A Schema Selection Framework for Data Warehouse Design

M. H. Peyravi

***Abstract*—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 …) has an important Impact on performance and usability of the designed data warehouse. One of the problems that exists 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. In this paper, we present a schema selection framework that is based on a decision tree for solving the problem of choosing right schema for a data warehouse. The main selection criteria that are used in the presented decision tree are query type, attribute type, dimension table type and existence of index. To evaluate correctness and soundness of this framework, we have designed a test bed that includes multiple data warehouses and we have created all the possible states in decision tree of schema selection framework. Then we designed all types of queries and performed the designed queries on these data warehouses. The results confirm the correct functionality of the schema selection framework.**

***Index Terms*—Data warehouse, framework, online transaction processing, schema selection.**

I. Introduction

One of the problems that exist related to data warehouse design, is lack of procedures to select appropriate schema. Available resources [1]-[3], investigated advantages and disadvantages of different schemas. Some of them [2]- [5], solve some of the problems related to schemas and some of others [6]- [8] 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.

In available resources, [3], [5] Schema selection is based on personal opinion and business requirements. Also, the tool is used, widely affected schema selection. Some of tools like oracle and MS SQL have higher efficiency with star schema; While DB2 works better with snowflake schema. Environment is one of the factors affected schema selection too. For example if data warehouse is composed of some data marts, using star schema is better. With this condition, finding the appropriate schema is time consuming and is based on try and error. In fact we should start from completely normal snowflake schema, in each time, renormalize one of the dimensions and measure the efficiency. This work is repeated until the optimal compound schema is obtained.

In fact, until now, above factors have affected schema selection in data warehouse design. These factors are necessary for schema selection, but aren't sufficient and may be lead to inappropriate schema selection and low efficiency. To solve these problems and represent the appropriate way to schema selection that improves the efficiency and usability of data warehouse, In this paper, the new framework to select data schema for data warehouse is reported. In next section, this framework is described and in the following section, all tests regarding to all classic schemas and some research developed schema [9] which show the framework is effective, are reported.

Manuscript received May 2, 2012; revised May 30, 2012. This work was supported in part by Islamic Azad University.

M. H. Peyravi is with Department Of Computer & Science, Sarvestan Branch, Islamic Azad University, Fars, Iran (email:Peyravi@iausarv.ac.ir).

II. Representing A Framework for Appropriate Schema Selection

In this section, we will represent a framework for appropriate schema selection in data warehouse design. For this purpose, Decision tree is used. The type of queries and attributes affected schema selection in this framework. The type of query depends on number of join operation needed to response it and type of attributes it access. The types of attributes are multi-valued attributes, single-valued attributes and indexed attributes.

The structure of framework is formed as a decision tree and represented in Fig 1. We can state all paths in this decision tree as IF, THEN statements. All these statements have been tested and correctness of them was confirmed. In the following, we will show these statements.

A. Case 1

If in some dimension tables, one attribute acts as a "parent" in two different hierarchies, Then If this attribute or one of its ancestors are queried frequently, the framework propose Improved Star Cluster schema [9]. Else Star Cluster schema [2] is used.

B. Case 2

If it is possible to normalize some of dimension tables, Then If the result tables from normalization these dimension tables are small, Then star schema and snowflake schema works equally. So with considering used tools, schema will be selected.

If used tool is oracle, MS SQL,… that works better with star schema, the framework propose star schema.

If used tool is DB2,… that works better with snowflake schema, the framework propose snowflake schema.

Else if the attribute is queried, is indexed, the framework propose star schema. Else with try and error, the appropriate schema is selected.

C. Case 3

If it isn't possible to normalize the rest of tables, the framework propose star schema.

D. Case 4

If there is multi-valued attribute in some dimension tables, i.e. There are multiple values for one attribute corresponding to single value for other attribute, then

If the number of multi-valued attributes is known, then

If in most times, 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. So with respect to used tool, we could select data schema. Therefore

If used tool is DB2,… that works better with snowflake schema, the framework propose snowflake schema

If used tool is used tool is oracle , MS SQL,… that works better with star schema, the framework propose extended star schema [4].

