WORKLOAD PERFORMANCE EVALUATION OF LARGE SPATIAL DATABASE FOR DSS BASED DISASTER MANAGEMENT

MUHAMMAD SYAIFUR ROHMAN

MASTER OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING AND INTELLIGENCE)

2017

Faculty of Information and Communication Technology

WORKLOAD PERFORMANCE EVALUATION OF LARGE SPATIAL DATABASE FOR DSS BASED DISASTER MANAGEMENT

Muhammad Syaifur Rohman

Master of Computer Science in Software Engineering and Intelligence

2017

WORKLOAD PERFORMANCE EVALUATION OF LARGE SPATIAL DATABASE FOR DSS BASED DISASTER MANAGEMENT

MUHAMMAD SYAIFUR ROHMAN

A thesis submitted
in fulfilment of the requirements for the degree of Master of Computer Science
in Software Engineering and Intelligence

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I declare that this thesis entitled "workload Performance Evaluation of Large Spatial Database for DSS based Disaster Management" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name : ....

Date : ....

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 (Software Engineering and Intelligence).

Signature	:	....
Supervisor Name	:	....
Date	:	....

DEDICATION

This thesis is dedicated to the Almighty Allah

To my beloved parents, Drs. Eko Hadi Rahmanto and Lilis Sintami Ertifa, S.Pd

To my lovely sister, Itsna and Tsalisa Noor Rahmawati

To my precious country, Indonesia

ABSTRACT

Workload performance evaluation can be implemented during Disaster Management and especially at the response phase to handle large spatial data in the event of an eruption and in this study it is involves the merapi volcano of Indonesia. Merapi volcano is known for its biggest eruption in the world. After the occurrence of an eruption, the affected areas are isolated, and thus it is difficult to be accessed by the rescuers. It is indeed very difficult to reach the isolated area as well as to rescue the victims from the affected areas. Although specific researches have resulted in solutions to the issue, other aspects that include the sending of workload to the database needs to be taken into consideration and it is viable to result in an effective and efficient process. Besides, the shortest route could be defined timely and accurately hence enabling the victims to leave the isolated area and to reach the evacuation point safely. This research intends to study on workload performance which is crucial to support the working mechanism of Database Management System (DBMS). Literature on recent studies has made it clear that research in this particular area of interest is scarce. Therefore, the general objective of this research is to evaluate and predict workload performance of spatial DBMS associated with PostgreSQL which is different from MySQL. Based on incoming workload, this research is able to predict the associated workload into OLTP and DSS workload performance types. From the SQL statements it is clear that the DBMS is able to obtain and record the process, measure the analyzed performances and the workload classifier in the form of snapshots from the DBMS. For example, it has been proven that Dijkstra Algorithm is able to determine the shortest and the safest path. Then, all the workload that are obtained to determine the processes are recorded into one excel file. The Case Based Reasoning (CBR) optimized with Hash Search Technique has been adopted in this study for the purpose of evaluating and predicting the workload performance of PostgreSQL DBMS. Data recorded in the shortest path analysis process reveals that the evaluation and the prediction on workload performance of shortest path analysis using Dijkstra algorithm has been well implemented. It has been proven that the proposed CBR using Hash Search technique has resulted in an excellent prediction of the accuracy measurement. Besides, the results of the evaluation using confusion matrix has resulted in excellent accuracy as well as improvement in execution time. Additionally, the results of the study indicated that the prediction model for workload performance evaluation using CBR that is optimized with Hash Search technique for determining workload data on Shortest Path analysis via the employment of Dijkstra algorithm can be useful for the prediction of incoming workload based on the status of the DBMS parameters. In this way, information is delivered to DBMS hence ensuring incoming workload information is very crucial for the purpose of determining the smooth works of PostgreSQL DBMS.

