EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 101 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 101
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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. Most Voted

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.

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (64%) A (36%)

by Rajokkiyam at *March 22, 2020, 6:35 a.m.*

Comments

Ganshank Highly Voted 3 years, 3 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 56 times

AzureDP900 7 months, 1 week ago

B is right answer

upvoted 2 times

tprashanth 3 years ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 102 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 102
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *C*

Community vote distribution
A (51%) C (49%)

by Rajokkiyam at *March 22, 2020, 6:37 a.m.*

Comments

MaxNRG Highly Voted 1 year, 7 months ago

Selected Answer: A
A - New correct answer
C - Old correct answer (for 2019)
upvoted 30 times

Yiouk 3 weeks, 2 days ago

There are still limitations on DML statements (2023) e.g. only 2 concurrent UPDATES and up to 20 queued hence not appropriate for this scenario:
https://cloud.google.com/bigquery/quotas#data-manipulation-language-statements
upvoted 1 times

NeoNitin 2 weeks, 6 days ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 103 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 103
Topic #: 1
[All Professional Data Engineer Questions]

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?

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. Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

C (100%)

by [deleted] at *March 22, 2020, 3:48 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer - C
upvoted 27 times

[Removed] 3 years, 4 months ago
highly available = multi-regional:
https://cloud.google.com/bigquery/docs/locations

recovery strategy of this data that minimizes cost = point-in-time snapshot:
https://cloud.google.com/solutions/bigquery-data-warehouse#backup-and-recovery
upvoted 16 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: C
Description: In multiregional, data is not lost and recovery time is ms. Regional, zone failure results in data loss

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 104 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 104
Topic #: 1
[All Professional Data Engineer Questions]

You used Cloud 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 Cloud Dataprep.

B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.

C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.

D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job. Most Voted

Hide Answer

Suggested Answer: *C*

Community vote distribution

D (100%)

by [deleted] at *March 22, 2020, 4:12 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: D
Description: Dataprep can be run on Dataflow using template and cloud composer will create dependency on previous job
upvoted 22 times

kino2020 Highly Voted 2 years, 10 months ago
Should be D

https://cloud.google.com/blog/products/data-analytics/how-to-orchestrate-cloud-dataprep-jobs-using-cloud-composer

We're happy to announce the latest release of Cloud Dataprep, which exposes orchestration APIs so you can integrate Cloud Dataprep within your schedulers or other orchestration solutions like Cloud Composer. This means you can expand your automation beyond Cloud Dataflow templates through direct integration in other tools to create repeatable data pipelines for your analytics and AI/ML initiatives—saving time and adding reliability. In addition, this API lets you use dynamic inputs and outputs through Cloud Dataprep variables or parameters—not available using Cloud Dataflow templates. As a result, you can re-use a single Cloud Dataprep flow to execute on a range of input/output values that are evaluated at runtime.
upvoted 12 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 105 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 105
Topic #: 1
[All Professional Data Engineer Questions]

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud 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 Most Voted
C. Cloud Scheduler
D. Workflow Templates on Cloud Dataproc

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (100%)

by Rajokkiyam at *March 22, 2020, 6:40 a.m.*

Comments

Ysance_AGS Highly Voted 1 year, 10 months ago
this website is it using Random() to answer the questions ??
upvoted 31 times

squishy_fishy 1 year, 10 months ago
I totally agree! LOL
upvoted 7 times

subhsubh 1 year, 6 months ago
Even Random() would have given more correct answers :-P
upvoted 6 times

Rajokkiyam Highly Voted 3 years, 4 months ago
Cloud Composer
upvoted 30 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 106 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 106
Topic #: 1
[All Professional Data Engineer Questions]

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.

