds the `sales_transaction_header` information as rows and the `sales_transaction_line` rows as nested and repeated fields.

upvoted 1 times

You created a new version of a Dataflow streaming data ingestion pipeline that reads from Pub/Sub and writes to BigQuery. The previous version of the pipeline that runs in production uses a 5-minute window for processing. You need to deploy the new version of the pipeline without losing any data, creating inconsistencies, or increasing the processing latency by more than 10 minutes. What should you do?

- A. Update the old pipeline with the new pipeline code.
- B. Snapshot the old pipeline, stop the old pipeline, and then start the new pipeline from the snapshot.
- C. Drain the old pipeline, then start the new pipeline.
- D. Cancel the old pipeline, then start the new pipeline.

**Correct Answer: D**

*Community vote distribution*

C (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C, draining the old pipeline solves all requests  
upvoted 1 times

 **raaad** 3 weeks, 3 days ago

**Selected Answer: C**

- Graceful Data Transition: Draining the old pipeline ensures it processes all existing data in its buffers and watermarks before shutting down, preventing data loss or inconsistencies.  
- Minimal Latency Increase: The latency increase will be limited to the amount of time it takes to drain the old pipeline, typically within the acceptable 10-minute threshold.

upvoted 4 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

C. Drain the old pipeline, then start the new pipeline.  
upvoted 2 times

Your organization's data assets are stored in BigQuery, Pub/Sub, and a PostgreSQL instance running on Compute Engine. Because there are multiple domains and diverse teams using the data, teams in your organization are unable to discover existing data assets. You need to design a solution to improve data discoverability while keeping development and configuration efforts to a minimum. What should you do?

- A. Use Data Catalog to automatically catalog BigQuery datasets. Use Data Catalog APIs to manually catalog Pub/Sub topics and PostgreSQL tables.
- B. Use Data Catalog to automatically catalog BigQuery datasets and Pub/Sub topics. Use Data Catalog APIs to manually catalog PostgreSQL tables.
- C. Use Data Catalog to automatically catalog BigQuery datasets and Pub/Sub topics. Use custom connectors to manually catalog PostgreSQL tables.
- D. Use customer connectors to manually catalog BigQuery datasets, Pub/Sub topics, and PostgreSQL tables.

**Correct Answer: C**

Community vote distribution

B (75%)

C (25%)

✉️  **datapassionate** 1 day ago

**Selected Answer: C**

Data Catalog is the best choice. But for cataloguing PostgreSQL it is better to use a connector when available, instead of using API. [https://cloud.google.com/data-catalog/docs/integrate-data-sources#integrate\\_unsupported\\_data\\_sources](https://cloud.google.com/data-catalog/docs/integrate-data-sources#integrate_unsupported_data_sources)

upvoted 1 times

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B - Data Catalog automatically maps out GCP resources and dev efforts are minimized by leveraging the data catalog API to do the same for postgresql db

upvoted 1 times

✉️  **GCP001** 3 weeks ago

B.  
-- Looks much better option as needed low development efforts.  
-- C not looking right as it will need lot of dev efforts for custom connectors.

upvoted 1 times

✉️  **raaad** 3 weeks, 3 days ago

**Selected Answer: B**

- It utilizes Data Catalog's native support for both BigQuery datasets and Pub/Sub topics.  
- For PostgreSQL tables running on a Compute Engine instance, you'd use Data Catalog APIs to create custom entries, as Data Catalog does not automatically discover external databases like PostgreSQL.

upvoted 2 times

✉️  **AllenChen123** 1 week, 2 days ago

Agree. <https://cloud.google.com/data-catalog/docs/concepts/overview#catalog-non-google-cloud-assets>

upvoted 1 times

You need to create a SQL pipeline. The pipeline runs an aggregate SQL transformation on a BigQuery table every two hours and appends the result to another existing BigQuery table. You need to configure the pipeline to retry if errors occur. You want the pipeline to send an email notification after three consecutive failures. What should you do?

- A. Use the BigQueryUpsertTableOperator in Cloud Composer, set the retry parameter to three, and set the email\_on\_failure parameter to true.
- B. Use the BigQueryInsertJobOperator in Cloud Composer, set the retry parameter to three, and set the email\_on\_failure parameter to true.
- C. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable email notifications.
- D. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable notification to Pub/Sub topic. Use Pub/Sub and Cloud Functions to send an email after three failed executions.

**Correct Answer: D**

*Community vote distribution*

B (67%)

D (33%)

 **datapassionate** 1 day ago

**Selected Answer: D**

D. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable notification to Pub/Sub topic. Use Pub/Sub and Cloud Functions to send an email after three failed executions

This method utilizes BigQuery's native scheduling capabilities for running the SQL job and leverages Pub/Sub and Cloud Functions for customized notification handling, including the specific requirement of sending an email after three consecutive failures.

upvoted 1 times

 **raaad** 3 weeks, 3 days ago

**Selected Answer: B**

- It provides a direct and controlled way to manage the SQL pipeline using Cloud Composer (Apache Airflow).
- The BigQueryInsertJobOperator is well-suited for running SQL jobs in BigQuery, including aggregate transformations and handling of results.
- The retry and email\_on\_failure parameters align with the requirements for error handling and notifications.
- Cloud Composer requires more setup than using BigQuery's scheduled queries directly, but it offers robust workflow management, retry log and notification capabilities, making it suitable for more complex and controlled data pipeline requirements.

upvoted 1 times

 **scaenrui** 3 weeks, 5 days ago

**Selected Answer: B**

B. Use the BigQueryInsertJobOperator in Cloud Composer, set the retry parameter to three, and set the email\_on\_failure parameter to true.

upvoted 1 times

You are monitoring your organization's data lake hosted on BigQuery. The ingestion pipelines read data from Pub/Sub and write the data into tables on BigQuery. After a new version of the ingestion pipelines is deployed, the daily stored data increased by 50%. The volumes of data in Pub/Sub remained the same and only some tables had their daily partition data size doubled. You need to investigate and fix the cause of the data increase. What should you do?

- A. 1. Check for duplicate rows in the BigQuery tables that have the daily partition data size doubled.
2. Schedule daily SQL jobs to deduplicate the affected tables.
3. Share the deduplication script with the other operational teams to reuse if this occurs to other tables.

- B. 1. Check for code errors in the deployed pipelines.
  - 2. Check for multiple writing to pipeline BigQuery sink.
  - 3. Check for errors in Cloud Logging during the day of the release of the new pipelines.
  - 4. If no errors, restore the BigQuery tables to their content before the last release by using time travel.
- C. 1. Check for duplicate rows in the BigQuery tables that have the daily partition data size doubled.
  - 2. Check the BigQuery Audit logs to find job IDs.
  - 3. Use Cloud Monitoring to determine when the identified Dataflow jobs started and the pipeline code version.
  - 4. When more than one pipeline ingests data into a table, stop all versions except the latest one.
- D. 1. Roll back the last deployment.
  - 2. Restore the BigQuery tables to their content before the last release by using time travel.
  - 3. Restart the Dataflow jobs and replay the messages by seeking the subscription to the timestamp of the release.

**Correct Answer: C**

*Community vote distribution*

C (88%)

13%

✉️  **raaad** Highly Voted 3 weeks, 3 days ago

**Selected Answer: C**

- Detailed Investigation of Logs and Jobs Checking for duplicate rows targets the potential immediate cause of the issue.
- Checking the BigQuery Audit logs helps identify which jobs might be contributing to the increased data volume.
- Using Cloud Monitoring to correlate job starts with pipeline versions helps identify if a specific version of the pipeline is responsible.
- Managing multiple versions of pipelines ensures that only the intended version is active, addressing any versioning errors that might have occurred during deployment.

=====

Why not B

While it addresses the symptom (excess data), it doesn't necessarily stop the problem from recurring. (The questions asked to investigate and upvoted 6 times)

✉️  **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: C**

Option C - agree with Raaad on the reasons

upvoted 1 times

✉️  **task\_7** 2 weeks, 4 days ago

**Selected Answer: B**

B. Check for code errors in the deployed pipelines, multiple writing to pipeline BigQuery sink, errors in Cloud Logging, and if necessary, restore tables using time travel.

Check for code errors

Check for multiple writes

Check Cloud Logging

Restore tables if necessary:

upvoted 1 times

You have a BigQuery dataset named "customers". All tables will be tagged by using a Data Catalog tag template named "gdpr". The template contains one mandatory field, "has\_sensitive\_data", with a boolean value. All employees must be able to do a simple search and find tables in the dataset that have either true or false in the "has\_sensitive\_data" field. However, only the Human Resources (HR) group should be able to see the data inside the tables for which "has\_sensitive data" is true. You give the all employees group the `bigrquery.metadataViewer` and `bigrquery.connectionUser` roles on the dataset. You want to minimize configuration overhead. What should you do next?

