**UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT**

Faculty of Computer Science and Information Technology

# Evaluation of MongoDB and Cassandra Database Performance on HealthCare Data

by

Mohammed Hussein Mohammed Musa

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

The NoSQL Database Management Systems (DBMS) has been emergence to meet the requirement of new application such as web 3.0 and very large scale application. These applications require a new features that doesn't available in the relational database management system. The performance is an important factor for deciding which DBMS will be used for enterprises and applications. Furthermore, NoSQL has different performance because the variety of the data Model used to store data. Therefore need to evaluate the NoSQL DBMS performance. The thesis compares different NoSQL databases, to evaluate their performance according to the typical use for storing and retrieving data. Moreover the goals of this thesis is reviewing the performance work related to NoSQL database by focusing on work related to performance comparison. In addition to build an evaluation testing application to evaluate the performance and scalability for the two of most NOSQL Database MongoDB which belong to Document-oriented database and Cassandra is a wide-column database. The performance parameter is measured by calculate the average of execution time ,and also the scalability is used the average of execution time with data size ,the both parameter is implemented on the four basic operation such as (insert, retrieve ,delete and update) which performed on the database ,the evaluation testing is performed on healthcare dataset in order to understanding which NoSQL database is offer a good performance and high scalability when used with healthcare data. The experiment results show that MongoDB achieve a good performance and high scalability better than Cassandra with insert and delete operation while Cassandra is offer a good performance and a high scalability better than MongoDB with retrieve and update therefore the Cassandra is good choice for to health care data because it's mostly execute the retrieve operation .

# المستخلص

نشأت قواعد البيانات الغير العلائقية لتلبية متطلبات التطبيقات الحديثة مثل تطبيقات الجيل الثالث من الويب والتطبيقات واسعة النطاق التى تحتاج الى خصائص لا تتوفر فى قواعد البيانات العلائقية التى تعتبر الحل الأمثل للتطبيقات التقليدية. أحد أغراض هذا البحث هى تقديم مقدمة عن قواعد البيانات الغير علائقية والأسباب التى أدت الى ظهورها بالأضافة الى ذلك من أهداف هذا البحث هى عمل عرض للدراسات السابقة المتعلقة بقواعد البيانات الغير العلائقية مع التركيز على الأعمال التى تناولت مقارنة للأداء بالاضافة الى ذلك تم بناء إطار عمل لأجراء عمليات الأختبار وذلك لتقييم معاملى الأداء قابلية التوسع لأثنين من أشهر انواع قواعد البيانات الغير علائقية هما قواعد بيانات MongoDB التى تعتبر من نوع المستند وال Cassandra التى تعتبر من نوع العمود . كما يتم قياس معامل الاداء بحساب متوسط زمن تنفيذ العملية بينما يتما قياس قابلية التوسع بحساب متوسط زمن تنفيذ العملية بالتناسب مع حجم البيانات ، يتم حساب نتائج معاملى الاداء وقابلية التوسع للعمليات الاساسية التى يمكن أجراءها على قواعد البيانات مثل (الأضافة ،الأسترجاع ،الحذف والتعديل )، جميع عمليات الأختبار يتم تنفيذها على بيانات صحية وذلك من أجل فهم أى من قواعد البيانات الغير علائقية التى تم اختبارها تقدم أداء أفضل وقابيلة توسع عالية عن أستخدامها مع بيانات طبية .

أظهرت نتائج التجارب ان MongoDB تقدم أداء عالى وقابلية توسع أفضل من Cassandraفى عمليتى الأضافة والحذف بينما تقدم Cassandra أداء عالى وقابلية توسع أفضل من MongoDB خلال تنفيذ عمليتى الاسترجاع والتعديل لذلك فأن Cassandra تعتبر خيار أفضل للبيانات الصحية وذلك بسبب تنفيذ عملية أسترجاع البيانات فيها بصورة دائمة اكثر من العمليات الاخرى

## 1.1 Introduction

Developers, Research Organization and Industries have been using the Relational Database Management Systems (RDBMS) for many decades. This database technology has been used by most traditional data-intensive storage applications and data retrieval applications. Furthermore, RDBMS are generally considered as most efficient databases but when it comes to high-performance, scalability, flexibility and availability then they are actually not Relational databases have limitations to deal with scalability for large volumes of data and variety types of Data(Unstructured ,semi structured) therefore the researchers suggest the non-relational database technologies, also known as NoSQL, were developed to better meet the needs of new Applications with large amounts of records. But there is a large amount of NoSQL candidates, and most have not been compared thoroughly yet. The thesis compares different NoSQL databases, to evaluate their performance according to the typical use for storing and retrieving data. The thesis evaluate NoSQL databases with Experimental testing using a mix of operations to better understand the capability of non-relational databases for handling different requests, and to understand how performance is affected by each database type and their internal mechanisms

## 1.2 Problem Background:

NoSQL database is considered a quite new technology in the database domain. However, the NoSQL are being developed on known and existing theory. NoSQL databases systems still have various limitations. There is no a common standard nor any common and familiar query language for querying NoSQL databases. Each database behaves in a different way and does things differently. Relatively these databases are immature and constantly evolving.

Organizations need to consider the following options when choosing the suitable NOSQL DBMS for their Requirements: Data Model, CAP (consistency, availability and partition tolerance) Support, Multi Data Center Support, Capacity , Performance, Query API, Reliability, Data Persistence and Business Support.

Among aforementioned issues the performance is consider as significant issue for NoSQL DBMS. Furthermore, a NoSQL database management system has 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 Standard Query language, and each of them 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 NoSQL DBMS performance

## 1.3 Problem Statement:

There are several types of NoSQL DBMS with different capabilities and features. The evaluation performance and scalability of NoSQL DBMS are a significant issue to identify the suitable NoSQL DBMS for specific applications.

## 1.4 Research Objectives:

1- Reviews related research papers that discussed the NOSQL database focusing on the performance comparison Issue.
2- Compare the NoSQL DBMS performance and scalability that helps to choose the most Suitable Database for the companies, developers, and applications.

## 1.5 Research Contribution:

The thesis contribution is composed from two part:

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

## 1.6 Scope:

Evaluate the NoSQL DBMS performance and scalability by testing the insert, read, update and delete operations.

## 1.7 Thesis Outline:

The thesis outlines organized as follows.

Chapter 1 provide an introduction about the thesis.

Chapter 2 present the literature review related to NOSQL and NOSQL Performance comparison.

Chapter 3 describe the research methodology which used in the thesis

Chapter 4 describe the test case dataset and operations testing interface implementation and the experiments and the results.

Chapter 5 illustrates conclusions and recommendations of the thesis