ABSTRAK

Pelaksanaan penilaian prestasi workload boleh pergi lebih jauh sebagai Pengurusan Bencana dalam fasa tindak balas untuk menangani data ruang yang besar dalam gunung berapi Merapi Indonesia. Seperti yang kita tahu, gunung berapi Merapi dikategorikan dalam letusan terbesar di dunia. Kawasan yang terjejas akan diasingkan selepas berlakunya bencana dan sukar untuk di akses oleh pihak penyelamat. Dan keadaan itu dapat dikatakan ia lebih sukar untuk sampai ke kawasan yang terpencil terutamanya membawa orang keluar dari kawasan tersebut. Kajian tertentu telah dibangunkan bagi menyelesaikan masalah tersebut. Bagaimanapun kami perlu mempertimbangkan aspek lain termasuk penghantaran workload kepada database yang boleh menjurus satu proses yang baik dan berkesan. Lantarannya, jalan terdekat dapat ditakrifkan bertepatan masa dan dengan tepat. Ini dapat membenarkan seseorang untuk dikeluar dengan selamat dari kawasan terpencil ke titik pemindahan. Penyelidikan ini bertujuan untuk meneliti prestasi workload yang mana amat penting menyokong DBMS mekanisme kerja. Daripada kajian baru-baru ini, ia menjadi jelas bahawa tidak ada begitu banyak penyelidikan terperinci yang telah dijalankan. Lantarannya, matlamat umum penyelidikan ini adalah untuk menilai dan meramal prestasi workload DBMS ruang dikaitkan dengan PostgreSQL yang mana berbeza dengan MySQL. Berdasarkan dalam workload yang datang, penyelidikan ini mampu mengelaskan workload bersekutu ke dalam prestasi workload OLTP dan jenis DSS. Daripada penyata SQL seperti yang dinyatakan ia adalah jelas bahawa DBMS merakamkan dan mendapat proses, mengukur prestasi dianalisis , dan pengelas workload dalam bentuk gambar dari DBMS. Sebagai contoh, algoritma Dijkstra telah terbukti mampu menentukan yang laluan paling pendek dan selamat kemudian semua workload diperolehi dalam menentukan proses direkodkan ke dalam satu fail excel. Untuk menilai dan meramal prestasi workload PostgreSQL DBMS, Case Based Reasoning (CBR) mengoptimum dengan teknik Hash Search telah diterima pakai dalam kajian ini. Hasilnya penilaian dan ramalan di prestasi workload analisis laluan paling pendek menggunakan algoritma Dijkstra telah dilaksanakan betul-betul pada data yang merekod dari proses analisis laluan yang paling pendek. Ia dibuktikan dengan hasil ketepatan pengukuran yang dicadangkan CBR menggunakan teknik carian Hash telah mendapat ramalan yang sangat baik dengan hasil penilaian menggunakan kekeliruan matriks menyediakan excellent ketepatan. Selain itu, berdasarkan hasil kajian ini, model ramalan untuk penilaian prestasi workload menggunakan CBR dioptimumkan dengan teknik Hash Search untuk menentukan data workload analisis terpendek dan jalan menggunakan algoritma Dijkstra boleh digunakan untuk meramalkan workload masuk berdasarkan DBMS parameter status dan memberi maklumat kepada DBMS untuk membuat persediaan yang baik. Untuk mengambil perhatian bahawa maklumat beban kerja yang diterima adalah sangat penting untuk menentukan kerja-kerja yang lancar PostgreSQL DBMS.

AKCNOWLEDGEMENTS

Alhamdullilahi Robbil'Alamnin, firstly I would like to express my deepest praise and gratitude to Allah SWT who has given the blessing and mercy to complete this thesis.

Secondly, I would like to express my deepest appreciation to my beloved supervisor Professor Nanna Suryana Herman for his guidance, suggestions, advice and encouragement throughout this study. I would also like to express my appreciation to the Dean of the Information and Communication Technology faculty, Prof. Dr. Burairah Bin Hussin for their support and help towards the completion of my postgraduate studies.

I am especially grateful to the Ministry of Education and Culture Indonesia, Ministry of Research Technology and Higher Education Indonesia and also Dian Nuswantoro Universisty (UDINUS) for providing me the Beasiswa Unggulan scholarship, without which I could not have secured this master study in UTeM Malaysia.

Last but not least, I would also like to express my deepest gratitude to my beloved parents, Drs. Eko Hadi Rahmanto and Lilis Sintami Ertifa, S.Pd, my beloved sisters, Itsna Noor Rahmawati and Tsalisa Noor Rahmawati, and my beloved family, Bani Haji Sulaiman, Bani Sukandar, Bani Sudhar, and RBrotherhood for their moral support in throughout my academic voyage.

Special thanks to all my peers in UTeM and Indonesia for the support rendered during the study. Lastly, thank you to everyone who had contributed towards the realisation of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligent
CBR	-	Case Based Reasoning
CPU	-	Control Processing Unit
CWN	-	Cognitive Wireless Network
DBAs	-	Database administrators
DBMS	-	Database Management system
DFT	-	Discrete Fourier Transform
DIY	-	Daerah Istimewa Yogyakarta
DSS	-	Decision Support System
DTN	-	Delay Tolerant Network
FBS	-	Fuzzy Based Scheduler
GeoJSON	-	Geographical JavaScript Object Notation
GIS	-	Geographic Information System
GPS	-	Global Positioning System
HPC	-	High Performance Computing
I/O	-	Input Output
KHM	-	Karmania Hazard Model
KM	-	Kilo Meters
MRDM	-	Mountain Risk and Disaster Management
OGC	-	Open Geospatial Consortium

OLTP	-	Online Transactional Processing
QGIS	-	Quantum GIS
RAM	-	Random Access Memory
SFSQL	-	Simple Features for SQL
SQL	-	Structured Query Language
TAC	-	Torrent and Avalanche Cadaster
TDWMS	-	Temporary Disaster Waste Management Sites
WFS	-	Web Feature Service

CHAPTER 1

INTRODUCTION

1.1 Background

A volcanic eruption is a natural disaster that is terrifying to mankind. Research stress that a volcanic eruption is the only important factor that is liable to eliminate human civilization. A volcano eruption spews some material in the form of hot gas and volcanic ash into the air and it could be reached at height of 20 km (kilometers) in less than ten minutes. As for the volcanic ash it is formed in an irregular arrangement, thus making it sharp, jagged, and devastate. Volcanic ash contains silica material, which is indicated as a kind of material used in the making of glass. It would be hazardous if it is inhaled. When this volcanic ash is thrown into the air, it quickly spreads to places around the erupted volcano. With the help of the blowing, volcanic ash that has a heat up to 800 degrees Celsius is capable of performing a moving speed of 100 km/h to 20 km radius from the center of the eruption.

In Firdaus et al. (2013) Merapi volcano is stressed to have resulted in many disadvantage which includes the biggest ever eruption in the world. Merapi volcano can be categorized to be very dangerous as it has been erupting (peak activity) every two to five years and is surrounded by a very dense settlement. Since 1548, the volcano has erupted 68 times. Earlier research and observation on volcanic eruption state that it is the major cause of property loss, environment damage and even loss of lives. In 2010, Merapi volcano experienced two eruptions that is on October 26, 2010 and November 5, 2010. Due to the eruption, disaster-prone areas of Merapi volcano suffered severe damages in terms of environment and infrastructure facilities. According to a report received by DIY Police Disaster Risk Identification Team, victims were 257 unidentified, and 360,557 victims were evacuated in DIY and Central Java. Record on the impact of Merapi Volcano eruption disaster reveal a total damage and loss of Rp 3.557 trillion.

