Impact of using Snowflake Schema and Bitmap Index on Data Warehouse Querying

Mohammed Benjelloun  
Laboratory Modeling and Information Theory Abdilmalek Essaadi University, Tétouan, Morocco

Mohamed El Merouani  
Laboratory Modeling and Information Theory Abdilmalek Essaadi University, Tétouan, Morocco

El Amin A. Abdelouarit  
Laboratory Modeling and Information Theory Abdilmalek Essaadi University, Tétouan, Morocco

ABSTRACT
Designing a data warehouse is the most important task realized by the designer; going to production is considered a remarkable phase to test the data warehouse performance based on it design.

Is known that hardware has developed in term of memory volume and speed, storage… but to minimize the execution time and the storage space still being the main objective for data warehousing specialists.

This paper shows the advantage of using bitmap index and partitioned fact tables in big data warehouses volumes based on a snowflake schema based on query execution time.

General Terms
Business Intelligence, Data Warehouse.

Keywords
Data Warehouse, DBMS, Indexes, Business Intelligence

1. INTRODUCTION
The performance of q data warehouse is considered the main preoccupation for designers. It is based in first on the query execution time and the data access complexity.

For that, the physical design is considered the most important task for data warehouse designers, including how to improve access to this data.

Toward minimizing the query execution time, the data warehouse design type should be optimized (how many joins in tables). For that, we will try in improving query execution query time using snowflake and bitmap index in comparison with the star schema.

2. ROLAP SYSTEMS
It uses a relational representation of a data cube, constituted from fact and dimension tables. The fact table contains in their attribute values of activity results and a foreign key to each dimension. The ROLAP has as advantage the use of existing databases which reduces it implementation cost [7][8].

In our study we compare between two schemas: Star and Snowflake schema.

3. SNOWFLAKE SCHEMA
The snowflake schema reflects the hierarchies associated with each dimension.

Each dimension table is split into a plurality of hierarchies. This diagram normalized dimensions, reducing the size of each of the connections, thus allowing to formalize the concept of hierarchy within a dimension [7][8].

Tables representing the finest hierarchy are directly linked to the fact table. The tables representing other hierarchies are linked to each other according to their level in the hierarchy.

Figure 2 shows a snowflake schema.

4. HYPOTHESIS
The bitmap index permits to a fast access to data [1][2][3][4][6], but the star schema is not efficient when the fact table contains a large data [5]: Less number of foreign keys and hence shorter query execution time (faster), to make it real, we look forward the use of partitioned tables, this will reduces the data volume in table: data will be distributed in many tables, so it conduces to less foreign key.

To do that, we have to divide the fact table in many partitions based on a dimension, and would be easier to take the time dimension to partition it (in months, years, decade…)

So I demonstrate that by using bitmap index and filtering and partitioned tables in a star schema is more efficient than using the snowflake one based on star join query optimization.

Note that we take two similar environment with SQL Server as RDBMS (Relational Data Base Management System).

5. ANALYSIS AND RESULTS
To satisfy our study we created the following big Data warehouse volume sample Schemas:
- Star schema: 1 fact table 3 dimension
- Snowflake schema: 1 fact table 3 dimension and 2 sub dimensions under dimension d1.
- 10 million row inserted in both fact tables with same data (by respecting it dimensions)
To get a deep analyze we will use a query for both data warehouse schema: star join query (Table 2) and snowflake join query (Table 3).

5.1 Technical specification of lab environment
This study has been done in similar technical environment that we present in the following table: (Table 1)

<table>
<thead>
<tr>
<th>Physical Memory</th>
<th>Storage Disk</th>
<th>Operation System</th>
<th>RDBMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 Gb</td>
<td>8 TB</td>
<td>Win 2008 Server</td>
<td>MS SQL SERVER</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2008 R2</td>
</tr>
</tbody>
</table>

The huge resources that we offer has for objective to make difference between old decision systems with lower resources and the new decision systems with high technology, high resources.

Table 1. Technical specification of lab environment

Table 2. Star Join Query Used

```
SELECT d1.att1, d3.att2, d1.att1, SUM(f.measure1) FROM Fact1 f 
INNER JOIN dim1 d1 on d1.dimkey1 = f.dimkey1 
GROUP BY d1.att1
```

Table 3. Snowflake Join Query Used

```
SELECT fact2.fact2_att1, 
fact2.fact2_att2,fact2.fact2_att3, 
d11.d11_att1, fact2.fact2_att3, 
fact2.fact2_att1 
FROM d11 LEFT JOIN 
ON d11.d11_att1 = d1.d11_att1 LEFT JOIN 
fact2 ON d11.d11_att1 = fact2.fact2_att1 
GROUP BY fact2.fact2_att1;
```

5.2 Used queries
These two queries are getting the same data in different way.

Our goal is to stress these 2 data warehouses scheme before and after applying the bitmap index and partitioning table to star schemas, this application will be called in our work as optimization tasks.

The comparison will be based on execution time and memory occupation. We focus on these two factors as per the execution time depends on memory consumption.

5.3 Results before applying the optimization tasks
Before applying our predictions, we will test the both queries on a standard star and snowflake schema we get results in the following table.

<table>
<thead>
<tr>
<th>Query</th>
<th>Memory Occupied by process</th>
<th>Real time Query execution in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Join + Bitmap</td>
<td>48%</td>
<td>40 min</td>
</tr>
<tr>
<td>Snowflake Join + Bitmap</td>
<td>69%</td>
<td>59 min</td>
</tr>
</tbody>
</table>

After analyzing these results, we may say that:
- Snowflake schema is better when dimension table is relatively big in size, because it reduces space.
- Query execution is slower when using star schema in a large data case.
- As per data redundancy, the star join occupies more memory than the snowflake.

5.4 Results after applying the optimization tasks
After partitioning table in star schema per time dimension d1 (per year) and added bitmap index, we got the following results:

<table>
<thead>
<tr>
<th>Query</th>
<th>Memory Occupied by process</th>
<th>Real time Query execution in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Join + Partitionned Tables + Bitmap</td>
<td>45%</td>
<td>33 min</td>
</tr>
<tr>
<td>Snowflake Join + Partitionned Tables + Bitmap</td>
<td>40%</td>
<td>21 min</td>
</tr>
</tbody>
</table>

After applying modification we found that in the big data warehouses case, the snowflake schema is recommended to use.

5.5 Results Analysis
Based on these results, the factors of memory occupation, and time execution size become efficient after partitioning tables and adding bitmap index because of:
- Less data when partitioning per table
- Many level of dimensions
- Bitmap index efficiency when the three is large data and less distinct values.

So, partitioning tables and using bitmap index is a tuning task that can qualify better the use of snowflake schema on big data warehouses volume.

6. CONCLUSION AND FUTURE WORKS
The use of the snowflake schema in our data warehouse has as objective is to conserve it advantages like fast access to data by many level of dimensions. Today, the majority of data warehouses that are used had the objective to analyze the organization production, commercial, finance activity. This means that in the most of cases data warehouses are more efficient if designed as data mart per activity. But the work that we present is to prove that snowflake schema can be used even if there is large data, and the bitmap index and partitioning fact tables are playing the role to create “data marts” time oriented.

As future work, we will discuss what a is given by this study on decision dashboard and business reporting, as they are considered the main window from where management can have a transparent look of what is going on their organization.

7. REFERENCES