- A. Create the "gdpr" tag template with private visibility. Assign the `bigrquery.dataViewer` role to the HR group on the tables that contain sensitive data.
- B. Create the "gdpr" tag template with private visibility. Assign the `datacatalog.tagTemplateViewer` role on this tag to the all employees group, and assign the `bigrquery.dataViewer` role to the HR group on the tables that contain sensitive data.
- C. Create the "gdpr" tag template with public visibility. Assign the `bigrquery.dataViewer` role to the HR group on the tables that contain sensitive data.
- D. Create the "gdpr" tag template with public visibility. Assign the `datacatalog.tagTemplateViewer` role on this tag to the all employees group, and assign the `bigrquery.dataViewer` role to the HR group on the tables that contain sensitive data.

**Correct Answer: D**

*Community vote distribution*

C (83%)

D (17%)

✉️  **raaad**  2 weeks, 6 days ago

**Selected Answer: C**

- The most straightforward solution with minimal configuration overhead.
- By creating the "gdpr" tag template with public visibility, you ensure that all employees can search and find tables based on the "has\_sensitive\_data" field.
- Assigning the `bigrquery.dataViewer` role to the HR group on tables with sensitive data ensures that only they can view the actual data in these tables.

upvoted 5 times

✉️  **scaenrui**  3 weeks, 5 days ago

**Selected Answer: D**

- D. Create the "gdpr" tag template with public visibility. Assign the `datacatalog.tagTemplateViewer` role on this tag to the all employees group, assign the `bigrquery.dataViewer` role to the HR group on the tables that contain sensitive data.

upvoted 1 times

You are creating the CI/CD cycle for the code of the directed acyclic graphs (DAGs) running in Cloud Composer. Your team has two Cloud Composer instances: one instance for development and another instance for production. Your team is using a Git repository to maintain and develop the code of the DAGs. You want to deploy the DAGs automatically to Cloud Composer when a certain tag is pushed to the Git repository. What should you do?

- A. 1. Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.

2. If the tests pass, use Cloud Build to copy the code to the bucket of the production instance.
- B. 1. Use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the code to the Google Kubernetes Engine (GKE) cluster of the development instance for testing.
  2. If the tests pass, use the KubernetesPodOperator to deploy the container to the GKE cluster of the production instance.
- C. 1. Use Cloud Build to build a container and the KubernetesPodOperator to deploy the code of the DAG to the Google Kubernetes Engine (GKE) cluster of the development instance for testing.
  2. If the tests pass, copy the code to the Cloud Storage bucket of the production instance.
- D. 1. Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.
  2. If the tests pass, use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the container to the Google Kubernetes Engine (GKE) cluster of the production instance.

**Correct Answer: C**

*Community vote distribution*

A (100%)

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A, DAGs are routinely stored in cloud storage buckets, Cloud Build act as a trigger for both the deployment process to test env and the test itself

<https://cloud.google.com/composer/docs/dag-cicd-integration-guide>

upvoted 2 times

✉️  **BIGQUERY\_ALT\_ALT** 2 weeks, 4 days ago

**Selected Answer: A**

The Answer is A. Given that there are two instances (development and production) already available, and the goal is to deploy DAGs to Cloud Composer not entire composer infra build.

Explanation:

- This approach leverages Cloud Build to manage the deployment process.
- It first deploys the code to the Cloud Storage bucket of the development instance for testing purposes.
- If the tests are successful in the development environment, the same Cloud Build process is used to copy the code to the Cloud Storage bucket of the production instance.

B. GKE-based approach is not standard for Cloud Composer. C. GKE used for testing is unconventional for DAG deployments. D. Involves unnecessary GKE deployment for production. Testing DAGs should use Composer instances directly, not Kubernetes containers in GKE.

upvoted 3 times

✉️  **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

I vote for A

upvoted 1 times

✉️  **GCP001** 3 weeks, 1 day ago

C.

It looks the correct choice, first build, test and verify everything on dev environment and then just copy the files on prod bucket.

<https://cloud.google.com/composer/docs/dag-cicd-integration-guide>

upvoted 1 times

Question #274

Topic 1

You have a BigQuery table that ingests data directly from a Pub/Sub subscription. The ingested data is encrypted with a Google-managed encryption key. You need to meet a new organization policy that requires you to use keys from a centralized Cloud Key Management Service (Cloud KMS) project to encrypt data at rest. What should you do?

- Use Cloud KMS encryption key with Dataflow to ingest the existing Pub/Sub subscription to the existing BigQuery table.
- Create a new BigQuery table by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.
- Create a new Pub/Sub topic with CMEK and use the existing BigQuery table by using Google-managed encryption key.
- Create a new BigQuery table and Pub/Sub topic by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.

**Correct Answer: D**

*Community vote distribution*

B (67%)

D (33%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D - I get the discussion about B and D, but also pub/sub has some data at rest, e.g. messages with retention period. To comply with the organisation policy, we need to adapt also pub/sub

upvoted 1 times

✉  **raaad** 2 weeks, 5 days ago

**Selected Answer: B**

- New BigQuery Table with CMEK: This option involves creating a new BigQuery table configured to use a CMEK from Cloud KMS. It directly addresses the need to use a CMEK for data at rest in BigQuery.  
- Migrate Data: Migrating data from the old table (encrypted with a Google-managed key) to the new table (encrypted with CMEK) ensures the existing data complies with the new policy.

upvoted 4 times

✉  **Matt\_108** 2 weeks, 2 days ago

But also pub/sub has some data at rest, e.g. messages with retention period.

To comply with the organisation policy, we need to adapt also pub/sub

upvoted 1 times

✉  **AllenChen123** 1 week, 2 days ago

No, "The ingested data is encrypted with a Google-managed encryption key", target is ingested data in BigQuery.

upvoted 1 times

✉  **GCP001** 3 weeks, 1 day ago

D.

We should use new CMSK for both pubsub topic and BQ tables along with migrating old data.

upvoted 2 times

✉  **Smakyel79** 3 weeks, 1 day ago

**Selected Answer: D**

This option ensures that both the ingestion mechanism (Pub/Sub) and the storage component (BigQuery) are aligned with the organization's policy of using CMEK, providing end-to-end encryption control.

upvoted 1 times

✉  **BIGQUERY\_ALT\_ALT** 2 weeks, 4 days ago

Requirement is encrypt bq data - " The ingested data is encrypted with a Google-managed encryption key" so pubsub encryption from ingestion is not needed. Option B is correct.

upvoted 1 times

✉  **raaad** 2 weeks, 6 days ago

Why not B??

upvoted 1 times

✉  **raaad** 2 weeks, 6 days ago

Configuring a Pub/Sub topic with a CMEK is not necessary for encrypting data at rest in BigQuery.

upvoted 1 times

✉  **Matt\_108** 2 weeks, 2 days ago

to me it's D because also pub/sub has some data at rest, e.g. messages with retention period.

To comply with the organisation policy, we need to adapt also pub/sub encryption

upvoted 1 times

✉  **KirkD** 3 weeks ago

I considered also A as they are asking about encryption at rest. The BigQuery is the one but Pub/Sub, not sure.

upvoted 1 times

✉  **raaad** 2 weeks, 6 days ago

I think option A is a partial solution as using Cloud KMS key in Dataflow for ingestion does not change the encryption of the data at rest the BigQuery table.

upvoted 1 times



You created an analytics environment on Google Cloud so that your data scientist team can explore data without impacting the on-premises Apache Hadoop solution. The data in the on-premises Hadoop Distributed File System (HDFS) cluster is in Optimized Row Columnar (ORC) formatted files with multiple columns of Hive partitioning. The data scientist team needs to be able to explore the data in a similar way as they used the on-premises HDFS cluster with SQL on the Hive query engine. You need to choose the most cost-effective storage and processing solution. What should you do?

- A. Import the ORC files to Bigtable tables for the data scientist team.
- B. Import the ORC files to BigQuery tables for the data scientist team.
- C. Copy the ORC files on Cloud Storage, then deploy a Dataproc cluster for the data scientist team.
- D. Copy the ORC files on Cloud Storage, then create external BigQuery tables for the data scientist team.

**Correct Answer: C**

*Community vote distribution*

D (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D - leverages BigQuery for SQL-based exploration on direct querying to cloud storage  
upvoted 1 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: D**

- It leverages the strengths of BigQuery for SQL-based exploration while avoiding additional costs and complexity associated with data transformation or migration.  
- The data remains in ORC format in Cloud Storage, and BigQuery's external tables feature allows direct querying of this data.  
upvoted 4 times

 **Smakye179** 3 weeks, 1 day ago

**Selected Answer: D**

This approach leverages BigQuery's powerful analytics capabilities without the overhead of data transformation or maintaining a separate cluster while also allowing your team to use SQL for data exploration, similar to their experience with the on-premises Hadoop/Hive environment.  
upvoted 2 times

You are designing a Dataflow pipeline for a batch processing job. You want to mitigate multiple zonal failures at job submission time. What should you do?

- A. Submit duplicate pipelines in two different zones by using the --zone flag.
- B. Set the pipeline staging location as a regional Cloud Storage bucket.
- C. Specify a worker region by using the --region flag.
- D. Create an Eventarc trigger to resubmit the job in case of zonal failure when submitting the job.

