INDEXING STRATEGIES OF MAPREDUCE FOR INFORMATION RETRIEVAL IN BIG DATA

MAZEN FARID EBRAHIM RAMADHAN

FSKTM 2016 25

INDEXING STRATEGIES OF MAPREDUCE FOR INFORMATION RETRIEVAL IN BIG DATA

By

MAZEN FARID EBRAHIM RAMADHAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Computer Science (Distributed Computing)

January 2016

DEDICATIONS

This thesis is dedicated to my parents and my wife who have supported me all the way since the beginning of my studies.

Also, this thesis is dedicated to all those who believe in the richness of learning.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Computer Science

INDEXING STRATEGIES OF MAPREDUCE FOR INFORMATION RETRIEVAL IN BIG DATA

By

MAZEN FARID EBRAHIM RAMADHAN

January 2016

Chair: Dr. Rohaya Latip

Faculty: Computer Science and Information Technology

In Information Retrieval (IR) the efficient strategy of indexing large dataset and terabyte-scale data is still an issue because of information overload as the result of increasing the knowledge, increasing the number of different media, increasing the number of platforms, and increasing the interoperability of platforms. Overall multiple processing machines MapReduce has been suggested as a suitable platform that use for distributing the intensive data operations. In this project, Sensei and Per-posting list indexing, Terrier will be analysed as they are the two most efficient MapReduce indexing strategies. The two indexing will be implemented in an existing framework of IR, and an experiment will be performed by using the Hadoop for MapReducing with the same large dataset, and try to analyse and verify the better efficient strategy between Sensei and Terrier. The experiment will measure the performance of retrieving when the size

and processing power enlarge. The experiment examines how the indexing strategies scaled and work with large size of dataset and distributed number of different machines. The throughput will be measured by using MB/S (megabyte/per second), and the experiment results analyzing the performance of delay, consuming time and efficiency of indexing strategies between Sensei and Per-posting list indexing ,Terrier.

Abstrak **t**esis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Komputer Sains

INDEXING STRATEGIES OF MAPREDUCE FOR INFORMATION RETRIEVAL IN BIG DATA

Oleh

MAZEN FARID EBRAHIM RAMADHAN

Januari 2016

Pengerusi: Dr. Rohaya Latip

Fakulti:-Sains Komputer Dan Teknologi Maklumat

Dalam Maklumat Retrieval (IR) strategi yang cekap pengindeksan dataset besar dan data berskala terabyte masih merupakan masalah kerana beban maklumat sebagai hasil daripada peningkatan pengetahuan, meningkatkan bilangan media yang berbeza, menambah bilangan plantar, dan meningkatkan antara operasi plantar. Secara keseluruhan mesin pemprosesan pelbagai MapReduce telah dicadangkan sebagai plantar yang sesuai yang digunakan untuk mengagihkan operasi data intensif. Dalam projek ini, Sensei pengindeksan dan senarai Per-posting pengindeksan, Terrier akan di kerana ianya adalah dua strategi MapReduce pengindeksan paling berkesan. Kedua-dua pengindeksan akan dilaksanakan dalam rangka kerja yang sedia ada IR, dan eksperimen akan dilakukan dengan menggunakan Hadoop untuk MapReduce dengan dataset besar yang sama, dan cuba untuk mencari dan mengesahkan strategi yang lebih baik berkesan antara Sensei and Terrier. Eksperimen akan mengukur prestasi mendapatkan semula apabila saiz dan kuasa pemprosesan di besarkan.

Eksperimen mengkaji bagaimana strategi pengindeksan diskalakan dan bekerja dengan saiz besar dataset dan nombor agilan mesin yang berbeza. Throughput ini akan diukur dengan menggunakan MB/S (megabait/sesaat), dan keputusan eksperimen menganalisis prestasi kelewatan, masa dan kecekapan strategi pengindeksan antara Sensei dan Per-posting senarai pengindeksan Terrier yang lama.

ACKNOWLEDGMENTS

I would like to thank all the people who contributed in some way to the work described in this thesis. I thank my academic advisor, Dr. Rohaya Latip who has been a great source of motivation and inspiration. Also for her unlimited support to get every results described in this thesis, supporting my attendance at ICDCB 2016 that will be on 14-15 April 2016.

