Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department

Cloud-based Data Warehouse Management Systems

A thesis submitted in partial fulfillment of the requirements for

the degree of MSc in Computer and Information Sciences

To

Information Systems Department,

Faculty of Computer and Information Sciences,

Ain Shams University

By

Mohammed Ezzat Megahed Shaaban

Under Supervision of

Prof. Dr. Mohamed Fahmy Tolba

Scientific Computing Department

Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Nagwa Lotfy Badr

Information Systems Department

Faculty of Computer and Information Sciences, Ain Shams University

Assoc. Prof. Dr. Rasha Mohamed Ismail

Information Systems Department

Faculty of Computer and Information Sciences, Ain Shams University

2017

Acknowledgement

First and above all, all the thanks and praises to Allah for everything we have in our life, for this thesis, our knowledge, and for everything.

I would like to gratefully thank Prof. Dr., Mohamed Fahmy Tolba, for his support and for sharing his pearls of wisdom and his deep knowledge with us during this research.

I would like to express my sincere gratitude to Dr. Nagwa Badr for the useful comments, remarks and engagement through the process of this master thesis.

Also, I would like to express my gratitude and appreciation to Dr. Rasha Ismail for the continuous support of my master's study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis.

I am using this opportunity to express my gratitude to everyone who supported me throughout the journey; my colleagues: Ghonemy, Hala, Emad and Salah who provided insights and expertise that greatly assisted the research.

Last but not the least; I would like to thank my family: my dad, my mom, and my sisters for supporting me spiritually throughout writing this thesis and my life in general. My deepest appreciation is expressed to them for their love, understanding, and inspiration.

Abstract

Cloud computing is becoming increasingly popular as it enables users to save both development and deployment time. It also reduces the operational costs of using and maintaining the systems. Moreover, it allows the use of any resources with elasticity instead of predicting workload which may be not accurate. There are many technologies that can be merged with Cloud computing to gain more benefits. One of these technologies is data warehousing, which can benefit from this trend when it's used to save large amounts of data with unpredictable sizes and if used in distributed environments. Large amounts of data are generated daily, according to the wide usage of social media websites, scientific data and all live matters; these amounts of data should be utilized.

The Big data storage is one of the most critical tasks as it will affect the data access, analysis, retrieve and also query answering process so it can help decision makers but the traditional database concepts are insufficient. Cloud computing is very essential in big data storage process, as big data will utilize the cloud features as the elasticity. One of the most beneficial algorithms to deal with the big data is the MapReduce algorithm.

In this thesis a system is presented which consists of: a cloud based view allocation algorithm which is presented to enhance the performance of the data warehousing system over a Peer-to-Peer architecture. The proposed approach improves the allocation of the materialized views on cloud peers. It also reduces the cost of the dematerialization process and furthermore. the proposed algorithm saves the transfer cost by distributing the free space based on the required space to store the views and on the placement technique. Furthermore, the proposed algorithm saves the transfer cost by distributing the free space on the peers based on the required space to store the views.

We also proposed an algorithm for big data allocation based on a peer-to-peer cloud architecture integrates the OLAP and MapReduce over cloud (considering workload balance) in order to enhance the performance of query processing over big data, this data is stored in a form of cubes segmented into chunks and query answering technique will run based on MapReduce approach. This process is done using the proposed allocation approach to save

resources and query processing times. The proposed system achieves enhancements as time saving in query processing, network transfer cost and in resources usage and utilization.

The thesis is organized as follows:

Chapter 1 (Introduction): *presents an introduction to the field of cloud computing and big data. Also this chapter presents the problems on data storing and allocation on peer-to-peer cloud computing and motivation behind the proposed work. In addition the objectives of the proposed work and our contributions are listed in this chapter.*

Chapter 2 (Background): *discusses cloud computing main concepts and provides an overview of big data allocation and processing. Also it presents an overview on the viewpoints of cloud computing. Data warehouse and big data cubes.*

Chapter 3 (Related Works): *presents a survey on the related works to our research.*