**Correct Answer: A**

*Community vote distribution*

C (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C: <https://cloud.google.com/dataflow/docs/guides/pipeline-workflows#zonal-failures>  
upvoted 4 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: C**

- Specifying a worker region (instead of a specific zone) allows Google Cloud's Dataflow service to manage the distribution of resources across multiple zones within that region  
upvoted 4 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: C**

<https://cloud.google.com/dataflow/docs/guides/pipeline-workflows#zonal-failures>  
upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

C. Specify a worker region by using the --region flag.  
upvoted 2 times

You are designing a real-time system for a ride hailing app that identifies areas with high demand for rides to effectively reroute available drivers to meet the demand. The system ingests data from multiple sources to Pub/Sub, processes the data, and stores the results for visualization and analysis in real-time dashboards. The data sources include driver location updates every 5 seconds and app-based booking events from riders. The data processing involves real-time aggregation of supply and demand data for the last 30 seconds, every 2 seconds, and storing the results in a low-latency system for visualization. What should you do?

- A. Group the data by using a tumbling window in a Dataflow pipeline, and write the aggregated data to Memorystore.
- B. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to Memorystore.
- C. Group the data by using a session window in a Dataflow pipeline, and write the aggregated data to BigQuery.
- D. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to BigQuery.

**Correct Answer: A**

*Community vote distribution*

B (100%)

 **raaad** 2 weeks, 6 days ago

**Selected Answer: B**

- Hopping Window: Hopping windows are fixed-sized, overlapping intervals.
- Aggregate data over the last 30 seconds, every 2 seconds, as hopping windows allow for overlapping data analysis.
- Memorystore: Ideal for low-latency access required for real-time visualization and analysis.

upvoted 4 times

 **scaenrui** 3 weeks, 5 days ago

**Selected Answer: B**

- B. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to Memorystore.

upvoted 2 times

Your car factory is pushing machine measurements as messages into a Pub/Sub topic in your Google Cloud project. A Dataflow streaming job, that you wrote with the Apache Beam SDK, reads these messages, sends acknowledgment to Pub/Sub, applies some custom business logic in a DoFn instance, and writes the result to BigQuery. You want to ensure that if your business logic fails on a message, the message will be sent to a Pub/Sub topic that you want to monitor for alerting purposes. What should you do?

- A. Enable retaining of acknowledged messages in your Pub/Sub pull subscription. Use Cloud Monitoring to monitor the subscription/num\_retained\_acked\_messages metric on this subscription.
- B. Use an exception handling block in your Dataflow's DoFn code to push the messages that failed to be transformed through a side output and to a new Pub/Sub topic. Use Cloud Monitoring to monitor the topic/num\_unacked\_messages\_by\_region metric on this new topic.
- C. Enable dead lettering in your Pub/Sub pull subscription, and specify a new Pub/Sub topic as the dead letter topic. Use Cloud Monitoring to monitor the subscription/dead\_letter\_message\_count metric on your pull subscription.
- D. Create a snapshot of your Pub/Sub pull subscription. Use Cloud Monitoring to monitor the snapshot/num\_messages metric on this snapshot.

**Correct Answer: D***Community vote distribution*

B (100%)

**raaad** Highly Voted 2 weeks, 6 days ago**Selected Answer: B**

- Exception Handling in DoFn: Implementing an exception handling block within DoFn in Dataflow to catch failures during processing is a direct way to manage errors.
  - Side Output to New Topic: Using a side output to redirect failed messages to a new Pub/Sub topic is an effective way to isolate and manage these messages.
  - Monitoring: Monitoring the num\_unacked\_messages\_by\_region on the new topic can alert you to the presence of failed messages.
- upvoted 5 times

**Matt\_108** Most Recent 2 weeks, 2 days ago**Selected Answer: B**

Option B - Raaad explanation is complete  
upvoted 1 times

**scaenruy** 3 weeks, 5 days ago**Selected Answer: B**

- B. Use an exception handling block in your Dataflow's DoFn code to push the messages that failed to be transformed through a side output to a new Pub/Sub topic. Use Cloud Monitoring to monitor the topic/num\_unacked\_messages\_by\_region metric on this new topic.
- upvoted 1 times

You want to store your team's shared tables in a single dataset to make data easily accessible to various analysts. You want to make this data readable but unmodifiable by analysts. At the same time, you want to provide the analysts with individual workspaces in the same project, where they can create and store tables for their own use, without the tables being accessible by other analysts. What should you do?

- A. Give analysts the BigQuery Data Viewer role at the project level. Create one other dataset, and give the analysts the BigQuery Data Editor role on that dataset.
- B. Give analysts the BigQuery Data Viewer role at the project level. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the project level.

C. Give analysts the BigQuery Data Viewer role on the shared dataset. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the dataset level for their assigned dataset.

D. Give analysts the BigQuery Data Viewer role on the shared dataset. Create one other dataset and give the analysts the BigQuery Data Editor role on that dataset.

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C

upvoted 1 times

✉️  **raaad** 2 weeks, 6 days ago

**Selected Answer: C**

- Data Viewer on Shared Dataset: Grants read-only access to the shared dataset.

- Data Editor on Individual Datasets: Giving each analyst Data Editor role on their respective dataset creates private workspaces where they can create and store personal tables without exposing them to other analysts.

upvoted 4 times

✉️  **Sofia98** 2 weeks, 6 days ago

**Selected Answer: C**

option C, because analysts can not see the individual datasets of other analysts

upvoted 1 times

✉️  **scaenruy** 3 weeks, 5 days ago

**Selected Answer: C**

C. Give analysts the BigQuery Data Viewer role on the shared dataset. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the dataset level for their assigned dataset.

upvoted 2 times

You are running a streaming pipeline with Dataflow and are using hopping windows to group the data as the data arrives. You noticed that some data is arriving late but is not being marked as late data, which is resulting in inaccurate aggregations downstream. You need to find a solution that allows you to capture the late data in the appropriate window. What should you do?

- A. Use watermarks to define the expected data arrival window. Allow late data as it arrives.
- B. Change your windowing function to tumbling windows to avoid overlapping window periods.
- C. Change your windowing function to session windows to define your windows based on certain activity.
- D. Expand your hopping window so that the late data has more time to arrive within the grouping.

**Correct Answer: D**

*Community vote distribution*

A (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A - <https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#watermarks>  
upvoted 1 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: A**

- Watermarks: Watermarks in a streaming pipeline are used to specify the point in time when Dataflow expects all data up to that point to have arrived.  
- Allow Late Data: configure the pipeline to accept and correctly process data that arrives after the watermark, ensuring it's captured in the appropriate window.  
upvoted 2 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#watermarks>  
upvoted 1 times

 **scaenruy** 3 weeks, 5 days ago

**Selected Answer: A**

A. Use watermarks to define the expected data arrival window. Allow late data as it arrives.  
upvoted 1 times

You work for a large ecommerce company. You store your customer's order data in Bigtable. You have a garbage collection policy set to delete the data after 30 days and the number of versions is set to 1. When the data analysts run a query to report total customer spending, the analysts sometimes see customer data that is older than 30 days. You need to ensure that the analysts do not see customer data older than 30 days while minimizing cost and overhead. What should you do?

- A. Set the expiring values of the column families to 29 days and keep the number of versions to 1.
- B. Use a timestamp range filter in the query to fetch the customer's data for a specific range.
- C. Schedule a job daily to scan the data in the table and delete data older than 30 days.
- D. Set the expiring values of the column families to 30 days and set the number of versions to 2.

**Correct Answer: D**

*Community vote distribution*

B (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Agree with others <https://cloud.google.com/bigtable/docs/garbage-collection>

upvoted 2 times

 **AllenChen123** 5 days, 3 hours ago

Agree. <https://cloud.google.com/bigtable/docs/garbage-collection#data-removed>

"Because it can take up to a week for expired data to be deleted, you should never rely solely on garbage collection policies to ensure that read requests return the desired data. Always apply a filter to your read requests that excludes the same values as your garbage collection rules. You can filter by limiting the number of cells per column or by specifying a timestamp range."

upvoted 1 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: B**

I will go for B too

upvoted 1 times

 **GCP001** 3 weeks, 1 day ago

B. Use a timestamp range filter in the query to fetch the customer's data for a specific range.

Always use query filter as garbage collectore runs on it's way - <https://cloud.google.com/bigtable/docs/garbage-collection>

upvoted 1 times

 **scaenrui** 3 weeks, 5 days ago

**Selected Answer: B**

B. Use a timestamp range filter in the query to fetch the customer's data for a specific range.

upvoted 1 times

You are using a Dataflow streaming job to read messages from a message bus that does not support exactly-once delivery. Your job then applies some transformations, and loads the result into BigQuery. You want to ensure that your data is being streamed into BigQuery with exactly-once delivery semantics. You expect your ingestion throughput into BigQuery to be about 1.5 GB per second. What should you do?

- A. Use the BigQuery Storage Write API and ensure that your target BigQuery table is regional.
- B. Use the BigQuery Storage Write API and ensure that your target BigQuery table is multiregional.
- C. Use the BigQuery Streaming API and ensure that your target BigQuery table is regional.
- D. Use the BigQuery Streaming API and ensure that your target BigQuery table is multiregional.