Hide Answer

Suggested Answer: *D*
Reference -
https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex

Community vote distribution
D (67%) A (33%)

by Rajokkiyam at *March 22, 2020, 6:42 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: D
Description: Graceful decommissioning will ensure that the data is processed by worker before it is removed by Yarn
upvoted 20 times

NicolasN 8 months, 2 weeks ago
Can please anybody explain in which way may a worker be removed when increasing cluster size?
upvoted 1 times

Rajokkiyam Highly Voted 3 years, 4 months ago
Answer D
upvoted 11 times

zellck Most Recent 8 months ago
Selected Answer: D

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 107 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 107
Topic #: 1
[All Professional Data Engineer Questions]

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 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 Stackdriver 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 API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review. Most Voted

Hide Answer

Suggested Answer: *A*

Community vote distribution

D (81%) A (19%)

by Rajokkiyam at *March 22, 2020, 6:44 a.m.*

Comments

Rajokkiyam Highly Voted 3 years, 4 months ago
Data Loss Prevention API does this job
upvoted 20 times

[Removed] Highly Voted 3 years, 4 months ago
Should be D
upvoted 14 times

zellck Most Recent 8 months ago
Selected Answer: D
D is the answer.
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 108 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 108
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *D*

Community vote distribution

A (100%)

by Rajokkiyam at *March 22, 2020, 6:48 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be A
upvoted 35 times

Rajokkiyam Highly Voted 3 years, 4 months ago
Create dependency in Cloud Composer and schedule it.
upvoted 21 times

MisuLava 9 months, 1 week ago
the jobs are not interdependent. just 3 individual jobs
upvoted 1 times

forepick Most Recent 2 months, 1 week ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 109 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 109
Topic #: 1
[All Professional Data Engineer Questions]

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 Stackdriver Log Viewer. 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. Most Voted

D. The subscriber code cannot keep up with the messages.

E. The subscriber code does not acknowledge the messages that it pulls. Most Voted

Hide Answer

Suggested Answer: *CD*

Community vote distribution

CE (100%)

by [deleted] at *March 22, 2020, 2:49 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago

Answer: C, E
Description: 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.
A: No problem with publisher rate as the observed result is a higher number of messages and not a lower number.
B: if messages exceed the 10MB maximum, they cannot be published.
D: Cloud Functions automatically scales so they should be able to keep up.

upvoted 34 times

[Removed] Highly Voted 3 years, 4 months ago

Should be - CE

upvoted 16 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 110 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 110
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

by [deleted] at *March 22, 2020, 2:51 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Correct - B
upvoted 16 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: B
Description: ParDo is used to do transformation and create side output
upvoted 11 times

midgoo Most Recent 4 months, 3 weeks ago
Selected Answer: B
A - SideInput is often used to validate data, however, we need to create the SideInput first. When using SideInput to filter data, it is actually another ParDo call.
C, D - This is common way to filter too, but we will need the key in order to partition or GroupByKey
B - ParDo is the most basic method, it can do anything to the PCollection
upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 111 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 111
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 1:15 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
should be A
upvoted 34 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Partition is the solution for reducing cost and time
upvoted 18 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 112 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 112
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (70%) 10% 10% 10%

by [deleted] at *March 22, 2020, 1:58 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be B
upvoted 31 times

Ganshank Highly Voted 3 years, 3 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 21 times

Yiouk 3 weeks, 2 days ago
C. Can you imagine changing the software in all sensors to use PupSub instead of the existing one? This is out of scope of the question.
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 113 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 113
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

D. AutoML Natural Language

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (77%) A (23%)

by rickywck at *March 17, 2020, 12:12 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
should be C, since we need to recognize both voice and intent
upvoted 26 times

AzureDP900 7 months, 1 week ago
C. Dialogflow Enterprise Editio
upvoted 1 times

Alasmindas Highly Voted 2 years, 9 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

hdmi_switch 2 years 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.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 114 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 114
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

C. Cloud Dataprep

D. Cloud Dataproc

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (100%)

by madhu1171 at *March 15, 2020, 4:16 a.m.*

Comments

madhu1171 Highly Voted 3 years, 4 months ago
Answer should be B
upvoted 29 times

[Removed] Highly Voted 3 years, 4 months ago
Answer - B
upvoted 12 times

forepick Most Recent 2 months, 1 week ago
Selected Answer: B
No other option is aimed for this purpose
upvoted 1 times

juliobs 4 months, 2 weeks ago
Selected Answer: B
Airflow
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 115 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 115
Topic #: 1
[All Professional Data Engineer Questions]

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. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.
Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

