Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 30 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 30
Topic #: 1
[All Professional Data Engineer Questions]

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.

Show Suggested Answer

by jvg637 at *March 15, 2020, 12:56 p.m.*

Comments

Type your comment...

Submit

HectorLeon2099 Highly Voted 4 years, 4 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 114 times

Tanzu 3 years, 3 months 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 10 times

cetanx 2 years, 3 months ago

Answer - B
mysql dump: This utility creates a logical backup and a flat file containing the SQL statements that can be run again to bring back the database to the state when this file was created. So this file can easily be processed by an ETL tool and loaded into BQ.

upvoted 2 times

ThorstenStaerk 2 years ago

So, you are saying that B takes the backup data from the nightly dumps? How can you be sure?

upvoted 2 times

cetanx 1 year, 11 months ago

I agree that B sounds like running ETL directly on the database. It doesn't say anything explicitly about using dumps.

However, by leveraging the Dataproc JDBC Connector, one can perform various operations such as querying, joining, filtering, and aggregating data from your SQL databases within your Dataproc jobs. This can be particularly useful when you want to combine data from multiple sources or perform complex data transformations before processing the data further.

So with D, you can run your analysis from a separate cloud-sql instance created from the dump and without affecting the production database.

upvoted 2 times

Preetmehta1234 1 year, 2 months ago

That's so true! This should be the first logic for elimination

upvoted 2 times

[Removed] Highly Voted 5 years, 1 month ago

Answer: D
Description: Easy and it won't affect processing

upvoted 41 times

dambilwa 4 years, 10 months ago

Agreed- Option[D] is most appropriate in this scenario

upvoted 6 times

StefanoG 3 years, 7 months ago

So I vote for B

upvoted 6 times

Alexej_123 4 years, 4 months ago

I think it is B and not D:
1) There are no info regarding date freshness required for analytics. So nightly backup might be not enough as a source because it will only provide info one tie a day.
2) Dataproc is recommended as easiest way for migration of hadoop processes. SO no reason to use Dataproc for designing a new analytics processes.
3) The solution is really very limited if you will extend it in the future and add new data sources or create new aggregate tables. Where they should be created?
4) There is no info on which version is on prem MySQL database (I am not an expert in MySql) but I can imagine there might be compartibility issue for backup / restore between different releases

upvoted 15 times

g2000 4 years, 3 months ago

but how about the impact on operation? D seems better match.

upvoted 2 times

StefanoG 3 years, 7 months ago

I also think the answer is D, because on B it is not written that the source is the backup but (directly) MYSQL. So wit this solution we add requests on MySQL and so, mpacting the operations-

upvoted 4 times

hellofrnds 3 years, 6 months ago

" Dataproc makes open source data and analytics processing fast, easy, and more secure in the cloud ". Please refer this google link.
https://cloud.google.com/blog/products/data-analytics/genomics-data-analytics-with-cloud-pt2

upvoted 1 times

sergio6 3 years, 6 months ago

The link titles "Genomics analysis with Hail, BIGQUERY, and Data Proc", the solution describes the use of Bigquery to do analytics

upvoted 1 times

Load full discussion...

StefanoG 3 years, 7 months ago

Google Cloud Dataproc is not an analytic tool

upvoted 10 times

Load full discussion...

vosang5299 Most Recent 2 weeks, 1 day ago

Selected Answer: B

B is correct

upvoted 1 times

willyunger 1 month, 2 weeks ago

Selected Answer: D

Option D has no impact on operations, uses backups which are already there. Option B with ETL could impact MySQL performance.

upvoted 1 times

Juanesdelacruz97 3 months, 1 week ago

I think it's B, today BigQuery has multiple connectors that can allow an easy connection to external data sources without impacting the database itself, even if the database was in a SQL instance, MySQL, Federated queries could be used. In my opinion it's B

upvoted 1 times

Augustax 3 months, 3 weeks ago

Selected Answer: D

Since the question mentions the nightly backup, why we cannot use it? ETL reduces the impact of the source system but still some impacts. D doesn't add any additional impact.