**Correct Answer: D**

*Community vote distribution*

A (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A

upvoted 1 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: A**

- BigQuery Storage Write API: This API is designed for high-throughput, low-latency writing of data into BigQuery. It also provides tools to prevent data duplication, which is essential for exactly-once delivery semantics.

- Regional Table: Choosing a regional location for the BigQuery table could potentially provide better performance and lower latency, as it would be closer to the Dataflow job if they are in the same region.

upvoted 3 times

 **AllenChen123** 5 days, 3 hours ago

Agree.

<https://cloud.google.com/bigquery/docs/write-api#advantages>

upvoted 2 times

 **Ed\_Kim** 3 weeks, 6 days ago

**Selected Answer: A**

Voting on A

upvoted 2 times

 **Smakyel79** 3 weeks, 1 day ago

This option leverages the BigQuery Storage Write API's capability for exactly-once delivery semantics and a regional table setting that can meet compliance and data locality needs without impacting the delivery semantics. The BigQuery Storage Write API is more suitable for your high-throughput requirements compared to the BigQuery Streaming API.

upvoted 2 times

You have created an external table for Apache Hive partitioned data that resides in a Cloud Storage bucket, which contains a large number of files. You notice that queries against this table are slow. You want to improve the performance of these queries. What should you do?

- A. Change the storage class of the Hive partitioned data objects from Coldline to Standard.
- B. Create an individual external table for each Hive partition by using a common table name prefix. Use wildcard table queries to reference the partitioned data.

C. Upgrade the external table to a BigLake table. Enable metadata caching for the table.

D. Migrate the Hive partitioned data objects to a multi-region Cloud Storage bucket.

**Correct Answer: C**

*Community vote distribution*

C (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C

upvoted 1 times

✉  **Sofia98** 2 weeks, 6 days ago

**Selected Answer: C**

agree with C

upvoted 1 times

✉  **raaad** 2 weeks, 6 days ago

**Selected Answer: C**

- BigLake Table: BigLake allows for more efficient querying of data lakes stored in Cloud Storage. It can handle large datasets more effectively than standard external tables.

- Metadata Caching: Enabling metadata caching can significantly improve query performance by reducing the time taken to read and process metadata from a large number of files.

upvoted 1 times

✉  **AllenChen123** 1 week, 2 days ago

Agree. [https://cloud.google.com/bigquery/docs/biglake-intro#metadata\\_caching\\_for\\_performance](https://cloud.google.com/bigquery/docs/biglake-intro#metadata_caching_for_performance)

upvoted 1 times

✉  **AllenChen123** 5 days, 3 hours ago

And <https://cloud.google.com/bigquery/docs/external-data-cloud-storage#upgrade-external-tables-to-biglake-tables>

upvoted 1 times

✉  **GCP001** 3 weeks, 1 day ago

C. Upgrade the external table to a BigLake table. Enable metadata caching for the table.

Check ref - <https://cloud.google.com/bigquery/docs/biglake-intro>

upvoted 1 times

You have a network of 1000 sensors. The sensors generate time series data: one metric per sensor per second, along with a timestamp. You already have 1 TB of data, and expect the data to grow by 1 GB every day. You need to access this data in two ways. The first access pattern requires retrieving the metric from one specific sensor stored at a specific timestamp, with a median single-digit millisecond latency. The second access pattern requires running complex analytic queries on the data, including joins, once a day. How should you store this data?

- A. Store your data in BigQuery. Concatenate the sensor ID and timestamp, and use it as the primary key.
- B. Store your data in Bigtable. Concatenate the sensor ID and timestamp and use it as the row key. Perform an export to BigQuery every day.
- C. Store your data in Bigtable. Concatenate the sensor ID and metric, and use it as the row key. Perform an export to BigQuery every day.
- D. Store your data in BigQuery. Use the metric as a primary key.

**Correct Answer: B**

*Community vote distribution*

B (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B - agree with raaad

upvoted 1 times

✉  **raaad** 2 weeks, 6 days ago

**Selected Answer: B**

- Bigtable excels at incredibly fast lookups by row key, often reaching single-digit millisecond latencies.
- Constructing the row key with sensor ID and timestamp enables efficient retrieval of specific sensor readings at exact timestamps.
- Bigtable's wide-column design effectively stores time series data, allowing for flexible addition of new metrics without schema changes.
- Bigtable scales horizontally to accommodate massive datasets (petabytes or more), easily handling the expected data growth.

upvoted 3 times

✉  **scaenrui** 3 weeks, 5 days ago

**Selected Answer: B**

- B. Store your data in Bigtable. Concatenate the sensor ID and timestamp and use it as the row key. Perform an export to BigQuery every day.

upvoted 2 times

✉  **Smakyel79** 3 weeks, 1 day ago

Based on your requirements, Option B seems most suitable. Bigtable's design caters to the low-latency access of time-series data (your first requirement), and the daily export to BigQuery enables complex analytics (your second requirement). The use of sensor ID and timestamp as the row key in Bigtable would facilitate efficient access to specific sensor data at specific times.

upvoted 1 times

You have 100 GB of data stored in a BigQuery table. This data is outdated and will only be accessed one or two times a year for analytics with SQL. For backup purposes, you want to store this data to be immutable for 3 years. You want to minimize storage costs. What should you do?

- A. 1. Create a BigQuery table clone.  
2. Query the clone when you need to perform analytics.
- B. 1. Create a BigQuery table snapshot.  
2. Restore the snapshot when you need to perform analytics.
- C. 1. Perform a BigQuery export to a Cloud Storage bucket with archive storage class.  
2. Enable versioning on the bucket.  
3. Create a BigQuery external table on the exported files.
- D. 1. Perform a BigQuery export to a Cloud Storage bucket with archive storage class.  
2. Set a locked retention policy on the bucket.  
3. Create a BigQuery external table on the exported files.

**Correct Answer: D***Community vote distribution*

D (100%)

**✉️**  **Matt\_108** 2 weeks, 2 days ago**Selected Answer: D**

Option D, clearly  
upvoted 1 times

**✉️**  **raaad** 2 weeks, 4 days ago**Selected Answer: D**

Straight Forward  
upvoted 4 times

**✉️**  **GCP001** 3 weeks, 1 day ago

D.  
For data keeping till last 3 years, use bucket lock with retention policy  
upvoted 1 times

You have thousands of Apache Spark jobs running in your on-premises Apache Hadoop cluster. You want to migrate the jobs to Google Cloud. You want to use managed services to run your jobs instead of maintaining a long-lived Hadoop cluster yourself. You have a tight timeline and want to keep code changes to a minimum. What should you do?

- A. Move your data to BigQuery. Convert your Spark scripts to a SQL-based processing approach.
- B. Rewrite your jobs in Apache Beam. Run your jobs in Dataflow.
- C. Copy your data to Compute Engine disks. Manage and run your jobs directly on those instances.
- D. Move your data to Cloud Storage. Run your jobs on Dataproc.

**Correct Answer: A**

*Community vote distribution*

D (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Clearly D

upvoted 2 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: D**

of course D

upvoted 2 times

 **GCP001** 3 weeks, 1 day ago

D. Move your data to Cloud Storage. Run your jobs on Dataproc.  
Dataproc is managed service and not needed much code changes.

upvoted 2 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: D**

D. Move your data to Cloud Storage. Run your jobs on Dataproc.

upvoted 2 times

You are administering shared BigQuery datasets that contain views used by multiple teams in your organization. The marketing team is concerned about the variability of their monthly BigQuery analytics spend using the on-demand billing model. You need to help the marketing team establish a consistent BigQuery analytics spend each month. What should you do?

- A. Create a BigQuery Enterprise reservation with a baseline of 250 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.
- B. Establish a BigQuery quota for the marketing team, and limit the maximum number of bytes scanned each day.
- C. Create a BigQuery reservation with a baseline of 500 slots with no autoscaling for the marketing team, and bill them back accordingly.
- D. Create a BigQuery Standard pay-as-you go reservation with a baseline of 0 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.

**Correct Answer: D**

Community vote distribution

C (67%)

B (33%)

✉️ **lipa31** 6 days, 12 hours ago

anybody for D ? <https://cloud.google.com/bigquery/docs/slots-autoscaling-intro>  
upvoted 2 times

✉️ **Sofia98** 2 weeks, 6 days ago

**Selected Answer: B**

<https://cloud.google.com/blog/products/data-analytics/manage-bigquery-costs-with-custom-quotas>  
upvoted 2 times

✉️ **raaad** 2 weeks, 6 days ago

**Selected Answer: C**

Reservations guarantee a fixed number of slots (computational resources) for BigQuery queries, ensuring a predictable monthly cost, address the marketing team's concern about variability.

upvoted 3 times

✉️ **AllenChen123** 1 week, 2 days ago

Why 500 slots?

upvoted 1 times

✉️ **AllenChen123** 5 days, 2 hours ago

But seems only C makes sense.

