View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universiti Utara Malaysia: UUM eTheses

A NEW AUDITING MECHANISM FOR OPEN SOURCE NOSQL DATABASE – A CASE STUDY ON OPEN SOURCE MONGODB DATABASE

HANY HEIDAR HUSSEIN MOHAMED

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

SCHOOL OF COMPUTING

COLLEGE OF ARTS AND SCIENCES

UNIVERSITI UTARA MALAYSIA

2015

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to :

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

MongoDB adalah satu contoh sistem pengurusan pangkalan data NoSQL yang agak baru di pasaran pangkalan data dan ia digunakan dalam banyak projek penting dan produk. Analisis Keselamatan untuk MongoDB mendedahkan bahawa ia tidak memberikan apa-apa kemudahan untuk tindakan audit dilakukan dalam pangkalan data. Baru-baru ini, syarikat MongoDB cuba untuk membetulkan jurang pengauditan dengan menyediakan MongoDB perusahaan baru versi 2.6 (8 April 2014). Sistem pengauditan boleh merakam operasi berikut: skema (DDL), set replika, pengesahan dan kebenaran, dan operasi umum. Tetapi malangnya ia masih tidak boleh merakam Data Manipulasi Bahasa (DML). Oleh itu, kajian ini bertujuan untuk meningkatkan fungsi pengauditan di MongoDB dengan membentangkan satu mekanisme baru bagi pengauditan pangkalan data NoSQL MongoDB untuk memasukkan Data Manipulasi Bahasa (DML) / CRUD (Membuat, Baca, Kemaskini dan memadam) operasi.

Kata Kunci: data big, NoSQL, MongoDB, MongoDB pengauditan

Abstract

MongoDB as a NoSQL database management system is relatively new on the database market and it is used in many important projects and products. Security analysis for MongoDB revealed that it doesn't provide any facilities for auditing actions performed in the database. Recently, MongoDB company tried to rectify the auditing gap by providing MongoDB new enterprise version 2.6 (8th of April 2014). The auditing system logs operations information including; schema data definition language operations and operations related to replica set in addition to operations of authentication and authorization, and eventually general operations. But unfortunately still cannot record Data Manipulation Language (DML). Thus, this study aims to improve the auditing functionality in MongoDB by presenting a new mechanism for auditing NoSQL MongoDB database to include Data Manipulation Language (DML)/ CRUD (Create, Read, Update and delete) operations.

Keywords: Big data, NoSQL, MongoDB, MongoDB auditing

Acknowledgement

All praises and thanks due to Almighty Allah, the most gracious and the most merciful for lightening my way throughout the completion of this valuable thesis. I adore His benevolence and mercy, without his kindness, I will not be able to complete this study especially as I was thousand miles away from my beloved country (Egypt). Also I would like to thanks to my Supervisor Dr. Massudi Mahmuddin. Without his patient support, enlightened guidance, it is impossible for me to complete and enhance the quality of my work.

My deepest and heartfelt gratitude, loves, thanks and appreciation for my dearest parents and my beloved siblings who are a part of my happiness, success, and the inspiration that led me for the quest for knowledge and self-empowerment through night and day. I hope I can put a smile on their faces for giving back their remendous support and encouragement, patience, unconditional love, and prayers for me. Thank you for giving me the strength to chase and reach my dreams.

Thank You All.

"This Thesis is only the beginning of my journey."

HANY HEIDAR HUSSEIN MOHAMED

UUM University, Kedah, Malaysia

Monday, January 12, 2015

Table of Contents

Permission to Use i
Abstrak ii
Abstract iii
Acknowledgement iiv
Table of Contents v
List of Tables vii
List of Figures viii
List of Abbreviations ix
CHAPTER ONE : INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement 3
1.3 Research Questions 5
1.4 Research Objectives 5
1.5 Rsearch Scope 6
1.6 Contributions 6
1.7 Report Organization 6
CHAPTER TWO: LITERATURE REVIEW 7
2.1 NoSQL Database 7
2.1.1 Overview 7
2.1.2 NoSQL Data Model 7
2.1.2.1 Key-value Data Model 11
2.1.2.2 Column Data Model 11
2.1.2.3 Document Data Model 12
2.1.2.4 Graph Data Model 13
2.2 Importance of NoSQL in Big Data Applications 13
2.3 NoSQL Database Security Issues 14
2.3.1 Threats Posed By Distributed Environments 15
2.3.2 Safeguarding Integrity 15
2.3.3 Communication between Nodes 15
2.3.4 Sharded Data/Fragmented Data 16
2.3.5 Compromised Clients 16
2.3.6 Protection of Data at Rest 17
2.3.7 Challenges in Enforcing Access Control 17
2.3.8 Administrative Data Access 17
2.3.9 Configuration and Patch Management 18