A (100%)

by jvg637 at *March 19, 2020, 10:20 a.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
A: By creating an authorized view one assures that the data is current and avoids taking more storage space (and cost) in order to share a dataset.
B and D are not cost optimal and C does not guarantee that the data is kept updated
upvoted 34 times

raf2121 2 years ago
Is Authorized View is for DataSet or Tables ? I believe its for Dataset.
Options A states Authorized view on BQ Table
upvoted 2 times

PM17 1 year, 10 months ago
Thanks for the explanation!
upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 116 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 116
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

C. Tier older data onto Cloud Storage files, and leverage extended tables.

D. Re-create the table using data partitioning on the package delivery date.

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (100%)

by rickywck at *March 17, 2020, 12:32 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
I think the answer is B. How come we cluster the table again with ingestion date given it is already partitioned with ingestion date?
upvoted 26 times

[Removed] Highly Voted 3 years, 4 months ago
Should be B
upvoted 14 times

Pime13 Most Recent 1 year, 1 month ago
Selected Answer: B
vote b
upvoted 1 times

VictorBa 1 year, 3 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 117 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 117
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by madhu1171 at *March 15, 2020, 4:23 a.m.*

Comments

madhu1171 Highly Voted 3 years, 4 months ago
Answer should be D
upvoted 26 times

[Removed] Highly Voted 3 years, 4 months ago
Answer - D
upvoted 13 times

NeoNitin Most Recent 1 day, 18 hours ago
Data proc is serverbased
Dataflow is serverless which is used to run pipelines which uses apache framework in the background. Just need to mention the number of workers needed.

so question saying we need scale automatically . so dataproc eliminate ho gaya
now Dataflow is correct , pub/sub is recommended for this scenario. D
upvoted 1 times

dconesoko 7 months, 1 week ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 118 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 118
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (100%)

by jvg637 at *March 18, 2020, 8:49 p.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
Hi, I also choose B for two reasons. One is that we want access at department level. In C & D, it is at project level. That means, one lead of one department will have all permissions for different department if all tables are in same project.
upvoted 17 times

Alasmindas Highly Voted 2 years, 8 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 119 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 119
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 12:41 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Timestamp at starting of rowkey causes bottleneck issues
upvoted 41 times

kichukonr Highly Voted 3 years, 3 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

Abhi16820 1 year, 8 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.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 120 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 120
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (92%) 8%

by [deleted] at *March 22, 2020, 12:46 p.m.*

Comments

dambilwa Highly Voted 3 years, 1 month 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 23 times

Barniyah 3 years, 1 month ago
Yes, you are right, it should be B:
Thank you
upvoted 4 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 121 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 121
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by [deleted] at *March 22, 2020, 11:22 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be C
upvoted 21 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: C
Description: Pubsub is global and dataflow can scale workers
upvoted 18 times

ga8our Most Recent 2 months, 1 week ago
Can anyone pls explain what's wrong with D, the load balancing solution?
upvoted 1 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 122 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 122
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export. Most Voted

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.

Hide Answer

Suggested Answer: *CE*

Community vote distribution

AB (84%) AD (16%)

by rickywck at *March 17, 2020, 12:49 p.m.*

Comments

Ganshank Highly Voted 3 years, 3 months ago

A,B
https://cloud.google.com/datastore/docs/export-import-entities
upvoted 36 times

salsabilsf 2 years, 3 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 123 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 123
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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. Most Voted

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.

Hide Answer

Suggested Answer: *AD*

Community vote distribution

AD (75%) BD (19%) 6%

by rickywck at *March 20, 2020, 4:21 a.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
I think AD is the answer. E will not improve performance.
upvoted 39 times

[Removed] Highly Voted 3 years, 4 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 2 years ago
My guess is append has better performance than update.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 124 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 124
Topic #: 1
[All Professional Data Engineer Questions]

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 Cloud Dataproc, BigQuery, and Compute
Engine.
⇨ A streaming data pipeline stores new data daily.
⇨ Peformance 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 Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.

B. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.

C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.