[https://cloud.google.com/bigquery/quotas#query\\_jobs](https://cloud.google.com/bigquery/quotas#query_jobs)

"There is no limit to the number of bytes that can be processed by queries in a project."

upvoted 1 times

✉️ **datapassionate** 21 hours, 27 minutes ago

"However, you can set limits on the amount of data users can query by creating custom quotas to control query usage per day or query usage per day per user."

<https://cloud.google.com/blog/products/data-analytics/manage-bigquery-costs-with-custom-quotas>

B would be correct

upvoted 1 times

✉️ **scaenruy** 3 weeks, 4 days ago

**Selected Answer: C**

C. Create a BigQuery reservation with a baseline of 500 slots with no autoscaling for the marketing team, and bill them back accordingly.

upvoted 1 times

✉️ **Sofia98** 2 weeks, 6 days ago

Provide, please, the reference

upvoted 1 times

You are part of a healthcare organization where data is organized and managed by respective data owners in various storage services. As a result of this decentralized ecosystem, discovering and managing data has become difficult. You need to quickly identify and implement a cost-optimized solution to assist your organization with the following:

- Data management and discovery
- Data lineage tracking
- Data quality validation

How should you build the solution?

- A. Use BigLake to convert the current solution into a data lake architecture.
- B. Build a new data discovery tool on Google Kubernetes Engine that helps with new source onboarding and data lineage tracking.
- C. Use BigQuery to track data lineage, and use Dataprep to manage data and perform data quality validation.
- D. Use Dataplex to manage data, track data lineage, and perform data quality validation.

**Correct Answer: A**

*Community vote distribution*

D (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Clearly D  
upvoted 2 times

 **Sofia98** 2 weeks, 6 days ago

**Selected Answer: D**

Agree with Dataplex option  
upvoted 2 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: D**

Straight forward  
upvoted 2 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: D**

D. Use Dataplex to manage data, track data lineage, and perform data quality validation.  
upvoted 1 times

You have data located in BigQuery that is used to generate reports for your company. You have noticed some weekly executive report fields do not correspond to format according to company standards. For example, report errors include different telephone formats and different country code identifiers. This is a frequent issue, so you need to create a recurring job to normalize the data. You want a quick solution that requires no coding. What should you do?

- A. Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.
- B. Use Dataflow SQL to create a job that normalizes the data, and that after the first run of the job, schedule the pipeline to execute recurrently.
- C. Create a Spark job and submit it to Dataproc Serverless.
- D. Use BigQuery and GoogleSQL to normalize the data, and schedule recurring queries in BigQuery.

**Correct Answer: B**

*Community vote distribution*

A (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Definitely A, cloud data fusion and wrangler to setup the clean up pipeline with no coding required  
upvoted 3 times

✉  **Sofia98** 2 weeks, 6 days ago

**Selected Answer: A**

Cloud Data Fusion and Wrangler  
upvoted 1 times

✉  **scaenrui** 3 weeks, 4 days ago

**Selected Answer: A**

A. Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.  
upvoted 1 times

You are designing a messaging system by using Pub/Sub to process clickstream data with an event-driven consumer app that relies on a push subscription. You need to configure the messaging system that is reliable enough to handle temporary downtime of the consumer app. You also need the messaging system to store the input messages that cannot be consumed by the subscriber. The system needs to retry failed messages gradually, avoiding overloading the consumer app, and store the failed messages after a maximum of 10 retries in a topic. How should you configure the Pub/Sub subscription?

- A. Increase the acknowledgement deadline to 10 minutes.
- B. Use immediate redelivery as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10.
- C. Use exponential backoff as the subscription retry policy, and configure dead lettering to the same source topic with maximum delivery attempts set to 10.
- D. Use exponential backoff as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10.

**Correct Answer: B**

*Community vote distribution*

D (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D - agree with other comments explanation  
upvoted 1 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: D**

- Exponential Backoff: This retry policy gradually increases the delay between retries, which helps to avoid overloading the consumer app.
- Dead Lettering to a Different Topic: Configuring dead lettering sends messages that couldn't be processed after the specified number of delivery attempts (10 in this case) to a separate topic. This allows for handling of failed messages without interrupting the regular flow of new messages.
- Maximum Delivery Attempts Set to 10: This setting ensures that the system retries each message up to 10 times before considering it a fail and moving it to the dead letter topic.

upvoted 2 times

 **GCP001** 3 weeks, 1 day ago

D. Use exponential backoff as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10

Best suitable options for graceful retry and storing failed messages  
upvoted 2 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: D**

D. Use exponential backoff as the subscription retry policy, and configure dead lettering to a different topic with maximum delivery attempts set to 10.  
upvoted 2 times

 **Smakyel79** 3 weeks, 1 day ago

Exponential backoff will help in managing the load on the consumer app by gradually increasing the delay between retries. Configuring dead lettering to a different topic after a maximum of 10 delivery attempts ensures that undeliverable messages are stored separately, preventing them from being retried endlessly and cluttering the main message flow.

upvoted 1 times

You designed a data warehouse in BigQuery to analyze sales data. You want a self-serving, low-maintenance, and cost- effective solution to share the sales dataset to other business units in your organization. What should you do?

- A. Create an Analytics Hub private exchange, and publish the sales dataset.
- B. Enable the other business units' projects to access the authorized views of the sales dataset.
- C. Create and share views with the users in the other business units.
- D. Use the BigQuery Data Transfer Service to create a schedule that copies the sales dataset to the other business units' projects.

**Correct Answer: A**

*Community vote distribution*

A (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Definitely A

upvoted 1 times

 **raaad** 2 weeks, 6 days ago

**Selected Answer: A**

Analytics Hub offers a centralized platform for managing data sharing and access within the organization. This simplifies access control management.

upvoted 3 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: A**

A. Create an Analytics Hub private exchange, and publish the sales dataset.

upvoted 1 times

You have terabytes of customer behavioral data streaming from Google Analytics into BigQuery daily. Your customers' information, such as their preferences, is hosted on a Cloud SQL for MySQL database. Your CRM database is hosted on a Cloud SQL for PostgreSQL instance. The marketing team wants to use your customers' information from the two databases and the customer behavioral data to create marketing campaigns for yearly active customers. You need to ensure that the marketing team can run the campaigns over 100 times a day on typical days and up to 300 during sales. At the same time, you want to keep the load on the Cloud SQL databases to a minimum. What should you do?

- A. Create BigQuery connections to both Cloud SQL databases. Use BigQuery federated queries on the two databases and the Google Analytics data on BigQuery to run these queries.
- B. Create a job on Apache Spark with Dataproc Serverless to query both Cloud SQL databases and the Google Analytics data on BigQuery for these queries.
- C. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.
- D. Create a Dataproc cluster with Trino to establish connections to both Cloud SQL databases and BigQuery, to execute the queries.

**Correct Answer: C**

*Community vote distribution*

C (100%)

 **raaad** 2 weeks, 4 days ago

**Selected Answer: C**

- Datastream: It's a fully managed, serverless service for real-time data replication. It allows to stream data from various sources, including Cloud SQL, into BigQuery.
- Reduced Load on Cloud SQL: By replicating the required tables from both Cloud SQL databases into BigQuery, you minimize the load on the Cloud SQL instances. The marketing team's queries will be run against BigQuery, which is designed to handle high-volume analytics workloads.
- Frequency of Queries: BigQuery can easily handle the high frequency of queries (100 times daily, up to 300 during sales events) due to its powerful data processing capabilities.
- Combining Data Sources: Once the data is in BigQuery, you can efficiently combine it with the Google Analytics data for comprehensive analysis and campaign planning.

upvoted 1 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: C**

- C. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.
- upvoted 2 times

 **Smakyel79** 3 weeks, 1 day ago

Datastream is a serverless, easy-to-use change data capture (CDC) and replication service. By replicating the necessary tables from the Cloud SQL databases to BigQuery, you can offload the query load from the Cloud SQL databases. The marketing team can then run their queries directly on BigQuery, which is designed for large-scale data analytics. This approach seems to balance both efficiency and performance, minimizing load on the Cloud SQL instances.

upvoted 1 times

Your organization is modernizing their IT services and migrating to Google Cloud. You need to organize the data that will be stored in Cloud Storage and BigQuery. You need to enable a data mesh approach to share the data between sales, product design, and marketing departments. What should you do?

- A. 1. Create a project for storage of the data for each of your departments.
- 2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
- 3. Create user groups for authorized readers for each bucket and dataset.

4. Enable the IT team to administer the user groups to add or remove users as the departments' request.
  - B. 1. Create multiple projects for storage of the data for each of your departments' applications.
  2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
  3. Publish the data that each department shared in Analytics Hub.
  4. Enable all departments to discover and subscribe to the data they need in Analytics Hub.
- C. 1. Create a project for storage of the data for your organization.
2. Create a central Cloud Storage bucket with three folders to store the files for each department.
3. Create a central BigQuery dataset with tables prefixed with the department name.
4. Give viewer rights for the storage project for the users of your departments.
- D. 1. Create multiple projects for storage of the data for each of your departments' applications.
2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
3. In Dataplex, map each department to a data lake and the Cloud Storage buckets, and map the BigQuery datasets to zones.
4. Enable each department to own and share the data of their data lakes.

