The Performance of Anchor Model in Data Warehousing

Gladly Caren Rorimpandey, Jan Pieter Zwart, Julyeta Paulina Amelia Runtuwene, Ferdinan Ivan Sangkop, Vivi Peggie Rantung, Parabelem Tinno Dolf Rompas, and Cindy Pamela C. Munaiseche
Universitas Negeri Manado, North Sulawesi, Indonesia

Keywords: Data Model, Data Warehouse, Anchor Model, Inferential Statistics, Post-Hoc Analysis, SQL Server

Abstract: Data model performance became one of essential to be future-proof criteria in data warehouse. The aim of this research is to proof that the performance of Anchor Model as good as the performance of traditional data models that used in data warehousing. The research method is inferential statistics which takes one scenario to generate the sample of data by using SQL Server as the RDBMS. The performance result is discussed by Post-hoc analysis. The experiment performed evidently shows that the Anchor Model has no significantly different with Optimal Normal Form and Data Vault but it has significantly different with star schema. It means the time execution in the SQL statement with more join tables will be shorter than the SQL statement with less join tables. So, the companies that will design and develop data warehouse can be consider to using Anchor Model as their data model in data warehousing.

1 INTRODUCTION

A data warehouse is a database used for reporting and making analyse. It focuses on the modelling and analysis of data for decision hence data warehouse typically provide a simple and concise view of particular subject issue by excluding data that are not useful in the decision support process (Kujur and Oraon, 2016). Nowadays, many companies are using data warehouse (Bassil, 2012). The reason are: a data warehouse can help to manage large amounts of data in a structured way, needing less time to read and analyse them compared to regular data architecture.

Data model is the starting point for designing and developing of data warehouses environment (Rönnbäck et al.., 2010). Data model made the designing of data warehouse become easier and clearer. Inmon said it is like a roadmap of the data warehouse development (Inmon, 2013). Data model is used to support developers of OLAP, data mining, and reporting system. Besides that, it acts as documentation for the final data warehouse. Therefore, data model performance is very important to support the efficiency of data warehouse.

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 generately 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 (Rönnbäck et al.., 2010). More than that, 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. Our research that publish in early 2018 also found that lack of redudancy has influence to the performance of data model in data warehouse, in terms of accessing it (Rorimpandey et al.., 2018). This reaseach is to extend the research of our group. The aim of this research is to look into deep the performance of Anchor Model in data warehousing and compare with the traditional data model, such as Star Schema, Data Vault and Optimal Normal Form (ONF) by using post-hoc analysis. The scenario and population method of this research will used same as the previous research.

2 METHODS

The method of this research is using inferential method is inferential statistics which takes one scenario to generate the sample of data by using SQL Server as the RDBMS. This research is start by designing queries for Anchor Model and others to

Rorimpandey, G., Zwart, J., Runtuwene, J., Sangkop, F., Rantung, V., Rompas, P. and Munaiseche, C.
The Performance of Anchor Model in Data Warehousing.
DOI: 10.5220/0009010102980302
In *Proceedings of the 7th Engineering International Conference on Education, Concept and Application on Green Technology (EIC 2018)*, pages 298-302
ISBN: 978-989-758-411-4

answer the 13 information needs. Then, the performance will be analyzed with post-hoc analysis because the previous research is done just ANOVA analysis. The post-hoc analysis should be done if there is significant different (Pereira, Afonso and Medeiros, 2015). Therefore in this research, it's needed to be tested because the previous research shown that there is significant difference of lack performance by star schema. The research design is described by Figure 1. The data of time execution will be automatically shown in SQL because the query that designed is also for counting the duration of time execution.

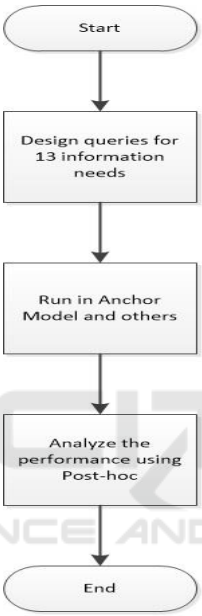

Figure 1: Research design.

3 RESULTS AND DISCUSSION

The scenario used in this research has taken from previous research of performance comparison. The original scenario has taken from Anchor Model Team that used in online modeller. The Anchor Model in this research is created from online tool of Anchor Model. The Anchor Model in this research is created from online tool of Anchor Model. From the FCO-IM point of view, all the nominalized fact types that absorb roles became anchors. The absorbed roles which are played by label type become static attributes. A knot is from a nominalized fact type with one role covered by one UC and has nominalized fact type that absorbed roles. If it is connected to the nominalized fact type that absorbed roles through a non-nominalized fact type with time validity, then it is a historized knot (Simanjuntak et al.., 2016). A non-nominalized fact type that connected with two nominalized fact types that absorbed roles become tie. If the non-nominalized fact type has roles with time validity, then it is a historized tie. Furthermore, if other role of the non-nominalized fact type which is played by label type, then make the role as anchor or knot (depend on modeller).

As mention in methods part, the first step of this research is design queries for 13 information needs. The testing is based on 13 information needs (suites) which implemented to the historized tables. So, for each model will be tested on historized 'tables' by thirteen information needs. Each model has different SQL query because of the difference of table structures for each model. The differences can be found in the table below which shown testing query for two information needs between Star Schema model and Anchor Model.

Table 1 : Queries between star schema & anchor model

No	Star Schema	Anchor Model
	SELECT FROM_DATE, SSN, USERNAME, BIRTHDATE, PROFESSIONAL_LEVEL, GENDER, ETHNICITY FROM STAR_HistorizationScenario..ACTOR_DIM	SELECT T8.AC_ETH_ChangedAt ,T2.AC_SSN_Actor_SocialSecurityNumber ,T3.AC_USR_Actor_Username ,T4.AC_BIR_Actor_Birthdate ,T5.AC_PLV_Actor_ProfessionalLevel ,T7.GEN_Gender ,T9.ETH_Ethnicity FROM AM_HistorizationScenario..AC_Actor T1 INNER JOIN AM_HistorizationScenario..AC_SSN_Actor_SocialSecurityNumber T2 ON T1.AC_ID = T2.AC_ID INNER JOIN AM_HistorizationScenario..AC_USR_Actor_Username T3 ON T1.AC_ID = T3.AC_ID INNER JOIN

No	Star Schema	Anchor Model
		AM_HistorizationScenario..AC_BIR_Actor_Birthdate T4 ON T1.AC_ID = T4.AC_ID INNER JOIN AM_HistorizationScenario..AC_PLV_Actor_ProfessionalLevel T5 ON T1.AC_ID = T5.AC_ID INNER JOIN AM_HistorizationScenario..AC_GEN_Actor_Gender T6 ON T1.AC_ID = T6.AC_ID INNER JOIN AM_HistorizationScenario..GEN_Gender T7 ON T6.GEN_ID = T7.GEN_ID INNER JOIN AM_HistorizationScenario..AC_ETH_Actor_Ethnicity T8 ON T1.AC_ID = T8.AC_ID INNER JOIN AM_HistorizationScenario..ETH_Ethnicity T9 ON T8.ETH_ID = T9.ETH_ID
	SELECT RATING, PERFORMANCE_CODE , COUNT(PERFORMANCE_CODE) AS TOTAL_RATING_PER_PCODE ,MAX(FROM_DATE) AS LASTDATE_RATING ,MIN(FROM_DATE) AS FIRSTDATE_RATING FROM	SELECT T5.RAT_Rating, T3.PE_COD_Performance_Code ,COUNT(T3.PE_COD_Performance_Code) AS TOTAL_RATING_PER_PCODE ,MAX(T4.PE_RAT_ChangedAt) AS LASTDATE_RATING ,MIN(T4.PE_RAT_ChangedAt) AS FIRSTDATE_RATING FROM

No	Star Schema	Anchor Model
	STAR_HistorizationScenario2..PERFORMANCE_DIM WHERE RATING <= 5 GROUP BY RATING, PERFORMANCE_CODE HAVING COUNT(PERFORMANCE_CODE) >= ALL (SELECT COUNT(PERFORMANCE_CODE) FROM STAR_HistorizationScenario2..PERFORMANCE_DIM WHERE RATING <= 5 GROUP BY RATING, PERFORMANCE_CODE)	AM_HistorizationScenario2..PE_AUD_Performance_Audience T1 INNER JOIN AM_HistorizationScenario2..PE_Performance T2 ON T1.PE_ID = T2.PE_ID INNER JOIN AM_HistorizationScenario2..PE_COD_Performance_Code T3 ON T1.PE_ID = T3.PE_ID INNER JOIN AM_HistorizationScenario2..PE_RAT_Performance_Rating T4 ON T1.PE_ID = T4.PE_ID INNER JOIN AM_HistorizationScenario2..RAT_Rating T5 ON T4.RAT_ID = T5.RAT_ID WHERE T5.RAT_Rating <= 5 GROUP BY T5.RAT_Rating, T3.PE_COD_Performance_Code HAVING COUNT(T3.PE_COD_Performance_Code)>= ALL (SELECT COUNT (PE_COD_Performance_Code) FROM AM_HistorizationScenario2..PE_RAT_Performance_Rating T11 INNER JOIN AM_HistorizationScenario2..PE_Performance T12 ON T11.PE_ID = T12.PE_ID INNER JOIN AM_HistorizationScen

No	Star Schema	Anchor Model
		ario2..PE_COD_Performance_Code T13 ON T11.PE_ID = T13.PE_ID INNER JOIN AM_HistorizationScenario2..RAT_Rating T15 ON T11.RAT_ID = T15.RAT_ID WHERE T15.RAT_Rating <= 5 GROUP BY T15.RAT_Rating, T13.PE_COD_Performance_Code)

From the Table 1, it shows that the differences of writing SQL query between the models are significant different. Anchor Model needs more join tables to get information need than star schema which the information need can be found in one table. Based on the testing queries, the Anchor Model performance is shown by Figure 2.

