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

Enhancing the Database Query Workload on Cloud Computing Environment

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

Eman Amin Maghawry Amin

Teaching Assistant at Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University

Under Supervision of

Prof. Dr. Mohamed Fahmy Tolba

Professor at Scientific Computing Department
Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Nagwa Lotfy Badr

Professor at Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University

Dr. Rasha Mohamed Ismail

Associate Professor at Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University

Cairo - 2016

Acknowledgement

First and above all, I praise Allah, the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

I would like to gratefully and sincerely thank Prof. Dr. Mohamed Fahmy Tolba for his leadership, professionalism, encouragement and valuable guidance have provided a significant basis for this thesis.

Also, I would like to express my deep and sincere thanks to Dr. Nagwa Badr for her enthusiasm, inspiring instructions, understanding, patience, motivation for me and her great help throughout this thesis.

Also, I would like to express my gratitude and deepest appreciation to Dr.Rasha Ismail for her caring, endless encouragement, continuous advices and her supportive supervision has been of great value for me.

A special gratitude and love goes to my family for their unfailing support. I warmly thank and appreciate my beloved parents and my mother and father-in-law for their spiritual support in all aspects of my life. Your prayers for me was what sustained me thus far. I also would like to thank my sisters, brother, and sister-in-law, for their assistance in numerous ways.

Finally, I must express my very profound gratitude to my beloved husband for his continued support, unending encouragement and understanding throughout my research. I can't thank you enough for encouraging me throughout this experience. Also, I thank my little and beloved son for making me so happy with his cute smile. My deepest appreciation is expressed to them for their love, understanding, and inspiration. Without their blessings and encouragement, I would not have been able to finish this work.

Abstract

Cloud computing is a promising computing model that provides a combination of parallel and distributed computing paradigms. It has the characteristics of on demand provisioning of a shared pool of configurable computing resources as a service. It provides a cost effective paradigm of computational, storage and database resources to users over the internet.

Cloud storage is an important service that is provided by the cloud computing system. It provides the data owners with high accessibility, availability and scalability in respect to increasing the amount of their data in cloud repositories. The increasing number of user's requesting data from deployed virtual instances can lead to increased loads on the cloud data management system.

Multiple queries compete for hardware resources causing resource contention with rapidly changing computational properties and efficient concurrent query executions on especially structured data has become an important challenge. Moreover, it offers numerous benefits regarding better utilization of virtual instances by exploiting parallel executions.

This thesis proposes an optimized concurrent queries execution to reduce node contention and handle the degradation in the query processing performance. Our proposed approach combines efficient query optimization and scheduling techniques. Furthermore, it considers the virtual instances load control based on database replication to improve the query processing performance and in addition, it involves a feedback loop comprising of observing, planning and responding to any overloaded node during the query execution.

The evaluation of the proposed query processing approach is conducted over a real world cloud using the Amazon EC2 infrastructure provisioning service. The results prove a significant benefit with regards to the overall query processing performance.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AQT	Abstract Query Tree
AQP	Adaptive Query Processing
APIs	Application programming interfaces
DaaS	Database as a Service
EC2	Elastic Compute Cloud
FIFO	First-In-First-Out
IaaS	Infrastructure as a Service
IT	Information Technology
PaaS	Platform-as-a-Service
QoS	Quality of Service
RDBMS	Relational Database Management System
SaaS	Software-as-a-Service
SQL	Structured Query Language
TPC	Transaction Processing Council

Chapter 1

Introduction

Chapter 1 Introduction

1.1 Overview

Cloud computing is becoming an emerging powerful model for hosting computing services. These services are delivered to the clients over the internet with their different expectations on the quality of the service [1]. Cloud computing unifies computing components to provide software, platforms and infrastructure as a service. At the lowest level, Infrastructure as a Service (IaaS) offers resources such as processing power or storage as a service and one level above, Platform-as-a-Service (PaaS) provides development tools to build applications based on the service provider's API. Finally, on the top-most level, Software-as-a-Service (SaaS) describes the model of deploying applications to clients on demand [2].

