EXAMTOPICS

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

Custom View Settings

Topic 1 - Single Topic

Question #1 *Topic 1*

Your company built a TensorFlow neutral-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 5 months, 4 weeks 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 featuers 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 8 times

Morock Most Recent 1 week, 5 days ago

Selected Answer: C

Dropout is a regularization method to remove random selection of fixed number of unit in a neural network layer. So pick C for this question.

upvoted 1 times

enghabeth 1 week, 5 days ago

Selected Answer: C

becouse it is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data.

upvoted 1 times

samdhimal 1 month ago

C. Dropout Methods

Dropout is a regularization technique that can be used to prevent overfitting of the model to the training data. It works by randomly dropping out a certain percentage of neurons during training, which helps to reduce the complexity of the model and prevent it from memorizing the training data. This can improve the model's ability to generalize to new data and reduce the risk of poor performance when tested against new data.

upvoted 2 times

samdhimal 1 month ago

A. Threading: it's not a method to address overfitting, it's a technique to improve the performance of the model by parallelizing the computations using multiple threads.

B. Serialization: it's a technique to save the model's architecture and trained parameters to a file, it's helpful when you want to reuse the model later, but it doesn't address overfitting problem.

D. Dimensionality Reduction: it's a technique that can be used to reduce the number of features in the data, it's helpful when the data contains redundant or irrelevant features, but it doesn't address overfitting problem directly.

upvoted 3 times

korntewin 1 month, 3 weeks ago

answer is likely to be C, but D (dimensionality reduction) can also be used to mitigate overfitting too! Not sure which one is the correct answer.

upvoted 1 times

Brillianttyagi 2 months ago

Selected Answer: C

Correct

upvoted 1 times

odacir 2 months, 2 weeks ago

Answer is C.
We are in an over feting problem with nerual-net model.
Read at least the abstract of this.
https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

upvoted 1 times

ovokpus 3 months, 1 week ago

Selected Answer: C

Correct

upvoted 2 times

Atnafu 4 months, 1 week ago

C is the correct answer.
Dropouts: are special layers that you can add to control some operations or add functionalities. If training has an issue you can take them out.
In the question, it fits training but not new data, i.e., overfitting.
Solution for overfitting is dropout

upvoted 1 times

vivi_p 4 months, 4 weeks ago

https://iq.opengenus.org/dropout-in-tensorflow/#:~:text=Let%20us%20understand%20what,good%20accuracy%20%26%20balance.&text=Let%20us,%26%20balance.&text=understand%20what,good%20accuracy

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?

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 (86%) 14%

serg3d Highly Voted 2 years, 9 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 32 times

dambilwa 2 years, 8 months ago

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

upvoted 4 times

jagadamba Highly Voted 2 years, 8 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

jin0 Most Recent 6 days, 15 hours 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 6 days, 15 hours 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 1 week, 5 days ago

Selected Answer: C

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

upvoted 1 times

samdhimal 1 month 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 the 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 overfitting and not generalize well to new users.

upvoted 2 times

korntewin 1 month, 3 weeks 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 month, 3 weeks ago

Correct answer is B

upvoted 1 times

javibadillo 3 months, 3 weeks ago

Selected Answer: B

B is the correct answer

upvoted 2 times

Maeel92 4 months, 1 week ago

B is the correct answer. Model training using existing data + new data will allow model to learn new behaviour of users while retaining some of the older behaviour to make model.

upvoted 1 times

Hikky_111 8 months, 3 weeks ago

B seems like the best option because we need to make the prediction based on the latest preferences by an individual also we need not forget how the user used to prefer before .It's highly unlikely that a person's fashion sense changes a lot.

upvoted 1 times

noob_master 8 months, 3 weeks ago

Selected Answer: B

"You know a user's fashion preference is likely to change over time" - It is necessary to understand old data and new inputs for a good model - B is the best option

upvoted 1 times

Subhranil1010 9 months ago

B would be correct ans

upvoted 1 times

nana1995 1 year ago

Selected Answer: B

Correct

upvoted 1 times

MaxNRG 1 year, 3 months ago

I would go with B (but A is also possible If the market is fast changing)
https://datascience.stackexchange.com/questions/12761/should-a-model-be-re-trained-if-new-observations-are-available:
Suppose that your model attempts to predict customers' behavior, e.g. how likely is a customer to purchase your product given an offer tailored for him. Clearly, the market changes over time, customers' preferences change, and your competitors adjust. You should adjust as well, so you need to retrain periodically. In such a case I would recommend to add new data, but also omit old data that is not relevant anymore. If the market is fast changing, you should even consider retraining periodically based on new data only.
https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html:
It is a good practice to continuously monitor the incoming data and retrain your model on newer data if you find that the data distribution has deviated significantly from the original training data distribution

upvoted 4 times

samdhimal 1 year, 1 month ago

Couldn't Agree More. The question seems little vast. It should've specified how a timeframe could play a role as well.

upvoted 2 times

Palanee 1 year ago

Answer B is fair, because the user fashion preference is likely to change overtime and question says we need to build the data pipeline to stream new data back to the model when it is available. So retraining the model with new and existing data would be fair.

upvoted 1 times

ML_Novice 1 year ago

it seems like a fair answer, although its says , the user is LIKELY to change preference. Its likely to not a fact when you continously retrain the model on the new data you are assuming that he s completely going to change his clothe preference

upvoted 1 times

anji007 1 year, 4 months ago

Ans: B

upvoted 2 times

sumanshu 1 year, 11 months ago

Vote for B.

In Recommendation System - Matrix stored in database - which store the users rating and other details like how much time user spent on that page.We generally recommend on basis of matrix. And In production that matrix updated (or model get retrained once a day or once a week) - Because there could be new users rating. or we can say new data.

So model retrained fully i.e. on Existing Data + New Data => and generate a new matrix (user/item matrix)

upvoted 3 times

sumanshu 1 year, 11 months ago

Amazon Recommend Method -

https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html

upvoted 3 times

sumaiya96 1 year, 11 months ago

It's 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 Highly Voted 2 years, 2 months ago
C - based on Google documentation, self-join is an anti-pattern: https://cloud.google.com/bigquery/docs/best-practices-performance-patterns
upvoted 23 times

awssp12345 1 year, 8 months ago
Super helpful! Thanks. :)
upvoted 1 times

[Removed] Highly Voted 2 years, 11 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 clinic while also increasing performance due to avoiding self-joins.
upvoted 12 times

Morock Most Recent 1 week, 5 days ago
Selected Answer: C
Should be C. Basic ER design...
upvoted 1 times

GCPpro 1 month, 2 weeks ago
c - is the correct one.
upvoted 1 times

testoneAZ 1 month, 3 weeks ago
C should be the correct answer
upvoted 1 times

Brillianttyagi 2 months ago
Selected Answer: C
C- Is the correct answer!
upvoted 1 times

Arkon88 12 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 ago
Selected Answer: C
Correct
upvoted 1 times

samdhimal 1 year, 1 month 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 month 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 1 year, 3 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 5 times

anji007 1 year, 4 months ago

Ans: C

upvoted 1 times

sumanshu 1 year, 7 months ago

Vote for C. I read this question in 'google 'sample question' of PDE.

upvoted 2 times

lbhhoya82 1 year, 11 months ago

Correct: C

upvoted 1 times

sid091 1 year, 12 months ago

C is correct . Self Joins increases the overhead

upvoted 1 times

naga 2 years ago

Correct C

upvoted 1 times

balseron99 2 years ago

A is incorrect because adding space won't solve the problem of query performance.
B is incorrect because there is nothing related to the report generation which is specified and sharding tables on date ranges is not a good option as it will create many tables.
C is CORRECT because the statement says "the scope of the project has expanded. The database must now store 100 times more patient records". As the data increases there would be difficulty in managing the tables and querying it. Hence creating different table is correct as per the need.
D is Incorrect as it Partitions on each clinic. We have to adjust the database design so that it performs optimally when generating reports. Also nothing is specified for generation of reports in the required statement.

upvoted 5 times

luatnc 2 years, 3 months ago

It should be C, to avoid self-join

upvoted 2 times

Question #4 *Topic 1*

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 virtualizations.

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

Khaled_Rashwan 2 days, 17 hours 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 1 times

Morock 1 week, 5 days ago

Selected Answer: D

The solution from the site is perfect.

upvoted 1 times

PolyMoe 1 month ago

Selected Answer: A

what is relevant here is to uncache Data Studio

upvoted 1 times

samdhimal 1 month 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 2 times

korntewin 1 month, 3 weeks 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 3 months, 1 week 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 3 months, 2 weeks 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, user 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 setup 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 3 months, 4 weeks 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 3 months, 4 weeks 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 4 months ago

Selected Answer: A

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

upvoted 1 times

kennyloo 4 months, 1 week ago

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

upvoted 1 times

max_c 4 months, 3 weeks 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 month ago

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

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: A

A. Disable caching by editing the report settings.

upvoted 3 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 2 years, 11 months ago
Agreed: D
upvoted 22 times

Radhika7983 Highly Voted 2 years, 3 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 11 times

jkhong 2 months, 3 weeks 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 1 times

Morock Most Recent 1 week, 5 days ago
Selected Answer: D
D. The question is asking pipeline, then let's build a pipeline.
upvoted 1 times

samdhimal 1 month 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 into 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 1 times

samdhimal 1 month 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 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.
upvoted 3 times

Besss 4 months, 2 weeks ago
Selected Answer: D
Agreed: D
upvoted 1 times

Dip1994 6 months, 4 weeks ago
The correct answer is D
upvoted 1 times

Arkon88 12 months ago

Selected Answer: D

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

upvoted 1 times

MaxNRG 1 year, 3 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 1 year, 4 months ago

Ans: D

upvoted 1 times

nickozz 1 year, 5 months ago

D seems to be correct. explained here how combined wth Pub/Sub, this can be achieved. https://cloud.google.com/pubsub/docs/handling-failures

upvoted 1 times

kubosuke 1 year, 6 months ago

D
this structure is known as "dead letter pattern".

https://medium.com/@rako/error-rate-monitoring-in-dead-letter-pattern-using-apache-beam-193fd03f8970

upvoted 2 times

fire558787 1 year, 6 months ago

Disagree a bit here. Could well be A. In one Coursera video course (https://www.coursera.org/learn/batch-data-pipelines-gcp/lecture/SkDus/how-to-carry-out-operations-in-bigquery), they do have a video about when to just use an SQL query to find wrong data without creating a Dataflow pipeline. The question says "SQL" as a language, not Cloud SQL as a service. Federated Sources is great because you can federate a CSV file in GCS with BigQuery. From the video: "In this section, we'll take a look at exactly how BigQuery can help with some of those data quality issues we just described. Let's start with validity, what do we mean by invalid? It can mean things like corrupted data maybe data that is missing a timestamp"

upvoted 4 times

kelvinksau 1 year, 6 months ago

https://cloud.google.com/bigquery/external-data-sources
Use cases for external data sources include:

For ETL workloads, loading and cleaning your data in one pass and writing the cleaned result into BigQuery storage.
Joining BigQuery tables with frequently changing data from an external data source. By querying the external data source directly, you don't need to reload the data into BigQuery storage every time it changes.

upvoted 2 times

amityleo 1 year, 9 months ago

D seems the correct answer: https://cloud.google.com/bigquery/docs/loading-data#dataflow

upvoted 1 times

lbhhoya82 1 year, 11 months ago

Correct : C

upvoted 1 times

sumanshu 1 year, 8 months ago

A. Use federated data sources, and check data in the SQL query. - WRONG (Because we are changing source itself, i.e. SQL, MySQL, PstgresSQL) instead of correcting the problem
B. Enable BigQuery monitoring in Google Stackdriver and create an alert. (WRONG - Because setting and creating an alert will not solve the corrupted data problem)
C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0. (wrong - here we are saying set max_bad_records = 0 (i.e let's load all bad records into bi-query)
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 - Dataflow is used for this pupose only i.e transform the data and dead letter queue pupose is to write any invalid records - so that it can be analyzed later (rather than ignoring))

upvoted 5 times

Hoham 4 months, 2 weeks ago

Correct - D
To be precise why the option C is wrong (Sumanshu and lbhhoya82), max_bad_records=0 doesn't mean that all bad records will be loaded, that means the max number of bad records allowed is 0 (https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv#:~:text=Number%20of%20bad%20records%20allowed), in the other hand, when you use this option with a value > 0 then the CSV file loading will be interrupted when the number of bad records it's greater than the vale defined with this message: {Too many errors encountered. Limit is: ..} (https://stackoverflow.com/questions/37329980/max-bad-records-in-google-bigquery), therefore with this option we'll never have the CSV file landing in BQ...

upvoted 2 times

sumanshu 1 year, 7 months ago

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

upvoted 1 times

SonuKhan1 1 year, 4 months ago

SQL u can write on Bigquery also. Not necessary MySQL etc

upvoted 1 times

sid091 1 year, 12 months ago

D is correct

upvoted 2 times

naga 2 years ago

Correct D

upvoted 3 times

Andrabi 2 years, 6 months ago

Agree with D

upvoted 2 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 server 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 2 years, 3 months ago

Correct answer is B. App engine create applications that use Cloud SQL database connections effectively. Below is what is written in google cloud documnetation.

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 39 times

llamaste Highly Voted 2 years, 7 months ago

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

upvoted 11 times

yafsong Most Recent 2 months, 2 weeks 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 3 times

hiromi 3 months, 1 week ago

Selected Answer: B

B is right

upvoted 1 times

shiv14 1 year ago

Selected Answer: B

According to the documentation

upvoted 1 times

samdhimal 1 year, 1 month 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 1 times

samdhimal 1 month 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 or a 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 1 times

deep_ROOT 1 year, 1 month ago
B is Correct; this question appeared in Cloud Architect exam also
upvoted 1 times

MaxNRG 1 year, 3 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

anji007 1 year, 4 months ago
Ans: B
upvoted 1 times

sumanshu 1 year, 10 months ago
Vote for B.

https://cloud.google.com/iot/docs/how-tos/exponential-backoff
upvoted 3 times

lbhhoya82 1 year, 11 months ago
Correct : B
upvoted 1 times

daghayeghi 1 year, 11 months ago
B:
https://cloud.google.com/sql/docs/mysql/manage-connections
upvoted 1 times

sid091 1 year, 12 months ago
B is correct answer.
https://cloud.google.com/sql/docs/mysql/manage-connections
upvoted 3 times

naga 2 years ago
Correct B
upvoted 2 times

PRABHUKKARTHI 2 years, 6 months ago
Option B: Its the best way to connect the database again in case of failure
upvoted 2 times

Nidie 2 years, 10 months ago
https://cloud.google.com/sql/docs/mysql/manage-connections
upvoted 4 times

[Removed] 2 years, 11 months ago
Correct: B

Exponential checking is best practice
upvoted 2 times

Question #7 *Topic 1*

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

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 2 years, 4 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 forecasting 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 or no, you will be using logistics regression.

upvoted 39 times

Anirkent 2 years, 1 month 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 just a distraction?

upvoted 6 times

muzammilnxs 2 years 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 17 times

[Removed] Highly Voted 2 years, 11 months ago

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

upvoted 9 times

Zosby Most Recent 3 days, 23 hours ago

predict housing prices = linear regresssion

upvoted 1 times

JJJJim 1 week, 5 days ago

Selected Answer: A

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

upvoted 1 times

lukas_xls 2 months, 2 weeks ago

Selected Answer: A

Must be A

upvoted 1 times

rowan_ 6 months, 2 weeks ago

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

upvoted 1 times

samdhimal 1 year, 1 month ago

correct answer -> Linear Regression

Linear regression is a statistical method that allows to summarize and study relationships between two continuous (quantitative) variables: One 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.

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

upvoted 4 times

samdhimal 1 month ago

Linear regression is a simple and computationally efficient algorithm that can be used to predict a continuous target variable based on one or 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 1 times

MaxNRG 1 year, 3 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 1 year, 4 months ago

Ans: A

upvoted 1 times

StefanoG 1 year, 5 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 1 times

sumanshu 1 year, 7 months ago

Vote for A

upvoted 1 times

lbhhoya82 1 year, 11 months ago

Correct : A

upvoted 1 times

daghayeghi 1 year, 11 months ago

A:
When a model is used to predict a continuous value, it is known as a regression model.

upvoted 1 times

davidbilla 1 year, 12 months ago

A. Linear Regression

upvoted 2 times

sid091 1 year, 12 months ago

Easy One. Linear Regression . Option A

upvoted 2 times

senthil836 2 years ago

Correct Answer is A

upvoted 2 times

naga 2 years ago

Correct A

upvoted 2 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] Highly Voted 2 years, 11 months ago
Answer: D
Description: Row Number equals 1 with partitioning will ensure only one record is fetched per partition
upvoted 18 times

[Removed] Highly Voted 2 years, 11 months ago
Correct D
upvoted 13 times

Zosby Most Recent 4 days, 16 hours ago
Correct D
upvoted 1 times

Morock 1 week, 4 days ago
Selected Answer: D
Row number gives the unique number ranking based on target column.
upvoted 2 times

odacir 2 months, 2 weeks ago
Selected Answer: D
It's the only valid option, try it your self with examples in QB.
upvoted 1 times

Ender_H 5 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 6 times

Davijde13 1 month, 1 week 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 1 times

Mamta072 8 months, 1 week ago
Ans is D as Row number is the clause to fetch unique record from duplicate
upvoted 1 times

Arkon88 12 months ago
Answer: D
upvoted 1 times

samdhimal 1 year, 1 month 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 1 times

samdhimal 1 month 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 1 times

samdhimal 1 month 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, so 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 1 times

nofaruccio 1 year, 3 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 1 year, 3 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 5 times

anji007 1 year, 4 months ago

Ans: D

upvoted 1 times

sumanshu 1 year, 10 months ago

Vote for D.

Explanation:

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

upvoted 11 times

awssp12345 1 year, 8 months ago

very helpful. :)

upvoted 1 times

lbhhoya82 1 year, 11 months ago

Correct : D

upvoted 1 times

daghayeghi 1 year, 11 months ago

D:
https://cloud.google.com/bigquery/streaming-data-into-bigquery#manually_removing_duplicates

upvoted 9 times

sid091 1 year, 12 months ago

D is correct

upvoted 3 times

senthil836 2 years ago

Correct answer is D.

upvoted 2 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 (100%)

Ender_H Highly Voted 5 months, 1 week ago

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

upvoted 12 times

Davijde13 1 month, 1 week ago

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

upvoted 1 times

Jasar 3 months, 2 weeks ago

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

upvoted 1 times

Zosby Most Recent 4 days, 16 hours ago

D is correct

upvoted 1 times

priluft 5 months, 3 weeks ago

Selected Answer: D

D. 'bigquery-public-data.noaa_gsod.gsod*`

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: D

D. 'bigquery-public-data.noaa_gsod.gsod*`

upvoted 1 times

Question #10 *Topic 1*

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 (59%) BDF (41%)

samdhimal Highly Voted 1 year, 1 month 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 19 times

samdhimal 1 month ago

I was WRONG. I am not sure why s o 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 2 times

samdhimal 1 month 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 3 times

SuperTee 1 week, 3 days ago

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

upvoted 1 times

[Removed] Highly Voted 2 years, 11 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 18 times

jin0 Most Recent 6 days, 1 hour 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 2 weeks, 6 days ago
Selected Answer: BDE
I don't choose stack driver because it suits more for audit

upvoted 2 times

Nirca 1 month, 2 weeks 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 1 times

desertlotus1211 1 month, 2 weeks ago
It's not about an audit... it's about violation for least privileges. this will aid in correcting

upvoted 1 times

Jackalski 2 months 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

upvoted 3 times

desertlotus1211 1 month, 2 weeks ago
E is an administrative nightmare... it's about given a user only the privileges to do their function. So it's all based on their service account.

upvoted 1 times

Nirca 2 months, 2 weeks ago
Selected Answer: BDF
are the right answers

upvoted 3 times

odacir 2 months, 2 weeks ago
Selected Answer: BDF
BDF
A. Disable writes to certain tables. ← No sense, to write it at user or group level no table level
B. Restrict access to tables by role. ← YES, Best recommended practices
C. Ensure that the data is encrypted at all times. ← Data is always encrypted at rest and in flight with BQ.
D. Restrict BigQuery API access to approved users. ← YES, best recommended practices
E. Segregate data across multiple tables or databases ← BigQuery not have multiple databases, it has data sets and tables, and it's nor necessary to segregate the data, you have ACL to control access to column level, etc.
F. Use Google Stackdriver Audit Logging to determine policy violations. ← Good practice, if your industry is regulated, have audit logging.

upvoted 2 times

Asheesh1909 3 months ago
Selected Answer: BDE
E seems more sensible to be as the question concentrates more on table access restriction than access violation, policy violations can only be determined through stackdriver how ever we cant restrict the access to tables.Probably option E should be considered as, by segregating the data into diferrent tables , we can restrict access to tables.

upvoted 2 times

odacir 2 months, 2 weeks ago
E says segregate across multiple tables or databases, this is not the pattern of BigQuery, in BQ there is only one database, and you can organize your data in datasets...

upvoted 1 times

hauhau 3 months ago
Option E doesn't mention about restrict access

upvoted 1 times

sfsdeniso 3 months ago
BDE is correct
F is not an option for 'ensure individual users have access only to the minimum amount of information required to do their jobs'
upvoted 1 times

Atnafu 3 months ago
Answer BDF
upvoted 1 times

NicolasN 3 months ago
Selected Answer: BDE
I disagree with [F]. It's too late for a "highly regulated industry" to detect access violations by audit logs.
[E] is a more reasonable answer, since it is a kind of row-level security, especially the times when BigQuery row-level security wasn't available.
It is a practice still recommended (even with row-level sec. available) for the extreme scenario that:
(Through repeated observation of query duration when querying tables with row-level access policies,) "a user could infer the values of rows that otherwise might be protected by row-level access policies"
"If you are sensitive to this level of protection, we recommend using separate tables to isolate rows with different access control requirements, instead."
Source:
https://cloud.google.com/bigquery/docs/best-practices-row-level-security#limit-side-channel-attacks
upvoted 3 times

MisuLava 4 months ago
Selected Answer: BDF
Correct: BDF
upvoted 1 times

rowan_ 6 months, 2 weeks ago
D B and F. Restrict in the first place; then you don't have write issues. Write was not even mentioned in the prompt; just access. This is more of a logical question. I believe E should be included
upvoted 1 times

prasanna77 1 year, 2 months ago
B,D,E,F looks like valid,
Since option E is Segregate data across multiple tables or databases here it is given databases which is not valid as there is only datasets in BigQuery
upvoted 3 times

MaxNRG 1 year, 3 months ago
It's an old question, old answers were DEF imho:
upvoted 2 times

MaxNRG 1 year, 3 months ago
Option D would help limit access to approved users only.
Option E as it would help segregate the data with the ability to provide access to users as per their needs.
Option F as it would help in auditing.
https://cloud.google.com/bigquery/docs/access-control
You share access to BigQuery tables and views using project- level LAM roles and dataset-level access controls. Currently, you cannot apply access controls directly to tables or views.
Project-level access controls determine the users, groups, and service accounts allowed to access all datasets, tables, views, and table data within a project.
Dataset-level access controls determine the users, groups, and service accounts allowed to access the tables, views, and table data in a specific dataset.
Option A is wrong as disabling writes does not prevent the users from reading and does not align with the least privilege principle.
Option B is wrong as access cannot be control on tables.
Option C is wrong as data is encrypted by default, however it does not align with the least privilege principle.
upvoted 3 times

MaxNRG 1 year, 3 months ago
now BDF is ok
upvoted 1 times

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

C (100%)

vetaal Highly Voted 3 months, 3 weeks 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 7 times

samdhimal Most Recent 3 weeks, 1 day 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 new 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 1 times

kennyloo 4 months, 1 week ago

Only C is feasible for this question

upvoted 1 times

AWSandeep 5 months, 4 weeks 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 (63%) BEF (38%)

[Removed] Highly Voted 2 years, 11 months ago
agreed B,D,F
upvoted 36 times

saurabh1805 Highly Voted 2 years, 6 months ago
My vota also goes for B,D,F
upvoted 9 times

samdhimal Most Recent 3 weeks, 1 day 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 1 times

Nirca 1 month, 2 weeks 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 1 times

jin0 6 days, 1 hour 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 month, 2 weeks 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 2 months ago
Please why is DEF not correct?
upvoted 1 times

Kyr0 2 months ago
Selected Answer: BDF
Agree BDF
upvoted 1 times

odacir 2 months, 2 weeks 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 trow service account, and not throw user level.

upvoted 3 times

cloudyy 1 day, 11 hours ago

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

upvoted 1 times

ler_mp 2 months ago

Yes, I also wonder why not E instead of D

upvoted 1 times

lalli117 5 months, 2 weeks ago

agreed, B,D,F

upvoted 1 times

hui521 6 months, 1 week ago

why is C not correct? Can someone explain?

upvoted 1 times

sfsdeniso 3 months ago

because of this: "with some allowing direct query access via Google BigQuery"

upvoted 1 times

Whoswho 4 months, 2 weeks ago

"only a(one) service account to access all datasets".which is wrong because we're actually trying to implement access control on clients datasets and data.

upvoted 3 times

jackjackington 9 months, 1 week ago

These questions rock!! They helped me pass the exam on the first attempt!! Make sure you study all of these questions carefully with their associated documentation.

upvoted 5 times

gamer5005 10 months, 2 weeks ago

Selected Answer: BDF

correct

upvoted 1 times

Arkon88 12 months ago

Selected Answer: BDF

agreed B,D,F

upvoted 1 times

anji007 1 year, 4 months ago

Ans: B, D and F.

upvoted 4 times

[Removed] 1 year, 4 months ago

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

What does it mean? Why we need to restrict to a 'approved' user?

upvoted 3 times

ishal 4 months, 4 weeks ago

Restrict a client's dataset to approved users means restricting the data only between approved users. Not the other way around.

upvoted 1 times

prasioso 1 year, 10 months ago

I agree with B, D, F. But could someone explain how A (loading different partitions) and C (different tables) would violate the stated requirements?

upvoted 2 times

thesysops 1 year, 9 months ago

I'm not so sure that C would violate the requirements anymore now that you can set access controls on the table level. I don't think that was possible when this question was created.

Answer A just isn't relevant to the goal. Partitioning data in BigQuery is more about improving performance and is not used for access controls.

upvoted 1 times

thesysops 1 year, 9 months ago

Actually, C still would probably be wrong because it doesn't say anything about setting the table level access which would still be required split customer's data by table.

upvoted 1 times

sergio6 1 year, 4 months ago

How could you put a dataset into a table ? A dataset is a container of tables

upvoted 3 times

sumanshu 1 year, 10 months ago

Some voted for 'E' i.e. E. Only allow a service account to access the datasets.
Not sure why ?

if we gave access ONLY to service account - Does not it mean - we need to access BigQuery using Some Code (by mentioning Service account credentials there) OR using some other resource like VM)
In this case - i think person can't even access the Big Query Service via UI (if we give access only to Service account). Correct me if there is option on UI as well

upvoted 8 times

awssp12345 1 year, 8 months ago

yes, that is precisely why we need to eliminate E.

upvoted 5 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

DeepakKhattar Highly Voted 2 years, 9 months ago
Initially, thinking D is the best answer but when question is re-re-read, A seems to be correct answer for following reasons
1. Is payment TRANSACTION -- DB should 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 mot 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 59 times

canon123 1 year, 3 months ago
CloudSql does not auto scale.
upvoted 9 times

BigQuery 1 year, 2 months ago
https://cloud.google.com/architecture/elastically-scaling-your-mysql-environment#objectives

Please read. It can be configured for autoscaling.
upvoted 2 times

nkunwar 5 months ago
C SQL doesn't AUTO SCALE, you need to manually edit , Please show where does it says AUTO SCALING
upvoted 1 times

hendrixlives 1 year, 2 months 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 3 times

MisuLava 4 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 1 times

Blobby 1 year, 6 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 2 times

Blobby 1 year, 5 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

jvg637 Highly Voted 2 years, 11 months ago

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

upvoted 34 times

BigQuery 1 year, 2 months 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 1 year, 2 months 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 questions 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 3 times

imsaikat50 2 months, 3 weeks 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 is being e-commerce.

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

upvoted 1 times

Zosby Most Recent 3 days, 23 hours ago

" transactions" # B and D are Wrong

upvoted 1 times

musumusu 4 days, 21 hours ago

Answer: D,
Key words: Point of sale (pos) you can compare it with fast payment methods ex, barcode scan of product or tickets and credit card scans.
These transactions doesn't need SQL database as it gets information from the card/barcode itself.
Right answer is Firestore as based on number of clients but alternative of Firestore is Datastore (old version of Firestore) you can say.
It has no limits of POS or Synchornization processing such as sensor.
If there a Question for fastest sync database, you can choose Firestore/datastore.

upvoted 1 times

techtitan 1 week, 1 day ago

Selected Answer: D

Datastore is a highly scalable NoSQL database for your applications. Datastore automatically handles sharding and replication, providing you with a highly available and durable database that scales automatically to handle your applications' load. Datastore provides a myriad of capabilities such as ACID transactions, SQL-like queries, indexes, and much more.

upvoted 2 times

chxnge 2 weeks, 6 days ago

Selected Answer: C

Whizlabs says C

upvoted 1 times

cucinareblog 3 weeks ago

Selected Answer: A

Transactional data i better to go into relational DB like CloudSql

upvoted 2 times

samdhimal 3 weeks, 1 day ago

A. Cloud SQL

Cloud SQL is a fully managed relational database service that is well suited for applications that require a structured and scalable database. It can handle high levels of concurrent connections and is designed to be highly available, so you do not need to manage infrastructure scaling. Since your payment transactions will likely require a structured database with tables and relationships, Cloud SQL is the best choice for this use case.

upvoted 2 times

samdhimal 3 weeks, 1 day ago

B. BigQuery is a data warehousing and analytics service designed for large-scale data processing and analysis. While it could be used to store payment transactions, it is not optimized for handling high levels of concurrent connections, making it less suitable for a point-of-sale application.

C. Cloud Bigtable is a NoSQL database service designed for handling massive amounts of semi-structured and unstructured data. It is not well suited for a point-of-sale application that requires a structured database with tables and relationships, making it a less ideal choice.

D. Cloud Datastore is a NoSQL document-oriented database service that is well suited for applications that require a flexible, scalable, and

highly available data store. While it could be used to store payment transactions, it is not optimized for handling high levels of concurrent connections, making it less suitable for a point-of-sale application.

upvoted 2 times

KGCP25 1 month, 2 weeks ago

Referring to https://cloud.google.com/datastore/docs/concepts/overview#other_storage_and_database_options --- If you need a relational database with full SQL support for an online transaction processing (OLTP) system, consider Cloud SQL.

upvoted 1 times

Nirca 1 month, 2 weeks ago

Selected Answer: C

BigTable is: "A fully managed, scalable NoSQL database service for large analytical and operational workloads with up to 99.999% availability"

* Cloud SQL is NOT auto-scale for Writes (inserts). only for reads (select)
* Datastore is not a stand-alone system. it is part on the Firestore/Data system. So I'm not expecting GCP to suggest this option any more (starting from 2022)
* BigQuery is only for analytics

upvoted 1 times

korntewin 1 month, 3 weeks ago

The answer should be C, as BigTable is the most scalable solution with transaction capability.

D is no sql DB with no strict schema which is not appropriate for fixed schema transaction workload.

IMO, A is not scalable enough!

upvoted 1 times

Jackalski 2 months ago

Selected Answer: C

high load cloud application -> bigtable is recommended.
A - cloudSQL will work but not scale as fast nor provide fast response time on high load
D - datastore is designed for simple use cases

upvoted 1 times

Brillianttyagi 2 months ago

Selected Answer: D

D- Cloud datastore is the right answer as it autoscale automatically but only storage scale automatically in clout sql not cpu, as load increase you also have to increase load.

You can find the answer in this guide.
https://towardsdatascience.com/how-i-passed-google-professional-data-engineer-exam-in-2020-2830e10658b6

upvoted 2 times

DGames 2 months, 2 weeks ago

Answer - A This question related to RDMS (Point-of sales) database and in GCP - cloud SQL is option. also need to maintain ACID property.

upvoted 1 times

permoon 2 months, 3 weeks ago

Selected Answer: A

I' asked ChatGPT about this question, and it told me the correct answer is A.

If you want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform and you don't want to worry about managing infrastructure scaling, you should use Google Cloud SQL. Cloud SQL is a fully-managed database service that makes it easy to set up, maintain, manage, and administer your relational databases on Google Cloud Platform.

upvoted 5 times

jkhong 2 months, 1 week ago

I think u need to revisit how chatgpt gives answers lol, I've asked it and it gave me Datastore as an answer

The correct answer is D. Cloud Datastore

Cloud Datastore is a fully managed, NoSQL document database service that provides automatic scalability, high performance, and durability for your applications. It is designed to handle large volumes of structured and unstructured data, and can support rapid growth without requiring you to manage infrastructure scaling. Cloud Datastore is a good choice for storing and retrieving large amounts of data quickly and efficiently, making it a suitable option for a point-of-sale application that may experience rapid growth in user base.

upvoted 2 times

Asheesh1909 3 months ago

Selected Answer: D

Datastore can autoscale and supports transactions, in case of exponential increase in sales too.

upvoted 2 times

Fabius 3 months, 1 week ago

Cloud Datastore is a document DB.
Cloud SQL is a correct answer

upvoted 2 times

Question #14 *Topic 1*

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

AC (47%) AD (39%) 14%

jvg637 Highly Voted 2 years, 11 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 59 times

BigQuery 1 year, 2 months 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%20detection%20has%20two%20basic,from%20the%20normal%20instances%20significantly.

upvoted 13 times

szefco 1 year, 2 months 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 17 times

jvg637 Highly Voted 2 years, 11 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

musumusu Most Recent 4 days, 20 hours ago

Right Answer is C & D:
Unsupervised learning can only justify the patten of data and make custer of them.
For example, It can make
1 cluster of Apple,
1 cluster of mango,
Now it can predict the Apple or mango (similar feature)
or Graphes in new cluster (different feature)

Option A: It doesn't matter there are few occurance or more to make the cluster. If there is only one Apple and Mango it will still make two cluster.

upvoted 1 times

charline 2 weeks ago

Selected Answer: AD

the future mutation should be similar to the tested mutations

upvoted 1 times

Zhaas 4 days, 19 hours ago

No, for ML 2 classes should be different

upvoted 1 times

certmonkey 2 weeks, 4 days ago

Selected Answer: AC

Anomaly detection unsupervised learning
The objective of Unsupervised Anomaly Detection is to detect previously unseen rare objects or events without any prior knowledge about these. The only information available is that the percentage of anomalies in the dataset is small, usually less than 1%.

From: https://paperswithcode.com/task/unsupervised-anomaly-detection#:~:text=The%20objective%20of%20Unsupervised%20Anomaly,%2C%20usually%20less%20than%201%25.

So it is A & C

upvoted 1 times

samdhimal 3 weeks, 1 day ago

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

dconesoko 1 month, 1 week ago

Selected Answer: AC

as i explained earlier

upvoted 2 times

dconesoko 1 month, 1 week ago

My two cents, we can not be very sure about how future mutations looks like, because thats a biological blind process, thus all we can do is to able to identity what is normal and no matter how the tissue mutates , so far as its different from the normal sample then its tagged as an anomaly. thus i vote for AC

upvoted 1 times

korntewin 1 month, 3 weeks ago

Selected Answer: AD

The answer should be AD.

A, anomaly should have a little amount, if there are many samples then we should do classification instead, because unsupervised will give a lot of false positive.

D, the future anomaly should be of the same distribution as present anomaly! or else our anomaly detection will not be generalize to the future feature.

upvoted 1 times

Kyr0 2 months ago

Selected Answer: AD

AD for me

upvoted 1 times

dragolemguty 2 months, 1 week ago

Selected Answer: AD

AD more sense

upvoted 1 times

Nirca 2 months, 2 weeks ago

Selected Answer: AC

Supervised and unsupervised learning have one key difference. Supervised learning uses labeled datasets, whereas unsupervised learning uses unlabeled datasets. By “labeled” we mean that the data is already tagged with the right answer.

upvoted 2 times

jkhong 2 months, 2 weeks ago

Selected Answer: AD

A makes sense

C and D compares future mutations to mutated samples in database

The question is pretty badly worded… If we were to run a full unsupervised anomaly detection over the entire dataset, C and D will be true, since some future mutations may be similar to current mutations and some will be significantly different to current mutations.

The question is suggesting "labelling" tissue samples using unsupervised anomaly detection, and subsequently using the labels with a supervised algorithm to classify future samples. If this interpretation of the question is correct, then D makes sense

upvoted 1 times

hauhau 3 months ago

Selected Answer: AD

C : if the similarity differ in future from your sample, how do you predict?

upvoted 3 times

dish11dish 3 months ago

Selected Answer: AC

Let's understand Unsupervised anomaly

Unsupervised anomaly Detection involves an unlabeled dataset. It assumes that the majority data points in the unlabeled dataset are “normal” and it looks for data points that differs from the “normal” data points.

the option A and C only supports Unsupervised anomaly Detection method

upvoted 2 times

gudiking 3 months, 1 week ago

Selected Answer: AD

A - anomaly detection is used for detecting rare events, meaning it is expected that there are much less of those than of normal ones.
D - you expect the future mutations to be similar to the mutations you already have, so that you can detect them (pattern recognition)

upvoted 1 times

thisisapen25 3 months, 2 weeks ago

Selected Answer: AD

AD for sure

upvoted 1 times

Question #15 *Topic 1*

You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight 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

[Removed] Highly Voted 2 years, 11 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 issues. 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 tio storage
upvoted 46 times

msaqib934 10 months, 2 weeks ago
How to estimate average latency speed.
upvoted 1 times

devaid 4 months, 3 weeks ago
Cloud Monitoring?
upvoted 1 times

[Removed] Highly Voted 2 years, 11 months ago
D - speaks about near real-time approach. None other.
upvoted 16 times

musumusu Most Recent 4 days, 20 hours 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 1 times

techtitan 1 week, 1 day ago
Selected Answer: D
near realtime
upvoted 1 times

donbigi 3 weeks, 2 days ago
Selected Answer: D
the answer is AD
upvoted 1 times

korntewin 1 month, 3 weeks 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 commited mode (which is not present in the choices).

upvoted 1 times

PrashantGupta1616 2 months ago

Selected Answer: D

D make more sense

upvoted 1 times

jkhong 2 months, 2 weeks ago

Selected Answer: D

There are assumptions over the quality of data acceptable. If slight variations of the analytics against actual can be accepted, then D would be a good choice.

Many people chose B, but this also requires some form of waiting for the late data to arrive.

I think a combination of D and B can be applied, but for an intial fix, delaying the aggregation queries with D seems to make more sense. If the variance is small and the some late data leakage is acceptable, and we can remain as D.

If problems arise, we can always proceed to attempt B

upvoted 1 times

DGames 2 months, 2 weeks ago

Selected Answer: D

D make more sense

upvoted 1 times

Nirca 2 months, 2 weeks ago

Selected Answer: B

Answer is B.
if you use the Committed mode, records are available for reading immediately as you write them to the stream. It is recommended to use his mode for streaming workloads that need minimal read latency. And the requirement is near rea-time.
https://cloud.google.com/bigquery/docs/write-api#overview_of_the

upvoted 1 times

kennyloo 4 months, 1 week ago

D cos consistency could be caused by latency

upvoted 1 times

devaid 4 months, 3 weeks ago

D is the closest to follow in real time

upvoted 1 times

Parth_P 4 months, 3 weeks ago

Selected Answer: D

D is correct. The problem requirement is doing analytics on real-time data. You cannot do batch processing because the business requires it to be real-time even if it makes your job simpler, so B is incorrect. Other options are not streaming.

upvoted 1 times

ducc 6 months ago

Selected Answer: D

I vote for D, it make more sense to me

upvoted 1 times

rowan_ 6 months, 2 weeks ago

D just made sense to me but I wish I could actually give a reason why

upvoted 1 times

[Removed] 6 months, 2 weeks ago

Selected Answer: D

if you convert the individual messages to batch load at 10k per min. you'll get in 10 min. to the limit of load jobs per day (100k).
Source: https://cloud.google.com/bigquery/quotas#load_jobs

upvoted 5 times

noob_master 8 months, 2 weeks ago

Selected Answer: D

Answer: D. The only that describe a way to resolve the problem, with buffering the data.

(the question is possible old, the best approach would be Pub/Sub + Dataflow Streaming + Bigquery for streaming data instead near-real time)

upvoted 5 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 IIAM) 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 (83%) C (17%)

[Removed] Highly Voted 2 years, 11 months ago
Correct: A.
B - Wrong, permissions are on dataset level
C & D - won't say anything to influence the access level audit.
upvoted 24 times

[Removed] Highly Voted 2 years, 11 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 16 times

Aparna_22 Most Recent 5 days, 1 hour 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 2 days, 22 hours ago
Mine is also on 26 FEB
upvoted 1 times

AshokPalle 4 days, 14 hours ago
Mine is tomorrow
upvoted 1 times

niketd 2 weeks ago
"Discover what everyone is doing" will happen through Audit logs, hence correct answer is A
upvoted 1 times

Nirca 2 months, 2 weeks 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 3 months, 2 weeks ago
Selected Answer: A
A is correct.
upvoted 1 times

nkunwar 5 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 5 months, 1 week ago
Selected Answer: A
A for sure
upvoted 1 times

lalli117 5 months, 2 weeks ago
need to discuss permissions on the dataset level for each user , so i think A is appropriate answer.

upvoted 1 times

ducc 6 months ago

Selected Answer: A

A is correct

upvoted 1 times

PhuocT 6 months, 1 week ago

Selected Answer: A

A - check audit logs, to see what everyone is doing.

upvoted 1 times

rowan_ 6 months, 2 weeks ago

A. Use audit logs. You can access this from Google Admin console.

upvoted 1 times

anand81 7 months ago

Correct A

As question related to bigquery and what everyone is doing. this will we could get it from stackdriver logs

upvoted 1 times

SnowLearner 8 months, 2 weeks ago

Selected Answer: A

Answer is A

upvoted 1 times

som_420 8 months, 2 weeks ago

Selected Answer: A

most voted for A

upvoted 1 times

noob_master 8 months, 2 weeks ago

Selected Answer: A

Answer: A

You need to understand who is using and what they are using in BigQuery before elaborate any kind of policy. Stackdriver audit is good for it

upvoted 1 times

noob_master 8 months, 2 weeks ago

Answer: A

You need to understand who is using and what they are using in BigQuery before elaborate any kind of policy. Stackdriver audit is good for it

upvoted 1 times

Question #17 *Topic 1*

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and 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] Highly Voted 2 years, 11 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 51 times

[Removed] Highly Voted 2 years, 11 months ago

Correct : D

upvoted 16 times

samdhimal Most Recent 1 month, 2 weeks 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 month, 3 weeks 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 2 months, 2 weeks ago

Selected Answer: D

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

upvoted 1 times

assU2 3 months, 3 weeks 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 3 months, 3 weeks 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 3 months, 3 weeks ago

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

upvoted 2 times

achafill 4 months, 2 weeks ago

Selected Answer: D

Correct Answer : D

upvoted 1 times

nkunwar 5 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 6 months, 1 week ago
Anwer: D
upvoted 1 times

Asheesh1909 8 months, 3 weeks ago
Isn't it A and D both dataflow for reusable jobs and gcs for data peraistance?
upvoted 1 times

kmaiti 10 months, 3 weeks 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 1 year, 1 month ago
Selected Answer: D
This question if from Practice Test by Google, they gave D as right answer
upvoted 1 times

MaxNRG 1 year, 3 months ago
D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster.
A is not correct because the goal is to re-use their Hadoop jobs and MapReduce and/or Spark jobs cannot simply be moved to Dataflow.
B is not correct because the goal is to persist the data beyond the life of the ephemeral clusters, and if HDFS is used as the primary attached storage mechanism, it will also disappear at the end of the cluster's life.
C is not correct because the goal is to use managed services as much as possible, and this is the opposite.
E is not correct because the goal is to use managed services as much as possible, and this is the opposite.
upvoted 10 times

anji007 1 year, 4 months ago
Ans: D
i) re-use hadoop jobs with minimum cluster management - Cloud Dataproc
ii) Persist data beyond the life of cluster - use Cloud Storage.

So Cloud Dataproc with Google Cloud Storage connector to output results to cloud storage.
upvoted 3 times

Jeyakaran 1 year, 8 months ago
why not B?
upvoted 1 times

sumanshu 1 year, 8 months ago
Because in Question it's mentioned - They also want to be able to persist data beyond the life of the cluster.
If we use 'Persistent disk of cluster' - So once cluster get stopped or deleted - data get lost
upvoted 5 times

naga 2 years ago
correct D
upvoted 3 times

DeepakSai010101 2 years, 1 month ago
I think it should be B. The question targets as is migration of hadoop jobs. So the HDFS storage should be retained as is in persistent disk.
upvoted 2 times

balseron99 2 years ago
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.

In Dataproc, best practice is as soon as the job is done, the cluster is shutdown/deleted which will remove the data also in the cluster. They want storage beyond the life of the cluster which B option won't provide.
upvoted 6 times

Question #18 *Topic 1*

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 (90%) 10%

jvg637 Highly Voted 2 years, 11 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 no AI and no interactions with data from the observer.
upvoted 60 times

sergio6 1 year, 6 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 2 times

hellofrnds 1 year, 4 months ago
yes. multiclass classification model should be applied
upvoted 4 times

[Removed] Highly Voted 2 years, 11 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 36 times

musumusu Most Recent 4 days, 19 hours ago
Anwer: BCD
Things to understand:
Supervised learning will only predict the column that is labeled. In this case, there is not Fraud or not Fraud column inside which he will train on. So Option A, wrong.
option D: Supervised learning for column (transaction location) is possible as column exist to train on.
Option C: Custering N-type is possible and also an unsupervised learning to make cluster of similar pattern.
Option B: Its a weaker point here, User should be able to know which clusters are fraud in history. As it doesn't give enough information about past analysis whether user knows potential frauds or not. Ignore this option, if question asked for 2 right options only.
upvoted 1 times

ThorstenStaerk 1 month, 1 week ago
How is a machine supposed to know if a transaction is fraudulent if it only has ID, type, location and amount?
upvoted 2 times

chxnge 2 weeks, 6 days ago
Exactly, I do not think A or B can be correct here
upvoted 1 times

korntewin 1 month, 3 weeks ago
Selected Answer: ABC
ABC make more sense to me;
A --> we can use supervise learning to predict fraudulent
B --> we can use anomaly detection to guess fraudulent
C --> Clustering can also be used for customer segmentation

We have no need to predict location of the transation because it is already in the data.
upvoted 2 times

Kyr0 2 months ago

Selected Answer: BCD

BCD makes more sens to me

upvoted 1 times

ler_mp 2 months ago

Selected Answer: BCD

BCD makes more sense. B, C should not be controversial. For D vs E, in this use case D fits better than usage of reinfocement learning

upvoted 1 times

Chuckq 2 months, 4 weeks ago

AB Discarded no data to do fraude detection. Al the others are possible to do, but predict location is clearly a stupid biased idea. Probably a bad or tricky designed question.

upvoted 2 times

Catweazle1983 2 months ago

I was thinking the same regarding A and B. But probably B is correct because you can do unsupervised anomaly detection. https://cloud.google.com/blog/products/data-analytics/bigquery-ml-unsupervised-anomaly-detection

upvoted 1 times

ducc 6 months ago

Selected Answer: BCD

BCD is correct

upvoted 1 times

SnowLearner 8 months, 2 weeks ago

Selected Answer: BCD

BCD makes sense

upvoted 1 times

alecuba16 10 months, 1 week ago

Selected Answer: BCD

A, E ,F discarded, BCD are the possible cases

upvoted 2 times

sraakesh95 1 year, 1 month ago

Selected Answer: BCD

comment from @jvg637

upvoted 1 times

morpho4444 1 year, 3 months ago

Selected Answer: BCD

D are locations that we know, that makes it supervised.

upvoted 2 times

StefanoG 1 year, 3 months ago

Selected Answer: BCD

As wrote by RP123
B - Not labelled as Fraud or not. So Unsupervised.
C - Clustering can be done based on location, amount etc.
D - Location is already given. So labelled. Hence supervised.

upvoted 6 times

dsyouness 1 year, 3 months ago

Selected Answer: BCD

BCD makes more sense

upvoted 1 times

ArturVemado 1 year, 3 months ago

Selected Answer: BCD

RL does not make any sense, since it is used for learning to take action according to rewards.

upvoted 2 times

MaxNRG 1 year, 3 months ago

A is not an option, you do not have the labels to solve a supervised fraud detection problem.
E is not an option too, RL use cases are totally different.
D suites instead of E.

upvoted 1 times

Question #19 *Topic 1*

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] Highly Voted 2 years, 11 months ago
Answer: A
Description: First rule of dataproc is to keep data in GCS
upvoted 38 times

[Removed] Highly Voted 2 years, 11 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

Nirca Most Recent 1 month, 2 weeks ago
Selected Answer: A
A - is the right answer.
upvoted 1 times

DGames 2 months, 2 weeks ago
Selected Answer: A
A - dataproc - storage - cost effective is cloud storage
upvoted 1 times

devaid 4 months, 3 weeks ago
Selected Answer: A
Cloud Storage
upvoted 1 times

sankar_s 9 months ago
Selected Answer: A
Cloud Storage is google recommended one
upvoted 1 times

anji007 1 year, 4 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 5 times

sumanshu 1 year, 8 months ago
Vote for 'A"
upvoted 2 times

sumanshu 1 year, 7 months ago
A is correct because Google recommends using Cloud Storage instead of HDFS as it is much more cost effective especially when jobs aren't 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 5 times

anudeepgupta42 1 year, 10 months ago

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

upvoted 1 times

naga 2 years ago

Correct A

upvoted 2 times

Alasmindas 2 years, 3 months ago

The correct answer is A - Put the data into google cloud storage. This is what Google always recommend.

upvoted 2 times

AaronLee 2 years, 5 months ago

Because the cluster is going to end-of-life. So maybe don't need to put hot data in the disk. All can be put in the Cloud Storage. The answer is A.

upvoted 4 times

aadaisme 2 years, 9 months ago

I would go for D. First rule of dataproc is to keep data in GCS. Use persistant disk for high I/O data (hot data).

upvoted 4 times

MadHolm 2 years, 7 months ago

That would be correct if we wanted to maximize performance without sacrificing cost too much, but we want to minimize the cost. That's the only requirement, hence answer A.

upvoted 16 times

Question #20 *Topic 1*

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 2 years, 11 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 68 times

[Removed] Highly Voted 2 years, 11 months ago
Answer D.
https://cloud.google.com/pubsub/docs/troubleshooting#dupes
upvoted 41 times

niketd Most Recent 5 days, 1 hour ago
Selected Answer: D
No acknowledgment -> Answer B. Moderators please update your answer
upvoted 1 times

GCPpro 1 month, 2 weeks ago
D is the correct answer.
upvoted 2 times

AzureDP900 1 month, 3 weeks ago
D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.
upvoted 1 times

connorscion 2 months, 1 week ago
D, with a push delivery method if not acknowledge the sensor will keep sending the message.
upvoted 1 times

DGames 2 months, 2 weeks ago
Selected Answer: D
without ack message deliver multiple time.
upvoted 1 times

assU2 3 months, 2 weeks ago
Selected Answer: D
When you do not acknowledge a message before its acknowledgement deadline has expired, Pub/Sub resends the message. As a result, Pub/Sub can send duplicate messages.
upvoted 1 times

Jasar 3 months, 2 weeks ago
wy the B answer is marked as correct? it clear its some Ack problem the message goes back to the queu
upvoted 1 times

crismo04 5 months, 2 weeks ago
Selected Answer: D
If the answer had been B the messages had not reached the HTTP page since the push subscription requires an SSL certificate (https://cloud.google.com/pubsub/docs/push#:~:text=Endpoint%20URL%20(required)), so the answer must be D.

I think it is due to a bad response due to overcharge in the HTTPS page, which is sending a non correct status code (https://cloud.google.com/pubsub/docs/push#:~:text=VPC%20Service%20Controls.-,Receive%20messages,-When%20Pub/Sub)

upvoted 1 times

t11 6 months, 2 weeks ago

It has to be D.

With an out of date SSL cert, no message should be pushed to the HTTPS endpoint.

upvoted 1 times

KundanK973 8 months ago

Selected Answer: D

A message is resent if subscriber takes more than ackDeadline to respond.

upvoted 1 times

rr4444 8 months, 3 weeks ago

Selected Answer: B

Definitely not D

This is push, and acknowledgment is via HTTP response code, which is messed up by wrong certs

Non http code ACK by client is only for pull, which would be B

But it is push

upvoted 3 times

alecuba16 10 months, 1 week ago

Selected Answer: D

https://cloud.google.com/pubsub/docs/troubleshooting#dupes

upvoted 2 times

mallain 10 months, 4 weeks ago

Selected Answer: D

https://cloud.google.com/pubsub/docs/troubleshooting#dupes

upvoted 1 times

jagan79 10 months, 4 weeks ago

https://cloud.google.com/blog/products/data-analytics/handling-duplicate-data-in-streaming-pipeline-using-pubsub-dataflow

Pub/Sub guarantees at least once delivery for every subscription. This means that a message may be delivered more than once by the same subscription if Pub/Sub doesn't receive acknowledgement within the acknowledgement deadline. The subscriber may acknowledge after the acknowledgement deadline or the acknowledgement may be lost due to transient network issues. In such scenarios the same message would be redelivered and subscribers may see duplicate data. It is the responsibility of the subscribing system (for example Dataflow) to detect such duplicates and handle accordingly.

upvoted 2 times

910 11 months, 1 week ago

Selected Answer: D

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 2 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 efficiency?

