

24 –Distributed OLAP Databases

1. Star Schema vs. Snowflake Schema

A. Star schema

- i. fact table and one level dimension table.
- ii. Query is more faster because less join operations are needed

B. Snowflake schema

- i. Fact table and one or more level dimension table
- ii. Take up less storage space

2. Execution Models

A. Push vs. Pull

- i. Push : send query to node that has data (shared nothing)
- ii. Pull : bring data to node that executes query (shared disk)

B. Query Fault Tolerance

- i. Save snapshots of intermediate result of query to redo the query fast if it fails

3. Query Planning

A. Physical operators

- i. Generate single query plan, and break it up into partition-specific fragments.
- ii. Most common solution

B. SQL

- i. Rewrite query into partition-specific query
- ii. Local optimization is possible

4. Distributed Join Algorithms

A. Scenarios

- i. One table is replicated every nodes
so we can join local data in every nodes, and combine results
- ii. Every tables are partitioned on the join key
so we can join local data in every nodes, and combine results
- iii. Every tables are partitioned on the different keys ranges
if one table is small, DBMS broadcasts that table to all nodes.
- iv. Every tables are not partitioned on the join keys
DBMS copies the table to shuffle them and join the tables.

B. Semi-Join

- i. Result : join key column from left table
- ii. Copying entire records is expensive,
Semi-Join need to send just join key, not entire records

5. Cloud Systems

A. Approaches

- i. Managed DBMSs : no significant changes of DBMS
- ii. Cloud-Native DBMS : cloud-specific system

B. Severless databases

- i. If tenants become idle, system evicts them and save there buffer pool to disk
- ii. Others can re-use evicted tenants' node, and all tenants can pay just for storage