upvoted 1 times

meh_33 8 months, 3 weeks ago

Believe me all questions were from Exam topic all were there yesterday in exam. But yes dont go with starting questions mainly focus questions after 200 and latest questions are at last page.

upvoted 1 times

Gayatri147 8 months, 3 weeks ago

How you accessed questions after question number 70 it is asking for pro subscription ?

upvoted 1 times

mark1223jkh 11 months, 3 weeks ago

Answer B:
I don't know why people are choosing D. It is two steps, first cloudsql and then dataproc, a lot of overhead. BigQuery is just perfect fit.

upvoted 1 times

0725f1f 1 year, 2 months ago

Selected Answer: D

This won't affect processing

upvoted 1 times

TVH_Data_Engineer 1 year, 5 months ago

Selected Answer: B

Based on these considerations, option B is likely the best approach. By using an ETL tool to load data from MySQL into Google BigQuery, you're leveraging BigQuery's strengths in handling large-scale analytics workloads without impacting the performance of the operational databases. This option provides a clear separation of operational and analytical workloads and takes advantage of BigQuery's fast analytics capabilities.

upvoted 2 times

axantroff 1 year, 5 months ago

Selected Answer: B

Do not spend much time on in - just B

upvoted 1 times

rocky48 1 year, 6 months ago

Selected Answer: B

Answer is B - Use an ETL tool to load the data from MySQL into Google BigQuery.

* Google BigQuery is a serverless, highly scalable data warehouse that can handle large-scale analytics workloads without impacting your MySQL cluster's performance.
* Using an ETL (Extract, Transform, Load) tool to transfer data from MySQL to BigQuery allows you to maintain a separate analytics environment, ensuring that your operational database remains unaffected.

Option C (connecting an on-premises Apache Hadoop cluster to MySQL and performing ETL) introduces complexity and may not be as scalable as a cloud-based solution.

Option D (mounting backups to Google Cloud SQL and processing the data using Google Cloud Dataproc) could be an option for historical data analysis but might not be the best choice for real-time analytics while the MySQL cluster is under heavy load. Additionally, the backups need to be restored and processed, which might introduce some delay.

upvoted 2 times

mk_choudhary 1 year, 6 months ago

It's GOOGLE exam where choosing the GCP service shall be first preference.
Now notice the problem statement "perform analytics with minimal impact on operations"
BigQuery is right option for analytic as well as Cloud SQL does provide easy export to GCS where we can query from BigQuery without loading into BQ to save storage cost.

upvoted 2 times

rtcpost 1 year, 6 months ago

Selected Answer: B

B. Use an ETL tool to load the data from MySQL into Google BigQuery.
* Google BigQuery is a serverless, highly scalable data warehouse that can handle large-scale analytics workloads without impacting your MySQL cluster's performance.
* Using an ETL (Extract, Transform, Load) tool to transfer data from MySQL to BigQuery allows you to maintain a separate analytics environment, ensuring that your operational database remains unaffected.

Option C (connecting an on-premises Apache Hadoop cluster to MySQL and performing ETL) introduces complexity and may not be as scalable as a cloud-based solution.

Option D (mounting backups to Google Cloud SQL and processing the data using Google Cloud Dataproc) could be an option for historical data analysis but might not be the best choice for real-time analytics while the MySQL cluster is under heavy load. Additionally, the backups need to be restored and processed, which might introduce some delay.

upvoted 3 times

melligeri 1 year, 6 months ago

Selected Answer: B

The question clearly says there is load on MYSQL already so doing analytics on it is bad idea. Its bad to run analytics on MYSQL but still a better option to run etl with it to load it to BigQuery.

upvoted 1 times

imran79 1 year, 7 months ago

B. Use an ETL tool to load the data from MySQL into Google BigQuery. This way, analytics is entirely separated from the operational database, and BigQuery is well-suited for large-scale analytics.

upvoted 2 times

emmylou 1 year, 7 months ago

The correct answer is to build a read replica :-) but since we can't do that then migrating to BigQuery will have to suffice.

upvoted 2 times

Fotofilico 1 year, 6 months ago

thanks! :3

upvoted 1 times

Load full discussion...