If in most times, queries need to access outer level tables, then the framework propose extended star schema [4]

If the number of multi-valued attributes is not known, then

If in most times, 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. So with respect to used tool, we could select data schema. Therefore

If used tool is DB2,… that works better with snowflake schema, Then the framework propose snowflake schema.

If used tool is used tool is oracle , MS SQL,… that works better with star schema, Then the framework propose extended star schema [4].

If in most times, queries need to access outer level tables, then the framework propose extended star schema [4].

E. Case 5

If conditions that Kimball states in [10] are true, then using snowflake schema is better. Kimball often prefers to use star schema because of its simplicity and efficiency. But he said in certain situations, snowflake schema is not only acceptable, but recommended [10]. These situations are the cases that there are many null values in large demurral dimension tables. In these situations, variations of snowflake schemas can be useful.

If multiple of above conditions are true, by combining the results of each condition, the final schema will be obtained. In the following, we will show the conditions related to every edges in this decision tree.

We assume if dimension table T is normalized, T1, T2,…,Tn will be resulted.

e_1: An attribute acts as a parent in two different dimensional hierarchies.

e_2: It is possible to normalize some of dimension tables.

e_3: It isn't possible to normalize the rest of tables.

e_4: There is multi-valued attribute in some dimension tables.

e_5: The conditions that Kimball states in [10] are true.

e_6: The attribute related to edge e1 or one of its ancestors isn't queried frequently.

e_7: The attribute related to edge e1 or one of its ancestors is queried frequently.

e_8: T1, T2,…,Tn are small.

e_9: T1, T2,…,Tn are large.

e_{10}: The number of multi-valued attributes is not known.

e_{11}: The number of multi-valued attributes is known.

e_{12}: The used tool is oracle, MS SQL… that works better with star schema.

e_{13}: The used tool is DB2… that works better with snowflake schema.

e_{14}: often the attribute is queried, is indexed.

e_{15}: often the attribute is queried, isn't indexed.

e_{16}: In most times, queries need to access T2,…,Tn that are outer level tables.

e_{17}: In most times, queries only need to access table T1 in first level of tables.

e_{18}: e_{16}

e_{19}: e_{17}

e_{20}: Try and error.

e_{21}: e_{12}

e_{22}: e_{13}

e_{23}: e_{12}

e_{24}: e_{13}

D: Star schema

F: Snowflake schema

G: Star Cluster

H: Improved Star Cluster schema

R: Star schema or snowflake schema

N: Extended star schema

P: Extended star schema

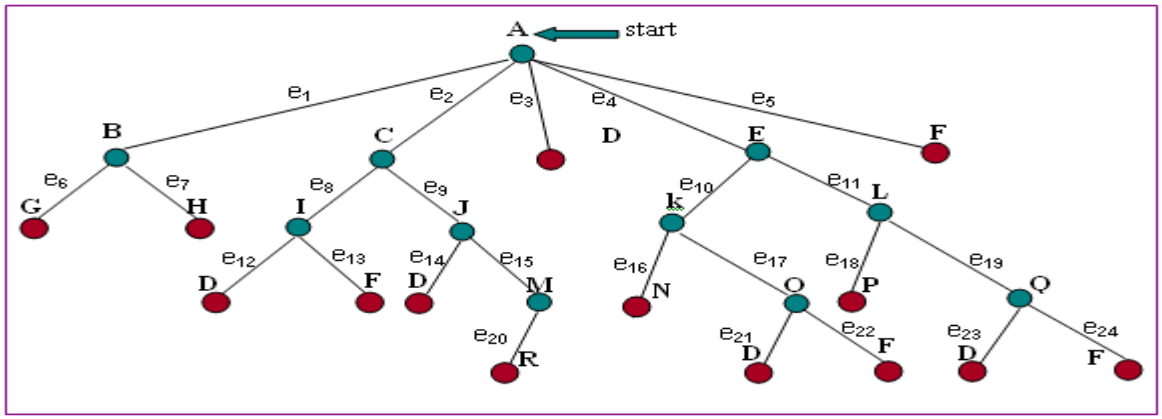

Fig. 1. Schema selection framework

III. Tests

In this section, all tests which show the framework is effective regarding to 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. The states that exist in decision tree were created in these data warehouse dimension tables and multiple types of query were run. The system on which queries run, has 2500Mhz CPU clock and 256 Mbyte RAM. To implement these data warehouses and run queries, SQL server 2000 and Query Analyzer were used. The required data is generated by a C#.Net application. Queries run in this test bed, are different from each other with respect to the number of join operation needed to response them. In most resources, query response time is the most important criteria to compare schemas in data warehouses. So in this paper, query response time is the criteria used to evaluate the framework and compare schemas.