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 (46%) D (39%) B (15%)

dg63 Highly Voted 2 years, 7 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 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 52 times

retax 2 years, 4 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 9 times

ralf_cc 1 year, 8 months ago

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

upvoted 8 times

omakin 1 year, 7 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 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?"
This means you need a "uniqueid" and timestamps to properly dedupe a data.

upvoted 6 times

Tanzu 1 year, 1 month 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 month 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 1 times

MaxNRG 1 year, 1 month ago

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

upvoted 1 times

MaxNRG 1 year, 1 month ago

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

upvoted 1 times

BigDataBB 1 year ago
Hi Max, I also think that the hash value would be worng because the timestamp is part of payload and is not written that the hash value is generated without the ts; but it 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 1 year, 2 months 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 timestmp.
D is better thatn B because maintaning a different table will reduce cost for hash computation for all historical data
upvoted 5 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
Description: Using Hash values we can remove duplicate values from a database. Hashvalues will be same for duplicate data and thus can be easily rejected.
upvoted 23 times

stefanop 10 months, 1 week ago
Hash values for same data will be the same, but in this case data contains also the timestamp
upvoted 2 times

DGames 2 months, 2 weeks ago
While calculating Hash value we exclude the timestamp.
upvoted 1 times

AshokPalle Most Recent 4 days, 14 hours ago
Just asked Chatgpt, it gave me option D
upvoted 1 times

musumusu 4 days, 19 hours ago
Answer B:
Option A: GUIDs can deduplicate the data but is expensive and good for multiple data processing.
Option B: Using hash function to authenticate the unique rows, this function can be applied directly in bigquery.
Option D, is complex and more expensive.

```
`CREATE TEMP FUNCTION hashValue(input STRING) AS (
CAST(FARM_FINGERPRINT(input) AS STRING)
);
```
upvoted 1 times

techtitan 1 week, 1 day ago
Selected Answer: A
A is prefereed way to generate unique identifier compared to hashing/indexing.
upvoted 1 times

hellobot 2 weeks ago
Best Answer = B
answer A, GUID will help in keeping all values unique but will not remove duplicates for transaction purpose.
answer D is not correct because maintaining external database is a good approach but it will not be efficient when there are large transactions that needs to be processed quickly and additional cost for database. I dont wanna pay for it.
Answer B is correct because: you just need to pass a check cloud function because i assume data has to be processed in real time, before loading the data to bigquery.
```import haslib
def compute_hash(data):
hash_value = hashlib.sha256(data.encode('utf-8'))
# detecting duplicates and ensure data integrity before storing in BQ
return hash_value.hexdigest()
```
upvoted 2 times

samdhimal 2 weeks, 5 days ago
D. Maintaining a database table to store the hash value and other metadata for each data entry would be the most efficient approach to deduplicate the data. This would allow you to compare the hash value of each new data entry with the values stored in the database, to quickly determine if it is a duplicate or not. Option A (Assigning GUID to each data entry) would be too time-consuming, Option B (computing the hash value and comparing it with all historical data) would not scale well with large amounts of data, and Option C (storing each data entry as the primary key in a separate database and applying an index) would not allow for efficient comparisons of the data.
upvoted 1 times

PolyMoe 1 month ago
Selected Answer: B
I would go for B. Seems to be most intuitive. If you want to know whether the input is already in the database, you need to compare it with the current content of the database. How ? by using hash
upvoted 2 times

Nirca 1 month, 2 weeks ago

Selected Answer: A

The simplest option is A. And simple is is an advantage.

upvoted 2 times

captainbu 1 month, 2 weeks ago

Answer D

upvoted 1 times

korntewin 1 month, 3 weeks ago

Selected Answer: A

I would vote on A, assuming that the re-transmission will submit the same GUID.
D can be confused and can cause duplication or incorrect data on two scenarios:
1. If the hash includes transmission timestamp, every data transmission will have different hash --> data re-transmission will have different hash --> duplicate
2. If the hash data doesn't include transmission timestamp --> the inventory data on different time can have the same hash --> no data with the same value can be inserted to the DB!

upvoted 2 times

Jackalski 2 months ago

Selected Answer: D

goal is to avoid duplicates in case resending the same information.
Assuming using key-value store to handle duplicated values correctly - would failed on insert etc (however not shared details here in this question)
tricky question is what are definition of GUID versus hash.
calculated hash on data sent first time and resent again can be the same assuming timestamp of datapoints inslide payload are not changed.
timestamp of transmission is another story and I dont assume it is consider as value here (assumption to resend).
GUID usually means to use UUID function which is random - no option to validate uniqness
if GUID is build on specific way - than it would require complex logic ...
Hash gives easy way to calculate it on full payload received
Question was about most effective ...

so I vote on D

upvoted 2 times

PrashantGupta1616 2 months ago

Selected Answer: D

Answer is D.

upvoted 1 times

jkhong 2 months, 1 week ago

Selected Answer: A

Initially misinterpreted the answers.. Assumed all of them to be very different approaches.

But essentially the question is a comparison of different ways to generate unique identifiers for entries. Ways to generate unique identifiers:
1. Hashing
2. indexing
3. GUID

There are storage and compute considerations that we have to make while choosing these methods.

GUID does not require calculation again compared to hashing or indexing. However storage of GUID is heavy compared to the two.

Since we are only concerned of efficiency, GUID is the fastest.

We also do not need to store as we can implement a deduplication when our window closes. We don't need to store the unique values in a database in this case, as aggregations will be performed after the windowing period ends.

upvoted 1 times

martarroyo 2 months, 2 weeks ago

Selected Answer: 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 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

DGames 2 months, 2 weeks ago

Selected Answer: D

Answer is D- Calculating Hash value (exclude timestamp) which help to find duplicate data already processed and eliminate.

upvoted 1 times

hauhau 3 months ago

Selected Answer: A

Guid +timestamp

upvoted 1 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 (100%)

Rajokkiyam Highly Voted 2 years, 11 months ago
Answer should be D.
upvoted 41 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
Description: Datalab provides Jupyter for this kind of work
upvoted 13 times

PolyMoe Most Recent 1 month ago
Selected Answer: D
B is not relevant at all
upvoted 1 times

AzureDP900 1 month, 3 weeks ago
D
Cloud Datalab is **packaged as a container and run in a VM (Virtual Machine) instance.**
VM creation**, running the container in that VM, and establishing a connection from your browser to the Cloud Datalab container, which allows you to open existing Cloud Datalab notebooks and create new notebooks**. Read through the introductory notebooks in the `/docs/intro` directory to get a sense of how a notebook is organized and executed.

Cloud Datalab uses **notebooks instead of the text files containing code. Notebooks bring together code, documentation written as markdown, and the results of code execution—whether as text, image or, HTML/JavaScript**.

Reference:
https://cloud.google.com/datalab/docs/quickstarts
upvoted 2 times

DGames 2 months, 2 weeks ago
Selected Answer: D
Answer - D, Its clear mention Data Scientist have task of visualization. that possible in Datalab
upvoted 2 times

Atnafu 3 months ago
D
Datalab before it get deprecated now Vertex AI
https://cloud.google.com/datalab/docs#:~:text=Datalab%20is%20deprecated,Google%20Charts%20APIs
upvoted 5 times

assU2 3 months, 1 week ago
Selected Answer: D
The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks.

Definitely not the Shell.
And its Vertex AI Workbench now.
upvoted 6 times

gudiking 3 months, 1 week ago

Selected Answer: D
D because of visualization tasks, they can't be done in the shell, right?
upvoted 3 times

wubston 3 months, 2 weeks ago
How in the hell isn't this D?
upvoted 1 times

crisimenjivar 6 months, 1 week ago
Answer: D
upvoted 1 times

Megmang 6 months, 2 weeks ago
Selected Answer: D
Datalab is the solution for this
upvoted 1 times

nexus1_ 8 months, 2 weeks ago
Selected Answer: D
Answer is D
upvoted 1 times

alecuba16 10 months, 1 week ago
Selected Answer: D
Cloud Datalab -> AI Notebooks -> Vertex AI Workbench
upvoted 2 times

MichaelOcto 12 months ago
Selected Answer: D
Datalab was made for this
upvoted 1 times

davidqianwen 1 year, 1 month ago
Selected Answer: D
Answer should be D.
upvoted 1 times

samdhimal 1 year, 1 month ago
correct answer -> Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

Google says:
Use Cloud Datalab to easily explore, visualize, analyze, and transform data using familiar languages, such as Python and SQL, interactively.
Reference: https://cloud.google.com/datalab/docs
This is exactly the tool needed in this scenario
But what about her laptop slowing her down?
Well, Cloud Datalab is packaged as a container and run in a VM (Virtual Machine) instance.
Reference: https://cloud.google.com/datalab/docs/concepts/key-concepts
upvoted 5 times

medeis_jar 1 year, 1 month ago
Selected Answer: D
But there is a new product that evolved from Cloud Datalab
Cloud Datalab -> AI Notebooks -> Vertex AI Workbench
upvoted 5 times

Question #23 *Topic 1*

You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and 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] Highly Voted 2 years, 11 months ago
Answer: B
upvoted 27 times

[Removed] 2 years, 11 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] Highly Voted 2 years, 11 months ago
Answer: B
Description: Pubsub for realtime, Dataflow for pipeline, Bigquery for analytics
upvoted 22 times

DGames Most Recent 2 months, 2 weeks ago
Selected Answer: B
GCP recommend best practice for streaming data pipeline as option B - pub/sub, dataflow & Bigquery
upvoted 1 times

Nirca 2 months, 2 weeks 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 2 months, 4 weeks ago
B aris tqve yleebo
upvoted 1 times

devaid 4 months, 3 weeks ago
Selected Answer: B
B of course
upvoted 1 times

Dip1994 6 months, 4 weeks ago
B is the correct answer
upvoted 1 times

noob_master 8 months, 2 weeks ago
Selected Answer: B
Answer: B

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

nexus1_ 8 months, 2 weeks ago
Definitely B
upvoted 1 times

vw13 10 months, 2 weeks ago
Selected Answer: B

B is the only option for real time process & analysis

upvoted 1 times

devric 10 months, 4 weeks ago

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

upvoted 1 times

samdhimal 1 year, 1 month 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 by Cloud Dataflow, a simplified stream and batch data processing solution, Cloud Datastore will save data to be displayed directly into the UI of the 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 1 times

MaxNRG 1 year, 3 months ago

B as the need to ingest it, transform and store the Cloud Pub/Sub, Cloud Dataflow, BigQuery is ideal stack to handle the IoT data.
https://cloud.google.com/solutions/mobile/mobile-gaming-analysis-telemetry
Option A is wrong as the Datastore is not an ideal ingestion service.
Option C is wrong as Cloud Storage is not an ideal ingestion service and Dataproc is not a data warehousing solution.
Option D is wrong as Cloud SQL is not a data warehousing solution.

upvoted 1 times

anji007 1 year, 4 months ago

Ans: B
All options are speaking about cloud storage, but need pub/sub in between to take streaming (IoT) data till storage/ into the cloud.

upvoted 2 times

sandipk91 1 year, 6 months ago

B for sure

upvoted 2 times

sumanshu 1 year, 8 months ago

Vote for B

upvoted 2 times

amarkan 2 years, 1 month ago

no doubt B

upvoted 3 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

jvg637 Highly Voted 2 years, 11 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 28 times

jkhong 2 months, 2 weeks 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 2 times

assU2 3 months, 1 week ago
"You want to minimize the migration effort without making future queries computationally expensive." Nothing about storage price.
upvoted 2 times

[Removed] Highly Voted 2 years, 11 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 20 times

Tanzu 1 year, 1 month ago
Due to the hard limitations of bq, Not E is the simple answer by the way!
upvoted 1 times

techtitan Most Recent 1 week, 1 day 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 1 times

Jackalski 2 months 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 3 times

Nirca 2 months, 2 weeks 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 2 months, 2 weeks ago

Selected Answer: D

D- Make more sense because then E because we are create new table again need to over head to update everywhere table name in application

upvoted 1 times

odacir 2 months, 3 weeks ago

Selected Answer: E

If I do this in my work, E is my answer 100% Otherwise is to complicate.

upvoted 1 times

beowulf_kat 3 months, 4 weeks ago

Selected Answer: C

C - as there are no structural changes to the original table, and no increase in storage/compute costs are there. whenever analysis is required, the view with casted datatypes will suffice.

upvoted 3 times

NicolasN 3 months, 3 weeks ago

We are interested in query performance ("without making future queries computationally expensive") so a common view (not materialized) will have the computational cost of performing the cast every time it is queried.

upvoted 6 times

KundanK973 8 months ago

Selected Answer: E

E is the best way to implement it.

upvoted 4 times

beowulf_kat 3 months, 4 weeks ago

The problem with E is, the new table will have to be re-created each time the original table is updated. It also increases storage costs as it is basically a replica with changed data types. I think C makes more sense hence.

upvoted 1 times

NicolasN 3 months, 3 weeks ago

Storage cost isn't a factor for this question. And even if it was, since "In the future, new data is loaded into the table NEW_CLICK_STREAM" the table CLICK_STREAM could be deleted. The DROP TABLE statement is free.

upvoted 1 times

thapliyal 8 months ago

D should be the correct answer. E is not because it is mentioned in the question that we have do less migration effort and D also supports SCD TYPE 2.

upvoted 1 times

JimJam1 9 months, 2 weeks ago

none of these, i would use cloud dataprep which is perfect for the task.

upvoted 3 times

richardchan66 10 months, 3 weeks ago

E : Export and load data into a new table
You can also change a column's data type by exporting your table data to Cloud Storage, and then loading the data into a new table with a schema definition that specifies the correct data type for the column. You can also use the load job to overwrite the existing table.

Advantages
You are not charged for the export job or the load job. Currently, BigQuery load and export jobs are free.
If you use the load job to overwrite the original table, you incur storage costs for one table instead of two, but you lose the original data.
Disadvantages
If you load the data into a new table, you incur storage costs for the original table and the new table (unless you delete the old one).
You incur costs for storing the exported data in Cloud Storage.

upvoted 4 times

devric 10 months, 4 weeks ago

E. D adds processing reading each row.

upvoted 1 times

Tanzu 1 year, 1 month ago

A does the job. Exporting from bq to gcs is free but storage costs. Sending from computer to bq, again, is ok but it is still hard to eliminate. A few days ... sounds like that way is not possible some how way.

B, adding a column w/o deleting the old one increases storage and computation costs further ops. Not valid

C, creating a view (means source table has to remain) is increasing storage and computation costs, too.

D, adding 2 column is non sense.

So E is much more sense. More complicated but less computational cost. The tradeoff in bq :)

upvoted 6 times

AlRaf 6 months, 2 weeks ago

uhh.. for C, "creating a view is increasing storage"? Isn't view just a query and results are not persisted on any disk? i was guessing C to be the right answer.. can anyone help, what am i missing?

upvoted 5 times

vetaal 3 months, 3 weeks ago

C won't increase storage, but C would be computationally expensive, as casting would have to be performed at runtime, each time the view is accessed.

upvoted 1 times

samdhimal 1 year, 1 month ago

I dont think any of the answer here are correct. Because google says that if you want to Change a column's data type then there are two options:
1)Using a SQL query: If needed simplicity and ease of use. Also if you are less concerned about costs.
2)Recreating the table: If More concerned about costs, and you are less concerned about simplicity and ease of use.

We aren't concern about current cost so we go a head and go with (1) Using a SQL query. This option basically creates a new table with type casted datatype for cols you want as well as there is the other original table. Yes we dont have to currently worry about cost but we do have to worry for future use. If we have two existing tables its costs more money then just having one so we have a conflict here.

upvoted 1 times

samdhimal 1 year, 1 month ago

Closest to correct answer is 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.

With this option there isn't any mention about deleting old table so that you don't have to pay for that table's storage anymore.... Not sure...

Reference: https://cloud.google.com/bigquery/docs/manually-changing-schemas#changing_a_columns_data_type

upvoted 1 times

Austinaws 1 year, 1 month ago

E. Column type can not be changed and column needs to casting and loaded into new table using either sql query or import/export features

upvoted 1 times

Nico1310 1 year, 1 month ago

Selected Answer: E

Answer is E

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%)

jvg637 Highly Voted 2 years, 11 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 42 times

MaxNRG Highly Voted 1 year, 3 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 13 times

Jackalski Most Recent 2 months ago

Selected Answer: D

"advanced log filter" is the key word here, all other options push all data ...

upvoted 2 times

Jasar 3 months, 2 weeks ago

Selected Answer: D

D is the best choice

upvoted 1 times

alecuba16 10 months, 1 week 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 10 months, 4 weeks ago

Selected Answer: D

D. Option B doesn't make sense

upvoted 2 times

samdhimal 1 year, 1 month 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 D.

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 for 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 out 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 1 times

santoshindia 1 year, 1 month ago
Selected Answer: D
explained by MaxNRG
upvoted 3 times

medeis_jar 1 year, 1 month ago
Selected Answer: D
as explained by MaxNRG
upvoted 3 times

rebootshen 1 year, 2 months ago
C is simple, the sink has filter option, "In the Build inclusion filter field, enter a filter expression that matches the log entries you want to include"
upvoted 1 times

StefanoG 1 year, 3 months ago
Selected Answer: D
as wrote MaxNRG
upvoted 3 times

anji007 1 year, 4 months ago
Ans: D
Notification to the monitoring can be sent via Pub/Sub and it is one off use case of Pub/Sub. To deliver log/ notification to Pub/Sub need a sink followed by a filter to look only for append data to BigQuery.
So
A is out as no Sink.
B is not correct as it is saying sink export to BigQuery.
C is not talking about filtering after setting up sink.
upvoted 3 times

fire558787 1 year, 6 months ago
It's definitely D. From: https://cloud.google.com/logging/docs/reference/v2/rest/v2/projects.sinks "a sink used to export log entries to one of the following destinations in any project: a Cloud Storage bucket, a BigQuery dataset, a Pub/Sub topic or a Cloud Logging log bucket. A logs filter controls which log entries are exported".
B is wrong because why would you export to BigQuery again if you want to be notified by your monitoring tool?
upvoted 5 times

raf2121 1 year, 8 months ago
I would choose D
Reference : https://cloud.google.com/logging/docs/reference/v2/rest/v2/projects.sinks
Only Key word that was sure initially is "Stackdriver API" (and I found the link for it too https://cloud.google.com/monitoring/api/v3)
upvoted 2 times

sumanshu 1 year, 8 months ago
Vote for 'D' - because we need to put filter for particular table, also we need monitoring tool
upvoted 2 times

jasper_430 1 year, 9 months ago
I would choose D since A and B don't make any notification and C doesn't make a filter of the result.
upvoted 1 times

Ette 1 year, 9 months ago
I would choose C: you can set a filter 'protoPayload.metadata.jobInsertion.job.jobConfig.queryConfig.destinationTable=...'
upvoted 4 times

moonlightbeamer 1 year, 8 months ago
i agree, in the link it reads: "Depending on the log sink's configuration, which includes filters and a destination" so you should be able to config the filter when sinking to destination - Pub/Sub. it would be silly it can't do simple things like this. whileas "project sink" in Answer D, what is that?
upvoted 1 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 (83%) B (17%)

jvg637 Highly Voted 2 years, 11 months ago
The Answer should be B. The Dataflow developer role will not provide access to the underlying data.
upvoted 64 times

cleroy 2 years, 11 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 2 years, 7 months ago
Service account is between applications and non human entry.
upvoted 30 times

Tanzu 1 year, 1 month ago
u can enable a service account as user so that externals can use to login
upvoted 1 times

Tanzu 1 year, 1 month 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 2 times

willymac2 8 months, 3 weeks 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:

ttps://cloud.google.com/dataflow/docs/concepts/access-control#example_role_assignment
upvoted 1 times

willymac2 8 months, 3 weeks 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 1 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: Provides the permissions necessary to execute and manipulate Dataflow jobs.
upvoted 16 times

NamitSehgal Most Recent 1 month ago
B is the answer , developer role is sufficient to secure data
upvoted 1 times

PolyMoe 1 month ago

Selected Answer: B
of course it is answer B. Providing dataflow developer role will not give access to the underlying data !
upvoted 1 times

desertlotus1211 1 month, 2 weeks ago
The external consultant will not access the data. The dataflow pipeline will retrieve the data wherever it's located based in a service account used by dataflow to access the data. So the consultant will never see the data... The consultant is not the one running the job...

So maybe Answer B is valid...
upvoted 1 times

KGCP25 1 month, 2 weeks ago
Selected Answer: B
Example role assignment
To illustrate the utility of the different Dataflow roles, consider the following breakdown:

The developer who creates and examines jobs needs the roles/iam.serviceAccountUser role.
For more sophisticated permissions management, the developer interacting with the Dataflow job needs the roles/dataflow.developer role.
They need the roles/storage.objectAdmin or a related role to stage the required files.
For debugging and quota checking, they need the project roles/compute.viewer role.
Absent other role assignments, this role lets the developer create and cancel Dataflow jobs, but not interact with the individual VMs or access other Cloud services.
upvoted 1 times

Aamir185 2 months, 1 week ago
Selected Answer: D
Privacy is the key here
upvoted 2 times

Asheesh1909 2 months, 3 weeks ago
Selected Answer: D
I think it should be D, if we give the consultant developer role he can print the data while developing jobs so he can have indirect access to the data so D sounds to be a good option.
upvoted 4 times

Naazid 3 months ago
should be B
upvoted 1 times

funnybunny11 3 months ago
Selected Answer: D
D make sense to me
upvoted 1 times

assU2 3 months, 1 week ago
Selected Answer: D
The question is - "How should you maintain users' privacy?" Not what permission external developers need.. So, D.
upvoted 3 times

beowulf_kat 3 months, 3 weeks ago
Selected Answer: D
The consultant will need access to the underlying data and transformed data to see if the transformation logic is coded correctly. This can be best handled by anonymizing the data.
upvoted 7 times

TOXICcharlie 5 months, 2 weeks ago
Selected Answer: D
The key ask here is maintaining the user's privacy. So only way to do is annonymize data.
Developer role as stated in option B will only limit the access for the consultant to use dataflow fully. That does not help to prevent him from sampling the actual dataset where sensitive data is. Also the other point about giving him access to a separate project is quite valid. In an internal project we do not want an external consultant to get access to other parts of GCP. So my answer is D.
upvoted 6 times

ducc 6 months, 1 week ago
Selected Answer: D
D is correct
B is not Correct because the Consultant need to test the pipeline
upvoted 1 times

aaristizabal 7 months ago
the answer should be C. The Dataflow developer role will not provide access to the underlying data.
https://cloud.google.com/dataflow/docs/concepts/access-control#example_role_assignment
upvoted 2 times

noob_master 8 months, 2 weeks ago

Selected Answer: D

Answer: D

It is possible to code an Apache Beam Pipeline without permission on GCP (ex: Dataflow developer to run the pipeline on cloud), but the external consultant needs access to the data to test his pipeline. The problem is with cloud access and sensitive information, that is resolved in D

upvoted 1 times

nexus1_ 8 months, 2 weeks ago

Selected Answer: D

Answer is D

upvoted 1 times

tavva_prudhvi 8 months, 2 weeks ago

Wow, could be a better idea if you could provide some explanation.

upvoted 3 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 (85%) D (15%)

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: Best Choice out of given options.
upvoted 27 times

[Removed] Highly Voted 2 years, 11 months ago
Should be B
upvoted 15 times

jin0 Most Recent 1 week ago
Selected Answer: D
wrong question. there are two answers B, D
upvoted 1 times

jin0 1 week ago
B. Combine highly co-dependent features into one representative feature.
-> Explainable feature should be dependent from each other feature. expecially not deep leanring. 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 1 times

AzureDP900 1 month, 3 weeks 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 1 week ago
But, if there are null datas more than 50% then, it should be eliminated because there are two ways to train the model. first, remove records 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 answer in this question B, D I think
upvoted 1 times

Thasni 3 months, 1 week ago
I have a doubt, instead of combining highly corelated features why cant we remove corelated features which may give much more simplified dataset?
upvoted 1 times

noob_master 8 months, 2 weeks ago
Selected Answer: B
Answer: B

Data that is co-dependent is high corelated is some kind of reduldant information in some cases. If the features x1, x2 and x3 are x2 = x1 + 1 and x3 = 2*x1, for example, x2 and x3 are reduldant because can be explained with x1 feature, so can be excluded of the the model. Other option is to group this features. There is a lot of ways to resolve, but the main ideia is to use data engineer in co-depedent features to reduce the number of features in the model
upvoted 2 times

Ishiske 8 months, 3 weeks 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 it's accuracy.

upvoted 1 times

Yad_datatonic 9 months, 2 weeks ago

Answer: B

upvoted 1 times

alecuba16 10 months, 1 week ago

Selected Answer: B

Co-dependent -> correlated -> correlated info = already present info in other variable.

upvoted 2 times

pamepadero 1 year 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 9 months, 1 week ago

Yes. But nearly 50% of the non-null data still seems to be a lot to ignore.

upvoted 1 times

exnaniantwort 1 year, 1 month 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 1 year, 1 month ago

Selected Answer: B

For me the best option is B

upvoted 1 times

aet2123 1 year, 2 months ago

Selected Answer: B

B is the correct option

upvoted 1 times

BigQuery 1 year, 2 months ago

Selected Answer: D

D is correct

upvoted 1 times

BigQuery 1 year, 2 months ago

Sorry B is correct. ignore my previous comment

upvoted 3 times

anji007 1 year, 4 months ago

Ans: B
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 8 times

sumanshu 1 year, 8 months ago

Vote for 'B'

combining features to createte a new feature is a step of "Feature construction"
or
decomposing or splitting features to create new features.

Ideally, PCA should be apply if we want to reduce the dimension.

Removing those columns / features - where Data is miss > 50% (may improve the speed) - but will decrease the accuracy as well. So instead of dropping features where we have missing data, we need to impute something

upvoted 6 times

sumanshu 1 year, 8 months ago

'D" could not be the answer - If There is NO missing value OR if all features have missing values > 50% does it means we drop everything - NO , right ?.

So, B is more appropriate

upvoted 2 times

awssp12345 1 year, 8 months ago

What is PCA?

upvoted 1 times

sumanshu 1 year, 7 months ago

PCA (principal component analysis) - a decomosition tehcnique apply on numerical data....this algo automatically check which features (columns) are more related with output and which are less related...and it keep only related one....(i.e. reduce dimension) or number of features...generally applied...when we have thouusands of columns (features)

upvoted 2 times

sumanshu 1 year, 7 months ago

Otheriwse...we need to manually check all columns (using plotting the graphs) or using mathematical formula..

upvoted 2 times

moonlightbeamer 1 year, 8 months ago

Answer: A

directly eliminate redundant feature

upvoted 1 times

moonlightbeamer 1 year, 8 months ago

sorry, change mind, B. if corelated to other feature, can remove, but A is corelate to label...

upvoted 2 times

Question #28 *Topic 1*

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being 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 (73%) D (27%)

arthur2385 Highly Voted 5 months, 4 weeks 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 8 times

maxdataengineer Highly Voted 4 months, 3 weeks 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 featueColum 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_
upvoted 5 times

maxdataengineer 4 months, 3 weeks ago
The answer is B
upvoted 1 times

jkh_goh Most Recent 1 month, 1 week 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 2 months, 4 weeks 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 4 weeks, 1 day 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 1 times

gcm7 4 months, 2 weeks ago
Selected Answer: B
reading only relevant cols
upvoted 3 times

devaid 4 months, 3 weeks ago
Selected Answer: D
Answer is D, apparently.
upvoted 1 times

Kowalski 5 months ago

Answer is Use .fromQuery operation to read specific fields from the table.

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).

Reference:
https://cloud.google.com/bigquery/docs/best-practices-costs#avoid_select_

upvoted 2 times

crismo04 5 months, 2 weeks ago

Selected Answer: D

D is correct, you want to increase the performance, not the cost

upvoted 2 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: B

B. Use .fromQuery operation to read specific fields from the table.

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: B

B is correct

upvoted 2 times

Question #29 *Topic 1*

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] Highly Voted 2 years, 11 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 sensorid and timestamp as rowkey
upvoted 29 times

[Removed] Highly Voted 2 years, 11 months ago
Answer D
upvoted 17 times

AzureDP900 Most Recent 1 month, 3 weeks 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 2 months, 2 weeks ago
Selected Answer: D
#<sensorid>#<timestamp> ------> low cardinality # high cardinality
This is current Bigtable Best Practice (to avoid Hotspots on the inserts)
upvoted 1 times

maxdataengineer 4 months, 3 weeks 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 will 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 5 months, 3 weeks 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 6 months, 1 week ago
ANSWER: D
upvoted 1 times

som_420 8 months, 2 weeks ago
Selected Answer: D
Answer is D
upvoted 1 times

samdhimal 1 year, 1 month 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 1 year, 4 months ago

Ans: D

upvoted 2 times

sumanshu 1 year, 8 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 1 year, 7 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 ago

Correct D

upvoted 2 times

NamitSehgal 2 years, 2 months 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 1 year, 1 month ago

reverse TS or hashing is not always first choice or better. never.

upvoted 1 times

Radhika7983 2 years, 3 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 becuase
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. Using 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 2 years, 3 months ago

answer would be D to avoid hotspoting..

upvoted 2 times

ch3n6 2 years, 8 months ago

correct: D
why not C? Using 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. https://cloud.google.com/bigtable/docs/schema-design#row-keys

upvoted 4 times

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

HectorLeon2099 Highly Voted 2 years, 2 months 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 85 times

Tanzu 1 year, 1 month 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 5 times

cetanx 1 month 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 1 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
Description: Easy and it won't affect processing
upvoted 37 times

StefanoG 1 year, 5 months ago
Google Cloud Dataproc is not an analytic tool
upvoted 6 times

dambilwa 2 years, 8 months ago
Agreed- Option[D] is most appropriate in this scenario
upvoted 5 times

StefanoG 1 year, 5 months ago
So I vote for B
upvoted 6 times

[Removed] 1 year, 4 months ago
27 up vote for a wrong ans!!
Why do you need dataproc for MySQL dump?!
upvoted 24 times

Lestrang Most Recent 4 weeks, 1 day ago
While I do not have a solid answer, I'd like to offer a different way to look at the problem.
Maybe we do not need to provide the analytical platform in the solution, after all, they did not mention what kind of analytics they used to use previously.

Maybe B and D are just targets, where data will be held.
What am trying to say is, just cause you processed the data via dataprocs, doesn't mean you can't write it to bigquery or any other analytical tool. Hell you can even analyze it with Spark if you want (not the best idea but if you want sure.)

I cannot seem to accept B as the answer, it will definitely violate the "doesn't affect operations" criteria.

but I am not experienced enough to say for sure.

upvoted 1 times

niketd 2 weeks ago
Since they take an overnight dump, the ETL process can happen using the dump. Moreover, google only recommends to use Dataproc when one wants to migrate their existing on-prem Hadoop/Spark workloads with minimal effort.
upvoted 1 times

desertlotus1211 1 month, 2 weeks ago
How could you say D is the answer? Dataproc is used to easily create, manage, and scale Hadoop and Spark clusters in GCP. this is SQL...

Answer is B
upvoted 2 times

noonting 1 month, 3 weeks ago
Selected Answer: D
Minimal impact on operations is key.
upvoted 1 times

Jackalski 2 months ago
Selected Answer: D
none of the option is really good ....
if need to choose - I vote on D to avoid impact on source system
option B (as all other options) will add load into mySQL to extract data from db directly - so rejected
upvoted 1 times

PrashantGupta1616 2 months ago
Selected Answer: B
its B...
upvoted 1 times

PrashantGupta1616 2 months ago
Selected Answer: B
Bigquery is most suitable for analytical purposes and the question is asking about 'minimal impact' on current DB
upvoted 3 times

Brillianttyagi 2 months ago
Selected Answer: D
D makes more sense to me as in B ETL consumes lot of MySQL resources, to read the data, as per question MySQL is under pressure already, In ques they are also talking about the mysql dump then dump can be used to process data or do analytics as we don't have to increase workload.
"They don't give name of ETL tool in B so it can also be Dataproc and in D There is also dataproc and in D we are using google services like Cloud sql and dataproc and we can also use bigquery although it is not mention in the ques but they also not denied that we can not use bigquery in option D.
So thing about it it a situation to confuse all of us by giving bigquery in option and analytics in question.
upvoted 2 times

Krish6488 2 months, 1 week ago
Its B. Close your eyes and choose BigQuery for analytical processing. If operations are not to be impacted, choose either a file extract from the source system or a jdbc connection to run a batch load to GCS/BQ directly in a window where operations are least impacted. Processing delta data in a nightly batch wont take more than 15-20 mins even if the changes dealt with are a million for a given window
upvoted 2 times

Rodolfo_Marcos 2 months, 2 weeks ago
Selected Answer: D
D of course
upvoted 2 times

odacir 2 months, 3 weeks ago
Selected Answer: B
The only option viable is B. Analytics in google World it's BigQuery. The only thing is the heavy load of moving data from on premise to BQ, but that is only once as a big task, after that it's only small chunks and all the analytical power will be out of the DB...
upvoted 1 times

Jay_Krish 3 months, 1 week ago
What about performance issues of choosing option B when you read heavily from the source DB to process data and ingest into BigQuery? As there's a nightly backup already taken every night won't this be an easier solution
upvoted 4 times

odacir 2 months, 3 weeks ago
I think the same as you, it's overload the system, then it's better to use backups. But I'm not .
upvoted 1 times

Jay_Krish 3 months, 1 week ago
Option D seems logical but i'm not sure

upvoted 2 times

assU2 3 months, 1 week ago

Selected Answer: B

Its BigQuery because this is Googles data warehouse (storage + query and analysis)

upvoted 1 times

kennyloo 4 months, 1 week ago

B - BigQuery is all the while for Analytics purpose

upvoted 1 times

maxdataengineer 4 months, 3 weeks ago

Selected Answer: B

Discarded
A -> You are not even using cloud services in this solution
C -> You are not even using cloud services in this solution.
D -> Dataproc is an ETL tool, not an analytical one.

B -> BEST CHOICE. BigQuery is a warehouse for analytical workloads, and MySQL is for transactional workloads. You take the data from MySQL to BigQuery apply the necessary transformations and make the analysis there.

upvoted 4 times

devaid 4 months, 3 weeks ago

Selected Answer: B

Answer: B. D don't make sense to me, cloud sql is not designed for analytics, and B mention the usage of an ETL tool to store clean and transformed data (ready for analysis) to a datawarehouse that is Big Query.

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

VishalB Highly Voted 2 years, 7 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 64 times

BigQuery 1 year, 2 months 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 correct 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 to 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 Json mapping.

In this case it is Code change to new version. so, Update with Json mapping. Simple as that.

All the Best Guys.
upvoted 21 times

BigQuery 1 year, 2 months 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 5 times

anji007 11 months, 1 week 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 3 times

maxdataengineer 4 months, 3 weeks ago
JSON Mapping is a way to solve compatibility issues when updating
upvoted 1 times

maxdataengineer 4 months, 3 weeks 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 3 months, 1 week ago
You do not want to lose any data when making this update - is the core problem. You are doing it ANYWAY.
upvoted 1 times

VishalB 2 years, 7 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 1 year, 1 month 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 4 times

sergio6 1 year, 4 months ago
A is incorrect because updating pipeline does not include any drain flag
upvoted 2 times

Tanzu 1 year, 1 month 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 4 months, 3 weeks ago
Yes but the compatibility problem will still be there, stopping the pipeline does not solve that
upvoted 1 times

sergio6 1 year, 5 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 1 year, 1 month ago
drain is in the guide ...stopping-a-pipeline. Just ...updating-a-pipeline is not enough to evaluate this question.

that's why drainnin 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 1 year, 4 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 1 year, 1 month 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 4 months, 3 weeks ago
B is a way to solve compatibility issues
upvoted 2 times

[Removed] Highly Voted 2 years, 11 months ago
Correct B - https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#preventing_compatibility_breaks
upvoted 22 times

[Removed] 2 years, 11 months ago
Changing the pipeline graph without providing a mapping. When you update a job, the Dataflow service attempts to match the transforms in 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 2 years, 10 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 8 times

sergio6 1 year, 4 months ago
Canceling the job will cause data loss, against the requirement
upvoted 1 times

musumusu Most Recent 4 days, 13 hours ago

Answer C:
According to latest docs you can update an existing batch pipleine,
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python_1

create new one with same subscription. And start getting the data as before and once its ready, remove earlier one.
I will go with C

upvoted 1 times

musumusu 4 days, 13 hours ago

You can't *update * the pipeline. Please check the docs

upvoted 1 times

SuperTee 6 days, 15 hours ago

Selected Answer: D

Explained below.

upvoted 1 times

SuperTee 6 days, 15 hours ago

This is a tricky question.
1. Because of incompatibility, you just cannot "UPDATE", plus, for update method, there is never a "DRAIN" flag.
2. When you talk about a flag, it is like a parameter attached to the "UPDATE" method, it is not the same as stopping the pipeline with Drain method.
3. Option D is not the best because you will need to reprocess duplicated old data, HOWEVER, you will not LOSE any data. It says must not lose any data, it did not say must not lose "in-flight" data.
4. I see many comments jumped on the "CANCEL" word in C&D, but don't forget, although you did not "DRAIN", the data will not be lost with new subscription.
5. In this particular question, I believe general consensus might not be a wise one.

upvoted 1 times

midgoo 6 days, 23 hours ago

Selected Answer: B

There is no such thing as 'drain flag' for update. It is for stopping the pipeline. It could work too but the answer has to be rephrased

upvoted 1 times

hellobot 2 weeks ago

Correct Option: A
But practically, D can be correct too we will not lose any data here as per requirement, only problem here is to update all connected services to use the new subscription that were getting messages from pub/sub topic.
Option B: is not good approach as json transform mapping is used for specifying the mapping between input data and output data, which does not address our issue because we an have version issue and it does not ensure that data will not be lost during update.

upvoted 1 times

waiebdi 1 month ago

Selected Answer: A

Draining is correct

upvoted 1 times

ler_mp 1 month, 2 weeks ago

Selected Answer: A

You do not want to lose any data, so you should drain the pipeline instead of just cancelling it. Since the only option that includes draining the pipeline, the answer should be A.

upvoted 1 times

korntewin 1 month, 3 weeks ago

Selected Answer: B

Update command can only be done if the replacement pipeline is compatible with existing pipeline. the problem state very clear that the replacement pipeline is incompatible, thus Option A is out of choice.

The problem does say that they dont want to lose any of the processing data, thus provide JSON mapping object to deal with incompatible is correct choice here.

upvoted 1 times

Brillianttyagi 2 months ago

Selected Answer: A

A - 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

adelynllllllllll 2 months, 1 week ago

B is the answer
When transfer forming mapping, the data will be buffered temporarily. it will be transferred to the new job after the update is completed.

upvoted 1 times

Krish6488 2 months, 1 week ago

Selected Answer: B

The answer seems to be B to me. Dataflow automatically performs any compatibility checks when the code is updated. Inflight data processing is also taken care. There could be a small amount of downtime while the new jobs start post the compatibility checks. However this process means that new incoming data may not be applied with the revised transformations. But since the question is about not losing the data, the answer is B

upvoted 1 times

Prakzz 2 months, 1 week ago

Selected Answer: B

B is right.

upvoted 1 times

Nirca 2 months, 2 weeks ago

Selected Answer: A

Correct Option : A

upvoted 1 times

DGames 2 months, 2 weeks ago

Selected Answer: D

Answer - D Why ? after reading all discussion point and then re-read the question again , might I am thinking incorrect, Here is below point are mention in question.
1 . New Pipeline - means this is not current pipeline where we need to apply drain option. Yes in question it misleading statement mention about data lost.
2. Option B - Why you need to update current pipeline with JSON, No sense to update.
3. Option C - Old subscription already in use with current Pipeline.
SO I will go Option D.

upvoted 1 times

odacir 2 months, 3 weeks ago

Selected Answer: A

There are two cores in this question.
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

Thasni 3 months, 1 week ago

The below two links shows the answer as A. https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline and https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

upvoted 2 times

Question #32 *Topic 1*

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 site of the BigDate 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 2 years, 5 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 41 times

[Removed] Highly Voted 2 years, 11 months ago

Correct A

upvoted 20 times

[Removed] 2 years, 11 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 16 times

Brillianttyagi Most Recent 2 months 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 3 times

hilel_eth 2 months, 2 weeks 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 2 times

Pime13 7 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 12 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 1 year, 1 month 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 1 year, 3 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 10 times

anji007 1 year, 4 months ago

Ans: A

upvoted 1 times

sumanshu 1 year, 7 months ago

Vote for A

upvoted 3 times

timolo 1 year, 7 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

naga 2 years ago

Correct A

upvoted 4 times

Radhika7983 2 years, 3 months ago

This pretty straightforward. https://cloud.google.com/bigtable/docs/performance#troubleshooting
One of the troubleshoot performance issue is Make sure you're reading and writing many different rows in your table. That is Cloud Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Cloud Bigtable distribute the workload across all of the nodes in your cluster.

upvoted 6 times

snamburi3 2 years, 3 months ago

A
mentioned in the below post "Ideally, both reads and writes should be distributed evenly across the row space of the table."
https://cloud.google.com/bigtable/docs/schema-design
https://cloud.google.com/bigtable/docs/schema-design

upvoted 4 times

atnafu2020 2 years, 6 months ago

A
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.

upvoted 5 times

dg63 2 years, 7 months ago

In my opinion D is that correct answer as hotspot issue may be happening.
Question does not indicate that application is doing frequent read and updates to limited number of rows.

upvoted 1 times

[Removed] 2 years, 11 months ago

Answer: A
Description: Cloud Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Cloud Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Cloud 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.

upvoted 9 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

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Stackdriver can be used to check the error like number of unack messages, publisher pushing messages faster
upvoted 31 times

tprashanth 2 years, 7 months ago
B.
Stack driver monitoring is for performance, not logging of missing data.
upvoted 21 times

jkhong 2 months, 3 weeks ago
Please refer to this PubSub specific Monitoring metrics https://cloud.google.com/pubsub/docs/monitoring#monitoring_the_backlog
upvoted 1 times

mikey007 2 years, 7 months ago
https://cloud.google.com/pubsub/docs/monitoring
upvoted 2 times

ritinhabb 8 months, 2 weeks ago
Exactly!
upvoted 1 times

snamburi3 2 years, 3 months ago
All messages are being published to Cloud Pub/Sub successfully. so Stackdriver might not help.
upvoted 9 times

kubosuke 1 year, 5 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 1 year, 1 month ago
to be more precise, first to publisher,
- then forwards to topic, and persistance for a while
- then forwards to subscribe,
- then to subscription..
- then acknowledgement happens

so in every steps, there is possibly for errors.
upvoted 1 times

jkhong 2 months, 3 weeks ago
PubSub doesn't forward from subscriber to subscription. A topic sends it over to subscription first, then to subscriber
upvoted 1 times

[Removed] 2 years, 11 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 2 years, 11 months ago
Should be B
upvoted 22 times

[Removed] 2 years, 11 months ago
confused with D as well.
upvoted 1 times

Tanzu 1 year, 1 month 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 2 times

Rajokkiyam 2 years, 11 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 1 year, 11 months ago
Push needs Https endpoint
upvoted 3 times

jvg637 2 years, 11 months ago
pushing is only for a https endpoint. So Dataflow just can pull messages
upvoted 4 times

midgoo Most Recent 5 days, 4 hours 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 1 times

ploer 3 weeks, 3 days 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 3 times

Lestrang 4 weeks, 1 day 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, hurray.

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 2 times

PolyMoe 1 month 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

Lestrang 4 weeks, 1 day 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 2 times

desertlotus1211 1 month, 1 week 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 2 times

Prakzz 2 months, 1 week 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 month, 1 week 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 9 second delay. Push will automatically push the data to any subscribers as soon as the data comes, and thus is closer to real-time.

upvoted 1 times

Krish6488 2 months, 1 week 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 2 months, 2 weeks 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 2 months, 2 weeks 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 3 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

Jay_Krish 3 months, 1 week 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 5 times

NicolasN 2 months, 2 weeks ago

+1, Sound and concise reasoning

upvoted 1 times

soume 6 months, 3 weeks ago

Answer : C
Based to the documentation, with can monitor the ack message from the subscription (subscription/ack_latencies)
https://cloud.google.com/monitoring/api/metrics_gcp#gcp-pubsub

upvoted 1 times

Rahul95 7 months, 1 week ago

Answer B, You can't se the missing records using stackdriver you can only see logs.

upvoted 2 times

thapliyal 8 months ago

Ansewer is C. Just read this google doc https://cloud.google.com/pubsub/docs/monitoring#monitoring_the_backlog

upvoted 1 times

Deepakd 11 months ago

Answer is B.
Stackdriver will tell the count of messages not delivered. It wont tell why they are not being delivered. To understand the RCA , the job needs to be executed with fixed set of data.

upvoted 2 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/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 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 he common data in the HDFS storage for a Google Cloud Dataproc cluster.

Correct Answer: *B*

Community vote distribution

midgoo 5 days, 1 hour 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 1 times

korntewin 1 month, 3 weeks 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 2 months, 2 weeks ago

Selected Answer: B

Answer B look ok , because in question they want to store common data which can use by both workload, and using big query and primary analytical tool that would be best option and easy to analysis common data.

upvoted 1 times

kelvintoys93 2 months, 4 weeks ago

"Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data" - BigQuery cant 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 1 times

tunstila 1 month, 3 weeks ago

I thought you can now store unstructured data in BigQuery via the object tables announced during Google NEXT 2022... If that's possib;e, does that make B a better choice?

upvoted 1 times

drunk_goat82 3 months, 1 week 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 1 times

wan2three 3 months, 1 week 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 3 months, 1 week ago

Also apparently they want all data at one place and want bigQ

upvoted 2 times

gudiking 3 months, 1 week 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 in Dataproc)

upvoted 1 times

solar_maker 3 months, 2 weeks 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 3 months, 3 weeks 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 5 months, 1 week ago

C is ans...avro data can be accessed by spark as well

upvoted 3 times

ducc 5 months, 4 weeks 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 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

Correct Answer: *C*

Community vote distribution

jvg637 Highly Voted 2 years, 11 months ago
I would say A.
I think Pub/Sub can't directly send data to Cloud SQL.
upvoted 35 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
upvoted 14 times

niketd Most Recent 1 week, 6 days ago
Selected Answer: A
Pub/Sub to scale streaming data, Dataflow to processes both structured and unstructured data and cloud storage to store common data
upvoted 1 times

PolyMoe 1 month 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 amount 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 month 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 month, 1 week 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 3 months, 1 week ago
Are scenario based questions still in the latest exam?? Are these still relevant?
upvoted 2 times

kastuarr 4 months, 1 week 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 1 times

Dhamsl 5 months, 2 weeks ago
This site make me feel that it intends to make users to be involved in discussion by suggesting wrong answer
upvoted 7 times

Megmang 6 months, 3 weeks ago

Selected Answer: A
Answer is clearly option A.
upvoted 3 times

rytizzle 7 months, 2 weeks ago
why are there so many incorrect answers? it's so hard to study this way
upvoted 6 times

ratnesh99 7 months, 3 weeks ago
Answer A : because Cloud Sql not suitable for Global
upvoted 1 times

CedricLP 10 months, 2 weeks 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

Arkon88 12 months ago

Selected Answer: A
A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
as explained by JayZeeLee :
B is incorrect, because local SSD wouldn't satisfy the needs.
C is incorrect, because one of the requirements is 'Global', Cloud SQL is well suited for regional applications. Cloud Spanner is a better suit in that regard.
D is incorrect, because Load Balancer is for web traffic, not messages.
upvoted 4 times

kped21 1 year, 1 month ago
I should be A
upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: A
as explained by JayZeeLee
upvoted 1 times

kishanu 1 year, 2 months ago
Its A
Ingest Data from Global Resouces - Pub/Sub
Process data in real-time - Dataflow
Store data reliably - Cloud Storage
upvoted 1 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

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
upvoted 34 times

Radhika7983 Highly Voted 2 years, 3 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 13 times

jin0 6 days, 20 hours 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 1 times

JJJJim Most Recent 2 weeks, 5 days ago
Selected Answer: C
Answer is C, creating view tables can easy and flexible to do the most cost-effetive way.
upvoted 1 times

PolyMoe 1 month 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 1 times

dconesoko 1 month, 1 week 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 month, 1 week ago
C is the correct answer
upvoted 1 times

Isaga 1 month, 2 weeks ago
Selected Answer: C
I mean C
upvoted 1 times

noonting 1 month, 3 weeks ago
Selected Answer: B
The answer is B. Because it is not specified as a materialized view
upvoted 1 times

ler_mp 1 month, 2 weeks ago
But you would want a logical view for this, not materialized view
upvoted 1 times

Rodolfo_Marcos 2 months 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 delta from query updates. This is possible by using smart caching.
upvoted 1 times

lukacs16 2 months, 2 weeks ago
The correct answer is C. It should be B if there wouldn't for the real-time tracking requirement.
upvoted 3 times

hauhau 3 months ago
Selected Answer: B
B
Not view for C because view will be billed every query if it is not data studio for caching and it will costly if view is complex
upvoted 1 times

Brillianttyagi 2 months ago
How do you insert new data in table(what for the visual for new data).
upvoted 1 times

Dan137 6 months ago
Selected Answer: B
I go with B becase according to views documentation: https://cloud.google.com/bigquery/docs/views-intro#view_pricing "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."
upvoted 4 times

rr4444 8 months, 2 weeks ago
Selected Answer: C
C. Create a view on the table to present to the VISUALIZATION tool.

"virtualization" is one of several typoes....
upvoted 1 times

sw52099 9 months, 3 weeks ago
Since the primary problem is the team are overwhelmed by all the data, so the solution should be reducing the number of columns. C. creating a view doesn't reduce the number of columns.
upvoted 1 times

CedricLP 10 months, 2 weeks ago
Selected Answer: C
As already said :
A. Data is too big to be virtualized using Google Sheets.
B. Creating the new table won't be able to keep up with the new data inserts.
C. This will help to select appropriate columns as well as will be able to deal with new data inserts.
D. You cannot restrict access to selected columns using IAM. Views should be used to do that.
upvoted 4 times

stefanop 9 months, 3 weeks ago
It is possible to set column-level policies to enable only specific users (or group of users) to access (https://cloud.google.com/bigquery/docs/column-level-security-intro). So I think option D will be the best one.
upvoted 1 times

devric 10 months, 4 weeks ago

Selected Answer: B

Everyone said C, but... it's a virtualization tool??

upvoted 1 times

anji007 1 year, 4 months ago

Ans: C

upvoted 2 times

Question #37 *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 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 Clod 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

Manue Highly Voted 1 year, 10 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 30 times

sfsdeniso 3 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 1 year, 8 months ago
lol this is a vendor exam...
upvoted 7 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
upvoted 23 times

JJJJim Most Recent 2 weeks, 5 days ago
Selected Answer: B
Answer is B, attach the timestamp and ID is necessary to analyze data easily.
upvoted 1 times

nidmed 10 months, 1 week ago
Selected Answer: B
Answer: B
upvoted 4 times

Arkon88 12 months ago

Selected Answer: B

we need package ID + Timestamp so B

upvoted 1 times

davidqianwen 1 year, 1 month ago

Selected Answer: B

Answer: B

upvoted 1 times

exnaniantwort 1 year, 1 month ago

Selected Answer: B

agree with humza

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: D

https://cloud.google.com/pubsub/docs/reference/rest/v1/PubsubMessage

upvoted 4 times

[Removed] 1 year, 4 months ago

D is enough.. we have publish timestamp which is enough for this requirement

upvoted 1 times

Tanzu 1 year, 1 month 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, sensors or any other reasons..
- you need to populate orderingKey field too, so that subscribers can get in ordered.

upvoted 3 times

Pinko1497 10 months, 3 weeks ago

Also, since this is a International company, adding timestamp on message receiving would help catch local time.

upvoted 1 times

Chelseajcole 1 year, 4 months ago

It is about processing time and event time.. Answer is B.

upvoted 1 times

Tanzu 1 year, 1 month 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 1 year, 4 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 1 year, 7 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 1 year, 7 months ago

Vote for B

upvoted 2 times

funtoosh 2 years 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 ago
Correct B
upvoted 3 times

Radhika7983 2 years, 3 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.
upvoted 6 times

snamburi3 2 years, 3 months ago
D. "PublishTime- It must not be populated by the publisher in a topics.publish call."
https://cloud.google.com/pubsub/docs/reference/rest/v1/PubsubMessage
upvoted 1 times

snamburi3 2 years, 3 months ago
We can order the messages: https://cloud.google.com/pubsub/docs/ordering
upvoted 1 times

daghayeghi 1 year, 11 months ago
but no one can guarantee that finally the message will receive in order, because during the transmission of message the order of them will change. then using an identifier and timestamp would be a must.
upvoted 1 times

Tanzu 1 year, 1 month ago
you can add timestamp as data , not as processTime.
upvoted 1 times

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 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.
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 2 years, 11 months ago

D. The maximum number of workers answers to the scale question

upvoted 25 times

Radhika7983 Highly Voted 2 years, 3 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 job is processed as efficiently as possible.

upvoted 23 times

GCPpro Most Recent 1 month, 1 week ago

D . is the correct answer

upvoted 1 times

jkh_goh 1 month, 1 week ago

Answer A provided is definitely wrong. Who comes up with these answers?

upvoted 1 times

Ender_H 5 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 the 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 1 year, 1 month ago

Selected Answer: D

@Radhika7983

upvoted 2 times

medeis_jar 1 year, 1 month 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 1 year, 3 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 3 times

anji007 1 year, 4 months ago
Ans: D
upvoted 1 times

sumanshu 1 year, 7 months ago
Vote for D
upvoted 3 times

Lodu_Lalit 1 year, 11 months ago
D, thats because scalability is directly corerlated to max number of workers, size determines the speed of functioning
upvoted 3 times

naga 2 years ago
Correct D
upvoted 4 times

daghayeghi 2 years, 1 month ago
A exactly is correct:
Because Dataflow is scalable and don't need to repeated like cluster in Spark. then the only problem that remain is the zone of data that is important.
upvoted 3 times

daghayeghi 1 year, 11 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

SteelWarrior 2 years, 5 months ago
Shouldn't the answer be B. This will create the workers with the minimum number of workers required to support the streaming data and scale up as required.
upvoted 1 times

sumanshu 1 year, 8 months ago
Dataflow is autoscaled, So we can't mention exact number of workers. But we can cap it, even though we have default value of CAP.