The risks of Merapi's eruption can be reduced by managing the disaster. Disaster Management plays an important role during the response phase that is after the disaster has occurred. In the after-disaster scenario, the government and rescuers provide assistance in the form of food, shelter and medical aid to the victims. Victims who need serious handling are brought to nearest healthcare centers. This condition could be worse in isolated area as it would be difficult to rescue the victims from the affected area. Furthermore, the government and rescue team must be deployed to remove the wreckage on the damaged roads, streets and living areas. Such a problem requires technology for the identification of the affected and isolated areas. Besides, help service providers for medical, fire and rescue, police and military forces, ambulance, nearest hospitals and blood banks need to be identified in a short time. A study by Kropat and Meyer-Nieberg, (2016) proposes the implementation of the shortest route analysis to reach the affected places as well as to reach the specific area in the affected place.

The use of technology may help in the management of disaster such as in Merapi volcano. Technology is assistive in the decision making process as well as in obtaining the required information timely. In this regard, technology could improve the mathematical function that could not be handled by man before due to the large amount of data in terms of space and time. With technology, direct conclusion could be made from related data that has been convert into information. This enable the realization of meaning in space and time and its translation into another form of technology like graphic and alpha numerical. Then, new solutions to disaster management that includes risk management could be implemented using Geographic Information System (GIS).

However, to understand the kind of technology that could be implemented after the identification of the exact problem. The understanding of GIS technology especially by the expert is useful in the improving effectiveness of the decision-making process. GIS offers many advantages (Qaddah and Abdelwahed, 2015). It is a powerful tool that integrates graphical interfaces and a variety of information to solve complicated problems, such as modeling the exact sites that have complex criteria and a variety of data. GIS had been used in decision making at different approaches of disaster management in terms of preparedness, mitigation, prevention, recovery, response and rehabilitation. Spatial database is one of the basic foundations of any GIS application. The Spatial data is necessary to conclude and fully function in GIS map to display all the data layers that were included in its design specially to manage disaster.

The number and size of spatial databases now is growing rapidly due to the large amount of data needed to be obtained from X-ray crystallography, satellite images, and other scientific equipment. The different amounts of spatial data in fact force research developments to handle large amounts of data for certain purposes. Large spatial database with the help of GIS technology has been applied in some research problems such as in disaster simulation, evacuation plan or even to make query to the databases itself. One research has been discussing about how large spatial database can be used to make decision support system in Xu et al., (2016). This research explores data mining, statistical analysis, and semantic analysis methods to obtain valuable information on public opinion and requirements based on Chinese microblog data. In this research dealing with disaster management for emergency is a typical big data scenario. Numerous sensors and monitoring devices continuously sample the states of the physical world, while the web data processing techniques make the Internet a big data repository which can reflect the states of the cyber world and the human world. Correspondingly, public opinion and requirements can be obtained from the spatial and temporal perspectives to enhance situation awareness and help the government offer more effective assistance.

The need to specifically manage and analyze large spatial data as well as to support such data which is determined by specialized systems, algorithms and techniques become urgent. System performance management includes identifying the causes of performance problems, measuring the performance, and applying the tools and techniques that are available to handle problems of large spatial databases. As being part of the GIS, the advantage and disadvantage of large spatial database structure has close relation to whether it can succeed or not for the entire study. Complexity and fussiness of large spatial database is different with conventional database. It make relational database management systems fail to full fill request and need of query mentioned from Cao et al., (2015). Rapid growth in the amount of data, changing behavior, and maximum functionality tends to result in database workload thus making it more tricky and complex. Every DBMS experiences complex workloads that is difficult to be managed by humans due to the current situation. Human experts take more time to manage database workload fast and accurately. Even in some cases it may result in failures and thus leading towards undernourishment.

