Comparison of Traditional SQL DBMS and NoSQL Performance

Table of Contents

[Abstract 1](#_Toc370042849)

[Acknowledgements 2](#_Toc370042850)

[Introduction 3](#_Toc370042851)

[Project Introduction 3](#_Toc370042852)

[Comparison of MySQL and MongoDB 3](#_Toc370042853)

[Aim 4](#_Toc370042854)

[Objectives 4](#_Toc370042855)

[Analysis 5](#_Toc370042856)

[Literature Review 5](#_Toc370042857)

[Research Questions 5](#_Toc370042858)

[Design 6](#_Toc370042859)

[Tools Used 6](#_Toc370042860)

[Hardware 6](#_Toc370042861)

[Data Set 6](#_Toc370042862)

[System Design 6](#_Toc370042863)

[Example of MongoDB document: 7](#_Toc370042864)

[Test Design 7](#_Toc370042865)

[Implementation 8](#_Toc370042866)

[Problems during implementation 8](#_Toc370042867)

[Results 9](#_Toc370042868)

[Evaluation 10](#_Toc370042869)

[Conclusion 11](#_Toc370042870)

[Table of Figures 12](#_Toc370042871)

[Table of Charts 13](#_Toc370042872)

[Appendices 14](#_Toc370042873)

[Appendix A. Project Initiation Document 14](#_Toc370042874)

[Appendix B. Ethical Checklist 14](#_Toc370042875)

[Appendix C. Initial Project Plan (Gannt Chart) 14](#_Toc370042876)

[Appendix D. Final Project Plan (Gannt Chart) 14](#_Toc370042877)

[Glossary 15](#_Toc370042878)

[Bibliography & Citations 16](#_Toc370042879)

# Abstract

# Acknowledgements

# Introduction

## Project Introduction

Over the last few years there has been an explosion in the amount of data and all this needs to be stored somewhere in order that it can be made use of. It has been reported that we create nearly 12 terabytes of data a day in Tweets alone, (Gobble, 2013), and this data flow is increasing daily with 90% of the data in the world today being created in the last 2 years. (Gobble, 2013).

This huge amount of data has to be stored but it also has to be accessed if it is to be of any use and this project is looking at how this “big data” can be stored in the most appropriate way to enable fast read access or fast write access, depending on how the data is to be used. Traditionally this data has been stored in relational databases accessed by a language called SQL, (Structured Query Language). There has been a change in recent years with the development of NoSQL, (the definition varies between No SQL and Not Only SQL, depending on where you look).

When data is stored in a traditional style of SQL database the question arises - “Is this the best way of storing this particular data?” now that there are new ways of storing it. In some cases the data needs to be stored rapidly and accessed at a slower rate so the database needs to be able to handle fast writes and slower reads; some data needs to be accessed at the similar rate as the data is being written and some may be accessed in bursts of activity. This project is being undertaken to try to find which methodology is the best in different scenarios.

Related work has been carried out in a field similar to this project, (Parker, Poe, & Vrbsky, 2013), but this work was based on a much smaller dataset with the testing not being done at the base level of a cli client. It is possible that this approach caused small changes in the timings. This project will be using a direct cli entry for commands to ensure that the timings received are not obscured by any other technology.

## Comparison of MySQL and MongoDB

A traditional relational database model, (see Figure i) looks similar to a MongoDB but they are not used in the same way.



Figure i

In a MongoDB database instance there is a different structure to a standard SQL format, as shown in Figure ii.



Figure ii

A collection can be seen as analogous to the table in a standard relational database and a document appears to fulfill the same function as a row. However, these two structures are vastly different to their relational counterparts.

## Aim

## Objectives

# Analysis

## Literature Review

By undertaking a literature review, an understanding about the current uses of NoSQL databases can be gained. There are a lot of papers available that cover the traditional database methodologies and the way in which they can be used but there are less available for the newer technologies at this point.

## Research Questions

What / why / how questions

# Design

## Tools Used

* TextPad (<http://www.textpad.com/>)
* EditPad Pro (<http://www.editpadpro.com/>)
* UltraEdit (<http://www.ultraedit.com/>)
* Large Text File Viewer (<http://www.swiftgear.com/ltfviewer/features.html>)
* MySQL client
* MongoDB client

## Hardware

## Data Set

The dataset chosen for this project has been obtained from the SNAP project at Stanford University. (Lescovec, 2012) The file consists of 7,911,864 reviews from 889,176 users for 253,059 movies. The median number of words per review is 101 and the data covers August 1997 to October 2012. The file, when word-wrap is disabled, has 71,205,230 lines: the format for a record follows this structure:

product/productId: B00006HAXW

review/userId: A1RSDE90N6RSZF

review/profileName: Joseph M. Kotow

review/helpfulness: 9/9

review/score: 5.0

review/time: 1042502400

review/summary: Pittsburgh - Home of the OLDIES review/text: I have all of the doo wop DVD's and this one is as good or better than the 1st ones. Remember once these performers are gone, we'll never get to see them again. Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE this DVD !!