**Correct Answer: B**

*Community vote distribution*

D (100%)

✉  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: D**

that's pure data mesh, which is what dataplex has been built for  
upvoted 1 times

✉  **raaad** 2 weeks, 4 days ago

**Selected Answer: D**

- Decentralized ownership: Each department controls its data lake, aligning with the core principle of data ownership in a data mesh.
- Self-service data access: Departments can create and manage their own Cloud Storage buckets and BigQuery datasets within their data lake, enabling self-service data access.
- Interdepartmental sharing: Dataplex facilitates data sharing by enabling departments to publish their data products from their data lakes, making it easily discoverable and usable by other departments.

upvoted 3 times

✉  **Sofia98** 2 weeks, 5 days ago

**Selected Answer: D**

For me, Dataplex looks more logical  
upvoted 1 times

✉  **GCP001** 3 weeks, 1 day ago

D. Dataplex looks more suitable for data mesh approach, Check the ref - <https://cloud.google.com/dataplex/docs/introduction>  
upvoted 1 times

You work for a large ecommerce company. You are using Pub/Sub to ingest the clickstream data to Google Cloud for analytics. You observe that when a new subscriber connects to an existing topic to analyze data, they are unable to subscribe to older data. For an upcoming yearly sale event in two months, you need a solution that, once implemented, will enable any new subscriber to read the last 30 days of data. What should you do?

- A. Create a new topic, and publish the last 30 days of data each time a new subscriber connects to an existing topic.
- B. Set the topic retention policy to 30 days.
- C. Set the subscriber retention policy to 30 days.
- D. Ask the source system to re-push the data to Pub/Sub, and subscribe to it.

**Correct Answer: C**

*Community vote distribution*

B (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Definitely B

upvoted 2 times

 **raaad** 2 weeks, 4 days ago

**Selected Answer: B**

- Topic Retention Policy: This policy determines how long messages are retained by Pub/Sub after they are published, even if they have not been acknowledged (consumed) by any subscriber.

- 30 Days Retention: By setting the retention policy of the topic to 30 days, all messages published to this topic will be available for consumption for 30 days. This means any new subscriber connecting to the topic can access and analyze data from the past 30 days.

upvoted 3 times

 **Sofia98** 2 weeks, 5 days ago

**Selected Answer: B**

<https://cloud.google.com/blog/products/data-analytics/pubsub-gains-topic-retention-feature>

upvoted 2 times

 **scaenrui** 3 weeks, 4 days ago

**Selected Answer: B**

B. Set the topic retention policy to 30 days.

upvoted 1 times

You are designing the architecture to process your data from Cloud Storage to BigQuery by using Dataflow. The network team provided you with the Shared VPC network and subnetwork to be used by your pipelines. You need to enable the deployment of the pipeline on the Shared VPC network. What should you do?

- A. Assign the compute.networkUser role to the Dataflow service agent.
- B. Assign the compute.networkUser role to the service account that executes the Dataflow pipeline.
- C. Assign the dataflow.admin role to the Dataflow service agent.
- D. Assign the dataflow.admin role to the service account that executes the Dataflow pipeline.

**Correct Answer: A***Community vote distribution*

A (70%)

B (30%)

  **Matt\_108** 2 weeks, 2 days ago**Selected Answer: A**

Option A, I do agree with Raaad, it's the dataflow service agent that needs the networkUser role, because it's the one that provisions the netw resources <https://cloud.google.com/dataflow/docs/guides/specifying-networks#shared>  
upvoted 3 times

  **task\_7** 2 weeks, 4 days ago**Selected Answer: B**

compute.networkUser to the service account that executes the Dataflow pipeline.  
upvoted 2 times

  **raaad** 2 weeks, 4 days ago**Selected Answer: A**

- Dataflow service agent is the one responsible for setting up and managing the network resources that Dataflow requires.  
- By granting the compute.networkUser role to this service agent, we are enabling it to provision the necessary network resources within the Shared VPC for your Dataflow job.  
upvoted 4 times

  **BIGQUERY\_ALT\_ALT** 2 weeks, 4 days ago**Selected Answer: B**

Option B is Correct.

Explanation:

You need to give compute networkuser role to service account that is processing the pipeline as it will need to deploy nessesary worker node of the shared vpc project.

Option A is incorrect as Dataflow Service Agent is Google MGS service account that will not responsible for running or deoplyling workers in shared vpc.

Option C and D is incorrect as dataflow.admin is elevated privlages to create and manage all of dataflow components not deploying resource: shared vpc.

upvoted 1 times

  **GCP001** 3 weeks, 1 day ago

B. Assign the compute.networkUser role to the service account that executes the Dataflow pipeline. See the ref - <https://cloud.google.com/dataflow/docs/guides/specifying-networks>

upvoted 1 times

  **raaad** 2 weeks, 4 days ago

Option A makes more sense:

- Assigning the compute.networkUser role to the pipeline's service account grants it unnecessary and possibly excessive permissions out: its core responsibility of data processing.

The question focused specifically on the deployment aspect (i.e., provisioning of network resources like VMs) rather than what the pipeline accesses or processes once it's running.

upvoted 1 times

  **GCP001** 1 week, 5 days ago

Yes , I agree, it should be A. Dataflow service account should be the one having this permission instaed of worker

upvoted 1 times

Your infrastructure team has set up an interconnect link between Google Cloud and the on-premises network. You are designing a high-throughput streaming pipeline to ingest data in streaming from an Apache Kafka cluster hosted on-premises. You want to store the data in BigQuery, with as minimal latency as possible. What should you do?

- A. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Use a Google-provided Dataflow template to read the data from Pub/Sub, and write the data to BigQuery.
- B. Use a proxy host in the VPC in Google Cloud connecting to Kafka. Write a Dataflow pipeline, read data from the proxy host, and write the data to BigQuery.
- C. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.
- D. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Write a Dataflow pipeline, read the data from Pub/Sub, and write the data to BigQuery.

**Correct Answer: B**

*Community vote distribution*

A (60%)

C (40%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A, leverage dataflow template for Kafka <https://cloud.google.com/dataflow/docs/kafka-dataflow>  
upvoted 3 times

 **AllenChen123** 1 week, 2 days ago

Agree. "Google provides a Dataflow template that configures a Kafka-to-BigQuery pipeline. The template uses the BigQueryIO connector provided in the Apache Beam SDK."  
upvoted 1 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: C**

C. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.  
upvoted 1 times

 **rahulvin** 1 month ago

**Selected Answer: C**

Dataflow has templates to read from Kafka. Other options are too complicated  
<https://cloud.google.com/dataflow/docs/kafka-dataflow>  
upvoted 1 times

 **Sofia98** 2 weeks, 5 days ago

so, this is the answer A, whe C?  
upvoted 2 times

 **Matt\_108** 2 weeks, 2 days ago

Yeah, the answer is A. C requires you to develop the pipeline yourself and ensure minimal latency, which means that you perform better than a pre-built template from Google...not really the case most of the times :)  
upvoted 1 times

 **saschak94** 18 hours, 21 minutes ago

but Option A introduces additional replication into Pub/Sub and the question states with minimal latency. In my opinion subscribing to Kafka via Dataflow has a lower latency than replicating the messages first to Pub/Sub and subscribing with Dataflow to it.  
upvoted 1 times

You migrated your on-premises Apache Hadoop Distributed File System (HDFS) data lake to Cloud Storage. The data scientist team needs to process the data by using Apache Spark and SQL. Security policies need to be enforced at the column level. You need a cost-effective solution that can scale into a data mesh. What should you do?

- A. 1. Deploy a long-living Dataproc cluster with Apache Hive and Ranger enabled.  
2. Configure Ranger for column level security.  
3. Process with Dataproc Spark or Hive SQL.
- B. 1. Define a BigLake table.  
2. Create a taxonomy of policy tags in Data Catalog.  
3. Add policy tags to columns.  
4. Process with the Spark-BigQuery connector or BigQuery SQL.
- C. 1. Load the data to BigQuery tables.  
2. Create a taxonomy of policy tags in Data Catalog.  
3. Add policy tags to columns.  
4. Process with the Spark-BigQuery connector or BigQuery SQL.
- D. 1. Apply an Identity and Access Management (IAM) policy at the file level in Cloud Storage.  
2. Define a BigQuery external table for SQL processing.  
3. Use Dataproc Spark to process the Cloud Storage files.