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.

https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline
upvoted 1 times

atnafu2020 2 years, 6 months ago
D
numWorkers- int- The initial number of Google Compute Engine instances to use when executing your pipeline. This option determines how many workers the Dataflow service starts up when your job begins. If unspecified, the Dataflow service determines an appropriate number of workers.
maxNumWorkers-int-The maximum number of Compute Engine instances to be made available to your pipeline during execution. Note that this can be higher than the initial number of workers (specified by numWorkers to allow your job to scale up, automatically or otherwise.
upvoted 4 times

haroldbenites 2 years, 6 months ago
D Obviously
upvoted 3 times

yoRob 2 years, 7 months ago
D. Clearly
upvoted 4 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 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.
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 resent 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?

A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.

B. 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.

C. 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.

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.

Correct Answer: *C*

Community vote distribution

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
upvoted 25 times

itche_scratche Highly Voted 2 years, 10 months ago
D; dataflow doesn't connect to datastore, and not really for reporting. BQ, and data studio is a better choice.
upvoted 11 times

DGames Most Recent 2 months, 2 weeks 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 8 months, 3 weeks 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 be must be pre-computed. Thus option D cannot work as it create the metric AFTER querying the data (we are also not sure if we can really compute it in a query).
upvoted 4 times

Raj0123 9 months, 3 weeks ago
Answer D
upvoted 1 times

CedricLP 10 months, 2 weeks ago
Selected Answer: D
DataStudio and BQ are the simpliest way to do it
upvoted 1 times

devric 10 months, 4 weeks ago
Selected Answer: D
They also can activate BI Engine feature to improve the response time.
upvoted 1 times

sraakesh95 1 year, 1 month 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 to it's OLAP nature and the wonderful integration provided by GCP for these 2
upvoted 2 times

medeis_jar 1 year, 1 month ago
Selected Answer: D
as explained by JayZeeLee
upvoted 1 times

JayZeeLee 1 year, 3 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 3 months, 2 weeks 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 1 year, 4 months ago
Visualization = Data Studio 360
upvoted 1 times

Chelseajcole 1 year, 4 months ago
Next question give you the answer: Question #40. They using Data Studio 360 and Bigquery as source
upvoted 1 times

anji007 1 year, 4 months ago
Ans: D
upvoted 1 times

sumanshu 1 year, 7 months ago
Vote for D
upvoted 3 times

zosoabi 1 year, 9 months ago
just check the next question (#40) to get an idea about correct answer
upvoted 3 times

BhupiSG 1 year, 11 months ago
D
Correct
upvoted 2 times

naga 2 years ago
Correct C
upvoted 1 times

RogerLoSa 2 years, 1 month ago
After reading it twice, it's C. It's telemetry data, unstructured data, can't use BigQuery.
upvoted 2 times

daghayeghi 1 year, 11 months ago
Analyzing telemetry data won't make us to use NoSQL database and it depend to goal of the project. when the main goal is analytical job, then BigQuery is the best choice.
upvoted 2 times

Jay3244 2 years ago
Agreed... however they are using Biguery for analysis right for the company... so using Data Flow for the telemetry data and then loading to BQ... wont help ?? This way D is most suitable right ??
upvoted 3 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 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.
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

[Removed] Highly Voted 2 years, 11 months ago
Answer: B E
upvoted 41 times

shanjin14 Highly Voted 1 year, 9 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 32 times

samstar4180 1 year, 6 months ago
Yes, the correct answer should be BC - since we can have table-level access and each region represents a table.
upvoted 8 times

Lestrang Most Recent 4 weeks, 1 day 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 1 times

PolyMoe 1 month 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 1 times

korntewin 1 month, 3 weeks 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 1 times

korntewin 1 month, 3 weeks 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 3 months, 4 weeks 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 datasets and different regions?
So i choose BE
upvoted 7 times

svkds 5 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 6 months ago
Selected Answer: BC
I voted for BC

upvoted 1 times

NR22 10 months, 1 week ago

Selected Answer: BC

can apply IAM policies at table level

upvoted 1 times

devric 10 months, 4 weeks 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 11 months ago

Selected Answer: BC

BC are correct

upvoted 1 times

exnaniantwort 1 year, 1 month ago

Selected Answer: BC

"ensure employees can view only the data associated with their region, so you create and populate a table for each region"
definitely C instead of E

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: BC

Agree with B as the table split is region / location based
C: table level access is possible in Cloud IAM, which can be used to our advatnage for individual regions.
https://cloud.google.com/bigquery/docs/table-access-controls-intro

upvoted 2 times

medeis_jar 1 year, 1 month ago

Selected Answer: BE

as explained by hendrixlives

upvoted 1 times

prasanna77 1 year, 2 months ago

B,E because Table takes region of dataset

upvoted 1 times

hendrixlives 1 year, 2 months ago

Selected Answer: BE

B/E: Even if now BQ offers table level access control, since the number of tables can be expected to be high, controlling access at the dataset level would ease operations. That is why I would still go for E instead of C.

upvoted 8 times

exnaniantwort 1 year, 1 month ago

1 table for each region
you can just expect there is a limited number of regions
this number should be manageable
not E

upvoted 1 times

devric 10 months, 4 weeks ago

But the option B means that you will generate a data set per region also. So the permissions can be assigned at the dataset level.

If you pick the option A also, then you should assign the permissions over tables but... what about views? You should assign permissions to them also.

That's why I pick B and E.

upvoted 1 times

nellyoaid 1 year, 2 months ago

Selected Answer: BC

Answer

upvoted 2 times

Jlozano 1 year, 2 months ago

Now, Bigquery allows table level access control, but if you have 1.000 tables in a regional dataset, focusing on dataset is the easy and quick way to manage it.

upvoted 2 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 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.
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

itche_scratche Highly Voted 2 years, 10 months ago

None, rowkey should be Device_Id+Date(reverse)

upvoted 66 times

Jlozano 1 year, 2 months 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#12345device" 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 devide id (12345, 12346, 12347...).

upvoted 17 times

maxdataengineer 4 months, 3 weeks ago

Actually it depends. You will only have 365 x 2 dates unique dates at a given tame, since is a 2 year history, while most likely have more devices than that. So it will make more sence to start with the date first instead of the the device

upvoted 4 times

Ankit267 2 years, 7 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 13 times

Rajuuu 2 years, 7 months ago

A is a better option then other ..though not perfect as you mentioned.

upvoted 5 times

jvg637 Highly Voted 2 years, 11 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 devcie and date.

upvoted 15 times

michaelkhan3 1 year, 5 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 all together

upvoted 10 times

Whoswho 2 months, 1 week ago

I remember seeing it as well. the answer should be A. (reversed)

upvoted 1 times

wan2three 3 months, 1 week ago

but it didnt say cant use date, date and timestamp are different

upvoted 3 times

PolyMoe Most Recent 1 month 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 up 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 1 times

GCPpro 1 month, 1 week ago

A is the answer.

upvoted 1 times

Jackalski 2 months 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 1 times

slade_wilson 2 months, 2 weeks ago

Answer : A

upvoted 1 times

DGames 2 months, 2 weeks 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 2 months, 4 weeks ago

Obv you can't have date as the key, that would mean getting yourself a nice hotspot. The answer is C.

upvoted 2 times

ejlp 3 months, 1 week ago

Selected Answer: A

focus on "for a given device for a given day"

upvoted 1 times

NM1212 7 months, 3 weeks 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 7 months, 4 weeks ago

Selected Answer: C

need to make a key of device id and date, date is a secondary key.

upvoted 1 times

rr4444 8 months, 2 weeks ago

Yup this Q&A is a mess

upvoted 1 times

CedricLP 10 months, 2 weeks ago

Selected Answer: A

Classic answer will be Device_Id#Timestamp, but it's said query by date.
Date as less values than a timestamp, so Date#Device_Id is a good key.
Moreover, all other answer can't work

upvoted 2 times

devric 10 months, 4 weeks ago

Selected Answer: C

Option C. The most common query retrieve data from a given device in a given day. There's no sense in not using device_id as a beginning of the rowkey.

upvoted 2 times

Arkon88 12 months ago

Selected Answer: C

C. Rowkey: device_id Column data: date, data_point

1) The most common query is for all the data for a given device for a given day so device id + date will be required
2) to avoid hot spotting device id need to be first
https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid

upvoted 2 times

exnaniantwort 1 year, 1 month ago

Selected Answer: A

clearly A

upvoted 1 times

Arkon88 12 months ago

2) to avoid hot spotting device id need to be first
https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid

upvoted 1 times

jin0 6 days, 15 hours ago

I agree yours. cause of hot spotting A should not be a answer.

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: A

avoids hotspotting to some extent to the combination of the device ID and the timestamp

upvoted 1 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 2 years, 11 months ago
I would say B since Apache Spark is faster than Hadoop/Pig/MapReduce
upvoted 30 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: Spark performs in-memory processing and faster, which results in optimization of job's processing time
upvoted 15 times

ler_mp Most Recent 1 month, 3 weeks ago
Wow, a question that does not recommend to use Google product
upvoted 4 times

Whoswho 2 months, 1 week ago
looks like he's trying to spark the company up.
upvoted 3 times

Krish6488 2 months, 1 week 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 1 times

Mamta072 8 months ago
Ans is B . Apache spark.
upvoted 2 times

alecuba16 10 months, 1 week ago
Selected Answer: B
SPARK > hadoop, pig, hive
upvoted 2 times

kped21 1 year ago
B - Apache Spark
upvoted 1 times

kped21 1 year, 1 month ago
B Spark for optimization and processing.
upvoted 1 times

sraakesh95 1 year, 1 month ago
Selected Answer: B
B: Spark is suitable for the given operation is much more powerful
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
as explained by pr2web

upvoted 1 times

pr2web 1 year, 2 months 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 1 year, 3 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 1 year, 3 months ago

With a Spark pipeline, you have two different kinds of operations, transforms and actions. Spark builds its pipeline used 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 1 year, 3 months ago

Option A is wrong as Pig is wrapper and would initiate Map Reduce jobs
Option C is wrong as it would increase the cost.
Option D is wrong Hive is wrapper and would initiate Map Reduce jobs. Also, reducing the size would reduce performance.

upvoted 3 times

kastuarr 4 months, 1 week ago

Wont Option B increase the cost ? Cost of re-writing the job in Spark + Cost of additional memory ?

upvoted 1 times

anji007 1 year, 4 months ago

Ans: B
Spark performs better than MapReduce due to in memory processing.

upvoted 2 times

[Removed] 1 year, 4 months ago

All are wrong

A D runs in map-reduce , so rejected
B - spark costlier than map-reduce as in-memory. Data also increased

C - adding Hadoop nodes also costlier..

upvoted 1 times

ManojT 1 year, 4 months ago

B: The objective is to not increase the cost at the sametime do the analyitics required. Mapreduce jobs are not efficient and fast as spark so it will avoid failing the jobs.

upvoted 1 times

squishy_fishy 1 year, 4 months ago

None of the given answers are correct.
A. Rewrite the job in Pig. -- Wrong, because of the developer time cost.
B. Rewrite the job in Apache Spark. -- Wrong, because of the developer time cost.
C. Increase the size of the Hadoop cluster. This will increase cost too.
D. Wrong.

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

[Removed] Highly Voted 2 years, 11 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 55 times

funtoosh 2 years ago
cannot be 'A'as it clearly says that you need to change the schema and data.
upvoted 11 times

exnaniantwort 1 year, 1 month 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 11 times

YorelNation 5 months, 4 weeks 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 1 times

exnaniantwort 1 year, 1 month 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 4 times

[Removed] 2 years, 11 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 6 times

lgdantas 2 years, 6 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 1 year, 9 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 and 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 1 year, 5 months ago
You are asked to modify the schema and data. By using a view, the underlined table remains intact.
upvoted 9 times

HarshKothari21 5 months, 3 weeks ago
good catch, yoshik.

upvoted 1 times

alecuba16 9 months, 3 weeks 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 2 times

ovokpus 3 months ago
the point of billing is extra storage costs for a new concatenated column
upvoted 1 times

beowulf_kat 3 months, 3 weeks 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 3 months, 1 week ago
Of course, you use UPDATE after creating the new column, that is what the option said
upvoted 2 times

SuperTee 6 days, 11 hours ago
What happen if there are new employees joining the company, update every single time?
upvoted 1 times

[Removed] 2 years, 11 months ago
Can't be A
upvoted 5 times

BhupiSG Highly Voted 1 year, 11 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 engineer 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 41 times

DGames 2 months, 2 weeks ago
But you need to maintain table means regularly you have to execute the update query whenever new data comes.
upvoted 1 times

lollo1234 1 year, 9 months ago
I will also add that B would imply changing upstream workloads to write the new field every time a records gets added
upvoted 5 times

lollo1234 1 year, 9 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 1 year, 1 month 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 month, 3 weeks ago
Storage is cheap compared to computation
upvoted 2 times

midgoo Most Recent 4 days, 23 hours ago
Selected Answer: A
I choose A as it is the easiest way.
Furthermore, B is not cheaper. When we SELECT based on full_name after concatenated, it is the same cost and SELECT the full_name on the VIEW.
upvoted 1 times

techtitan 1 week, 1 day ago
Selected Answer: B
A vs B - I'd choose B because it is ideal to do bulk updates and minimize the cost that comes from view/concatenation.
upvoted 1 times

Tarek_74 1 week, 2 days ago
Selected Answer: B
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. Most Voted
upvoted 1 times

JJJJim 2 weeks, 4 days ago

Selected Answer: A
Answer should be A. create View is easy and cost-effective way
upvoted 2 times

tsontson 1 week, 3 days ago
How is it cost effective ? You will have to compute the view each time thus inducing costs.
upvoted 1 times

Tarek_74 3 weeks, 2 days ago
Selected Answer: B
its B to minimizing Cost
upvoted 1 times

chatocoral 3 weeks, 2 days ago
Selected Answer: A
Is not B because you are increasing costs in storage for the new column.
upvoted 1 times

donbigi 3 weeks, 2 days ago
Selected Answer: A
A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.

This option is the most cost-effective and efficient, as creating a view in BigQuery doesn't result in any data duplication, and the view is automatically updated whenever data in the underlying table changes. Additionally, querying a view is as efficient as querying a table, so performance will not be impacted.
upvoted 1 times

PolyMoe 1 month ago
Selected Answer: A
A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName. This approach allows the IT member to query the FullName field without modifying the existing Users table, and it also doesn't require any additional storage or data processing costs.

B. Adding a new column called FullName to the Users table and running an update statement also will work, but it will require additional storage and cost for the new column and also the cost of running the update statement.
upvoted 1 times

Nirca 1 month, 3 weeks ago
Selected Answer: A
A is the only right answer!!!!
upvoted 1 times

Jackalski 2 months ago
Selected Answer: A
to minimizing cost - I vote on A - simple and done in 2 minutes
B - will not work for new data/ employees
C,D are just overkill
upvoted 2 times

Catweazle1983 2 months ago
Selected Answer: A
I vote A because with B the Fullname field will not be filled when a new row is added.
upvoted 1 times

DGames 2 months, 2 weeks ago
Selected Answer: C
Answer C - here is some point to think on each option
option A - View - every time we need execute query adding computation cost.
Option B - Create another Table but nothing mention about to maintain table when new data come again we need to execute Update and adding cost.
Option C - it s best option design pipeline so it fix the issue
Option D- CSV doesn't make sense here.
upvoted 1 times

solar_maker 3 months, 2 weeks ago
Selected Answer: B
Best approach is B
upvoted 1 times

tikki_boy 3 months, 3 weeks ago
Selected Answer: D
Has to be D. It's the cheapest. BigQuery export/Load is free
upvoted 1 times

connorscion 2 months, 1 week ago

You are paying for processing power (for the dataproc job), storage (for the staging folder of the DP job) and storage from the second table instead of just storage from the second table from a view (from option A).

Also you need to re-run that job to prevent data staleness. A view is the easiest and cheapest option.

upvoted 2 times

MisuLava 3 months, 4 weeks ago

Selected Answer: B

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

a view would be very expensive being billed for all the data in the table when you only need one record, for example.

upvoted 3 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?

A. Manually configure the index in your index config as follows:

```
Indexes:
-kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

B. Manually configure the index in your index config as follows:

```
Indexes:
    -kind: Movie
        Properties:
        -name: actors
        -name: tags
-name: date_published
```

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 (75%) D (25%)

jkhong 2 months, 1 week 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 2 months, 1 week 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 is 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 2 months, 2 weeks 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 5 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 5 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 5 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 5 months, 1 week 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

Wasss123 5 months, 2 weeks ago

Selected Answer: A

Correct answer is A
Read in reference : https://cloud.google.com/datastore/docs/concepts/indexes#index_limits
n 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 4 times

soichirokawa 5 months, 3 weeks 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

Wasss123 5 months, 2 weeks ago

Correct answer is A
In the same reference you provided
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 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: A

A. Manually configure the index in your index config as follows:

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 in 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%)

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
upvoted 21 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Scheduler for adhoc jobs – 3 jobs free and $0.10 per job
upvoted 12 times

jin0 Most Recent 5 days, 19 hours 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 1 times

captainbu 1 month, 2 weeks 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 2 times

Ender_H 5 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 8 months, 1 week ago
Selected Answer: C
Answer: C
upvoted 1 times

anji007 1 year, 4 months ago
Ans: C
upvoted 2 times

Chelseajcole 1 year, 4 months ago
I know probably this question is testing on if you know corn.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 8 months, 3 weeks 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? This would be opportunity cost.

upvoted 2 times

Chelseajcole 1 year, 5 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 1 year, 8 months ago

Vote for 'C'

upvoted 2 times

naga 2 years ago

Correct C

upvoted 3 times

Radhika7983 2 years, 3 months ago

Answer is C. https://cloud.google.com/appengine/docs/flexible/nodejs/scheduling-jobs-with-cron-yaml

upvoted 5 times

haroldbenites 2 years, 6 months ago

C Correct

upvoted 4 times

Question #46 *Topic 1*

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

mmarulli Highly Voted 2 years, 11 months ago
this is one of the sample exam questions that google has on their website. The correct answer is B
upvoted 35 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: B is correct because regional storage is cheaper than BigQuery storage.
upvoted 11 times

funtoosh 2 years 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 8 times

musumusu Most Recent 4 days, 10 hours 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

AzureDP900 1 month, 3 weeks ago
B is right
upvoted 1 times

Krish6488 2 months, 1 week 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 it an overkill
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 for this use case
upvoted 2 times

ler_mp 1 month, 2 weeks ago
D is also overkill for this use case, so I'd pick B
upvoted 1 times

jkhong 2 months, 1 week 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 2 months, 2 weeks ago

Selected Answer: B

B is most inexpensive approach.

upvoted 1 times

odacir 2 months, 2 weeks 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 3 months, 1 week 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 4 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 3 months, 3 weeks ago

Correct, it cannot be B because BQ federated queries only work with Cloud SQL or Spanner

upvoted 1 times

gudiking 3 months, 1 week ago

It seems to me that they do: https://cloud.google.com/bigquery/docs/external-data-cloud-storage

upvoted 1 times

ducc 6 months ago

Selected Answer: B

I voted for B

upvoted 2 times

FrankT2L 8 months, 2 weeks ago

Selected Answer: C

The average prices of these goods are updated every 30 minutes
--> No Cloud Storage

upvoted 1 times

paulino_gauna 8 months, 3 weeks ago

Selected Answer: C

need to udpate the data

upvoted 2 times

Yad_datatonic 9 months ago

Selected Answer: B

Answer: B is correct because regional storage is cheaper than BigQuery storage.

upvoted 1 times

Vip777 9 months, 4 weeks ago

you cannot use federated queries with google cloud storage. how can B would be correct answer?

upvoted 4 times

Julionga 4 months, 2 weeks ago

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

upvoted 1 times

Arkon88 12 months ago

Selected Answer: B

B is correct.

we need it as cheaer as possible so cloud storage + external tables. Data wil be overwritten every 30 min

https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: B

@MaxNRG

upvoted 2 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

jvg637 Highly Voted 2 years, 11 months ago
You want to optimize the data schema + Machine Learning --> Bigquery. So A
upvoted 46 times

yoshik 1 year, 5 months ago
BigQuery is a datawarehouse, not a transactional db. You need to store transactional data as a requirement.
upvoted 18 times

alexmirmao 1 year, 4 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 4 times

yoshik 1 year, 4 months ago
In other questions they talk about 'transactional log data' when referring to past transactions, but you could be right, agree. In that case ok A BigQuery. Nevertheless, the question is formulated ambiguously.
upvoted 4 times

alecuba16 7 months, 2 weeks ago
Biquery 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 2 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: Should be D - Datastore
upvoted 23 times

GeeBeeEl 2 years, 3 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 16 times

BigQuery 1 year, 2 months ago
BQML is there. But, In question do they want to do ML on BQ?? Its saying just ML Based Company.
upvoted 4 times

jin0 Most Recent 5 days, 16 hours ago
1. Strong Schema. the table should be normalized
2. Transaction DB
The conclusion is absolutely B. Cloud SQL..
upvoted 1 times

Tarek_74 3 weeks, 1 day ago
Selected Answer: B

Transactions so B cloud Sql

upvoted 2 times

vraj_007 4 weeks, 1 day ago

ANSWER IS 'D'
Bigtable is optimized for high volumes of data and analytics,igtable is indexed by a single Row Key
Datastore is optimized to serve high-value transactional data to applications

upvoted 1 times

PolyMoe 1 month ago

Selected Answer: D

D. Datastore
Cloud Datastore is a NoSQL document database service that is well-suited for storing structured and semi-structured data, such as the user profile, user account information, and order information described in the question. Cloud Datastore can provide several options to optimize the data schema, such as denormalization, partitioning and sharding

upvoted 1 times

andye12 3 weeks, 6 days ago

do not just copy the answer from chatGPT. This is very structured data, so no need for datastore

upvoted 3 times

samdhimal 1 month, 1 week ago

I am Seriously super confused about this one. People are literally just saying A,B,C,D.

upvoted 2 times

samdhimal 1 month, 1 week ago

So I am pretty sure D is wrong for sure because I am looking at the doc and it's saying that Datastore doesnt have schema. Hence, this option is of no use.

upvoted 1 times

samdhimal 1 month, 1 week ago

But what about others?
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 it's 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

Let's discuss this only based on docs. Please make your your reply contains supporting google doc links. Let's get to the bottom of this.

upvoted 2 times

desertlotus1211 1 month, 2 weeks ago

If you want to optimize the data schema for a transactional database on the Google Cloud Platform, you should use Cloud Spanner... but that's not a choice...

Then it would make Cloud SQL the answer. I'm on the fence between BQ and Cloud SQL

Definitely not BigTable - that NoSQL.... Datastore is also NoSQL... not needed

upvoted 1 times

Nirca 1 month, 3 weeks ago

Selected Answer: A

You need Analytics. A

upvoted 1 times

SelvaJG 1 month, 4 weeks ago

Answer is C: ChatGPT says it is BigTable and the explanation makes sense to me as it can be used for storing transactional data and be used for ML models.

upvoted 1 times

Krish6488 2 months, 1 week ago

Selected Answer: A

It should be BQ. Transactional data term probably is to mislead the user. BQ tables when used as a DwH do consist a lot of transactions sourced from different source systems designed as fact tables. We are not talking about ingestion here rather its more about consumption/serving (ML).
So although Cloud SQL is good for OLTP, this is still a analytics use case.
Answer should be A (BigQuery)

upvoted 2 times

Asheesh1909 2 months, 2 weeks ago

Selected Answer: A

Bigquery stores transactional data as a fact table,support schema optimization which is not possible with datastore (not suggestable), So i support option A

upvoted 1 times

DGames 2 months, 2 weeks ago

Selected Answer: A

Straight Answer A - Big Query, for ML and transaction capability (work as data warehouse)
Easy to built ML model using BQML to forecast what same would like to eat.

upvoted 1 times

lukas_xls 2 months, 2 weeks ago

Selected Answer: B

Optimizing for transactional data, not for machine learning.

upvoted 3 times

odacir 2 months, 2 weeks ago

Selected Answer: A

As NicolasN says. The answer is A.

upvoted 1 times

NicolasN 3 months ago

Selected Answer: A

The answer comes from the official "Best practices for implementing machine learning on Google Cloud".
🔗 https://cloud.google.com/architecture/ml-on-gcp-best-practices#avoid-storing-data-in-block-storage

The interesting points:
- Store tabular data in BigQuery / Store image, video, audio and unstructured data on Cloud Storage
- Avoid storing data in block storage, like Network File Systems or on virtual machine (VM) hard disks
- Similarly, avoid reading data directly from databases like Cloud SQL.
- Instead, store data in BigQuery and Cloud Storage.

upvoted 5 times

nkit 3 months, 1 week ago

Selected Answer: A

Machine learning -> BQ
(the word "transactional database" is used to confuse you

upvoted 2 times

Question #48 *Topic 1*

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.

B. The CSV data has invalid rows that were skipped on import.

C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Correct Answer: *C*

Community vote distribution

C (100%)

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Bigquery understands UTF-8 encoding anything other than that will result in data issues with schema
upvoted 23 times

YAS007 Highly Voted 1 year, 6 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 12 times

samdhimal Most Recent 1 month, 1 week 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 1 times

samdhimal 1 month, 1 week 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 1 times

samdhimal 1 month, 1 week 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 to tell bigquery that orelse it will assume it is UTF-8. Hence, Imported data is not matching byte-to-byte to the source file. CORRECT ANSWER!
upvoted 1 times

samdhimal 1 month, 1 week 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 1 times

medeis_jar 1 year, 1 month 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 5 times

MaxNRG 1 year, 3 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 2 months, 2 weeks 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 1 year, 4 months ago

Ans: C

upvoted 3 times

sumanshu 1 year, 8 months ago

Vote for 'C'

upvoted 3 times

sumanshu 1 year, 7 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 ago

Correct C

upvoted 2 times

haroldbenites 2 years, 6 months ago

C is correct

upvoted 3 times

saurabh1805 2 years, 6 months ago

C is correct answer, Refer below link for more informaiton.
https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv#details_of_loading_csv_data

upvoted 6 times

[Removed] 2 years, 11 months ago

Answer: C

upvoted 10 times

Question #49 *Topic 1*

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

Toto2020 Highly Voted 2 years, 2 months ago
E cannot be: Transfer Service is recommended for 300mbps or faster
https://cloud.google.com/storage-transfer/docs/on-prem-overview

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

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

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

So, given the alternatives, I think answer is CD
upvoted 41 times

musumusu 4 days, 9 hours ago
50mbps is so slow, why you think bandwidth is ok! For parallel upload you need good internet ?
upvoted 1 times

vholti 1 year, 4 months ago
D is incorrect. gsutil with -m option uses multiprocessing/multithreading. It means it will copy the file in parallel. The benefit of multiprocessing/multithreading is significantly high when working with large number of files, instead of file size. The important point of multiprocessing/multithreading is sending multiple files in parallel. Hence file size doesn't give impact to gsutil with -m option. Gsutil with -m option doesn't split a big file into multiple chunks and transfer it in parallel. So in my opinion the answer is A and C.
upvoted 3 times

vholti 1 year, 4 months ago
Here is the docs which support my opinion: https://cloud.google.com/storage/docs/gsutil/addlhelp/TopLevelCommandLineOptions
upvoted 3 times

awssp12345 1 year, 7 months ago
This should be the correct answer.
upvoted 2 times

[Removed] Highly Voted 2 years, 11 months ago
Should be AC
upvoted 31 times

GeeBeeEl 2 years, 3 months ago
support this with a link....

upvoted 3 times

gcppde 2 years ago
Here you go: https://cloud.google.com/storage-transfer/docs/overview#gsutil
upvoted 1 times

BhupiSG 1 year, 11 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 9 months, 4 weeks ago
This link does support for C, but what about A? any supported links?
upvoted 1 times

musumusu Most Recent 1 week, 6 days 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 1 times

manigcp 2 weeks, 5 days ago
-- From ChatGPT --

B. Redesign the data ingestion process to use gsutil tool to send the CSV les to a storage bucket in parallel.
A. Introduce data compression for each file to increase the rate of file transfer.

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

A. Introducing data compression for each file will also help improve the rate of file transfer. This is because compressed data takes up less space and can be transferred faster, thereby reducing the time required to transfer the data.
upvoted 1 times

manigcp 2 weeks, 5 days 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

Leeeeee 3 months ago
Selected Answer: CD
https://cloud.google.com/storage/docs/parallel-composite-uploads
upvoted 2 times

abhineet1313 10 months, 1 week 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 10 months, 1 week 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 8 months, 2 weeks ago
E is wrong, as Bandwidth already low, so storage Transfer service will not help here

upvoted 2 times

Jojo9400 11 months, 1 week ago

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

upvoted 1 times

OmJanmeda 11 months, 2 weeks ago

Selected Answer: CD

CD is correct option

upvoted 1 times

Tanzu 1 year ago

20k files * 24 hours = 480k files x2 * 4kilobyte= 3.8gb everyday and must be processed in 10 hours

- either C or E (not both)= total size is less than 1tb per day, so C (gsutil) is the right tool
- in gsutil, the process can be paralleled (so you can utilize bandwidth throughput) and using rsync it is also possible to compress (to increase the transfer rate). and you do not need to decompress at target (to decrease process time at target such as untar or decompress). So A is better than E.!

the following cmd does the job.
gsutil -m rsync -az sourceDir gs://targetBucket

upvoted 1 times

rbeeraka 1 year, 1 month ago

Selected Answer: AC

A - would minimize over all size during seasonal times
C - gcp standard for skinny files

upvoted 1 times

rbeeraka 1 year, 1 month ago

A,C
A - would minimize over all size during seasonal times
C - gcp standard for skinny files

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: CD

@Toto2020

upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: CD

as explained by hendrixlives

upvoted 1 times

hendrixlives 1 year, 2 months ago

Selected Answer: CD

CD is correct.
See: https://jbrojbrojbro.medium.com/parallel-uploads-for-smaller-files-387ff86afc74

A: size is small enough that compressing each file will not help (indeed, it may even add overhead).

B: bandwidth is not a problem, no need to increase.

C: Parallel uploading the files with -m will increase speed in general.

D: many individual small files are a problem, since each file adds overhead to the processing and upload to GCS, and the upload sped of GCS is proportional to the size. If we pack all the small files in a bigger single TAR, it will improve the overall performance.

E: Storage Transfer Service is intended to move 100s of TB, and requires a 300Mbps connection as minimum (the doc even states that if your connection is less than 300Mbps is better to use gsutil). See: https://cloud.google.com/storage-transfer/docs/on-prem-overview#requirements

upvoted 7 times

abhineet1313 10 months, 1 week ago

dont see 300 limit mentioned anywhere, it says you can limit the bandwidth based on your isp https://cloud.google.com/storage-transfer/docs/managing-on-prem-jobs#control-bandwidth

upvoted 1 times

MaxNRG 1 year, 3 months ago

Selected Answer: CD

compression individual files won't help and increase bandwidth neither since there is no shortage there.
C looks good. sending files in parallel problem solved
D also looks good, since we'll save many connections

upvoted 3 times

Untamables 1 year, 3 months ago

A, D are correct. Sending tar.gz is efficient. B, C and E are not effective because the network latency is so high.

upvoted 1 times

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 (92%) 8%

jvg637 Highly Voted 2 years, 11 months ago
BDE. Hive is not for NoSQL
upvoted 35 times

sergio6 1 year, 5 months ago
Redis is also NoSQL
upvoted 1 times

vholti 1 year, 4 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

awssp12345 Highly Voted 1 year, 7 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 DATABSE.
upvoted 30 times

SuperTee 1 week ago
HDFS is. Hadoop Distributed File System. HDFS is storage and HIVE is for processing.
upvoted 1 times

musumusu Most Recent 1 week, 6 days 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 MongoBD
upvoted 1 times

MisuLava 6 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 1 year, 1 month ago

Selected Answer: BDE
@hendrixlives
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: BDE
as explained by hendrixlives
upvoted 1 times

hendrixlives 1 year, 2 months 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 highly 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 9 times

MaxNRG 1 year, 3 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, excelent 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 1 year, 2 months ago
Latency in Hive is usually quite high, and one of the requirements is "low latency"
upvoted 2 times

MaxNRG 1 year, 1 month ago
agreed on BDE
upvoted 1 times

MaxNRG 1 year, 1 month ago
good point!
upvoted 1 times

anji007 1 year, 4 months ago
Ans: B, D and E
upvoted 2 times

sumanshu 1 year, 8 months ago
vote for BDE
upvoted 2 times

BhupiSG 1 year, 11 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 1 year, 10 months ago
Hive is not for low latency queries. It is for analytics.
upvoted 5 times

daghayeghi 1 year, 11 months ago
BDE:
These are NoSQL DB, Hive is not for NoSQL.
upvoted 2 times

Rayleigh 2 years 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 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 ago

Correct BDE

upvoted 3 times

apnu 2 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 2 years, 3 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 (92%) 8%

madhu1171 Highly Voted 2 years, 11 months ago
it should be ACE
upvoted 59 times

[Removed] Highly Voted 2 years, 11 months ago
Should be ACE
upvoted 17 times

[Removed] 2 years, 11 months ago
prevent overfitting: less variables, regularisation, early ending on the training
upvoted 14 times

jin0 Most Recent 5 days 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 month, 2 weeks 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 the 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 month, 2 weeks 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 weigh heavily.

A & C are correct... the third one --- not sure on
upvoted 1 times

RoshanAshraf 1 month, 2 weeks ago
Selected Answer: ACE
A -The training data is causing the overfiting 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 month, 3 weeks 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 2 times

DGames 2 months, 2 weeks ago

Selected Answer: ADE

Answer ADE

upvoted 1 times

MisuLava 4 months ago

Selected Answer: ACE

100% sure ACE

https://elitedatascience.com/overfitting-in-machine-learning

upvoted 1 times

MisuLava 6 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 6 months, 2 weeks ago

Selected Answer: ACE

im vote for ACE

upvoted 1 times

Dip1994 6 months, 3 weeks ago

It should be ACE

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: ACE

@medeis_jar

upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 2 months 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

MaxNRG 1 year, 2 months ago

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 2 times

JG123 1 year, 3 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?
Ans: ACE

upvoted 5 times

anji007 1 year, 4 months ago

Ans: A, C and E.
To avoid overfitting use more training samples, with lesser features and use regularization.

upvoted 3 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 (82%) B (18%)

digvijay Highly Voted 2 years, 11 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 and read from bucket
upvoted 29 times

dambilwa 2 years, 8 months ago
Hence - Contextually, Option [C] looks to be the right fit
upvoted 15 times

retep007 1 year, 5 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 2 years, 11 months ago
Should be C
upvoted 27 times

PolyMoe Most Recent 1 month 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 job 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 1 times

Mkumar43 1 month, 3 weeks ago
Selected Answer: B
B works for the given requirement
upvoted 1 times

Krish6488 2 months, 1 week 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 2 months, 2 weeks 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 service account it’s similar just running job and doing rest of tasks read and writing as well.
upvoted 1 times

ThomasChoy 10 months, 1 week 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 1 year, 1 month ago

Selected Answer: C

Data owners cant 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 1 year, 1 month 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 1 year, 2 months 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 1 year, 2 months ago

Selected Answer: C

C,
Project Owner role to a service account - is too much

upvoted 1 times

JG123 1 year, 3 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 1 year, 4 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 1 year, 5 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 1 year, 8 months ago

Vote for 'C"

upvoted 4 times

sumanshu 1 year, 7 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

kino2020 1 year, 10 months ago

If the answer is "Workflow using Cloud Schedule" and the question is about cost effectiveness, then it's Workflow using Cloud Schedule compared to "Cloud Composer".
However, since the only answer is "workflow template", scheduling is not possible.
So the answer is C's "Cloud Composer".
https://cloud.google.com/dataproc/docs/tutorials/workflow-scheduler

upvoted 2 times

BhupiSG 1 year, 11 months ago

C
B seems to much

upvoted 2 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

[Removed] Highly Voted 2 years, 11 months ago

Should be D

upvoted 24 times

itche_scratche Highly Voted 2 years, 10 months ago

D; Purple is reading, Blue is writing. so majority is reading.

upvoted 20 times

squishy_fishy 1 year, 4 months ago

I have been looking for the color code descriptions for a while. Thank you!

upvoted 1 times

PolyMoe Most Recent 1 month 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 1 times

AzureDP900 1 month, 3 weeks ago

D is right

upvoted 1 times

Krish6488 2 months, 1 week 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 3 months, 2 weeks ago

Selected Answer: A

A is the best option becouse the color bar show the high number of reads and i think its not a skew becouse biguery was build to compute the data fast

upvoted 2 times

arpitagrawal 5 months, 3 weeks ago

The query would throw the error because you're using a group by clause on country but not aggregating city or state.

upvoted 7 times

MisuLava 6 months ago

Selected Answer: D

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

upvoted 1 times

Paul_Oprea 10 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 first place, unless I'm missing something.
upvoted 14 times

Arkon88 12 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 1 year, 1 month ago
Selected Answer: D
https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d
upvoted 3 times

sraakesh95 1 year, 1 month ago
https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d
upvoted 1 times

medeis_jar 1 year, 1 month 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 2 times

morpho4444 1 year, 3 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 1 year, 2 months ago
A Cant be answer. Since users whenever running queries facing the problems.
upvoted 1 times

JG123 1 year, 3 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 4 times

anji007 1 year, 4 months ago
Ans: D
Data is skewed.
upvoted 2 times

timolo 1 year, 7 months ago
Correct is D: https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

Partition skew, sometimes called data skew, is when data is partitioned into very unequally sized partitions. This creates an imbalance in the amount of data sent between slots. You can't share partitions between slots, so if one partition is especially large, it can slow down, or even crash the slot that processes the oversized partition.
upvoted 9 times

vholti 1 year, 4 months ago
The query from question doesn't have partition filter (no WHERE clause). So regardless there is data skew, it's not the root cause of the low performance in this case. Option A is more reasonable I think.
upvoted 1 times

Question #54 *Topic 1*

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

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

jvg637 Highly Voted 2 years, 11 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 48 times

[Removed] 1 year, 10 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 12 times

donbigi 3 weeks, 2 days 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 1 times

AzureDP900 1 month, 3 weeks ago
Agree with B
upvoted 1 times

ralf_cc 1 year, 8 months ago
Yep, Pub/Sub doesn't have FIFO yet, B is the one that keeps the right order
upvoted 4 times

Tanzu 1 year ago
it is not a queue, and that is not a issue :)
upvoted 3 times

Tanzu 1 year ago
in a distributed environment, you can not handle this problem wit a queue by the way !
upvoted 3 times

Ganshank Highly Voted 2 years, 10 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 ot Pub-Sub and consuming via Dataflow.
upvoted 28 times

Tanzu 1 year 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

musumusu Most Recent 1 week, 6 days ago
Anser is D:
many users are confused here in PUSH and PULL from pub/sub. PUSH is used when you want monitor data quickly on dashboard or in reports

for analysis. PULL is used when data is big, and duplicate (like same amount of bids in this case is possible) which is collected securely and processed at a given time. It doesn't mean pub/sub did not receive data in real time. Here we need to be careful in data collection and all bid time record than showing it on dashboard. PUSH is not a good approach here.
Option B: is more expensive and less compatible with Pub/Sub, you need to handle same amount bids as the occured.. i dont see any real time solution here.

upvoted 1 times

donbigi 3 weeks, 2 days ago

Selected Answer: D

This approach provides a scalable and highly available solution for processing bid events in real time. By using Cloud Pub/Sub, the application servers can publish the bid events as they occur, and the data can be pulled from the Pub/Sub topic by a single Cloud Dataflow pipeline. This ensures that bid events are not duplicated and can be processed in the order they are received. Additionally, Cloud Dataflow provides a flexible and scalable mechanism for processing large amounts of real-time data, which makes it well-suited for this use case.

upvoted 2 times

r9765345 3 weeks, 3 days ago

Selected Answer: D

2 minutes in to this video by Google explains why the answer is D:
https://www.youtube.com/watch?v=KObJkda4ZfY

upvoted 1 times

GCPpro 1 month, 1 week ago

B is the correct answer

upvoted 1 times

Jackalski 2 months ago

Selected Answer: B

I vote on B
option D does not answer where data is stored and it is pull with latency

upvoted 2 times

Krish6488 2 months, 1 week ago

Selected Answer: B

I would go with B as D is pull + 'give bid to user who processed first'. Since the requirement is to give bid on the basis of event and not processing ts, B is a close match

upvoted 1 times

Sonika11 2 months, 2 weeks ago

Selected Answer: D

Dataflow is needed to see which user sent bid first. Cannot track this with Pub/Sub alone.

upvoted 1 times

DGames 2 months, 2 weeks ago

Selected Answer: C

Option C look ok because in question mentioned You want to collate those bid events into a single location in real time to determine which user bid first

upvoted 1 times

balajir27 2 months, 3 weeks ago

Selected Answer: D

It is a global application hence cloud sql can not be an option

upvoted 3 times

ducc 6 months ago

Selected Answer: B

I voted for B

upvoted 2 times

MisuLava 6 months ago

Selected Answer: B

this is a bidding system. reading data periodically is not eficient.
you need to push data as soon as it is available, and storing it to CLoudSQL is fine. pubSub and Push are the key features here.

upvoted 2 times

changsu 8 months ago

Selected Answer: B

You have to compare the timestamp before choosing the winner. There is latency to publish a message and to pull it. The first processed one by the dataflow is not always the first bid. You have to store all data to compare them via timestamp. Therefore not D, but B.

upvoted 1 times

alecuba16 7 months, 2 weeks ago

That's not true, if you have a single partition in kafka is like a queue, so yes, the message will be ordered. In that case we can assume that the partition for the topic bids is 1, so only a single susbcriber = queue.

Then point is if you have multiple partitions in pub/sub even using push, the messages can arrive at the wrong order, so the first is not necessary the first message to be pushed.

Dataflow can be considered as realtime if the window is seconds.

upvoted 1 times

Preemptible_cerebrus 8 months, 2 weeks ago

Selected Answer: B

This questions have ambigous answers, by the way, if we want to fit into the requirement, the answer that aproximates to it, is answer B because we have 2 facts "they need to have a location where to RECEIVE the events in REAL TIME". We have a push subscription, we have an endpoint that would be Cloud SQL, therefore, this gives the detail that they do not need to have automated the answer of which bid was the first to come, placing answer D in 2nd place and also, because it has a pull subscription, and in addition it doesnt mention where and which would be the sink. A and C aren't near to accomplish the goal to reach.

upvoted 1 times

AmirN 8 months, 3 weeks ago

B is not correct. Cloud SQL is not a global system, and it is not scalable for use with pubsub.
Push subscription to Cloud SQL is disaster because it wont be able to handle it. And also after the data gets into Cloud SQL what next? All the processing required to determine the fastest bid will cause more latency.

upvoted 3 times

snehalpatil 8 months, 2 weeks ago

it is proceed in a single region.

upvoted 1 times

Arkon88 12 months ago

Selected Answer: B

B.
- for realtime pub/sub push is critical. pull creates latency so not D

upvoted 2 times

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 ƛ€eventsƛ€

Correct Answer: *AE*

Community vote distribution

CD (62%) | CE (23%) | DE (15%)

jvg637 Highly Voted 2 years, 11 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 42 times

[Removed] Highly Voted 2 years, 11 months ago

Answer: CD

upvoted 9 times

musumusu Most Recent 1 week, 6 days 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 1 times

musumusu 4 days, 8 hours ago

Go for B, create a new view from the table, If you modify the syntex in option A, its also mean you created a new view on table :P

upvoted 1 times

PolyMoe 1 month 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, but is not required for the ODBC connection to work.

upvoted 2 times

samdhimal 1 month, 1 week 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 connectt ODBC we need a view described by standard SQL.

upvoted 3 times

samdhimal 1 month, 1 week 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 ƛ€eventsƛ€
- This option is of no use in this scenario

upvoted 1 times

GCPpro 1 month, 1 week ago

CE is the correct answer

upvoted 1 times

MisuLava 6 months ago

Selected Answer: CD

needed a service account for ODBC drivers
standard SQL vs legacy SQL.

upvoted 1 times

Smaks 7 months, 1 week 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 3 times

Smaks 7 months, 1 week ago

typo - D; E

upvoted 4 times

Arkon88 12 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 1 year, 1 month 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 1 year, 2 months 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 3 times

MaxNRG 1 year, 2 months 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 1 year, 3 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 6 times

anji007 1 year, 4 months ago

Ans: C & D.

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#logical_views

upvoted 2 times

BhupiSG 1 year, 11 months ago

Correct: CD

upvoted 3 times

Tanmoyk 2 years, 5 months ago

Should be C,D

upvoted 3 times

haroldbenites 2 years, 6 months ago

C y D is correct

upvoted 3 times

dambilwa 2 years, 8 months ago

Options [C&D] are most appropriate

upvoted 3 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:

```
SELECT
  user_dim.geo_info.city,
  COUNT(user_dim.geo_info.city) as city_count
FROM
TABLE_DATE_RANGE([firebase-analytics-sample-data:a
TABLE_DATE_RANGE([firebase-analytics-sample-data:i
GROUP BY
  user_dim.geo_info.city
ORDER BY
  city_count DESC
```

Community vote distribution

A (100%)

damaldon Highly Voted 5 months, 1 week ago
A. is correct according to this link:
https://cloud.google.com/bigquery/docs/reference/legacy-sql
upvoted 6 times

samdhimal Most Recent 1 month, 1 week ago
A Is correct.

TABLE_DATE_RANGE() : Queries multiple daily tables that span a date range.
upvoted 1 times

samdhimal 1 month ago
Example:
SELECT *
FROM TABLE_DATE_RANGE(app_events_, TIMESTAMP_SUB(CURRENT_TIMESTAMP(), INTERVAL 30 DAY), CURRENT_TIMESTAMP())
upvoted 1 times

DipT 2 months, 1 week ago
Selected Answer: A
https://cloud.google.com/bigquery/docs/reference/legacy-sql
TABLE_DATE_RANGE() Queries multiple daily tables that span a date range.
upvoted 1 times

skp57 3 months, 2 weeks ago

A. is Correct...
from...https://cloud.google.com/blog/products/management-tools/using-bigquery-and-firebase-analytics-to-understand-your-mobile-app
SELECT
user_dim.app_info.app_platform as appPlatform,
user_dim.device_info.device_category as deviceType,
COUNT(user_dim.device_info.device_category) AS device_type_count FROM
TABLE_DATE_RANGE([firebase-analytics-sample-data:android_dataset.app_events_], DATE_ADD('2016-06-07', -7, 'DAY'),
CURRENT_TIMESTAMP()),
TABLE_DATE_RANGE([firebase-analytics-sample-data:ios_dataset.app_events_], DATE_ADD('2016-06-07', -7, 'DAY'), CURRENT_TIMESTAMP())
GROUP BY
1,2
ORDER BY
device_type_count DESC

upvoted 1 times

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 (90%) 10%

[Removed] Highly Voted 2 years, 11 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 57 times

samdhimal 1 month 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

jvg637 Highly Voted 2 years, 11 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 11 times

midgoo Most Recent 23 hours, 3 minutes 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 1 week, 6 days 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 2 months, 1 week 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 5 months, 2 weeks ago

Selected Answer: D

Answer is D

upvoted 1 times

TOXICcharlie 5 months, 2 weeks 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 1 times

FrankT2L 8 months, 3 weeks 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 10 months, 2 weeks ago

Why A won't cause an error?

upvoted 3 times

910 11 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 1 year, 1 month ago

Selected Answer: D

D.
https://beam.apache.org/documentation/programming-guide/#windowing

upvoted 3 times

JG123 1 year, 3 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 1 year, 4 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, a trigger strategy, or both for each collection, Beam generates an IllegalStateException error at pipeline construction time."

...as it says one of windowing or triggering or both are required. But question itself says windowing and transformations are already. So may be this particular pipeline requires both but triggering strategy is missing. So I will go with Option B.

upvoted 3 times

zxing233 2 years, 5 months ago

Why not A? D can still work if you add a trigger

upvoted 4 times

VishalB 2 years, 7 months ago

Correct Answer: D
Explanation:-This option is correct as with unbounded (Streaming) Pub/Sub collection you need to apply the non-global windowing function.

upvoted 8 times

itche_scratche 2 years, 10 months ago

D, global is default and for batch not streaming.

upvoted 7 times

haroldbenites 2 years, 6 months ago

Right. D is correct

upvoted 3 times

[Removed] 2 years, 11 months ago

Should be D

upvoted 5 times

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 2 years, 5 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 48 times

Yiouk 1 year, 6 months ago

B. different MR jobs execute in series, adding 1 more job makes sense in this case.

upvoted 6 times

[Removed] Highly Voted 2 years, 11 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 mapreduce job and storing/passing the values to next job

upvoted 19 times

Jphix 1 year, 9 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

samdhimal Most Recent 1 month 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 month 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 2 months, 1 week ago

Selected Answer: B

It is much cleaner approach

upvoted 1 times

DGames 2 months, 2 weeks 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 2 months, 3 weeks 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 patron and has a lot of problems
D It's overkilling, a don't have sense in this scenario.

upvoted 1 times

ZIMARAKI 1 year, 1 month ago

Selected Answer: B

SteelWarrior explanation is correct :)

upvoted 3 times

lord_ryder 1 year, 1 month ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

hendrixlives 1 year, 2 months ago

Selected Answer: B

SteelWarrior's answer is correct

upvoted 1 times

anji007 1 year, 4 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 1 year, 8 months ago

Vote for 'B' (introduce new job) over 'A', (instead of modifying existing job)

upvoted 5 times

YuriP 2 years, 6 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

[Removed] 2 years, 11 months ago

It's between A or B.
Should choose A

upvoted 4 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

PolyMoe 1 month 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 1 times

samdhimal 1 month 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

samdhimal 1 month 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 month, 2 weeks 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 1 times

DipT 2 months, 1 week ago

Selected Answer: B

https://cloud.google.com/bigquery/docs/partitioned-tables

upvoted 1 times

DipT 2 months, 1 week 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 queries
https://cloud.google.com/bigquery/docs/reference/legacy-sql

upvoted 3 times

DGames 2 months, 2 weeks 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 BI tool its possible with RDBMS db.

upvoted 1 times

zellck 2 months, 3 weeks 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 2 times

odacir 2 months, 3 weeks 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 3 months ago

Selected Answer: B

BigQuery for Analytics and BI

upvoted 1 times

Leelas 3 months, 3 weeks 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 3 months, 4 weeks ago

Cloud SQL supports transactions as well as analysis through a BI tool. Firestore/Datastore does not support SQL syntax typically needed to do 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

MisuLava 4 months ago

Selected Answer: B

it is obvious.

upvoted 1 times

MisuLava 4 months ago

I meant to choose A, BugQuery :)

upvoted 1 times

Chavoz 4 months, 4 weeks ago

Selected Answer: B

I'd say B. Big query is not suitable for transactional use case, and Cloud SQL supports transactions as well as analysis through a BI tool.

upvoted 2 times

Chavoz 4 months, 4 weeks ago

Changed my mind. It's A. Even if they mention transactional, BI it's Big Query. Also, they don't mention it must be relational, in that case would be Cloud SQL. If the main focus is Business Inteligence, Big Query.

upvoted 1 times

DiegoGonL 5 months, 2 weeks ago

Selected Answer: A

BigQuery for analysing

upvoted 2 times

sedado77 5 months, 2 weeks ago

Selected Answer: A

BigQuery for anything Analytics

upvoted 1 times

badrisrinivas9 5 months, 2 weeks ago

Selected Answer: A

Bigquery

upvoted 1 times

TOXICcharlie 5 months, 2 weeks ago

Selected Answer: A

I feel BigQuery would make more sense here because data needs to be aggregated and there would be analytical workloads, which would benefit from bigquery's caching.

upvoted 1 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] Highly Voted 2 years, 11 months ago
should be B
https://cloud.google.com/bigquery/docs/creating-partitioned-tables#converting_date-sharded_tables_into_ingestion-time_partitioned_tables
upvoted 36 times

Rajuuu 2 years, 7 months ago
The above link does mention about shard ing benefits but only about partition tables.
A is correct.
upvoted 3 times

g2000 2 years, 1 month ago
keyword is single
upvoted 5 times

Chelseajcole 1 year, 5 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. This 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 6 times

vholti 1 year, 4 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 ago
https://cloud.google.com/bigquery/docs/partitioned-tables provides that info you are looking for. Shortly, partitioning performs better than 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] Highly Voted 2 years, 11 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 22 times

lgdantas 2 years, 5 months ago
Can you please share the reference?
upvoted 2 times

Tumri 1 year, 5 months ago
https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard
upvoted 6 times

PolyMoe Most Recent 1 month 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 partioned). So we still facing the issue of max 1000 tables. In B. we have only ONE table (partioned)

upvoted 2 times

samdhimal 1 month 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 1 times

RoshanAshraf 1 month, 2 weeks 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 month, 2 weeks 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 2 months, 1 week ago

Selected Answer: B

https://cloud.google.com/bigquery/docs/partitioned-tables

upvoted 1 times

DGames 2 months, 2 weeks 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 2 months, 3 weeks 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 4 months, 2 weeks ago

Selected Answer: B

Convert MANY sharded tables into a single ONE (partitioned) table

upvoted 2 times

rrr000 6 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 8 months, 2 weeks 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 for every SINGLE DAY "LOGS_YYYYMMDD" partitioning every table will end on scanning all the records of each table when you query them by date ranges using the wildcards, there will be no difference on time-partitioning each table versus consuming them as described.

upvoted 2 times

AmirN 8 months, 3 weeks 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 6 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 9 months, 4 weeks 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 1 year, 1 month ago

Selected Answer: B

Partitioning > table sharding:
https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

upvoted 2 times

kishanu 1 year, 2 months ago

Its A, as google, recommends the best practice to use partition over sharding.

upvoted 1 times

hendrixlives 1 year, 2 months ago

Selected Answer: B

B
Partitioning is recommended over table sharding:
https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

upvoted 3 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

B (100%)

jvg637 Highly Voted 2 years, 11 months ago
B. (Hadoop/Spark jobs are run on Dataproc, and the pre-emptible machines cost 80% less)
upvoted 42 times

rickywck Highly Voted 2 years, 11 months ago
I think the answer should be B:

https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms
upvoted 16 times

samdhimal Most Recent 1 month 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 2 times

Rodolfo_Marcos 1 month, 3 weeks 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 2 months, 1 week ago
Selected Answer: B
Using preemtible machines are cost effective , and because is suitable for a job mentioned here as it is fault tolerant .
upvoted 1 times

DGames 2 months, 2 weeks ago
Selected Answer: B
User Pre-emptible VM machine and save process cost, and question want simple solution.
upvoted 1 times

odacir 2 months, 3 weeks 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 standar, so this is the answer
C and D are more expensive.
upvoted 1 times

Remi2021 5 months, 3 weeks ago

Selected Answer: B
B is right way to go
upvoted 1 times

FrankT2L 9 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 11 months, 1 week ago
B is teh right answer. examtopics update your answers or make your site free again.
upvoted 4 times

OmJanmeda 11 months, 2 weeks 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 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 1 year, 1 month ago
Ans : B,
here we are checking on reducing cost, so pre-emptiable machines are best choice
upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 2 months ago
Selected Answer: B
A is not valid, for apache spark jobs dataproc y 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…thats the one. B
upvoted 1 times

hendrixlives 1 year, 2 months ago
Selected Answer: B
B
Use pre-emptible virtual machines (VMs) for the cluster: this will reduce costs of the dataproc cluster. The other options will increase the costs. The work is required to run weekly and takes 15 min, so the performance is not an issue.

The option "A" of migrating to dataflow will impose a migration of the technology from spark to beam, which will be costly and even difficult if the developers are not experienced in Beam. This can be considered for future evolution, but it is out of the scope of the requirements stated in the question.
upvoted 2 times

JG123 1 year, 3 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 2 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 (87%) 7%

[Removed] Highly Voted 2 years, 11 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 36 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
upvoted 16 times

samdhimal Most Recent 1 month 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 1 times

samdhimal 1 month 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 be late.
upvoted 2 times

desertlotus1211 1 month, 1 week ago
Answer is C:

There is no such thing as a sliding windows using by dataflow.
upvoted 1 times

DeeData 1 month, 1 week ago
I highly doubt, DataFlow windowing is divided into three(3) types:

1. Fixed
2. Sliding
3. Session
upvoted 1 times

AzureDP900 1 month, 3 weeks 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 2 months, 2 weeks ago

Selected Answer: C

Watermark is use for late date,

upvoted 2 times

[Removed] 3 months, 2 weeks 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-partionned table using the timestamp of the element)

upvoted 3 times

ovokpus 3 months, 1 week ago

with watermarks, when the late data arrives, it goes into its rightful window and gets in order

upvoted 1 times

ovokpus 3 months, 1 week ago

C even says watermarks AND timestamps.

upvoted 1 times

FrankT2L 9 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 9 months ago

delete this answer. The answer belongs to another question

upvoted 4 times

Tanzu 1 year ago

Selected Answer: A

That's why we have watermarks in apache beam.

upvoted 1 times

VishalBule 1 year 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 1 year, 1 month 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 1 year, 2 months ago

Selected Answer: C

A is a direct No, if data don't have timestamp, we'll only have the procesing 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 1 year, 2 months 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 1 year, 2 months 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 1 year, 2 months ago

Selected Answer: C

"Expert Verified" but >50% questions have random answer. "Sliding window" really? Please, this can be fixed easyly with our most voted answer. Of course, the correct answer is C.

upvoted 4 times

JG123 1 year, 3 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 1 year, 4 months ago

Ans: C

upvoted 2 times

safiyu 1 year, 6 months ago

Answer should be C. sliding windows are meant for calculating running average and not lagging data. Watermark is best for this purpose

upvoted 7 times

sumanshu 1 year, 8 months ago

vote for 'C"

upvoted 4 times

Question #63

Topic 1

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. X2+Y2

B. X2

C. Y2

D. cos(X)

Correct Answer: *D*

Community vote distribution

A (81%) B (19%)

jvg637 Highly Voted 2 years, 11 months ago
For fitting a linear classifier when the data is in a circle use A.
upvoted 36 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
upvoted 14 times

samdhimal Most Recent 1 month ago
A. X2+Y2

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 X2+Y2. By 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.

X2 and Y2 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 1 times

GCPpro 1 month, 1 week ago
A is the correct answer as graph of circle is x^2 + y^2
upvoted 1 times

desertlotus1211 1 month, 1 week ago
Answer is A:
The answer reflects 'x' to the 2nd power + 'y' the 2nd power.
I guess they can't use carots in the exam answers!
upvoted 1 times

AzureDP900 1 month, 3 weeks ago
A is right
Reference:
https://medium.com/@sachinkun21/using-a-linear-model-to-deal-with-nonlinear-dataset-c6ed0f7f3f51
upvoted 1 times

DipT 2 months, 1 week ago

Selected Answer: A

https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture

upvoted 1 times

DGames 2 months, 2 weeks ago

Selected Answer: A

linear circle X2+Y2 https://www.stat.cmu.edu/~cshalizi/dm/20/lectures/08/lecture-08.html

upvoted 2 times

mvww11 8 months, 1 week 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 9 months, 3 weeks ago

Answer: A.
X^2+Y^2 is the equation of a circle.

upvoted 1 times

diagniste 10 months ago

Selected Answer: A

C'est A

upvoted 2 times

Tanzu 1 year ago

Selected Answer: A

only A is draw a circle

upvoted 1 times

sraakesh95 1 year, 1 month ago

Selected Answer: A

Equation of circle as represented in the question

upvoted 1 times

moumou 1 year, 1 month ago

F(x) as A B C will have always a positive values as result, for A will need a third dimenssion Z to represent data, only D:cos(x) can be presented as the shown classification. this is a math question

upvoted 1 times

NR22 10 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 1 year, 1 month 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 1 year, 2 months ago

Selected Answer: A

A is the equation of a circle which looks like a good separation for the black vs grey dots.
http://playground.tensorflow.org

upvoted 1 times

StefanoG 1 year, 3 months ago

Selected Answer: B

I think the right answer is B, we have 3 shades that are specular on Y axis so the classificatin for -x is the same for +x so i think that we must use the tuple(X^2 , y) to classify the value

upvoted 3 times

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 files system, and use those credentials to access the BigQuery dataset

Correct Answer: *C*

Community vote distribution

C (100%)

[Removed] Highly Voted 2 years, 11 months ago
Correct: C
upvoted 25 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Service Account are best option when granting access from tools/appllications
upvoted 15 times

samdhimal Most Recent 1 month 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 1 times

samdhimal 1 month 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 files 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 1 times

DGames 2 months, 2 weeks ago
Selected Answer: C
Service account approach is secure in GCP to communicate with between service or application
upvoted 1 times

NicolasN 3 months, 3 weeks 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 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 ago

Selected Answer: C

Correct: C

upvoted 2 times

anji007 1 year, 4 months ago

Ans: C

upvoted 2 times

daghayeghi 1 year, 11 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 2 years, 3 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 to 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 1 year, 4 months ago

The answer is C.
When access data through application, Google recommendation is using service account.

upvoted 1 times

kavs 2 years, 3 months ago

Yes A seems to be right

upvoted 4 times

kavs 2 years, 3 months ago

It says individually don't want to authorise service account could be right too

upvoted 3 times

Ankush_j 2 years, 5 months ago

Ans is C, Service account is best for secure data

upvoted 3 times

haroldbenites 2 years, 6 months ago

C is correct

upvoted 3 times

Rajuuu 2 years, 7 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 Highly Voted 2 years, 11 months ago
real-valued can not be null N/A or empty, have to be “0”, so it has to be B.
upvoted 35 times

[Removed] Highly Voted 2 years, 11 months ago
Should be B
upvoted 16 times

samdhimal Most Recent 1 month 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 1 times

GCPpro 1 month, 1 week ago
B is the correct answer.
upvoted 1 times

AzureDP900 1 month, 3 weeks 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 need to be replaced with a number. And Cloud dataprep is best casual tool out of given options.
upvoted 1 times

DGames 2 months, 2 weeks ago
Selected Answer: B
real value 0
upvoted 1 times

byash1 1 year, 1 month ago
Selected Answer: B
It is B
upvoted 2 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
Dataprep + real value (0)
upvoted 1 times

MaxNRG 1 year, 2 months 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 1 year, 4 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 1 year, 8 months ago
Vote for 'B'
upvoted 4 times

Ral17 1 year, 5 months ago
why not D?
upvoted 1 times

szefco 1 year, 2 months 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 thus preferred option.
upvoted 3 times

Sush12 2 years ago
It is B as dataprep has a feature to convert values as desired
upvoted 2 times

NamitSehgal 2 years, 2 months ago
Should be B
DataPrep runs a dataflow job at background.
upvoted 4 times

Tanmoyk 2 years, 5 months ago
Casual approach Dataprep ...and convert null value to numerical value to 0 ...answer B
upvoted 5 times

atnafu2020 2 years, 6 months ago
B
is the correct answer with Trifacta that is data ranging software
upvoted 2 times

atnafu2020 2 years, 6 months ago
Dataprep by Trifacta
upvoted 2 times

haroldbenites 2 years, 6 months ago
B is correct
upvoted 2 times

haroldbenites 2 years, 6 months ago
B is correct
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 1 year, 4 months ago
Dear Admin, almost every answer is incorrect . Please check the comments and update your website.
upvoted 26 times

[Removed] Highly Voted 2 years, 11 months ago
correct: B
upvoted 18 times

[Removed] 2 years, 11 months ago
https://cloud.google.com/security/encryption-at-rest/
upvoted 4 times

tprashanth 2 years, 7 months ago
Based on the info at the link you referred, it seems C is the right answer
upvoted 4 times

baubaumiaomiao 1 year, 2 months ago
If you create it locally, you can't rotate keys. Answer should be B
upvoted 2 times

samdhimal Most Recent 1 month 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 running your Redis and Kafka clusters. This ensures that your data is protected even when it is stored on disk.
upvoted 1 times

samdhimal 1 month 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 1 times

DGames 2 months, 2 weeks 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 3 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%2C%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 1 year, 1 month ago

Selected Answer: B

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

upvoted 1 times

MaxNRG 1 year, 2 months 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 7 times

lg1234 1 year, 4 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 1 year, 5 months ago

Answer is B : both clusters are on GCP, so we can use KMS to manage the keys.

upvoted 2 times

Sakshi_12 1 year, 6 months ago

Well there are two things that a user 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 an 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 a option to create a key yourself in KMS.

https://cloud.google.com/kms/docs
Ans - B

upvoted 3 times

gaco 1 year, 7 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 1 year, 7 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 optionB, GCP KMS creates and manages the key for you.

GCP certification does not always mean use all gcp managed services.

upvoted 3 times

awssp12345 1 year, 7 months ago

Please see https://cloud.google.com/kms/docs/faq#import_keys

upvoted 1 times

awssp12345 1 year, 7 months ago

https://www.testpreptraining.com/tutorial/managing-customer-managed-encryption-keys-with-cloud-kms/

upvoted 2 times

sumanshu 1 year, 7 months ago

https://cloud.google.com/kms/docs/creating-keys#create_a_key_ring

upvoted 2 times

sumanshu 1 year, 8 months ago

Vote for B

upvoted 1 times

daghayeghi 1 year, 11 months ago

B:
https://cloud.google.com/security/encryption-at-rest/
based on information from this document it seems to be CMEK, and not CSEK, since in question said: "you can create, rotate, and destroy" and in the CMEK description said as well.

upvoted 1 times

bobby8521 2 years, 1 month ago

I will recommend B.
Most of the enterprise companies will not recommend to create a Encryption key locally. It doesnt mentioned as On-premise. KMS does everything https://cloud.google.com/kms/docs/destroy-restore
So i will go with B.

upvoted 2 times

xfoursea 2 years, 3 months ago

very bad question, confusing, do not know what it is intended to ask...

upvoted 3 times

aleedrew 2 years, 4 months ago

Confused about this one too. B would be correct according to this doc https://cloud.google.com/kms/docs/cmek?authuser=1. C works too.

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 2 years, 11 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 32 times

[Removed] Highly Voted 2 years, 11 months ago

Answer: C
Description: Why to build own model, Video API with Bigtable is best solution

upvoted 14 times

samdhimal Most Recent 1 month, 1 week 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 for 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 1 times

AzureDP900 1 month, 3 weeks 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 1 year, 2 months ago

Selected Answer: C

C.
The cloud video intelillence api does the label generation without the need of building any model, A and B are excluded. Now, the bbdd 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 1 year, 8 months ago

Vote for C

upvoted 4 times

timolo 1 year, 11 months ago

Answer: C
Reference https://cloud.google.com/video-intelligence/docs/feature-label-detection

upvoted 2 times

daghayeghi 1 year, 11 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 2 years, 2 months ago
Answer: C
upvoted 3 times

Alasmindas 2 years, 3 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 2 years, 5 months ago
Answer is D : BigTable doesnt support JOIN and not built for transactions - https://cloud.google.com/bigtable/docs/overview
upvoted 2 times

Surjit24 2 years, 4 months ago
There are no joins but filtering based on condition.
upvoted 4 times

karthik89 2 years ago
but the requirement involves join as well, it is stated in the problem.
upvoted 2 times

sumanshu 1 year, 7 months ago
Where? Though it's mention - " very fast filtering suggestions" - which means something like dictionary in python OR Key: Value (which is Bigtable)
upvoted 1 times

sraakesh95 1 year, 1 month ago
I think "based on other customer preferences" from the questions requires a join before a filter is applied for collaborative filtering
upvoted 1 times

Deepakd 11 months ago
Recommendation based on other customer"s views cannot be achieved through simple joins. A class pf machine learning algorithms called collaborative filtering is required for that. You need big table to run these algorithms.
upvoted 1 times

haroldbenites 2 years, 6 months ago
Correct C
upvoted 2 times

dg63 2 years, 7 months ago
I doubt if C can be an answer. Will Bigtable allow filtering on labels?
upvoted 2 times

tprashanth 2 years, 7 months ago
Yes, if its part of the rowkey
upvoted 3 times

Rajuuu 2 years, 7 months ago
Answer is C.
upvoted 4 times

Ganshank 2 years, 10 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 (100%)

madhu1171 Highly Voted 2 years, 11 months ago
Answer should be C
upvoted 33 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C - best suitable for the purpose with autoscaling and google recommended transform engine between pubsub n bq
upvoted 22 times

AbdullahAnwar Most Recent 4 days, 2 hours ago
Answer should be C
upvoted 1 times

samdhimal 1 month 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 1 times

zellck 2 months, 3 weeks ago
Selected Answer: C
C is the answer.
upvoted 1 times

odacir 2 months, 3 weeks 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 4 times

zellck 2 months, 3 weeks ago
you need to look at community vote distribution and comments, and not the suggested answer.
upvoted 4 times

medeis_jar 1 year, 1 month ago
Selected Answer: C
C only as referred by MaxNRG
upvoted 3 times

MaxNRG 1 year, 2 months 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 8 times

StefanoG 1 year, 3 months ago

Selected Answer: C

Dataflow, Stackdriver and autoscaling

upvoted 3 times

victorlie 1 year, 3 months ago

Admin, please take a look on the comments. Almost all answers are wrong

upvoted 4 times

nguyenmoon 1 year, 5 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 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 1 year, 6 months ago

Option C is the answer

upvoted 3 times

sumanshu 1 year, 8 months ago

Vote for C

upvoted 1 times

apnu 2 years, 1 month ago

B , it is correct , as it says minimum service cost, dataflow is more expansive than dataproc.

upvoted 2 times

daghayeghi 1 year, 11 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 1 year, 10 months ago

You have to transform the JSON messages. Hence, you need to use dataflow.

upvoted 1 times

apnu 2 years, 1 month ago

B , it is correct as it says minimum service cost, data flow is more expensive than dataproc.

upvoted 2 times

NamitSehgal 2 years, 2 months ago

Should be C. Only concern here is we do not have Window which can monitor size of incoming messages in dataflow.

upvoted 2 times

Tanmoyk 2 years, 5 months ago

Answer should be C

upvoted 2 times

Question #69 *Topic 1*

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

VishalB Highly Voted 2 years, 7 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 58 times

asksathvik 1 year, 6 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 4 times

yoshik 1 year, 5 months ago

logs like stuff goes better on buckets

upvoted 4 times

triipinbee 1 year, 6 months ago

all the option clearly says "storage bucket", once you master that, you'll realize the correct option is A

upvoted 37 times

HarshKothari21 5 months, 3 weeks ago

good one :)

upvoted 1 times

henryCho 2 years, 1 month ago

What about ANSI SQL?

upvoted 6 times

Jphix 1 year, 9 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 up the log data transfer into Google Cloud?", not "How should you set up the querying." ANSI SQL is a distraction!

upvoted 5 times

nadavw 9 months ago

use external table for it

upvoted 1 times

tainangao 1 year, 5 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 7 times

Ganshank Highly Voted 2 years, 10 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 17 times

dambilwa 2 years, 8 months ago
Agreed! - All Options look wrong
upvoted 1 times

dambilwa 2 years, 8 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 1 year, 5 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 2 times

TNT87 2 years, 4 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 8 times

TNT87 2 years, 4 months ago
https://cloud.google.com/bigquery-transfer/docs/youtube-channel-transfer
this link will help to cement the answer.
upvoted 3 times

ayush_1995 Most Recent 1 month 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 month 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 1 times

samdhimal 1 month 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's infrastructure.
upvoted 1 times

samdhimal 1 month 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 1 times

desertlotus1211 1 month, 1 week ago
Big Query is only multi-region in US and EU... question ask for worldwide...
Hence Answer is: A
upvoted 1 times

RoshanAshraf 1 month, 2 weeks ago
Selected Answer: A
Ans - A
Keys - Multi region, TRANSFER FROM "LOCAL FILES" TO GCP.
Bigquery transfer is from Youtube App to BQ.
But the tricky part is Analysis for the log files and SQL which is BQ.
Answers are little confusing
upvoted 1 times

lool 1 month, 2 weeks ago
Selected Answer: A

BQ can't transfer from on-prem infrastructure yet, and it can only transfer to BQ and not a storage bucket. The main purpose of the question is to transfer logs into a bucket, and not on how to analyse it.
https://cloud.google.com/bigquery/docs/dts-introduction

upvoted 1 times

Mkumar43 1 month, 3 weeks ago

Selected Answer: C

Correct answer:C

upvoted 1 times

juliansierra 1 month, 4 weeks ago

The possible sources in Storage Transfer Service are: Cloud Storage, S3, Azure Blob, HTTP and HTTPs URLs and File System. How do you transfer the data from YouTube channels using these possible sources in Storage Transfer Service? For me there is not an exactly correct answer. The best answer could be use BigQuery Data Transfer Service (that accept YouTube Channels data how source) and storage these data in BigQuery datasets and tables, not in Cloud Storage.

upvoted 1 times

slade_wilson 2 months, 1 week ago

Correct Answer : A

The BigQuery Data Transfer Service automates data movement into BigQuery on a scheduled, managed basis. Your analytics team can lay 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. You can also initiate data backfills to recover from any outages or gaps. Currently, you cannot use the BigQuery Data Transfer Service to transfer data out of BigQuery.

upvoted 1 times

lukas_xls 2 months, 2 weeks ago

Selected Answer: C

It says 'Google Cloud', not 'Google Cloud Storage'!

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: A

A is the answer.

upvoted 1 times

Jay_Krish 3 months, 1 week ago

I'm not sure of the answer but from the comments here and logically looking it seems to be A for me. All listed option clearly says "storage bucket" as the final destination, so it is better to use a Storage Transfer service instead of BigQuery transfer. Don't see the benefit of C over A. ANSI SQL is just to divert attention I feel and as many have commented here you could have a table created to read data from the storage.

upvoted 1 times

Lynix 4 months, 1 week ago

Selected Answer: B

https://cloud.google.com/bigquery/docs/dts-introduction
Currently, you cannot use the BigQuery Data Transfer Service to transfer data out of BigQuery.

upvoted 2 times

rr4444 8 months ago

Selected Answer: C

"to perform ANSI SQL and other types of analysis" so it must be C and Bigquery, not GCS object store

upvoted 1 times

GCP_data_engineer 8 months, 3 weeks ago

C is the answer

upvoted 1 times

sw52099 9 months, 3 weeks ago

According to the docs, storage transfer service is for on-premise, another cloud storage bucket or data source from other cloud platform.
https://cloud.google.com/storage-transfer/docs/overview#what-is-storage-transfer

Youtube data is one of the data source of Bigquery Data Transfer service, from https://cloud.google.com/bigquery-transfer/docs/introduction#supported_data_sources

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

Ganshank Highly Voted 2 years, 10 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 the input locations. This can be met using GCS for storage and Bigquery permanent tables with external data source in GCS.
upvoted 47 times

atnafu2020 2 years, 7 months ago
why GCS as external since Bigquery can be used as storage as well?
upvoted 9 times

jkhong 2 months, 1 week ago
The question focuses on "designing storage", rather than designing a data warehouse.
upvoted 1 times

atnafu2020 2 years, 7 months ago
A seems correct for me
upvoted 9 times

atnafu2020 2 years, 6 months ago
Since its best practice, i go by with B not A
upvoted 4 times

gopinath_k 1 year, 11 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] Highly Voted 2 years, 11 months ago
Should be A
upvoted 12 times

tavva_prudhvi 8 months, 2 weeks ago
Not A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue which can be more expensive than Google Storage.
upvoted 3 times

samdhimal Most Recent 1 month 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 1 times

samdhimal 1 month 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 2 months, 1 week 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 1 times

odacir 2 months, 2 weeks ago

Selected Answer: B

C and D are discarted.
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 recommended practices.
BigQuery only support parallel load from storage, storage support parallel load from CLI. So the only option is B.

upvoted 2 times

zellck 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 1 times

nkit 2 months, 3 weeks 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 2 months, 3 weeks 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 2 times

cloudmon 3 months, 3 weeks 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 1 times

Lui1979 9 months, 3 weeks 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 4 times

cloudmon 3 months, 3 weeks ago

Your comment supports A more than B

upvoted 2 times

Didine_22 10 months, 2 weeks ago

Selected Answer: B
B
Because they are talking about the parallel loading from input locations.
upvoted 1 times

devric 10 months, 4 weeks ago
Selected Answer: B
B. The objetive is to follow the best practices.
upvoted 2 times

devric 10 months, 4 weeks ago
Sory I mean A not B :-)
upvoted 1 times

danielttt 11 months, 1 week ago
Selected Answer: B
B fits all the requirements
upvoted 1 times

cloudmon 3 months, 3 weeks ago
So does A
upvoted 1 times

Arkon88 11 months, 3 weeks 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 billing storage costs are accrued to your account since you only get charged for the data that is stored in the data lake, which is comparatively cheaper than storing it in BigQuery
https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro
upvoted 2 times

sfsdeniso 3 months ago
GCP and BigQuery storage costs are the same ~20$ per 1TB per month
upvoted 1 times

rbeeraka 12 months ago
B is better than A as there is no need for Analytics
upvoted 1 times

cloudmon 3 months, 3 weeks ago
They want ANSI SQL queries. In any case, both A and B support analytics.
upvoted 1 times

VishalBule 1 year ago
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 billing storage costs are accrued to your account since you only get charged for the data that is stored in the data lake, which is comparatively cheaper than storing it in BigQuery

A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue which can be more expensive than Google Storage
upvoted 3 times

cloudmon 3 months, 3 weeks ago
Who cares if it takes more time? There are no load time requirements in the question. Putting it in GCS just for that reason is unnecessary in this case. Store it in BQ, query it in BQ. (Answer A)
upvoted 1 times

Mr_Stark 11 months, 2 weeks ago
Yeah, I agree with you.
upvoted 1 times

bury 1 year ago
Selected Answer: A
A, as per medeis_jar
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 2 years, 7 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 30 times

AzureDP900 1 month, 3 weeks 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 2 years, 11 months ago

should be A

upvoted 12 times

samdhimal Most Recent 1 month 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 month, 3 weeks 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 2 months, 3 weeks 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 3 months, 3 weeks 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 7 months, 2 weeks 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 9 months, 3 weeks ago

Selected Answer: A

Vote for A

upvoted 1 times

bury 1 year ago

Selected Answer: A

a is correct

upvoted 1 times

JayZeeLee 1 year, 3 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 1 year, 8 months ago

Vote for A

upvoted 3 times

haroldbenites 2 years, 6 months ago

A is correct

upvoted 4 times

[Removed] 2 years, 11 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 1 year, 11 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 39 times

[Removed] Highly Voted 2 years, 11 months ago

Should be C

upvoted 29 times

samdhimal Most Recent 1 month 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 1 times

samdhimal 1 month 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 month, 2 weeks 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 1 times

rivua 2 months, 2 weeks ago

The 'correct' answers on this platform are ridiculous

upvoted 4 times

zellck 2 months, 3 weeks 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 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, a table is created in a special dataset and will be available for approxi- mately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse

upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 2 months 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 temorary.
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 1 year, 2 months 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 1 year, 2 months ago

Answer is A. While C seems the most reasonable answer there are 2 points to notics: 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 1 year, 3 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%2Dtime%2C%20ad%2Dhoc%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20(ETL)%20processes.,-Querying%20Cloud%20Bigtable

upvoted 1 times

tsoetan001 1 year, 4 months ago

C is the answer.

upvoted 1 times

Ysance_AGS 1 year, 5 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 1 year, 5 months ago

Vote for C

upvoted 1 times

gcp_learner 1 year, 7 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

Yiouk 1 year, 6 months ago

https://cloud.google.com/bigquery/docs/writing-results#temporary_and_permanent_tables

upvoted 1 times

triipinbee 1 year, 6 months ago

Storage cost of data for BQ is the same as standard cloud storage, actually less for long term storage as it automatically moves to nearline storage.

https://cloud.google.com/bigquery/pricing#storage
https://cloud.google.com/storage#section-10

upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for C

upvoted 2 times

sumanshu 1 year, 7 months ago

Vote for 'D'

BigQuery uses temporary tables to cache query results.
https://cloud.google.com/bigquery/docs/writing-results

upvoted 1 times

sumanshu 1 year, 7 months ago

There are no storage costs for temporary tables, but if you write query results to a permanent table, you are charged for storing the data.

upvoted 1 times

triipinbee 1 year, 6 months ago

even if it's a temp table, the storage cost for active storage in cloud bucket and the standard storage in BQ is the same. So no point creating temp tables.

upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for 'C'

upvoted 1 times

sumanshu 1 year, 7 months ago

Vote for D

BigQuery uses temporary tables to cache query results.

https://cloud.google.com/bigquery/docs/writing-results

upvoted 1 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] Highly Voted 2 years, 11 months ago
Answer: C
Description: Spanner allows transaction tables to scale horizontally and secondary indexes for range queries
upvoted 26 times

[Removed] Highly Voted 2 years, 11 months ago
Correct: C
upvoted 9 times

PolyMoe Most Recent 1 month 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 the 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 month 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 the 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 month 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

samdhimal 1 month 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

zellck 2 months, 3 weeks 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 5 months, 2 weeks ago

Selected Answer: C

As sumanshu said

upvoted 1 times

tsoetan001 1 year, 4 months ago

Answer: C

upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for C

upvoted 4 times

sumanshu 1 year, 7 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 1 year, 11 months ago

Answer: C
https://cloud.google.com/spanner/docs/secondary-indexes

upvoted 1 times

Nileshk611 2 years, 2 months ago

Correct: C

upvoted 3 times

arghya13 2 years, 3 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

zellck Highly Voted 2 months, 3 weeks 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+ APIs and offers a Bigtable HBase client in Maven, so it is easy to use Bigtable with Dataproc.

upvoted 6 times

Atnafu 2 months, 1 week ago
Hbase concept here us beautiful

upvoted 1 times

samdhimal Most Recent 1 month ago
Why not Biquery?

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 it 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 month 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 1 times

samdhimal 1 month 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 1 times

samdhimal 1 month ago

Can someone clarify why Bigtable and Not Bigquery? Super Confused.

upvoted 1 times

desertlotus1211 1 month, 1 week ago

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

upvoted 2 times

Yazar97 3 months, 1 week ago

Time series data = Bigtable... So it's A

upvoted 3 times

Jay_Krish 3 months, 1 week ago

Selected Answer: A

Option A seems right

upvoted 1 times

drunk_goat82 3 months, 1 week ago

Selected Answer: A

Big Table has a HBase compliant API and is transactional unlike GCS.

upvoted 1 times

solar_maker 3 months, 1 week 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 3 months, 3 weeks ago

Selected Answer: C

It must be C because of the existing Hadoop jobs

upvoted 3 times

cloudmon 3 months, 3 weeks ago

On 2nd thought, it's Bigtable: https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-bigtable

upvoted 6 times

pluiedust 4 months, 3 weeks ago

Selected Answer: C

I think it is C

upvoted 2 times

maia01 5 months ago

Selected Answer: C

Use Datarproc with Cloud Storage in combo with HDFS
https://cloud.google.com/dataproc/docs/concepts/dataproc-hdfs

upvoted 2 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: A

A. Cloud Bigtable

upvoted 4 times

YorelNation 5 months, 4 weeks ago

Selected Answer: A

A but not sure about the existing hadoop jobs

upvoted 4 times

Question #75 *Topic 1*

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

[Removed] Highly Voted 4 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 8 times

musumusu Most Recent 1 week, 6 days ago

I will go with A, as i wanna save cost, dont need to create separate dataset for permanent storage.

upvoted 1 times

samdhimal 1 month 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 month 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 1 times

Rupendra06 1 month ago

Selected Answer: B

Ensure the analysis cost for other projects is assigned to those projects indicates B is the correct answer.

upvoted 1 times

GCPpro 1 month, 1 week ago

B is the correct answer

upvoted 1 times

Kyr0 2 months ago

Selected Answer: A
I would say 1 too
upvoted 1 times

slade_wilson 2 months, 1 week ago
Selected Answer: B
B is the correct approach.
upvoted 1 times

wan2three 2 months, 2 weeks ago
I think its B, as projects need to be charged separately
upvoted 1 times

zellck 2 months, 3 weeks 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 3 times

Wonka87 2 months 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? In option A, isn't is 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

Gudwin 2 months, 4 weeks ago
That's ambiguous. While A is correct, B is the recommended approach:
"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."

Bit it doesn't say "authorised view" in B.
upvoted 3 times

Wonka87 2 months 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? In option A, isn't is 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

dish11dish 3 months ago
Selected Answer: A
As they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects.

optaion A is most feasible solution as it just create authorized view with aggregated data so storage cost can me minimize also this won't compromise the access.
upvoted 3 times

gudiking 3 months, 1 week ago
Selected Answer: B
B, as Joey_chicity explains
upvoted 1 times

sedado77 5 months, 2 weeks ago
Selected Answer: A
A is correct
upvoted 3 times

AWSandeep 5 months, 3 weeks ago
Selected Answer: A
A. Create and share an authorized view that provides the aggregate results.
upvoted 1 times

Question #76 *Topic 1*

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

Mitra123 Highly Voted 1 year, 11 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 40 times

[Removed] Highly Voted 2 years, 11 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 19 times

awssp12345 1 year, 7 months ago
The question has no mention of ANALYZE.. BQ is not correct. I would go with D.
upvoted 10 times

sraakesh95 1 year, 1 month ago
There is no option for archiving with BQ
upvoted 1 times

tavva_prudhvi 10 months, 2 weeks 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 7 months, 2 weeks 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 2 times

aleixfc96 Most Recent 4 weeks, 1 day ago
It is so clear that is B lol
upvoted 1 times

NamitSehgal 1 month ago
B bigquery for a record set store
upvoted 1 times

PolyMoe 1 month 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 1 times

samdhimal 1 month 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

samdhimal 1 month 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 not provide an auditable record of access to the data.

upvoted 1 times

GCPpro 1 month, 1 week ago

D is the correct answer

upvoted 1 times

RoshanAshraf 1 month, 2 weeks 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 2 months ago

Selected Answer: D

I will go with D

upvoted 1 times

DGames 2 months, 2 weeks 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

zellck 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 2 times

Wonka87 2 months ago

are you assuming that data is in BQ compatible format?

upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: A

Just store it in GCS. Cloud Audit logging is enforced, and will provide the required auditability logs. D is overkill.

upvoted 1 times

z83j 3 months, 2 weeks ago

In what world is user-supplied encryption keys and seperate decryption key to each user account for audibility records? It's also just bad practice to give everyone a personal key and hope they don't get stolen somewhere down the road from any of the users.

upvoted 2 times

cloudmon 3 months, 3 weeks ago

Reference: https://cloud.google.com/storage/docs/audit-logging#available-logs

upvoted 1 times

clouditis 5 months, 1 week ago

Its B, question is asking where will the data set reside

upvoted 2 times

crismo04 5 months, 2 weeks ago

Selected Answer: B

"record of ACCESS to certain types of data." It's not the data, it's the access of it, so no image or video data is stored, I go with B

upvoted 2 times

Azlijaffar 5 months, 3 weeks ago

Certain types of data is subjected to this mandate. Data can be audio files etc. You need to keep these files on Cloud Storage and not BigQuery. D says the data is stored in a bucket in GCS (correct) and users that access these data will be provided with a link to the storage bucket of that data after their user information is logged for accessing said data.
Should be D.

upvoted 2 times

Azlijaffar 4 months, 1 week ago

"where should you store DATA that is subject to that mandate?". Where to store the DATA, not the record or the logs. The DATA that could be of CERTAIN TYPES. At least that's how I understand this tricky question. Will go with D.

upvoted 1 times

ducc 6 months, 1 week ago

Selected Answer: D

There are audio, image,...

upvoted 2 times

alecuba16 8 months ago

Selected Answer: B

-Create a dataset with the permissions
-Create a federated table with GCS as source (the logs are already in GCS)
-Set audit on

upvoted 1 times

ducc 6 months ago

Record from government, there might be Audio, Photo,...

upvoted 1 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

MingSer 5 days, 8 hours ago

Selected Answer: B

all other are wrong

upvoted 1 times

PolyMoe 1 month ago

Selected Answer: B

of course !

upvoted 1 times

samdhimal 1 month ago

B. Subsampling your training dataset can decrease the amount of data the model needs to process and can speed up training time. However, it 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 1 times

GCPpro 1 month, 1 week ago

B is the answer as we are bothered about speed not the accuracy.

upvoted 1 times

ler_mp 1 month, 2 weeks 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 2 months, 1 week 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 the model.

upvoted 3 times

DGames 2 months, 2 weeks 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 to go.

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 1 times

pluiedust 4 months, 3 weeks ago

Selected Answer: B

It is B. D would improve the accuracy, not speed.

upvoted 4 times

Chavoz 5 months ago

It's B. D Would be for increase performance

upvoted 1 times

crismo04 5 months, 2 weeks ago

if you Increase the number of layers, you increase the training time, right?

upvoted 1 times

HarshKothari21 5 months, 3 weeks ago

Both B and D seems correct.

upvoted 1 times

jkhong 2 months, 2 weeks ago

Increasing D will increase training time

upvoted 2 times

YorelNation 5 months, 4 weeks ago

Selected Answer: B

Only valid awnser

upvoted 2 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 (100%)

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: Pig is scripting language which can be used for checkpointing and splitting pipelines
upvoted 21 times

[Removed] Highly Voted 2 years, 11 months ago
Should be A
upvoted 14 times

musumusu Most Recent 1 week, 5 days 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, restart 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 1 times

samdhimal 1 month 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 month 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 1 times

vishal0202 5 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 5 months, 3 weeks ago
Selected Answer: A
Description: Pig is scripting language which can be used for checkpointing and splitting pipelines
upvoted 1 times

BigDataBB 1 year ago
Why not D?
upvoted 1 times

rbeeraka 1 year, 1 month ago
Selected Answer: A
PigLatin supports checkpoints
upvoted 1 times

davidqianwen 1 year, 1 month ago
Selected Answer: A
Answer: A
upvoted 1 times

maddy5835 1 year, 4 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 1 year, 8 months ago
Vote for A
upvoted 1 times

kdiab 2 years 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 2 years, 5 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 2 years, 4 months ago
Did you take the exam? I am ready to do it this month
upvoted 1 times

MaxNRG 1 year, 2 months ago
seems like a very old question :)
not sure it's actual
upvoted 1 times

haroldbenites 2 years, 6 months ago
A is correct
upvoted 1 times

saurabh1805 2 years, 6 months ago
A should be correct answer here.
upvoted 2 times

dg63 2 years, 7 months ago
A, C and D - all are valid answers. You can do checkpointing, splitting pipelines and merging pipelines with all three options.
upvoted 8 times

norwayping 2 years, 8 months ago
A
https://books.google.no/books?id=uL59BAAAQBAJ&pg=PA263&lpg=PA263&dq=pig+checkpoint+hadoop&source=bl&ots=8mqgffDBTQ&sig=ACfU3U0NDfWpLKsQcdtjwIor3f85_m8zJA&hl=en&sa=X&ved=2ahUKEwihhq61maTqAhXSwqYKHS7JDXkQ6AEwAHoECAkQAQ#v=onepage&q=pig%20checkpoint%20hadoop&f=false
upvoted 2 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

A (50%) C (50%)

[Removed] Highly Voted 2 years, 11 months ago
correct: C
upvoted 15 times

[Removed] Highly Voted 2 years, 11 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 12 times

Subhajeetpal Most Recent 2 days, 12 hours ago
Answer is C
upvoted 1 times

mahdiaqim 6 days, 19 hours 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 month 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 1 times

zellck 2 months, 3 weeks ago
Selected Answer: C
C is the answer.
upvoted 1 times

Nirca 4 months, 2 weeks 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 8 months ago
"The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. "

This makes zero sense. Is it to GCS or GCE? Question had to make up its mind. Nonsense, literally.
upvoted 3 times

Thierry_1 1 year, 3 months ago
Vote for C. Mostly because other options seems useless
upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for 'C'

You want to maximize transfer speeds.

upvoted 3 times

Tolberic 2 years ago

https://cloud.google.com/compute/docs/machine-types#n1_machine_types

upvoted 2 times

snamburi3 2 years, 3 months ago

"The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP."
This is confusing, as the upload is through a data transfer server on GCP - not directly to Storage. In this case, maybe A?

upvoted 1 times

snamburi3 2 years, 3 months ago

I am going with C as I found a doc: https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#increasing_network_bandwidth. still a confusing question...

upvoted 4 times

SureshKotla 2 years, 5 months ago

I guess option A because, increasing processing power for parallel uploads is the first thing to try. And if it doesn't work, then go to bandwidth issues.

upvoted 1 times

bbozz_ 2 years, 7 months ago

Answer: CA

upvoted 2 times

Callumr 2 years, 8 months ago

B. Increasing the disk size will increase the write performance and will enable better parallel throughput - this is an easy first thing to try

upvoted 3 times

ayush955 2 years, 6 months ago

I don't think so, parallel throughput depends upon processor rather than the disk size

upvoted 2 times

atnafu2020 2 years, 6 months ago

increasing disk or resizing it will have downtime

upvoted 1 times

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 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.
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 2 years, 11 months ago
correct: C
upvoted 17 times

atnafu2020 Highly Voted 2 years, 7 months ago
C
Bigquery and Big table =PB storage capacity
Bigtable=to read scan rows Big query select row to read
upvoted 7 times

dconesoko Most Recent 1 month, 1 week ago
Selected Answer: C
bigquery and bigtable
upvoted 1 times

sandipk91 1 year, 6 months ago
C is correct, no doubt
upvoted 1 times

sumanshu 1 year, 8 months ago
Vote for C
upvoted 2 times

daghayeghi 1 year, 11 months ago
C:
They need to do aggregations over their petabyte-scale datasets: Bigquery
They need to scan specific time range rows with a very fast response time (milliseconds): Bigtable
upvoted 5 times

Insane7 2 years, 4 months ago
Why not D? Biqquery and GCS.
Also Big Table is no sql, where as BQ is SQL
upvoted 1 times

Gcpyspark 2 years, 2 months ago
With GCS you can only scan the rows from BigQuery using External federated Datasources, with that millisecond latency is not possible. Also "scan specific time range rows with a very fast response time" is a natural fit use case for Cloud Bigtable.
upvoted 4 times

haroldbenites 2 years, 6 months ago
C is correct.
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 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.
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 (62%) B (38%)

[Removed] Highly Voted 2 years, 11 months ago

Should be B

upvoted 26 times

PolyMoe Most Recent 1 month 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 2 months, 4 weeks ago

Selected Answer: B

B
But can someone explain the question and selection clearly?

upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: B

It's B. All the other choices are unreasonable.

upvoted 4 times

edwardlin421 3 months, 3 weeks ago

ACD-Design for each possible combination of criteria, so if your team has new requirements, you must design new charts.
So, answer shoud be B.

upvoted 1 times

ducc 6 months ago

Selected Answer: D

the key is " You want to avoid creating and updating new visualizations each month."
only D work for that phrase

upvoted 2 times

wan2three 2 months 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 8 months ago

must be D

upvoted 1 times

ealpuche 8 months ago

Selected Answer: D

The answer is B

upvoted 1 times

rr4444 8 months ago
This Q feels very disconnected from GCP products.....
upvoted 1 times

sw52099 9 months, 3 weeks 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 2 months ago
But q says the data sources only have latest 6 weeks data, so current data means latest?
upvoted 1 times

RRK2021 1 year ago
B is optimal to avoid creating and updating new visualizations each month
upvoted 1 times

ManojT 1 year, 4 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 1 year, 6 months ago
Yes it's B
upvoted 2 times

sumanshu 1 year, 8 months ago
vote for B
upvoted 1 times

reima990 2 years, 4 months ago
Yes agreed. B seems to be the only reasonable choice here.
upvoted 4 times

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 2 years, 11 months ago
Correct - B
upvoted 17 times

piotrpiskorski Most Recent 3 months ago
Selected Answer: B
always partion large tables
upvoted 1 times

Thierry_1 1 year, 3 months ago
B for sure
upvoted 3 times

nguyenmoon 1 year, 5 months ago
Correct is B
upvoted 3 times

sandipk91 1 year, 6 months ago
Option B for sure
upvoted 2 times

awssp12345 1 year, 7 months ago
https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard - Supports B
upvoted 2 times

sumanshu 1 year, 8 months ago
Vote for 'B' Partitioned Table for Faster Query and Low cost (because it will process less data)
upvoted 1 times

alonsoRios 1 year, 11 months ago
B is correct
upvoted 2 times

fabenavideso 2 years, 1 month ago
Correct - B
upvoted 2 times

ceak 2 years, 2 months ago
should be C
upvoted 1 times

lammingtons 2 years, 1 month ago
They're using BigQuery so partitioning is the better choice here. B
upvoted 3 times

haroldbenites 2 years, 6 months ago
B is correct
upvoted 3 times

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 2 years, 11 months ago
Seems like A..Data should ingest from multiple sources which might be real time or batch .
upvoted 22 times

navemula 1 year, 7 months ago
How is it possible to query in real time with option A. It needs Dataflow
upvoted 2 times

navemula 1 year, 7 months ago
To use Dataflow SQL it needs BigQuery
upvoted 1 times

mikey007 Highly Voted 2 years, 5 months ago
Repeated Question see ques 35
upvoted 11 times

awssp12345 1 year, 7 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 2 years, 1 month ago
Well Done mikey007, Many people have already answered as A.
upvoted 2 times

Kyr0 Most Recent 2 months ago
Selected Answer: A
Answer is A
upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: A

It's A

upvoted 1 times

ducc 6 months ago

Selected Answer: A

A is the answer

upvoted 1 times

RRK2021 1 year 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 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: A

PubSub (for global ingestion from multiple sources) + Dataflow (for process and query) + reliable (gcs).

upvoted 1 times

lifebegins 1 year, 3 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 1 year, 4 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 1 year, 4 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

cualquiernick 8 months, 1 week 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 1 year, 5 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 1 year, 8 months ago

Vote for 'A'

SQL - will not handle the volume

upvoted 2 times

daghayeghi 1 year, 11 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 2 years, 5 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 "Flowlogistic'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 2 years, 5 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 2 years, 4 months ago
Dataflow SQL cannot output to cloud storage only BigQuery...so I am confused on this one.
upvoted 2 times

Jay3244 1 year, 12 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 1 year, 11 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 1 year, 5 months ago
Answer should be C then?
upvoted 2 times

kuntal8285 2 years, 5 months ago
should be E
upvoted 1 times

Tanmoyk 2 years, 5 months ago
Should be A ...data need to feed to the propriority system and for that dataflow is required.
upvoted 2 times

Question #84 *Topic 1*

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 2 years, 11 months ago
C is the only way which all records will be compared.
upvoted 27 times

odacir 2 months, 3 weeks ago
Agree with your argument
upvoted 1 times

[Removed] Highly Voted 2 years, 11 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 14 times

musumusu Most Recent 1 week, 5 days ago
key words here- Hash or collect value on "EACH table", after sorting the table.
Option C
upvoted 1 times

samdhimal 4 weeks, 1 day 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 4 weeks, 1 day 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 1 times

zellck 2 months, 3 weeks ago
Selected Answer: C
C is the answer.
upvoted 2 times

odacir 2 months, 3 weeks ago
All records need to be checked to be sure, so C is the answer
upvoted 1 times

Leeeeee 3 months ago
Selected Answer: C
All records
upvoted 1 times

hfuihe 4 months, 2 weeks ago

Selected Answer: B

B is the only way which all records will be compared.

upvoted 1 times

cloudmon 3 months, 3 weeks ago

You must have meant to say C

upvoted 2 times

medeis_jar 1 year, 1 month ago

Selected Answer: C

HASH() to compare data skipping dates and timestamps

upvoted 1 times

stefanop 10 months ago

The hash in answer C is used to select a sample of the table, not to compare them

upvoted 1 times

stefanop 10 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 1 year, 2 months 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, but 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 1 year, 3 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. So C.

upvoted 1 times

u_t_s 1 year, 4 months ago

Since there is no PK and it is possible that set of values is commons in some records which result in same hashkey for those records. But still Anwer is C

upvoted 3 times

sumanshu 1 year, 8 months ago

Vote for 'C"

upvoted 1 times

daghayeghi 2 years 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 1 year, 5 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 2 years, 6 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 on your data, and measure performance from your local computer to Cloud Storage.

upvoted 3 times

haroldbenites 2 years, 6 months ago

C is correct

upvoted 4 times

haroldbenites 2 years, 6 months ago

It Says: "...that they are identical." , You must not use sample.

upvoted 3 times

Rajuuu 2 years, 7 months ago

C is correct.

upvoted 4 times

dambilwa 2 years, 8 months ago

Option [C] is most appropriate

upvoted 4 times

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

sedado77 Highly Voted 5 months, 1 week ago

Selected Answer: C
I got this question on sept 2022. Answer is C
upvoted 5 times

samdhimal Most Recent 3 weeks, 6 days ago

C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.
upvoted 1 times

samdhimal 3 weeks, 6 days ago

Switching to flat-rate pricing would allow you to ensure a consistent level of service and avoid running into the on-demand slot quota per project. Additionally, by establishing a hierarchical priority model for your projects, you could allocate resources based on the specific needs and priorities of each business unit, ensuring that the most critical queries are executed first. This approach would allow you to balance the needs of each business unit while maximizing the use of your BigQuery resources.
upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: C
C is the answer.

https://cloud.google.com/bigquery/docs/reservations-intro
The benefits of using BigQuery Reservations include:
- Workload management. After you purchase slots, you can allocate them to workloads. That way, a workload has a dedicated pool of BigQuery computational resources available for use. At the same time, if a workload doesn't use all of its allocated slots, any unused slots are shared automatically across your other workloads.

- Centralized purchasing: You can purchase and allocate slots for your entire organization. You don't need to purchase slots for each project that uses BigQuery.

upvoted 1 times

hybridpro 2 months, 4 weeks ago

C. https://cloud.google.com/bigquery/quotas - 2000 is the max no. of slots

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 2 years, 10 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 29 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: Question says mirroring and avoid kafka connect plugins
upvoted 9 times

samdhimal Most Recent 3 weeks, 2 days 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 3 weeks, 2 days 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 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 1 times

samdhimal 3 weeks, 2 days 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 3 days, 18 hours 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 3 weeks, 2 days 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 the 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 3 weeks, 1 day ago

Sorry. I messed up. The answer is probably A. My badd....

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: A

A is the answer.

upvoted 1 times

Afonya 4 months, 1 week ago

Selected Answer: A

"The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins."

upvoted 1 times

somnathmaddi 4 months, 2 weeks ago

D is the right answer

upvoted 1 times

clouditis 5 months, 1 week ago

D is the right answer

upvoted 1 times

hendrixlives 1 year, 2 months 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 1 year, 4 months ago

Going with "D"

Refer: https://stackoverflow.com/questions/55277188/kafka-to-google-pub-sub-using-sink-connector

upvoted 3 times

baubaumiaomiao 1 year, 2 months ago

"avoid deployment of Kafka Connect plugins"

upvoted 1 times

sup12345 1 year, 2 months ago

you're actually an idiot

upvoted 2 times

ML_Novice 12 months ago

stop with the swearig shame on you, you should be banned

upvoted 2 times

JACKYLU 7 months, 1 week ago

GHETTO TALK

upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for A

upvoted 1 times

daghayeghi 1 year, 11 months ago

Answer: A
Description: Question says mirroring to avoid kafka connect plugins

upvoted 3 times

Allan222 2 years ago

Correct is D

upvoted 1 times

sumanshu 1 year, 7 months ago
As per question - "avoid deployment of Kafka Connect plugins."
upvoted 1 times

vakati 2 years, 5 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 2 years, 5 months ago
D should be the correct answer. Configure pub/sub as sink
upvoted 4 times

haroldbenites 2 years, 6 months ago
C is correct.
https://docs.confluent.io/current/connect/kafka-connect-gcp-pubsub/index.html
upvoted 2 times

haroldbenites 2 years, 6 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 1 year, 8 months ago
Question says : avoid deployment of Kafka Connect plugins.
upvoted 2 times

clouditis 2 years, 6 months ago
its D, why would google prefer Kafka in their own cert questions! :)
upvoted 3 times

Ral17 1 year, 5 months ago
Because the questions mentions to avoid deployment of Kafka connect plugins
upvoted 3 times

Archy 2 years, 7 months ago
answer is D, as on-prime kafka support sink connector for outging data.
upvoted 1 times

Question #87 *Topic 1*

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 (58%) A (35%) 8%

rickywck Highly Voted 2 years, 11 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 56 times

raf2121 1 year, 7 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 SDD for Master nodes)
https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs
upvoted 8 times

rr4444 8 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 1 year, 7 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 2 months, 3 weeks 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 1 times

odacir 2 months, 3 weeks ago
https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance
upvoted 1 times

diluvio 1 year, 4 months ago
It is A . please read the links above
upvoted 4 times

madhu1171 Highly Voted 2 years, 11 months ago
Answer should be D
upvoted 12 times

jvg637 2 years, 11 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 14 times

ch3n6 2 years, 8 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 2 years, 7 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 2 years, 6 months ago
File sizes are already within the expected range for GCS (128MB-1GB) so not C.
D seems most feasible
upvoted 3 times

SuperTee Most Recent 4 days, 6 hours ago
Selected Answer: A
Cost sensitive is the keyword.
upvoted 1 times

musumusu 1 week, 5 days ago
this question asked by Google, So option C is not correct otherwise, good approach to use initial data in hdfs and swtich from HDD to SDDs for 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 1 times

PolyMoe 1 month 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 questions asks to continue using preemtibles ==> not D
- C. is a good solution.
upvoted 2 times

ayush_1995 1 month 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 the boot disk size of the preemptible VMs may not be necessary or cost-effective for this particular workload.
upvoted 1 times

slade_wilson 2 months, 1 week 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 1 times

odacir 2 months, 3 weeks 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 1 times

zellck 2 months, 3 weeks 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 3 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 1 times

dish11dish 3 months, 1 week 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 2 times

piotrpiskorski 3 months, 1 week 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 2 months, 3 weeks 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 3 months, 1 week ago

Selected Answer: A
https://www.dremio.com/blog/tuning-parquet/
upvoted 1 times

cloudmon 3 months, 3 weeks 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 3 months, 2 weeks 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 1 times

josrojgra 4 months, 1 week ago

Selected Answer: D
I choose D because the files already fit at the recommended size (https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#manage_cloud_storage_file_sizes).

And to improve the performance with a shuffle-heavy workload is recommended to increase the boot disk of the preemptible VMs (https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#use_preemptible_vms)
upvoted 1 times

clouditis 5 months, 1 week ago
A it is, SSD makes it expensive & defeats the question's purpose of no increased cost requirement
upvoted 1 times

John_Pongthorn 5 months, 1 week ago

Selected Answer: D
D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size

Read the section, it is pretty straightforward.
https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance
upvoted 1 times

ducc 6 months, 1 week ago

Selected Answer: D

agree with gcd56

upvoted 1 times

Question #88 *Topic 1*

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 (55%) C (45%)

jonathanthezombieboy 1 week, 5 days 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 a 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 1 times

musumusu 1 week, 5 days 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 lets follow standard sideOutput(withoutputs in the code)
syntax be like in pipeline:
'ProcessData' >> beam.ParDo(DoFn).with_outputs

upvoted 1 times

musumusu 3 days, 17 hours 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 2 weeks, 5 days ago

Selected Answer: D

blahblahblahblahblahblahblahblah

upvoted 1 times

waiebdi 4 weeks, 1 day 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 sent 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 2 months, 3 weeks ago

Selected Answer: C

C is the answer.

upvoted 1 times

hauhau 2 months, 4 weeks ago

C
D: dataflow to pub/sub is weird
````

upvoted 1 times

Atnafu 3 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 2 times

sfsdeniso 3 months ago

Answer is D

upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: C

It's C.
In D, "storing to PubSub later" doesn't really make sense.