I certify that a Thesis Examination Committee has met on 18 January 2016 to conduct the final examination of Mazen Farid Ebrahim Ramadhan on his thesis entitled “**Indexing Strategies of MapReduce for Information Retrieval in Big Data”** in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the master degree.

Members of the Thesis Examination Committee were as follows:

Dr. Rohaya Latip, Ph.D.
Computer Science and Information Technology
Universiti Putra Malaysia
(Supervisor)

Dr. Masnida Ph.D.
Computer Science and Information Technology
Universiti Putra Malaysia
(Assessor)

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

MAZEN FARID EBRAHIM

Date:18-January-2016

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF ABBREVIATIONS

HDFS	Hadoop Distributed File System
WAN	Wide Area Network
IR	Information retrieval
MapReduce	Framework for processing large data sets in cloud.

CHAPTER 1

INTRODUCTION

1.1 Background

The Web is a large place to store documents, and has the main challenge for Information Retrieval systems (IR), which used by search engines of the Web or Web IR researchers. Index should be created in order to increase the efficiency of documents information retrieval. Index is considered the special data structure which used in this regard. The dataset usually contains many documents which are stored in one hard disk and sometimes in more than one hard disk. The indexing, therefore, should cover different or many hard disks in which the documents are stored. The IR system works through inverted index in which every term has a posting list. This posting list represents the documents through numbers or integer documents (IDs), and they also contain the terms as stated by (McCreadie et al., 2012). Every document has a representing score which stored in the posting list. The importance of this process is to figure out the information sufficient statistics. The main goal here is to find the sufficient statistics of information, e.g. the terms frequencies which are occurred, and the information of position.

The textual indexes usually stored with lexicon which are considered additional structures. These lexicons include pointers which are important for the posting list that added to the inverted index. In the final stage the result of the search can be shown through using the information of the documents such as its name and its length. This is used to display the documents in a specific order for the user. This is what is called indexing and an important point is that it should be in the mode of offline before the process of searching is done.

The single pass indexing used in Terrier system that released with Terrier 2.0 (Macdonald et al., 2012) this term is used to describe the idea of building the document of single pass central structure over the collection. In addition, single pass use low memory consumption when creating the index. This process is achieved through compressing the inverted files as well as creating a temporary posting list that used in one single machine only. According to memory consumption, this type of indexing is considered the most efficient and the faster.

In distributed systems is that Terrier supports huge dataset indexing through using the functions of Hadoop's MapReduce by using a single pass indexer. There are three organizations for the output of functions of the Map. The first organization is information saved about the document during its run. The second organization is the indices of each document for every map task. The third organization is the list of the term as well as its posting list.

When the space of the memory and power of processing is limited, the sharding strategy can be used. Sensei used this strategy through the creation of the fast MapReduce job. This is achieved through taking utilizing the data from Hadoop with the given schema. Sensei cannot support the JOINS, but it can create a single index. Therefore, the aim of this study is to explain and clarify the advantages of the process of indexing large datasets through utilizing MapReduce.

1.2 Problem Statement

In Information Retrieval (IR) (McCreadie et al., 2012), the efficient strategy of indexing large dataset and terabyte-scale data is still an issue because of information overload as the result of increasing the knowledge, increasing the number of different media, increasing the number of platforms, and increasing the interoperability of platforms. MapRduce is the efficient framework that has been suggested to work with large dataset and large number of machines.

1.3 Research Objectives

The main objective is to evaluate the indexing strategies through the below goals:

1. Performing experiments over large dataset to analyze and determine the efficient indexing strategy.
2. Testing the performance of indexing strategies that applying on dataset.
3. To report and support the most efficient indexing strategy in MapReduce for large number of machines.

1.4 Research Scope

The study concentrates on Terrier and Sensei indexing architecture by using their open source frameworks. Both of these strategies provide new data structure called indexing with Hadoop features. This study investigates which of these two strategies can index the large datasets with less time consuming. This can be done by applying MapReduce job on 10 machines, 12 map tasks and 4 reduce tasks.