System performance management includes identifying the causes of performance problems, measuring performance, and applying the tools and techniques available to handle problems of large spatial databases. Performance prediction on database is useful (Flores-Contreras et al., 2015) for different purposes such as for capacity planning, load balancing, and resource usage optimization among others, since it allows the estimations of response time of a system under a certain workload. Besides, performance prediction methods also provide insights for resource provisioning, workload management, and scheduling. In some cases, some scheduling algorithms, or resource managers use them as an auxiliary technique, for example, to improve the resource usage of the system. However, analyzing the performance of these systems under varying workloads and hardware configurations can be costly and time consuming (Molka and Casale, 2015).

Previously, it has been explained that database performance evaluation results in several effects. For example, in healthcare units, hospital databases store very important information about patients' clinical status, administrative information and other relevant information related to healthcare services, and thus they are of paramount importance. However, database availability is a complex feature. There are several problems that influence database availability, such as the database can be inaccessible due to network problems or a virus; the database can be too slow and therefore does not satisfy the user's requests (Molka and Casale, 2015). When similar problems occur in the Disaster Management phase, the process of decision-making can be slow and it may result in more losses and even increase the number of deaths. Predicting workload performance is a way to avoid contention rate that affects database performance by modeling how a workload reacts to changes in resource availability, users can make informed purchasing decisions and providers can better meet their users' expectations. Hence, this study would be useful for the parties who are required to make decisions on hardware configurations based on database workload prediction. By performing workload prediction, workload management can dramatically impact its database performance (Molka and Casale, 2015).

Sarwat (2015) upon discussing about spatial data elucidates that delays are not tolerated in fundamental spatial data management system as queries need to be executed accurately and timely. Instead, the user is required to observe the specific information quickly and interactively change any query if it is necessary especially during the response phase of the disaster management. Then, the primary spatial database system is required to Figure out efficient and effective ways to process user request as workload. Therefore, this study proposes specific data based performance evaluation that could handle several types of workload besides processing user workload effectively and efficiently.

1.2 Statement of the Purpose

The purpose of this study is to propose a preliminary spatial data workload performance evaluation framework for large spatial database which is specifically used in disaster management besides determining the shortest path in the delivery of goods as well as for medical assistant and to render urgent help to the needy. This logistical distribution process needs a complex decision making process. Therefore, the knowledge on workload performance evaluation of a spatial database especially before executing the process plays a very important role in ensuring that the decision-making process runs smoothly.

1.3 Problem Statement and Approach Used

The implementation of workload management can be further utilized in handling spatial data of the Indonesian Merapi volcano (Marhaento, 2016). Spatial data for this study is located in two different provinces, namely the Central Java and Yogyakarta of Indonesia, with an altitude of 2,914 meters' elevation above sea level as depicted in the satellite images in Figure 1.1. The Spatial data has grown to become large as it receives such amounts of data from satellite images, geographical area information, GPS position, and representations of geographic feature using primitives like polygons, lines and points to spatial data repository.

As the growth in the amount of spatial data in this study is advanced management of workload is needed to handle several requests from DBMS such as to find the shortest route to the nearest place. Basically, Database administrators (DBAs) tune a DBMS based on their knowledge of the system and its workload. The type of the workload, specifically whether it is Online Transactional Processing (OLTP) or Decision Support System (DSS), is an important criterion for workload tuning (Chiba and Onodera, 2016). Memory resources, for example, are allocated very differently for OLTP and DSS workloads.

Dijkstra Algorithm can be employed for the identification of the shortest route to and isolated area that is affected by disaster and also identity the nearest meeting point to the point of evacuation. Chen et al., (2014) posit that Dijkstra Algorithm approach is the most effective and efficient to make shortest path analysis since improvements have been implemented by some researchers. The processes of workload encompassed in Dijkstra algorithm include sorting, scanning and joining. Elnaffar et al., (2007) state that are mainly the sort, join and scan are mainly the kind of workload type associated with Decision Support System (DSS). For this workload, DBAs typically allocate more memory to the sort heap and the buffer pools.

