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

TABLE OF FIGURES