Faculty of Information and Communication Technology

ACCURACY OF QUERIES FOR STORAGE SPACE OPTIMISATION IN GREEN DATA CENTRE

Rabatul Aduni Binti Sulaiman

Master of Computer Science in Database Technology

2014

ACCURACY OF QUERIES FOR STORAGE SPACE OPTIMISATION IN GREEN DATA CENTRE

RABATUL ADUNI BINTI SULAIMAN

A thesis submitted
in fulfilment of the requirements for the degree of Master of Computer Science in Database Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2014

TABLE OF CONTENT

DEDICATION

My effort I dedicate to my loving

Father and mother,

Those that not miss with the advice and prays for me day and night to make me able to get such a success and honour in life and career.

Also to person that always correct my mistakes in this thesis writing

Puan Azilah Binti Draman @ Muda,

Dr Nurul Akmar Binti Emran

Thanks.

ABSTRACT

Proxy based approach is the technique that can be used to make sure the storage space of the database can be minimized towards the development of the green data centre. The idea of proxy based technique is via the schema modification of datasets by dropping some attribute under functional dependency relationships. When the attribute is dropped from the table, it can cause the information loss and can make the query process error. But, when the droppable attribute can be retrieved by using another table, the problem can be solved. Another table in this situation is referred to the proxy map table. The query is needed in the retrieval process. But, it needs to make sure that the query is more simple and able to retrieve the data. The query transformation is another technique that can help proxy based to minimize the storage space. Query transformation will make sure the data can be retrieve successfully. Accuracy of queries help to analyze the result of query transformation and compare with result before query transformation is applied. The result of an attributes selection, query process and the comparison of storage space will be shown in this project. This project will be concluded with the basic characteristic of the proxy map that can help to minimize the storage space and contribute in the development of green data centre.

ABSTRAK

Proksi adalah teknik yang boleh digunakan untuk memastikan ruang simpanan pangkalan data dapat dikurangkan ke arah membangunkan pusat data hijau. Idea teknik proksi adalah melalui pengubahsuain skema kumpulan data dengan menghapuskan sebilangan ciri-ciri didalam pangkalan data bergantung kepada konsep fungsi kebergantungan. Apabila ciri-ciri didalam pangkalan data dihapuskan, ia akan menyebabkan kehilangan informasi can menyebabkan permintaan tidak berjaya. Tetapi, ciri-ciri ini boleh dicapai didalam pangkalan data lain menggunakan proksi. Permintaan data juga diperlukan didalam proses mencapai data. Tetapi, pengguna perlu memastikan bahawa permintaan adalah mudah dan boleh digunakan untuk mendapatkan data. Permintaan transformasi adalah teknik lain yang akan membantu teknik proksi dalam memastikan ruang simpanan pangkalan data dikurangkan. Ketepatan permintaan membantu menganalisis keputusan permintaan transformasi dan boleh membandingkan keputusan sebelum dan selepas permintaan transformasi dilakukan. Hasil daripada pemilihan ciri-ciri, proses permintaan dan perbandingan di dalam ruang simpanan akan ditunjukkan di dalam projek ini. Projek ini juga akan diakhiri dengan ciri-ciri asas pemetaan proksi yang membantu didalam mengurangkan ruang simpanan dan menyumbangkan didalam perkembangan pusat simpanan data hijau.

ACKNOWLEGMENT

First and foremost, I would like to thank to my supervisor Puan Azilah Binti Deraman @ Muda, the person that supports me through her concern to make sure everything is done correctly. I really appreciate the advices and the ideas that she gives to improve writing skills and development of project.

I also would like to thank to my parent, Sulaiman Bin Abdul Rahman and Raja Ramlah Binti Raja Mamat because give me the opportunity and fully support in this project fulfilment.

I also do not forget to thank to Dr Akmar Emran that helping me contribute an idea to develop the project. In addition, all the staff of the Faculty of Information and Communication Technology (FTMK) that helping me in terms of rules to finish the master project thesis.

Finally, I would like to thank to all my friends that supporting in studies and also helping me to give new ideas in project development.

APPROVAL

I hereby declare that I have read this dissertation/report and in my opinion this dissertation/report is sufficient in terms of scope and quality as a partial fulfilment of Master of Computer Science in Database Technology.

Signature :…….....

Supervisor Name :…….....

Date :…….....

DECLARATION

I declare that this thesis entitle “Accuracy of queries for storage space optimisation in green data centre” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and in not concurrently submitted in candidature of any other degree.

Signature : .....

Name : .....

Date : .....

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CMR	-	Comprehensive Microbial Resource
CO_2	-	Carbon Dioxide
RAD	-	Rapid Application Development

CHAPTER 1

INTRODUCTION

1.0 Overview

This chapter reviews the basic concept of green data centre and the techniques that can be used to develop it. The green data centre is the process to store and manage the data whereas the usage of electrical and computer system process is minimized. The implementation of the green data centre can contribute in minimizing energy usage, manage the cost of computer hardware and decrease usage of cooling systems. The aim of the green data centre is to optimize the usage of data centre because it possibly affects our environment with the productions of carbon footprint. The objective of the green data centre is to optimize the storage space of the data centre. It is because the huge storage space requires a lot of energy to process data servers and make the production of carbon footprint increases. The production of carbon footprint can harm the environment and contribute to global warming.

