Research Article [Srivastava,1(4): Jun., 2012]
ISSN: 2277-9655

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

An Optimized algorithm to select the appropriate Schema in Data Warehouses

Rahul Kumar Shrivastava[*1], **Mehul Mahrishi**[2], **Pooja Parnami**[3]

[*1,3]Dept. Of Computer Science & Engg., VIT, Jaipur

[2]Dept. Of Computer Science & Engg., SKIT, Jaipur

shrivastava.rahul07@gmail.com

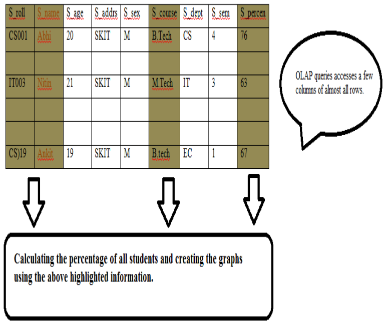
Abstract

The successful organization is the result of the successful decisions made by the top management. These are the collaborated decisions and require a lump sum data or the consolidated view of organization, which is provided by data warehousing system. These systems are used as an organization repository to support strategic decision-making. Data schema represents the arrangement of fact table and dimension tables and the relations between them. In data warehouse development, selecting a right and appropriate data schema (Snowflake, Star, Star Cluster etc.) has an important Impact on performance and usability of the designed data warehouse. One of the problems that exist in data warehouse development is lack of a comprehensive and sound selection framework to choose an appropriate schema for the data warehouse at hand by considering application domain-specific conditions. This research work, presents a stepwise algorithm for schema selection, which solves the problem of choosing right schema for a data warehouse. The main selection criteria are query type, attribute type, dimension table type and existence of index. Authors also tried to described efficient way of answering queries that are coming from multiple classes of users.

Keyword:- Data warehouse, query execution, schema, decision-making

I. Introduction

The successful organization is the result of the successful decisions made by the top management. Decision maker must make effective decisions in time, for survival, to get competitive advantages and to increase profitability of an organization. In the mid of 1990's a new era of data management arises which was query specific and involves large complex data volumes. Example of such query specific DBMS are OLAP and Data mining.



S_roll	S_name	S_age	S_addrs	S_sex	S_course	S_dept	S_sem	S_percen
CS001	Abhi	20	SKIT	M	B.Tech	CS	4	76
IT003	Nitin	21	SKIT	M	M.Tech	IT	3	63
CS)19	Ankit	19	SKIT	M	B.tech	EC	1	67

OLAP queries accesses a few columns of almost all rows.
Calculating the percentage of all students and creating the graphs using the above highlighted information.

Figure 1: OLAP Access

OLAP tool summarizes the data from large data volumes and represents the query into results using 2-D or 3-D graphics to Visualize the answer. The OLAP query is like "Give the % comparison between the marks of all students in B. Tech and in M. Tech". The answer to this query would be generally in the form of graph or chart. Such 3-D and 2-D visualization of data is called as "Data Cubes".

Figure 1 represents the access pattern of OLAP, which requires a few attributes to be process and access to huge volume of data. It must be noted that the execution of number of queries per second in OLAP is very less in comparison to OLTP. Currently, Data warehouses are used as an organizational repository to support decision making.

II. Literature Review

Data warehouse is centralized data repository maintained separately from organization's operational databases to help organization in corporate decision-making process. William Inmon has described data warehouse as "A subject-oriented, integrated, time-variant, non volatile collection of data in support of management decisions"[1], [2] "Data warehouse is a set of materialized views over data sources [3], [4], [5] Ralph Kimball ET. Al. Defined "A data warehouse is a copy of transaction data specially structured for query and analysis" [6]. "A data warehouse combines various data sources into a single source for end user access. End user can perform ad hoc querying, reporting, analysis, data mining and visualization of warehouse information. The goal of data warehouse is to establish a data repository that makes operational data accessible in a form that is readily acceptable for decision support and other application" [7].

A data warehouse is a logical collection of information gathered from many different operational databases used to create business intelligence that supports business analysis activities and decision-making tasks. It is used for providing the basic infrastructure for decision making by extracting, cleansing and storing huge amount of data. Data warehouses support business decisions by collecting, consolidating, and organizing data for reporting and analysis with tools such as online analytical processing (OLAP) and data mining.

The normal size of data warehouse varies from hundreds of gigabytes to terabytes. Different scans, joins, and aggregates are performed while querying the data warehouse. The queries on data warehouse are ad hoc and multi-faced. Throughput of query determines the success of data warehousing project. The query response time is also important factor in data warehouse success. The allocation of facts and dimensions in a certain schema also effect query success.

One of the problems that exist related to data warehouse design, is lack of procedures to select appropriate schema. Available resources ([11], [12], [13]), investigated advantages and disadvantages of different schemas. Some of them ([12], [14], [15]) solve some of the problems related to schemas and some of others ([16], [17], [18]) improved query response time. But none of these resources have represented the appropriate framework to select appropriate schema based on type of queries and type of attributes.

