EXAMTOPICS

MENU

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 34 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 34

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:

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 wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to

BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

A. Store the common data in BigQuery as partitioned tables.

B. Store the common data in BigQuery and expose authorized views.

C. Store the common data encoded as Avro in Google Cloud Storage.

D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

Show Suggested Answer

by ducc at *Sept. 3, 2022, 12:44 a.m.*

Comments

Type your comment...

Submit

rtcpost Highly Voted 1 year, 6 months ago

Selected Answer: C

C. Store the common data encoded as Avro in Google Cloud Storage.

This approach allows for interoperability between BigQuery and Hadoop/Spark as Avro is a commonly used data serialization format that can be read by both systems. Data stored in Google Cloud Storage can be accessed by both BigQuery and Dataproc, providing a bridge between the two environments. Additionally, you can set up data transformation pipelines in Dataproc to work with this data.

upvoted 6 times

iooj Most Recent 9 months, 1 week ago

Selected Answer: C

in BigQuery we can use BigLake tables based on Avro for historical data, and Spark stored procedures

upvoted 1 times

dhvanil 10 months, 3 weeks ago

Data lake,fully managed, data analytics. Stores structured and unstructured data are keywords,so answer is GCS, OPTION C

upvoted 1 times

JOKKUNO 1 year, 5 months ago

Given the scenario described for Flowlogistic's requirements and technical environment, the most suitable option for storing common data that is used by both Google BigQuery and Apache Hadoop/Spark workloads is:

C. Store the common data encoded as Avro in Google Cloud Storage.

upvoted 4 times

nescafe7 1 year, 9 months ago

Selected Answer: D

To simplify the question, Apache Hadoop and Spark workloads that cannot be moved to BigQuery can be handled by DataProc. So the correct answer is D.

upvoted 3 times

Mathew106 1 year, 9 months ago

Selected Answer: B

B is the right answer. Common data will lie in BigQuery but will be accessible via the views with SQL in Hadoop workloads.

upvoted 2 times

midgoo 2 years, 2 months ago

Selected Answer: B

C should be the correct answer. However, please note that Google just released the BigQuery Connector for Hadoop, so if they ask the same question today, B will be the correct answer.
A could be correct too, but I cannot see why it has to be partitioned

upvoted 4 times

res3 1 year, 10 months ago

If you check the https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery, it unloads the BQ data to GCS, utilizes it, and then deletes it from the GCS. Storing common data twice (at BQ and GCS) will not be the best option compared to 'C' (using GCS as the main common dataset).

upvoted 3 times

korntewin 2 years, 3 months ago

Selected Answer: C

I would vote for C as it can be used for analysis with Bigquery. Furthermore, Hadoop workload can also be transferred to dataproc connected to GCS.

upvoted 1 times

DGames 2 years, 4 months ago

Selected Answer: B

Answer B look ok , because in question they want to store common data which can use by both workload, and using big query and primary analytical tool that would be best option and easy to analysis common data.

upvoted 1 times

kelvintoys93 2 years, 5 months ago

"Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data" - BigQuery cant take unstructured data so A and B are out.
Storing data in HDFS storage is never recommended unless latency is a requirement, so D is out.

That leaves us with GCS. Answer is C

upvoted 3 times

tunstila 2 years, 4 months ago

I thought you can now store unstructured data in BigQuery via the object tables announced during Google NEXT 2022... If that's possib;e, does that make B a better choice?

upvoted 1 times

drunk_goat82 2 years, 5 months ago

Selected Answer: C

BigQuery can use federated queries to connect to the avro data in GCS while running spark jobs on it. If you duplicate the date you have to manage both data sets.

upvoted 2 times

wan2three 2 years, 5 months ago

A
They wanted BigQuery. And connector is all you need to perform Hadoop or spark. Hadoop migration can be done using dataproc.

upvoted 1 times

wan2three 2 years, 5 months ago

Also apparently they want all data at one place and want bigQ

upvoted 2 times

gudiking 2 years, 5 months ago

Selected Answer: C

C as it can be used as an external table from BigQuery and with the Cloud Storage Connector it can be used by the Spark workloads (running in Dataproc)

upvoted 1 times

solar_maker 2 years, 5 months ago

Selected Answer: C

C, as both capable of AVRO, but the customer does not know what they want to do with the data yet.

upvoted 1 times

Leelas 2 years, 6 months ago

Selected Answer: D

In Technical requirements it Was clearly mentioned that they need to Migrate existing Hadoop Cluster for which Data Proc Cluster is a replacement.

upvoted 1 times

vishal0202 2 years, 7 months ago

C is ans...avro data can be accessed by spark as well

upvoted 4 times

ducc 2 years, 8 months ago

Selected Answer: C

The answer is C

upvoted 3 times

EXAMTOPICS

Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses