Data Model Performance in Data Warehousing

IOP Conf. Series: Materials Science and Engineering **306** (2018) 012044 doi:10.1088/1757-899X/306/1/012044

Data Model Performance in Data Warehousing

G C Rorimpandey[1*], F I Sangkop[1], V P Rantung[1], J P Zwart[2], O E S Liando[1] and A Mewengkang[1]

[1] Department of Informatics, Universitas Negeri Manado, Jl. Kampus FT-Unima Tondano 95618, North Sulawesi, Indonesia
[2] HAN University of Applied Scieces. Ruitenberglaan 31, 6826 CC Arnhem, Netherlands

*gladlycrorimpandey@unima.ac.id

Abstract. Data Warehouses have increasingly become important in organizations that have large amount of data. It is not a product but a part of a solution for the decision support system in those organizations. Data model is the starting point for designing and developing of data warehouses architectures. Thus, the data model needs stable interfaces and consistent for a longer period of time. The aim of this research is to know which data model in data warehousing has the best performance. The research method is descriptive analysis, which has 3 main tasks, such as data collection and organization, analysis of data and interpretation of data. The result of this research is discussed in a statistic analysis method, represents that there is no statistical difference among data models used in data warehousing. The organization can utilize four data model proposed when designing and developing data warehouse.

1. Introduction

Data model has constant change in storing data therefore consistent and stable interfaces are needed for information that spans for a longer period of time [1]. Because of this, performance in data model became one of the essential criteria of the data warehouse that to be future-proof. It is about response time when the data warehouse is executed to analyze a large amount of data. In the beginning, data warehouses have been designed using 2 approaches. First approach is settled by Ralph Kimbal. He defined the Data Warehouse as a copy of transaction data specifically structured for query and analyzes [2]. Second approach is settled by Inmon. He focuses on building a centralized Enterprise Data Warehouses of which data marts sources their information from [3]. A data warehouse stores all of data in a form that prioritizes query performance, rather than transactional performance or storage volumes [4].

This research is trying to extend a preliminary performance comparison between four data models that has been done in January 2012 by Rael Rutto. Its aim is to measure performances and analyze the results statistically to be able to draw well founded conclusions about differences in performance between the various models [5]. Furthermore, Anchor Model which the one of data model used in data warehouses is a technique recently advocated by Lars Rönnbäck. It uses 6 Normal Form (6NF) databases which are generate expected to perform badly. But, in October 2010 Lars Rönnbäck and friends performed the result of their research that Anchor Model performs substantially better than databases constructed using traditional modeling techniques [6]. Furthermore, they claim however that query optimizers (SQL Server) are so powerful that performance issues are no longer important as for

as table designs are concerned. Query performance has an important role to increase the performance of data model in data warehouse [7].

We have decided to test the claim against Star Schema [8], Optimal Normal Form (ONF) [9] and Data Vault [10] by using SQL Server 2008 and with the same facts for the four models. Compared to the previous research where ONF model is populated by 100000 rows and transformed them to the three models by using ETL processes [5], this scenario is populated with 200000 rows for each historized tables and some less number for the small and intermediate size of tables as described. The experiment will be done using processor speed Intel® Core ™ i3 CPU M350 @2.27GHz, 2 GB RAM and 300 GB Space of Hard Drive. The aim of this research is to measure performances and analyze the results statistically to be able to draw well founded conclusions about differences in performance among the various models.

2. Methods

Basically, statistic is the science of data which concerned with processing data, analysing data and collecting, presenting and transforming data to assist decision maker [11]. Therefore, it contains of 3 main tasks, collection and organization; analysis of data; and interpretation of data. In collection and organization task, there are several methods of organizing data, such as graphically and numerically. In analysis of data task, the computation of various quantities associated with data can be done after data is organized. While in interpretation of data task, the information from two tasks before can be used to make assertions about the real world, such as the average of the calculated data. There are 4 kinds of data, which are nominal, ordinal, interval and ratio. Each type requires its own statistical approach or tests. There are two branches of statistics, they are:

a) Descriptive statistics is collecting, summarizing and presenting data in order to describe the situation from which is the data drawn [12, 13].
b) Inferential statistics is drawing conclusions about a population based only on sample data in order to make predictions, generalizations, or other interferences about a larger set of data [14].

In this research, the writer used branches of descriptive statistic which is determined the mean and SD of sample [15]. Also, because of this research is done for more than 2 groups in multiple comparisons, then Analysis of Variances (ANOVA) method will be used as the format of the performance comparison. There are four data models that will be used a performance comparison experiment, which are: Anchor Model, Optimal Normal Form (ONF), Star Schema and Data Vault. The four models are populated with the same facts, and each are queried differently to answer 13 information needs. The information needs are designed in a query for each model as the measurement for those models because each model has different table structures.

3. Results and discussion

There are a number of steps which followed to get the result of this research. First, the four models should be populated with the same facts. Then, the testing queries of each model should be created by having the same result even in the different model. There is one scenario that will be used in this research. This scenario is taken from previous research with extend some part in the elementary IGD of FCO-IM. The domain tables of scenario have been defined. Null value is not allowed in this scenario and the validity in time is included. Therefore, the Historized attributes are defined and this will be shown in FCO-IM model by the fact types with time validity. Based on the GLR-IGD, there are 9 tables will be generated. To implement the business rules in the population, the number of population of these 9 tables will be different. The population of the scenario is grouped based on the explanation in Table 1.

The testing will be done in all the historization tables. Those tables will be tested by using 13 information needs which called suite, which are:

1) Information need 1 (N1): Testing to get all the attributes in the table without any condition for gathering information about the Actor Ethnicity.

2) Information need 2 (N2): Testing to get all the attributes in the table without any condition for gathering information about the Performance Rating Validity.
3) Information need 3 (N3): Testing to get all the attributes in the table without any condition for gathering information about the Actor Performance Earnings.
4) Information need 4 (N4): Testing all the attributes in the table with where condition for gathering information about the Actor Ethnicity.
5) Information need 5 (N5): Testing all the attributes in the table with where condition for gathering information about the Performance Rating Validity.
6) Information need 6 (N6): Testing all the attributes in the table with where condition for gathering information about the Actor Performance Earnings.
7) Information need 7 (N7): Testing some attributes in the table with where condition and having clause for gathering information about the Actor Ethnicity.
8) Information need 8 (N8): Testing some attributes in the table with where condition and having clause for gathering information about the Performance Rating Validity.
9) Information need 9 (N9): Testing some attributes in the table with where condition and having clause for gathering information about the Actor Performance Earnings.
10) Information need 10 (N10): Testing by using more aggregate function in the select part.
11) Information need 11 (N11): Testing by using more aggregate function in select part and more sub query in where clause.
12) Information need 12 (N12): Testing by using operator 'LIKE' in where clause.
13) Information need 13 (N13): Testing by joining tables with nested query.

Table 1. Number of rows in population scenario.

Tables	Group of Tables		
	Small Tables (in rows)	Intermediate Size (in rows)	Historization Tables (in rows)
Gender	2		
Audience	3		
Ethnicity	10		
Rating	10		
Actor		2000	
Performance		10000	
Actor Ethnicity			200000
Performance Rating Validity			200000
Actor Performance Earnings			200000

Moreover, using ETL technique will spend shorter time than using generated queries for each data model. Because all data models used in experiment have totally different structures. So, this project only generated queries for 1 model (ONF model) and then using ETL technique for populating the three models (Star Schema, Anchor Model and Data Vault). The population was checked by using some queries to make sure that all the models have the same facts. Based on the checking result, the generated queries and ETL processes were worked proper to populate the four models. Then, we decided to populate the four models by extended into 200000 rows. It verified that the methods used to populate the models were easy to extend about the number of rows. The queries were executed 12 and 6 times so as to see how the output varies from the first output. The result of the SQL Server 2008 was exported to the GraphPad Prism program to be analyse in the statistic method. All the runs were entered as data and the program calculated the average for the measurements. The runs of the duration were done with several runs (12, 6). Because there is some scatter in the duration of several runs of the same query, every query was run several times and took the averages as a measurement. The scatter was computed, and the SD deviation was typically about 5%. As a check, SPSS yielded exactly the

same ANOVA results when only these averages were entered as data. The average for each model for each information needs is shown in table 2.

Table 2. The averages between four models for each information needs.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13
AM	3539	3659.67	4890.83	678.5	374.67	669.17	397.5	304.33	3511	666.33	3899.17	6564	32294
ONF	3655..83	3477.67	4820.17	644.83	300.83	668.5	724.17	140.33	5602.33	1209	9668	9359.67	19405.67
STR	3592.5	3614.5	5256.17	642	323.67	725.83	731.33	216.83	15362.5	1076.83	23296.33	24545.17	63414.83
DV	3533	4135.5	4832	630.83	411.33	788.33	1222.33	2038	2219.17	8166.33	7815.17	6154.5	34623.5

For the statistical analysis, figure 1 shows the performance comparison between the four models based on one-way ANOVA. The results of One-way ANOVA analysis has given F(3,36) = 3.32 and p-value = 0.03. It means the null-hypothesis is rejected with a p-value of 0.03, a significant result. Any result of a one-way ANOVA with paired measurements are always reported with their F value, in this experiment the value of F=3.32. The number of (3,36) in F are the degrees of freedom. The number 3 means 4 models – 1 and the number 36 because (13 information needs – 1)*(4 models-1). The rejected null-hypothesis means it was established that some difference between these four models does indeed exist. No information about what difference that is can be obtained in the way however. For that, the post-hoc analysis is needed in this experiment.

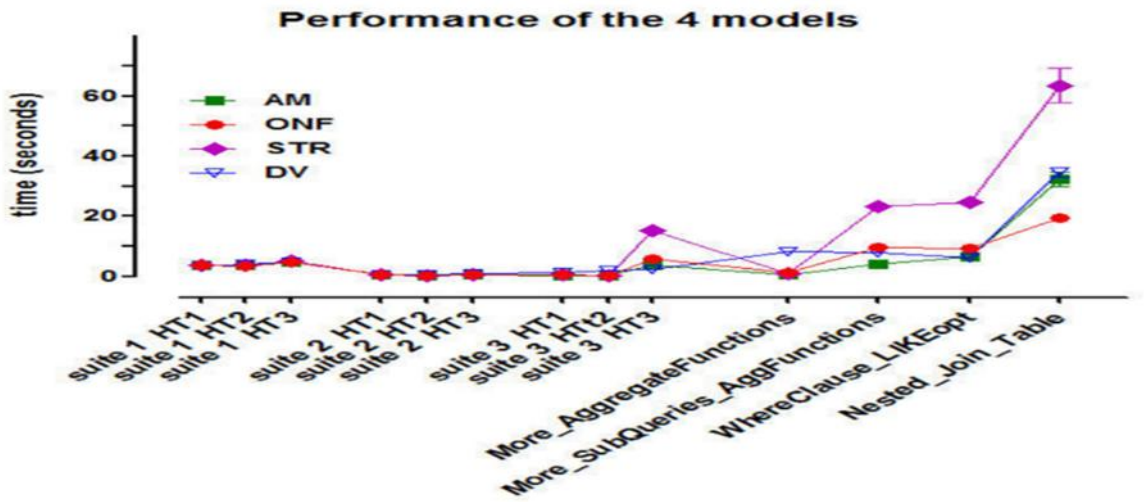

Figure 1. Performance Comparison based on one-way ANOVA.