**Correct Answer: A**

*Community vote distribution*

B (86%)

14%

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

Option B, agree with comments explanation  
upvoted 1 times

✉️  **raaad** 2 weeks, 4 days ago

**Selected Answer: B**

- BigLake Integration: BigLake allows you to define tables on top of data in Cloud Storage, providing a bridge between data lake storage and BigQuery's powerful analytics capabilities. This approach is cost-effective and scalable.
- Data Catalog for Governance: Creating a taxonomy of policy tags in Google Cloud's Data Catalog and applying these tags to specific columns in your BigLake tables enables fine-grained, column-level access control.
- Processing with Spark and SQL: The Spark-BigQuery connector allows data scientists to process data using Apache Spark directly against BigQuery (and BigLake tables). This supports both Spark and SQL processing needs.
- Scalability into a Data Mesh: BigLake and Data Catalog are designed to scale and support the data mesh architecture, which involves decentralized data ownership and governance.

upvoted 3 times

✉️  **Jordan18** 3 weeks, 1 day ago

**Selected Answer: B**

BigLake leverages existing Cloud Storage infrastructure, eliminating the need for a dedicated Dataproc cluster, reducing costs significantly.  
upvoted 2 times

✉️  **scaenrui** 3 weeks, 4 days ago

**Selected Answer: C**

C.  
1. Load the data to BigQuery tables.  
2. Create a taxonomy of policy tags in Data Catalog.

Question #298

Topic 1

One of your encryption keys stored in Cloud Key Management Service (Cloud KMS) was exposed. You need to re-encrypt all of your CMEK-protected Cloud Storage data that used that key, and then delete the compromised key. You also want to reduce the risk of objects getting written without customer-managed encryption key (CMEK) protection in the future. What should you do?

- A. Rotate the Cloud KMS key version. Continue to use the same Cloud Storage bucket.
- B. Create a new Cloud KMS key. Set the default CMEK key on the existing Cloud Storage bucket to the new one.
- C. Create a new Cloud KMS key. Create a new Cloud Storage bucket. Copy all objects from the old bucket to the new one bucket while specifying the new Cloud KMS key in the copy command.
- D. Create a new Cloud KMS key. Create a new Cloud Storage bucket configured to use the new key as the default CMEK key. Copy all objects from the old bucket to the new bucket without specifying a key.

**Correct Answer: C**

*Community vote distribution*

D (88%)

13%

✉️  **Medmah** 5 days, 13 hours ago

I don't understand why only Matt select A

<https://cloud.google.com/sdk/gcloud/reference/kms/keys/update>

This seems to do the job, am I wrong ?

upvoted 1 times

✉️  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Definitely A

upvoted 1 times

✉️  **raaad** 3 weeks, 2 days ago

**Selected Answer: D**

- New Key Creation: A new Cloud KMS key ensures a secure replacement for the compromised one.
- New Bucket: A separate bucket prevents potential conflicts with existing objects and configurations.
- Default CMEK: Setting the new key as default enforces encryption for all objects in the bucket, reducing the risk of unencrypted data.
- Copy Without Key Specification: Copying objects without specifying a key leverages the default key, simplifying the process and ensuring consistent encryption.
- Old Key Deletion: After copying, the compromised key can be safely deleted.

upvoted 4 times

✉️  **rahulvin** 1 month ago

**Selected Answer: D**

Wrong:

A - rotating external key doesn't trigger re-encryption of data already in GCS: <https://cloud.google.com/kms/docs/rotate-key#rotate-external-coordinated>

C - Setting key during copy doesn't take care of objects that are later uploaded to the bucket, that will still use the default key

upvoted 3 times

Question #299

Topic 1

You have an upstream process that writes data to Cloud Storage. This data is then read by an Apache Spark job that runs on Dataproc. These jobs are run in the us-central1 region, but the data could be stored anywhere in the United States. You need to have a recovery process in place in case of a catastrophic single region failure. You need an approach with a maximum of 15 minutes of data loss (RPO=15 mins). You want to ensure that there is minimal latency when reading the data. What should you do?

1. Create two regional Cloud Storage buckets, one in the us-central1 region and one in the us-south1 region.
2. Have the upstream process write data to the us-central1 bucket. Use the Storage Transfer Service to copy data hourly from the us-central1 bucket to the us-south1 bucket.
3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in that region.
4. In case of regional failure, redeploy your Dataproc clusters to the us-south1 region and read from the bucket in that region instead.

- B. 1. Create a Cloud Storage bucket in the US multi-region.
- 2. Run the Dataproc cluster in a zone in the us-central1 region, reading data from the US multi-region bucket.
- 3. In case of a regional failure, redeploy the Dataproc cluster to the us-central2 region and continue reading from the same bucket.
- C. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.
- 2. Enable turbo replication.
- 3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the us-south1 region.
- 4. In case of a regional failure, redeploy your Dataproc cluster to the us-south1 region and continue reading from the same bucket.
- D. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.
- 2. Enable turbo replication.
- 3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the same region.
- 4. In case of a regional failure, redeploy the Dataproc clusters to the us-south1 region and read from the same bucket.

**Correct Answer: B**

*Community vote distribution*

D (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: D**

Option D, answers all needs from the request

upvoted 1 times

✉  **raaad** 3 weeks, 2 days ago

**Selected Answer: D**

- Rapid Replication: Turbo replication ensures near-real-time data synchronization between regions, achieving an RPO of 15 minutes or less.
- Minimal Latency: Dataproc clusters can read from the bucket in the same region, minimizing data transfer latency and optimizing performance.
- Disaster Recovery: In case of regional failure, Dataproc clusters can seamlessly redeploy to the other region and continue reading from the same bucket, ensuring business continuity.

upvoted 3 times

✉  **scaenruy** 3 weeks, 4 days ago

**Selected Answer: D**

D.

- 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.
- 2. Enable turbo replication.
- 3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the same region.
- 4. In case of a regional failure, redeploy the Dataproc clusters to the us-south1 region and read from the same bucket.

upvoted 1 times

You currently have transactional data stored on-premises in a PostgreSQL database. To modernize your data environment, you want to run transactional workloads and support analytics needs with a single database. You need to move to Google Cloud without changing database management systems, and minimize cost and complexity. What should you do?

- A. Migrate and modernize your database with Cloud Spanner.
- B. Migrate your workloads to AlloyDB for PostgreSQL.
- C. Migrate to BigQuery to optimize analytics.
- D. Migrate your PostgreSQL database to Cloud SQL for PostgreSQL.

**Correct Answer: A***Community vote distribution*

B (67%)

D (33%)

**✉️**  **datapassionate** 19 hours, 52 minutes ago**Selected Answer: D**<https://cloud.google.com/alloydb#all-features>

The requirement is to minimize cost and complexity. Cloud SQL would be the best choice.

upvoted 1 times

**✉️**  **Vaisnavi** 1 week, 4 days ago**Selected Answer: D**

Database Migration Service makes it easier for you to migrate your data to Google Cloud. This service helps you lift and shift your PostgreSQL workloads into Cloud SQL.

upvoted 1 times

**✉️**  **raaad** 3 weeks, 2 days ago**Selected Answer: B**

- AlloyDB is a fully managed, PostgreSQL-compatible database service with industry-leading performance.

upvoted 3 times

**✉️**  **AllenChen123** 1 week, 2 days ago

Why not D

upvoted 1 times

**✉️**  **scaenruy** 3 weeks, 4 days ago**Selected Answer: B**

B. Migrate your workloads to AlloyDB for PostgreSQL.

upvoted 1 times

You are architecting a data transformation solution for BigQuery. Your developers are proficient with SQL and want to use the ELT development technique. In addition, your developers need an intuitive coding environment and the ability to manage SQL as code. You need to identify a solution for your developers to build these pipelines. What should you do?

- A. Use Dataform to build, manage, and schedule SQL pipelines.
- B. Use Dataflow jobs to read data from Pub/Sub, transform the data, and load the data to BigQuery.
- C. Use Data Fusion to build and execute ETL pipelines.
- D. Use Cloud Composer to load data and run SQL pipelines by using the BigQuery job operators.

**Correct Answer: B**

*Community vote distribution*

A (100%)

✉️  **raaad** Highly Voted 3 weeks, 2 days ago

**Selected Answer: A**

- Aligns with ELT Approach: Dataform is designed for ELT (Extract, Load, Transform) pipelines, directly executing SQL transformations within BigQuery, matching the developers' preference.
  - SQL as Code: It enables developers to write and manage SQL transformations as code, promoting version control, collaboration, and testing.
  - Intuitive Coding Environment: Dataform provides a user-friendly interface and familiar SQL syntax, making it easy for SQL-proficient developers to adopt.
  - Scheduling and Orchestration: It includes built-in scheduling capabilities to automate pipeline execution, simplifying pipeline management.
- upvoted 5 times

✉️  **Matt\_108** Most Recent 2 weeks, 2 days ago

**Selected Answer: A**

Definitely A

upvoted 2 times

✉️  **scaenruy** 3 weeks, 4 days ago

**Selected Answer: A**

- A. Use Dataform to build, manage, and schedule SQL pipelines.

upvoted 3 times

✉️  **rahulvin** 1 month ago

**Selected Answer: A**

Dataform = transformations in SQL

upvoted 1 times

You work for a farming company. You have one BigQuery table named sensors, which is about 500 MB and contains the list of your 5000 sensors, with columns for id, name, and location. This table is updated every hour. Each sensor generates one metric every 30 seconds along with a timestamp, which you want to store in BigQuery. You want to run an analytical query on the data once a week for monitoring purposes. You also want to minimize costs. What data model should you use?

- A. 1. Create a metrics column in the sensors table.
2. Set RECORD type and REPEATED mode for the metrics column.
3. Use an UPDATE statement every 30 seconds to add new metrics.