The breakdown of the data:

* **product/productId**: [asin](http://en.wikipedia.org/wiki/Amazon_Standard_Identification_Number), e.g. [amazon.com/dp/B00006HAXW](http://amazon.com/dp/B00006HAXW/)
* **review/userId**: id of the user, e.g. [A1RSDE90N6RSZF](http://www.amazon.com/gp/cdp/member-reviews/A1RSDE90N6RSZF)
* **review/profileName**: name of the user
* **review/helpfulness**: fraction of users who found the review helpful
* **review/score**: rating of the product
* **review/time**: time of the review (unix time)
* **review/summary**: review summary
* **review/text**: text of the review

## System Design

Each record needs to be edited so that it will be saved in the databases properly as there are many reviews that have double quotation marks within the reviews. This will need to be edited so that these marks are escaped, enabling them to be displayed properly in the query results. This work will be carried out on the original dataset so that it can be used without further editing by both database technologies being used in this project.

There will be no editing of the actual content of each record; the views are of the original author and not mine. Spelling and grammatical errors within these records are not being corrected as these errors would not have any bearing on the final results.

### Example of MongoDB document:

>

> db.amazonReviews.findOne()

{

"\_id" : ObjectId("52591a8ac2892db741bd3f3b"),

"productId" : "B003AI2VGA",

"userId" : "A141HP4LYPWMSR",

"profileName" : "Brian E. Erland \"Rainbow Sphinx\"",

"helpfulness" : "7/7",

"score" : 3,

"date" : 1182729600,

"summary" : "\"There Is So Much Darkness Now ~ Come For The Miracle\"", "text" : "Synopsis: On the daily trek from Juarez, Mexico to El Paso, Texas an ever increasing number of female workers are found raped and murdered in the surrounding desert. Investigative reporter Karina Danes (Minnie Driver) arrives from Los Angeles to pursue the story and angers both the local police and the factory owners who employee the undocumented aliens with her pointed questions and relentless quest for the truth. Her story goes nationwide when a young girl named Mariela (Ana Claudia Talancon) survives a vicious attack and walks out of the desert crediting the Blessed Virgin for her rescue. Her story is further enhanced when the \"Wounds of Christ\" (stigmata) appear in her palms. She also claims to have received a message of hope for the Virgin Mary and soon a fanatical movement forms around her to fight against the evil that holds such a stranglehold on the area. Critique: Possessing a lifelong fascination with such esoteric matters as Catholic mysticism, miracles and the mysterious appearance of the stigmata, I was immediately attracted tothe '05 DVD release `Virgin of Juarez'. The film offers a rather unique storyline blending current socio-political concerns, the constant flow of Mexican migrant workers back and forth across the U.S./Mexican border and the traditional Catholic beliefs of the Hispanic population. I must say I was quite surprised by the unexpected route taken by the plot and the means and methods by which the heavenly message unfolds. `Virgin of Juarez' is not a film that youwould care to watch over and over again, but it was interesting enough to merit at least one viewing. Minnie Driver delivers a solid performance and Ana Claudia Talancon is perfect as the fragile and innocent visionary Mariela. Also starring Esai Morales and Angus Macfadyen (Braveheart)."

}

## Test Design

The data

# Implementation

How is the project implemented?

## Problems during implementation

The dataset was in excess of 71,000,000 lines and so this meant that a tool that would handle very large files and display them was needed. The initial requirements appeared to be that the file would need to be viewed only; however, after working on the dataset, it became obvious that the file needed to be viewed and edited within the same program. It was not feasible to copy and paste some text from one program into another, from Large Text Viewer into TextPad or Notepad++. Both of the programs that would allow macros to be programmed for the replacement of text, for example the quotation mark character or the phrase ‘product/productId’. The initial tests were done using copy and paste from the Large Text Viewer, (LTV), program into TextPad and although the macros ran successfully the copying of a few thousand lines at a time into the program proved to be too labour intensive.

A new approach was required and the programs tested were EditPad Lite, EditPad Pro and UltraEdit. The final choice was EditPad Pro; the lite version did not allow files of over 2Gb to be loaded so this was discarded. UltraEdit was tested against EditPad Pro and it was found, in my opinion, that the find / replace functions were more cumbersome in UltraEdit and I decided to use EditPad Pro. Initially a 30 day demo was used in order to test the usability. The initial tests were on replacing the text product/productId with “productId” and these took around 95 minutes to replace nearly 8,000,000 matches.

Although this was a long time, it was deemed a successful test. The alternative was

1. To copy and paste a portion of the text into NotePad++.
2. Run a find and replace on the text using my own pre-recorded macros.
3. Save the edited text.
4. Copy and paste another portion of text and start again.

