Exam Professional Data Engineer All Questions
View all questions & answers for the Professional Data Engineer exam

Go to Exam

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 83 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 83
Topic #: 1
[All Professional Data Engineer Questions]

Flowlogistic Case Study -

Company Overview -
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -
The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -
Flowlogistic wants to implement two concepts using the cloud:
⇨ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
⇨ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -
Flowlogistic architecture resides in a single data center:

Homogenous architecture resides in a single data center.

✑ Databases

- 8 physical servers in 2 clusters
- SQL Server `" user data, inventory, static data
- 3 physical servers
- Cassandra `" metadata, tracking messages

10 Kafka servers `" tracking message aggregation and batch insert

✑ Application servers `" customer front end, middleware for order/customs

- 60 virtual machines across 20 physical servers
- Tomcat `" Java services
- Nginx `" static content
- Batch servers

✑ Storage appliances

- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) `" SQL server storage

Network-attached storage (NAS) image storage, logs, backups

✑ 10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

✑ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

✑ Build a reliable and reproducible environment with scaled panty of production.

✑ Aggregate data in a centralized Data Lake for analysis

✑ Use historical data to perform predictive analytics on future shipments

✑ Accurately track every shipment worldwide using proprietary technology

✑ Improve business agility and speed of innovation through rapid provisioning of new resources

✑ Analyze and optimize architecture for performance in the cloud

✑ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

✑ Handle both streaming and batch data

✑ Migrate existing Hadoop workloads

✑ Ensure architecture is scalable and elastic to meet the changing demands of the company.

✑ Use managed services whenever possible

✑ Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments

are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.
You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage

B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD

C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage

D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

E. Cloud Dataflow, Cloud SQL, and Cloud Storage

Show Suggested Answer

by [deleted] at *March 22, 2020, 6:13 p.m.*

Comments

Type your comment...

Submit

digvijay Highly Voted 3 years, 7 months ago
Seems like A..Data should ingest from multiple sources which might be real time or batch .
upvoted 25 times

navemula 2 years, 3 months ago
How is it possible to query in real time with option A. It needs Dataflow
upvoted 3 times

navemula 2 years, 3 months ago
To use Dataflow SQL it needs BigQuery
upvoted 2 times

mikey007 Highly Voted 3 years, 1 month ago
Repeated Question see ques 35
upvoted 11 times

awssp12345 2 years, 4 months ago
These exams make people over analyse. People who vote A earlier in 35 seem to be confused here.. haha
upvoted 1 times

StelSen 2 years, 9 months ago
Well Done mikey007, Many people have already answered as A.
upvoted 2 times

Kyr0 Most Recent 10 months, 1 week ago
Selected Answer: A
Answer is A
upvoted 1 times

cloudmon 12 months ago
Selected Answer: A
It's A

upvoted 1 times

ducc 1 year, 2 months ago

Selected Answer: A

A is the answer

upvoted 1 times

RRK2021 1 year, 8 months ago

ingest data from a variety of global sources - cloud pub/sub
process and query in real-time - cloud Dataflow
store the data reliably - Cloud Storage

upvoted 2 times

medeis_jar 1 year, 10 months ago

Selected Answer: A

PubSub (for global ingestion from multiple sources) + Dataflow (for process and query) + reliable (gcs).

upvoted 1 times

lifebegins 1 year, 11 months ago

Selected Answer: A

using Dataflow you can apply the propriety analytics and you can push the data in to Cloud storage

upvoted 1 times

gcp_k 2 years ago

Also read the technical requirements section. Not just the last 3 lines of the question.

When you do that, you'll know the answer is PubSub (for global ingestion) + Dataflow (for process and query) + reliable (gcs).

Answer is: A

upvoted 1 times

ManojT 2 years ago

Answer C:
Look the 3 requirement in the question "ingest data from a variety of global sources, process and query in real-time, and store the data reliably"
Ingest data from global sources: Pub-Sub
Process and Query in realtime: Cloud SQL
Store reliably: Cloud storage
I can understand Databflow is required in case you need to analyze and transform data but question does not refer it.

upvoted 1 times

cualquiernick 1 year, 4 months ago

Cloud SQL, is not suitable and efficient for storing real time data ingested from PUB/SUB, so A is the answer

upvoted 1 times

nguyenmoon 2 years, 1 month ago

Correct is A.
Kafka --> replace by PubSub, Streaming then Dataflow, store data reliably and not mention any other condition then Cloud Storage

upvoted 1 times

sumanshu 2 years, 4 months ago

Vote for 'A'

SQL - will not handle the volume

upvoted 2 times

daghayeghi 2 years, 7 months ago

Dataflow SQL cannot output to cloud storage:
https://cloud.google.com/dataflow/docs/guides/sql/data-sources-destinations
but the main problem is that Cloud SQL can't do process, then response is A or C.

upvoted 1 times

kino2020 3 years, 1 month ago

A
I don't expect this question to come up, but if I had to write the answer, it would be A.
The problem statement "Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.
As it says, "we cannot determine the data volume", but it doesn't say that we can't calculate it either.

Requirement definition: The system must be able to
ingest data from a variety of global sources

ingest data from a variety of global sources
process and query in real-time
Store the data reliably.

It says above, if you look at the Google page.

Logging to multiple systems. for example, a Google Compute Engine instance can write logs to a monitoring system, to a database for later querying, and so on.
https://cloud.google.com/pubsub/docs/overview#scenarios

stream processing with Dataflow
https://cloud.google.com/pubsub/docs/pubsub-dataflow?hl=en-419

The answer is A, since it is stated above.

upvoted 4 times

vakati 3 years, 1 month ago

A. SQL queries can be written in Dataflow too.
https://cloud.google.com/dataflow/docs/guides/sql/dataflow-sql-intro#running-queries

upvoted 3 times

aleedrew 3 years ago

Dataflow SQL cannot output to cloud storage only BigQuery...so I am confused on this one.

upvoted 2 times

Jay3244 2 years, 8 months ago

https://cloud.google.com/pubsub/docs/pubsub-dataflow... It is possible to load the data to Cloud Storage. Can refer to above docs.

upvoted 1 times

daghayeghi 2 years, 7 months ago

he said correct, DataflowDataflow SQL cannot output to cloud storage:
https://cloud.google.com/dataflow/docs/guides/sql/data-sources-destinations

upvoted 1 times

Ral17 2 years, 2 months ago

Answer should be C then?

upvoted 2 times

kuntal8285 3 years, 1 month ago

should be E

upvoted 1 times

Tanmoyk 3 years, 1 month ago

Should be A ...data need to feed to the propriority system and for that dataflow is required.

upvoted 2 times

Load full discussion...

EXAMTOPICS

Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses

© 2024 ExamTopics