D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Hide Answer

Suggested Answer: *C*

Community vote distribution

D (80%) B (20%)

by jvg637 at *March 18, 2020, 8:30 p.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
D (since pdf cannot be stored in BigQuery, and also questions asks for availability)
upvoted 34 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: D
Description: Multi-region increases high availability and pdf can be stored in gcs
upvoted 23 times

hiromi Most Recent 8 months, 2 weeks ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 125 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 125
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (77%) D (15%) 8%

by [deleted] at *March 22, 2020, 12:07 p.m.*

Comments

Rajokkiyam Highly Voted 3 years, 4 months ago
Answer C.
upvoted 34 times

AJKumar Highly Voted 3 years, 1 month 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 22 times

AzureDP900 7 months, 1 week ago
Agree with C
upvoted 1 times

vamgcp Most Recent 1 week, 4 days ago
Selected Answer: B
It can be C or D , but I will go with C as storing the full dataset in BigQuery and a compressed copy of the data in Cloud Storage is a good way to balance performance and cost.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 126 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 126
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

B. Use Cloud Vision AutoML, but reduce your dataset twice. Most Voted

C. Use Cloud Vision API by providing custom labels as recognition hints.

D. Train your own image recognition model leveraging transfer learning techniques.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (53%) B (32%) C (15%)

by [deleted] at *March 22, 2020, 10:52 a.m.*

Comments

Callumr Highly Voted 3 years, 1 month ago
B - You only need a PoC and it has be done quickly
upvoted 50 times

[Removed] Highly Voted 3 years, 4 months ago
Correct - A
upvoted 20 times

musumusu Most Recent 5 months, 2 weeks ago
Whats wrong with C, its fast, cheap and add your 750 labels which is not big work.
AutoML is good to train on big dataset and costly as compared to APIs
upvoted 2 times

knith66 1 week, 5 days ago
it is a labeled dataset and why do you need to label it once again? So no C

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 127 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 127
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *B*

Community vote distribution

D (51%) C (37%) 12%

by [deleted] at *March 22, 2020, 11:01 a.m.*

Comments

dhs227 Highly Voted 3 years, 4 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 62 times

aiguy Highly Voted 3 years, 4 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 2 years, 4 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

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 128 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 128
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution
C (52%) D (48%)

by [deleted] at *March 22, 2020, 11:11 a.m.*

Comments

Callumr Highly Voted 3 years, 1 month 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 59 times

NeoNitin 1 day, 16 hours ago
Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.
upvoted 1 times

hellofrnds 1 year, 10 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 129 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 129
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (88%) 12%

by [deleted] at *March 22, 2020, 11:14 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be B
upvoted 22 times

Ganshank Highly Voted 3 years, 3 months ago
B
The questions is specifically about organizing the data in BigQuery and storing backups.
upvoted 12 times

Lanro Most Recent 1 week ago
Selected Answer: D
From BigQuery documentation - Benefits of using table snapshots include the following:

- Keep a record for longer than seven days. With BigQuery time travel, you can only access a table's data from seven days ago or more recently. With table snapshots, you can preserve a table's data from a specified point in time for as long as you want.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 130 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 130
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by madhu1171 at *March 15, 2020, 4:01 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
Should be D.

https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery
upvoted 30 times

Rajuuu 3 years ago
There is no mention about merge or limit in the link provided.
upvoted 3 times

Chelseajcole 1 year, 10 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

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 131 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 131
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

C. Create distinct groups for various teams, and specify groups in Cloud IAM policies. Most Voted

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.

Hide Answer

Suggested Answer: *AC*

Community vote distribution

by [deleted] at *March 22, 2020, 10:12 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: B, C
Description: Google suggests that we should provide access by following google hierarchy and groups for users with similar roles
upvoted 31 times

sipsap 2 years, 8 months ago
"Each project requires unique access control configurations" rules out hirearchy
upvoted 11 times

AJKumar Highly Voted 3 years, 1 month ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 132 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 132
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (85%) A (15%)

by [deleted] at *March 22, 2020, 10:15 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago

Answer: B
Description: All the criteria meets for Spanner
upvoted 26 times

sumanshu Highly Voted 2 years, 1 month 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

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 133 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 133
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by rickywck at *March 20, 2020, 3:49 a.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
I think the key reason for pick D is the 100ms requirement.
upvoted 26 times

