

 [Google Discussions](#)

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 35 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 35

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 into. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

⇒ Databases

8 physical servers in 2 clusters

- SQL Server `` user data, inventory, static data

3 physical servers

- Cassandra `` metadata, tracking messages

10 Kafka servers `` tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

- Tomcat `` Java services

- Nginx `` static content

- Batch servers

⇒ Storage appliances

- iSCSI for virtual machine (VM) hosts

- Fibre Channel storage area network (FC SAN) `` SQL server storage

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

⇒ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

⇒ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

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

⇒ Aggregate data in a centralized Data Lake for analysis

⇒ Use historical data to perform predictive analytics on future shipments

⇒ Accurately track every shipment worldwide using proprietary technology

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

⇒ Analyze and optimize architecture for performance in the cloud

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

Technical Requirements -

⇒ Handle both streaming and batch data

⇒ Migrate existing Hadoop workloads

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

⇒ Use managed services whenever possible

⇒ Encrypt data flight and at rest

⇒ Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

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

are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.

Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

- A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
- B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Show Suggested Answer

by [jvg637](#) at March 15, 2020, 1:18 p.m.

Comments



Type your comment...

Submit

  **jvg637** Highly Voted 4 years, 7 months ago

I would say A.
I think Pub/Sub can't directly send data to Cloud SQL.

   upvoted 38 times

  **[Removed]** Highly Voted 4 years, 7 months ago

Answer: A

   upvoted 15 times

  **billalltf** Most Recent 5 months, 2 weeks ago

Selected Answer: A

A is right answer

   upvoted 1 times

  **JOKKUNO** 11 months ago

Given the requirements for ingesting data from global sources, processing and querying in real-time, and storing the data reliably for the real-time inventory tracking system, the most suitable combination of Google Cloud Platform (GCP) products is:

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

Explanation:

Cloud Pub/Sub: It is a messaging service that allows you to asynchronously send and receive messages between independent applications.

Cloud Dataflow: It can handle both streaming and batch data, making it suitable for real-time processing of data from various sources.

Cloud Storage: Cloud Storage can be used to store the processed and analyzed data reliably. It provides scalable, durable, and globally accessible object storage, making it suitable for storing large volumes of data.

   upvoted 1 times

  **rtcpost** 1 year ago

Selected Answer: A

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

Here's why this combination is suitable:

Cloud Pub/Sub: It is used for ingesting real-time data from various global sources. It's a messaging service that can handle large volumes of data and is highly scalable.

Cloud Dataflow: It's a stream and batch data processing service that allows you to process and analyze the data in real-time. It can take data from Pub/Sub and perform transformations or aggregations as needed.

Cloud Storage: It provides reliable storage for the data. You can store the processed data in Cloud Storage for further analysis, and it is a scalable and durable storage solution.

Option B is not ideal because Local SSDs are not a suitable storage option for persisting data that needs to be reliably stored. Option C includes Cloud SQL, which is not typically used for ingesting and processing real-time data. Option D includes Cloud Load Balancing, which is not relevant to the use case of ingesting and processing data for the inventory tracking system.

   upvoted 2 times

  **Vipul1600** 1 year, 3 months ago

Since Cloud SQL is fully managed service & Dataflow is serverless hence we should opt for dataflow as it is thumb rule for google that we should choose serverless product over fully managed service.

   upvoted 1 times

  **Mathew106** 1 year, 3 months ago

Selected Answer: A

The technical requirements mention that the pipeline should handle both streaming and batch data. The solution should include DataFlow and not Cloud SQL. the answer is A.

   upvoted 2 times

  **niketd** 1 year, 8 months ago

Selected Answer: A

Pub/Sub to scale streaming data, Dataflow to processes both structured and unstructured data and cloud storage to store common data

   upvoted 1 times

  **PolyMoe** 1 year, 9 months ago

Selected Answer: A

Option B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD is not a good option as Local SSD is not a scalable solution and could not handle large amount of data

Option C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage is not a good option as Cloud SQL is a relational database and is not suitable for real-time processing and querying large amounts of data

Option D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage is not a good option as Cloud Load Balancing is used for distributing traffic across multiple instances, it doesn't handle data processing and storage.

   upvoted 1 times

  **PolyMoe** 1 year, 9 months ago

Selected Answer: A

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage is the best combination of GCP products for the use case described. Cloud Pub/Sub can be used to ingest data from a variety of global sources, as it allows for easy integration with external systems through its publish-subscribe messaging model.

Cloud Dataflow can be used to process and query the data in real-time, as it is a fully managed service for creating data pipelines that can handle both batch and streaming data.

Cloud Storage can be used to store the data reliably, as it is a fully managed object storage service that can handle large amounts of data and is highly durable and available.

   upvoted 1 times

  **jkh_goh** 1 year, 9 months ago

Selected Answer: A

Answer is A. Cloud Dataflow for batch + streaming, Cloud Pub/Sub for streaming ingestion, Cloud Storage for long term data storage.

   upvoted 1 times

  **Jay_Krish** 1 year, 11 months ago


Are scenario based questions still in the latest exam?? Are these still relevant?

   upvoted 2 times

  **kastuarr** 2 years ago

Selected Answer: C



Existing inventory data is in SQL, data ingested from Kafka will need to update inventory at some point. Existence of SQL in current estate indicates SQL must be present in the Cloud estate

   upvoted 2 times

  **DhamsI** 2 years, 1 month ago

This site make me feel that it intends to make users to be involved in discussion by suggesting wrong answer

   unvoted 9 times

  **Megmang** 2 years, 2 months ago

Selected Answer: A

Answer is clearly option A.

   upvoted 3 times

  **[Removed]** 2 years, 3 months ago

why are there so many incorrect answers? it's so hard to study this way

   upvoted 6 times

  **ratnesh99** 2 years, 3 months ago

Answer A : because Cloud Sql not suitable for Global

   upvoted 1 times

[Load full discussion...](#)



Platform

> [Home](#)

> [Examtopics PRO](#)

> [All Exams](#)

> [Training Courses](#)