Chapter 4 (An Enhanced Cloud based View Materialization Approach for Peer-to-Peer Architecture): *describes the proposed cloud allocation approach applied over the materialized views of the data warehouse. Also the results and analysis of the proposed approach is presented.*

Chapter 5 (A Peer-to-Peer Architecture for Cloud Based Data Cubes Allocation): *presents the architecture of the big data cubes allocation over the peer-to-peer cloud computing. Also it shows the experimental evaluation, and the results and analysis of the proposed algorithms.*

Chapter 6 (An Enhanced Peer-to-Peer Cloud based Chunks Allocation for Big Data Cubes): *introduces big data allocation approach over the cloud but after segmented into data chunks instead of data cubes and the experimental evaluation, and the results and analysis of the proposed algorithms.*

Chapter 7 (The Whole Architecture Evaluation): *the system evaluation includes evaluation of each algorithm and the comparisons and charts of the results.*

Chapter 8 (Conclusion and Future Work): *gives a conclusion of the proposed work. A summary of the proposed approaches and algorithms is presented in this chapter. Also, this chapter presents the directions of our future research.*

Table of Contents

Chapter **Page**

Acknowledgement..... vii

Abstract..... vii

Table of Contents..... vii

List of Figures..... vii

List of Tables..... viii

1. Introduction…… 1

1.1 Overview…… 1

1.2 Problem Definition…… 3

1.3 Objective…… 5

1.4 Contribution…….. 5

2. Background ……..... 8

2.1 Cloud Computing…… 8

2.2 Data Warehouse……. 14

2.3 View……. 15

2.4 View Materialization…….. 16

2.5 Materialized View…… 16

2.6 Big data on Cloud Computing……. 17

3. Related Works…… 20

3.1 Introduction……... 20

3.2 Cloud computing and materialized views allocation…….. 20

3.3 Cloud computing and big data cubes allocation…… 26

3.4 Cloud computing and data chunks allocation…….. 28

4. A Peer-to-Peer Architecture for Cloud Based Data Cubes Allocation……. 33

4.1 The Proposed Cloud View Allocation Architecture….. 33

4.1.1 The architecture consists of the following modules…… 33

4.1.1.1 Module One: The Materialization module…. 33

4.1.1.2 Module Two: The Placement Module…… 34

4.1.1.3 Module Three: The Resource Allocation Module……. 35

4.2 Algorithm Objectives…….. 36

5. A Peer-to-Peer Architecture for Cloud Based Data Cubes Allocation……. 43

5.1 The Proposed Data Streams Processing on Cloud Computing. 44

5.1.1 The Proposed Cloud based Big Data Cube Allocation and Processing Architecture……. 44

5.1.1.1 Module 1: Cube Construction……. 44

5.1.1.2 Module 2: Cube allocation……. 44

5.1.1.3 Module 3: Query Processing……. 44

5.1.1.4 Module 4: Query Answering…… 44

5.2 The Proposed Cloud based Big Data Cubes allocation and processing Algorithm……. 45

5.2.1 The algorithm consists of three main functions……. 45

5.2.1.1 The first function (Cube Construction)……. 45

5.2.1.2 The second part (Random allocation and Re-allocation processes)…….. 46

5.2.1.3The third part (query processing and answering)…… 47
6. An Enhanced Peer-to-Peer Cloud based Chunks Allocation for Big Data Cubes…… 51
6.1 The Proposed Cloud based Chunks Allocation and Processing Architecture……. 51
6.1.1 The first approach 4 Modules……. 51
6.1.1.1 Module 1: Cube Construction…… 51
6.1.1.2 Module 2: Cube allocation…… 51
6.1.1.3 Module 3: Query Processing……. 52
6.1.1.4 Module 4: Query Answering……. 52
6.1.2 The second approach also consists of 4 steps…… 52
6.1.2.1 Module 1: Cube segmentation…….. 52
6.1.2.2 Module 2: Chunks allocation…… 53
6.1.2.3 Module 3: Query Processing……. 53
6.1.2.4 Module 4: Query Answering……..…… 53
6.2 The Proposed Cloud based Chunks allocation and processing Algorithms…… 55
6.2.1 The algorithm consists of three main functions……. 55
6.2.1.1 The first part (Cube Construction)…… 55
6.2.1.2 The second part (Re-allocation)…… 55
6.2.2 The algorithm consists of three main parts…… 58
6.2.2.1 The first part (Cube Construction and segmentation into chunks)……. 58
6.2.2.2 The second part (Random allocation and Re-allocation processes)…….. 59

6.2.2.3 The third part (query processing and answering)…… 60
7. Evaluation……. 64
7.1 Introduction…….. 64
7.2 The View materialization cloud allocation……. 64
7.3 The cube cloud allocation……. 70
7.4 The chunks/cubes allocation…… 76
8. Conclusion and Future Work …… 85
References…….. 89
Arabic Summary…….. 114

List of Figures

Fig. 2.1 The 3 cloud architectures……... 15

Fig. 4.1 The Proposed Dynamic View Allocation Architecture……. 36

Fig. 4.2 Cloud Based Dynamic View Allocation Algorithm…….. 41

Fig. 5.1 Proposed Cloud based Big Data Cubes Allocation and Processing Architecture……..……..……..……... 43

Fig. 6.2 The Proposed Cloud based Chunks Allocation and Processing Architecture……..……..……..……..……... 54

Fig. 7.1 Comparisons between the three policies against the transfer cost.. 65

Fig. 7.2 Measurements of the saved transfer cost……..…….. 66

Fig. 7.3 Comparisons between the three policies against the processing cost……..……..……..……..……..…... 67

Fig. 7.4 Measurements of the saved processing time…….……. 68

Fig. 7.5 Comparisons between the three policies against the space usage... 69

Fig. 7.6 Results of Space Usage……..……..……..…… 70

Fig. 7.7 Resource Saving Experiment……..……..……. 72

Fig. 7.8 Workload Balance……..……..……..…… 73

Fig. 7.9 Processing Time……..……..……..……... 74

Fig. 7.10 Average time (Query Processing) ……..……... 75

Fig. 7.11 Resource Saving Experiment for chunks/cubes allocation and reallocation……..……..……..…….. 78

Fig. 7.12 Processing Time for chunks allocation and reallocation……... 79

Fig. 7.13 Average time (Query Processing) for chunks allocation and

reallocation…… 80
Fig. 7.14 Processing Time for cubes allocation and reallocation…… 82
Fig. 7.15 Average time (Query Processing) for cubes allocation and reallocation……..……..……..……..…… 82

List of Tables

Table 7.1	Comparisons between the three policies against the transfer cost (time in milliseconds) ……..……..……	65
Table 7.2	Comparisons between the three policies against the processing cost (time in milliseconds) ……..……..……..…..	67
Table 7.3	Comparisons between the three policies against the space usage (space in kilobytes) ……..……..……..…….	69
Table 7.4	Number of visits for each peer……..……..……...	72
Table 7.5	Workload Balance……..……..……..……	72
Table 7.6	Query Processing Time……..……..……..……	74
Table 7.7	Number of visits for each peer……..……..……..	77
Table 7.8	Query Processing Time for chunks allocation and reallocation…	79
Table 7.9	Query Processing Time for cubes allocation and reallocation…..	81

CHAPTER 1

INTRODUCTION

Chapter 1 Introduction

1.1 Overview

Cloud computing is the term of using any computer resource (processing, storage or network bandwidth) as a service over the internet based on a virtualization concept and is a very successful service oriented computing paradigm [1-3]. The definition of the National Institute of Standards and Technology (NIST) for Cloud computing is, "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interactions" [4]. The most used cloud computing types based on how the service is provided are: Infrastructure (IaaS), Platform (PaaS) and Software as a Service (SaaS) [5,6]. Cloud computing has many architectures; one of these architectures is the Peer-to-Peer cloud architecture which is most suitable for distributed environments as it enables data travelling between the peers, data communication directly from each peer to any other peer in the architecture, it also allows data marts distribution and communication which allows query processing over all the peers [6-8].

Data warehouse is considered one of the solutions for storing large amounts of data to be used in decision making or supporting analysis. Data warehouse has a very large size and can be geographically distributed leading to the occurrence