upvoted 2 times

devaid 4 months, 2 weeks ago

Selected Answer: C

Answer is C. You need to reprocess all the failling 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 5 months ago

Answer C

upvoted 4 times

nickyshil 5 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 why 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 4 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: D

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.

upvoted 3 times

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_GEOGPOINT` 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

PolyMoe 1 month 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 bucketing 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

zellck 2 months, 3 weeks 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 2 times

dish11dish 3 months, 1 week 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 3 times

AHUI 5 months ago
Ans C, use L1 regularization becuase we know the feature is a strong feature. L2 will evenly distribute weights

upvoted 3 times

crismo04 5 months, 2 weeks ago

https://medium.com/riga-data-science-club/geographic-coordinate-encoding-with-tensorflow-feature-columns-e750ae338b7c#:~:text=to%20the%20rescue!-,Feature%20Crosses,-Combining%20features%20into

upvoted 2 times

crismo04 5 months, 2 weeks ago

Feature cross seems to be the right feature option

upvoted 1 times

crismo04 5 months, 2 weeks ago

So it's B option

upvoted 2 times

[Removed] 5 months, 3 weeks ago

Selected Answer: C

Regularization + location into one

upvoted 1 times

AWSandeep 5 months, 4 weeks 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 4 times

nwk 5 months, 4 weeks ago

C or D?
https://medium.com/riga-data-science-club/geographic-coordinate-encoding-with-tensorflow-feature-columns-e750ae338b7c

upvoted 1 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 (79%) A (21%)

Barniyah Highly Voted 2 years, 10 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 25 times

fire558787 1 year, 5 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 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 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
upvoted 11 times

[Removed] Highly Voted 2 years, 11 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 1 year, 6 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

slade_wilson Most Recent 2 months, 1 week 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 2 months, 3 weeks ago
Selected Answer: D
D is the answer.

https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/third-party/mariadb
upvoted 3 times

wan2three 2 months ago
For supplement, 'Stackdriver agent' now called as Ops agent, 'Operations Suite'
upvoted 2 times

dish11dish 3 months, 1 week 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 2 times

girgu 4 months ago

Selected Answer: D

https://cloud.google.com/monitoring/agent/ops-agent/third-party/mariadb

upvoted 2 times

clouditis 5 months, 1 week ago

C is the answer, fluentd plug in is needed as the DB is on GCE

upvoted 2 times

ducc 6 months ago

Selected Answer: D

go for D

upvoted 1 times

eRaymox 6 months, 1 week 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 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

upvoted 2 times

Kriegs 8 months, 1 week 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 minimal 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

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 10 months, 1 week 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 10 months, 2 weeks 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 9 months, 1 week 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 1 year, 1 month ago

Selected Answer: A

Opencensus Agent is right one

upvoted 2 times

nguyenmoon 1 year, 5 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] 1 year, 1 month ago

agree; https://cloud.google.com/monitoring/agent/ops-agent/third-party/mariadb#configure-instance

upvoted 2 times

safiyu 1 year, 6 months ago

Answer : D
mariaDB is an extension of mysql and mysql plugin must work fine to extract the metrics of mariaDB.

https://cloud.google.com/monitoring/agent/plugins/mysql

upvoted 6 times

sumanshu 1 year, 8 months ago

Vote for C

upvoted 2 times

bobby8521 2 years ago

should be D
https://cloud.google.com/monitoring/api/metrics_agent#agent-mysql - This plugins collects the replication status and other metrics can be collected from Stackdriver agent. And also MariaDB was not listed in that as this can be used as MySQL plugin.

upvoted 6 times

Question #91 *Topic 1*

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

GHN74 Highly Voted 2 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 33 times

sergio6 1 year, 4 months ago
Information in social profiles are public
upvoted 1 times

sergio6 1 year, 4 months ago
according to the privacy settings and shareable informations
upvoted 1 times

sumanshu Highly Voted 1 year, 10 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 14 times

szefco 1 year, 2 months 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 16 times

musumusu Most Recent 1 week, 5 days 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 month ago
Selected Answer: B
default rates can be predicted with linear regression.
upvoted 1 times

Whoswho 2 months ago
Answer should actually be a logistic Regression model
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: B
B is the answer.
upvoted 1 times

ladistar 2 months, 4 weeks 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 1 times

woyaolai 3 months, 3 weeks ago

I used to be a Credit Risk modeler and I think this question is stupid.

upvoted 4 times

Azlijaffar 4 months, 1 week 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 wont have considered profiles of applicants that normally would not be granted those loans in the first place. Am I missing something here?

upvoted 1 times

Azlijaffar 4 months, 1 week ago

So answer should be C i think.

upvoted 1 times

devaid 4 months, 2 weeks 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

John_Pongthorn 5 months, 1 week ago

Selected Answer: B

Predict default rate, it seem to be numberic value , it is not y Or n
so it is regression
default risk score = default rate

upvoted 1 times

ducc 6 months ago

Selected Answer: B

I did work for a real project
We always use linear regression or AUC for predicting default in loan processes.
So I go for B

upvoted 2 times

ducc 6 months, 1 week ago

Selected Answer: B

I vote for B

upvoted 1 times

changsu 7 months, 3 weeks ago

Selected Answer: B

B or D, and I think B because the credit score already means a profile and no more data given but approved applications and defaults.

upvoted 3 times

cualquiernick 8 months, 1 week ago

A is incorrect, there is no need of more data
B Possiby correct answer is the default rates are numeric (int)
C Not realy a solution
D) Maybe the right answer, but not sure what social profiles has to be in relation of default payments.....
D

upvoted 1 times

hello_world404 8 months, 2 weeks ago

I agree it's B

upvoted 1 times

Georgesaju 9 months, 1 week ago

Selected Answer: B

B is correct

upvoted 1 times

Question #92

Topic 1

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 2 years, 11 months ago
Answer - D
upvoted 22 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
Description: Cloud SQl cheap and relational DB.
upvoted 13 times

zellck Most Recent 2 months, 3 weeks ago
Selected Answer: D
D is the answer.
upvoted 1 times

Nirca 4 months, 2 weeks ago
Selected Answer: D
Cloud SQL is relational DB (pg mssql, mysql)
upvoted 1 times

Dhass 8 months, 3 weeks ago
Answer - D
upvoted 1 times

homaj 9 months, 3 weeks ago
Selected Answer: D
answer D
upvoted 1 times

sumanshu 1 year, 8 months ago
Vote for D
upvoted 3 times

daghayeghi 1 year, 11 months ago
D:
https://cloud.google.com/sql/docs/features
upvoted 1 times

GypsyMonkey 2 years, 2 months ago
D, cloud SQL is a relational database; if > 10tb, then choose spanner
upvoted 3 times

atnafu2020 2 years, 6 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 1 year, 3 months ago
64TB AS OF TODAY
upvoted 5 times

haroldbenites 2 years, 6 months ago

D is correct. Obviously

upvoted 3 times

Barniyah 2 years, 10 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

taepyung 2 years, 10 months ago

At this moment, Cloud SQL is providing up to 30,720GB(about 30TB)
So I think it's D.

upvoted 12 times

Barniyah 2 years, 9 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 2 years, 2 months 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

Question #93 *Topic 1*

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

[Removed] Highly Voted 5 months, 3 weeks 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 11 times

[Removed] 3 months, 2 weeks 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 2 months, 3 weeks ago
Option D didnt 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 5 months, 3 weeks ago
Agreed :)
upvoted 1 times

musumusu Most Recent 1 week, 5 days 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 3 weeks, 1 day ago
Selected Answer: C
I am going for C?
upvoted 1 times

slade_wilson 2 months, 1 week 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

zellck 2 months, 3 weeks 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 2 months, 3 weeks 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-serve

upvoted 2 times

sfsdeniso 3 months ago

Answer is C

upvoted 1 times

dish11dish 3 months, 1 week 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 2 times

piotrpiskorski 3 months, 1 week ago

Selected Answer: C

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."

It is C.

upvoted 2 times

gudiking 3 months, 1 week 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-serve

upvoted 1 times

cloudmon 3 months, 3 weeks 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 4 months ago

Selected Answer: B

It is B . Add one cluster means you have multi clusters so you use Multi-cluster routing

upvoted 2 times

MisuLava 4 months ago

Selected Answer: B

I meant to vote B not C

upvoted 3 times

MisuLava 4 months ago

Selected Answer: C

how can you use multi-cluster routing when you want to go to old instance with the regular load and to the new one with the analytical node ? MULTI will go to the nearest one...

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.

upvoted 4 times

somnathmaddi 4 months, 2 weeks ago

Selected Answer: C

Answer is C

upvoted 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: C

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.

upvoted 2 times

Question #94 *Topic 1*

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 (100%)

rickywck Highly Voted 2 years, 11 months ago
Why not C? Without BigQueryIO how can data be written back to BigQuery?
upvoted 29 times

xq 2 years, 11 months ago
C should be right
upvoted 7 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Sideinput for Bigquery data
upvoted 11 times

Siadd Most Recent 1 month, 3 weeks ago
A is the Answer.
A. Batch job, PubSubIO, side-inputs
upvoted 1 times

zellck 2 months, 3 weeks 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 you 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 3 times

sedado77 5 months, 1 week ago
Selected Answer: C
I got this question on sept 2022. Answer is C
upvoted 2 times

chrismayola 4 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 ago
Selected Answer: C
Answer: C
upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 2 months 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 1 year, 3 months ago

Ans: C

upvoted 1 times

pals_muthu 1 year, 6 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 1 year, 6 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 y BigQueryIO 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 1 year, 11 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 ago

Correct A. batch is cost-effective and no need to go for streaming

upvoted 1 times

funtoosh 2 years ago

How you are going to write back to BQ?

upvoted 1 times

someshsehgal 2 years ago

Correct A:
There are two points to defend it.
a. Its batch hence free of cost
b. static reference data hence no need to opt for streaming

upvoted 2 times

arghya13 2 years, 4 months ago

After so much confusion I think correct answer is B..reason
1.PubSub is always used for streaming input..no reason to mention
2.JdbcIO can be used by apache beam to connect bigquery to use a lookup
3.Sideoutput as a result to put in bigquery

upvoted 2 times

haroldbenites 2 years, 6 months ago

C is correct

upvoted 2 times

AJKumar 2 years, 7 months ago

Questions says - with static reference data---so there is no need for streaming, Choice A seems correct.

upvoted 2 times

zonc 2 years, 6 months ago

But A answer doesn't have BigQueryIO.

upvoted 3 times

AJKumar 2 years, 6 months ago

"The reference data is small enough to fit in memory on a single worker" does not require BQIO. This is a trick question, A is the obvious answer.

upvoted 2 times

zonc 2 years, 6 months ago

I don't think so. There are an input for streaming data from BigQuery and a side-input for BigQuery.

upvoted 1 times

dragon123 2 years, 6 months ago

Without BQIO how could you send data back to BQ?

upvoted 2 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

jvg637 Highly Voted 2 years, 11 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/performance
D –> since Google recommends adding nodes when storage utilization is > 70% https://cloud.google.com/bigtable/docs/modifying-instance#nodes

upvoted 49 times

sergio6 1 year, 4 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 2 years, 4 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 fail until you add nodes to each cluster that is over the limit. Also, if you try to remove nodes from a cluster, and the change would cause the cluster to exceed 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 2 years, 4 months ago

https://cloud.google.com/bigtable/docs/monitoring-instance

upvoted 3 times

Barniyah Highly Voted 2 years, 10 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 8 times

ch3n6 2 years, 8 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 12 times

musumusu Most Recent 1 week, 5 days ago

why not B ?

upvoted 1 times

musumusu 3 days, 17 hours 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 1 times

RoshanAshraf 1 month, 1 week ago

Selected Answer: CD
Key visualizer is Metrics for Performance issues. Ruled out
Storage and Write Operations ; C and D
upvoted 1 times

zellck 2 months, 3 weeks 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 2 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: CD
Well-designed row key : A B are not nessary
Write : CD both are involved in the question the most.
upvoted 1 times

Fezo 7 months, 4 weeks ago
Answer: CD
https://cloud.google.com/bigtable/docs/scaling
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: CD
as explained by MaxNRG
upvoted 2 times

MaxNRG 1 year, 2 months 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 1 year, 2 months ago
Selected Answer: CD
CD.

I agree with jvg637
upvoted 1 times

StefanoG 1 year, 3 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 1 year, 3 months ago
I am Voting for CD
upvoted 2 times

u_t_s 1 year, 4 months ago
Answer should be D & E
upvoted 1 times

tavva_prudhvi 10 months, 3 weeks ago
Why are you monitoring read operations, when youre supposed to write? why E?
upvoted 1 times

sergio6 1 year, 4 months ago
D--> 70% is the the recommended percentage of the cluster's storage capacity that is being used, If you are using more than 70% storage limit, 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 1 times

hauhau 1 year, 6 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 1 year, 8 months ago

B, C , D - all three looks okay to me

upvoted 2 times

sumanshu 1 year, 8 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 1 year, 4 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 1 year, 11 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 2 years, 11 months ago
Answer: C
Description: Social media posts can images/videos which cannot be stored in bigquery
upvoted 41 times

Shawvin 1 year, 4 months ago
Yes, the raw data needs to be archived too
upvoted 1 times

Devx198912233 2 years, 7 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 1 year, 6 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 2 years, 10 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 posts to the API.
upvoted 16 times

zellck Most Recent 2 months, 3 weeks ago
Selected Answer: C
C is the answer.
upvoted 1 times

sedado77 5 months, 1 week ago
Selected Answer: C
I got this question on sept 2022. Answer is C
upvoted 5 times

Erso 5 months, 3 weeks ago
Selected Answer: C
C is the correct one
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: C
Only C make sense.
upvoted 2 times

MaxNRG 1 year, 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 1 year, 2 months ago

SAY MY NAME!

upvoted 4 times

StefanoG 1 year, 2 months ago

Selected Answer: C

Analysis BQ
Storage GCS

upvoted 2 times

fire558787 1 year, 6 months ago

I believe the API accesses data only from GCS Buckets not BigQuery (but I'm not entirely sure)

upvoted 1 times

sumanshu 1 year, 8 months ago

Vote for C

upvoted 2 times

DPonly 2 years, 1 month ago

Answer should be C because we need to consider storage archival

upvoted 2 times

arghya13 2 years, 3 months ago

I'll go with option C

upvoted 2 times

Alasmindas 2 years, 3 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 fulfils 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 2 years, 5 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 2 years, 5 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

Ravivarma4786 2 years, 6 months ago

Bigquery is suitable for social media post ans should be 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 (93%) 7%

cleroy Highly Voted 2 years, 11 months ago
Use Dataprep ! It's THE tool for this
upvoted 52 times

rickywck Highly Voted 2 years, 11 months ago
Yes B.

Honest speaking, sometime I thought the answers being posted here were intentionally to mislead people whose do not have proper knowledge on the subject, but just memorizing answers to pass the exam.
upvoted 50 times

[Removed] 2 years, 1 month ago
True.. might be legal issue?!
upvoted 6 times

Besss Most Recent 1 month, 2 weeks ago
Selected Answer: B
It'B. DataPrep it's the right tool.
https://cloud.google.com/dataprep
upvoted 1 times

zellck 2 months, 3 weeks 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 5 months, 1 week ago
I got this question on sept 2022.
upvoted 3 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: B
B. Actually there are two tools to fix this problem.
Dataprep rely on dataflw
Datafusion rely on dataproc
upvoted 2 times

diagniste 9 months, 3 weeks ago
Selected Answer: A
A is the best answer !
upvoted 1 times

desertlotus1211 1 month, 1 week ago
dataflow IS Apache Beam...
upvoted 1 times

Venkat007 1 year ago

Selected Answer: B
B Dataprep
upvoted 2 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
https://cloud.google.com/dataprep/
upvoted 2 times

MaxNRG 1 year, 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 7 months, 1 week ago
max you rock man!
upvoted 1 times

GirijaSrinivasan 1 year, 4 months ago
Answer is B. Data prep. The keyword here is no programming skills required.
upvoted 3 times

nguyenmoon 1 year, 5 months ago
B- Dataprep
upvoted 2 times

pass_gcp 1 year, 6 months ago
Use Dataprep ....is the answer
upvoted 2 times

sumanshu 1 year, 8 months ago
Vote for B
upvoted 2 times

sumanshu 1 year, 8 months ago
Cloud Dataprep - almost fully automated
upvoted 1 times

AnilKr 2 years, 1 month ago
Most of the answers are wrong, this is simple even.
upvoted 1 times

mbiy 2 years, 2 months ago
Answer is B (Dataprep), for remaining option programming and/or SQL is involved.
upvoted 2 times

arghya13 2 years, 3 months ago
Dataprep-B
upvoted 2 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 2 years, 11 months ago
Should be A
upvoted 22 times

itche_scratche Highly Voted 2 years, 10 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 12 times

Besss Most Recent 1 month, 2 weeks ago
Selected Answer: A
A is correct
upvoted 1 times

zellck 2 months, 3 weeks 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
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 3 times

somnathmaddi 4 months, 2 weeks ago
Selected Answer: A
Should be A
upvoted 2 times

medeis_jar 1 year, 1 month 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 1 year, 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 sais 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 not detailed in the answer.
https://cloud.google.com/storage/docs/gsutil/commands/rsync
upvoted 6 times

am2005 1 year, 2 months ago
I am confused . which one is correct A or B ???
upvoted 1 times

hendrixlives 1 year, 2 months ago
Selected Answer: A
A is correct.

upvoted 1 times

Chelseajcole 1 year, 4 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
upvoted 2 times

manocha_01887 1 year, 6 months ago
How rsynch will handle private network?
"..The security rules don't allow access from external IPs to their on-premises resources.."
upvoted 3 times

sumanshu 1 year, 8 months ago
Vote for A
upvoted 2 times

daghayeghi 1 year, 11 months ago
A:
https://cloud.google.com/solutions/migration-to-google-cloud-transferring-your-large-datasets#options_available_from_google
upvoted 4 times

maurodipa 1 year, 2 months 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 1 year, 2 months ago
Yes. There is no access to on-prem from external IPs, but on prem can talk to external
upvoted 2 times

Ravivarma4786 2 years, 6 months ago
gsutil rsync will be used to transfer the files ANS A
upvoted 4 times

haroldbenites 2 years, 6 months ago
A is correct
upvoted 3 times

VishalB 2 years, 7 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 2 years, 7 months ago
option A
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 Highly Voted 2 years, 11 months ago
should be C:

https://cloud.google.com/bigquery/docs/best-practices-costs
upvoted 40 times

[Removed] Highly Voted 2 years, 11 months ago
Correct - C
upvoted 17 times

zellck Most Recent 2 months, 3 weeks 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
lustered 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 2 times

Fezo 7 months, 4 weeks ago
Selected Answer: C
C is the answer
https://cloud.google.com/bigquery/docs/best-practices-costs
upvoted 3 times

medeis_jar 1 year, 1 month ago
Selected Answer: C
C only make sense
upvoted 2 times

MaxNRG 1 year, 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, and the query counts against your free tier quota.
A and D doesnt 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 4 times

Anilgcp980 1 year, 2 months ago
this is a trap to make people fail by giving wrong answer as B.

upvoted 3 times

snadaf 1 year, 2 months ago
It's D, here is the link
https://cloud.google.com/bigquery/docs/best-practices-costs
upvoted 1 times

maurodipa 1 year, 2 months ago
Well, you mean C, isn't it?
upvoted 1 times

Crudgey 1 year, 3 months ago
Are they having a laugh at us by giving so many bad answers?
upvoted 5 times

tsoetan001 1 year, 4 months ago
Answer: B
Note: minimal change to sql
upvoted 1 times

szefco 1 year, 3 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 ans clustering amount of bytes scanned will decrease
upvoted 3 times

Ysance_AGS 1 year, 5 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 1 year, 3 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 1 year, 4 months ago
D would just block your query. The answer is C.
upvoted 1 times

nguyenmoon 1 year, 5 months ago
C - create partition table
upvoted 2 times

sumanshu 1 year, 8 months ago
Vote for C
upvoted 4 times

felixwtf 2 years, 2 months 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 2 years, 3 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 1 year, 7 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 2 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 1 year, 4 months ago
Recreating table is recommended by Google.
upvoted 1 times

arghya13 2 years, 3 months ago
should be C:
upvoted 2 times

gyclop 2 years, 5 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

Question #100

Topic 1

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 2 years, 11 months ago
I think we need a pipeline, so it's B to me.
upvoted 26 times

[Removed] Highly Voted 2 years, 11 months ago
Correct - B
upvoted 15 times

zellck Most Recent 2 months, 3 weeks ago
Selected Answer: B
B is the answer.
upvoted 1 times

Prasha123 3 months, 2 weeks ago
Selected Answer: B
Need pipeline so its B
upvoted 1 times

sedado77 5 months, 1 week ago
Selected Answer: B
I got this question on sept 2022. Answer is B
upvoted 5 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
omg. B only
upvoted 2 times

MaxNRG 1 year, 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 6 times

hendrixlives 1 year, 2 months ago
Selected Answer: B
B, streaming with dataflow
upvoted 3 times

JG123 1 year, 3 months ago
Wrong answer shown again by examtopics.com
Ans: B
upvoted 1 times

Ysance_AGS 1 year, 5 months ago
B => with dataflow you can parallelize data ingestion
upvoted 2 times

szefco 1 year, 3 months ago
And make it streaming
upvoted 1 times

sandipk91 1 year, 6 months ago
B is the right answer
upvoted 2 times

sumanshu 1 year, 8 months ago
Vote for B
upvoted 4 times

felixwtf 2 years, 2 months ago
You need a pipeline because this type of operation can be easily parallelized, as the ingestion can be divided between into chunks (PCollections) and handled by many workers.
upvoted 6 times

felixwtf 2 years, 2 months ago
so, B is the right answer.
upvoted 3 times

Tanmoyk 2 years, 5 months ago
You need a pipeline to feed streaming data , Answer : B
upvoted 3 times

haroldbenites 2 years, 6 months ago
B is correct
upvoted 2 times

dambilwa 2 years, 8 months ago
Option [B] is correct. 'C' is incorrect because there's a different method -'tabledata.insertAll' for streaming inserts
upvoted 2 times

Callumr 2 years, 8 months ago
The following limits apply when you load data into BigQuery.
Load jobs per table per day — 1,500 (including failures)
Load jobs per project per day — 100,000 (including failures)

so batch loading will not be able handle the expected capacity - so answer mut be B
upvoted 3 times

krizt9 2 years, 6 months ago
One day has 60 * 24 = 1,440 minutes = 1,440 batch jobs. This is still not exceeding the limit of 1,500. However, I still think it is 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

Ganshank Highly Voted 2 years, 10 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 47 times

AzureDP900 1 month, 4 weeks ago
B is right answer
upvoted 1 times

tprashanth 2 years, 7 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

BigQuery 1 year, 2 months ago
SAY MY NAME!
You need to Transfer Sensitive Patient information, over public ISP you shouldn't do that.
upvoted 1 times

TNT87 2 years, 5 months ago
Answer is B,gsutil has a limit of 1TBaccording to Google documentation,if data is morethan 1TBthen we have to use Transfer Appliance.
upvoted 16 times

Yiouk 1 year, 6 months ago
The answer is clearly seen here: https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options
upvoted 5 times

SSV Highly Voted 2 years, 7 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 7 times

VASI 2 years, 1 month ago
5Tb "for individual objects". Create smaller AVRO files.
upvoted 1 times

VASI 2 years, 1 month ago
AVRO compression can reduce file size to a tenth
upvoted 2 times

Acheto Most Recent 1 month ago
Selected Answer: B

gsutil is recommended for data less than 1 TB. In this question, the data is 10 TB, so A is ruled out as well.
In my opinion, B is the correct answer.

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

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

upvoted 1 times

cajica 3 months, 1 week ago

Selected Answer: A

For me it's A, at the beginning I thought it was B but I don't really think transfer appliance is time-efficient, as I don't think I classify gsutil as an insecure transfer method. It obviously depends on the bandwidth but it would be very weird to me consider a really low bandwidth as the standard.

upvoted 1 times

wan2three 1 month, 3 weeks ago

“ Export the records from the database as an Avro file”, here it mentioned 'an', unless Avro can compress 10TB to less than 1 then we cant use GSutil given the limit is 1TB

upvoted 1 times

Atnafu 3 months, 2 weeks ago

According to this it's B
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#options_available_from_google

upvoted 1 times

clouditis 5 months, 1 week ago

B, as gsutil is only for 1 Tb or less data

upvoted 1 times

John_Pongthorn 5 months, 1 week ago

Selected Answer: B

In sense of speek(time-effective)
https://cloud.google.com/transfer-appliance/docs/4.0/overview#transfer-speeds
ransfer Appliance, you can receive the appliance and capture 300 terabytes of data in under 25 days. Your data can be accessed in Cloud Storage within another 25 days, all without consuming any outbound network bandwidth.

upvoted 1 times

MounicaN 5 months, 2 weeks ago

Selected Answer: B

less than 1TB . gsutil
more than 1TB, transfer service
when the internet speeds aren't good, transfer appliance

upvoted 2 times

YorelNation 5 months, 3 weeks ago

"time efficient solution" 7 days transfer appliance is pretty long if you ask me.

upvoted 1 times

ducc 6 months ago

Selected Answer: B

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

upvoted 1 times

ducc 6 months ago

Selected Answer: B

B is most secured way
Because it is patient information, it is needed to be very secured

upvoted 1 times

t11 6 months, 1 week ago

A is better than B in speed. It takes two plus weeks for B. Both methods encrypted data during transfer.

upvoted 1 times

rytizzle 7 months ago

"Acquiring a Transfer Appliance is straightforward. In the 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 is 20 days."

sounds like a long time to me. I'm going with A.

upvoted 2 times

BJPJowee 7 months, 3 weeks ago

A is the only valid answer! The key word is "time-efficient". Although gsutil has a limit of 1TB per file smaller files can be created. So splitting it into multiple files will be an idea. B is not time-efficient, you will have to ship to google cloud. What if your data center is remote with no fast delivery service?

upvoted 1 times

IJFIJD 8 months ago

Selected Answer: B

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
its Transfer appliace : B

upvoted 2 times

mahx 8 months, 1 week ago

Answer is B , gsutil can support only small transfer sizes (up to 1 TB)

upvoted 1 times

Question #102 *Topic 1*

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

A (57%) C (43%)

MaxNRG Highly Voted 1 year, 1 month ago

Selected Answer: A

A - New correct answer
C - Old correct answer (for 2019)

upvoted 25 times

haroldbenites Highly Voted 2 years, 6 months ago

C is correct.
It says “update Every hour”
And need “ accuracy”

upvoted 22 times

musumusu Most Recent 3 days, 16 hours ago

Answer : A
Don't be confused anymore.
Current case: Data updates in EVERY HOUR, according to that we need to update the table immediately.
A. Update statement/ query as soon as new data landed, current case is good. If data is frequently updating every second then Option C was the right answer.
B. Partition table is good to increase query performance but this is not the demand.
Option D, this is also not batch load condition, and unnecssory work

upvoted 1 times

techtitan 1 week, 6 days ago

Selected Answer: C

New inventory table is the realtime inventory updates table. Data has to be streamed in this realtime. The historical inventory balance table can be updated nightly.

upvoted 1 times

desertlotus1211 1 month, 1 week ago

Answer is A: https://cloud.google.com/blog/topics/developers-practitioners/inventory-management-bigquery-and-cloud-run/

leverage UPDATE statement .... there is another part to use Cloud Run. The update statement is what's important here. Inventory changes every hour. Nightly inventory updates doesn't help.

upvoted 4 times

desertlotus1211 1 month, 1 week ago

It also says ' near real-time inventory dashboard'... and to have accurate data. That eliminates answer C.

upvoted 1 times

jkhong 2 months, 1 week ago

Selected Answer: C

What this question is about is actually Change Data Capture. I'm surprised nobody has mentioned this as of yet..
(https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#before_you_begin)

3 methods of CDC is:
i. Immediate consistency approach -> view table which joins main table and delta table (expensive if queries are made frequently as join operations need to be executed)
ii. Cost-optimized approach -> merge delta and main periodically, insert / delete from main and delete delta entries in delta table (cheaper queries, but not immediate)
iii. Hybrid approach

upvoted 1 times

jkhong 2 months, 1 week ago

Hybrid approach would be a combination of (i) and (ii).

What C and D is recommending is essentially the Hybrid approach. A staging table will be maintained (i.e. the data streamed for the daily inventory movement). The immediate consistency approach representing the view and cost-optimized approach representing nightly inventory balance update. I'd pick C because streaming is still important to keep data updated. For anyone who has had experience in processing transactions, A would be a nightmare for managing data warehouses... BigQuery's suggested approach for updates is always append, and is not built for high load UPDATEs

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: C

C is the answer.

https://cloud.google.com/bigquery/docs/best-practices-performance-patterns#dml_statements_that_update_or_insert_single_rows

Best practice: Avoid point-specific DML statements (updating or inserting 1 row at a time). Batch your updates and inserts.

For example, you could load your set of replacement records into another table, then write the DML statement to update all values in the original table if the non-updated columns match.

upvoted 1 times

hauhau 2 months, 4 weeks ago

Selected Answer: C

C streaming is correct. updata has limit

upvoted 1 times

maksi 3 months, 2 weeks ago

C is Correct

According to Documentation, you can perform 5 UPDATE operations per 10 seconds, so as a result during an hour we will be able to perform only 1800 UPDATE operations. I'm not sure, in the Description it's mentioned that we need to perform SEVERAL thousands operations, not 1.8.
https://cloud.google.com/bigquery/quotas#standard_tables

upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: C

C it is

upvoted 2 times

NicolasN 3 months, 3 weeks ago

Selected Answer: C

I'm afraid that even the DML statements per day are unlimited nowadays (2022), the answer [C] remains the best option for the near real-time constraint, since updating row-by-row in BigQuery is considered by Google an anti-pattern.
Quoting from https://cloud.google.com/bigquery/docs/best-practices-performance-patterns#dml_statements_that_update_or_insert_single_rows
:
Best practice: Avoid point-specific DML statements (updating or inserting 1 row at a time). Batch your updates and inserts.
... BigQuery DML statements are intended for bulk updates. UPDATE and DELETE DML statements in BigQuery are oriented towards periodic rewrites of your data, not single row mutations.
... If your use case involves frequent single row inserts, consider streaming your data instead.

Through this prism: [C] (streaming) > [D] (batch) > A [row-by-row]

upvoted 5 times

John_Pongthorn 5 months ago

Selected Answer: C

C
This link give reason to support the answer
https://cloud.google.com/bigquery/docs/write-api#overview

upvoted 2 times

John_Pongthorn 5 months ago

Another link
Streaming inserts and timestamp-aware queries
https://cloud.google.com/blog/products/bigquery/performing-large-scale-mutations-in-bigquery#Streaming%20inserts%20and%20timestamp-aware%20queries

upvoted 1 times

clouditis 5 months, 1 week ago

C it is

upvoted 1 times

alecuba16 6 months, 1 week ago

Selected Answer: C
2022
Cannot be A there are limits:

Table operations per day 1500 operations
Your project can make up to 1,500 table operations per table per day, whether the operation appends data to the table or truncates the table. This limit includes the combined total of all load jobs, copy jobs, and query jobs that append to or overwrite a destination table or that use a DML DELETE, INSERT, MERGE, TRUNCATE TABLE, or UPDATE statement to write data to a table.

DML statements count toward partition limits, but aren't limited by them. For more information about DML limits, see Data manipulation language statements.

1.000 updates per hour * 24 = 24000 updates > 1500 daily limit.

https://cloud.google.com/bigquery/quotas
upvoted 5 times

MisuLava 4 months ago
February 11, 2019
The following DML quotas have changed:

The maximum number of combined UPDATE, DELETE, and MERGE statements per day per table has increased from 200 to 1,000.
The limit of 10,000 combined UPDATE, DELETE, and MERGE statements per day per project has been removed. There is no longer a project-level limit on DML statements.
upvoted 1 times

boulathone 6 months ago
I think that you are confusing jobs/operations and DML statement.
Google Cloud do have a quota for the jobs, but not for the DML anymore, so a single operation/job can do more than 1500 updates in 24h, but you can't do more than 1500 jobs/operations in 24h.
Since there is no DML quotas anymore, the correct answer today is A.
upvoted 2 times

DataEngineer_WideOps 7 months, 1 week ago
"ensure that the data is accurate"
Means leverage update statement to update as they are changing.

Answer A.
upvoted 1 times

alecuba16 7 months, 4 weeks ago
Selected Answer: C
Cannot be A:

Rows that were written to a table recently by using streaming (the tabledata.insertall method or the Storage Write API) cannot be modified with UPDATE, DELETE, or MERGE statements. The recent writes are those that occur within the last 30 minutes. All other rows in the table remain modifiable by using UPDATE, DELETE, or MERGE statements. The streamed data can take up to 90 minutes to become available for copy operations.

https://cloud.google.com/bigquery/docs/reference/standard-sql/data-manipulation-language
upvoted 4 times

rytizzle 7 months ago
it can be A. the data doesn't have rows that are streamed in using tabledata.insertall

it's historical data that sits in BQ and is updated, not streamed.
upvoted 1 times

alecuba16 6 months, 1 week ago
You cannot do A because it is not a recommended practise:

https://cloud.google.com/blog/products/bigquery/performing-large-scale-mutations-in-bigquery
upvoted 4 times

alecuba16 6 months, 1 week ago
Provide a link that justifies your answer, bigquery states to avoid frequent updates directly on BQ, for continuous insertion states the usage of steaming insert and insertall methods.
upvoted 2 times

cualquiernick 8 months, 1 week ago
Ans is A, as Bq quota for updates in BQ does not longer have quota limits
upvoted 2 times

alecuba16 6 months, 1 week ago
Table operations per day 1500 operations
Your project can make up to 1,500 table operations per table per day, whether the operation appends data to the table or truncates the table. This limit includes the combined total of all load jobs, copy jobs, and query jobs that append to or overwrite a destination table or that use a DML DELETE, INSERT, MERGE, TRUNCATE TABLE, or UPDATE statement to write data to a table.

DML statements count toward partition limits, but aren't limited by them. For more information about DML limits, see Data manipulation language statements.

upvoted 1 times

Question #103 *Topic 1*

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?

A. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

B. 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.

C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

D. 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

kostol 3 days, 5 hours ago

Selected Answer: D

https://cloud.google.com/bigquery/docs/table-snapshots-scheduled

upvoted 1 times

desertlotus1211 1 month, 1 week 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 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."
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 1 times

YatMoh 2 months, 1 week 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 1 times

jkhong 2 months, 1 week 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)

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 2 months, 2 weeks ago

Even when the dataset/table is regional, google provides high availability with data available in 2 zones. Isn't A correct answer?

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: C

C is the answer.

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 1 times

hauhau 2 months, 4 weeks ago

Selected Answer: D

point in time snapshots only have data from past 7 days
upvoted 1 times

desertlotus1211 1 month, 1 week ago
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 2 months, 1 week 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
upvoted 1 times

Atnafu 3 months, 2 weeks 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 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.
upvoted 1 times

cloudmon 3 months, 3 weeks ago
Selected Answer: C
C
Multi-region for HA
Snapshot for recovery
upvoted 1 times

max_c 5 months, 1 week 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
upvoted 3 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: C
C and D are high avaiability
But C use buid-in feature in BQ
https://cloud.google.com/bigquery/docs/table-snapshots-intro#table_snapshots

While D it is quite put too much effeot and not minimize cost
upvoted 1 times

MounicaN 5 months, 2 weeks ago
Selected Answer: D
point in time snapshots only have data from past 7 days
upvoted 2 times

MisuLava 3 months, 4 weeks 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

devaid 4 months, 3 weeks 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
upvoted 2 times

bigquery1102 5 months, 3 weeks ago
Should be D while point-in-time query only access data fron recent 7 days
upvoted 3 times

jkhong 2 months, 1 week 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

upvoted 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: C

C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

upvoted 1 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

musumusu 3 days, 16 hours ago
Answer C: Use Recipe Template feature of dataprep. Don't need to change the service.
upvoted 1 times

jroig_ 1 month, 2 weeks ago
Selected Answer: C
Why not C?
upvoted 1 times

jkhong 2 months, 1 week 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 3 times

zellck 2 months, 3 weeks ago
Selected Answer: D
D is the answer.
upvoted 1 times

anicloudgirl 2 months, 3 weeks 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 3 times

anicloudgirl 2 months, 3 weeks 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 2 months, 1 week 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 times
upvoted 1 times

cloudmon 3 months, 3 weeks ago
Selected Answer: D
It's D
upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: D
Dataprep and Dataflow are same famitly
upvoted 2 times

AWSandeep 5 months, 3 weeks ago
Selected Answer: D
D. Export the Dataprep job as a Dataflow template, and incorporate it into a Composer job.
Reveal Solution

upvoted 3 times

damaldon 5 months, 4 weeks ago

It's D, use composer to schedule tasks

upvoted 2 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%)

niketd 5 days, 22 hours ago

Selected Answer: B

Multi-step sequential pipelines -> Cloud Composer

upvoted 1 times

AzureDP900 1 month, 4 weeks ago

Cloud composer B is right

upvoted 2 times

odacir 2 months, 3 weeks ago

Selected Answer: B

Cloud Composer (Airflow) is the answer to chain different steps from different apps...

upvoted 1 times

MisuLava 4 months ago

Selected Answer: B

" multiple dependencies on each other. You want to use managed service"
= Cloud Composer

upvoted 1 times

John_Pongthorn 5 months, 2 weeks 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 5 months, 3 weeks ago

Selected Answer: B

composer :)

upvoted 1 times

YorelNation 5 months, 3 weeks ago

Selected Answer: B

Composer

upvoted 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: B

B. Cloud Composer

upvoted 1 times

damaldon 5 months, 4 weeks 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

AzureDP900 1 month, 4 weeks ago
D is right
https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scaling-clusters#using_graceful_decommissioning
upvoted 3 times

Prakzz 2 months, 1 week 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 month, 3 weeks ago
nope, they are not only for non-preeemtible workers
upvoted 1 times

yafsong 2 months, 1 week 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 1 times

odacir 2 months, 3 weeks 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 3 months, 2 weeks 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 5 months, 2 weeks 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 4 months ago

This weird.
The question mentions that increase cluster, but Graceful Decommissioning use in downscale the cluster

upvoted 1 times

odacir 2 months, 3 weeks 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 5 months, 3 weeks ago

Selected Answer: D

D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

upvoted 1 times

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 2 months ago

Selected Answer: D

The cloud function with DLP seems the best option

upvoted 1 times

HarshKothari21 5 months, 3 weeks ago

Selected Answer: D

D option

upvoted 1 times

AWSandeep 5 months, 3 weeks 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 month, 4 weeks 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 2 years, 11 months ago
Should be A
upvoted 33 times

Rajokkiyam Highly Voted 2 years, 11 months ago
Create dependency in Cloud Composer and schedule it.
upvoted 20 times

MisuLava 3 months, 4 weeks ago
the jobs are not interdependent. just 3 individual jobs
upvoted 1 times

AzureDP900 Most Recent 1 month, 4 weeks ago
A is correct
upvoted 1 times

dconesoko 2 months ago
Selected Answer: A
Cloud composer's DAG would manage the dependencies
upvoted 1 times

zellck 2 months, 3 weeks 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 7 months, 3 weeks ago
This should be A
upvoted 2 times

medeis_jar 1 year, 1 month ago
Selected Answer: A
https://cloud.google.com/composer/docs/how-to/using/writing-dags
upvoted 3 times

MaxNRG 1 year, 1 month 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 1 year, 2 months 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 1 year, 1 month ago
nice article thanks!
upvoted 1 times

JG123 1 year, 3 months ago
Selected Answer: A
Cloud Composer
upvoted 3 times

JG123 1 year, 3 months ago
Correct: A
upvoted 3 times

sandipk91 1 year, 6 months ago
should be option A
upvoted 5 times

sumanshu 1 year, 8 months ago
Vote for A
upvoted 4 times

someshsehgal 2 years ago
COrrect A: Couldn't understand why a option with no connection with actual problem has been given as correct option (D)
upvoted 3 times

arghya13 2 years, 3 months ago
I'll go for A
upvoted 2 times

Tanmoyk 2 years, 5 months ago
Should be A
upvoted 2 times

atnafu2020 2 years, 6 months ago
A
is correct Answer
upvoted 4 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 (100%)

TNT87 Highly Voted 5 months, 2 weeks ago

Answer C E
By not acknowleding 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 ig 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 6 times

musumusu Most Recent 1 week, 4 days ago

Answer D&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 1 times

desertlotus1211 1 month, 1 week 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 1 times

AzureDP900 1 month, 4 weeks ago

C, E seems correct

upvoted 1 times

MounicaN 5 months, 2 weeks ago

D might also be right?
Subscriber might not be provisioned enough

upvoted 3 times

AWSandeep 5 months, 3 weeks 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] Highly Voted 2 years, 11 months ago
Correct - B
upvoted 16 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: ParDo is used to do transformation and create side output
upvoted 10 times

AzureDP900 Most Recent 1 month, 4 weeks ago
B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
upvoted 1 times

zellck 2 months, 3 weeks 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 1 times

Pime13 7 months, 3 weeks 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 2 times

medeis_jar 1 year, 1 month 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 month, 4 weeks ago
I agree with B
upvoted 1 times

MaxNRG 1 year, 1 month 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 5 times

sumanshu 1 year, 8 months ago
vote for 'B', ParDo can discard the elements.

https://beam.apache.org/documentation/programming-guide/

upvoted 4 times

DeepakKhattar 2 years, 1 month 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 2 years, 3 months ago
B is correct

upvoted 3 times

SteelWarrior 2 years, 5 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 which can both identify valid/invalid records and also generate two PCollections one with valid records and other with invalid records.

upvoted 7 times

haroldbenites 2 years, 6 months ago
B is correct

upvoted 3 times

Archy 2 years, 7 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 2 years, 7 months ago
Looks like C it is
https://beam.apache.org/documentation/programming-guide/

upvoted 2 times

atnafu2020 2 years, 6 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 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 2 years, 2 months 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 2 years, 7 months ago
Correct answer should be "C". A Pardo transform will allow the processing to happening in parallel using multiple workers. Partition transform will allow data to be partitions in two different Pcollections according to some logic. Using partition transform once can split the corrupted data and finally discard it.

upvoted 5 times

Rajuuu 2 years, 7 months ago
Correct B.

upvoted 4 times

norwayping 2 years, 8 months ago
Correct - B

upvoted 5 times

Question #111 *Topic 1*

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"€λ 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] Highly Voted 2 years, 11 months ago
should be A
upvoted 31 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: Partition is the solution for reducing cost and time
upvoted 18 times

willbot 2 years, 9 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 1 year, 4 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 Most Recent 2 months, 3 weeks 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 2 months, 3 weeks 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 2 times

AzureDP900 1 month, 4 weeks ago
A is right
upvoted 1 times

John_Pongthorn 5 months, 1 week 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 5 months, 1 week ago

Selected Answer: A

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 5 months, 2 weeks ago

Selected Answer: A

A AND D , they are the most likely choiced but the questionn 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 wildcarf

upvoted 1 times

Didine_22 10 months, 1 week ago

Selected Answer: A

answer A

upvoted 2 times

medeis_jar 1 year, 1 month ago

Selected Answer: A

https://cloud.google.com/bigquery/docs/partitioned-tables

upvoted 2 times

MaxNRG 1 year, 1 month ago

Selected Answer: A

A. Partiotioning
https://cloud.google.com/bigquery/docs/partitioned-tables

upvoted 1 times

Tomi1313 1 year, 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 5 months ago

Partitioning is recommended over table sharding, because partitioned tables perform better
This is a google recommendation nowaday.

upvoted 1 times

StefanoG 1 year, 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 wrost.
Is not mandatory make the queries while you make the modify so the right answer is A

upvoted 4 times

JayZeeLee 1 year, 3 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 1 year, 8 months ago

Vote for A

upvoted 5 times

arghya13 2 years, 3 months ago

I will go with Option A

upvoted 5 times

Alasmindas 2 years, 3 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 mot mentioned, we will land up having multiple - 30-90 days data in separate tables + the master table.

upvoted 3 times

karthik89 2 years ago

but how will you append the data that is older than 90days in to the master table?

upvoted 2 times

Cloud_Enthusiast 2 years, 3 months ago

Answer is A. Recreating the DDL with new parition 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%) | 14% | Other

[Removed] Highly Voted 2 years, 11 months ago

Should be B

upvoted 29 times

Ganshank Highly Voted 2 years, 10 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 a 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 20 times

serg3d 2 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 2 years, 3 months ago

the question also talks about a cost effective way...

upvoted 3 times

sh2020 2 years, 8 months ago

I agree, C is the only choice that addresses the problem. The problem is caused by leased line. How come pub/sub service can resolve it? Pub/sub will still use the leased line

upvoted 5 times

awssp12345 1 year, 7 months ago

DEFINITELY NOT COST EFFECT. C IS THE WORST CHOICE.

upvoted 6 times

musumusu Most Recent 1 week, 4 days 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 1 times

musumusu 3 days, 14 hours 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 deal with current situation. Or reprogramme IOT if they have connected with sensors.

upvoted 1 times

ayush_1995 1 month 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 3 times

desertlotus1211 1 month, 1 week ago

Are they talking about GCP in this question?
Where is the event processing infrastructure?

Answer A, might be correct!

upvoted 1 times

PrashantGupta1616 2 months 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 2 months, 2 weeks 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 1 times

desertlotus1211 1 month, 1 week 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 1 times

desertlotus1211 1 month, 1 week 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 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 2 times

AzureDP900 1 month, 4 weeks ago

yes it is B. Have the data acquisition devices publish data to Cloud Pub/Sub.

upvoted 1 times

piotrpiskorski 3 months, 1 week 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 like 3k$)..

It's C.

upvoted 1 times

jkhong 2 months, 3 weeks 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 a 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 2 months, 3 weeks 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 2 months, 3 weeks ago

ALSO, the problem it's no t 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 5 months, 2 weeks ago

Selected Answer: B

Cloud Pub/Sub, it supports batch & streaming , push and pull capabilities
Answer B

upvoted 1 times

t11 6 months, 1 week ago

It has to be B.

upvoted 1 times

rr4444 6 months, 2 weeks 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 7 months, 2 weeks 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 fully reliable Cloud network. That should solve the problem here.

upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: B

skip processing and sending through multiple endpoints and ingest raw messages from sensors in GCP

upvoted 1 times

MaxNRG 1 year, 1 month ago

Selected Answer: B

B: the most cost-effective way - Pub/Sub

upvoted 2 times

MaxNRG 1 year, 1 month ago

improve event delivery reliability

upvoted 1 times

sergio6 1 year, 3 months ago

Selected Answer: C

A. On edge solution, it doesn't resolve the network latency problem
B. like A
D: It's impossible define the minimum gap duration of the session window

upvoted 1 times

mjb65 1 year, 3 months ago

I've reas the text a fifth time now and all of your comments, folks... mmmh, they do not mention that the processing infrastructure is loacated at GCP, in this case A seems technically correct as buffering solution, otherwise I would go with B as modern actuators can use MQTT --> Cloud IoT Core --> PubSub (???)

upvoted 2 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 (74%) A (26%)

rickywck Highly Voted 2 years, 11 months ago
should be C, since we need to recognize both voice and intent
upvoted 23 times

AzureDP900 1 month, 4 weeks ago
C. Dialogflow Enterprise Editio
upvoted 1 times

Alasmindas Highly Voted 2 years, 3 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 17 times

HarshKothari21 5 months, 3 weeks 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 1 times

hdmi_switch 1 year, 7 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 6 times

techtitan Most Recent 1 week, 6 days ago
Selected Answer: C
https://cloud.google.com/dialogflow/es/docs/integrations/aog
upvoted 1 times

desertlotus1211 1 month, 1 week ago
I think the answer is A: Speech to Text.
You want to interpret what a user say... Diagflow is text to speech, not what the question asked for...

Thoughts?
upvoted 1 times

PrashantGupta1616 2 months 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 2 months, 3 weeks 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 2 months, 1 week 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 4 times

zellck 2 months, 3 weeks 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 1 times

TNT87 5 months, 2 weeks 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 1 times

Smaks 7 months, 1 week 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 10 months, 1 week ago

dialog

upvoted 1 times

Vip777 10 months, 1 week ago

speech

upvoted 1 times

PJG_worm 11 months, 1 week ago

It should be D. INTERPRET customer voice commands and issue an order to the backend systems. Option C is usually applied for conversation. But in this case, it is not a conversation.

upvoted 1 times

rytizzle 7 months ago

it can totally be a conversation. if the system is attempting to interpret voice commands, it may need clarifying questions, which would in turn be 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 1 year, 1 month 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 1 year, 1 month 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 interactions 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 send 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 1 year, 2 months ago

Selected Answer: A

Speech-to-text is mentioned in Googles 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

upvoted 3 times

sergio6 1 year, 3 months ago

Selected Answer: C

Dialogflow is able to understand the customer's voice commands and send an order to the backend systems

upvoted 2 times

manocha_01887 1 year, 6 months ago

D
A Dialogflow Agent is a virtual agent that participates in a conversation. Conversations can be audio or text based.

