Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 50 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 50
Topic #: 1
[All Professional Data Engineer Questions]

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100
TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID).
However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis

B. HBase

C. MySQL

D. MongoDB

E. Cassandra

F. HDFS with Hive

Show Suggested Answer

by jvg637 at *March 15, 2020, 1:59 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 4 years, 1 month ago

BDE. Hive is not for NoSQL

upvoted 39 times

sergio6 2 years, 7 months ago

Redis is also NoSQL

upvoted 2 times

vholti 2 years, 6 months ago

Redis is limited to 1 TB capacity quota per region. So it doesn't satisfy the requirement.
https://cloud.google.com/memorystore/docs/redis/quotas

upvoted 3 times

ckanaar 7 months, 2 weeks ago

Memorystore, Google's managed Redis service is. But OS Redis is not. Though it is hard to find a 100GB RAM machine

upvoted 1 times

awssp12345 Highly Voted 2 years, 10 months ago

Answer is BDE -
A. Redis - Redis is an in-memory non-relational key-value store. Redis is a great choice for implementing a highly available in-memory cache to decrease data access latency, increase throughput, and ease the load off your relational or NoSQL database and application. Since the question does not ask cache, A is discarded.
B. HBase - Meets reqs
C. MySQL - they do not need ACID, so not needed.
D. MongoDB - Meets reqs
E. Cassandra - Apache Cassandra is an open source NoSQL distributed database trusted by thousands of companies for scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data.
F. HDFS with Hive - Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. HIVE IS NOT A DATABSE.

upvoted 32 times

[Removed] 1 year, 2 months ago

HDFS is. Hadoop Distributed File System. HDFS is storage and HIVE is for processing.

upvoted 1 times

sravi1200 Most Recent 4 months, 2 weeks ago

Selected Answer: BDE

Option A: Redis cannot handle large scale data it is NOSQL db to store small amount of key value pairs,
Option B: HBase NOSQL db built on Hadoop does not support ACID Properties. Correct answer
Option C: Mysql Does not store telemetry IOT data. Mysql is a relational database structured data only stored.
Option D, E: NOSQL Databases, Option F: HDFS with hive used for batch processing not real time streaming data.
Option

upvoted 1 times

musumusu 1 year, 2 months ago

BDE
Faster Database are NoSql db than SQL, Cassandra is the fastest one in market now than Hbase and then others, in given list MongoBD

upvoted 1 times

MisuLava 1 year, 8 months ago

"Which three databases meet your requirements? "
Hive is not a database server.
HBase, Mongo and Cassandra are and meet the criteria.
BDE is the right answer

upvoted 1 times

sraakesh95 2 years, 3 months ago

Selected Answer: BDE

@hendrixlives

upvoted 1 times

medeis_jar 2 years, 4 months ago

Selected Answer: BDE

as explained by hendrixlives

upvoted 1 times

hendrixlives 2 years, 4 months ago

Selected Answer: BDE

BDE:
A. Redis is a key-value store (and in many cases used as in-memory and non persistent cache). It is not designed for "100TB per year" of highly available storage.
B. HBase is similar to Google Bigtable, fits the requirements perfectly: highly available, scalable and with very low latency.
C. MySQL is a relational DB, designed precisely for ACID transactions and not for the stated requirements. Also, growth may be an issue.
D. MongoDB is a document-db used for high volume data and maintains currently used data in RAM, so performance is usually really good. Should also fit the requirements well.
E. Cassandra is designed precisely for highly available massive datasets, and a fine tuned cluster may offer low latency in reads. Fits the requirements.
F. HDFS with Hive is great for OLAP and data-warehouse scenarios, allowing to solve map-reduce problems using an SQL subset, but the latency is usually really high (we may talk about seconds, not milliseconds, when obtaining results), so this does not complies with the requirements.

upvoted 14 times

MaxNRG 2 years, 5 months ago

Selected Answer: BEF

Very strange question, seems outdated and irrelevant to me as it doesn't contain any GCP products :)

Anyway, I would choose BEF.
Redis is in-memory key value, not good
HBase yes, excelent case for linear growth and a column-oriented database
mysql not good, too big and no need for transactionality
Mongodb, document db with flexible schema ??
Yes Cassandra, good use case
Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis.
https://www.wikiwand.com/en/Apache_Hive

upvoted 1 times

hendrixlives 2 years, 4 months ago

Latency in Hive is usually quite high, and one of the requirements is "low latency"

upvoted 2 times

MaxNRG 2 years, 3 months ago

good point!

upvoted 1 times

MaxNRG 2 years, 3 months ago

agreed on BDE

upvoted 2 times

anji007 2 years, 6 months ago

Ans: B, D and E

upvoted 2 times

sumanshu 2 years, 10 months ago

vote for BDE

upvoted 2 times

BhupiSG 3 years, 1 month ago

BEF
B: HBASE is based upon BigTable
E: Cassandra is low latency columnar distributed database like BigTable
F: HDFS is low latency distributed file system and Hive will help with running the queries

upvoted 2 times

Manue 3 years ago

Hive is not for low latency queries. It is for analytics.

upvoted 5 times

daghayeghi 3 years, 1 month ago

BDE:
These are NoSQL DB, Hive is not for NoSQL.

upvoted 2 times

Rayleigh 3 years, 2 months ago

The answer is ADE, the statement says they require a NoSQL with high availability and low latency, they do not require consistency.
C. it is not NoSQL.
F. it is not NoSQL.
B. it is NoSQL but focused on strong consistency and based on HDFS, you need HDFS for Hbase.
Therefore the answer is ADE

upvoted 1 times

daghayeghi 3 years, 2 months ago

BDF:
Redis and Cassandra have only Rowkey and couldn't be indexed, and MySQL isn't NoSQL, Then B D and E is correct answer.

upvoted 1 times

naga 3 years, 2 months ago

Correct BDE

upvoted 3 times

apnu 3 years, 4 months ago

it should be BDE because Hive is a sql based datawarehouse , it is not a nosql DB

upvoted 3 times

Load full discussion...