



UNIVERSITY OF SCIENCE AND TECHNOLOGY
COLLEGE OF GRADUATE STUDIES AND ACADEMIC
ADVANCEMENT

Faculty of Computer Science and Information Technology

A performance evaluation of SQL and NOSQL Database on
HealthCare Data

by

Taj Eldeen Abubaker

A Thesis

Submitted to the College of graduate Studies and Academic Advancements
in Partial Fulfillment of the Requirement for the Degree of Master of Science in
Information Technology

Supervisor

Dr: Mohammed Bakri Bashir

November 2015

Abstract

Until now SQL Database management system (DBMS) able to handle large amount of database efficiently while NOSQL Database management system (DBMS) offer to achieve high availability and more scalability. The variety of database management system make need to know which suitable Database for specific use hence the performance become essential to evaluate . The research aimed to reviews research papers that it's presented in performance comparison between SQL and NOSQL databases. In addition, the data model, classification, and type of SQL and NOSQL databases were reviewed in this study. The research design interface to evaluate the experiment of set of operation (Insertion , Retrieve, Delete , Update one , Update multi) implemented in SQL by Oracle and Mysql database and NOSQL use MongoDB and Cassandra database on Electronic Health Record (EHR) . The performance parameter is measured by calculate the average of response time and the scalability is used average of response time with data size. The both parameter is implemented on for basic operation, which performed in healthcare dataset in order to understand which DBMS offer a good performance and high scalability when used with healthcare data. The experiment conducted over different data size (from 1000 to 10000) record. In insert operation the experiment result indicate that the NOSQL databases (MongoDB and Cassandra), showed better response time and scalability than the SQL (MySQL, Oracle). In retrieve operation the NOSQL databases across Cassandra present better response time and scalability than (MongoDB, Oracle, and MySQL). In delete operation MySQL showed better response time and scalability than (Oracle, MongoDB, Cassandra) databases. In update operation SQL databases (Oracle, MySQL) present better response time and more scalability than the NOSQL(MongoDB, Cassandra) . Finally the experimental result revealed that the NOSQL Database perform better than SQL databases when was used the EHR , because the nature EHR used insert and retrieve operations more than other operations.

المستخلص

لاتزال قواعد البيانات (SQL) حتى الآن قادرة على معالجة كمية كبيرة من البيانات بكفاءة وفعالية عالية إلا إن قواعد البيانات غير العلائقية (NOSQL) ظهرت لتقدم إتاحة عالية وتكون أكثر توسعية . إن تنوع وتباين نظم ادارة قواعد البيانات أدى الى الحاجة الماسة لمعرفة ماهي قواعد البيانات الأمثل للاستخدام المعين لذلك أصبح من المهم تقييم الاداء. هدفت هذه الدراسة لإجراء دراسة للأوراق البحثية التي قُدمت في مجال مقارنة الاداء بين قواعد البيانات العلائقية (SQL) وقواعد البيانات غير العلائقية (NOSQL) ، وكذلك دراسة نماذج وتصنيفات وأنواع قواعد البيانات ، كما تم في هذا البحث تصميم واجهة لتقييم التجارب التي تمت باستخدام مجموعة من عمليات (الإدخال والاسترجاع والحذف والتعديل لسجل واحد و مجموعة من السجلات) حيث استخدمت عملية التقييم Oracle و MySQL في قواعد البيانات (SQL) كما استخدمت قواعد البيانات MongoDB و Cassandra في قواعد البيانات NOSQL وتم تطبيق هذه التجارب على سجلات بيانات طبية (EHR) ، كما ان تقييم الاداء تم بناءً على قياس زمن الاستجابة للعملية من خلال تكرار زمن الاستجابة و حساب متوسط زمن الاستجابة ، وكذلك قياس التوسعية البيانات من خلال حساب متوسط زمن الاستجابة مع زيادة حجم البيانات ، حيث تم إجراء التجربة على العمليات الأساسية (الإدخال والاسترجاع والحذف والتعديل) وذلك باستخدام بيانات صحية وذلك لمعرفة أي من قواعد البيانات التي تقدم أداء عالي وتوسعية أفضل من الأخرى ، تم إجراء التجربة أحجام مختلفة من البيانات من (1000 الى 10000 سجل) في عملية الإدخال أظهرت النتيجة أن قاعدة البيانات NOSQL من خلال (Cassandra و MongoDB) تؤدي بشكل أسرع وتوفر أفضل وتوسعية من قواعد البيانات العلائقية ((Oracle,MySQL) ، أيضاً تشير التجربة في عملية الاسترجاع أن قاعدة البيانات (Cassandra) كانت أسرع و أفضل توسعية من قواعد البيانات (MongoDB و MySQL و Oracle) وفي عملية الحذف نجد أن قواعد البيانات MySQL كانت أسرع وأفضل توسعية من قواعد البيانات (Oracle و MongoDB و Cassandra) أما في عملية التعديل فإن قواعد البيانات SQL من خلال (Oracle و MySQL) تؤدي بشكل أسرع من قواعد البيانات NOSQL (MongoDB و Cassandra)، توصلت الدراسة الى أن قواعد البيانات (NOSQL) تؤدي بطريقة أفضل من قواعد البيانات (SQL) عند استخدام البيانات الطبية (EHR) وذلك نسبة لطبيعية البيانات الطبية التي تستخدم عملية الإدخال والاسترجاع أكثر من العمليات الأخرى.