REFERENCES

Abdulkarem, M., & Latip, R. (2015). Data Transmission Performance Analysis in Cloud and Grid. *ARPN Journal of Engineering and Applied Sciences* .Vol. 10, no. 18.

Chen, C. H., Hsieh, S. H., Su, Y. S., Hsu, K. P., Lee, H. H., & Lai, F. (2012). Design and Implementation of Web-based Discharge Summary Note Based on Service-oriented Architecture. *Journal of Medical Systems*, *36*(1), 335-345.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, *51*(1), 107-113.

Elsayed, T., Ture, F., & Lin, J. (2010). Brute-Force Approaches to Batch Retrieval: Scalable Indexing with MapReduce, or Why Bother?. *Technical Report HCIL-2010-23,* University of Maryland, College Park, Maryland.

He, B., & Ounis, I. (2006). Query Performance Prediction. *Information Systems*, *31*(7), 585-594.

He, C., Weitzel, D., Swanson, D., & Lu, Y. (2012). Hog: Distributed Hadoop MapReduce on the Grid. In *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:* (pp. 1276-1283). IEEE.

Heinz, S., & Zobel, J. (2003). Efficient Single-pass Index Construction for Text Databases. *Journal of the American Society for Information Science and Technology*, *54*(8), 713-729.

http://senseidb.github.io/sensei/overview.html.

http://Terrier.org.

Macdonald, C., McCreadie, R., Santos, R. L., & Ounis, I. (2012). From Puppy to Maturity: Experiences in Developing Terrier. *Open Source Information Retrieval*, 60.

McCreadie, R., Macdonald, C., & Ounis, I. (2009). Comparing Distributed Indexing: To MapReduce or Not?. *Proc. LSDS-IR*, 41-48.

McCreadie, R., Macdonald, C., & Ounis, I. (2012). MapReduce Indexing Strategies: Studying Scalability and Efficiency. *Information Processing & Management*, *48*(5), 873-888.

Mo, X., & Wang, H. (2012). Asynchronous Index Strategy for High Performance Real-time Big Data Stream Storage. In *Network Infrastructure and Digital Content (IC-NIDC), 2012 3rd IEEE International Conference on* (pp. 232-236). IEEE.

Mohamed, H., & Marchand-Maillet, S. (2013). MRO-MPI: MapReduce Overlapping Using MPI and an Optimized Data Exchange Policy. *Parallel Computing*, *39*(12), 851-866.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of the OSIR Workshop* (pp. 18-25).

Patel, A. B., Birla, M., & Nair, U. (2012). Addressing Big Data Problem Using Hadoop and Map Reduce. In *Engineering (NUiCONE), 2012 Nirma University International Conference on* (pp. 1-5). IEEE.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on* (pp. 1-10). IEEE.

Yu, G., Xie, X., & Liu, Z. (2010). The Design and Realization of Open-source Search Engine Based on Nutch. In *Anti-Counterfeiting Security and Identification in Communication (ASID), 2010 International Conference on* (pp. 176-180). IEEE.

Yu, W., Wang, Y., & Que, X. (2014). Design and Evaluation of Network-Levitated Merge for Hadoop Acceleration. *Parallel and Distributed Systems, IEEE Transactions on*, *25*(3), 602-611.

Zhang, Ji, Xunfei Jiang, Wei-Shinn Ku, and Xiao Qin.(2015). Efficient Parallel Skyline Evaluation using MapReduce, *IEEE Transactions on Parallel and Distributed Systems*, 2015.

BIODATA OF STUDENT

Mazen Farid Ebrahim was born in Yemen on 24th April 1979. He obtained Degree in Computer Science and Engineering from Aden University College on 2004. He peruses his Master of Computer Science and Information technology majoring in Distributed Computing Department at Universiti Putra Malaysia by focusing in Retrieving Information in BIG DATA using different indexing strategies of MapReduce on 2015 for his final project. His professional working experience includes 8 years of service as a Lecturer at Aden University for computer’s science. Recently, he did some publication in this area.