Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 54 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 54
Topic #: 1
[All Professional Data Engineer Questions]

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

A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.

B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.

C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

Show Suggested Answer

by jvg637 at *March 15, 2020, 4:17 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 5 years, 1 month ago

I'd go with B: real-time is requested, and the only scenario for real time (in the 4 presented) is the use of pub/sub with push.

upvoted 65 times

Tanzu 3 years, 2 months ago

B.
- for realtime pub/sub push is critical. pull creates latency. (eliminates D)
- process by event-time, not by process -time (eliminates D)

upvoted 6 times

godot 3 years, 1 month ago

no push avail: https://cloud.google.com/dataflow/docs/concepts/streaming-with-cloud-pubsub#streaming-pull-migration

upvoted 1 times

jin0 2 years, 2 months ago

The dataflow is designed for realtime processing. and this case should be needed to use dataflow because there is no option to order the data if not using dataflow. So D is answer I think

upvoted 2 times

AzureDP900 2 years, 4 months ago

Agree with B

upvoted 1 times

[Removed] 4 years ago

i would go with option B, Cause option D states "Give the bid for each item to the user in the bid event that is processed first" . The requirement is to get the first bid based on event time not processed first in dataflow.

upvoted 28 times

donbigi 2 years, 3 months ago

This approach is not ideal because it requires a custom endpoint to write the bid event information into Cloud SQL. This adds additional complexity and potential points of failure to the architecture, as well as adding latency to the processing of bid events, since the data must be written to both Pub/Sub and Cloud SQL. Additionally, it can be more challenging to ensure that bid events are processed in the order they were received, since the data is being written to multiple databases. Finally, using a single database to store bid events could limit scalability and availability, and can also result in slow query performance.

upvoted 3 times

Load full discussion...

Ganshank Highly Voted 5 years ago

D
The need is to collate the messages in real-time. We need to de-dupe the messages based on timestamp of when the event occurred. This can be done by publishing ot Pub-Sub and consuming via Dataflow.

upvoted 34 times

unnamed12355 2 years, 1 month ago

D isnt correct, Pub/sub can send messages out of order, it is no guaranty that the event with lowest timestamp will be processed first
B is correct

upvoted 4 times

Tanzu 3 years, 2 months ago

Yeap, that's why B is the right one. It has pub/sub push, more real time than pub/sub pull. You need to aware at some point , something has to be pulled which adds a latency.

upvoted 2 times

dcruzado Most Recent 2 months, 1 week ago

Selected Answer: B

Answer is B
This sentence invalidates D
"Give the bid for each item to the user in the bid event that is processed first."

upvoted 1 times

manikolbe 3 months, 4 weeks ago

Selected Answer: B

process by event-time, not by process -time (eliminates D

upvoted 3 times

Ronn27 4 months ago

Selected Answer: D

Writing directly to Cloud SQL in real time can cause bottlenecks as Cloud SQL is not designed for high-frequency, low-latency writes from multiple sources.

Answer D is right as Dataflow and Pubsub has the realtime capability

upvoted 1 times

DGames 4 months, 2 weeks ago

Selected Answer: D

Bid event time and Pull Subscription is important part.

upvoted 1 times

baimus 7 months, 2 weeks ago

Selected Answer: D

It feels like it depends what's actually in the dataflow pipeline. D I believe is the answer they intend, even if messages are pulled out of order.

upvoted 1 times

manel_bhs 9 months, 3 weeks ago

Selected Answer: D

While using Cloud Pub/Sub for real-time event streaming is a good choice, pushing events to a custom endpoint that writes to Cloud SQL introduces additional complexity.
Custom endpoints need to be maintained, and the process of writing to Cloud SQL might not be as efficient as using a purpose-built data processing service.

upvoted 1 times

Snnnnneee 10 months ago

Selected Answer: B

In D the user gets it where the data is ingested first. That can be wrong for a global auction solution

upvoted 2 times

yassoraa88 12 months ago

Selected Answer: D

This is the most suitable solution for the requirements. Google Cloud Pub/Sub can handle high throughput and low-latency data ingestion. Coupled with Google Cloud Dataflow, which can process data streams in real time, this setup allows for immediate processing of bid events. Dataflow can also handle ordering and timestamp extraction, crucial for determining which bid came first. This architecture supports scalability and real-time analytics, which are essential for a global auction system.

upvoted 2 times

teka112233 12 months ago

Selected Answer: D

the Answer should be D for the following
Real-time Processing
Centralized Processing
Winner Determination
also, B is unsuitable as While Pub/Sub can ingest data, Cloud SQL is a relational database not designed for real-time processing at this scale. Maintaining a custom endpoint adds complexity.

upvoted 3 times

I__SHA1234567 1 year, 1 month ago

Selected Answer: D

Google Cloud Pub/Sub is a scalable and reliable messaging service that can handle high volumes of data and deliver messages in real-time. By having each application server publish bid events to Cloud Pub/Sub, you ensure that all bid events are collected centrally.

Using Cloud Dataflow with a pull subscription allows you to process the bid events in real-time. Cloud Dataflow provides a managed service for stream and batch processing, and it can handle the real-time processing requirements efficiently.

By processing the bid events with Cloud Dataflow, you can determine which user bid first by applying the appropriate logic within your Dataflow pipeline. This approach ensures scalability, reliability, and real-time processing capabilities, making it suitable for handling bid events from multiple application servers.

upvoted 3 times

philli1011 1 year, 3 months ago

B should be the answer, because it writes the bid into Cloud SQL to a distributed system. This way the customer know if they get the bid or not, immediately

get the bid or not, immediately.
Also, push requests are faster than pull requests, hence they are better for realtime experience.

upvoted 1 times

arpana_naa 1 year, 4 months ago

Selected Answer: D

pub/sub for entry time stamp + event time
dataflow for processing and dataflow is better for real time

upvoted 1 times

Nandababy 1 year, 4 months ago

To accurately determine who bid first in a globally distributed auction application, utilizing a push mechanism instead of a pull mechanism is generally considered the more reliable approach.
B should be correct answer.

upvoted 1 times

Zepopo 1 year, 5 months ago

Selected Answer: B

key words is "single location in real time"

upvoted 3 times

rocky48 1 year, 5 months ago

Selected Answer: D

Answer : D
We need to de-dupe the messages based on timestamp of when the event occurred. This can be done by publishing ot Pub-Sub and consuming via Dataflow.
D sounds like a complete answer. B does not.

upvoted 2 times

Load full discussion...