

# **CERTIFICATE**

This is to certify that the research project entitled “**ENHANCEMENT OF RDB-MINER ALGORITHM FOR MULTI-CLUSTER HADOOPDB**” submitted by **TERERAI TINASHE MAPOSA** (2K13/SWE/25), to the Department of Computer Science and Engineering, Delhi Technological University, New Delhi for the award of the degree of **Master of Technology in Software Engineering** is a record of research work carried out by him under my supervision.

**Mr Manoj Sethi**

Project Guide  
Associate Professor  
Department of Computer Science and Engineering  
Delhi Technological University  
Shahbad Daultpur, Bawana Road, Delhi-110042

# **ACKNOWLEDGEMENTS**

I would like to extend my heartfelt gratitude to my guide Mr. Manoj Sethi for his unparalleled wisdom and expertise in guiding me towards the production of this research work. I would also like to thank all the staff members in the Department of Computer Science and Engineering for imparting their knowledge to us throughout the duration of the MTech and also to all the DTU staff at large for their assistance whenever we requested for such.

I would also like to appreciate my wife Portai Maposa, who was with me here in India for the greater part of my studies. Thank you for your support love. I would also like express my admiration to my parents and family for their support. **I dedicate this research work to my son Nashe!**

Also, not forgetting all my Indian friends (too many to mention by name) and my fellow countrymen for your companionship. God bless you all. Most of all I would like to honor my Lord Jesus Christ and our Father the Almighty God for His grace and mercy which enabled me to undertake this mammoth task. All glory and honor belongs to you!

**Tererai T. Maposa**

Roll No: 2K13/SWE/25

M.Tech (Software Engineering)

Department of Computer Science and Engineering

Delhi Technological University

## **ABSTRACT**

In recent times, volumes of data repositories have astronomically skyrocketed. Data is now regarded as big data and is now measured in the magnitudes of Yotabytes and Petabytes. This has been catalyzed by the rapid development of the Internet and the central role it has claimed in our daily lives. Commercial companies have taken this opportunity to gather as much data about their clients as possible. Governments have also invested heavily in data gathering and analytics. This obsessive collection of data has resulted in an enormous sea of data of all formats and structures. Data has become the most valuable asset for most organizations as it allows them to glean some business intelligence from it and enables them to have a competitive advantage over their rivals. However, it has become more complicated to extract or discover any meaningful patterns, associations or business intelligence from this large pool. Many technologies, platforms and techniques have been developed in order to aid in the data mining process of this BIG DATA. Unfortunately, most projects have been aimed at unstructured and semi-structured data. A blind eye has been turned on the structured data in big data. This project aims to enhance the RDB-Miner algorithm which is an association rule mining algorithm for relation data model (structured data). The project intends to make it applicable and adoptable to structured data in big data. We aim to parallelize the algorithm and implement it in a multi-cluster HadoopDB environment.

# **TABLE OF CONTENTS**

Certificate.....	i
Acknowledgements.....	ii
Abstract.....	iii
Table of Contents.....	iv
Table of Figures.....	vii
<b>CHAPTER 1 – INTRODUCTION.....</b>	<b>1</b>
1.1 Background.....	4
1.2 Big Data .....	4
1.2.1 Characteristics of Big Data .....	6
1.3 Data mining with big data.....	9
1.3.1 HACE Theorem [12].....	9
1.3.2 Challenges in mining big data.....	10
1.4 Structure of thesis.....	14
<b>CHAPTER 2 – LITERATURE REVIEW.....</b>	<b>16</b>
2.1 Cloud Computing.....	16
Advantages of Cloud Computing.....	18
Disadvantages of Cloud Computing.....	18
2.1.1 Apache Hadoop.....	19
2.1.2 Apache Hive.....	20
2.1.3 NoSQL Databases.....	21
2.1.4 Apache Pig.....	22
2.1.5 HBase.....	23
2.1.6 Hybrids of Hadoop.....	23
2.2 An Overview of HadoopDB.....	25
2.2.1 HadoopDB Architecture.....	26
2.3 Data Mining Techniques for Big Data.....	28
2.3.1 Distributed Data Mining (DDM).....	28
2.3.2 Parallel Data Mining.....	32
2.3.3 Algorithms for parallel and distributed data mining.....	34
2.3.4 Challenges in developing parallel algorithms for distributed environment.....	37