Based on the one-way ANOVA analysis shown on figure 2, the null hypothesis said there is no difference in performance among the four models. The results of this research experiment presented no statistical significant difference between Anchor Model, ONF, and Data Vault. SQL Server will process the FROM clause first in the SQL statement. When, table the number of join tables are increasing then the rows to be evaluated in the next step (WHERE clause and SELECT clause) become smaller. It means the time execution in the SQL statement with more join tables will be shorter than the SQL statement with less join tables. Furthermore, the experiment has done by comparing all other models with the ONF. The reason is that the ONF is a maximally grouped, redundancy free table structure, which can serve as the point of reference for all other models. This analysis yielded:

1) ONF and Star Schema is significantly different by p-value = 0,03
2) ONF and Anchor Model is not significantly different by p-value > 0.09
3) ONF and Data Vault is not significantly different by p-value > 0.09

Based on the result, the null hypothesis that ONF and Anchor Model do not differ significantly cannot be rejected. So, the claim of the Anchor Model team (Lars Rönnbäck and friends) is not a big issue anymore.

Figure 2. One-way analysis by Graph Prism.

	1way ANOVA			
1	Table Analyzed	models and quiries in 6 fold , models in columns		
2				
3	Repeated Measures ANOVA			
4	P value	0.0304		
5	P value summary	*		
6	Are means signif. different? (P < 0.05)	Yes		
7	Number of groups	4		
8	F	3.323		
9	R squared	0.2169		
10				
11	Was the pairing significantly effective?			
12	R squared	0.7495		
13	F	11.46		
14	P value	< 0.0001		
15	P value summary	***		
16	Is there significant matching? (P < 0.05)	Yes		
17				
18	ANOVA Table	SS	df	MS
19	Treatment (between columns)	350800000	3	116900000
20	Individual (between rows)	4840000000	12	403300000
21	Residual (random)	1267000000	36	35190000
22	Total	6458000000	51	

4. Conclusion

Based on the results of this research the main question can be answered: "Are there any significant performance differences between the four models Anchor Model, Optimal Normal Form (ONF), Star Schema and Data Vault overall, based on all sorts of queries executed on these models?" The results of this research experiment presented no statistical significant difference between Anchor Model, ONF, and Data Vault. However, there is a significant difference between Star Schema and the other three models, with Star Schema performing worse. It can be concluded that lack of redundancy has significant influence to the performance of data model in data warehouse, in terms of accessing it. Generally, the main question and the objectives of this research were met. However, the performance between the four models can be further compared by using more varied scenarios. Because in the world of data warehousing or world of Business Intelligence, each company would face different situation and the data set that were used are not always ideal or stable. For example, a scenario has transactional data are not kept in one database. Other further research for the performance comparison can be experimented in other RDBMS used in data warehousing, such as Oracle.

References

[1] Ribeiro A, Silva A, and Silva A R 2015 *J Soft Eng & App*. **8** 617-634.
[2] Kimbal R, 2002. *The Data Warehouse Toolkit* (Canada: John Willey and Sons, Inc)
[3] Inmon, William H, 2000**.** *Building the Data Warehouses: Getting Started* (Winston Churchill).
[4] Team Ca, 2008 *Computer Associates* **1** 2
[5] Rutto R 2012 A Performance Comparison between four data models used in data warehouse *Thesis Rev* (*HAN University*)
[6] Rönnbäck L 2010 *Data and Knowledge Engineering* **69** 12 1229-1253
[7] Khan F A, Ahmad A, Muhammad I, Alharbi M, Rehman M, and Jan B 2017 J *SCS* 722
[8] Garani G 2012 *International Journal of Data Warehousing and Mining* **8** 4 22-40

[9] Sunar B and Koc C K 2001 *IEEE Transac on Comp* **50** 1 83-87
[10] Jovanovic V and Bojicic I 2012 *P Southern Ass for Inf Sys Conf*.
[11] Lindstedt D 2011 *Super Change your Data Vault*. Avalaible at: LearnDataVault.com
[12] Hannu 1983 *Statistics and Probability Letters* **1** 6 327-332
[13] Petter J and Erich L 2012 *Springer US* 465-471
[14] Mashall G and Jonker L 2011 *Radiography* **17** e1-e6
[15] Bannister H, Goldys B and Penev S W W 2016 *Automatica* **73** 15-26