HADOOP CONFIGURATION TUNING FOR PERFORMANCE OPTIMIZATION

By

Christian
11302024

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat no. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

August 2017

Revision after the Thesis Defense on 19 July 2017

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Christian

_____ _____

Student Date

Approved by:

Dipl.-Inf. Kho I Eng

_____ _____

Thesis Advisor Date

Ir. Heru Purnomo Ipung, M.Eng

_____ _____

Thesis Co-Advisor Date

Dr. Ir. Gembong Baskoro, M.Sc

_____ _____

Dean Date

DEDICATION

I dedicate this work to myself and Swiss German University.

ACKNOWLEDGEMENTS

First of all, I would like to thank to my advisors Kho I Eng and Heru Purnomo Ipung for giving me the opportunity to work on this research and advising me each step of the way. Furthermore, I would like to thank to all my lecturers for any recommendations and guidance regarding my research. Finally, I give special thanks to my parents for any assistance and support to complete this research.

ABSTRACT

HADOOP CONFIGURATION TUNING
FOR PERFORMANCE OPTIMIZATION

By

Christian
Dipl.-Inf. Kho I Eng, Advisor
Ir. Heru Purnomo Ipung, M.Eng, Co-Advisor

SWISS GERMAN UNIVERSITY

Configuration parameter tuning is an essential part of the implementation of Hadoop clusters. Each parameter in a configuration plays a role that impacts the overall performance of the cluster. However, we need to learn the characteristics of said parameter and understand the impact in hardware utilization in order to achieve optimal configuration.

Several configuration changes includes mapper count, reduces count, HDFS block size, and MapReduce compression codec selection. The experiment also includes the rebuilding Hadoop source to produce the 64bit version over the 32bit version release from Apache.

To prove any performance gain, we performed benchmark test every experiment we conducted. The benchmark includes TeraGen, TeraSort, and TeraValidate. We used 1GB, 10GB, 50GB of data size that we generated initially using TeraGen which will be used throughout all benchmarks. TeraSort is the program that runs the benchmark, we measure the time needed to complete the sort of the set of data and the CPU utilization during the benchmark. TeraValidate only validates the output of TeraSort to ensure that the output is correct.

From the experiments that we conducted, we have observed significant performance improvements. However, the results may vary between different cluster configuration.

Keywords: apache hadoop, computer cluster, configuration tuning, terasort benchmark

TABLE OF CONTENTS