III. Selecting Proper Schema for Design

This research work, presents a stepwise algorithm for schema selection, which solves the problem of choosing right schema for a data warehouse. The main selection criteria are query type, attribute type, dimension table type and existence of index. The type of query depends on number of join operation needed to response it and type of attributes it access. All types of attributes such as simple, multi-valued and indexed attributes are used in the research work.

Table wise Checker:

Step1: If table is un-normalized and can be normalized

Then

Step1.1 convert to appropriate normal form

Step2: If the result tables after appropriate normal form is small in size,

Step3: Then star schema and snowflake schema works equally.

Step4: So with considering used tools, schema will be selected.

```
    Step 4.1: If database used is oracle, MS SQL
        Then
            Step 4.1.1: Use star schema
        Elseif
    Step4.2: If database is DB2
         Then
            Step4.2.2: Use snowflake schema
        Elseif
    Step4.3: Use star schema.
        Else
      Step4.4: with try and error, the appropriate  schema is selected.
Step5: If the table cannot be normalized
Then
    Step 5.1: Use star schema.
Exit
```

Attribute wise Checker:

```
Step1: If attribute is composite
Then
    Step1.1 Improved Star Cluster schema:
    ElseIf
Step2: any attribute or its any predecessors are queried frequently, Star Cluster schema otherwise.
Step3: If attribute is multivalued
Then
    Step3.1: If the number of multi-valued attributes is  known,
         Then
            Step3.1.1: If, queries only need to access table T1 in first level of tables that resulted from normalizing this dimension, then there is no difference between star schema and snowflake schema.
        Then
```

GOTO Table wise Checker Step

Step 3.2: If multiple of above conditions are true, by combining the results of each condition, the final schema will be obtained.

IV. Tests

This section shows the effectiveness of a framework tested on all classic and research developed schemas [9] within different kind of queries are presented. The test bed used in this section includes multiple data warehouses. To implement these data warehouses and run queries, MySQL 5.0 and Query Analyzer were used. Queries run in this test bed, are different from each other with respect to the number of join operations. Query response time is always been the most important criteria to compare schemas in data warehouses, so we have also selected the same criteria to evaluate the results of our research.

A. *Testing for different queries*

This test, includes 3 types of query and relates to the case
The results of this test have been shown in table 2, 3 and 4. These results show when condition of case 1 is true, whether Star Cluster schema or snowflake schema is better.

TABLE 1: RESULT FOR QUERY TYPE 1

Average Response time (s)	Query type	Schema type
126.38	1	Snowflake
126.67	1	Star Cluster
129.28	2	Snowflake
124.78	2	Star Flake
34.06	3	Snowflake
29.36	3	Star Flake
34.31	4	Snowflake
16.81	4	Star Flake

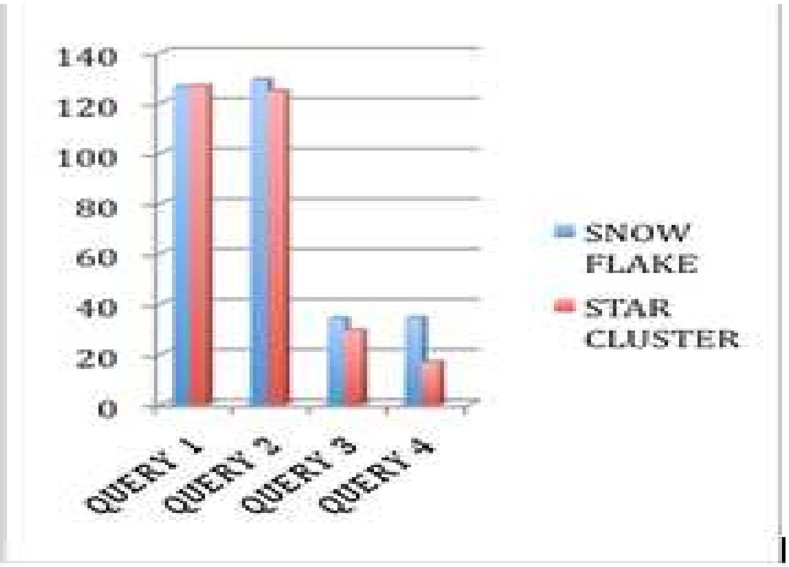


Figure 2: Test Result

TABLE 2: RESULT FOR QUERY TYPE 2

Average Response time (s)	Query type	Schema type
164.28	1	Star Cluster
157.1	1	Improved Star Cluster [19]
7.97	2	Star Cluster
3.68	2	Improved Star Cluster