These services are offered and maintained by various cloud computing vendors over the Internet. It prevents the clients from the operational costs such as purchasing, maintaining hardware and set up costs. Service providers offer on-demand resources provisioning to the clients in a pay-only-for-what-you-use pricing model [3]. Some of the cloud service providers are Amazon [4], Microsoft [5], Google Apps [6] and Sales-force [7]. Cloud systems offer Service Level Agreements (SLA) that represent a signed contract between the customer and the service provider about the Quality of Service (QoS). SLA considers service pricing and penalties in case of agreement violations. Flexible and reliable management of SLA agreements is of paramount importance for Cloud users [8].

One of the prominent services offered in cloud computing is the cloud data storage where clients store their own data on the service provider's instances instead of their servers. Clients can deploy their database on the clouds, which are virtual machines with pre-installed and pre-configured database systems. A major component of many cloud services is query processing on data stored in the underlying cloud cluster. Therefore the cloud storage has resulted in an increasing demand to handle the highly amount of concurrent queries access to the resources [9].

Major commercial cloud providers are supported by collections of physical resources within distributed data centers that are spread over large geographically regions. The unprecedented scale of such environments raises new challenging research issues in organizing and managing the cloud resources. Querying distributed data sources is the problem that businesses will encounter as cloud computing grows in popularity. Such a database also needs to deliver high availability and recover any failure [10]. Cloud's users and providers are interested in performance objectives such as minimizing requests response time and maximizing the utilization of cloud resources. In this thesis, a collection of techniques are proposed in an integrated methodology in order to enhance the performance of the query processing.

1.2 Motivation

As the continuous of data growing, cloud enables the remote clients to store their data to the cloud storage environment. By hosting their data in the cloud, clients can avoid the initial investment of expensive infrastructure setup and daily maintenance cost. Cloud environment consists of heterogeneous resources, and they process tasks and workloads in parallel. When a client submits a query, master nodes dispatches the query into worker nodes to be processed concurrently and merging the results returned from the distributed nodes. Some of these resources may execute faster while some may be slower.

Therefore, resources heterogeneity can be resulted in load imbalance that may occur during execution. Therefore the cloud storage is essentially required to handle the highly amount of concurrent queries access the resources. It can cause degradation in performance where queries are evaluated in environments with quickly changing in computational properties, such as available memory or loads. In addition, it can cause the system resources contention and slow the query response time that affects on the system performance. Therefore, an optimized query processing technique is required which take into consideration continuous monitor, assessment and a faster response according to the progress of execution.

1.3 Objective

This thesis focuses on presenting techniques for performing optimizing query processing to enhance the overall performance of query executions over cloud environments. The proposed approach focuses on minimizing the query response time and maximizing the utilization of cloud nodes. Previous researches on query processing over a cloud focused only on issues of query optimization, resource allocation or query scheduling topics.

Therefore in this paper, we focus on presenting integrated techniques for optimizing, scheduling, allocating queries and load management that are presented and implemented in an efficient arrangement for performing the overall query processing. The approach enhances the overall performance of query executions over cloud environments. The presented approach is evaluated in terms of the query processing performance.

1.4 Contributions

This thesis improves the query processing efficiency by contributing with the following components:

- Query optimization techniques that exploit and optimize the shared data among the submitted queries through optimizing and merging the related queries to improve query processing efficiency over cloud resources. It also considers different delay times of submitted queries in order to apply the queries merging step in case of a positive impact on the query execution performance.
- Scheduling technique that determines the efficient ordering of the queries execution to reduce their response time. It also determines the scheduling decisions with taking into account saving the waiting time of the queries within the execution queue.
- Allocation technique to assign and distribute the queries across the virtual nodes in an efficient manner.
- Workload Management technique that recovers the failure or imbalance that may occur during the execution. This is achieved by