Sullivan, Dan. Official Google Cloud Certified Professional Data Engineer Study Guide (p. 274). Wiley. Kindle Edition.

upvoted 4 times

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 2 years, 11 months ago
Answer should be B
upvoted 27 times

[Removed] Highly Voted 2 years, 11 months ago
Answer - B
upvoted 11 times

PrashantGupta1616 Most Recent 2 months ago
Selected Answer: B
Cloud Composer is Airflow
upvoted 1 times

odacir 2 months, 3 weeks ago
Selected Answer: B
Cloud Composer is Airflow, It's made for this job.
upvoted 1 times

zellck 2 months, 3 weeks 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 2 times

AzureDP900 1 month, 4 weeks ago
Cloud composer is right
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
https://cloud.google.com/composer/
upvoted 1 times

MaxNRG 1 year, 1 month 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 Eves 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 1 year, 1 month 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 1 year, 3 months ago

Selected Answer: B

Answer: B

upvoted 3 times

sandipk91 1 year, 6 months ago

Cloud composer

upvoted 3 times

sumanshu 1 year, 8 months ago

Vote for B

upvoted 5 times

daghayeghi 1 year, 11 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 ago

Correct B: without any doubt.

upvoted 3 times

arghya13 2 years, 3 months ago

B-Cloud Composer works on a multicloud environment

upvoted 4 times

Alasmindas 2 years, 3 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 2 years, 3 months ago

Composer is the obvious answer. so B

upvoted 4 times

Darlee 2 years, 4 months ago

How come the `Correct Answer` so ridiculous WRONG?

upvoted 7 times

squishy_fishy 1 year, 4 months ago

Ha ha.. I know.

upvoted 1 times

haroldbenites 2 years, 6 months ago

B is correct obviously.

upvoted 3 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 Highly Voted 4 months, 4 weeks 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 with access to that view.

upvoted 10 times

musumusu Most Recent 1 week, 4 days 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 3 days, 14 hours 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) by sharing BQ dataset.

upvoted 1 times

lool 1 month 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 combine 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 1 times

AzureDP900 1 month, 4 weeks ago

A. Use Analytics Hub to control data access, and provide third party companies with access to the dataset.

upvoted 1 times

odacir 2 months, 3 weeks ago

Selected Answer: A

https://cloud.google.com/analytics-hub

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: A

A is the answer.

https://cloud.google.com/bigquery/docs/analytics-hub-introduction
nalytics 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. Listings let you share data without replicating the shared data. You can build a catalog of analytics-ready data sources with granular permissions that let you deliver data to the right audiences. You can also manage subscriptions to your listings.

upvoted 2 times

Atnafu 3 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 bucket for third-party companies to use.

upvoted 1 times

ajayrtk 4 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 access to that view.

upvoted 2 times

AWSandeep 5 months, 3 weeks 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 5 months, 4 weeks 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. Listings let you share data without replicating the shared data. You can build a catalog of analytics-ready data sources with granular permissions that let 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 the 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 (74%) 11% Other

YorelNation Highly Voted 5 months, 3 weeks 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. (all statements comes from the staging table)

upvoted 8 times

musumusu Most Recent 3 days, 14 hours ago

B&D
Tricks here: Always choose google recommended approach, Use data first in Staging table then merge with original tables.

upvoted 1 times

dconesoko 2 months 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 the 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 1 times

odacir 2 months, 3 weeks 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 1 times

zellck 2 months, 3 weeks 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
Delta tables contain all change events for a particular table since the initial load. Having all change events available can be valuable for identifying 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 1 times

NicolasN 3 months, 1 week ago

I really can't find a correct combination of answers. I'm between the following alternatives, but with no one fitting:
1️⃣ [B] and [D]: That's a proposed solution, but as a cost-optimized approach (along with an extra step to "Periodically DELETE outdated records 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.
2️⃣ [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.
3️⃣ [A] standalone: Provides immediate results but is far from compute-optimized.

upvoted 1 times

NicolasN 3 months, 1 week 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 1 times

NicolasN 3 months, 1 week ago

The previous guidelines were here:
🔗 https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#immediate_consistency_approach
There were two approaches:
1️⃣ Immediate consistency approach
2️⃣ Cost-optimized approach
For approach 1️⃣, 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' no answer that fits
For approach 2️⃣ 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 2 times

beanz00 3 months, 4 weeks ago

B and E. Typically in a Data Warehouse you don't delete date. 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 4 months, 2 weeks ago

https://www.striim.com/blog/oracle-to-google-bigquery/

upvoted 1 times

TNT87 5 months, 2 weeks ago

https://cloud.google.com/architecture/database-replication-to-bigquery-using-change-data-capture#data_latency

upvoted 1 times

TNT87 5 months, 2 weeks ago

https://docs.streamsets.com/platform-datacollector/latest/datacollector/UserGuide/Destinations/GBigQuery.html

upvoted 1 times

Wasss123 5 months, 2 weeks 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 5 months, 2 weeks 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 5 months, 3 weeks 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

changsu 5 months, 3 weeks ago

Selected Answer: CE

It costs more to update/delete.

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: BE

BE is correct
Big Query only need to capture change, no need DELETE, UPDATE

upvoted 1 times

Question #117 *Topic 1*

You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at 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 2 years, 11 months ago
Answer should be D
upvoted 25 times

[Removed] Highly Voted 2 years, 11 months ago
Answer - D
upvoted 13 times

dconesoko Most Recent 2 months ago
Selected Answer: D
google's preferred choice
upvoted 1 times

zellck 2 months, 3 weeks ago
D is the answer.
upvoted 1 times

Pime13 7 months, 3 weeks ago
Selected Answer: D
Answer should be D
upvoted 1 times

VictorBa 10 months, 1 week ago
Selected Answer: D
It cannot be C because Dataproc is more suitable for Hadoop jobs.
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: D
Pub/Sub + Dataflow
upvoted 1 times

MaxNRG 1 year, 1 month 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 3 times

hendrixlives 1 year, 2 months 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 1 year, 3 months ago
Correct: D
upvoted 3 times

Chelseajcole 1 year, 5 months ago
rule of thumb: If you see Kafka and Pub/Sub, always go with Pub/Sub in Google exam
upvoted 4 times

hendrixlives 1 year, 2 months 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 all messages without time limit. So no Pub/Sub there.
This being a Google cert does not mean that they always force Google solutions.

upvoted 3 times

sandipk91 1 year, 6 months ago

Answer is D

upvoted 2 times

awssp12345 1 year, 7 months ago

https://cloud.google.com/architecture/migrating-from-kafka-to-pubsub#comparing_features

upvoted 2 times

sumanshu 1 year, 8 months ago

Vote for D

Scaling - Dataflow.
Delivery of confimed atleast 1 message - Pub/Sub

upvoted 3 times

Sush12 2 years ago

Answer is D

upvoted 2 times

Alasmindas 2 years, 3 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 4 times

szefco 1 year, 3 months ago

To make us think of each question while studying, not just trying to memorize answers :) it looks that "correct" answers are chosen randomly :)

upvoted 1 times

Cloud_Enthusiast 2 years, 3 months ago

D for GCP native solution

upvoted 3 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 (69%) D (31%)

nwk Highly Voted 5 months, 3 weeks 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 5 months, 3 weeks 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 2 times

AWSandeep Highly Voted 5 months, 3 weeks 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 5 times

musumusu Most Recent 3 days, 14 hours ago
Answer B:
Why not D, mentioned in question: Data lead will create tables in dataset. Imagine, other department leads are creating unnecessory 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 seperate dataset and do whatever you want in that dataset.
upvoted 1 times

xj_kevin 2 weeks, 4 days ago
Vote B, both BD can fullfill 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 1 times

desertlotus1211 1 month, 1 week ago
Interestingly enough - I know believe the answer is A...
Deleting is not the same as modify...
upvoted 1 times

desertlotus1211 1 month, 1 week 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 month, 1 week ago
I apologize - I thought B said Owner...
This questions makes no sense now...
upvoted 1 times

odacir 2 months, 3 weeks 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 dataset.
upvoted 2 times

jkhong 2 months, 1 week 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 they 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 provide access on a table level...
upvoted 5 times

Wonka87 2 months ago
and to supplement why does it need viewer role on the project the table is in?
upvoted 1 times

Atnafu 3 months, 2 weeks ago
Wow B is an answer
https://cloud.google.com/bigquery/docs/access-control-basic-roles#dataset-basic-roles
upvoted 3 times

MisuLava 3 months, 4 weeks 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 month, 1 week ago
Ummm owner is a predefined role
https://cloud.google.com/bigquery/docs/access-control
BigQuery Data Owner
(roles/bigquery.dataOwner)
upvoted 1 times

josrojgra 4 months, 2 weeks 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 bigquery.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 4 months, 2 weeks 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

ducc 5 months, 4 weeks ago
Selected Answer: B
B is correct
upvoted 3 times

damaldon 5 months, 4 weeks ago

Ans.B
Leads needs to create tables, meaning they need a dataset and analysts only need read permissions

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%)

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: Timestamp at starting of rowkey causes bottleneck issues
upvoted 38 times

kichukonr Highly Voted 2 years, 10 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 12 times

karthik89 2 years ago
it can start with stock symbol concated with timestamp can be a good row key design
upvoted 6 times

Yonghai 1 year, 2 months ago
for a given company, the data poits starts with the same stock symbol. The dataset is not distrubuted. It is not a good option.
upvoted 3 times

taepyung 2 years, 10 months ago
I agree with u
upvoted 2 times

Abhi16820 1 year, 3 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 7 times

musumusu Most Recent 3 days, 14 hours ago
Answer A:
Trick to remember: Row-key adjustment always be like in decending 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 1 times

zellck 2 months, 3 weeks 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 3 times

AzureDP900 1 month, 4 weeks ago
I agree with you . A is right
upvoted 1 times

MaxNRG 1 year, 1 month ago
Selected Answer: A
A: https://cloud.google.com/bigtable/docs/schema-design-time-series#prefer_rows_to_column_versions
upvoted 3 times

JG123 1 year, 3 months ago
Correct: A
upvoted 1 times

JayZeeLee 1 year, 3 months ago
A and B would both work, since both would distribute the work. This question is not framed properly.
upvoted 1 times

sumanshu 1 year, 8 months ago
Vote for A
upvoted 3 times

Jay3244 1 year, 12 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 2 years, 3 months ago
I think A
upvoted 2 times

kavs 2 years, 3 months ago
Catch here is current Rowley starts with timestamp which should not be in the starting or end position so symbolmshould be prefixed before timestamp
upvoted 1 times

Cloud_Enthusiast 2 years, 3 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 2 years, 5 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
upvoted 2 times

Diqtator 2 years, 5 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 2 years, 5 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 2 years, 6 months ago
B is correct.
A is incorrect. The docuemntation don´t recommend constants in the row key because the balance is not efficiente. The are 2 methods to avoid hotspoting. Promotion Field (use a UserId BEFORE the timestamp) and Salting (use timestamp-hash divided by 3 y put it before the timestamp)
upvoted 1 times

atnafu2020 2 years, 6 months ago
Aggree with promotion field and salting. But, there is no constant. A stock symbol is a unique series of letters assigned to security for trading purposes.
upvoted 2 times

daghayeghi 2 years ago
A is correct:
You deny your sentence, User ID means Stock Symbol, then B is correct.

upvoted 1 times

daghayeghi 2 years ago

, then A is correct.

upvoted 1 times

Rajuuu 2 years, 7 months ago

Answer is A ..Prefix the timestamp of the rowdy to avoid hot spotting.

upvoted 3 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 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 (80%) A (20%)

dambilwa Highly Voted 2 years, 8 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 it's rate of change
upvoted 20 times

Barniyah 2 years, 8 months ago
Yes, you are right, it should be B:
Thank you
upvoted 4 times

marioferrulli 1 year, 2 months 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_bytes remain constant while the num_undelivered_messages increases? I don't get why the destination's used bytes should decrease.
upvoted 1 times

szefco 1 year, 2 months 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 5 times

baubaumiaomiao 1 year, 2 months 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] Highly Voted 2 years, 11 months ago
Correct - B
upvoted 20 times

musumusu Most Recent 3 days, 14 hours 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 1 times

atlan 1 month, 2 weeks 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

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 month, 1 week ago
look here: https://cloud.google.com/monitoring/api/metrics_gcp

instance/storage/used_bytes GA
Storage used.
upvoted 1 times

AzureDP900 1 month, 3 weeks 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 month, 4 weeks ago
B is right
upvoted 1 times

Catweazle1983 2 months 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

zellck 2 months, 3 weeks ago
Selected Answer: B
B is the answer.

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 2 times

A1000 6 months, 1 week ago
Selected Answer: B
Increase subscription/num delivered message
decrease instance/storage/used bytes
upvoted 1 times

Pime13 7 months, 3 weeks ago
Selected Answer: B
Correct - B
upvoted 1 times

JG123 1 year, 3 months ago
Correct: B
upvoted 2 times

Abhi16820 1 year, 3 months ago
isn't B and C are same.
upvoted 2 times

JayZeeLee 1 year, 3 months ago
B.
It's useful to monitor the source that keeps sending data while the destination that doesn't take anything in.
upvoted 2 times

squishy_fishy 1 year, 4 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
upvoted 2 times

squishy_fishy 1 year, 5 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 11 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

ref: https://cloud.google.com/monitoring/api/metrics_gcp#pubsub/subscription/num_undelivered_messages

upvoted 1 times

sumanshu 1 year, 8 months ago

Looks B

upvoted 3 times

gcper 1 year, 12 months ago

A

"An alert based on a decrease of subscription/num_undelivered_messages for the source"
The more we have undelivered messages, the worse. Thus we want to be alerted when the ratio goes down as the denominator goes up.

"A rate of change increase of instance/storage/used_bytes for the destination"
An increase in the rate of change of how much we are storing per instance storage.

upvoted 5 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 2 years, 11 months ago
Should be C
upvoted 20 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: C
Description: Pubsub is global and dataflow can scale workers
upvoted 18 times

musumusu Most Recent 3 days, 13 hours 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 1 times

musumusu 1 week, 3 days 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 2 months, 3 weeks 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 3 times

AzureDP900 1 month, 4 weeks ago
Agree with your explanation
upvoted 1 times

MisuLava 3 months, 4 weeks 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 7 months, 1 week ago
Selected Answer: C
Answer is C
upvoted 1 times

Prasanna_kumar 1 year ago
Answer is option C
upvoted 1 times

ivanhsiav 1 year, 7 months ago

Answer c
kafka cluster in on-premise for streaming msgs
pub/sub for streaming msgs in cloud
upvoted 4 times

sumanshu 1 year, 7 months ago
Vote for C
upvoted 4 times

Allan222 2 years ago
Should be C
upvoted 4 times

daghayeghi 2 years 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 ago
Answer - C
upvoted 4 times

Alasmindas 2 years, 3 months ago
Easy Question : ANswer is Option C.
Alterative 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 2 years, 6 months ago
C
the issue is with a single Kafka cluster is the need to scale automatically with Dataflow
upvoted 4 times

haroldbenites 2 years, 6 months ago
C is correct
upvoted 4 times

Rajuuu 2 years, 7 months ago
Answer is C. Pub/Sub is the messaging tool for Global.
upvoted 5 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

Ganshank Highly Voted 2 years, 10 months ago
A,B
https://cloud.google.com/datastore/docs/export-import-entities
upvoted 33 times

salsabilsf 1 year, 10 months ago
"while keeping the costs"

should be A,D
upvoted 6 times

MrCastro 1 year, 6 months ago
Big query streaming inserts ARE NOT cheap
upvoted 9 times

hellofrnds 1 year, 5 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 2 years, 6 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. Tho this statment 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 22 times

AzureDP900 1 month, 4 weeks ago
A, B is perfect
upvoted 1 times

aparna4387 1 year, 3 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 4 times

Yiouk 1 year, 6 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/time-travel#limitation
upvoted 1 times

Chelseajcole 1 year, 5 months ago
you wanna say invalid?
upvoted 1 times

tavva_prudhvi 11 months, 1 week 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 1 times

zellck Most Recent 2 months, 3 weeks ago
Selected Answer: AB
AB is the answer.
upvoted 3 times

NicolasN 2 months, 4 weeks 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 in a different environment" (an objection to the word "namespace", I think it should be just "name")
upvoted 6 times

NicolasN 2 months, 4 weeks 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 or the 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
❌ [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
❌ [E] Cloud Source Repositories are for source code and not a suitable storage for this case.
upvoted 4 times

John_Pongthorn 5 months ago
Selected Answer: AB
https://cloud.google.com/datastore/docs/export-import-entities
upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: AB
https://cloud.google.com/datastore/docs/export-import-entities
upvoted 1 times

John_Pongthorn 5 months, 2 weeks 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 7 months, 1 week 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 are not created for achieving purposes.
upvoted 1 times

AmirN 8 months, 2 weeks ago
Option B is 36 times more expensive than C
upvoted 1 times

Nico1310 1 year, 1 month ago
Selected Answer: AB
AB. for sure streaming to BQ its quite expensive!

upvoted 2 times

MaxNRG 1 year, 1 month 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 10 months, 1 week ago

D is wrong, BQ Streaming inserts costs are high!

upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 4 months ago

Answer is A, B.
https://cloud.google.com/datastore/docs/export-import-entities#exporting_specific_kinds_or_namespaces

upvoted 1 times

sergio6 1 year, 4 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 3 times

Chelseajcole 1 year, 5 months ago

Vote A B . What's the purpose load into bigquery?

upvoted 1 times

Chelseajcole 1 year, 5 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 request must include one or more kind names in the entity filter.

You have to specify an entity fliter before you can load from datastore to BQ. It didn't mention that at all. So C is incorrect

upvoted 3 times

fire558787 1 year, 6 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 the 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 two options that allow data to be restored to Datastore are A and B.

upvoted 7 times

DeepakS227 1 year, 7 months ago

Answer should be A, B
Here the requirement is yo have PIT recovery, by periodicaaly copy data to a differenent datastore , recovery can be faster.

upvoted 2 times

Question #123 *Topic 1*

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 must 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 (91%) 9%

rickywck Highly Voted 2 years, 11 months ago
I think AD is the answer. E will not improve performance.
upvoted 36 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A, D
Description: Denormalization will help in performance by reducing query time, update are not good with bigquery
upvoted 19 times

awssp12345 1 year, 7 months ago
My guess is append has better performance than update.
upvoted 3 times

odacir Most Recent 2 months, 3 weeks 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 1 times

zellck 2 months, 3 weeks 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 3 times

AzureDP900 1 month, 4 weeks ago
A, C is correct I agree
upvoted 1 times

NicolasN 3 months, 3 weeks 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 significantly 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 2 months, 1 week 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 3 months, 3 weeks ago
(mistakenly voted AC instead of AB)
upvoted 1 times

DerickTW 5 months, 1 week ago
Selected Answer: AC
The DML quota limit is removed since 2020, I think C is better than D now.
upvoted 1 times

devaid 4 months, 2 weeks 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 performant. 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 3 times

MaxNRG 1 year, 1 month 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
D: BigQuery append
upvoted 3 times

medeis_jar 1 year, 1 month 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 1 year, 3 months ago
I think AD. E is not valid because it use external table which is not good for performance
upvoted 1 times

sumanshu 1 year, 7 months ago
A - correct (denormlization 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 1 year, 8 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 team 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 1 year, 4 months ago
It didn't mention cost is a concern
upvoted 1 times

retep007 1 year, 5 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 1 year, 11 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 entire 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 storage

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, you can use BigQuery's native support for date-partitioned tables and partition expiration. In other words, BigQuery can automatically delete older data.
https://cloud.google.com/solutions/bigquery-data-warehouse#handling_change

upvoted 4 times

Hithesh 1 year, 11 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 1 year, 7 months ago
C not possible, maximum 1500 updates possible in a day

upvoted 1 times

raf2121 1 year, 7 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 1 year, 7 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 20. 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 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 ago
you can update bigquery 1500 times in a day

upvoted 3 times

daghayeghi 1 year, 11 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

upvoted 3 times

Nams_139 2 years, 3 months ago
A,D Since the requirements are both performance and usability.

upvoted 5 times

federicohi 2 years, 3 months ago
i tink may be ita AC becuase appending its worst to increase dataset size. THe question seems to put like a problem the size of dataset and performance to datascience so inserting more rows decrease performace for them.

upvoted 3 times

Ram459 2 years, 3 months ago
AD looks like a good fit

upvoted 4 times

Question #124 *Topic 1*

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%)

AzureDP900 1 month, 4 weeks ago

D is right

upvoted 1 times

dconesoko 2 months ago

Selected Answer: D

vote for D

upvoted 1 times

zellck 2 months, 3 weeks ago

Selected Answer: D

D is the answer.

upvoted 1 times

jkhong 4 months, 1 week 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 the answer

upvoted 2 times

devaid 4 months, 2 weeks ago

Selected Answer: D

D of course

upvoted 2 times

kenanars 5 months, 3 weeks 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 2 years, 11 months ago
Answer C.
upvoted 34 times

AJKumar Highly Voted 2 years, 8 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 20 times

AzureDP900 1 month, 4 weeks ago
Agree with C
upvoted 1 times

jkhong Most Recent 2 months, 1 week ago
Selected Answer: C
D → does not guarantee 100% queryable or accessible/available
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: C
C is the answer.
upvoted 1 times

Smaks 7 months, 3 weeks 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 7 months, 3 weeks ago
ignore this comment, please
upvoted 6 times

Aslkdup 12 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 1 year, 1 month ago
Selected Answer: C
both requirements are full filled.
upvoted 2 times

MaxNRG 1 year, 1 month ago
Selected Answer: D
D: BigQuery + Cloud Storage
upvoted 1 times

jkhong 2 months, 1 week ago
D → does not guarantee 100% queryable or accessible/available
upvoted 1 times

medeis_jar 1 year, 1 month 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 tools in other cloud providers?"
Analytics -> BQ
Exposing -> GCS

upvoted 4 times

JG123 1 year, 3 months ago
Correct: C

upvoted 2 times

xiaofeng_0226 1 year, 6 months ago
vote for C

upvoted 3 times

sumanshu 1 year, 7 months ago
Vote for 'C'

A - Only Half requirement fulfil, expose as a file not getting fulfiled
B - Not a warehouse
C. Both requirements fulfiled...Bigquery and GCS
D. Both requirement fulfiled...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 1 year, 12 months ago
C

BigQuery for analytics processing and Cloud Storage for exposing the data as files

upvoted 3 times

daghayeghi 2 years ago
answer A:
with BigQuery Omni it is now possible to read data from other cloud providers without transferring the data to GCP and thereby by saving on egress charges.
https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-omni

upvoted 4 times

salsabilsf 1 year, 10 months ago
the question says :
"expose the dataset as files" wich means Cloud Storage

upvoted 2 times

Ram459 2 years, 3 months ago
C is the right answer

upvoted 3 times

Cloud_Enthusiast 2 years, 3 months ago
Answer is D. However with BigQuery Omni it is now possible to read data from other cloud providers without transferring the data to GCP and thereby by saving on egress charges.

https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-omni

upvoted 3 times

snamburi3 2 years, 3 months ago
But, Omni does not mention storage for batch analysis. the data will be still be available through Bigquery (I am no sure about it)....

upvoted 1 times

DeepakKhattar 2 years, 5 months ago
Choice is between C and D, C - makes ticks both the boxes.
1. Perform data warehouse-style analytics (you need historical data, data can be stored on files in 20/80 ratio but other cloud providers will not have access to 20% data, if we have 5 year worth data then 20% means other cloud providers will be missing out 1 year data)
2. expose the dataset as files for batch analysis tools in other cloud providers. (A and B does not meet this requirement)

upvoted 2 times

Question #126 *Topic 1*

You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-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

Callumr Highly Voted 2 years, 8 months ago
B - You only need a PoC and it has be done quickly
upvoted 45 times

[Removed] Highly Voted 2 years, 11 months ago
Correct - A
upvoted 20 times

musumusu Most Recent 1 week, 3 days 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 1 times

techtitan 2 weeks, 6 days ago
A - https://cloud.google.com/vertex-ai/docs/beginner/beginners-guide Target at least 1000 examples per target
upvoted 1 times

techtitan 2 weeks, 6 days ago
The quick POC part can be achieved by using Auto ML instead of creating and training your own model
upvoted 1 times

odacir 2 months, 3 weeks 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
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 2 times

AzureDP900 1 month, 4 weeks ago
A is correct
upvoted 2 times

zellck 2 months, 3 weeks ago
Selected Answer: A
A is the answer.

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

odacir 2 months, 3 weeks ago
Agreed !
upvoted 1 times

gudiking 3 months, 1 week ago
A - https://cloud.google.com/vision/automl/docs/beginners-guide#include_enough_labeled_examples_in_each_category

upvoted 1 times

MarielaYBird 3 months, 2 weeks 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 dataset twice.

https://cloud.google.com/vision/automl/object-detection/docs/prepare#how_big_does_the_dataset_need_to_be

upvoted 2 times

josrojgra 4 months, 1 week 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 and the problem is the quality of the data and not the amount of it.

upvoted 3 times

John_Pongthorn 5 months ago

Selected Answer: A

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 5 months ago

The more labels, the more accurate the result.

upvoted 1 times

changsu 5 months, 3 weeks ago

Selected Answer: B

750*1000 are a lot.

upvoted 1 times

ducc 6 months ago

Selected Answer: A

It is labeled, so A is correct

upvoted 1 times

civilizador 6 months, 1 week ago

It's A.
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 4 times

civilizador 6 months, 1 week 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 6 months, 4 weeks ago

Option A & B are quite close. Refer: 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 9 months ago

Selected Answer: B

B
cant choose A because model needs to pass through the dataset several times for a proof of concept, existing data set samples might not be all seen in several working days causing over generalization

upvoted 1 times

Kriegs 9 months, 1 week 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

[Removed] 9 months, 1 week ago

Selected Answer: A

AutoML work well with more than 100 training images/label but ideally with 1000 of each.

upvoted 1 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

dhs227 Highly Voted 2 years, 11 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
upvoted 53 times

aiguy Highly Voted 2 years, 11 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 42 times

gopinath_k 1 year, 11 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 5 times

tavva_prudhvi 11 months, 1 week 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

musumusu Most Recent 3 days, 12 hours 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

zellck 2 months, 3 weeks ago
Selected Answer: D
D is the answer.

https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus
Cloud TPUs are optimized for specific workloads. In some situations, you might want to use GPUs or CPUs on Compute Engine instances to run your machine learning workloads. In general, you can decide what hardware is best for your workload based on the following guidelines:
CPUs
- Models that are dominated by custom TensorFlow operations written in C++
upvoted 3 times

AzureDP900 1 month, 4 weeks ago
D is right
upvoted 1 times

odacir 2 months, 3 weeks ago
But GPU is not a step between CPU and TPU in cost and performance? Why not use GPU in this case?
upvoted 2 times

Azlijaffar 4 months ago
CPUs
-Quick prototyping that requires maximum flexibility

-Simple models that do not take long to train
-Small models with small effective batch sizes
-Models that are dominated by custom TensorFlow operations written in C++
-Models that are limited by available I/O or the networking bandwidth of the host system
GPUs
-Models for which source does not exist or is too onerous to change
-Models with a significant number of custom TensorFlow operations that must run at least partially on CPUs
-Models with TensorFlow ops that are not available on Cloud TPU (see the list of available TensorFlow ops)
-Medium-to-large models with larger effective batch sizes
TPUs
-Models dominated by matrix computations
-Models with no custom TensorFlow operations inside the main training loop
-Models that train for weeks or months
-Larger and very large models with very large effective batch sizes

upvoted 6 times

Azlijaffar 4 months ago

Hi. I just did the exam. A similar question came up. It said that the model partially runs on custom TensorFlow ops. They need to speed up the time it takes for the model to complete its run. What type of accelerators to use etc. I chose GPUs as supported above. Hope this helps.

upvoted 3 times

Prakzz 2 months, 1 week ago

That's a different question in this series. This is a diffferent question/scenario

upvoted 1 times

Azlijaffar 4 months ago

-Use TPUs if model takes weeks or months. This model only takes days. Since model is DOMINATED by custom TF ops in C++, can stick with CPUs. Ans should be D.

upvoted 1 times

AmirNaik204 6 months, 1 week ago

Vote for C. GPU as the question says "performing bulky matrix multiplications" and "low cost"

upvoted 3 times

Pime13 7 months, 2 weeks ago

Selected Answer: D

https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus

upvoted 2 times

VGalan 9 months, 2 weeks ago

Selected Answer: D

https://cloud.google.com/tpu/docs/tpus

CPUs
Models that are dominated by custom TensorFlow operations written in C++
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 2 times

wences 11 months ago

Selected Answer: D

D according to documentation

upvoted 1 times

tavva_prudhvi 10 months, 3 weeks ago

Which documentation, please share the reference.

upvoted 2 times

PJG_worm 11 months, 1 week ago

B: TPU: performing bulky matrix multiplications
GPU: support CPP

upvoted 1 times

Aslkdup 12 months ago

Answer is B.
TPUs are optimized to perform fast, bulky matrix multiplication and developed model is using bulky matrix multiplication.

upvoted 1 times

AM10593 1 year ago

Selected Answer: D

D: CPU is recommended for models that use custom TensorFlow operations written in C++ [Dan Sullivan]

upvoted 5 times

Bhawantha 1 year, 1 month ago

D.
"Cloud TPUs are not suited to 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."
because TPUs are not suited for this application.

The battle between GPU vs CPU.
Specially mentioned that the C++ part in the description. Since we can go with CPU.

upvoted 3 times

eagle_fang 1 year, 1 month ago

D
https://cloud.google.com/tpu/docs/tpus

upvoted 1 times

MaxNRG 1 year, 1 month ago

Selected Answer: D

D:
https://cloud.google.com/tpu/docs/tpus
Cloud TPUs are optimized for specific workloads. In some situations, you might want to use GPUs or CPUs on Compute Engine instances to run your machine learning workloads. In general, you can decide what hardware is best for your workload based on the following guidelines:

upvoted 7 times

MaxNRG 1 year, 1 month ago

CPUs:
• Quick prototyping that requires maximum flexibility
• Simple models that do not take long to train
• Small models with small effective batch sizes
• Models that are dominated by custom TensorFlow operations written in C++
• Models that are limited by available I/O or the networking bandwidth of the host system

upvoted 4 times

MaxNRG 1 year, 1 month ago

GPUs:
• Models that are not written in TensorFlow or cannot be written in TensorFlow
• Models for which source does not exist or is too onerous to change
• Models with a significant number of custom TensorFlow operations that must run at least partially on CPUs
• Models with TensorFlow ops that are not available on Cloud TPU (see the list of available TensorFlow ops)
• Medium-to-large models with larger effective batch sizes

upvoted 3 times

MaxNRG 1 year, 1 month ago

TPUs:
• Models dominated by matrix computations
• Models with no custom TensorFlow operations inside the main training loop
• Models that train for weeks or months
• Larger and very large models with very large effective batch sizes
TPUs are optimized to perform fast, bulky matrix multiplication.
Cloud TPUs are not suited to 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.
https://cloud.google.com/tpu/docs/tpus

upvoted 3 times

medeis_jar 1 year, 1 month ago

Selected Answer: D

Only D -> https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus

upvoted 2 times

kishanu 1 year, 2 months ago

Selected Answer: B

Confused between B and D. Refer to the below link
https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus
Custom Tensorflow operations - CPUs
Bulky Matrix calc - TPUs

upvoted 3 times

Question #128 *Topic 1*

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 of batch normalization) to avoid overfitting.

D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

Correct Answer: *D*

Community vote distribution

D (88%) 13%

Callumr Highly Voted 2 years, 8 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 45 times

hellofrnds 1 year, 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 is twice of training error. So, it is clearly overfitting. Isn't it?
upvoted 3 times

odacir 2 months, 3 weeks ago
NO, its underfitting.
upvoted 2 times

hellofrnds 1 year, 4 months ago
So, answer should be C
upvoted 1 times

tavva_prudhvi 10 months, 3 weeks 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

alecuba16 6 months, 1 week ago
It's overfitting.

Overfitting->low rmse in train / high accuracy-f1 score in train for classification.

Underfitting -> high rmse / low f1score or accuracy in train, you don't have to look into test set if there is an underfitting problem.
upvoted 1 times

velliger 1 year, 3 months ago
High rmse: The model is underfitting the train data. To reduce overfitting, we increase the number of layers in the model or we change the type of layer.
upvoted 1 times

velliger 1 year, 3 months ago
*underfitting
upvoted 1 times

[Removed] Highly Voted 2 years, 11 months ago
should be D
upvoted 18 times

musumusu Most Recent 1 week, 3 days ago
Answer D
RMSE shows the variation from actual value and predictive value. If it is higher on RMSE, then you should definitely reject the model. In this case it is saying, RMSE is higher in training and lower in test. Which means model could be better trained and the guy was sleeping doing this work. So, it means, it released the underfit model or poor model in simple term. So what you should do, to reduce RMSE, you increase complexity of model by adding more Neurons and Hidden layers.

If it was overfit, means very high RMSE in test while showing almost 0 during training, it means it was overfit.
So how you fix overfit, Feature Engineering, Regularisation, Hypertunning, dropout (reduce feature) .. bla bla bla

upvoted 1 times

jkh_goh 1 month ago

Obviously the question meant to say "error twice as high on the test set as on the train set", thus overfitting.

upvoted 1 times

desertlotus1211 1 month ago

This will help in understanding overfitting vs. underfitting
https://daviddalpiaz.github.io/r4sl/regression-for-statistical-learning.html

upvoted 1 times

Astrophile 1 month, 3 weeks ago

Selected Answer: C

There are 100M labeled examples in the dataset and the train and test ratio is 90:10, hence the model is capturing all the noises with a pattern in the data. It is sufficient enough to indicate the overfitting the model as "Training error is small and test error is big" .

upvoted 1 times

Azlijaffar 4 months ago

The wording is tricky. "RMSE twice as high on train set as on test set" means Train RMSE is higher than Test RMSE. Means train data has more errors than test data. More errors = underfitting.
FYI: i have twice as many apples as you means i have more apples than you. hahaha.

upvoted 3 times

Azlijaffar 4 months ago

Answer should be D.

upvoted 1 times

YorelNation 5 months, 3 weeks ago

I don't get how in an underfit context you can have twice the RMSE on the train set (meaning twice less on the test set ==> You have doubled performance on your testing set ?) The question seems very odd to me.

upvoted 2 times

jkhong 4 months, 1 week ago

Many cases, it can just be a coincidence that your test data yields a better performance than your train, where your underfitted training data is just the right fit for this specific test data

upvoted 1 times

VGalan 9 months, 2 weeks ago

Selected Answer: D

The lower the RMSE, the better a given model is able to “fit” a dataset

Overfitting: Model perform well on train set and poorly on test set. In this case, model performs better in test set. Cant not be Overfitting.
The options B and C are for overfitting cases.

We need to improve the train perform so in my opinion, D is the correct answer

upvoted 2 times

MaxNRG 1 year, 1 month ago

Selected Answer: D

D:
A is incorrect since test sample is large enough.
B is incorrect since dataset is pretty large already, and having more data typically helps with overfitting and not with underfitting.
C is incorrect since regularization helps to avoid overfitting and we have a clear underfitting case.
D is correct since increasing model complexity generally helps when you have an underfitting problem.

upvoted 5 times

MaxNRG 1 year, 1 month ago

https://developers.google.com/machine-learning/crash-course/generalization/peril-of-overfitting
https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d
https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/l2-regularization
https://developers.google.com/machine-learning/crash-course/training-neural-networks/best-practices

upvoted 1 times

Pupina 1 year, 2 months ago

Can you add the link to say is underfit or overfit. I understand that is overfit because in the training set there are not significant error but there are in the testing set. That means that there are noise in the training data set.

upvoted 1 times

lifebegins 1 year, 3 months ago

"Training error is small and test error is big" is an indication of overfitting. Answer is C:

Dont add more complexity.

upvoted 3 times

lifebegins 1 year, 3 months ago

RMSE of test > RMSE of train => OVER FITTING of the data.
RMSE of test < RMSE of train => UNDER FITTING of the data.

upvoted 5 times

sumanshu 1 year, 7 months ago

Vote for D

upvoted 6 times

Sumanth09 1 year, 11 months ago

question says 100M labels and saying high error on testing data - if it's Underfitting model, model might get high error on train data.
1) Testing data might not representing the training data. if it is the case, need to add more data
2) Error might be due to more labels. if it is the case, need to apply regularization
Answer might be B or C

upvoted 3 times

arghya13 2 years, 3 months ago

Problem of underfitting..D is the correct answer

upvoted 2 times

snamburi3 2 years, 3 months ago

D. this question is in the sample exam and the correct answer shows as D.

upvoted 8 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 (100%)

[Removed] Highly Voted 2 years, 11 months ago

Should be B

upvoted 21 times

Ganshank Highly Voted 2 years, 10 months ago

B

The questions is specifically about organizing the data in BigQuery and storing backups.

upvoted 12 times

zellck Most Recent 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 1 times

John_Pongthorn 5 months, 1 week ago

Selected Answer: B

B

https://cloud.google.com/architecture/dr-scenarios-for-data#BigQuery

upvoted 2 times

MaxNRG 1 year, 1 month 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 5 times

MaxNRG 1 year, 1 month 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 pricing and the 90 day countdown starts again.

upvoted 3 times

MaxNRG 1 year, 1 month 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 that 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 1 year, 1 month 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 month, 4 weeks ago
Agreed
upvoted 1 times

lifebegins 1 year, 3 months ago
https://cloud.google.com/bigquery/docs/time-travel

Answer should be B only, because snapshot decorator u can time travel only for 7 days
upvoted 2 times

vintop95 1 year, 4 months ago
B:
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.
https://cloud.google.com/architecture/dr-scenarios-for-data#managed-database-services-on-gcp
upvoted 4 times

Mcloudgirl 3 months, 2 weeks ago
Thanks for explaining this - I previously was unsure why separate tables was necessary.
upvoted 2 times

Ysance_AGS 1 year, 5 months ago
B since the question is about saving costs when storing backups : use Cloud Storage.
upvoted 3 times

raf2121 1 year, 7 months ago
Point for discussion

Is A an Option ? One of the information is "BQ is Centralized analytics Platform" - Will not be good to store the data in single table rather separate table for each month (for example, if One needs to analyze a trend for last two years, will not be complicated to have 24 tables in the query)
upvoted 1 times

p111111111111 1 year, 7 months ago
why not A
upvoted 3 times

sumanshu 1 year, 7 months ago
A & B - Not sure (what to choose)
C - elimnated (because of cost)
D - eliminated (because of error detected in 2 weeks)
upvoted 1 times

shanjin14 1 year, 9 months ago
How about C?
https://medium.com/google-cloud/how-to-backup-google-big-query-5f078138cedc
Sounds like alternative 2 in above article
upvoted 1 times

z8zhong 1 year, 8 months ago
Because "your backups should be optimized for storage costs"
upvoted 2 times

someshsehgal 2 years ago
Why nobody supporting C. Data management will be easy with data reside only within BQ. Also long term storage cost in BQ is almost similar like GCS
upvoted 4 times

ABM9 2 years, 2 months ago
Should be B
Using snapshot decorators , recovery is valid only for a period of 7 days. Here it says 2 weeks so "D" is ruled out.

You can undelete a table within seven days of deletion, including explicit deletions and implicit deletions due to table expiration. After seven days, it is not possible to undelete a table using any method, including opening a support ticket.
https://cloud.google.com/bigquery/docs/managing-tables
upvoted 8 times

aleedrew 2 years, 4 months ago
It says sometimes it is discovered in two weeks, not all of it is discovered in two weeks. I am inclined to choose D.
upvoted 2 times

kino2020 2 years, 5 months ago

B
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 pricing and the 90 day countdown starts again.
https://cloud.google.com/solutions/dr-scenarios-for-data#managed-database-services-on-gcp

upvoted 7 times

patitonav 1 year, 7 months ago

The link you provided supports B!

upvoted 2 times

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?

A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.

B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.

C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.

D. 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 2 years, 11 months ago
Should be D.

https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery
upvoted 27 times

Rajuuu 2 years, 7 months ago
There is no mention about merge or limit in the link provided.
upvoted 2 times

Chelseajcole 1 year, 5 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 2 years, 11 months ago
Should be D
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 15 times

AzureDP900 1 month, 4 weeks ago
D is right
upvoted 1 times

AACHB 1 year, 2 months ago
I had it in the exam (14/12/2021)
upvoted 3 times

GCPLearning2021 1 year, 1 month ago
Does examtopics questions help?
upvoted 1 times

nellyoaid 1 year, 2 months ago
Please what was the answer? @AACHB
upvoted 1 times

Atnafu Most Recent 3 months, 2 weeks 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 1 times

MisuLava 4 months ago

there is no update quota anymore.
but i would say D

upvoted 1 times

amitsingla012 9 months, 4 weeks ago

Option D is the right answer

upvoted 1 times

tavva_prudhvi 10 months, 3 weeks 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 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 2 times

MaxNRG 1 year, 1 month ago

Selected Answer: D

D:
BigQuery is primarly designed and suit 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 arhitect 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 1 year, 1 month ago

Selected Answer: D

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 1 year, 3 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 1 year, 7 months ago

Vote for D

upvoted 4 times

daghayeghi 1 year, 11 months ago

D:
https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery

upvoted 3 times

daghayeghi 2 years ago

D:
https://cloud.google.com/blog/products/bigquery/performing-large-scale-mutations-in-bigquery

upvoted 3 times

SteelWarrior 2 years, 5 months ago

D should be the answer. Avoid updates in Datawarehousing environment instead use merge to create a new table.

upvoted 3 times

VIncent9261111 2 years, 5 months ago

Quota Exceed Error Solution from BQ Documentation -
View the message property of the error object for more information about which quota was exceeded. To reset or raise a BigQuery quota, contact support. To modify a custom quota, submit a request from the Google Cloud Console page. If you receive this error using the BigQuery sandbox, you can upgrade.

upvoted 1 times

haroldbenites 2 years, 6 months ago

D is correct.
https://cloud.google.com/bigquery/quotas#queries.
The error is because of the number of statements over the table , but not because of the number of records.

upvoted 5 times

Chelseajcole 1 year, 5 months ago

What is number of statements mean?

upvoted 1 times

haroldbenites 2 years, 6 months ago

D is correct. Obviously.

upvoted 4 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

[Removed] Highly Voted 2 years, 11 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 29 times

sipsap 2 years, 3 months ago
"Each project requires unique access control configurations" rules out hirearchy
upvoted 11 times

AJKumar Highly Voted 2 years, 8 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 18 times

hauhau 1 year, 6 months ago
Question mention minimize IAM policies,but E should create complex policies
upvoted 5 times

zellck Most Recent 2 months, 3 weeks ago
Selected Answer: BC
BC is the answer.
upvoted 2 times

FrankT2L 8 months, 3 weeks 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 4 times

annie1196 1 year, 1 month ago
A and C is correct, same question I encountered on Udemy.
upvoted 2 times

tavva_prudhvi 11 months, 1 week 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 month ago

you explanation for D is incorrect....
upvoted 1 times

MaxNRG 1 year, 1 month 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 descendents 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 month, 4 weeks ago
BC is the answer
upvoted 1 times

MaxNRG 1 year, 1 month 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

medeis_jar 1 year, 1 month 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

MaxNRG 1 year, 1 month ago
good point, AC looks better, agreed
upvoted 1 times

MaxNRG 1 year, 1 month 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 1 year, 1 month ago
So, I stay with BC :)))
upvoted 2 times

hellofrnds 1 year, 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 question. When 2 project communicate, service account should be used
upvoted 8 times

sumanshu 1 year, 7 months ago
Vote for B & C
upvoted 4 times

daghayeghi 1 year, 11 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 2 years, 2 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 4 times

sumanshu 1 year, 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 1 year, 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 1 year, 7 months ago
Agree with Sumanshu
upvoted 1 times

ceak 2 years, 2 months ago
C & E are the correct answers.
upvoted 3 times

sumanshu 1 year, 7 months ago
E is a long process and we need to simply the process...So E is ruled out
upvoted 1 times

vito9630 2 years, 3 months ago
Answer: B, C
upvoted 1 times

kavs 2 years, 3 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 2 years, 4 months ago
C and E
upvoted 3 times

atnafu2020 2 years, 6 months ago
B & C
the correct answer, as google suggest
upvoted 1 times

haroldbenites 2 years, 6 months ago
C, E is correct.
upvoted 3 times

Question #132 *Topic 1*

Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend 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

[Removed] Highly Voted 2 years, 11 months ago
Answer: B
Description: All the criteria meets for Spanner
upvoted 24 times

sumanshu Highly Voted 1 year, 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 16 times

AzureDP900 Most Recent 1 month, 4 weeks ago
B is the answer
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: B
B is the answer.
upvoted 1 times

Remi2021 5 months, 2 weeks 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 2 times

JamesKarianis 7 months, 1 week ago
Selected Answer: B
Spanner is globally available and meets all the requirements
upvoted 2 times

devric 10 months, 3 weeks 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 specifing the Data Location it's gonna be stored in "US Multiregion" by default
upvoted 2 times

MaxNRG 1 year, 1 month 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 4 times

Tanmoyk 2 years, 5 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 7 times

haroldbenites 2 years, 6 months ago

B is correct

upvoted 2 times

Archy 2 years, 7 months ago

B, as Cloud Spanner has three types of replicas: read-write replicas, read-only replicas, and witness replicas.

upvoted 6 times

VishalB 2 years, 7 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 2 years, 6 months ago

but bigquery does so why not A?

upvoted 1 times

WizzzardLlama 2 years 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] 1 year, 10 months ago

The requirement is for transactional application serving customers , not analytical so BQ is ruled out.

upvoted 1 times

mAbreu 2 years, 5 months ago

wrong, cloud spanner can have read-only replicas
https://cloud.google.com/spanner/docs/replication?hl=pt-br

upvoted 3 times

norwayping 2 years, 8 months ago

I was wrong Bigtable doesn not support ansi SQL. B instead

upvoted 2 times

norwayping 2 years, 8 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 2 years, 11 months ago

Answer B

upvoted 4 times

[Removed] 2 years, 11 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 (100%)

rickywck Highly Voted 2 years, 11 months ago
I think the key reason for pick D is the 100ms requirement.
upvoted 25 times

AzureDP900 1 month, 4 weeks 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] Highly Voted 2 years, 11 months ago
Answer: D
Description: Bigtable provides lowest latency
upvoted 12 times

musumusu Most Recent 1 week, 3 days ago
what is wrong with B ? View can be precomputed and cached and it can definitely satisfy the 100 miliseconds request. create a pipeline to send data to bigtable .. don't you think its too much to run a simple prediction query ?
upvoted 1 times

hiromi 3 months, 1 week ago
Selected Answer: D
Vote for D
upvoted 1 times

HarshKothari21 5 months, 3 weeks ago
Selected Answer: D
Option D
upvoted 1 times

sumanshu 1 year, 7 months ago
Vote for D, requirement to serve predictions with in 100 ms
upvoted 6 times

Tanmoyk 2 years, 5 months ago
D is correct , 100ms is most critical factor here.
upvoted 5 times

sh2020 2 years, 5 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 2 years, 1 month ago
Predictions are computed in advance for all users and written to BigTable for low-latency serving.
upvoted 3 times

haroldbenites 2 years, 6 months ago

D is correct

upvoted 4 times

Rajokkiyam 2 years, 11 months ago

Answer D.

upvoted 6 times

[Removed] 2 years, 11 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 (95%) 5%

[Removed] Highly Voted 2 years, 11 months ago
should be B
upvoted 22 times

itche_scratche Highly Voted 2 years, 4 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 11 times

jkhong 2 months, 2 weeks 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

seiyassa 2 years, 2 months ago
I think pubsub doesn't have good connection to dataproc, so D is not the answer
upvoted 3 times

jkhong 2 months, 2 weeks ago
As of Dec 2022,there is the PubSub Lite connector to Dataproc
upvoted 1 times

AzureDP900 Most Recent 1 month, 4 weeks ago
B. Cloud Pub/Sub, Cloud Storage, BigQuery
upvoted 1 times

AzureDP900 1 month, 4 weeks ago
https://cloud.google.com/solutions/stream-analytics/
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: B
B is the answer.
upvoted 3 times

John_Pongthorn 5 months, 1 week 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 5 months, 2 weeks 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 11 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 10 months, 3 weeks ago

You gonna use batch historical export for Spanner? It's B!

upvoted 2 times

Prasanna_kumar 1 year ago

Answer is B

upvoted 1 times

MaxNRG 1 year, 1 month ago

Selected Answer: B

Cloud Pub/Sub, Cloud Dataflow, BigQuery
https://cloud.google.com/solutions/stream-analytics/

upvoted 4 times

medeis_jar 1 year, 1 month 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 9 times

JG123 1 year, 3 months ago

Correct: B

upvoted 1 times

AdrianMonter26 1 year, 3 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 1 year, 7 months ago

Vote for B

upvoted 3 times

sumanshu 1 year, 7 months ago

Vote for B

upvoted 1 times

rgpalop 2 years, 4 months ago

I think B

upvoted 2 times

haroldbenites 2 years, 6 months ago

B is correct

upvoted 2 times

VishalB 2 years, 7 months ago

Correct Answer: B

Streaming Data -- Pub/Sub
ANSI SQL access -- Big Query
Batch historical exports -- Cloud Storage

upvoted 9 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 2 years, 11 months ago
Correct - D
upvoted 42 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: D
Description: All the requirements meet with D
upvoted 16 times

OberstK Most Recent 3 weeks, 3 days ago
Selected Answer: D
Correct - D
upvoted 1 times

desertlotus1211 1 month ago
I believe this was also on the GCP PCA exam as well! ;)
upvoted 1 times

AzureDP900 1 month, 4 weeks ago
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.
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: D
D is the answer.
upvoted 1 times

mbacelar 3 months, 2 weeks ago
Selected Answer: D
For sure D
upvoted 1 times

clouditis 5 months, 1 week ago
D it is!
upvoted 1 times

Prasanna_kumar 1 year ago
Answer is D
upvoted 2 times

MaxNRG 1 year, 1 month 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 1 year, 1 month ago

Selected Answer: D

OMG only D

upvoted 1 times

JG123 1 year, 3 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 10 times

sandipk91 1 year, 6 months ago

Answer is D for sure

upvoted 3 times

sumanshu 1 year, 7 months ago

Vote for D

upvoted 3 times

daghayeghi 2 years ago

D:
because we have to be able to query over historical 2 years data only BigQuery address this issue.

upvoted 3 times

daghayeghi 2 years ago

and because we have lots of input data we have to use Dataflow for processing

upvoted 2 times

atnafu2020 2 years, 6 months ago

D is correct

upvoted 2 times

haroldbenites 2 years, 6 months ago

D is correct

upvoted 3 times

Question #136 *Topic 1*

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 2 years, 11 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 46 times

AzureDP900 1 month, 4 weeks ago

Agreed

upvoted 1 times

sumanshu Highly Voted 1 year, 7 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 12 times

zellck Most Recent 2 months, 3 weeks ago

Selected Answer: AB

AB is the answer.

upvoted 1 times

mbacelar 3 months, 2 weeks ago

Selected Answer: AB

Scale in and Scale Out

upvoted 1 times

FrankT2L 8 months, 3 weeks 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 1 year, 1 month ago

Selected Answer: AB

A & B, other options don't make sense

upvoted 4 times

medeis_jar 1 year, 1 month ago

Selected Answer: AB
Only A & B make sense for improving pipeline performance.
upvoted 2 times

Mjvsj 1 year, 2 months ago
Selected Answer: AB
Should be A & B
upvoted 2 times

daghayeghi 2 years ago
B, E:
B: Dataflow manage number of worker 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 don't have to move data to specific zone and can be remain in same zone of EU, then E is correct.
upvoted 2 times

Vasu_1 1 year, 9 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 2 years, 3 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 perf apanee ruled out not sure of bigtable
upvoted 3 times

Alasmindas 2 years, 3 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 many and does not solve the purpose of the question.
upvoted 3 times

SureshKotla 2 years, 5 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 2 years, 5 months ago
On second thought, A B is looking right
upvoted 2 times

sumanshu 1 year, 7 months ago
automatically taking care of workers up to 3 (as the maximum worker is 3 set as per questions)
upvoted 1 times

atnafu2020 2 years, 6 months ago
AB
is correct
upvoted 2 times

haroldbenites 2 years, 6 months ago
A , E is correct
upvoted 4 times

Ravivarma4786 2 years, 6 months ago
Worker nodes increase and large instance will give good performance
upvoted 2 times

FARR 2 years, 6 months ago
AB
https://cloud.google.com/dataflow/docs/guides/common-errors#tsg-resource-exhausted-no-disk-space
upvoted 2 times

kino2020 2 years, 7 months ago
postscript
maxNumWorkers int The maximum number of Compute Engine instances to be made available to your pipeline during execution. Note that this can be higher than the initial number of workers (specified by numWorkers to allow your job to scale up, automatically or otherwise.
https://cloud.google.com/dataflow/docs/guides/specifying-exec-params

In these examples, the single-cluster instance can handle twice the write throughput that the replicated instance can handle, even though each instance's clusters have a total of 6 nodes.
https://cloud.google.com/bigtable/docs/performance
upvoted 2 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 (100%)

arpitagrawal Highly Voted 5 months, 3 weeks ago

Selected Answer: BC

It should be B and C

upvoted 6 times

ducc Highly Voted 5 months, 4 weeks ago

Selected Answer: BC

BC is correct

Why the comments is deleted?

upvoted 5 times

NamitSehgal Most Recent 4 weeks, 1 day ago

BC only makes sense here , no mention of data, no mention of keeping cost low

upvoted 1 times

AzureDP900 1 month, 4 weeks ago

B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions Most Voted
C. Increase the number of nodes in the Bigtable cluster

upvoted 1 times

ovokpus 3 months, 1 week 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

AzureDP900 1 month, 4 weeks ago
C. Create a Directed Acyclic Graph in Cloud Composer
upvoted 2 times

saurabhsingh4k 2 months, 1 week 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 month, 3 weeks 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 1 times

zellck 2 months, 3 weeks ago
Selected Answer: C
C is the answer.