This has produced an unexpected recommendation from this project; the program to use to edit large text files is EditPad Pro, available from <http://www.editpadpro.com/>

# Results

# Evaluation

# Conclusion

# Table of Figures

[Figure i 2](#_Toc369455368)

[Figure ii 2](#_Toc369455369)

# Table of Charts

# Appendices

## Appendix A. Project Initiation Document



## Appendix B. Ethical Checklist

## Appendix C. Initial Project Plan (Gannt Chart)



## Appendix D. Final Project Plan (Gannt Chart)

# Glossary

* Big Data – A collection of data-sets that is so large that it has become difficult to process. This can due to the sheer amount of data or the rate at which it is being collected.
* Data Model – An integrated collection of concepts for describing & manipulating data, relationships between data and constraints on the data in an organisation. (Connolly & Begg, 2010)
* Database – A shared collection of logically related data and its description, designed to meet the information needs of an organization. (Connolly & Begg, 2010)
* DBMS – A software system that enables users to define, create, maintain, and control access to the database. (Connolly & Begg, 2010)
* NoSQL – This definition was initially said to mean “**No SQL**” but it is now generally accepted to mean “**N**ot **O**nly **SQL**”
* SQL – **S**tructured **Q**uery **L**anguage. A specialised programming language that is designed to manage data that is held in relational database systems. There are a number of different SQL languages that closely follow a standard definition. However there are enough differences between these different versions that queries written in one are not guaranteed to work in another.

# Bibliography & Citations

Chodorow, K. (2013). *MongoDB - The Definitive Guide* (2nd ed.). Sebastopol, CA: O'Reilly.

Connolly, T., & Begg, C. (2010). *Database Systems - A Practical Approach to Design, Implementation and Management* (5th ed.). Boston, MA, USA: Pearson.

Fotache, M., & Cogean, D. (2013). NoSQL and SQL Databases for Mobile Applications. Case Study: MongoDB versus PostgreSQL. *Informatica Economica, 17*(2), 41-58. doi:10.12948/issn14531305/17.2.2013.04

Francis, M. (2012, April 2). *Future telescope array drives development of exabyte processing.* Retrieved October 3rd, 2013, from ars technica: http://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/

Gobble, M. M. (2013, Jan/Feb). Big Data: The Next Big Thing in Innovation. *Research Technology Management, 56*(1), 64-66. doi:10.5437/08956308x5601005

Higginbotham, S. (2010, September 13). *Sensor Networks Top Social Networks for Big Data.* Retrieved October 2nd, 2013, from Gigaom: http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/

Kimbal, R., & Ross, M. (2002). *The Data Warehouse Toolkit* (2nd ed.). New York: Wiley.

Leavitt, N. (2010, Febrary). Will NoSQL Databases Live Up to Their Promise? *Computer*, 12-14. doi:10.1109/MC.2010.58

Lescovec, J. (2012). *Stanford Large Network Dataset Collection.* Retrieved October 11th, 2013, from Stanford University: http://snap.stanford.edu/data/index.html

Liu, Y., Wang, Y., & Jin, Y. (2012). Research on The Improvement of MongoDB Auto-Sharding in Cloud Environment. *The 7th International Conference on Computer Science & Education* (pp. 851-854). Melbourne: IEEE.

Nunan, D., & Di Domenico, M. (2013, July 1st). Market research and the ethics of big data. *International Journal of Market Research, 55*(4), 2-13. Retrieved October 2nd, 2013

Parker, Z., Poe, S., & Vrbsky, S. V. (2013). Comparing NoSQL MongoDB to an SQL DB. *Proceedings of the 51st ACM Southeast Conference - Article No 5.* New York: ACM.

Plugge, E., Membury, P., & Hawkins, T. (2010). *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing.* (F. Pohlmann, & M. M. Lowman, Eds.) New York, New York State, USA: Apress.

Pratt, P. J., & Last, M. Z. (2006). *A Guide to MySQL.* Boston: Course Technology.

Ruflin, N., Burkhart, H., & Rizzotti, S. (2011). Social-Data Storage-Systems. *Databases and Social Networks (DBSocial '11)* (pp. 7-12). New York: ACM. doi:10.1145/1996413.1996415

The Open University. (2011, July 15). *The database development life cycle.* Retrieved September 27, 2013, from Open Learn: http://www.open.edu/openlearn/science-maths-technology/computing-and-ict/information-and-communication-technologies/the-database-development-life-cycle/content-section-0

Tiwari, S. (2011). *Professional NoSQL.* Indianapolis, USA: Wiley.

Tudorica, B. G., & Bucur, C. (2011). A comparison between several NoSQL databases with comments and notes. *Roedunet International Conference (RoEduNet), 2011 10th* (pp. 1,5, 23-25). Iasi: ARNIEC/RoEduNet. doi:10.1109/RoEduNet.2011.5993686

Vaish, G. (2013, April). *Comparative Study of NoSQL Products.* Retrieved from Packt Publishing: http://www.packtpub.com/article/comparative-study-nosql-products

van der Veen, J. S., & Meijer, R. J. (2012). Sensor Data Storage Performance: SQL or NoSQL, Physical or Virtual. *2012 ieee 5th international conference on cloud computing* (pp. 431-438). IEEE. doi:10.1109/CLOUD.2012.18

Welling, L., & Thomson, L. (2005). *PHP and MySQL Web Development* (4th ed.). Indianapolis: Sams Publishing.

Witten, I. H., Eibe, F., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools & Techniques* (3rd ed.). Burlington, MA: Morgan Kaufman.