1.1 Introduction:

Relational databases (RDB) allow data storage, extraction and manipulation using a Structured Query Language (SQL). The developers, research organization and industries have been using the RDBMS (Relational Database Management System) for many decades. This database technology has been also used by most traditional data-intensive storage applications and data retrieval applications [1].

Relational databases exhibit a variety of limitations, the limitations of scalability and storage, and efficiency losing of query due to the large volumes of data, and the storage and management of larger databases become challenging, In order to overcome these limitations, a new database model was developed with a set of new features, known as NoSQL databases (Not Only SQL). The thesis compare relational database and NoSQL databases, to evaluate their performance according to the typical use for storing and retrieving data. The thesis evaluate the databases with experimental testing using a mix of operations to understand the capability of relational and non-relational databases for handling different requests, and to understand how performance is affected by each database type and their internal mechanisms [2].

1.2 Research Motivation:

The internet is a familiar technology used across the globe and big data is a real challenge in present era. The organizations like Facebook, Google, and Amazon etc. are facing the big data problem. To overcome the big data issue, experts proposed a new system of data management commonly known as NOSQL to meet the data management demands of present era [4]. According to aforementioned issues that lead to important needs to face the users requirement.

1.3 Problem background:

The features and architecture of database system differ from one to another. These differences make the performance of database differ and vary, because the variety of the mechanisms used to store and retrieve data and each of them has different features and characteristic.

The performance is considered as significant issue for database management system (DBMS). Furthermore, NOSQL database management systems have different performance because the variety of the data Model used to store data such as (key-value, column –oriented, document-store, etc.). In addition, there is no Structure query language. Furthermore, each DBMS has different features and characteristic used in concurrency control, data storage medium, replication, and transaction mechanisms of the systems. The performance is an important factor for deciding which database will be used for enterprises and applications. therefore there is need to evaluate the performance SQL and NoSQL DBMS's[2].

1.4 Problem Statement:

There are varietyof several database management system performance that make the evaluation performance and scalability of SQL and NoSQL DBMS are a significant issue to identify the suitable DBMS for specific applications.

1.5 Objective:

1. Reviewing the related researches studies which tackled in this field as well as how it'smeasured the performance factor. Additionally, performing an experiments testing to compare the MongoDB and Cassandra DBMS performance versus to Oracle and MySQL DBMS in order to helps to choose the most suitable Database.

1.6 Research Contribution:

1. The analysis study of related work that discussed and evaluated the SQL and NoSQL DBMS software.
2. Guide the user to select appropriate database management system for our application by evaluate their performance.

1.7 Scope:

Evaluate the SQL, NOSQL DBMS performance by testing the insertion, retrieve, delete, and update operations.

1.8 Thesis structure:

The structure of the research consists of the five chapters:

1. Chapter one introduction about the thesis.
2. Chapter two present the literature review related to SQL and NoSQL Performance comparison.
3. Chapter three discuss the methodology that used in the thesis
4. Chapter four contains the practical implementation of Experiments and Analysis.
5. Chapter five contains Conclusions and the Recommendation