AzureDP900 7 months, 1 week 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 3 years, 4 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 134 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 134
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (96%) 4%

by [deleted] at *March 22, 2020, 10:43 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
should be B
upvoted 22 times

itche_scratche Highly Voted 2 years, 10 months ago
D, not ideal but only option that work. You need pubsub, then a processing layer (dataflow or dataproc), then storage (some sql database).
upvoted 12 times

jkhong 7 months, 3 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

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 135 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 135
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *A*

Community vote distribution

D (100%)

by [deleted] at *March 22, 2020, 10:49 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Correct - D
upvoted 43 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: D
Description: All the requirements meet with D

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 136 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 136
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

B. Use a larger instance type for your Dataflow workers Most Voted

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

Hide Answer

Suggested Answer: *BE*

Community vote distribution

AB (100%)

by jvg637 at *March 18, 2020, 4:39 p.m.*

Comments

jvg637 Highly Voted 3 years, 4 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 48 times

AzureDP900 7 months, 1 week ago

Agreed
upvoted 2 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 137 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 137
Topic #: 1
[All Professional Data Engineer Questions]

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud 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 Cloud Dataflow pipeline to use local execution

B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions Most Voted

C. Increase the number of nodes in the Cloud Bigtable cluster Most Voted

D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable

E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

Hide Answer

Suggested Answer: *DE*
Reference:
https://www.slideshare.net/LucasArruda3/how-to-build-an-etl-pipeline-with-apache-beam-on-google-cloud-dataflow

Community vote distribution
BC (83%) Other

by rickywck at *March 17, 2020, 1:50 p.m.*

Comments

Rajokkiyam Highly Voted 3 years, 4 months ago
Answer BC
upvoted 38 times

haroldbenites Highly Voted 2 years, 11 months ago
B, C is correct
upvoted 14 times

Oleksandr0501 Most Recent 3 months ago
Selected Answer: BC
I choose with BC, after having read discussion here.
More likely BC

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 138 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 138
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Hide Answer

Suggested Answer: *A*
Reference:
https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows

Community vote distribution
C (100%)

by jvg637 at *March 18, 2020, 5:08 p.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
Option C seems correct.
https://airflow.apache.org/docs/stable/concepts.html
upvoted 22 times

nadavw 1 year, 2 months ago
C. as workflow doesn't have scheduling : https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions
upvoted 1 times

arnabbis4u Highly Voted 3 years, 3 months ago
Answer A. Composer might be an overkill in this case. It is useful when multiple different types of jobs are involved. Since all are dataproc jobs, DataProc Workflows should be sufficient.
upvoted 13 times

rsamant 2 years ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 139 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 139
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *A*
Reference:
https://cloud.google.com/appengine/docs/standard/go/mail/sending-receiving-with-mail-api

Community vote distribution
B (100%)

by rickywck at *March 17, 2020, 1:49 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
Will pick B
upvoted 23 times

[Removed] Highly Voted 3 years, 4 months ago
Should be B
upvoted 9 times

MaxNRG Most Recent 1 year, 6 months ago
Selected Answer: B
B: PUB/SUB
upvoted 3 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 140 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 140
Topic #: 1
[All Professional Data Engineer Questions]

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) Most Voted

D. The original order identification number from the sales system, which is a monotonically increasing integer

Hide Answer

Suggested Answer: *C*
Reference:
https://www.uuidgenerator.net/version4

Community vote distribution
C (100%)

by rickywck at *March 17, 2020, 1:53 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
C is correct

https://cloud.google.com/spanner/docs/schema-and-data-model
upvoted 17 times

[Removed] Highly Voted 3 years, 4 months ago
Should be C
https://cloud.google.com/spanner/docs/schema-and-data-model#choosing_a_primary_key
upvoted 8 times

zellck Most Recent 8 months, 1 week ago
Selected Answer: C
C is the answer.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 141 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 141
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by [deleted] at *March 22, 2020, 8:58 a.m.*

Comments

SteelWarrior Highly Voted 2 years, 10 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 25 times

AzureDP900 7 months, 1 week ago
D is right
upvoted 1 times

[Removed] Highly Voted 3 years, 4 months ago

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 142 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 142
Topic #: 1
[All Professional Data Engineer Questions]

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 Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes

B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project Most Voted

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 Stackdriver 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 Stackdriver Monitoring dashboard based on the custom metric

Hide Answer

Suggested Answer: *D*

Community vote distribution

B (100%)

by rickywck at *March 17, 2020, 2:02 p.m.*

Comments

rickywck Highly Voted 3 years, 4 months ago
Just tried and seems B can do the job ...
upvoted 18 times

shilpa 2 years, 6 months ago
Option B, refer to https://cloud.google.com/monitoring/api/metrics_gcp
upvoted 4 times

sumanshu Highly Voted 2 years, 1 month ago
Vote for B

A - Eliminated (it will not tell, anything about slots, it will show, which query scan how many data)
B - Correct METRIC given slots/allocated_for_project GA (which is used to tell Slots used by project) Number of BigQuery slots currently allocated for query jobs in the project.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 143 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 143
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

Hide Answer

Suggested Answer: *A*
Reference:
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

Community vote distribution
D (63%) A (38%)

by jvg637 at *March 18, 2020, 4:34 p.m.*

Comments

jvg637 Highly Voted 3 years, 4 months ago
Answer is D: 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.
upvoted 32 times

jsr2017 1 year, 11 months ago
It is A, with A you do not lose data. here does not say anything about mayor changes to ensure that the data is treated with the new algorithm
upvoted 5 times

sergio6 1 year, 10 months ago
Updating pipeline and changing the windowing or trigger strategies will not affect data that is already buffered or otherwise in-flight. So option A will cause data loss.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 144 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 144
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by [deleted] at *March 22, 2020, 9:36 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over 100TB
upvoted 22 times

[Removed] Highly Voted 3 years, 4 months ago
Correct - A
upvoted 10 times

vaga1 Most Recent 2 months, 4 weeks ago
Selected Answer: A
2,000,000,000,000,000 bytes = 2 Petabytes
20,000,000 bytes = 20 Megabytes

Once we do the math:
2 Petabytes / 20 Megabytes = 100,000,000 seconds forecasted to migrate the data.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 145 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 145
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *D*

Community vote distribution

A (100%)

by madhu1171 at *March 15, 2020, 5:07 p.m.*

Comments

madhu1171 Highly Voted 3 years, 4 months ago
A should be the answer
upvoted 34 times

[Removed] Highly Voted 3 years, 4 months ago
Answer: A
Description: Dataprep is used by non developers
upvoted 17 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 146 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 146
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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. Most Voted

D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones. Most Voted Most Voted

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.

Hide Answer

Suggested Answer: *BC*

Community vote distribution

AD (47%) CD (40%) 13%

by [deleted] at *March 22, 2020, 8:31 a.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be B C
upvoted 17 times

Sid19 Highly Voted 1 year, 7 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.

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 147 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 147
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (83%) A (17%)

by madhu1171 at *March 15, 2020, 5:11 p.m.*

Comments

mario_ordinola Highly Voted 2 years, 4 months ago
if someone are not sure that D is the answer, I suggest to don't take the exam
upvoted 38 times

madhu1171 Highly Voted 3 years, 4 months ago
D should be the answer
upvoted 23 times

AzureDP900 Most Recent 7 months, 1 week ago
D is right
upvoted 2 times

zellck 8 months, 1 week ago
Selected Answer: D

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 148 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 148
Topic #: 1
[All Professional Data Engineer Questions]

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. Most Voted

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.

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (100%)

by madhu1171 at *March 15, 2020, 5:16 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 months ago
Should be B.
upvoted 32 times

[Removed] Highly Voted 3 years, 4 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 22 times

[Removed] 3 years, 4 months ago
Correct : B
upvoted 12 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 149 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 149
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *C*

Community vote distribution

A (100%)

by madhu1171 at *March 15, 2020, 5:19 p.m.*

Comments

someshsehgal Highly Voted 2 years, 5 months ago
Correct A: A . Now it is feasible to provide table level access to user by allowing user to query single table and no other table will be visible to user in same dataset.
upvoted 37 times

jits1984 1 year, 9 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 7 times

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 150 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 150
Topic #: 1
[All Professional Data Engineer Questions]

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 Most Voted

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

Hide Answer

Suggested Answer: *A*

Community vote distribution

B (100%)

by rickywck at *March 17, 2020, 2:58 p.m.*

Comments

[Removed] Highly Voted 3 years, 4 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.