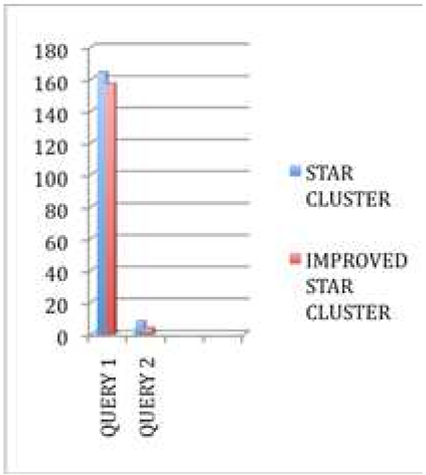


Figure 3: Test Result

TABLE 3: RESULT FOR QUERY TYPE 3

Average Response time (s)	Query type	Schema type
22.19	1	Snowflake
21.83	1	Extended Star Cluster
32.86	2	Snowflake
27.2	2	Extended Star Cluster

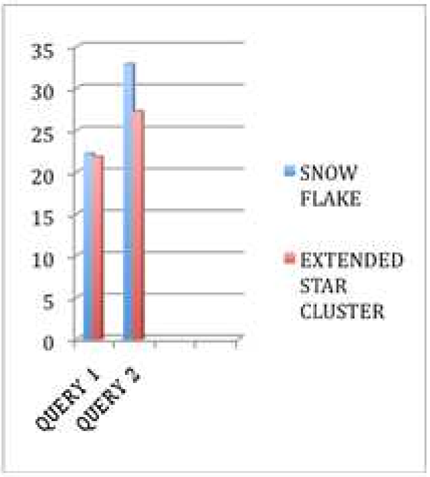


Figure 4: Test Result

V. Conclusions

By using the represented framework, data warehouse builders can choose the best schema for their data warehouse based on the specified criteria and characteristics of the application domain. Also, data warehouse researchers can use this framework to evaluate, compare and extend existing data schemas. This framework could be extended too.

VI. References

[1] Thomas Connolly, Carolyn Begg, Database Systems: A Practical Approach to Design, Implementation and Management, 4th Edition, Addison-Wesley, 2003

[2] William Inmon, *Building the Data Warehouse*, 2nd Edition, New York: Wiley publisher. Inc, 1996

[3] Z. Bellahsene, Schema, "Evolution in Data Warehouses", *Knowledge and Information Systems*, Springer-Verlag, pp 283-304, 2002

[4] E.A. Rundensteiner, A. Koeller, X. Zhang, "Maintaining Data Warehouses over Changing Information Sources", *Communications of the ACM*, Volume, 43, New York, NY, USA, pp 57-62, 2000

[5] Ralph Kimball, M. Joy and T. Warren, *the Data warehouse Toolkit: with SQL server and Microsoft Business Intelligence Toolset*, 2nd Edition, New York: Wiley publisher. Inc, 2006

[6] Efraim Turban, Jay E. Aronson and Narasimha Bolloju, *Decision Support Systems and Intelligent Systems*, 7th edition, Prentice Hall College Div, 2001

[7] Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden, Modern *database management*, Sixth Edition, Pearson Education Publishers, Singapore

[8] Online Analytical Processing (OLAP) and Data Warehousing, academic2.bellevue.edu/~jwright/CIS605/Lesson10/OLAP.ppt Accessed Data: Dec 5, 2008

[9] Daniel L. Moody, Mark A.R. Kortink, "From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design", June 5- 6, 2000, Stockholm, Sweden

[10] Mohammad Rifaie, Erwin J. Blas, Abdel Rahman M. Muhsen, Terrance T. H. Mok, Keivan Kianmehr, Reda Alhajj, Mick J. Ridley, "Data warehouse Architecture for GIS Applications",

[11] B. Heinsius, E.O.M. Data, Hilversum. The Netherlands, "Querying Star and Snowflake Schemas in SAS", SAS Conference Proceedings: SUGI26, paper 123-26, 22-25 April, Long Beach, California, 2001.

[12] D. Moody, M. Kortnik, "From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design",

[13] T. Martyn, "Reconsidering Multi-Dimensional Schemas", SIGMOD Record, Vol. 33,No. 1, pp. 83-88, March

2004.

[14] B. Seyed-Abbassi, “Teaching Effective Methodologies to Design a Data Warehouse”

[15] V. Peralta, R. Ruggia,"Using Design Guidelines to Improve Data Warehouse Logical Design", Proceedings of the International

[16] A. Tsois, N. Karayannidis, T. Sellis, R. Pieringer, V. Markl, F.Ramsak, R.Fenk, K. Elhardt, R. Bayer, "Processing Star Queries On Hierarchically-Clustered Fact Tables",

[17] P.Lane, V. Schupmann, “Oracle9i Data Warehousing Guide, Release 2 (9.2)”, Oracle Corporation, 2000.

[18] V. Markl, R. Bayer, “Processing Relational OLAP Queries with UB-Trees and Multidimensional Hierarchical Clustering”

[19] A. Ghane, “ Comparing the data schemas in data warehouse and representing the improved data schema”