A BIG DATA ANALYTIC FRAMEWORK OVER FEDERATED DATA CENTERS FOR INTELLIGENT TRAVEL RECOMMENDERS

H.P.I. Udayanthi

(188067X)

Degree of Master of Science

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

February 2021

A BIG DATA ANALYTIC FRAMEWORK OVER FEDERATED DATA CENTERS FOR INTELLIGENT TRAVEL RECOMMENDERS

Hewa Pathirange Ishara Udayanthi

(188067X)

Thesis submitted in partial fulfilment of the requirements for the degree Master of Science in Computational Mathematics

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

February 2021

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement made is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books)

Signature: Date:

The above candidate has carried out the research for the Master's thesis (By research) under my supervision.

Signature of the supervisor: Date:

Dedicated to

My beloved Father and Mother

Acknowledgements

Many people have helped their best to successfully complete this research. I acknowledge all of them for their valuable thoughts and constant encouragement gave me to make my project a reality.

I would first like to acknowledge my supervisor, Senior Lecturer Thushari Silva, for accepting me as his research student and giving excellent support and advice. Dr. Thushari Silva is a great mentor whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Exceptional thanks goes to Dr. A. Shehan Perera for his valuable comments and guidance as my examiner of progress review panels.

I wish to extend my sincere thanks for all the academic and non-academic staff of the Department of Computational Mathematics, the University of Moratuwa for their kind-hearted help to fulfil my research work.

I acknowledge the sacrificial dedication of my family members for their encouragement and cooperation by managing all the works while I was busy with my research work.

Abstract

Big Data is a series of enormous and complex data sets that are nearly impossible to store and process using traditional data storing and processing methods. The emergence of heterogeneous data in different domains causes significant challenges in data manipulation and decision making. In recent years, the requirement for analysis of heterogeneous data on distributed data storages has been increased and has gained a lot of researchers' attention. Distributed data storage systems and parallel data processing techniques are typically used for data-intensive computing today. Due to the rapid growth of data, a single-cluster environment is inadequate to process that much data. At the same time, there are heterogeneous data sources on different platforms, which need to inter-connect to derive meaningful analysis. The MapReduce software paradigm has surfaced to fill the gap, and it has been successfully operating on systems. However, only single cluster environments are supported by the current implementation of MapReduce and cannot be applied to federated heterogeneous data centers. Hence, it does not have enough capabilities to process heterogeneous data sources. This research presents a big data analytic framework that supports the integration of heterogeneous data sources on distributed computing models across different data centers. The architecture of this framework is based on a master/slave distributed computing model and Map - Reduce - Merge - GlobalReduce is presented as the programming model. Besides, the performance of the novel approach is measured under different cluster configurations, and experimental evaluations had shown promising results for the proposed framework compared to a single cluster environment.

Keywords: Big Data, Heterogeneous data centers, Hierarchical MapReduce, Cloud computing, Multi-cluster

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

HDFS – Hadoop Distributed File System

YARN – Yet Another Resource Negotiator

GPS – Global Positioning System

RDMS – Relational Database Management System

RAM – Random Access Memory

OS – Operating System

CPU – Central Processing Unit

i.e., That is

NLTK – Natural Language Toolkit

NLP – Natural Language Processing

LAN – Local Area Network

URL – Uniform Resource Locator