A. Test 1

This test includes 4 types of query and relates to the e_1 edge in figure 1. The results of this test have been shown in Table I. These results show when condition of e1 edge is true, whether Star Cluster schema or snowflake schema is better.

TABLE I: Test I Results

Average response time(s)	Query type	Schema type
129.78	1	Snowflake
129.67	1	Star Cluster
135.58	2	Snowflake
128.68	2	Star Cluster
37.06	3	Snowflake
33.66	3	Star Cluster
37.31	4	Snowflake
16.81	4	Star Cluster

B. Test 2

This test includes 2 types of query and evaluates e_1e_6 and e_1e_7 path in figure 1. The results of this test have been shown in Table II.

TABLE II: Test 2 Results

Average response time(s)	Query type	Schema type
173.28	1	Star Cluster
165.1	1	Improved Star Cluster[1]
12.97	2	Star Cluster
6.56	2	Improved Star Cluster

C. Test 3

This test includes 3 types of query and relates to e_4 edge in figure 1. The results of this test have been shown in Table III.

TABLE III: Test 3 Results

Average response time(s)	Query type	Schema type
26.19	1	Snowflake
25.83	1	Extended Star[2]
36.86	2	Snowflake
31.2	2	Extended Star
37.8	3	Snowflake
33.23	3	Extended Star

D. Test 4

This test includes 4 types of query and relates to e_2e_8 path in figure 1. The results of this test have been shown in Table IV. The results show when dimension tables are small, there is no important difference between star schema and snowflake schema.

TABLE IV: Test 4 Results

Average response time(s)	Query type	Schema type
9.26	1	Snowflake
9.39	1	Star
8.09	2	Snowflake
8.96	2	Star
8.14	3	Snowflake
8.82	3	Star
7.99	4	Snowflake
8.91	4	Star

E. Test 5

This test includes 1 type of query and relates to $e_2e_9e_{14}$ path in figure 1.The query of this test is the same query of type 3 in test 1 except one of the attributes was indexed in test 5. The results of this test have been shown in Table V. Comparing these results and the results of query 3 in table 1 shows indexing in star schema lead to higher efficiency than in snowflake schema.

TABLE V: Test 5 Results

Average response	Query type	Schema type
35.82	1	Snowflake
24.41	1	Star Cluster

The results of Tables I to V have been represented in Fig. 2, 3, 4, 5, 6 respectively.

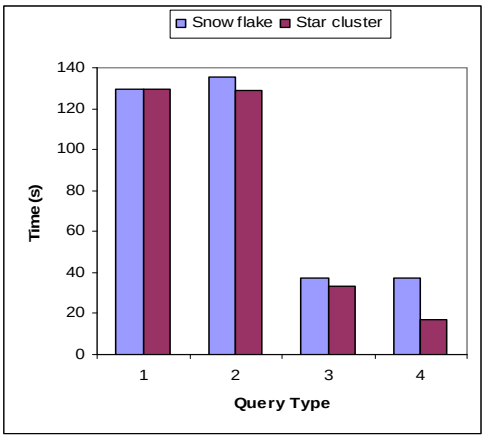

Fig. 2. Test 1 results

[1] This schema was developed during this research work and details available at [9].

[2] Details of this schema are available at [4].