One storage space optimization technique is called as a proxy based approach. This technique can help to optimize the usage of storage space in the green data centre process. The proxy based technique provides the modification of database schema and offers the space savings. The basic concept of proxy based technique is by dropping attributes from main table and insert into another small table. The droppable attributes is depends on the

functional dependency concept between each attribute in the table. This technique will make some of information loss but it can be retrieve if we have the queries that help user to get the data from these two tables. The query transformation is requires to make sure the query that is used is accurate and not affects the storage space and able to retrieve the data. Accuracy of queries is needed to handle the query result after implemented proxy based approach and compared with the result before implementing proxy based approach.

1.1 Background of study

The fast development of data volume increased the usage of electrical energy and carbon footprint. A lot of energy possibly affected the air in the form of diesel exhaust. Cost of handling the database development data centre keep increasing followed by the administration and maintenance expense (Corporation 2009). Data centre provided the expanding of database storage and raises the amount of server used. A recent study of Datacentre Dynamics states that the global data centre nowadays used 31GW of energy per year (Robles 2008).

Data centre manager faced many challenges such as the power limitation, cooling demands and the constraint spacing (Jenkins 2011). To overcome these problems, the green data centre need to focused on energy, environmental efficiency and materials (Bauer et. al, 2011). Moreover, the green data centres focusing on minimizing the usage of electricity cost in operating of the data centre. Besides, it does not use many servers and becomes an option to handle the Carbon Dioxide (CO2) emission. One characteristic of the green data centre is it only uses 40% of electrical energy less than the normal data centre used. With less electrical energy usage can improve the environment and human lives well.

The development of the green data centre requires the efficiency of energy and cooling system. It needs to be redesigned to handle the servers in one place and manage the recycle servers that will transform the old server to the new one in order to achieve the best performance. The illustration of the green data centre is like Figure 1.1 below:

Figure 1.1: Green data centre illustration (Consciousness, 2009).

In fact, the idea to optimize the storage space is not a new issue. A recent study stated that the famous database vendor like Oracle proposed a way to optimize the storage space (Eaton 2006). The compression and deduplication tools are implemented by Oracle and Microsoft in order to optimize storage space and remove repeatable values. It also has several deficiencies that require the whole data tables to be used and manipulated (Eaton 2006). In addition, the technique that is used has limitations in database (E.lai 2012). So, it is not handling the storage space problem due to the constraint issues. By reducing the amount of storage space requirement, we can minimise the number of new additional servers in data centre. By having small amount of server, the amount of CO_2 can be

reduced and emission for cooling purposes which support the objective of the green data centre.

The proxy based approach is a database space optimization technique which can support the establishment of green data centres. The basic process of this technique is to modify the database schema by dropping attributes depends on functional dependency. In this approach, it will map the droppable attributes to its proxy values. In terms of the reference table, it is called as proxy map table. The proxy values that have a functional dependency relationship with droppable attribute are also known as identifier attribute or proxy candidate. Identifier attribute is important in the proxy table because it will be used in the query process to retrieve the droppable attribute information.

Query statement is important in developing proxy map table (T.nagar 2013). In proxy based approach, it requires query transformations to handle the data in the proxy map table. It is because when the attribute is dropped from the main table, it can cause the information loss and affects the queries that are used to extract the data from this droppable attribute. The information that cannot be extracted by user will cause error in the query process. It need query transformation to solve this problem. Query transformation can help to retrieve the data from main and proxy map table.

The accuracy of queries can be used to measure the result of query from proxy table whether it highly accurate or not. The query transformation needs to be analyzed because we want to know the exact accuracy of queries that query transformation was done. The query results before and after proxy based approach is applied will be compared and analyzed in terms of different percentage of query results. The accuracy of queries can also be affected by proxy candidates attribute selection. Therefore, the entire attribute in table needs to be analyzed to make sure the attribute that is chosen is depends on the functional

dependency relationship. In addition, the space accuracy trade off that proxy also help to process selecting attributes what will be deleted from the main table. The accuracy of a query is an indicator of reliability of the proxy based approach. The accuracy of queries can be used by the Database Administrator (DBA) as a criterion to decide on the proxy selection and thus which attributes can be dropped from a table's schema.

The main advantage of proxy based approach is contributed in optimizing the storage space in order to develop the green data centre. The problem is how the usage of proxies can affect the queries. Therefore, the potential of proxy based approach in supporting the establishment of the green data centre in not known. In this project, we want to analyze whether the use of proxy based can contribute in the green data centre or not.