2.3.10 Firewalls……..18
2.3.11 Authentication Clients……..18
2.3.12 Audit and Logging…….19
2.3.13 Monitoring, Filtering, and Blocking…….19
2.3.14 API security…….19
2.4 NoSQL MongoDB Database 20
2.4.1 Overview……..……...20
2.5 NoSQL Database Auditing 25
2.5.1 Database Auditing Definition…….……...25
2.5.2 Importance of NoSQL Database Auditing…….…….……...26
2.5.3 NoSQL DBMS Auditing……..…….…….……..…...27
2.5.4 MongoDB Database Auditing…….……...29
2.6 Related Work 31
2.7 Conclusion 34
CHAPTER THREE: RESEARCH MOTHODOLOGY 37
3.1 Introduction 37
3.2 Research Methodology 37
3.2.1 Stage 1: Identifying MongoDB auditing features.…….…...38
3.2.1.1 Analysis of Logging Techniques in MongoDB..……39
3.2.1.2 MongoDB Auditing Features……..……41
3.2.1.3 Auditing Events and Filters……....……41
3.2.2 Stage 2: Develop Auditing Mechanism……....…….…...44
3.2.3 Stage 3: Evaluation……...…….…...51
3.3 Conclusion 53
CHAPTER FOUR: RESULTS AND EVALUATION 54
4.1 Introduction 54
4.2 Auditing Mechanism Prototype 54
4.3 The Experiment Results 56
4.3.1 Results of Auditing the CRUD/DML Operations..…….…....61
4.4 Performance Evaluation 62
4.5 Summary 75
CHAPTER FIVE: CONCLUSION AND FUTURE WORK 77
5.1 Introduction 77
5.2 Conclusion 77
5.3 Limitations 78
4.4 Future Work 79
REFERENCES 80
APPEDIX A 87

List of Tables

Table 2.1: List of companies using NoSQL database with its categories ….....8

Table 2.2: NoSQL data stores …..…..9

Table2.3: MongoDB vs SQL terms …..…..23

Table 2.4: Auditing types and descriptions …..…..26

Table 2.5: Auditing in NoSQL databases …..…..28

Table 2.6: Examples of DML/CRUD operations …..…..30

Table 2.7: Sample of MongoDB related work …..…..33

Table 3.1: Description of event message fields …..…..40

Table 3.2: MongoDB auditing system records the following actions…..…..41

Table 3.3: DML/CRUD operations: compare MongoDB vs MySQL …..…..51

Table 3.4: Description of data sets used in the evaluation stage ……..…..52

Table 4.1: Time of the select operations in both data sets (ms)……..…..63

Table 4.2: Time of the insert operations in both data sets (ms).……..…..66

Table 4.3: Time of the remove/delete operations in both data sets (ms).…….. …..69

Table 4.4: Time of the update operations in both data sets (ms)…….. …..64

List of Figures

Figure 2.1: MongoDB architecture 24

Figure 3.1: Stages of research methodology 35

Figure 3.2: MongoDB security architecture 38

Figure 3.3: Sample of MongoDB auditing system messages 39

Figure 3.4: The architecture of auditing mechanism 42

Figure 3.5: MongoDB Database 43

Figure 3.6: Sample of audit trail records 44

Figure 3.7: Sample of data recorded in MongoDB database. 47

Figure 3.8: MongoDB auditing algorithm flowchart 49

Figure 3.9: Libraries used in the C# code 50

Figure 4.1: Auditing mechanism prototype 55

Figure 4.2: Fields of prototype records 55

Figure 4.3: Details field of the prototype output 55

Figure 4.4: Data Set 1 “AuditData” 56

Figure 4.5: Sample of data in “Department” collection in data set one “AuditData” . 57

Figure 4.6: Sample of data in “Employee” collection in data set one “AuditData” 57

Figure 4.7: Sample of data in “Department” collection in data set one “AuditData”.. 57

Figure 4.8: Dataset 2 “AuditData2” description 57

Figure 4.9: Sample of data in “Course” collection in data set two “AuditData2” 58

Figure 4.10: Sample of data in “Lecturer” collection in data set two “AuditData2” ... 58

Figure 4.11: Sample of data in “Student” collection in data set two “AuditData2” 58

Figure 4.12: Auditing of the query (select) operation for data set 1 59

Figure 4.13: Auditing of the update operation for data set 1 59

Figure 4.14: Auditing of the insert operation for data set 1 60

Figure 4.15: Auditing of the remove (delete) operation for data set 1 62

Figure 4.16: Auditing of the query (select) operation for data set 2. 61
Figure 4.17: Auditing of the update operation for data set 2. 61
Figure 4.18: Auditing of the insert operation for data set 2. 61
Figure 4.19: Auditing of the remove operation for data set 2. 62
Figure 4.20: Time of the select operation for Data Set 1 before and after applying the proposed auditing mechanism..... 64
Figure 4.21: Time of the select operation for Data Set 2 before and after applying the proposed auditing mechanism..... 65
Figure 4.22: Time of the insert operation for Data Set 1 before and after applying the proposed auditing mechanism..... 67
Figure 4.23: Time of the insert operation for Data Set 2 before and after applying the proposed auditing mechanism..... 68
Figure 4.24: Time of the remove/delete operation for Data Set 1 before and after applying the proposed auditing mechanism..... 70
Figure 4.25: Time of the remove/delete operation for Data Set 2 before and after applying the proposed auditing mechanism..... 71
Figure 4.26: Time of the update operation for Data Set 1 before and after applying the proposed auditing mechanism..... 73
Figure 4.27: S Time of the update operation for Data Set 2 before and after applying the proposed auditing mechanism..... 74

List of Abbreviations

ACID	Atomicity, Consistency, Isolation, Durability
BSON	Binary JavaScript Object Notation
CRUD	Create Read Update Delete
DDL	Data Definition Language
DML	Data Manipulation Language
DBA	Database Administrators
JSON	JavaScript Object Notation
NoSQL	Not Only SQL
RBAC	Role Based Access Control
RDMS	Relational Database Management System
RFID	Radio Frequency Identification
RPC	Remote Procedure Call
OS	Operating System

SQL	Structured Query Language
TCP/IP	Transmission Control Protocol /Internet Protocol

CHAPTER ONE
INTRODUCTION

1.1 Background

The term NoSQL is used first time by Mr. Carlo Strozzi (1998) to name his lightweight open source relational database. The system did not expose the standard SQL (Structure Query Language) interface. There is a series of database following NoSQL (Not Only SQL) standards. The term "Not Only SQL" is also used for these databases that provide storage and retrieval mechanism with less constrained consistency models than traditional relational databases (Mohamed, Altrafi, & Ismail, 2014).

The last three decades were ruled by the traditional relational database management systems such as DB2, MS SQL Server and Oracle (Bonnet, Laurent, Sala, Laurent, & Sicard, 2011). They have the standard SQL. Due to the growing web scale applications such as Facebook, mobile applications and RFID (Radio Frequency Identification) the Internet has become an essential part of the world today.

Everyday zettabytes of data are being generated due to these applications. Due to changing need of applications and databases, the traditional relational databases are proved to be weak in distributed environment. This made NoSQL databases to get importance and preference. Being schema free, elastic and scalable, NoSQL databases appeared to be effective(Kanade, Gopal, & Kanade, 2013).

The contents of the thesis is for internal user only

REFERENCES

Apache CouchDB. (2014).Retrieved 12 March, 2014 from Apache CouchDB: http: //couchdb. apache.org/

Apache HBase. (2014). HBase - Apache HBase™ Home. Retrieved 12 March, 2014 from http://hbase.apache.org

Boicea, A., Radulescu, F., & Agapin, L. I. (2012). MongoDB vs Oracle-Database Comparison. *2012 Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)* (pp. 330-335).

