1. Difference Between Data Warehouse and Database

Sunday, August 11, 2024 11:14 AM

Data Warehouse: A centralized repository designed to store large volumes of structured and unstructured data from various sources, optimized for query and analysis. It supports complex queries and is used for reporting and analytics.

Database: A system designed to store and manage data in a structured format, optimized for transactional processing (CRUD operations). It supports daily operations and typically handles a single application.

2. Difference Between Data Warehouse and Data Mart

Sunday, August 11, 2024 12:08 PM

Data Warehouse: A large, centralized repository that stores integrated data from various sources, used for enterprise-wide reporting and analysis.
Data Mart: A subset of a data warehouse, usually focused on a specific business area or department. It is smaller in scope and more targeted for specific business needs.

3. Difference Between OLTP vs. OLAP

Sunday, August 11, 2024 12:08 PM

OLTP (Online Transaction Processing): Optimized for transaction-oriented tasks such as insert, update, and delete. It handles a large number of short online transactions and is used in day-to-day operations.

OLAP (Online Analytical Processing): Optimized for complex queries and analysis, involving large volumes of data. It is used for business intelligence and supports decision-making processes.

4. Why Hive Metadata is Stored in SQL?

Sunday, August 11, 2024 12:09 PM

Hive Metadata: Stores schema information, tables, partitions, columns, etc. SQL databases are used because they provide transaction management, reliability, and support for complex queries, which are essential for managing the metadata.

5. Which SQL is the Default Database for Hive?

Sunday, August 11, 2024 12:10 PM

Default Database: Apache Derby is the default SQL database for Hive in standalone mode. However, in production environments, MySQL or PostgreSQL is commonly used for better scalability and performance.

6. What is a Managed Table?

Sunday, August 11, 2024 12:12 PM

Managed Table: A Hive table where Hive manages both the metadata and the data. When you drop a managed table, Hive deletes both the metadata and the data.

7. What is an External Table?

Sunday, August 11, 2024 12:13 PM

External Table: A Hive table where Hive only manages the metadata, while the data is stored externally (e.g., in HDFS). Dropping an external table will only remove the metadata, leaving the data intact.

8. When Do We Use an External Table?

Sunday, August 11, 2024 12:14 PM

Use Cases for External Table: When the data is managed by another process, needs to be shared between multiple applications, or should not be deleted when the table is dropped. It's also used when the data resides outside of the Hive warehouse directory.

9. Difference Between Managed and External Table

Sunday, August 11, 2024 12:14 PM

Managed Table: Hive controls both data and metadata. Dropping the table deletes both.

External Table: Hive controls only the metadata. Dropping the table does not delete the data.

10. What Happens if You Don’t Provide Location to an External Table?

Sunday, August 11, 2024 12:16 PM

Default Behavior: If no location is provided for an external table, Hive will assume the data resides in the default Hive warehouse directory, but it won’t delete the data upon table deletion.

11. Performance optimization in hive?

Sunday, August 11, 2024 12:16 PM

Optimization Techniques:
Partitioning: Dividing large tables into smaller, more manageable pieces.
Bucketing: Grouping data into buckets based on a hash function.
Indexing: Creating indexes to speed up data retrieval.
File Formats: Using optimized file formats like ORC or Parquet.
Vectorization: Processing multiple rows simultaneously to reduce CPU cycles.

12. Explain Partitioning? Where Did You Use It with Example

Sunday, August 11, 2024 12:18 PM

Partitioning: The technique of dividing a table into smaller parts based on the value of a specific column, such as date or region. This allows for more efficient querying by only scanning the relevant partitions.

Example: In an e-commerce application, a sales table can be partitioned by date. This allows queries on specific dates to be faster, as only the relevant partitions are scanned.

13. Explain Bucketing? Where Did You Use It with Example

Sunday, August 11, 2024 12:19 PM

Bucketing: A method of dividing data into equal-sized buckets based on a hash of the bucket column. This is used to ensure even data distribution and improve join operations.

Example: If a table is bucketed by the customer ID, all rows with the same customer ID will be stored in the same bucket, optimizing joins on this column.

14. Explain Transactional Table and Implement Merge to Load Incremental Data

Sunday, August 11, 2024 12:20 PM

Transactional Table: A Hive table that supports ACID operations like insert, update, and delete. It allows fine-grained modifications to data, which is essential for handling updates and incremental loads.

```
MERGE INTO target_table AS T
USING source_table AS S
ON T.id = S.id
WHEN MATCHED THEN UPDATE SET T.column1 = S.column1, T.column2 =
S.column2
WHEN NOT MATCHED THEN INSERT (id, column1, column2) VALUES (S.id,
S.column1, S.column2);
```

Use Case: Useful for loading incremental data into a table, where existing records are updated, and new records are inserted.