UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

# A MAPREDUCE TOOL FOR DATA MINING AND DATA OPTIMIZATION: CASE OF TEACHERS' WEB PORTAL

BY

SILAS KIBET MAIYO: P58/61616/2010

SUPERVISOR: MR. CHRISTOPHER MOTURI

August 2012

Project report submitted in partial fulfillment of the requirements for the Master of Science in Computer Science of the University of Nairobi

# DECLARATION

I declare that this research project as my original work and has not been submitted to the University of Nairobi and any other university to the best of my knowledge for the same purpose in the same scope and area of research case study.

NAME: SILAS KIBET MAIYO

P58/61616/2010

SIGNATURE........................................ DATE.....................................26 Sep 2012.

This project report is submitted for presentation with my approval as the University Project Supervisor

Name: Mr. Christopher Moturi

SIGNATURE........................................ DATE 26 Sep 2012.......................................

# ABSTRACT

The problem of limited resources in computing and related data management and processing operations is a paramount challenge that is affecting the functions of Ministry of education and in general the Government of Kenya core functions and will remain for a little longer as a catastrophic phenomenon if not considered as a priority concern now.

MapReduce programming technique in cluster computing was studied and whose primary advantage was that, it allows automatic parallelization of applications written in a functional programming style. This allows researcher with no specific knowledge of parallel programming to attain parallelism in distributed cluster environment. Various optimization techniques are considered during the design stage of the system to make it highly efficient on shared memory system architectures, that is, cluster computing environment.

As a result, a MapReduce tool was formulated, designed and developed which consisted of two core sections, the API for MapReduce programming environment and MapReduce runtime system entirely implemented on Hadoop Distributed File system (HDFS). The implementation of the MapReduce runtime system was specifically tailored for shared memory multi-core systems: Case for teachers' web portal as a source of data.

The MapReduce tool provided a considerable performance on any multi-core architecture where large dataset on distributed cluster computing is operational. The research revealed that in making keen evaluations, the MapReduce tool improved the high performance in data optimization as the numbers of computing nodes and data size increases hence scalability of cluster computing is considerably improved.

Researcher's evaluation concluded that the tool can improve on the general security of the data in the system since all data is replicated on nodes hence making them available at all times across the system.

It is therefore envisioned that the study will be of a considerable benefit to the strategist and policy makers in formulations of policies for effective implementation where technology resources are utilized by integrating various new and existing technologies in cluster computing for resource optimizations.