Bonnet, L., Laurent, A., Sala, M., Laurent, B., & Sicard, N. (2011). Reduce, you say: What nosql can do for data aggregation and bi in large repositories. *In Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on* (pp. 483-488). IEEE.

Buerli, M., & Obispo, C. P. S. L. (2012). The Current State of Graph Databases. Retrieved 7 May, 2014 from http://www.cs.utexas.edu: http://www.cs.utexas.edu~cannata/dbms/Class%20Notes/09%20Graph_Databases_Survey.pdf

Couchbase Server the NoSQL document database. (2014). Couchbase Server Distributed, Non-Relational Database Couchbase. Retrieved from http://www.couchbase.com/couchbase-server/overview

Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. Communications of the ACM, 53(1), 72-77.

Ezumalai, R., & Aghila, G. (2009). Combinatorial approach for preventing SQL Injection attacks. In *Advance Computing Conference, 2009. IACC 2009. IEEE International* (pp. 1212-1217). IEEE.

Geer, D. (2005). Malicious bots threaten network security. *Computer*, *38*(1), 18-20.

Dijcks, J. P. (2012). Oracle: Big data for the enterprise. Oracle White Paper.

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East. IDC iView: IDC Analyze the Future. Retrieved 12 March, 2014 from www.emc.com: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). The Google file system. In ACM SIGOPS Operating Systems Review (Vol. 37, No. 5, pp. 29-43). ACM.

Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and Applications*, (2), *2-22*.

Hecht, R., & Jablonski, S. (2011). NoSQL Evaluation. *International Conference on Cloud and Service Computing*, (pp. 337-341).

Hsu, W. C., Huang, J. Y., Chen, C. H., Su, C. Y., Shih, H. C., Liao, T. Y., & Liao, I. E. (2013). A cloud service for the evaluation of company's financial health using XBRL-based financial statements. In *Big Data, 2013 IEEE International Conference on* (pp. 10-14). IEEE.

Kadebu, P., & Mapanga, I. (2014). A Security Requirements Perspective towards a Secured NOSQL Database Environment. *International Conference of Advance Research and Innovation (ICARI-2014), (3), 472-480.*

Kanade, A., Gopal, A., & Kanade, S. (2014, February). A study of normalization and embedding in MongoDB. In *Advance Computing Conference (IACC), 2014 IEEE International* (pp. 416-421). IEEE.

Kanade, A. S., Gopal, A., & Kanade, S. (2013). Cloud Based Databases-A Changing Trend. International Journal of Management, IT and Engineering, 3(7), 273-287.

Lawrence, R. (2014). Integration and Virtualization of Relational SQL and NoSQL Systems Including MySQL and MongoDB. In Computational Science and Computational Intelligence (CSCI), 2014 International Conference on (Vol. 1, pp. 285-290). IEEE.

Li, Y., & Manoharan, S. (2013). A performance comparison of SQL and NoSQL databases. In Communications, Computers and Signal Processing (PACRIM), 2013 IEEE Pacific Rim Conference on (pp. 15-19). IEEE.

Liang, J., & Mizuno, O. (2011). Analyzing Involvements of Reviewers Through Mining A Code Review Repository. In *Software Measurement, 2011 Joint Conference of the 21st Int'l Workshop on and 6th Int'l Conference on Software Process and Product Measurement (IWSM-MENSURA)* (pp. 126-132). IEEE.

Liu, L., & Huang, Q. (2009). A framework for database auditing. In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on (pp. 982-986). IEEE.

Liu, Y., Wang, Y., & Jin, Y. (2012). Research on the improvement of MongoDB Auto-Sharding in cloud environment. In *Computer Science & Education (ICCSE), 2012 7th International Conference on* (pp. 851-854). IEEE.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Mapanga, I., & Kadebu, P. (2013). Database Management Systems: A NoSQL Analysis. International journal of Modern Communication Technologies and Research, 1(7), 12-18.