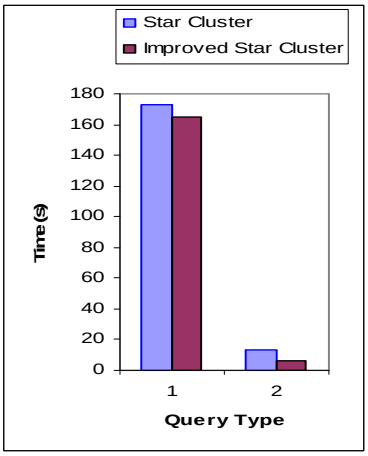

Fig. 3. Test 2 results

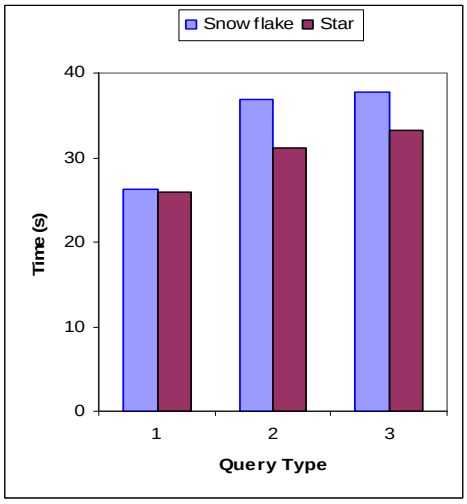

Fig. 4. Test 3 results

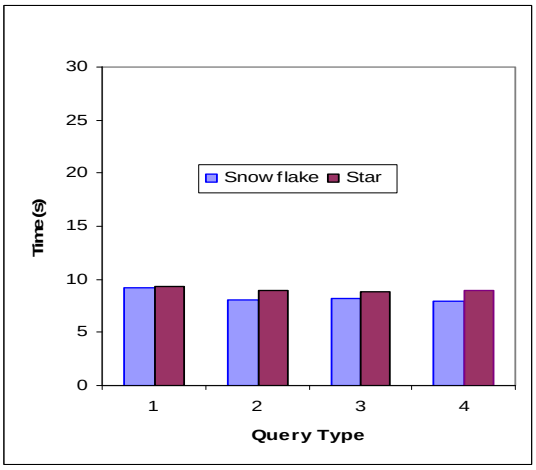

Fig. 5. Test 4 results

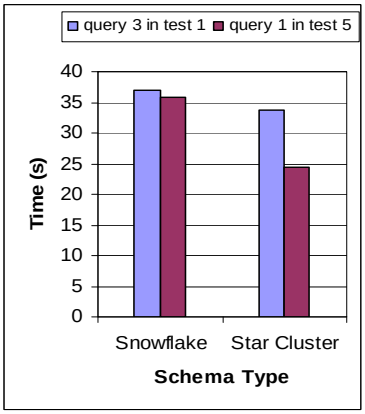

Fig. 6. Test 5 results

IV. 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 extending too.

References

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

[2] D. Moody, M. Kortnik, *"From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design*", *Proc of the International Workshop on Design and Management of Data Warehouses*, 5.1-5.12, Sweden, 2000.

[3] B. Seyed-Abbassi, "*Teaching Effective Methodologies to Design a Data Warehouse*", *Proc of the 18th Annual Information Systems Education Conferenc*e, November 1-4, CD#35C, 2001

[4] V. Markl, R. Bayer, "*Processing Relational OLAP Queries with UB-Trees and Multidimensional Hierarchical Clustering*", *Proc of the International Workshop on Design and Management of Data Warehouses*, Stockholm, Sweden, 1.1- 1.10, 5-6 June, 2000

[5] A. Tsois, N. Karayannidis, T. Sellis, R. Pieringer, V. Markl, F.Ramsak, R.Fenk, K. Elhardt, R. Bayer, "*Proc Star Queries On Hierarchically-Clustered Fact Tables", proces of the 28th Very Large Data Bases Conference*, pp.730-741, Hong Kong, China, 2002.

[6] V. Peralta, R. Ruggia,"Using *Design Guidelines to Improve Data Warehouse Logical Design*", *Proc of the International Workshop on Design and Management of Data Warehouses*, Berlin, 2003

[7] A. Ghane, "*Comparing the data schemas in data warehouse and representing the improved data schema*", *M.SC Thesis, Amirkabir University of Technology*, Tehran, 2005 (in Persian).

[8] T. Martyn, "*Reconsidering Multi-Dimensional Schemas*", SIGMOD Record, Vol. 33,No. 1, pp. 83-88, March 2004.

[9] R. Kimball, "*A Trio of Interesting Snowflakes*", Intelligent Enterprise Magazine, 21 June, 2001.

[10] P. Lane, V. Schupmann, "*Oracle9i Data Warehousing Guide*, Release 2 (9.2)", Oracle Corporation, 2000.