The job of DSS to access the whole databases to perform Decision support and to join one Table with another Table to fulfill the needs of Dijkstra algorithm to be performed. Another kind workload besides DSS is OLTP (Online Transaction Processing). In the first place, OLTP performs workload that include daily activities like updating, inserting and deleting contain numbers of small transactions and that involve retrievals of individual records based on key values and updates. The OLTP workload does not include processing of whole databases as it only deals with specific Tables and an entity of fields. There are relatively few sorts and joins for this workload. DBAs therefore typically allocate memory to areas such as the buffer pools and log buffers while minimizing the sort heap. Besides memory resources that are involved in the workload processes, there are also CPU Utilization and I/O activity trend. Workload tuning itself is one of the responsibility of DBA to reduce the maintenance costs of DBMS (Database Management System) and to produce efficient workload. Since DSS and OLTP workload affect memory resources, CPU utilization and I/O activity, DBA should be aware about the problem and predict the types that can be performed by monitoring them.

However, there is no clear availability of the spatial database workload performance evaluation framework as well as to the parameters that may be suiTable and applicable for the evaluation of spatial database workload performance. MySQL and PostgreSQL have their own parameters. It has also been identified that different parameters of workload performance from several databases are generated based on the capability of each DBMS to monitor. In this regard, even in using same database different researchers have used different parameters to evaluate workload performance. In other words, selection of suiTable parameters as well as the availability of a framework are considered to be one of the crucial steps that may affect spatial database work load performance evaluation. It has been identified that different parameters of workload performance from several databases are generated based on the capability of each DBMS to monitor.

In MySQL, 291 variables are used to monitor workload. Mostly, previous researchers did not use all the variables to make workload prediction. They had just used some of the variables that are related to the difference between the status of these variables and their status before entertaining the workload which include the actual cost of the workload. It is the detection of this change that is referred to as database workload detection (Zewdu et al., 2009). After taking the differences between the variables used for workload prediction into account, Zewdu et all took 4 of 10 status variable before and after the experiment of another research. Abdul et al., (2014) has come up with 3 variables that give more information about the type of workload. The paper states that 3 variables can be gained from the process of key write, key read and Table lock.

To deal with such Large Spatial Data, MySQL had lack of speed to extracts overlapping regions than PostgreSQL (Khushi, 2015). Based on paper experience from Matuszka and Kiss, (2014) for benchmarking large spatial data, PostgreSQL over performs MySQL even other the most widespread databases (for instance, Oracle, Microsoft SQL, Jena and Sesame database) while PostgreSQL proved to be the best in case of query response time and good for loading time. Based on experiment by Zhang et al., (2015), MySQL spatial provides a basic implementation of OGC (Open Geospatial Consortium)s SFS for SQL standard, but query and analysis operations utilize bounding rectangles instead of true geometries. Then explained MySQL Spatial has some substantial disadvantages. OGC recently standardized the set of spatial operations that should be provided in any geospatial database. However, many products are not yet OGC-compliant. For example, MySQL lacks support for spatial operations like ST_Distance, ST_Dwithin, ST_Intersection and many more. However, MySQL only supports part of the simple spatial queries, which can lead to many false positives in the result set. Open source oriented, another spatial database PostgreSQL with spatial extension is able to store spatial data and support what MySQL does not support specially, to generate Dijkstra algorithm, PGrouting as PostgreSQL extension provides this function. These requirements lead us to select PostgreSQL with PostGIS extension as our target database. As mentioned earlier, there is no clear idea about the kind of parameter that may be suiTable and applicable to evaluate spatial database workload performance. This leads to workload performance evaluation in which the different parameter of each database that had to be monitored had different benchmark. PostgreSQL does not include the parameters for evaluating workload performance as much as MySQL. In benchmarking PostgreSQL to evaluate like MySQL workload performance, three variables could also be monitored with variable in PostgreSQL that include buffers activity, block read and the amount of specific lock.