- B. 1. Create a metrics column in the sensors table.  
 2. Set RECORD type and REPEATED mode for the metrics column.  
 3. Use an INSERT statement every 30 seconds to add new metrics.
- C. 1. Create a metrics table partitioned by timestamp.  
 2. Create a sensorId column in the metrics table, that points to the id column in the sensors table.  
 3. Use an INSERT statement every 30 seconds to append new metrics to the metrics table.  
 4. Join the two tables, if needed, when running the analytical query.
- D. 1. Create a metrics table partitioned by timestamp.  
 2. Create a sensorId column in the metrics table, which points to the id column in the sensors table.  
 3. Use an UPDATE statement every 30 seconds to append new metrics to the metrics table.  
 4. Join the two tables, if needed, when running the analytical query.

**Correct Answer: B**

*Community vote distribution*

C (100%)

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: C**

Option C  
upvoted 2 times

 **raaad** 3 weeks, 2 days ago

**Selected Answer: C**

Option C  
upvoted 3 times

 **raaad** 3 weeks, 2 days ago

Partitioned Metrics Table: Creating a separate metrics table partitioned by timestamp is a standard practice for time-series data like sensor readings. Partitioning by timestamp allows for more efficient querying, especially when you're only interested in a specific time range (like wee monitoring).

Reference to Sensors Table: Including a sensorId column that references the id column in the sensors table allows you to maintain a relations between the metrics and the sensors without duplicating sensor information.

INSERT Every 30 Seconds: Using an INSERT statement every 30 seconds to the partitioned metrics table is a standard approach for time-ser data ingestion in BigQuery. It allows for efficient data storage and querying.

Join for Analysis: When you need to analyze the data, you can join the metrics table with the sensors table based on the sensorId, allowing fo comprehensive analysis with sensor details.

upvoted 4 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: C**

C.  
 1. Create a metrics table partitioned by timestamp.  
 2. Create a sensorId column in the metrics table, that points to the id column in the sensors table.  
 3. Use an INSERT statement every 30 seconds to append new metrics to the metrics table.  
 4. Join the two tables, if needed, when running the analytical query.

upvoted 1 times



You are managing a Dataplex environment with raw and curated zones. A data engineering team is uploading JSON and CSV files to a bucket asset in the curated zone but the files are not being automatically discovered by Dataplex. What should you do to ensure that the files are discovered by Dataplex?

- A. Move the JSON and CSV files to the raw zone.
- B. Enable auto-discovery of files for the curated zone.
- C. Use the bg command-line tool to load the JSON and CSV files into BigQuery tables.
- D. Grant object level access to the CSV and JSON files in Cloud Storage.

**Correct Answer: D**

*Community vote distribution*

A (70%)

B (30%)

 **GCP001** 1 week, 5 days ago

**Selected Answer: A**

Should be A. Curated zone need Parquet, Avro, ORC format not CSV or JSON. Check the ref - <https://cloud.google.com/dataplex/docs/add-zone#curated-zones>

upvoted 4 times

 **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: B**

I'd go for Option B, auto-discovery is enabled by default for any zone, including curated ones, so if a file is not automatically discovered it's disabled auto-discovery

upvoted 2 times

 **Sofia98** 2 weeks, 5 days ago

**Selected Answer: A**

I will go with A, check the ref. Curated zones only store Parquet, Avro, and ORC in CS, and well-defined schema and Hive-style partitions in BigQuery:

<https://cloud.google.com/dataplex/docs/add-zone#curated-zones>

upvoted 2 times

 **raaad** 3 weeks, 2 days ago

**Selected Answer: B**

- Auto-Discovery Feature: Dataplex has an auto-discovery feature that, when enabled, automatically discovers and catalogs data assets within zone.

- Appropriate for Both Raw and Curated Zones: This feature is applicable to both raw and curated zones, and it should be tailored to the specific data governance and cataloging needs of the organization.

upvoted 1 times

 **scaenruy** 3 weeks, 4 days ago

**Selected Answer: A**

A. Move the JSON and CSV files to the raw zone.

upvoted 1 times

You have a table that contains millions of rows of sales data, partitioned by date. Various applications and users query this data many times a minute. The query requires aggregating values by using AVG, MAX, and SUM, and does not require joining to other tables. The required aggregations are only computed over the past year of data, though you need to retain full historical data in the base tables. You want to ensure

that the query results always include the latest data from the tables, while also reducing computation cost, maintenance overhead, and duration. What should you do?

- A. Create a materialized view to aggregate the base table data. Include a filter clause to specify the last one year of partitions.
- B. Create a materialized view to aggregate the base table data. Configure a partition expiration on the base table to retain only the last one year of partitions.
- C. Create a view to aggregate the base table data. Include a filter clause to specify the last year of partitions.
- D. Create a new table that aggregates the base table data. Include a filter clause to specify the last year of partitions. Set up a scheduled query to recreate the new table every hour.

**Correct Answer: D**

*Community vote distribution*

A (86%)

14%

✉  **Matt\_108** 2 weeks, 3 days ago

**Selected Answer: A**

- . Create a materialized view to aggregate the base table data. Include a filter clause to specify the last one year of partitions.  
upvoted 2 times

✉  **raaad** 2 weeks, 4 days ago

**Selected Answer: A**

- Materialized View: Materialized views in BigQuery are precomputed views that periodically cache the result of a query for increased performance and efficiency. They are especially beneficial for heavy and repetitive aggregation queries.
- Filter for Recent Data: Including a clause to focus on the last year of partitions ensures that the materialized view is only storing and updating the relevant data, optimizing storage and refresh time.
- Always Up-to-date: Materialized views are maintained by BigQuery and automatically updated at regular intervals, ensuring they include the latest data up to a certain freshness point.

upvoted 3 times

✉  **Sofia98** 2 weeks, 5 days ago

**Selected Answer: A**

- To preserve the historical data  
upvoted 1 times

✉  **scaenrui** 3 weeks, 4 days ago

**Selected Answer: B**

- B. Create a materialized view to aggregate the base table data. Configure a partition expiration on the base table to retain only the last one year of partitions.  
upvoted 1 times

✉  **Sofia98** 2 weeks, 5 days ago

Don't agree, it is said that we need to store the historical data, so answer A is correct

upvoted 2 times

✉  **raaad** 2 weeks, 4 days ago

Why not B

- Configuring partition expiration on the BASE TABLE is a way to manage storage and costs by automatically dropping old data. However, question specifies the need to retain full historical data, making this approach not suitable since it doesn't keep all historical records.

upvoted 1 times

Your organization uses a multi-cloud data storage strategy, storing data in Cloud Storage, and data in Amazon Web Services' (AWS) S3 storage buckets. All data resides in US regions. You want to query up-to-date data by using BigQuery, regardless of which cloud the data is stored in. You need to allow users to query the tables from BigQuery without giving direct access to the data in the storage buckets. What should you do?

- A. Setup a BigQuery Omni connection to the AWS S3 bucket data. Create BigLake tables over the Cloud Storage and S3 data and query the data using BigQuery directly.
- B. Set up a BigQuery Omni connection to the AWS S3 bucket data. Create external tables over the Cloud Storage and S3 data and query the data using BigQuery directly.
- C. Use the Storage Transfer Service to copy data from the AWS S3 buckets to Cloud Storage buckets. Create BigLake tables over the Cloud Storage data and query the data using BigQuery directly.
- D. Use the Storage Transfer Service to copy data from the AWS S3 buckets to Cloud Storage buckets. Create external tables over the Cloud Storage data and query the data using BigQuery directly.

**Correct Answer: D**

*Community vote distribution*

A (100%)

✉  **Matt\_108** 2 weeks, 2 days ago

**Selected Answer: A**

Option A - clearly explained in comments  
upvoted 2 times

✉  **raaad** 3 weeks, 2 days ago

**Selected Answer: A**

- BigQuery Omni: This is an extension of BigQuery that allows you to analyze data across Google Cloud, AWS, and Azure without having to manage the infrastructure or move data across clouds. It's suitable for querying data stored in AWS S3 buckets directly.
- BigLake: Allows you to create a logical abstraction (table) over data stored in Cloud Storage and S3, so you can query data using BigQuery without moving it.
- Unified Querying: By setting up BigQuery Omni to connect to AWS S3 and creating BigLake tables over both Cloud Storage and S3 data, you can query all data using BigQuery directly.

upvoted 3 times

✉  **AllenChen123** 4 days, 20 hours ago

Agree. <https://cloud.google.com/bigquery/docs/omni-introduction>

"To run BigQuery analytics on your external data, you first need to connect to Amazon S3 or Blob Storage. If you want to query external data, you would need to create a BigLake table that references Amazon S3 or Blob Storage data."

upvoted 2 times

✉  **rahulvin** 1 month ago

**Selected Answer: A**

A - BigLake tables work for S3 and GCS

upvoted 2 times

✉  **rahulvin** 1 month ago

[https://cloud.google.com/bigquery/docs/external-data-sources#external\\_data\\_source\\_feature\\_comparison](https://cloud.google.com/bigquery/docs/external-data-sources#external_data_source_feature_comparison)

upvoted 1 times