1.2 Problem statement

Proxy based approach require schema modification that can help to reduce storage (Emran et.al, 2013). The droppable attributes actually can cause the information loss from the schema. Information losses can cause the process of retrieving data from the table failed and SQL execution possibly error. It needs an identifier attribute in the query process to retrieve the data from droppable attribute. The varied types of data in the database make user difficult to choose which attributes can be mapped on proxy table.

Suitable candidates are too important in proxy based approach because it will prevent the error in accuracy of queries. The suitable selected candidates able to make sure the accuracy of queries can retrieve the data without any problems. In proxy based technique, the accuracy of query is important for each process. The query will map the droppable

attributes from schema by using the identifier attribute from the schema modification. It is included in the query transformation process which is one of the query optimizer types.

The problems addressed in this study are listed as follows:

1. The variety types of attributes make it difficult to select the suitable proxy candidate to be inserted into proxy map table.
2. Data from main table are loss after implemented the proxy based approach.
3. Difficulties to calculate the total database storage that the proxy based approach can save.

1.3 Research questions

Addresses the study problem, the following research questions are formulated:

1. Which proxy candidate is accepted in terms of space saving and accuracy criteria?
2. How to calculate the total database storage that proxy based can save?
3. How the data can be retrieved after implementing proxy based approach?

1.4 Objectives of study

Candidate selection in proxy based approach is important to make sure the process of storage optimization to develop the green data centre successfully. Candidate selections are chosen based on schema modification under functional dependency rules. In proxy based approach, it dropped attributes and caused query failed because information from this table already loss. Therefore, it needs accuracy of queries to make sure the data is still can be retrieve although the attributes already dropped. In this process, the query transformation is

needed to help the retrieval data process from proxy table. The storage space also requires tools to make DBA easier to use proxy table and choose candidates. The tool which consists of proxy candidate selections and accuracy of queries transformation help the storage space optimization being done.

The objectives of this study are listed as follows:

- To propose proxy candidate that can be implemented in proxy map table.
- To develop tool for proxy based approach on saving the database storage.
- To propose the query transformations steps that can be used to retrieve data after proxy based approach is implemented.

1.5 Scopes of study

This project handles the green data centre and optimizes the storage space. The scope of this research is based on the study about storage space optimization and query transformation in proxy based techniques. Due to a storage space problem, the proxy based approach technique is proposed and will be tested whether it can save the storage space or not. This technique involves the best candidate selection and accuracy of queries that will retrieve the data from proxy table.

The sample data is from the Comprehensive Microbial Resource (CMR) database and the project is focused on the genomic data which keeps information about the gene. The whole project is working on this certain datasets from huge CMR database and the process of optimization is also being implemented. In addition, the CMR need to be looked up to make sure that the optimization process will work like usual. The technique that is used will make sure the green data centre can be implemented in the CMR database.

1.6 Significant of study

In this study, the green data centre can give impact towards the environment and also the computing system. People nowadays need to focus on the problem that arises which is related to the environmental pollution. A technique which has reduced the storage space can contribute to the environment pollution. By reducing the storage space, the usage of electricity and the computing hardware can be minimized. This study is based on the importance on the way the database is managed efficiently.

The use of the proxy based technique can make the information in the database being minimized. It is because some of information from table is already dropped and automatically can improve the storage. Besides, the technique implemented can help to handle the huge amount of database or data centre. This project also can help DBA to make decision making in storage space planning that contributes to the establishment of the green data centre by the aid of proxy selection tool.

In addition, this project also will show the advantages of the proxy based approach that works on the CMR dataset in order to minimize the storage space. The usability and benefits of proxy based can be analyzed in this project by using the result of the accuracy of queries and percentage of storage that can be saved. Accuracy of queries helps to retrieve the information from the droppable attribute by mapping with identifier attributes in proxy table. The efficient queries in the proxy based approach lead the saving time to retrieve the information and because it only included certain attributes. It also process proxy candidate attributes which are droppable and identifier attributes.

The experimental result from this project enable the identification of proxy characteristic that will be useful for proxy selection. The experimental tools also help the identification to choose the proxy characteristic. The tools that consist of the selection of

proxy candidates, accuracy of query process and query transformation steps can contribute in the process of identification proxy based whether it suitable to be used for development of the green data centre or not. The tools can produce the output of the space saving amount and the data that user retrieved from the schema. The spacing of proxy based will be calculated by using the formula that is based on the number of droppable attributes and the size of the schema. The information loss from droppable attributes will be retrieved and solve user problem with data.

1.7 Chapter outline

This thesis contains of the following chapters:

Chapter 1

This chapter consists of an overview of the introduction of the study. It covers the introduction, problem statements, research questions, objectives, scopes and significant of study. It also consists of the description about the idea of storage space optimization by using proxy based approach.

Chapter 2

This chapter covers the reviews of the proxy based approach and query transformations. The literature review discusses other works that are related to the new ideas of the green data centre and the space optimization techniques. The literature review can decides solutions to the problem and also make it as the real situation based on previous studies.