Mohamed, M.A., Altrafi, O.G., & Ismail, M. O. (2014). Realtional vs. NoSQL A survey. *International Journal of Computer and Information Technology*, 3(3), 589-601

MongoDB. (2014). Retrieved 1 March, 2014 from http://www.mongodb.org/

Mullins, C. S. Retrieved 9 May, 2014 from www.oowidgets.com: http://www.oowidgets.com/Database%20Auditing%20Essentials.pdf

Murugesan, P., & Ray, I. (2014). Audit Log Management in MongoDB. In Services (SERVICES), 2014 IEEE World Congress on (pp. 53-57). IEEE.

Narde, R. (2013). A Comparison of NoSQL systems (Doctoral dissertation, Rochester Institute of Technology).

Neo4j. (2014). Neo4j - The World's Leading Graph Database. Retrieved 12 March, 2014 from http://www.neo4j.org/

Ohlhorst, F. J. (2012). Big data analytics: turning big data into big money. John Wiley & Sons.

Okman, L., Gal-Oz, N., Gonen, Y., Gudes, E., & Abramov, J. (2011). Security issues in nosql databases. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on (pp. 541-547). IEEE.

Pavlenko, D. (2014). MongoDB Audit Logging or How to Log Data Changes Using MongoDB. Retrieved 12 March, 2014 from sysgears.com: http://sysgears.com/articles/mongodb-audit-logging-or-how-log-data-changes-using-mongodb/

PCI Security Standards Council. (2010). Payment card industry (pci) data security standard – requirements and security assessment pro-cedures version 2. 0. Wakefield, MA, USA: Author. Retrieved 17 March, 2014 from:https://www.pcisecuritystandards.org/documents/pcidss v2.pdf

Pozzani, G. (2013). Introduction to NoSQL. Retrieved 21 March, 2014 from profs.sci.univr.it:http://profs.sci.univr.it/~pozzani/Materiale/nosql/01%20-%20introduction.pdf

Stonebraker, M., Madden, S., Abadi, D. J., Harizopoulos, S., Hachem, N., & Helland, P. (2007). The end of an architectural era :(it's time for a complete rewrite). In Proceedings of the 33rd international conference on Very large data bases (pp. 1150-1160). VLDB Endowment.

Rutishauser, N. (2012). *TPC-H applied to MongoDB: How a NoSQL database performs* . Retreived 21 March,2014 from www.ifi.uzh.ch: http://www.ifi.uzh.ch/dbtg/teaching/thesesarch/VertiefungRutishauser.pdf

Truică, C. O., Boicea, A., & Trifan, I. (2013). CRUD Operations in MongoDB. Paper presented at the International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013). (pp. 347-250).

White, T. (2009). Hadoop: The Definitive Guide. O'Reilly Media, Inc.

Tudorica, B. G., & Bucur, C. (2011). A comparison between several NoSQL databases with comments and notes. In Roedunet International Conference (RoEduNet), 2011 10th (pp. 1-5). IEEE.

US Department of Health and Human Services. (2013). The Health Insurance Portability and Accountability Act of 1996: health information privacy. US Department of Health and Human Services website. Retrieved 30 July, 2013.from:http://www.hhs.gov/ocr/privacy/.

Valley Programming. (2014). Big data datasets (large dataset examples) Boulder, Colorado. Retrieved 21 March, 2014, from www.valleyprogramming.com: http://www.valleyprogramming.com/blog/big-data-datasets-large-examples-boulder-colorado-hadoop-mongodb

Van der Veen, J. S., Van der Waaij, B., & Meijer, R. J. (2012). Sensor data storage performance: Sql or nosql, physical or virtual. In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on (pp. 431-438). IEEE.

Venable, J. & Kuechler B, (2006), The Role of Theory and Theorising in Design Science Research, First International Conference on Design Science Research in Information Systems and Technology, Claremont, California, pp. 1-18.

Wisseman, S., Wilson, B., & Wichers, D. (1996). Trusted Database Management System Interpretation of the Trusted Computer System Evaluation Criteria. Diane Publishing Co.

Zagarese, Q., Canfora, G., Zimeo, E., & Baude, F. (2012). Enabling advanced loading strategies for data intensive web services. In *Web Services (ICWS), 2012 IEEE 19th International Conference on* (pp. 480-487). IEEE.