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 1 times

devaid 4 months, 2 weeks 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 3 times

louisgcpde 4 months, 2 weeks ago
Selected Answer: C
So that I thing the answer should be C (Composer).

upvoted 1 times

louisgcpde 4 months, 2 weeks ago

To me, the point is "automate" the process, so that Composer DAG is needed and can be used with Dataproc Workflow Template.

upvoted 1 times

dmzr 4 months, 3 weeks ago

Selected Answer: A

Ans A makes more sense, since a question is regarding Dataproc jobs only

upvoted 2 times

LP_PDE 4 months, 4 weeks ago

Correct answer is A. https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows

upvoted 3 times

HarshKothari21 5 months, 3 weeks ago

Selected Answer: C

Option c

upvoted 1 times

ducc 5 months, 4 weeks 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%)

musumusu 1 week, 3 days ago

key words here: job generators (pushlish message on pub/sub) and job runners(subscribe message for further analysis). You may add as much as pushlishing job and subscribing job to same topic. So Answer B.
Using API , app engine is also good approach but its more complex than pub/sub.

upvoted 1 times

jkhong 2 months, 3 weeks 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 (this can be solved by having multiple subscriptions, with each relating to a unique application) to a central topic

upvoted 4 times

AzureDP900 1 month, 4 weeks ago

Yes it is
B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them

upvoted 2 times

zellck 2 months, 4 weeks ago

Selected Answer: B

B is the answer.

upvoted 2 times

Atnafu 3 months, 2 weeks ago

A
Since it's application i will go with

upvoted 1 times

arpitagrawal 5 months, 3 weeks ago

Selected Answer: B

use pubsub

upvoted 3 times

YorelNation 5 months, 3 weeks ago

Selected Answer: B

I would tend to think B , one of the use of pub/sub is decoupling app

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: B

I choose B

upvoted 2 times

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

Community vote distribution

C (80%) B (20%)

jkhong 2 months, 1 week 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 1 times

odacir 2 months, 3 weeks ago

Selected Answer: C

A and D are not valid, because they monotonically increase.
C avoid hotspots for sure, but It's nor relate with querys. So for writing performance it's perfect that the reason for chose 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, its 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

upvoted 2 times

Remi2021 5 months, 2 weeks 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 4 times

AzureDP900 1 month, 4 weeks ago

Agree with C

upvoted 1 times

YorelNation 5 months, 3 weeks 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 5 months, 3 weeks 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 rate 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 2 years, 5 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

upvoted 22 times

AzureDP900 1 month, 4 weeks ago

D is right

upvoted 1 times

[Removed] Highly Voted 2 years, 11 months ago

Correct: D
https://cloud.google.com/iam/docs/roles-audit-logging#scenario_external_auditors

upvoted 12 times

Rajuuu 2 years, 7 months ago

The above link shows BigQuery as a sink for aggregated exports and not Cloud Storage.

upvoted 3 times

daghayeghi 1 year, 11 months ago

https://cloud.google.com/iam/docs/job-functions/auditing#scenario_operational_monitoring

upvoted 3 times

zellck Most Recent 2 months, 4 weeks ago

Selected Answer: D

D is the answer.

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 might aggregate and route audit log entries from all the folders contained by an organization to a Cloud Storage bucket.

upvoted 2 times

dffffff 10 months ago

D is correct

upvoted 1 times

MaxNRG 1 year, 1 month ago

Selected Answer: D

D: 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 1 year, 5 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
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 1 year, 7 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 needs 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 4 times

sumanshu 1 year, 7 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) - If I am not wrong... Thus it's INCORRECT
upvoted 2 times

at99 1 year, 1 month ago
Sinks are different from Aggregate Sinks, refer https://cloud.google.com/logging/docs/export/configure_export_v2#api
upvoted 1 times

septiandy 1 year, 10 months ago
what is the difference between C and D? I think it's same.
upvoted 3 times

haroldbenites 2 years, 6 months ago
D is correct
upvoted 3 times

saurabh1805 2 years, 6 months ago
D is correct answer, refer below link for more information.
upvoted 3 times

VishalB 2 years, 7 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] 2 years, 11 months ago
Answer D
upvoted 4 times

Question #142 *Topic 1*

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 (100%)

musumusu 1 week, 3 days 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 2 months, 1 week 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 2 times

dn_mohammed_data 5 months, 1 week ago
vote for B
upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago
B ,the another is related to the question as well.
https://cloud.google.com/bigquery/docs/reservations-monitoring#viewing-slot-usage
upvoted 3 times

John_Pongthorn 5 months, 2 weeks ago
Selected Answer: B
B the below is related to the question.
https://cloud.google.com/blog/topics/developers-practitioners/monitoring-bigquery-reservations-and-slot-utilization-information_schema
upvoted 4 times

AzureDP900 1 month, 4 weeks ago
B. Create a Cloud Monitoring dashboard based on the BigQuery metric slots/allocated_for_project
upvoted 1 times

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

odacir Highly Voted 2 months, 3 weeks 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 mayor changes. Stop and drain and then launch the new code is a lot is the 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

upvoted 8 times

maggieee 2 months, 2 weeks 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 1 times

musumusu Most Recent 1 week, 3 days 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 yourpipeline is running. Its safe any fast, you can continuously make some change 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 3 days, 11 hours ago