The result of this performance comparison is based on time execution in milliseconds. Time execution is called duration which took time between start time and end time. Start time is the time when the query starts to execute, while end time is the time when the query finishes the execution. The result of the previous research, 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 (Rorimpandey et al.., 2018).

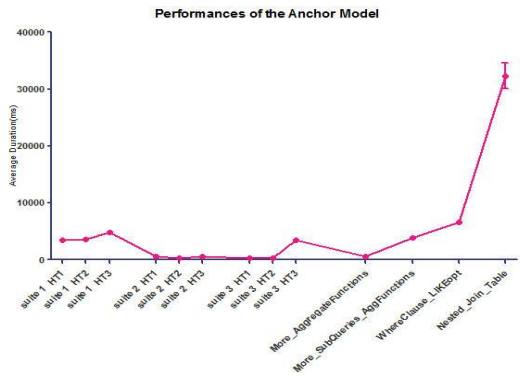

Figure 2: Performance of anchor model.

The post-hoc analysis was done using Sidak's multi comparisons test in GraphPad Prism 6. This is essentially a series of paired tests, taking multi comparison aspects into account somehow. The Figure 3 below has shown the post-hoc analysis.

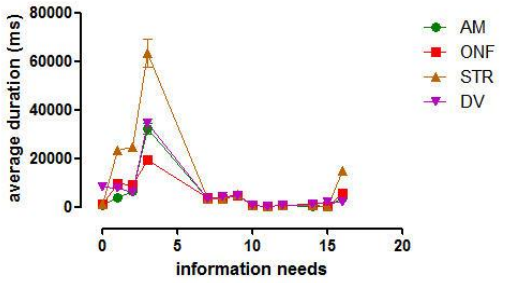

Figure 3: Post-hoc analysis.

Based on this result of the experiment, the analysis by post-hoc, data models with the lack of redundancy has significant influence to the performance of data model in data warehousing, in terms of accessing it. Furthermore, data models in data warehousing is a part of online transaction processing (OLTP) systems which have many writes, or frequent updates, may not be able to take advantage of indexed views because of the increased maintenance cost associated with updating both the view and underlying base tables. It evidently shows by the result of information need 12. So, the indexes will short the performances of data models in accessing some of the information needs in data models of data warehousing but not all information needs. The performance between the ONF, the Anchor Model, and the Data Vault is not significantly different. So, the Anchor Model performance is good as Data Vault and ONF but can be better than Star Schema.

4 CONCLUSIONS

Based on the results of this research the main question can be answered: "Is anchor model will perform better than the traditional data model in data warehousing?" The analysis testing shown that anchor model has same performance with Data Model and ONF but, Star Schema has bad performance. This also proof that the query execution will be shorter if there are many join tables than less join tables. So, the companies that will design and develop data warehouse can be consider to using Anchor Model as their data model in data warehousing.

REFERENCES

Bassil, Y., 2012. A Data Warehouse Design for A typical University Information System. *Journal of Computer Science & Research*., Vol 6(2), pp 12-17.

Inmon, W., . *Building the data warehouse*, Wiley Publishing. Canada, 4th edition.

Kujur, A G P., Oraon, A., 2016, A Data Warehouse Design and Usagel. *International Research Journal of Engineering and Technology*. IRJE, Vol 3(11), pp 335-337

Pereira, D G., Afonso, A., 2015. *Overview of friedman's Test and Post hoc Analysis*. Communication in statistics-simulation and computation, Vol 44(10), pp 2636-2653

Ronnback, L., Johannesson, P., Regardt, O., Wohed, P., 2010. Anchor Modeling-Agile Information modelling in Envolving Data Environment. *Journal of Data and Knowledge Engineering*. Research Gate, Vol 69(12), pp 1229-1253

Rorimpandey, G C., Sangkop, F I., Rantung, V P., Zwart, J P., Liando, O E S., Mewengkang, A., 2018. Data Model Performance in Data Warehousing. *IOP Conf. Series: Materials Sciences and Engineering*. IOP Publishing, Vol 306., pp 1-6