<b>CHAPTER 3 – PROBLEM DEFINITION</b> .....	39
3.1 Motivation.....	39
3.1.1 Misconception of big data.....	39
3.1.2 Market Basket data format.....	40
3.1.3 Results from previous work.....	41
3.2 Problem Statement.....	42
<b>CHAPTER 4 – PROPOSED METHODOLOGY</b> .....	43
4.1 Proposed Implementation Environment.....	43
4.2 Proposed Parallel RDB-Miner Algorithm.....	45
4.2.1 The original RDB-Miner Algorithm.....	45
4.2.2 Improvement strategy.....	47
4.2.3 Pseudo Code for the proposed PRDB-Miner Algorithm.....	48
4.2.4 Assumptions made for the proposed algorithm.....	51
4.3 Test Data.....	51
4.4 Performance Factors to be considered.....	52
4.5 Algorithm Implementation.....	52
<b>CHAPTER 5 – RESULTS AND ANALYSIS</b> .....	57
5.1 Results for varying minSup.....	57
5.1.1 Observation.....	57
5.1.2 Explanation.....	58
5.1.3 Comparison to RDB-Miner.....	58
5.2 Results for varying data volumes.....	59
5.2.1 Observation.....	59
5.2.2 Explanation.....	59
5.2.3 Comparison to RDB-Miner.....	60
5.3 Results for varying number of cluster nodes.....	60
5.3.1 Observation.....	61
5.3.2 Explanation.....	61
5.1.3 Optimum number of nodes.....	61

<b>CHAPTER 6- CONCLUSION AND FUTURE WORK</b> .....	63
<b>CHAPTER 7- PUBLICATIONS FROM THIS THESIS</b> .....	64
7.1 Research paper accepted by journal.....	64
<b>REFERENCES</b> .....	65

## **TABLE OF FIGURES**

Figure 1: Dimensions of Big Data [13].....	5
Figure 2: Characteristics of Big Data [13].....	5
Figure 3: Rate at which data is accumulating [11].....	7
Figure 5: Classification of cloud computing services.....	16
Figure 6: HDFS Architecture [6].....	19
Figure 7: Idea behind HadoopDB [3].....	25
Figure 8: The architecture of HadoopDB [3].....	25
Figure 9: Distributed data mining framework [24].....	28
Figure 10: BODHI: Agent-based distributed data mining system [21].....	31
Figure 11: Methods of Parallelism [21].....	32
Figure 12: Mining frequent itemsets using parallel Eclat algorithm [12].....	36
Figure 13: Heat Map for types of data generated by various sectors [12].....	38
Figure 14: Results for volume of data.....	40
Figure 15: results for various minSup.....	41
Figure 16: Proposed multi-cluster HadoopDB configuration.....	42
Figure 17: Proposed cluster network topology.....	43
Figure 18: RDB-Miner Flow Chart.....	46
Figure 19: Proposed PRDB-Miner Flow Chart.....	49
Figure 20: User Interface.....	52
Figure 21: Rules and execution time.....	53
Figure 22: Results for varying minSup for PRDB-Miner.....	56
Figure 23: Results for varying minSup for original RDB-Miner.....	57
Figure 24: Results for varying volumes of data for PRDB-Miner.....	58
Figure 25: Results for varying data volumes for original RDB-Miner.....	59
Figure 26: Results for varying number of cluster nodes for PRDB-Miner.....	.60
Figure 27: Optimum number of nodes.....	61