Answer D: as per latest documents 02/2023 google has removed update flag.
```

upvoted 1 times

jkhong 2 months, 1 week ago

Selected Answer: D

agree with odacir

upvoted 3 times

hauhau 2 months, 3 weeks 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 2 months, 4 weeks 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 2 months, 3 weeks ago

Are mayor changes. It's not safe to update. I vote D.

upvoted 1 times

Atnafu 3 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 3 months ago

Changed my mind to A
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python_2:~:text=Set%20the%20%2D%2Djob_name,%2D%2Dtransform_name_mapping%20option.

upvoted 1 times

drunk_goat82 3 months ago

Selected Answer: D

Changing windowing algorithm may break the pipeline.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

upvoted 3 times

ovokpus 3 months, 1 week ago

Selected Answer: A

No, do not drain the current job.

upvoted 1 times

dish11dish 3 months, 1 week ago

Selected Answer: D

in this scenario pipline is streaming pipline with windowing algorithm and triggering strategy changes to new one without loss of data,so better to 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

Referances:-
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 3 months, 1 week ago

If the pipeline was batch then ans would been A

upvoted 1 times

Mcloudgirl 3 months, 3 weeks 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

upvoted 2 times

cloudmon 3 months, 3 weeks ago

Selected Answer: A

It's A (https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#UpdateMechanics)
D would stop the pipeline, leading to loss of new data that would be sent into the pipeline before you start the new pipeline.

upvoted 2 times

Azlijaffar 4 months ago

The google documentation did mention that if the windowing or triggering algorithm changes, you might have unpredictable results. This question did mention change in windowing and triggering algorithm.
So A might lead to unpredictable results.
Safe bet is D.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

upvoted 2 times

devaid 4 months, 2 weeks ago

Selected Answer: D

D of course. Also you can only update for minor changes on windowing/triggering. Question say a different strategy.

upvoted 3 times

josrojgra 4 months, 2 weeks ago

Selected Answer: A

The answer is A.
The question says "update", and on the documentation (https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#UpdateMechanics) says explicitly on the second paragraph this:
The replacement job preserves any intermediate state data from the prior job, as well as any buffered data records or metadata currently "in-flight" from the prior job. For example, some records in your pipeline might be buffered while waiting for a window to resolve.

So the update process prevent data loss.

upvoted 4 times

Duckjai 4 months, 2 weeks ago

Agree. It is A

upvoted 1 times

dn_mohammed_data 5 months, 1 week ago

D : drain to ensure that there is no data loss during the code update

upvoted 2 times

MounicaN 5 months, 2 weeks ago

Selected Answer: D

drain for graceful shutdown

upvoted 2 times

sanjithj 5 months, 3 weeks ago

Selected Answer: D

Drain option is the correct one

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 ג€"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] Highly Voted 2 years, 11 months ago
Answer: A
Description: Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over 100TB
upvoted 18 times

[Removed] Highly Voted 2 years, 11 months ago
Correct - A
upvoted 9 times

musumusu Most Recent 1 week, 3 days 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 GCS and empty the appliance.
upvoted 1 times

AzureDP900 2 months ago
This is no brainer question, A is right
upvoted 1 times

zellck 2 months, 4 weeks 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 1 times

jkhong 4 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 2 times

sumanshu 1 year, 7 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 2 years, 5 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 2 years, 6 months ago

A is correct

upvoted 3 times

Rajuuu 2 years, 7 months ago

Correct - A

upvoted 3 times

Rajokkiyam 2 years, 11 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 2 years, 11 months ago
A should be the answer
upvoted 33 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: Dataprep is used by non developers
upvoted 17 times

Dhruv28 Most Recent 2 days 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

zellck 2 months, 4 weeks 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 2 times

arpitagrawal 5 months, 3 weeks ago
Selected Answer: A
non-developer analysts
upvoted 2 times

devdimidved 9 months ago
Selected Answer: A
Dataprep is for non developers
upvoted 1 times

amitsingla012 9 months, 2 weeks ago
Selected Answer: A
Option A -- Dataprep is the right answer
upvoted 1 times

Prasanna_kumar 1 year ago

Answer is A
upvoted 1 times

MaxNRG 1 year, 1 month ago
Selected Answer: A
A: https://cloud.google.com/dataprep/
upvoted 2 times

medeis_jar 1 year, 1 month ago
Selected Answer: A
Cloud Dataprep is a tool to do the job.
upvoted 1 times

JG123 1 year, 3 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 5 times

duytran_d 7 months ago
this comment is being repeated and i really appreciate this feeling :D
upvoted 1 times

sandipk91 1 year, 6 months ago
vote for option A
upvoted 4 times

sumanshu 1 year, 7 months ago
Vote for 'A', because of requirement - Enabling non-developer analysts to modify transformations
upvoted 5 times

haroldbenites 2 years, 6 months ago
A is correct
upvoted 3 times

SSV 2 years, 7 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

Archy 2 years, 7 months ago
spark is not graphical tool.
upvoted 5 times

atnafu2020 2 years, 6 months ago
A
you can use dataprep for continuously changing target schema
In general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema information can include:

Names of columns
Order of columns
Column data types
Data type format
Example rows of data
A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator (or schema tag) is displayed.
https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049
upvoted 8 times

Rajokkiyam 2 years, 11 months ago
Answer A
upvoted 5 times

[Removed] 2 years, 11 months ago
Should be A.
upvoted 3 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 do 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

[Removed] Highly Voted 2 years, 11 months ago
Should be B C
upvoted 17 times

Sid19 Highly Voted 1 year, 2 months 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.
Explaination
A and B cannot be true as gsutil can copy data to master node and the 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 connecter stores all the BQ data to GCS as temp and then processes it to HDFS. As data is already in GCS we donot need to load it to bq and use a connector then unloads it back to GCS and then processes it.
upvoted 10 times

[Removed] 1 year, 1 month ago
option D is mentioned in the new release 2019: https://cloud.google.com/blog/products/data-analytics/new-release-of-cloud-storage-connector-for-hadoop-improving-performance-throughput-and-more
upvoted 2 times

Deepakd 1 year ago
How can master node store data ? C is wrong.
upvoted 1 times

hauhau Most Recent 2 months, 3 weeks ago
Selected Answer: AD
C: master node doesn't make sense
upvoted 2 times

hauhau 2 months, 3 weeks ago
B: from the Cloud Storage bucket to any node of the Dataproc cluster
-> still on cloud not maxize the speed
upvoted 1 times

zellck 2 months, 4 weeks 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 2 times

tikki_boy 4 months ago

I'll go with DE

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: CD

CD is correct

upvoted 1 times

BigDataBB 1 year 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 masternode needs to be copied on datanode.
B for managed hive table option, C for external hive table

upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: CD

as explained by Sid19

upvoted 3 times

bkovari 1 year, 2 months 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 1 year, 2 months ago

Prefer A & D. typo, moderator can you fix it?

upvoted 2 times

BigQuery 1 year, 2 months ago

Guys Listen. Here's how you solve this problem. In Question "You need to replicate some data to the cluster's local Hadoop Distributed File System" So, Reading all options in process of elimination. A,B,C are out Since they transfer all ORC Files from GCS to DataProc. So D is best ans because its possible and data is in GCS. 2nd best is E. So, Ans is D and E. Upvote if you agree.

upvoted 5 times

Sid19 1 year, 2 months ago

Nope E is not the answer

upvoted 1 times

LORETOGOMEZ 1 year, 7 months ago

I vote for A, D

upvoted 5 times

daghayeghi 1 year, 11 months ago

You need to replicate some data to the cluster's local Hadoop Distributed File System
(HDFS) to maximize performance
This requirement looks like the key for this question. so, i think answer is B,C

upvoted 3 times

apnu 2 years, 1 month ago

D and E are externam table which will have performance issue and which wil not fullfill the reqt, so D and E are eliminated. in ABC , A is eliminated because we cannot copy data directly from GCS bucket to HDFS through utility. so A is also eliminated. so only B and C are correct answer.

upvoted 6 times

ashuchip 2 years, 2 months ago

DE seems to be correct, BC looks irrelevant

upvoted 3 times

Nams_139 2 years, 3 months ago

D,E As we don't want to replicate all files to hdfs - to the point.

upvoted 4 times

IT3008 2 years, 3 months ago

As per the condition given in question - "data has to be moved to hdfs from GCS" . Depending on that BC seems to be right choice. B for managed hive table option, C for external hive table .

upvoted 2 times

arghya13 2 years, 4 months ago

I think correct answer is DE..
Sourec:- https://docs.cloudera.com/HDPDocuments/DLM1/DLM-1.4.0/administration/content/dlm_replication_on_premise_to_gcs_hive.html

upvoted 1 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 (80%) A (20%)

mario_ordinola Highly Voted 1 year, 11 months ago
if someone are not sure that D is the answer, I suggest to don't take the exam
upvoted 30 times

madhu1171 Highly Voted 2 years, 11 months ago
D should be the answer
upvoted 23 times

AzureDP900 Most Recent 2 months ago
D is right
upvoted 2 times

zellck 2 months, 4 weeks 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 2 times

DataEngineer_WideOps 7 months, 1 week ago
Selected Answer: A
Cloud Composer for sure.
upvoted 1 times

GoReplyGCPExam 5 months ago
Composer is D
upvoted 1 times

nadavw 8 months, 3 weeks 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 8 months, 3 weeks ago
https://cloud.google.com/blog/topics/developers-practitioners/choosing-right-orchestrator-google-cloud
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: D
Cloud Composer
upvoted 2 times

JG123 1 year, 3 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 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 2 years, 2 months ago
D seems to be quiet relevant , because using composure you can do all things which are being asked to perform, even retry property is there in composure.
upvoted 3 times

Alasmindas 2 years, 3 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 2 years, 3 months ago
A seems to be right
upvoted 1 times

mumukshu 2 years, 2 months 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 1 year, 2 months ago
You forgot the "These jobs have many interdependent steps" which can be handled only though Composer
upvoted 1 times

Abby1356 2 years, 3 months ago
should be A
upvoted 1 times

arghya13 2 years, 4 months ago
Answer should be A..Cloud scheduler..cloud composer is an workflow manager. Can't run unix,bigquery jobs
upvoted 2 times

Tanmoyk 2 years, 5 months ago
D should be the best option
upvoted 3 times

haroldbenites 2 years, 6 months ago
D is correct
upvoted 3 times

lgdantas 2 years, 7 months ago
D!
"Cloud Scheduler is a fully managed enterprise-grade cron job scheduler"
https://cloud.google.com/scheduler
upvoted 6 times

Rajokkiyam 2 years, 11 months ago
Answer D
upvoted 6 times

Question #148

Topic 1

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants 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 (100%)

[Removed] Highly Voted 2 years, 11 months ago
Should be B.
upvoted 31 times

[Removed] Highly Voted 2 years, 11 months ago
AutoML is used to train model and do damage detection
Auto Vision is used is a pre trained model used to detect objects in images
upvoted 20 times

[Removed] 2 years, 11 months ago
Correct : B
upvoted 12 times

Deepakd 11 months, 2 weeks ago
Damage is an object in the image . So Auto Vision API can be used.
upvoted 1 times

AzureDP900 Most Recent 2 months ago
B is right
upvoted 1 times

Atnafu 3 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 1 times

cloudmon 3 months, 3 weeks 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 2 times

John_Pongthorn 5 months, 1 week ago
https://cloud.google.com/solutions/visual-inspection-ai#all-features
upvoted 2 times

cloudmon 3 months, 3 weeks 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 do AutoML. It looks like they assume that you have your own corpus of images.
upvoted 2 times

Deepakd 1 year 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 dont have data to train the model.

upvoted 1 times

Deepakd 11 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 1 year, 1 month ago
Selected Answer: B
AutoML is used to train model and do damage detection Auto Vision is used is a pre trained model used to detect objects in images
upvoted 2 times

medeis_jar 1 year, 1 month 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

JG123 1 year, 3 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 1 times

kino2020 2 years, 4 months ago
AutoML is right from the actual blogs I use.
The answer is B.
https://www.fourcast.io/blog/google-automl-how-i-built-my-own-image-classification-algorithm
upvoted 4 times

kavs 2 years, 3 months ago
B there is similar usecase in video for turbine image based fault detection..catch is if it is a pretraine or custom label...as damages has to be trained by us it has to be custom labels..on the other hand product search like the component usecase images are going to be same Pretrained hence vision api was right for the other question on 750 component..any thoughts
upvoted 3 times

nehaxlpb 2 years, 5 months ago
Answer is B.
Cloud Vision API detects lot of things for not damages. The description of Damages can be different for each business . So we need to train the model with test and training data to give our definition of Damages, so we need ML capabilities so answer is B, AutoML.
upvoted 7 times

atnafu2020 2 years, 6 months ago
C is the correct answer
upvoted 2 times

haroldbenites 2 years, 6 months ago
B is correct.
upvoted 6 times

clouditis 2 years, 6 months ago
C it is ! period :)
upvoted 3 times

Prakzz 2 years, 6 months ago
Should be B
upvoted 5 times

Archy 2 years, 7 months ago
AutoML is the way to go 'B'
upvoted 6 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 (100%)

someshsehgal Highly Voted 2 years 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 35 times

Shiv_am 1 year, 6 months ago
A is not at all possible
upvoted 1 times

squishy_fishy 1 year, 4 months ago
It is possible for about a year now. https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case
upvoted 6 times

jits1984 1 year, 4 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 4 times

alecuba16 7 months, 1 week 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

madhu1171 Highly Voted 2 years, 11 months ago
D should be the answer
upvoted 22 times

ducc 6 months ago
It is updated, now A is correct
upvoted 1 times

squishy_fishy 1 year, 4 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 2 times

musumusu Most Recent 1 week, 2 days 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 i would go for C. If it's Table level request only in question, then A is simple answer
upvoted 1 times

zellck 2 months, 4 weeks 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

upvoted 1 times

gudiking 3 months, 1 week ago

Selected Answer: A

A - table level access control now exists: https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case

upvoted 1 times

Transscend 3 months, 2 weeks ago

A.
Please see:
https://cloud.google.com/bigquery/docs/table-access-controls-intro#example_use_case

upvoted 1 times

John_Pongthorn 5 months, 1 week 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 obviousely.
Assume that User see certain table so he can see everything in such a table

upvoted 1 times

John_Pongthorn 5 months, 2 weeks 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 1 times

ducc 6 months ago

Selected Answer: A

Vote for A

upvoted 2 times

ducc 6 months ago

Selected Answer: A

Vote for A

upvoted 2 times

soichirokawa 6 months, 1 week ago

C can not be correct because: "authorized views should be created in a different dataset from the source data".

upvoted 1 times

Remi2021 7 months, 3 weeks ago

D should be the answer. When we set rights on the dataset level then all tables are affected. Imagine having hundreds of tables how difficult would it be to assign permissions directly.

upvoted 2 times

alecuba16 7 months, 1 week ago

Thinking about A and D, now I have realised that both A and D are the same more or less in terms of changes, but D has a big drawback, if the table changes, the view has to change.

upvoted 1 times

kped21 1 year ago

A - as per latest
BigQuery table ACL lets you set table-level permissions on resources like tables and views. Table-level permissions determine the users, groups, and service accounts that can access a table or view. You can give a user access to specific tables or views without giving the user access to the complete dataset. For example, grant the role BigQuery Data Viewer (roles/bigquery.dataViewer) to a user to let that user query just the table or view without full dataset access.

upvoted 3 times

medeis_jar 1 year, 1 month ago

Selected Answer: A

A, since we just deal with permission to data and not give access to a subset of data.

upvoted 3 times

kishanu 1 year, 2 months ago

I would go with D.
Though A seems to be feasible, there could be numerous tables under a dataset, and to provide access to the users could be a cumbersome process. Instead, create datasets for multiple groups(so-called teams), and assign access policies to them.

upvoted 3 times

AACHB 1 year, 2 months ago

Answer is A : it is now possible to assign roles at the table/view level

upvoted 1 times

Pupina 1 year, 2 months ago

Tables are in one dataset but the autorized view can be in a different new dataset. One for each user simplifying the number of binding policies used or all the views in one dataset. Both eschemas has pros and cons. Need to understand whtat this question is wanting us to answer. Cannot be A because it is not a best practice. I prefer D over than C because reduce administration but the on the other hand the views cannot be re-used.

upvoted 3 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 (100%)

[Removed] Highly Voted 2 years, 11 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 30 times

Rajokkiyam Highly Voted 2 years, 11 months ago
Answer B
Its google recommended approach to use LocalDisk/HDFS to store Intermediate result and use Cloud Storage for initial and final results.
upvoted 11 times

Chelseajcole 1 year, 5 months ago
Any link to support this recommended approach?
upvoted 1 times

zellck Most Recent 2 months, 4 weeks ago
Selected Answer: B
B is the answer.

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 2 times

John_Pongthorn 5 months, 1 week ago
Selected Answer: B
https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing_primary_disk_options
upvoted 1 times

rrr000 6 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 2 times

rrr000 6 months ago
Further the question states that original on prem machines has 100gb ram.
8-core nodes with 100-GB RAM
upvoted 1 times

SoerenE 1 year, 1 month ago
B should be the right answer: https://cloud.google.com/compute/docs/disks/performance#optimize_disk_performance
upvoted 1 times

medeis_jar 1 year, 1 month ago
Selected Answer: B
https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs
upvoted 3 times

JG123 1 year, 3 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 1 year, 11 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 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.
Its B
https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs
upvoted 3 times

ashuchip 2 years, 2 months ago
yes B is correct
upvoted 2 times

Alasmindas 2 years, 3 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

haroldbenites 2 years, 6 months ago
B is correct . I/O
upvoted 3 times

Archy 2 years, 7 months ago
B, this job is high on I/O, local HDFS on dish is best option
upvoted 4 times

sh2020 2 years, 7 months ago
How can adding Disk help with job becoming faster? Adding memory will facilitate in memory operations so that makes it faster. Adding CPU power will increase processing speed so that can make it faster. But I didnt quite follow how adding disk can make the process faster. I will vote for A
upvoted 2 times

sh2020 2 years, 7 months ago
Sorry a correction. I just re-read the question and answers. Option B seems to be correct for me
upvoted 2 times

FARR 2 years, 6 months ago
adding more disk space so that it doesn't runs out of space when intermediate data is kept on it
upvoted 2 times

PRC 2 years, 8 months ago
B is correct answer. For I/O intensive jobs, increasing the disk size resolves the issue. In case the issue is with compute capacity then increasing CPUs may help.
upvoted 4 times

[Removed] 2 years, 11 months ago
Should be A
upvoted 1 times

[Removed] 2 years, 11 months ago
Could be B - As the job is I/O intensive
upvoted 6 times

dambilwa 2 years, 8 months ago

Since the job is Disk I/O intensive, adding PDs would help.. If we add memory, we may have to re-write the job to use memory instead of PD

upvoted 2 times

rickywck 2 years, 11 months ago

After further study, I would pick A. According to Google, using persistent disk HDFS will not necessarily faster than using Cloud storage even though it gives more consistent performance.

upvoted 4 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 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 (100%)

zellck 2 months, 4 weeks ago

Selected Answer: C

C 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.

Integrated — Dataproc has built-in integration with other Google Cloud Platform services, such as BigQuery, Cloud Storage, Cloud Bigtable, Cloud Logging, and Cloud Monitoring, so you have more than just a Spark or Hadoop cluster—you have a complete data platform. For example, you can use Dataproc to effortlessly ETL terabytes of raw log data directly into BigQuery for business reporting.

upvoted 3 times

zellck 2 months, 3 weeks ago

https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery

You can use a BigQuery connector to enable programmatic read/write access to BigQuery. This is an ideal way to process data that is stored in BigQuery. Command-line access is not exposed. The BigQuery connector is a library that enables Spark and Hadoop applications to process data from BigQuery and write data to BigQuery using its native terminology.

upvoted 2 times

Atnafu 3 months, 1 week ago

C
Dataproc PySpark job, to invoke Spark ML functions

upvoted 1 times

gudiking 3 months, 1 week ago

Selected Answer: C

C - https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery

upvoted 1 times

dn_mohammed_data 5 months, 1 week ago

C : dataproc provide MLLIB for trainign ML models + dataproc can connect to bigquery

upvoted 2 times

[Removed] 5 months, 3 weeks ago

Selected Answer: C

Data proc has bigquery connector

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: C

C is correct

upvoted 2 times

Question #152 *Topic 1*

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 2 years, 11 months ago
Answer: A
Description: Geospatial and ML functionality is with bigquery
upvoted 19 times

[Removed] Highly Voted 2 years, 11 months ago
Answer : A
upvoted 15 times

musumusu Most Recent 1 week, 2 days ago
Answer B: BigTable,
Catchup words: Telemetry (sensor- semi structured data) as data is bigger than 500GB, datastore is not a good option.
GEOJSON , bigquery has geospatical capabilites but still not quick enough for semi structure geojson data.
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 senstive prediction.
upvoted 1 times

musumusu 3 days, 10 hours ago
answer A: You are just looking for a storage solution not a workflow
upvoted 1 times

zellck 2 months, 4 weeks 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 mailer for 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 1 times

Atnafu 3 months, 2 weeks 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 1 year, 3 months ago
Answer: C
upvoted 1 times

Chihhanyu 1 year, 3 months ago
GeoJson + Native functionality for prediction -> BigQuery
upvoted 2 times

singh_payal_1404 1 year, 3 months ago

Answer : A

upvoted 1 times

PM17 1 year, 4 months ago

This is more of a question that 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 2 times

haroldbenites 2 years, 6 months ago

A is correct

upvoted 3 times

FARR 2 years, 6 months ago

A
https://cloud.google.com/bigquery/docs/gis-intro

upvoted 3 times

Rajokkiyam 2 years, 11 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] Highly Voted 2 years, 11 months ago
Should be A
upvoted 24 times

[Removed] Highly Voted 2 years, 11 months ago
Correct: A

Dataflow can connect with Kafka and sliding window is used for taking averages
upvoted 15 times

zellck Most Recent 2 months, 4 weeks 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 by 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 3 times

medeis_jar 1 year, 1 month ago
Selected Answer: A
as explained by Alasmindas
upvoted 2 times

AACHB 1 year, 2 months ago
Selected Answer: A
Correct Answer: A
upvoted 2 times

JG123 1 year, 3 months ago
Correct: A
upvoted 1 times

Chelseajcole 1 year, 4 months ago
A is enough
upvoted 1 times

daghayeghi 2 years ago
A:
the correct answer is between A and B, But because used "Moving Average" then we should go for A.
upvoted 1 times

apnu 2 years, 1 month ago
yes , using KafkaIO , we can connect to Kafka cluster.
upvoted 2 times

ashuchip 2 years, 2 months ago
yes A is correct , because sliding window can only help here.
upvoted 3 times

Alasmindas 2 years, 3 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 5 times

Alasmindas 2 years, 3 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 4 times

atnafu2020 2 years, 6 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 2 years, 6 months ago
I don't think its A or B. Dataflow can't connect directly to kafka.
upvoted 1 times

FARR 2 years, 6 months ago
Yes, via KafkaIO. See the link in above comment
upvoted 3 times

kino2020 2 years, 7 months ago
"You operate an IoT pipeline built around Apache Kafka"
The statement in question states. Therefore, building with kafka is the requirement definition for this problem.

Just in case you are wondering, a case along with this problem is listed on google by the architects.
"Using Cloud Dataflow to Process Outside-Hosted Messages from Kafka"
https://cloud.google.com/solutions/processing-messages-from-kafka-hosted-outside-gcp

Therefore, A is the correct answer.
upvoted 4 times

SPutri 2 years, 3 months ago
the link that you share above is saying, "..illustrates a popular scenario: you use Dataflow to process the messages, where Kafka is hosted either on-premises or in another public cloud such as Amazon Web Services (AWS)."
but in this case, we are processing data coming from IoT pipeline, not from on-premise or other cloud. so, i don't think A is the proper solution. I consider option C instead.
upvoted 2 times

Rajuuu 2 years, 7 months ago
C and D is incorrect . It moves data from one queue(Kafka) to another queue (Pub/Sub).
A is Correct.
upvoted 3 times

Ral17 1 year, 5 months ago
Why not B?
upvoted 1 times

saurabhsingh92 2 years, 7 months ago
A, As we need to capture the moving average that's sliding window.
Option D also looks good but complex and costly, using dataflow with kfka IO a better option
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 (88%) 13%

madhu1171 Highly Voted 2 years, 11 months ago
A should be correct answer
upvoted 28 times

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

[Removed] Highly Voted 2 years, 11 months ago
Correct: A

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

musumusu Most Recent 1 week, 2 days ago
Answer A, key words to remember, High Scale use extra read replica. High availablity use extra failure replica. Both should be in different zone but in same region.
upvoted 2 times

desertlotus1211 1 month 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 1 times

desertlotus1211 1 month ago
The questions asks to ensure high availability in the event of a zone failure
upvoted 1 times

zellck 2 months, 4 weeks 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 and 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 4 months, 1 week 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

dn_mohammed_data 5 months, 1 week ago
vote for A
upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago

Selected Answer: B

B https://cloud.google.com/sql/docs/mysql/high-availability#read_replicas
This is the trick To find out the Answer without Much knowledge
If you try count the words Read replica asb compared to with faiuare repkica , The former is more than another one. It must be anwser.

upvoted 1 times

PhuocT 5 months, 4 weeks ago

Selected Answer: A

should be A

upvoted 1 times

Zorro27 7 months, 2 weeks ago

Selected Answer: A

Replicas can be created within the same region as the primary instance.
And Failover replica stored in other zone within the same region is used in case of any failure of the primary instance.

upvoted 1 times

watiobil 9 months, 2 weeks ago

Can't be B "Read replicas cannot be made highly available like primary instances. During a zonal outage, traffic to read replicas in that zone stops"
https://cloud.google.com/sql/docs/mysql/high-availability
A seems to be correct

upvoted 1 times

medeis_jar 1 year, 1 month ago

Selected Answer: A

https://cloud.google.com/sql/docs/mysql/high-availability
"Multiple zones (Highly available)
Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost."

upvoted 2 times

JG123 1 year, 3 months ago

Correct: A

upvoted 1 times

chpoola 1 year, 3 months ago

A should be the right answer even though it is legacy
I got confused with option : D , but D just stores the data as back up but not reroute the request

upvoted 1 times

hdmi_switch 1 year, 7 months ago

Seems to depend on the date the question was published.

A) is mentioned in legacy config "The legacy configuration for high availability used a failover replica instance."
https://cloud.google.com/sql/docs/mysql/configure-legacy-ha
B) Read replica is mentioned here https://cloud.google.com/sql/docs/mysql/high-availability scroll down to the diagrams
C) external read replica seems to be wrong, never heard of it
D) Aligns which Google steps, but automatic backup is not mentioned for HA https://cloud.google.com/sql/docs/mysql/configure-ha#ha-create just regional instance mentioned

I would pick D now, since regional instance is described as HA, although the automatic backup seems like a bonus.

upvoted 7 times

ralf_cc 1 year, 7 months ago

D - This is not RDS or Aurora, read replica cannot be used as a failover replica.

upvoted 3 times

baudrual 1 year, 11 months ago

Option A - regarding this : "The HA (High Availability) configuration, sometimes called a cluster, provides data redundancy. A Cloud SQL instance configured for HA is also called a regional instance and is located in a primary and secondary zone within the configured region. Within a regional instance, the configuration is made up of a primary instance and a standby instance." On that link : https://cloud.google.com/sql/docs/mysql/high-availability.

upvoted 4 times

Question #155 *Topic 1*

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%)

YorelNation Highly Voted 5 months, 3 weeks ago

Selected Answer: A

A I think it's the only technology that met the requirements

upvoted 7 times

dn_mohammed_data Highly Voted 5 months, 1 week ago

vote for A: topics, offsets --> apache kafka

upvoted 5 times

musumusu Most Recent 1 week, 2 days 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 1 topic.
Per key ordering. means message with same key can be process or assigned to a user in kafka.

upvoted 1 times

aquevedos91 5 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] Highly Voted 2 years, 11 months ago
Correct: A

Ask for cost effective so persistent disk are HDD which are cheaper in comparison to SSD.
upvoted 26 times

[Removed] Highly Voted 2 years, 11 months ago
Confused between A and B. For r/w intensive jobs need to use SSDs. But questions doesnt state anything about the nature of the jobs. So better to start with a default option.
Choose A
upvoted 14 times

baubaumiaomiao 1 year, 2 months ago
"You need to ensure that the deployment is as cost-effective as possible"
hence, no SSD unless stated otherwise
upvoted 2 times

zellck Most Recent 2 months, 4 weeks 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 5 months, 2 weeks ago
Selected Answer: A
it says cost effective , hence no SSD
upvoted 1 times

JG123 1 year, 3 months ago
Correct: A
upvoted 2 times

LORETOGOMEZ 1 year, 7 months ago
Correct : A
Option B is usefull if you use HDFS, and in this case as you use preemtible machines it isn't worth use SSD disks.
upvoted 2 times

ArunSingh1028 2 years ago
Answer - B
upvoted 1 times

StelSen 2 years, 1 month 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 on most cases. But after reading the whole page, they also mentioned

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 ready heavy process?).

upvoted 4 times

NM1212 6 months, 3 weeks ago

Caution about the link you provided as reference. It's intendedfor BigTable which is GC'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 2 years, 3 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 recommendations.

upvoted 4 times

beedle 2 years, 2 months ago

where is the proof...show me the link?

upvoted 2 times

Raangs 1 year, 12 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 4 times

Ravivarma4786 2 years, 6 months ago

Ans is B, for long running SDD suitable. HDD maintenance will be additional charge for long running jobs

upvoted 2 times

Rajuuu 2 years, 7 months ago

Answer is A…Cloud Dataproc for Managed Cloud native application and HDD for cost-effective solution.

upvoted 6 times

Rajokkiyam 2 years, 11 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%)

aadaisme Highly Voted 2 years, 8 months ago
Seems to be A. Preprocessing/scaling should be done with input features, instead of predictions (output)
upvoted 40 times

FARR Highly Voted 2 years, 6 months ago
A
Deep LEarning is not always the best solution
D talks about fudgin the output which is wrong
upvoted 11 times

musumusu Most Recent 1 week, 2 days 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 month ago
Selected Answer: A
a is the correct answer
upvoted 1 times

Dan137 3 months, 3 weeks ago
Also a good read is: https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview
upvoted 1 times

medeis_jar 1 year, 1 month 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 1 year, 3 months ago
Correct: A
upvoted 1 times

Spider7 1 year, 3 months ago
0.89 it's already not bad but by performing tuning rather then using the model default parameters there's a way to increase the overall model performance --> A.
upvoted 3 times

Spider7 1 year, 3 months ago
0.87 precisely
upvoted 1 times

hdmi_switch 1 year, 7 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 1 year, 11 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 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 2 years, 4 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 2 years, 4 months ago

Definitely not D
https://developers.google.com/machine-learning/crash-course/classification/check-your-understanding-roc-and-auc

upvoted 3 times

saurabh1805 2 years, 6 months ago

A for me, read below link for more details.

https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

upvoted 4 times

Callumr 2 years, 8 months ago

B: SVM is a single layer solution that tries to find a boundary between the two sets of data points. Deep neural networks can use many layers to define a more complex model to better fit the solution

upvoted 2 times

zonc 2 years, 7 months ago

Neural Network can be better, but not always. That isn't true that more complex architecture/method are always better, especially we can't say that when we know nothing about dataset. Generally SVM is a quite good ml method for many problems.

A is the correct answer.
Why B is a wrong one I've just described above.
C is wrong because we can't assume that our model will be work better with new data and generally models don't work better.
D doesn't make sense, it's not about preprocessing features (which is usefull), but only about a kind of postprocessing.

upvoted 3 times

[Removed] 2 years, 11 months ago

Answer: D
Description: Preprocessing of data needs to be done to improve AUC

upvoted 1 times

[Removed] 2 years, 11 months ago

Looks like D

upvoted 1 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] Highly Voted 2 years, 11 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 26 times

AzureDP900 1 month, 4 weeks ago

Thank you for detailed explanation. C is right

upvoted 1 times

rickywck Highly Voted 2 years, 11 months ago

Should be C:

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

upvoted 10 times

musumusu Most Recent 1 week, 2 days 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 1 times

zellck 2 months, 4 weeks ago

Selected Answer: C

C is the answer.

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 7 months, 1 week 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 1 year, 1 month ago

Selected Answer: C

https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only

upvoted 2 times

Prabusankar 1 year, 2 months 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 2 times

JG123 1 year, 3 months ago
Correct: C
upvoted 1 times

clouditis 2 years, 6 months ago
c it is!
upvoted 2 times

Rajokkiyam 2 years, 11 months ago
Should be C
upvoted 2 times

[Removed] 2 years, 11 months ago
Should be C
upvoted 2 times

jvg637 2 years, 11 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
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 (63%) C (38%)

[Removed] Highly Voted 2 years, 11 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 27 times

AzureDP900 1 month, 4 weeks ago
A. Cloud SQL There is no need of Spanner
upvoted 1 times

Rajuuu 2 years, 7 months ago
C:- Cloud SQL is not fully managed as that is one of the requirement.
upvoted 4 times

hightech 3 months ago
Cloud Sql is a fully managed service by Google
upvoted 2 times

zxing233 2 years, 5 months ago
https://cloud.google.com/sql it is fully managed
upvoted 2 times

dem2021 1 year, 9 months ago
Have you really worked on GCP?
upvoted 13 times

google_learner123 2 years, 6 months ago
C - CloudSQL does not scale automatically.
upvoted 6 times

zxing233 2 years, 5 months ago
Cloud SQL can automatically scale up storage capacity when you are near your limit
upvoted 3 times

dmzr 4 months, 3 weeks ago
it does not say about type of scaling, Cloud SQL scale up automatically with storage, that should works
upvoted 3 times

[Removed] Highly Voted 2 years, 11 months ago
Should be C.
upvoted 11 times

[Removed] 2 years, 11 months ago
May be A

upvoted 8 times

musumusu Most Recent 1 week, 2 days 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 1 times

certmonkey 1 week, 5 days ago

Selected Answer: A

Cloud SQL is enough here. Internet scale is not required. So Cloud spanner is incorrect.

upvoted 1 times

desertlotus1211 1 month ago

Answer is C: https://cloud.google.com/architecture/autoscaling-cloud-spanner

upvoted 2 times

hauhau 2 months, 3 weeks ago

Selected Answer: C

C because only cloud spanner fit requirement: Able to automatically scale up

upvoted 2 times

zellck 2 months, 4 weeks ago

Selected Answer: A

A is the answer.

upvoted 2 times

NicolasN 3 months, 1 week ago

Selected Answer: A

The whole discussion is around the bullet "Able to automatically scale up"
If this was referring explicitly to memory or CPU scaling. answer [A] would be incorrect.
But the subsequent bullet "Able to scale up to 6 TB" makes it clear (at least for me) that the former is referring to storage scaling.

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: A

It not required scale up horizontally and global end point
So A is correct

upvoted 2 times

BISeniorConsultant 11 months, 2 weeks ago

Selected Answer: A

A - Because you don't need to scale horizontally

upvoted 1 times

alecuba16 7 months, 1 week ago

Scale up -> more computation resources, cloud sql only scales in storage, not in performance/cpu

upvoted 2 times

rbeeraka 12 months ago

Selected Answer: A

Cloud SQL can scale up and fully managed DB service.

upvoted 2 times

medeis_jar 1 year, 1 month ago

Selected Answer: C

- Able to automatically scale-up
- Transactionally consistent

Only Spanner, Cloud SQL scale only storage automatically.

upvoted 4 times

JG123 1 year, 3 months ago

Correct: A

upvoted 3 times

alvinlxw 1 year, 3 months ago

Cloud SQL can easily scale up by adding up to 64 processor cores and 400+ GB of RAM to support up to 30TB of storage. What's more, Cloud SQL can automatically scale up your storage capacity when you get close to your limit.

upvoted 1 times

alecuba16 7 months, 1 week ago
What you are indicating is MANUAL scaling, the service has to be turn off for the scaling process.
upvoted 1 times

alecuba16 7 months, 1 week ago
"In order to change the machine type, you have to stop the Compute Engine instance and, after the change, restart it."

https://cloud.google.com/architecture/elastically-scaling-your-mysql-environment

So it is MANUAL scaling with the drawback of the service being interrupted.

Spanner doesn't require to turn off the service for scaling.
upvoted 2 times

MisuLava 4 months ago
you don't have to change the machine type. it simply scales up by itself increasing the storage. A is the correct answer
upvoted 2 times

gcp_k 1 year, 4 months ago
Answer could be : A

Both Cloud Spanner and Cloud SQL does not have auto scaling feature natively. You need to build automation around it based on metrics.

Both Cloud SQL and Cloud Spanner supports SQL.

With Cloud SQL, you can go up to 10 TB of storage which also satisfies the other requirement.

Consistency - Of course, with Cloud SQL, you have single master and read replicas. So technically the data will be consistent across all instances so to speak.

The reason why I didn't choose Spanner is - There is no requirement for HA DR, multi region, secondary indexes, etc.. So I choose "A"
upvoted 8 times

dotchi 1 year, 4 months ago
Cloud SQL
Cloud SQL is a good option when you need relational database capabilities but don’t need storage capacity over 10TB or more than 4000 concurrent connections. You also need to be skilled at on-premise management. Cloud Spanner
Cloud Spanner is a good option when you plan to use large amounts of data (more than 10TB) and need transactional consistency. It is also good if you want to use sharding for higher throughput and accessibility.If you know or think that you might eventually need to be able to horizontally scale your Google Cloud database, Cloud Scanner is a better option than Cloud SQL. If you start with Cloud SQL and need to eventually move to Cloud Spanner, be prepared to re-write your application in addition to migrating your database.
upvoted 2 times

triipinbee 1 year, 6 months ago
A: Cloud SQL can manage upto 10TB of data, the storage is automatically increase if you have selected the option and is fully managed.
upvoted 3 times

MisuLava 4 months ago
now, Cloud SQL can manage upto 64 TB.
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 (100%)

jvg637 Highly Voted 2 years, 11 months ago
A. Cloud SQL (30TB)
upvoted 30 times

vindahake 2 years, 9 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 2 months, 3 weeks ago
https://cloud.google.com/sql/docs/quotas#storage_limits
64TB
upvoted 2 times

Gcpyspark 2 years, 2 months 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 month ago
you can always call GCP to add quota.. .Spanner is for global reach, ideally...
upvoted 1 times

dagoat 1 year, 5 months ago
65 TB now in Sept 2021
upvoted 12 times

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

Rajuuu Highly Voted 2 years, 7 months ago
A as limit is now 30 TB for Cloud SQL
upvoted 6 times

musumusu Most Recent 1 week, 2 days 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 month, 4 weeks ago
A. Cloud SQL
upvoted 1 times

zellck 2 months, 4 weeks 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 3 times

Jay_Krish 3 months, 1 week 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 to see which is simple and doesn't come with extra frills.

upvoted 1 times

Atnafu 3 months, 2 weeks ago

A
65 TB now in Nov 2022

upvoted 2 times

WZH 7 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 A looks weird imho.

upvoted 1 times

Dan226 7 months, 1 week ago

Cloud SQL can store 64 Tb, but in the intial set up the operation are 20tb. It will reach the limitation soon if you choose Cloud SQL

upvoted 1 times

gcp_k 1 year, 4 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 1 year, 4 months ago

Hi everyone, Can i purchase this exam? is it worthable?

upvoted 2 times

Alasmindas 2 years, 3 months ago

Option A - Cloud SQL is the correct answer. Cloud SQL can store upto 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

SureshKotla 2 years, 5 months ago

SQL is sufficient

upvoted 2 times

Ravivarma4786 2 years, 6 months ago

Ans: A Cloud SQL limit increased to 30 TB

upvoted 3 times

Prakzz 2 years, 6 months ago

A Cloud SQL is correct as the requirement is to kinda lift n shift the db from on prem to cloud

upvoted 4 times

arnabbis4u 2 years, 9 months ago

Answer D : Datastore is a transactional database and can easily handle 20 tb. Supports ACID transactions.

upvoted 3 times

arnabbis4u 2 years, 9 months ago

A - After Further Study. Cloud SQL can now Scale upto 30 TB.

upvoted 8 times

Rajokkiyam 2 years, 11 months ago

Should be C.

upvoted 2 times

Rajokkiyam 2 years, 7 months ago

Limit for Cloud SQL changed from 10 to 30 TB. So answer A

upvoted 9 times

Question #161 *Topic 1*

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 (100%)

psu Highly Voted 2 years, 10 months ago

Answer C

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

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

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

upvoted 28 times

AzureDP900 1 month, 4 weeks 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 8 months, 3 weeks 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 2 years, 11 months ago

C correct answer

upvoted 19 times

zellck Most Recent 2 months, 4 weeks ago

Selected Answer: C

C is the answer.

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

upvoted 2 times

Remi2021 5 months, 3 weeks ago

Selected Answer: C

time series = narrow table

upvoted 1 times

_8008_ 10 months, 2 weeks 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 2 times

medeis_jar 1 year, 1 month 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 1 year, 3 months ago
Correct: C
upvoted 2 times

squishy_fishy 1 year, 4 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 1 year, 6 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 D is out of the picture.
upvoted 3 times

DeepakS227 1 year, 7 months ago
BQ is optimized for large-scale, ad-hoc SQL-based analysis. i Think it should be A
upvoted 1 times

koupayio 1 year, 8 months ago
First C & D won't cause hotspoting 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 1 year, 9 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 1 year, 8 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 1 times

Sumanth09 1 year, 11 months ago
Should be A
question did not talk about latency
without query cost -- BigQuery Cache
flexible schema - BigQuery (nested and repeated)
upvoted 5 times

VM_GCP 2 years, 2 months 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 2 years, 2 months ago
I will go for C
upvoted 2 times

Prakzz 2 years, 6 months ago
both C and D row key will cause hotspotting
upvoted 3 times

dg63 2 years, 7 months ago
I believe A. Use case requires support for Adhoc queries, and avoiding duplicate query costs.
upvoted 3 times

Callumr 2 years, 8 months ago
A - BigQuery caching will reduce query costs, and low latency of big table is not needed
upvoted 3 times

Question #162 *Topic 1*

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

dhs227 Highly Voted 2 years, 11 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 33 times

[Removed] 1 year, 1 month 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)
upvoted 2 times

mikey007 2 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 3 times

[Removed] Highly Voted 2 years, 11 months ago
Answer: A
Description: AAD is used to decrypt the data so better to keep it outside GCP for safety
upvoted 11 times

musumusu Most Recent 1 week, 2 days ago
answer A: KMS + AAD is more secure than CSEK
upvoted 1 times

zellck 2 months, 4 weeks 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 2 times

AzureDP900 1 month, 4 weeks ago
Agree with A
upvoted 1 times

Jay_Krish 3 months, 1 week ago

Selected Answer: D

CSEK with only security team having access seems to be right approach. Not sure how A can be better.

upvoted 1 times

cloudmon 3 months, 3 weeks 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

devaid 4 months, 3 weeks ago

Selected Answer: A

Answer: A

upvoted 1 times

clouditis 5 months, 1 week ago

D it is

upvoted 1 times

DataEngineer_WideOps 7 months, 1 week 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

A is the Answer.

upvoted 2 times

BigDataBB 1 year ago

Selected Answer: A

A because: "keep the AAD outside of Google Cloud"

upvoted 2 times

medeis_jar 1 year, 1 month ago

Selected Answer: D

As it is Google best practice.

upvoted 2 times

ramen_lover 1 year, 2 months ago

I don't see the correct answer.
D is a good practice; however, we have to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data.
If you upload the key in the cloud, the cloud provider staff eventually decrypt the file.
We have to keep the key anywhere else.
A prevents the cloud provider staff from decrypting the file but I don't think it's a good practice; use CMEK and keep your key on your side is a better practice.

upvoted 3 times

JG123 1 year, 3 months ago

Correct: D

upvoted 2 times

squishy_fishy 1 year, 4 months ago

The answer is A. In answer D, the security team can access the encryption key stored in the project, the requirement is trust no one.

upvoted 5 times

apnu 2 years, 1 month ago

here Symmetric key can never be the answer, because same key used for both encryption and decryption, which will be tough to hide because we have to share the information for data encryptiion. so A and B are eliminated. since in D , we are storing key in separate project , so it seems to be much secured way

upvoted 2 times

lna34fs 1 year ago

CSP can access the other project...

upvoted 1 times

federicohi 2 years, 2 months ago

People D its no possible because fo this "your data to prevent the cloud provider staff from decrypting your data." Any project in cloud give access to cloud staff

upvoted 2 times

SteelWarrior 2 years, 5 months ago

Should be A. As the question says Trust No One, this is the only approach where the additional authenticated data (AAD) is not available in GCP without which the files can not be decrypted.

upvoted 7 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 5 months, 2 weeks ago

Selected Answer: A

A . Your preference is to use managed products or features of the platform

upvoted 4 times

pluiedust 5 months, 3 weeks ago

Selected Answer: A

Should be A

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: A

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

upvoted 2 times

PhuocT 5 months, 4 weeks 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 (70%) D (30%)

cajica 3 weeks, 1 day ago

Selected Answer: D

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

upvoted 1 times

dconesoko 1 month, 3 weeks 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_ENCODER function. It just acts as an analytical funciton.

upvoted 2 times

zellck 2 months, 4 weeks ago

Selected Answer: B

B is the answer.

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 3 months ago

Selected Answer: B

The Cloud Data Fusion method will add unecessary 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

upvoted 4 times

Atnafu 3 months, 2 weeks 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 month, 3 weeks ago

Does it come with an out of the box one hot encoding template ?

upvoted 1 times

NicolasN 3 months, 3 weeks 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] obiously 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 searcing 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 2 times

cloudmon 3 months, 3 weeks 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 2 times

cloudmon 3 months, 3 weeks ago

D does have the ordinality problem, but at least it's actually something that's possible/relevant to do.

upvoted 1 times

Remi2021 5 months, 2 weeks ago

Selected Answer: B

key is in 'binary value'. It says low code not 'no code'. B is right one.

upvoted 3 times

pluiedust 5 months, 3 weeks ago

Selected Answer: B

I was also majoring in Econometrics and I think it is B

upvoted 2 times

damaldon 5 months, 3 weeks ago

Ans. B
https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing

upvoted 2 times

YorelNation 5 months, 3 weeks ago

Selected Answer: B

Coming for an econometric background i can assure you that for linear regression you can't encode a non ordinal feature as a sequence of number (awnser D).

The most appropriate answer is B, this is how to proceed with non-ordinal categorical features in a linear regression context.

upvoted 2 times

YorelNation 5 months, 3 weeks ago

And i discovered that BQ ML does the one-hot encoding for you:

https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing

upvoted 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: D

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.

upvoted 2 times

AmirNaik204 5 months, 3 weeks ago

D. The Key is "linear regression model". Linear regression model is used if you are predicting something numeric. So option is to write a no code transform that assign each city a number and use the number in the model. Predict the number and translate back to its corresponding city.

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Online points to D but not sure why B is incorrect.

upvoted 2 times

ducc 5 months, 4 weeks ago

It is least amount of coding, so D is correct.
Thank you for other explanations too.

upvoted 1 times

Question #165 *Topic 1*

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

devaid Highly Voted 4 months, 3 weeks ago

Selected Answer: B

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

upvoted 8 times

AzureDP900 1 month, 4 weeks ago

B is right

upvoted 1 times

Jay_Krish 3 months, 1 week ago

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

upvoted 7 times

musumusu Most Recent 3 days, 9 hours ago

Answer B:
locking read-write = for data accuracy
state read = for speed up or latency

upvoted 1 times

musumusu 1 week, 2 days 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 1 times

cajica 3 weeks, 1 day 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 can we 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 2 times

jkhong 2 months, 1 week 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 2 times

desertlotus1211 1 month ago

Argentina is South America...

upvoted 1 times

NicolasN 2 months, 3 weeks 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 preferrable 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, in a real case I would definitely prefer the advantages of Spanner.

upvoted 1 times

zellck 2 months, 4 weeks 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 3 months, 3 weeks 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 - Fulfills 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+ - Fulfills 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

NicolasN 2 months, 3 weeks 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

zellck 2 months, 3 weeks ago

North America has many regions, and the requirement is throughout North America, so Cloud Spanner will be more suitable to support multi-regions.

upvoted 2 times

NicolasN 2 months, 3 weeks ago

You are correct of course. The final sentence was totally inaccurate.

upvoted 1 times

cloudmon 3 months, 3 weeks 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 4 months, 1 week 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 no mention of a requirement for analytics

upvoted 1 times

MounicaN 5 months, 3 weeks ago

why not spanner?

upvoted 2 times

pluiedust 5 months, 3 weeks ago

Selected Answer: D

D is correct

upvoted 1 times

YorelNation 5 months, 3 weeks ago

Selected Answer: D

D no indication for multi region availability or very big volume of data. so no spanner.

upvoted 2 times

zellck 2 months, 3 weeks ago

North America has many regions, and the requirement is throughout North America, so Cloud Spanner will be more suitable to support multi-regions.

upvoted 2 times

Mcloudgirl 3 months, 2 weeks ago

+ CloudSQL is cheaper than Spanner

upvoted 2 times

nwk 5 months, 3 weeks ago

Why not B? There is no requirement for analysis and hence bq.
https://cloud.google.com/spanner/docs/transactions#read-write_transactions

upvoted 1 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: D

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

Cloud SQL or Cloud Spanner is required for ACID-based transactions in real-time. BigQuery cannot be used as table & row-level locking is required. BigQuery federated queries can directly access operational databased for ad-hoc analysis.

upvoted 2 times

adarifian 4 months, 2 weeks ago

there's no requirement for analysis, though

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: D

D is correct
Because ACID = (atomicity, consistency, isolation, and durability)

Sorry for other false explanation

upvoted 1 times

ducc 5 months, 4 weeks ago

C. Store transaction data in BigQuery. Disabled the query cache to ensure consistency.
It is storage system, not a processing system, so Bigquery is the choice
Here is my short-cut for choosing:
Storage system (or even "store transaction") -> Big Query
Process, consistency -> SQL (if global/ scale horizontally -> Spanner)

upvoted 1 times

Question #166 *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 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 (100%)

zellck 2 months, 4 weeks 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 2 times

AzureDP900 1 month, 4 weeks ago

Yes it is B. Implement clustering in BigQuery on the package-tracking ID column.

upvoted 1 times

cloudmon 3 months, 3 weeks ago

Selected Answer: B

B. https://cloud.google.com/bigquery/docs/clustered-tables

upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago

D is not correct becsuse This Is problem Is The Real Time so ingested date is the same as delivery date.

upvoted 1 times

kenanars 5 months, 3 weeks ago

why not D ?

upvoted 1 times

jkhong 2 months, 1 week ago

The table has already been partitioned

upvoted 1 times

John_Pongthorn 4 months, 3 weeks ago

There are several rows that represent movement of life cycle of 1 package-tracking ID
package delivery date = ingestion date , i suppose

upvoted 1 times

pluiedust 5 months, 3 weeks ago

Selected Answer: B

B;
As the table has already created with ingest-date partitioning.

upvoted 2 times

AWSandeep 5 months, 3 weeks ago

Selected Answer: B

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

upvoted 1 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

zellck Highly Voted 2 months, 4 weeks ago

Selected Answer: B

B is the answer.

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

https://cloud.google.com/bigquery/docs/migration/hive#data_migration
Migrating Hive data from your on-premises or other cloud-based source cluster to BigQuery has two steps:
1. Copying data from a source cluster to Cloud Storage
2. Loading data from Cloud Storage into BigQuery

upvoted 6 times

AzureDP900 1 month, 4 weeks ago

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

upvoted 1 times

John_Pongthorn Most Recent 4 months, 3 weeks ago

Selected Answer: B

B is most likely
1. migrate job and infrastructure to dataproc on clound
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 5 months, 2 weeks ago

Selected Answer: D

Answer D

upvoted 1 times

dn_mohammed_data 5 months, 1 week ago

you sould migrate spark to apache beam which is not the case here

upvoted 1 times

TNT87 5 months ago

apache beam for what???

upvoted 1 times

adarifian 4 months, 2 weeks ago

dataflow uses apache beam

upvoted 1 times

TNT87 1 month, 2 weeks ago

@adarifian Why use apache beam yet there is Dataflow an inhouse gcp solution to solve the problem? hence i said apache beam for what

upvoted 1 times

GyaneswarPanigrahi 5 months, 2 weeks 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 2 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 2 times

TNT87 5 months, 3 weeks ago

Selected Answer: B

Ans B
-cost saving
-time factor
-Spark -Data proc

upvoted 2 times

TNT87 5 months, 3 weeks 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 5 months, 3 weeks 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 5 months, 3 weeks 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 5 months, 3 weeks ago

It is B or D, still confusing

upvoted 2 times

AWSandeep 5 months, 4 weeks 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

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

AWSandeep Highly Voted 5 months, 4 weeks 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 6 times

jonathanthezombieboy Most Recent 2 days, 9 hours ago

Selected Answer: B

Option B is a better solution because it uses BigQuery column-level security to restrict access to the SSN column. This allows customer service representatives to view the original SSN when necessary, while still preventing unauthorized users from accessing the sensitive data. Additionally, this solution does not require any additional steps like encryption or data masking, which can simplify the data pipeline and reduce security risks.

upvoted 1 times

musumusu 1 week, 2 days ago

Answer B:
c&D will permanently change the value in the column, but they govt wants to see original value whenver they want

upvoted 1 times

kostol 1 week, 4 days ago

Selected Answer: D

D - the requirement is to redact but B only restricts the access

upvoted 1 times

donbigi 3 weeks, 2 days ago

Selected Answer: D

the requirement to allow customer service representatives to view the original data is the reason why option B (using BigQuery column-level security) is not the best solution. This method would only restrict access to the data, but not allow customer service representatives to view the original values when necessary. Option D (using Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token) would provide the necessary level of redaction while still allowing authorized personnel to access the original data when needed.

upvoted 1 times

zanhsieh 3 weeks, 4 days ago

Selected Answer: D

D. The question does not say SSN data as a column, which might be a string just contains SSN. In this case, option A and B should be out since both them assumed SSN must be a column. Option C is out since hash function is one-way and can't revert back. So option D is the answer. There are couple similar questions in Security Engineer exam for this.

upvoted 2 times

jkhong 2 months, 1 week ago

Selected Answer: B

Problem: Handle PII data Considerations: Redact PII data, but still allow re-identification
C -> this hashing does not ensure reidentification
A, D -> this requires that the token and keys to be managed separately. I would be paranoid of staff disclosing these tokens and keys. A more secure way would be to rely on IAM.

We can mask data using column level security

- https://cloud.google.com/bigquery/docs/column-data-masking-intro
- only authorised personnel would be able to access

I'd be more comfortable with B since it relies on IAM roles, rather than having the potential leak of token and keys.

upvoted 2 times

Atnafu 2 months, 1 week ago

B is answer
redact- hiding the column i will use bq ACL
mask- i use DLP

upvoted 2 times

NicolasN 2 months, 2 weeks ago

Selected Answer: A

It's [A]!
After thinking a lot about this question, I see that [A] is the most straightforward solution to satisfy both requirements, data redaction and ability to view original data when needed.

[B] works but it doesn't cover the "when necessary" condition. Customer service representatives will always have access to PII data.
[C] is rejected as it is a non reversible de-identification method.
[D] doesn't describe the re-identification process for customer service representatives and what special rights should be granted to them.

The following official document describes the same solution with [A] but uses Cloud KMS instead of a separate table for keys:
https://cloud.google.com/bigquery/docs/column-key-encrypt

upvoted 2 times

NicolasN 2 months, 2 weeks ago

In case of any doubts, what [A] proposes can be easily done with the following steps:
1. Load customer data in a table, let's say CUSTOMER
2. Create keys for every customer id with KEYS.NEW_KEYSET, and store the key and the Customer Id in a new table, named CUST_KEYS
3. Join the two tables on Customer Id and update the PII field by calling AEAD.ENCRYPT(CUST_KEYS.KEY, ...)
4. Limit access to CUST_KEYS table only to customer service representatives.
5. Whenever needed, a representative will have access to raw info by calling AEAD.DECRYPT_STRING on a query joining the two tables

upvoted 1 times

hauhau 2 months, 3 weeks ago

Selected Answer: B

B just hide but not redact
D redact but cheange the original value
there is no correct answer

upvoted 1 times

zellck 2 months, 3 weeks ago

for D, encrypted values can be re-identified using the original cryptographic key.

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 1 times

zellck 2 months, 4 weeks ago

Selected Answer: D

D is the answer.

https://cloud.google.com/dlp/docs/pseudonymization#supported-methods
Cloud DLP supports three pseudonymization techniques, all of which use cryptographic keys. Following are the available 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 2 times

Wonka87 2 months ago

why would you not consider B which is simpler than maintaining keys with access restriction and then decrypting it for only authorized people. It can be achieved using B also and redact doesn't necessarily mean changing / removing values, restricting to view is also redacting in my opinion. Kindly suggest.

upvoted 1 times

NicolasN 3 months, 3 weeks ago

B seems indeed to be a correct answer. Documentation (https://cloud.google.com/bigquery/docs/column-level-security-intro) is referring indirectly to a similar case: "For example, a policy can enforce access checks such as: You must be in group:high-access to see the columns containing TYPE_SSN"
But I can't argue why answer A couldn't fit except for being more complex.
If anyone knows please provide your arguments.

upvoted 1 times

NicolasN 2 months, 2 weeks ago
My voting answer explains why finally I prefer [A]
upvoted 1 times

devaid 4 months, 2 weeks ago
Selected Answer: D
Answer: D
upvoted 1 times

clouditis 5 months, 1 week ago
Its D, B does not make sense, they are asking to redact, not hide it away completely
upvoted 4 times

devaid 4 months, 3 weeks ago
Agree, the key here is "to see the original values", that means it already shows a protected value
upvoted 1 times

him98934 5 months, 2 weeks ago
I will go with D, https://cloud.google.com/bigquery/docs/column-level-security-intro
upvoted 2 times

TNT87 5 months, 3 weeks ago
Selected Answer: B
Ans B
https://cloud.google.com/bigquery/docs/column-level-security
Remember the custpomer service group has to view so thry need access from time to time
upvoted 1 times

PhuocT 5 months, 3 weeks ago
Selected Answer: D
I will go with D, https://cloud.google.com/bigquery/docs/column-level-security-intro
upvoted 3 times

PhuocT 5 months, 3 weeks ago
mistake, I did mean B
upvoted 3 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 (92%) 8%

AWSandeep Highly Voted 5 months, 3 weeks ago

Selected Answer: A

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

upvoted 5 times

Prakzz Most Recent 2 months, 1 week ago

Selected Answer: B

Should be B
The clustering should be according to the filtering needs

upvoted 1 times

Atnafu 2 months, 1 week 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 4 times

zellck 2 months, 4 weeks 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 5 months, 3 weeks ago

https://cloud.google.com/bigquery/docs/querying-clustered-tables

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: A

A
The question mention that the query is 30 days recently

upvoted 3 times

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 (100%)

AWSandeep Highly Voted 5 months, 4 weeks 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

upvoted 5 times

jkhong 2 months, 3 weeks ago

Don't think we need to configure subscription to retain ack messages. It is defaulted to retain for 7 days

upvoted 1 times

musumusu Most Recent 1 week, 2 days 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

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 1 times

zellck 2 months, 4 weeks 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 3 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.

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

upvoted 1 times

TNT87 5 months, 3 weeks ago

Selected Answer: B

Answer B

upvoted 1 times

PhuocT 5 months, 3 weeks ago

Selected Answer: B

should be B.

upvoted 1 times

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

Kvk117 1 month, 1 week ago

Selected Answer: A

A is the correct answer

upvoted 1 times

jkhong 2 months, 1 week 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 1 times

zellck 3 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 3 times

GCPSharon 4 months ago

Selected Answer: C

Stew prediction time by remove the preprocessing!

upvoted 1 times

TNT87 5 months, 3 weeks ago

Selected Answer: A

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

upvoted 4 times

ducc 5 months, 4 weeks 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

AWSandeep 5 months, 4 weeks 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 4 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 5 months, 4 weeks 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 6 times

zellck Most Recent 3 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 5 months, 3 weeks ago

Selected Answer: D

Moving average ——> sliding window

upvoted 2 times

Question #173 *Topic 1*

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%)

musumusu 3 days, 9 hours ago
why not c ? as data is arriving hourly why we can use batch processing rather than streaming with 1 hour fixed window?
upvoted 1 times

Atnafu 2 months, 1 week ago
D

TUMBLE=> fixed windows.
HOP=> sliding windows.
SESSION=> session windows.
upvoted 4 times

zellck 3 months ago
Selected Answer: D
D is the answer.

https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#tumbling-windows
upvoted 2 times

devaid 4 months, 2 weeks ago
Selected Answer: D
Answer D
Tumbling Windows = Fixed Windows
upvoted 2 times

TNT87 5 months, 1 week ago
Selected Answer: D
Answer D
upvoted 2 times

AWSandeep 5 months, 4 weeks 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%)

zellck 3 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 2 times

SMASL 5 months, 1 week 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 initally automate the 70 % of the requests that are simpler before automating the more complicated ones.

upvoted 4 times

AWSandeep 5 months, 3 weeks 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 (86%) 14%

waiebdi 2 weeks, 1 day 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 1 times

pcadolini 2 months, 2 weeks 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. It 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 2 times

zellck 3 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 2 times

NicolasN 3 months, 3 weeks 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 :
"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 2 times

josrojgra 4 months, 1 week 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

devaid 4 months, 3 weeks 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 5 months, 2 weeks ago

I think Denormalizing here means ,using big queries native data representation and that is using nested and repeated columns .Thats is the best practice in GCP
https://cloud.google.com/bigquery/docs/nested-repeated#example

upvoted 1 times

[Removed] 5 months, 3 weeks ago

Selected Answer: D

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

Reall01 5 months, 1 week 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 all depends on what context you take things in as "denormalization"

upvoted 2 times

NicolasN 3 months, 3 weeks 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 :
"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 5 months, 4 weeks ago

D. Partition the data by transaction date.

Star schema is already denormalized.

upvoted 3 times

PhuocT 5 months, 4 weeks 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%)

AzureDP900 1 month, 3 weeks 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 3 months ago

Selected Answer: C

C is the answer.

upvoted 3 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: C

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

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

upvoted 4 times

PhuocT 5 months, 4 weeks ago

Selected Answer: C

C is correct

upvoted 2 times

Question #177 *Topic 1*

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

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

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

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

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

Correct Answer: *D*

Community vote distribution

devaid Highly Voted 4 months, 3 weeks 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 5 times

musumusu Most Recent 1 week, 2 days ago

Answer C: need to setup SQL based job means transformation in not very complex. And Biqquery 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 month 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 B "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 the 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

Atnafu 3 months ago

C
D-is incorrect because you are rebuild your batch pipeline for structured data on Google Cloud.

upvoted 1 times

Atnafu 3 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 5 months, 1 week 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 5 months, 2 weeks 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 1 times

John_Pongthorn 5 months, 2 weeks ago

C is the most likely , bigquery is severless and sql
D is dataflow severless but it is wrong at using python sdk but using sql beam rthen it will be correct

upvoted 1 times

TNT87 5 months, 3 weeks ago
Answer C
upvoted 2 times

ducc 5 months, 4 weeks ago
Selected Answer: A
A
- You have to maintain PySpark Code -> Proc
upvoted 1 times

ducc 5 months, 4 weeks ago
After thinking a while, I think the question is not clear enough. To be honest
upvoted 1 times

ducc 5 months, 4 weeks ago
A or C. I go for C because they said they want to use SQL syntax...
upvoted 1 times

AWSandeep 5 months, 4 weeks 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 1 times

ducc 5 months, 3 weeks ago
The question said "use SQL syntax"
C might still correct
upvoted 1 times

AWSandeep 5 months, 4 weeks 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 3 months ago
Dataproc is diffrent than Dataproc Serveless. 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 2 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 (79%) 13% 8%

John_Pongthorn Highly Voted 5 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 11 times

musumusu Most Recent 1 week, 2 days ago

Answer: B,
reducing the batch size improve the speed performance also improve cpu utilisation.
Dead letter queues are generated for messages that are errorrly acknowledged and its 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 its file type issue in compression, or batch size.

upvoted 1 times

Atnafu 2 months, 1 week ago

D
Flatten will just merge all results into a single PCollection. To join them you can use CoGroupByKey

upvoted 1 times

zellck 3 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
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 (grouping) 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 4 times

TNT87 4 months, 3 weeks 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 1 times

devaid 4 months, 2 weeks ago

that is for Big Query isn't?

upvoted 1 times

TNT87 1 month, 2 weeks ago
datflow can use avro format sir. streaming or batching to bigquery in avro format it can
upvoted 1 times

TNT87 4 months, 3 weeks ago
https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion
upvoted 1 times

devaid 4 months, 3 weeks 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 5 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

csd1fggfhfgvh234 5 months, 1 week ago
A switching to avro. No serialisation
upvoted 1 times

TNT87 5 months, 3 weeks ago
Switch to Avro format
Answer A
upvoted 2 times

TNT87 5 months, 1 week ago
https://docs.confluent.io/platform/current/schema-registry/serdes-develop/serdes-avro.html
upvoted 1 times

YorelNation 5 months, 3 weeks 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 5 months, 4 weeks ago
Selected Answer: B
B. Reduce the batch size.
upvoted 3 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 button.

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

musumusu 1 week, 2 days 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 month, 1 week ago
Selected Answer: D
The answer is D
upvoted 1 times

nkit 2 months ago
Selected Answer: D
I believe "Format preserving token" in option D makes it easier choice for me
upvoted 1 times

PrashantGupta1616 2 months ago
Selected Answer: D
D looks right
upvoted 1 times

jkhong 2 months, 1 week 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 1 times

Atnafu 2 months, 1 week 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
Part1- 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 1 times

zellck 3 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 3 times

dish11dish 3 months, 1 week ago

Selected Answer: D

Though both option A nad D maintains referential integrity,question is why you wnat to keep untokenize data in GCS,best way is option D which even support 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 3 months, 3 weeks 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 2 times

cloudmon 3 months, 3 weeks ago

And, answer A does not include format preservation, which would lose referential integrity.

upvoted 1 times

NicolasN 3 months, 3 weeks ago

I think that this isn't true.
Look at the table 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 3 months, 3 weeks 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 locked-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

NicolasN 3 months, 3 weeks ago

The table in https://cloud.google.com/dlp/docs/transformations-reference#transformation_methods shows which transformation preserves referential integrity

upvoted 1 times

PoCk3T 4 months, 1 week ago

All answers are obsolete since https://cloud.google.com/bigquery/docs/column-data-masking-intro

upvoted 2 times

devaid 4 months, 2 weeks ago

Selected Answer: D

Answer D makes more sense to me.
Source: https://cloud.google.com/dlp/docs/pseudonymization

upvoted 3 times

Ray0506 5 months, 2 weeks ago

Selected Answer: D

D is correct

Refer: https://cloud.google.com/dlp/docs/transformations-reference#fpe

upvoted 1 times

Thobm 5 months, 2 weeks ago

Selected Answer: D

Format-preserving token makes more sense, no bucket needed

upvoted 1 times

Thobm 5 months, 2 weeks ago

Never mind, indeed A for the data that's still in GCS

upvoted 1 times

learner2610 5 months, 2 weeks ago
Why isn't it D ? we have to maintain referential integrity so isnt it D ?
upvoted 1 times

ducc 5 months, 4 weeks ago
https://cloud.google.com/dlp/docs/pseudonymization#fpe-ffx
upvoted 1 times

ducc 5 months, 4 weeks ago
Selected Answer: A
AWSandeep is correct
upvoted 3 times

Ray0506 5 months, 2 weeks ago
can you specify why he is correct? why A and not D?
upvoted 1 times

AWSandeep 5 months, 4 weeks ago
Selected Answer: A
A. Create a pseudonym by replacing the PII data with cryptogenic tokens, and store the non-tokenized data in a locked-down button.
upvoted 4 times

learner2610 5 months, 2 weeks ago
Any reason why it is A and not D ?
upvoted 1 times

Atnafu 3 months ago
A
replacing with cryptographic token not with format preserve stuff
https://cloud.google.com/dlp/docs/pseudonymization#:~:text=with%20its%20corresponding%20token
upvoted 1 times

Atnafu 2 months, 1 week ago
Changed my mind - This should be D
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 1 week, 2 days ago

C
if data is time based or sequential, find partition and cluster option
if data is not time based,
always look for denomalize / nesting option.

upvoted 1 times

AzureDP900 1 month, 3 weeks ago

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

upvoted 1 times

zellck 3 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 3 times

Atnafu 3 months ago

C
Best practice: Use nested and repeated fields to denormalize data storage and increase query performance.

upvoted 1 times

dish11dish 3 months, 1 week 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 1 times

Thobm 5 months, 2 weeks ago

Selected Answer: C

https://cloud.google.com/bigquery/docs/best-practices-performance-nested

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: C

C is correct

upvoted 2 times

AWSandeep 5 months, 4 weeks 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 (92%) 8%

John_Pongthorn Highly Voted 5 months ago

Selected Answer: D

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

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

upvoted 9 times

NicolasN 3 months, 3 weeks ago

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

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

Atnafu 3 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 3 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

musumusu Most Recent 1 week, 2 days 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),
messege 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 month 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 is 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 2 times

zellck 3 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 month, 3 weeks ago

D. Batch the job into ten-second increments.

upvoted 1 times

Atnafu 3 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 5 months ago

Ans C: reference https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls

upvoted 1 times

John_Pongthorn 5 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 norma case but if you need to stand for backpressure,
in last sector 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 4 times

SMASL 5 months, 1 week 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 1 times

Thobm 5 months, 2 weeks 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 5 months, 2 weeks ago

C , It is straight foward , 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 enrichmen

upvoted 1 times

TNT87 5 months, 3 weeks 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 5 months, 3 weeks ago
https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls
To support choice C
upvoted 1 times

YorelNation 5 months, 3 weeks ago
Selected Answer: C
i think you are right gg
upvoted 1 times

nwk 5 months, 3 weeks ago
How about C?
https://cloud.google.com/architecture/e-commerce/patterns/batching-external-calls
upvoted 2 times

AWSandeep 5 months, 3 weeks ago
Selected Answer: D
D. Batch the job into ten-second increments.
upvoted 2 times

ducc 5 months, 3 weeks ago
Selected Answer: D
D is correct IMO
upvoted 1 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 3 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
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 4 times

musumusu 1 week, 2 days ago

what is wrong with A, there is no cost constraint

upvoted 1 times

AzureDP900 1 month, 3 weeks ago

Agreed

thx for sharing link

upvoted 1 times

TNT87 5 months, 3 weeks ago

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil_for_smaller_transfers_of_on-premises_data
Answer C

upvoted 3 times

YorelNation 5 months, 3 weeks ago

Selected Answer: C

C seems legit

upvoted 3 times

AWSandeep 5 months, 4 weeks 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 4 times

Question #183 *Topic 1*

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

A (54%) B (43%) 4%

musumusu 3 days, 8 hours ago

Answer A:
I checked on other website and chatgpt also suggested A, to use index per stock and timestamp.

upvoted 1 times

niketd 5 days, 3 hours 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 1 times

kostol 1 week, 2 days ago

Selected Answer: A

https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid

upvoted 1 times

desertlotus1211 4 weeks, 1 day 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

nkit 2 months ago

Selected Answer: A

Reversing TS won't help

upvoted 2 times

PrashantGupta1616 2 months ago

Selected Answer: A

I go for A.

upvoted 2 times

kelvintoys93 2 months, 1 week ago

How does reversing timestamp help in query speed? its still gonna scan through the whole table to output the result.

According to Google,
"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."

Therefore I go for A.

upvoted 2 times

zellck 3 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 4 times

Atnafu 3 months ago
A
Row keys to avoid
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.
https://cloud.google.com/bigtable/docs/schema-design#row-keys-avoid
upvoted 1 times

Atnafu 3 months ago
plus
If you usually retrieve the most recent records first, you can use a reversed timestamp that is not row key design
Reverse is with domain on schema design
If you're storing data about entities that can be represented as domain names, consider using a reverse domain name (for example, com.company.product) as the row key.
upvoted 1 times

jkhong 3 months, 2 weeks ago
C, D -> Bigtable has a limitation of 1000 tables per instances. Since we are concerned of stock data, there has bound to be thousands... It is also an anti pattern to store similar schemas in separate tables.
https://cloud.google.com/bigtable/docs/schema-design#tables

Don't know why is everybody harping on the fact that timestamp or reverse timestamp is the only row key, or that they are the primary key. Im interpreting it as having timestamp as part of the design. We'll need reverse timestamp since we are only concerned with recent data so B, with only 1 unique table for all stock indexes
upvoted 2 times

Teraflow 3 months, 3 weeks ago
Selected Answer: A
Indexes need to be unique, reverse timestamp is not necessarily unique
upvoted 3 times

kastuarr 4 months, 1 week ago
Selected Answer: D
D will avoid hot spotting , allow a very fast query and the reverse time stamp will allow recent records to be retrieved easily
upvoted 1 times

devaid 4 months, 3 weeks ago
B. reverse timestamp rowkey = most newer registry
upvoted 1 times

devaid 4 months, 2 weeks ago
2nd tought: I'm agree with B, but the stock symbol as part of the row key is needed, like: "stock#reverseTimeStamp". So, in that case makes more sense D because it create one table for every index, but that is not efficient on Spanner so i keep with B ...
upvoted 2 times

jkhong 3 months, 2 weeks ago
There's no mention of Spanner here...
upvoted 1 times

Amey_Yeole 5 months ago
B is correct. By reversing the timestamp, you can design a row key where the most recent event appears at the start of the table instead of the end. As a result, you can get the N most recent events simply by retrieving the first N rows of the table.
upvoted 4 times

John_Pongthorn 5 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 idea, We will go with B, C
https://cloud.google.com/bigtable/docs/schema-design#time-based

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

If you usually retrieve the most recent records first, you can use a reversed timestamp in the row key by subtracting the timestamp from your programming language's maximum value for long integers (in Java, java.lang.Long.MAX_VALUE). With a reversed timestamp, the records will be ordered from most recent to least recent.
upvoted 4 times

Mcloudgirl 3 months, 2 weeks 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

Remi2021 5 months, 2 weeks ago
Selected Answer: B
reversing the timestamp
upvoted 3 times

DiegoGonL 5 months, 2 weeks ago

Selected Answer: B

I think its B because it ask that we need to ensure that you can access the data with the simplest query, not to create the rowkey as the best practices (which should include the index first and then the timestamp)

upvoted 2 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

nwk Highly Voted 5 months, 3 weeks 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 8 times

jkhong 2 months, 1 week 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 4 months, 3 weeks ago

A and B are discarded because the UPDATE statement, is not performance efficient. Neither appending more and more values to the stagging table. It's better cleaning the stagging table, and merging with the master dataset.

upvoted 2 times

YorelNation 5 months, 3 weeks ago

They don't seems concerned too much by data accuracy in the question

upvoted 1 times

NicolasN Highly Voted 3 months, 3 weeks 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 7 times

NicolasN 3 months, 3 weeks 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 7 times

AzureDP900 1 month, 3 weeks ago

Agreed C is right

upvoted 1 times

hauhau Most Recent 2 months, 3 weeks 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 1 times

hauhau 2 months, 3 weeks ago

B just say "update", not specificlly mention DML. update can be merge

upvoted 1 times

zellck 3 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 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 5 times

Atnafu 3 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 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.
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 4 months, 3 weeks 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 4 months, 3 weeks ago

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 5 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

upvoted 1 times

TNT87 5 months, 3 weeks ago

Selected Answer: D

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.

upvoted 2 times

TNT87 5 months, 3 weeks ago

Ans D
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.

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: D

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.

upvoted 2 times

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%)

zellck 3 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
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 3 times

Atnafu 3 months ago

C
Code error

upvoted 1 times

pluiedust 5 months, 2 weeks ago

Selected Answer: B

B is correct

upvoted 2 times

ducc 5 months, 3 weeks ago

Selected Answer: B

B is correct

upvoted 1 times

AWSandeep 5 months, 4 weeks 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 4 times

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 (73%) D (27%)

maci_f 1 month 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 logs (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 1 times

zellck 2 months, 1 week ago

Selected Answer: A

A is the answer.

upvoted 1 times

Atnafu 3 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 2 months, 1 week 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 3 months, 1 week ago

Selected Answer: A

A because you filter data daily by view not just once by cloud logging

upvoted 1 times

devaid 4 months, 3 weeks 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 loggs in Big Query or other.

upvoted 3 times

devaid 4 months, 1 week 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 obsolet. 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 3 months, 1 week 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

upvoted 1 times

AHUI 5 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 5 months, 2 weeks ago

Ans is C

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

upvoted 1 times

jkhong 3 months, 2 weeks 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 month, 2 weeks ago

so wjhaytsa your argument about daily generartion of data??

upvoted 1 times

TNT87 5 months, 2 weeks ago

On the other hand Ans A makes sense
https://cloud.google.com/logging/docs/export/bigquery#overview

upvoted 1 times

changsu 5 months, 2 weeks 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

pluiedust 5 months, 2 weeks ago

Why not C?

upvoted 1 times

Wasss123 5 months, 2 weeks ago

why not D ?

upvoted 1 times

Remi2021 5 months, 2 weeks ago

Challenging. B is right one but with B you do not automate wich 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

John_Pongthorn 5 months, 2 weeks ago

B is possible? https://cloud.google.com/logging/docs/view/logs-explorer-interface?cloudshell=true

upvoted 1 times

John_Pongthorn 5 months, 2 weeks ago

Is it Possible to B ? This is the Easy way

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: A

A seem good for me

upvoted 2 times

AWSandeep 5 months, 4 weeks ago

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

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

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

upvoted 4 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 (76%) C (24%)

AWSandeep Highly Voted 5 months, 4 weeks 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 9 times

TNT87 5 months, 1 week ago
WHy do you have to put the development team at the access perimeter???
upvoted 1 times

maci_f Most Recent 1 month 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 the perimeter, so it can access BQ, but is prohibited from accessing the Cloud Storage
upvoted 2 times

tav1nald3rs0n 1 month, 1 week 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 1 times

zellck 3 months ago
Selected Answer: D
D is the answer.
upvoted 1 times

TNT87 1 month, 2 weeks ago
Answer C, i dnt know if you have studied cloud security? if you have you will know
upvoted 1 times

Atnafu 3 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 3 months ago
I meant D Not C
upvoted 2 times

cloudmon 3 months, 3 weeks ago
Selected Answer: D
D makes the most sense to me
upvoted 4 times

cloudmon 3 months, 3 weeks 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 3 months, 3 weeks 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_ 3 months, 3 weeks 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 3 months, 2 weeks 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 3 months, 3 weeks ago
Selected Answer: C

Answer C
https://cloud.google.com/vpc-service-controls/docs/overview#isolate

upvoted 1 times

TNT87 4 months, 3 weeks ago

Selected Answer: C

Answer C
https://cloud.google.com/vpc-service-controls/docs/overview#isolate

upvoted 4 times

Wasss123 5 months, 3 weeks ago

Shoud 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 4 times

TNT87 5 months, 3 weeks ago

Ans C
https://cloud.google.com/vpc-service-controls/docs/overview#isolate

upvoted 2 times

nwk 5 months, 3 weeks ago

Vote C
https://cloud.google.com/vpc-service-controls/docs/vpc-accessible-services

upvoted 3 times

ducc 5 months, 4 weeks ago

A is more straight forward
Anyone can confirm?

upvoted 1 times

soichirokawa 5 months, 3 weeks ago

A. can not be correct. Roles are always inherited, and there is no way to explicitly remove a permission for a lower-level resource that is granted at a higher level in the resource hierarchy. Given the above example, even if you were to remove the Project Editor role from Bob on the "Test project", he would still inherit that role from the "Department Y" folder, so he would still have the permissions for that role on "Test project".
https://cloud.google.com/resource-manager/docs/cloud-platform-resource-hierarchy

upvoted 4 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

AWSandeep Highly Voted 5 months, 4 weeks 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 6 times

musumusu Most Recent 3 days, 7 hours ago

Answer D:
Cost effective transactional database Cloud SQL. Spanner is good case for data more than 30 GB

upvoted 1 times

odacir 2 months, 2 weeks 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 2 times

TNT87 5 months, 1 week ago

Selected Answer: A

https://cloud.google.com/spanner/docs/jdbc-drivers
Ans A
https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configurations
The last part of the question makes it easy

upvoted 3 times

TNT87 5 months, 2 weeks ago

Yes Spanner is expensive , but the question expresslty states that "after securing funding you want to have a global presence" the word globally is repeatedly stated there.
Answer is A.

upvoted 1 times

TNT87 5 months, 2 weeks ago

https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configurations
Ans A

upvoted 1 times

badrisrinivas9 5 months, 2 weeks 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 5 months, 3 weeks 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 wich is cheaper than multi region.

upvoted 2 times

YorelNation 5 months, 3 weeks ago

Selected Answer: D

Spanner is expensive can't be A

Would choose D

upvoted 2 times

TNT87 5 months, 2 weeks ago

The fact that its global cloud spanner is the answer. Secondly Option D, the fact that it has to be highly avaible 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

upvoted 1 times

YorelNation 5 months, 3 weeks 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 4 months, 3 weeks ago

no its A

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: A

Spanner still support JDBC
https://cloud.google.com/spanner/docs/jdbc-drivers

upvoted 2 times

Question #189 *Topic 1*

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

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

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 (83%) B (17%)

musumusu 1 week, 1 day 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 month, 3 weeks ago
A is right
upvoted 1 times

AzureDP900 1 month, 3 weeks 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

zellck 3 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-of-on-premises-data
Like gsutil, Storage Transfer Service for on-premises data enables transfers from network file system (NFS) storage to Cloud Storage. Although gsutil can support small transfer sizes (up to 1 TB), Storage Transfer Service for on-premises data is designed for large-scale transfers (up to petabytes of data, billions of files).
upvoted 2 times

Atnafu 3 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 3 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 3 times

devaid 4 months, 3 weeks ago
Selected Answer: A
Well it doesn't mentions anything about not enough bandwidth to meet your project deadline. I guess you can assume they have 200GBps+ of bandwith, otherwise it shouldn't take only a few hours.
upvoted 4 times

pluiedust 5 months, 2 weeks ago
Selected Answer: A
A is correct.
upvoted 1 times

bigquery1102 5 months, 2 weeks 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

upvoted 1 times

MounicaN 5 months, 2 weeks ago

Selected Answer: A

Huge data can be migrated over cloud interconnect

upvoted 1 times

GyaneswarPanigrahi 5 months, 2 weeks ago

1. Transfer over a secure connection.
2. Should take only a few hours.

Hence, Storage Transfer Service instead of Transfer Appliance.

upvoted 1 times

TNT87 5 months, 3 weeks ago

Answer A
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets

upvoted 1 times

YorelNation 5 months, 3 weeks ago

Selected Answer: A

Same A

upvoted 1 times

nwk 5 months, 3 weeks ago

Vote A
Questions states "transfer over a secure connection" , not offline with Transfer Appliance

upvoted 3 times

ducc 5 months, 4 weeks ago

Selected Answer: B

I vote for B

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: B

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

upvoted 1 times

Question #190 *Topic 1*

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 (100%)

musumusu 1 week, 1 day ago

Answer C,
Datafusion is costly and current transformation is just a cast transformation in a column.
I guess no one wanna pay for datafusion for this little transformation and Staging table processing handles such minor cleaning.

upvoted 1 times

maci_f 1 month 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 1 times

Mike422 1 month, 3 weeks ago

Apparently chatGPT thinks C is the correct answer just sayin (for the same reason that @saurabhsingh4k wrote)

upvoted 1 times

saurabhsingh4k 2 months, 1 week 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)

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 choice.

upvoted 1 times

Atnafu 2 months, 1 week 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%2C%20without%20having%20to%20manage%20the%20infrastructure.

upvoted 1 times

zellck 3 months ago

Selected Answer: A

A is the answer.

https://cloud.google.com/data-fusion/docs/concepts/overview
loud 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 month, 3 weeks ago

thx for sharing link

upvoted 1 times

samirzubair 3 months ago

The Correct Ans is C

upvoted 2 times

jkhong 2 months, 1 week ago

although this is my preferred answer. this doesn't satisfy how this becomes a pipeline.

upvoted 1 times

hiromi 3 months, 1 week ago

Selected Answer: A

Data Fusion

upvoted 1 times

TNT87 5 months, 2 weeks ago

Selected Answer: A

Ans A
https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported_transformations

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: A

A is correct for me

upvoted 1 times

AWSandeep 5 months, 4 weeks 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 (92%) 8%

AzureDP900 1 month, 3 weeks ago
B. Increase the size of the training dataset, and decrease the number of input features.
upvoted 1 times

John_Pongthorn 5 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 4 times

John_Pongthorn 5 months ago
https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html
upvoted 2 times

pluiedust 5 months, 2 weeks ago
Selected Answer: B
B is correct
upvoted 2 times

TNT87 5 months, 3 weeks ago
Answer B
https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/
upvoted 2 times

HarshKothari21 5 months, 3 weeks 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 2 times

YorelNation 5 months, 3 weeks ago
Selected Answer: B
Best option is not mentioned: generalize you neural net by decreasing the complexity of it's 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 5 months, 3 weeks 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 5 months, 3 weeks ago
Selected Answer: D
D is correct

upvoted 1 times

AWSandeep 5 months, 4 weeks ago

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

upvoted 1 times

ducc 5 months, 4 weeks 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-cade option, and you want to be able to easily train the chatbot to provide answers to keywords. What should you do?

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

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

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

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

Correct Answer: *D*

Community vote distribution

D (81%) C (19%)

PhuocT Highly Voted 5 months, 4 weeks 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 7 times

zellck Most Recent 3 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 3 times

AzureDP900 1 month, 3 weeks ago

Agree with D

upvoted 1 times

Atnafu 3 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

devaid 4 months, 1 week ago

Selected Answer: D

D definitely, as the documentation says (specially that you can call the detect Intect 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 3 times

TNT87 5 months, 3 weeks ago

Answer D

https://cloud.google.com/dialogflow/es/docs/how/detect-intent-tts

upvoted 3 times

nwk 5 months, 3 weeks ago

https://cloud.google.com/dialogflow/es/docs/how/detect-intent-stream
Vote D

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: C

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

This seem the best answer here but not the best answer in real world.
But with the Question, the answer must be the combination of both Diagflow and Speech API

upvoted 3 times

Question #193 *Topic 1*

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 (72%) B (28%)

beanz00 Highly Voted 3 months, 4 weeks ago
This has to be D. How could it even be B? The source is a proprietary format. Dataflow wouldn't have a built-in template to ead the file. You will have to create something custom.
upvoted 8 times

devaid Highly Voted 4 months, 1 week 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 7 times

musumusu Most Recent 1 week, 1 day ago
Answer is D ,
Why not B, changing data format before uploading to bigquery is good approach.
upvoted 1 times

cetanx 1 month 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 month ago
sorry the answer should be D
upvoted 1 times

AzureDP900 1 month, 3 weeks ago
D is right
upvoted 1 times

zellck 2 months, 3 weeks ago
Selected Answer: D
D is the answer.
upvoted 3 times

TNT87 1 month, 2 weeks ago
There is dataflow connector and D isnt cost effective
upvoted 1 times

hauhau 2 months, 3 weeks ago
Selected Answer: B
B is the most efficient
upvoted 2 times

TNT87 4 months, 3 weeks ago
https://cloud.google.com/spanner/docs/change-streams/use-dataflow#core-concepts
upvoted 1 times

TNT87 5 months, 3 weeks 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 5 months, 2 weeks ago
https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion
upvoted 1 times

learner2610 5 months, 2 weeks 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 5 months, 2 weeks ago
Option D streams, thats not cost effective. We need something that is cost effectictive, hence B is the option
upvoted 1 times

TNT87 5 months, 1 week ago
I mean that consumes fewer resources
upvoted 1 times

TNT87 5 months, 2 weeks ago
Do you mind reading the links i provided and revisiting the question, then you will understand why D isnt the best. Why use Apache beam yet there is Dataflow
upvoted 1 times

[Removed] 5 months, 1 week ago
Can Bigquery handle a proprietary file format?
upvoted 1 times

TNT87 5 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 5 months, 4 weeks 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

ducc 5 months, 4 weeks ago
Selected Answer: B
B is the most efficient for me.
upvoted 2 times

ducc 5 months, 3 weeks ago
Sorry, D is correct
upvoted 2 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%

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

AzureDP900 1 month, 3 weeks ago
A. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to the Python API.
upvoted 1 times

zellck 3 months ago
Selected Answer: A
A is the answer.
upvoted 3 times

GCPCloudArchitectUser 3 months, 2 weeks 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 3 times

TNT87 5 months ago
https://cloud.google.com/functions/docs/calling/pubsub
upvoted 1 times

TNT87 5 months, 3 weeks ago
Selected Answer: A
Ans A
https://cloud.google.com/functions/docs/calling/pubsub#deployment
upvoted 3 times

YorelNation 5 months, 3 weeks ago
Selected Answer: A
A because D is stupidly high latency
upvoted 2 times

nwk 5 months, 3 weeks ago
Vote A, can't see the need for composer
upvoted 1 times

soichirokawa 5 months, 3 weeks ago
A might be enough. Cloud composer will be an overkill
upvoted 2 times

AWSandeep 5 months, 4 weeks 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 5 months, 4 weeks ago

Selected Answer: D

D is a more recommend way by Google, IMO.

upvoted 1 times

PhuocT 5 months, 4 weeks ago

A. more sense to me.

upvoted 1 times

ducc 5 months, 4 weeks ago

Composer support exception and retry for complex pipeline.
D might be correct

upvoted 1 times

Question #195

Topic 1

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 5 months, 4 weeks ago

Selected Answer: B

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

upvoted 6 times

AzureDP900 Most Recent 1 month, 3 weeks ago

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

upvoted 2 times

zellck 3 months ago

Selected Answer: B

B is the answer.

upvoted 3 times

TNT87 5 months, 2 weeks ago

Answer B.
https://cloud.google.com/bigquery/docs/query-overview

upvoted 3 times

ducc 5 months, 4 weeks ago

Selected Answer: B

B is correct

upvoted 2 times

Question #196 *Topic 1*

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%)

musumusu 1 week, 1 day ago
answer C,
To avoid confustion: Install Storage Transfer Service is always on EXTERNAL OR NON GOOGLE service or data centre to connect google service.
upvoted 1 times

zellck 3 months ago
Selected Answer: C
C is the answer.

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#storage-transfer-service-for-large-transfers-of-on-premises-data
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

NicolasN 3 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 2 times

Atnafu 3 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 5 months, 1 week ago
why can't it be D?
upvoted 1 times

zellck 3 months ago
Installation is done for the source which is in your data centre, and not in GCP.
upvoted 2 times

Wasss123 5 months, 2 weeks ago
Selected Answer: C

I vote for C

upvoted 2 times

TNT87 5 months, 2 weeks ago

Ans C
https://cloud.google.com/storage-transfer/docs/overview

upvoted 2 times

MounicaN 5 months, 2 weeks ago

can you help with difference between c and d ?

upvoted 3 times

gudiking 3 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 1 times

AWSandeep 5 months, 4 weeks 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 5 months, 4 weeks 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 (100%)

NicolasN Highly Voted 3 months, 3 weeks 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 15 times

Mcloudgirl 3 months, 2 weeks ago

Your explanation is perfect, thanks

upvoted 2 times

musumusu Most Recent 1 week, 1 day 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 Highavailablity option means the same extra replicas instances are available for emergency.

upvoted 1 times

AzureDP900 1 month, 3 weeks ago

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

upvoted 1 times

zellck 3 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 3 months ago

I voted for B

upvoted 1 times

John_Pongthorn 5 months ago

Selected Answer: B

B it is exact anwer.

upvoted 1 times

TNT87 5 months, 2 weeks ago

Selected Answer: B

Ans B
Postgres is highly ACID compliant as compared to Mysql

upvoted 2 times

Remi2021 5 months, 2 weeks ago

Selected Answer: B

cloud sql with high availability enabled is enough

upvoted 2 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: B

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

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: B

I voted for B

upvoted 1 times

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 (100%)

zellck 3 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

musumusu 3 days, 7 hours ago

why not option C? if composer in host project it will be easier to connect one or more service project with it.

upvoted 1 times

AzureDP900 1 month, 3 weeks ago

Agreed

upvoted 1 times

Atnafu 3 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 the host project.
https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc#:~:text=This%20page%20describes,the%20service%20project.

upvoted 2 times

AWSandeep 5 months, 4 weeks ago

Selected Answer: D

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

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

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

upvoted 3 times

ducc 5 months, 4 weeks 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 the 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

upvoted 2 times

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 (94%) 6%

AWSandeep Highly Voted 5 months, 4 weeks ago

Selected Answer: A

A. Use BigQuery BI Engine with materialized views.

upvoted 6 times

zellck Most Recent 3 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 2 times

LPIT 3 months, 4 weeks 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 2 times

Julionga 5 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%20thereby%20reducing%20query%20time.%20You%20should%20create%20materialized%20views%20to%20improve%20performance%20and%20to%20reduce%20processed%20data%20by%20using%20aggregations%2C%20filters%2C%20inner%20joins%2C%20and%20unnests.

upvoted 2 times

MounicaN 5 months, 2 weeks ago

Selected Answer: A

use materialized views is better option here

upvoted 3 times

ducc 5 months, 4 weeks ago

Selected Answer: C

By integrating BI Engine with BigQuery streaming, you can perform real-time data analysis over streaming data without sacrificing write speeds or data freshness.

https://cloud.google.com/bigquery/docs/bi-engine-intro

upvoted 1 times

ducc 5 months, 3 weeks 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

A (54%) D (46%)

NicolasN Highly Voted 2 months, 3 weeks 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 9 times

Aamir185 Most Recent 1 week, 6 days ago

Selected Answer: D

D it is

upvoted 1 times

cetanx 1 month 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 managable and better than A

upvoted 2 times

AzureDP900 1 month, 3 weeks ago

D is right

upvoted 1 times

Amar2022 2 months ago

Selected Answer: A

A is the correct one

upvoted 1 times

jkhong 2 months, 1 week 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 2 months, 1 week ago

Selected Answer: A

A is the answer, as NicalsN says.
https://cloud.google.com/iam/docs/service-accounts#groups

upvoted 1 times

Andrix2405 2 months, 1 week ago

Selected Answer: A

Avoid using groups for granting service accounts access to resources -> D

upvoted 2 times

Andrix2405 2 months, 1 week ago

Sorry A

upvoted 1 times

zellck 3 months ago

Selected Answer: D

D is the answer.

upvoted 1 times

hauhau 3 months, 1 week ago

Why GCS?

upvoted 1 times

Wasss123 5 months, 2 weeks ago

Selected Answer: D

I vote D

upvoted 4 times

AWSandeep 5 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.
Reveal Solution

upvoted 2 times

ducc 5 months, 4 weeks ago

Selected Answer: D

D is the best for me

upvoted 2 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%)

AWSandeep Highly Voted 5 months, 4 weeks ago

Selected Answer: A

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

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

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

upvoted 7 times

AzureDP900 Most Recent 1 month, 3 weeks ago

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.

upvoted 1 times

zellck 3 months ago

Selected Answer: A

A is the answer.

https://cloud.google.com/memorystore/docs/redis/about-importing-exporting
The import and export feature uses the native RDB snapshot feature of Redis to import data into or export data out of a Memorystore for Redis instance. The use of the native RDB format prevents lock-in and makes it very easy to move data within Google Cloud or outside of Google Cloud. Import and export uses Cloud Storage buckets to store RDB files.

upvoted 3 times

gudiking 3 months ago

Selected Answer: A

A
https://cloud.google.com/memorystore/docs/redis/import-data

upvoted 1 times

Atnafu 3 months ago

A
Import and export uses Cloud Storage buckets to store RDB files.
https://cloud.google.com/memorystore/docs/redis/about-importing-exporting#:~:text=Import%20and%20export%20uses%20Cloud%20Storage%20buckets%20to%20store%20RDB%20files.

upvoted 1 times

ducc 5 months, 4 weeks ago

Selected Answer: A

A
https://cloud.google.com/memorystore/docs/redis/general-best-practices

upvoted 2 times

Question #202 *Topic 1*

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

musumusu 1 week ago
Answer C,
What is wrong with B ? Key words = Daily transfer .. so no to transfer appliance,
upvoted 1 times

zellck 3 months ago
Selected Answer: C
C is the answer.

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#storage-transfer-service-for-large-transfers-of-on-premises-data
Storage Transfer Service for on-premises data enables transfers from network file system (NFS) storage to Cloud Storage.

https://cloud.google.com/bigquery/docs/cloud-storage-transfer-overview
The BigQuery Data Transfer Service for Cloud Storage lets you schedule recurring data loads from Cloud Storage buckets to BigQuery.
upvoted 1 times

AzureDP900 1 month, 3 weeks ago
yes, It is C
upvoted 1 times

Atnafu 3 months ago
C
D-no answer because bq transfer service don't support from on-prem
upvoted 1 times

Atnafu 3 months ago
B-is not answer because you want transfer appliance for one time bulk transfer but the question is You want to use Google Cloud products to query and explore the platform data.

query and explore is the key
upvoted 1 times

John_Pongthorn 4 months, 3 weeks 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 4 months, 3 weeks 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 4 months, 3 weeks ago
Sorry I am wrong
(large files > 1 TB + bandwidth available on internal IP address communication + daily scheduling)
upvoted 1 times

Wasss123 5 months, 2 weeks ago
Selected Answer: C
I will go with C
upvoted 3 times

MounicaN 5 months, 2 weeks ago
I will g with C

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer-options
upvoted 1 times

John_Pongthorn 5 months, 3 weeks ago
C is correct, b is suitable for weekly .
https://cloud.google.com/transfer-appliance/docs/4.0/overview
upvoted 2 times

John_Pongthorn 4 months, 3 weeks 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 5 months, 3 weeks ago
Selected Answer: C
Ans C
https://cloud.google.com/storage-transfer/docs/on-prem-agent-best-practices
upvoted 1 times

HarshKothari21 5 months, 3 weeks ago
I would go with option C.
You need a service to transfer data from on-premises to cloud storage. so "Transfer service" is the best option & additionally you can easily configure the network so that data flows through private network.

cloud scheduler on other hand is used mostly for automation. You can schedule a service but in my view cannot be used solo to transfer data.
upvoted 1 times

nwk 5 months, 3 weeks 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 5 months, 4 weeks 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%)

AzureDP900 1 month, 3 weeks ago
C. Train the model using a VM with a GPU hardware accelerator.
upvoted 1 times

jkhong 2 months, 1 week 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 1 times

Atnafu 2 months, 4 weeks ago
C
https://cloud.google.com/tpu/docs/tpus#when_to_use_tpus:~:text=Models%20with%20a%20significant%20number%20of%20custom%20TensorFlow%20operations%20that%20must%20run%20at%20least%20partially%20on%20CPUs
upvoted 1 times

Atnafu 2 months, 1 week ago
The model has custom TensorFlow operations that must run partially on a CPU. is the key for GPU
upvoted 2 times

zellck 3 months ago
Selected Answer: C
C is the answer.

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 2 times

gudiking 3 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
upvoted 1 times

gudiking 2 months, 4 weeks ago
C is not cost-effective, so I stand corrected. I do not know the answer.
upvoted 1 times

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%)

AzureDP900 1 month, 3 weeks 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 2 months, 2 weeks ago

Selected Answer: D

D is the best option to sanitize the data to its D

upvoted 2 times

jkhong 2 months, 3 weeks 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 2 months, 3 weeks ago

Selected Answer: D

D is the answer.

upvoted 1 times

vidts 2 months, 4 weeks ago

Selected Answer: D

It's D

upvoted 2 times

Atnafu 2 months, 4 weeks 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%)

charline 2 weeks, 1 day ago

Selected Answer: B

b is ok

upvoted 1 times

AzureDP900 1 month, 3 weeks 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

zellck 2 months, 3 weeks ago

Selected Answer: B

B is the answer.

upvoted 2 times

hauhau 2 months, 3 weeks ago

Selected Answer: B

Maybe B

ref: https://cloud.google.com/architecture/managing-infrastructure-as-code

upvoted 2 times

Atnafu 2 months, 4 weeks ago

C is correct answer

upvoted 1 times

Atnafu 2 months, 4 weeks ago

Sorry I meant B

upvoted 3 times