

**Developing a Scalable and Accurate Job Recommendation System with
Distributed Cluster System using Machine Learning Algorithm**

By

Timothy Dicky
11502006

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

July 2019
Revision after the Thesis Defense on 17 July 2019

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.

Timothy Dicky

Student

Date

Approved by:

Alva Erwin, ST., M.Sc., MTI

Thesis Advisor

Date

Ir. Heru Purnomo Ipung, M.Eng

Thesis Co-Advisor

Date

Dr. Maulahikmah Galinium, S.Kom, M.Sc

Dean

Date

Timothy Dicky

ABSTRACT

DEVELOPING A SCALABLE AND ACCURATE JOB RECOMMENDATION SYSTEM WITH DISTRIBUTED CLUSTER SYSTEM USING MACHINE LEARNING ALGORITHM

By

Timothy Dicky

Alva Erwin, ST., M.Sc., MTI, Advisor

Ir. Heru Purnomo Ipung, M.Eng, Co-Advisor

SWISS GERMAN UNIVERSITY

The purpose of this research is to develop a job recommender system based on the Hadoop MapReduce framework to achieve scalability of the system when it processes big data. Also, a machine learning algorithm is implemented inside the job recommender to produce an accurate job recommendation. The project begins by collecting sample data to build an accurate job recommender system with a centralized program architecture. Then a job recommender with a distributed system program architecture is implemented using Hadoop MapReduce which then deployed to a Hadoop cluster. After the implementation, both systems are tested using a large number of applicants and job data, with the time required for the program to compute the data is recorded to be analyzed. Based on the experiments, we conclude that the recommender produces the most accurate result when the cosine similarity measure is used inside the algorithm. Also, the centralized job recommender system is able to process the data faster compared to the distributed cluster job recommender system. But as the size of the data grows, the centralized system eventually will lack the capacity to process the data, while the distributed cluster job recommender is able to scale according to the size of the data.

Keywords: Machine Learning, Distributed Cluster System, Hadoop MapReduce



© Copyright 2019
by Timothy Dicky
All rights reserved

SWISS GERMAN UNIVERSITY

DEDICATION

I dedicate this works for the future of the country I loved: Indonesia



ACKNOWLEDGEMENTS

First and foremost, I grateful to God for His blessing and health He gave so that I am able to complete my thesis work.

My advisor, Mr. Alva Erwin, that inspires me to do this thesis work, for all his feedback and guidance he gave and I appreciate all of the expertise he shares.

My co-advisor, Mr. Heru Purnomo Ipung, for all encouragement he provides for me to finish my thesis work.

Not forget to mention my parents for their motivation support for me, encouragement, prayer and financial support they provide for me during this thesis work.

Last I want to appreciate all my friends and colleagues that keep me motivated to complete this thesis.

SWISS GERMAN UNIVERSITY

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES.....	10
LIST OF TABLES.....	12
LIST OF FORMULAS.....	13
CHAPTER 1 - INTRODUCTION.....	14
1.1 Background.....	14
1.2 Research Problems.....	16
1.3 Objectives.....	16
1.4 Significance of Study.....	16
1.5 Research Question.....	17
1.6 Hypothesis.....	17
1.7 Scope.....	17
1.8 Research Limitation.....	17
CHAPTER 2 - LITERATURE REVIEW.....	19
2.1 Theoretical Perspectives.....	19
2.1.1 Recommender System.....	19
2.1.2 Similarity Measure.....	21
2.1.3 Apache Hadoop.....	23

2.1.3.1 Hadoop Distributed File System (HDFS).....	23
2.1.3.2 Hadoop MapReduce.....	25
2.2 Previous Studies.....	26
2.2.1 Analysis of Jobs and Candidates Data to Produce an Accurate Job Recommendation Result.....	26
2.2.2 Movie Recommendation System using Collaborative Filtering with Various Similarity Measures.....	27
2.2.3 Apache Hadoop Hive to Reduce Time Consumptions to Query a Big Data	29
CHAPTER 3 – RESEARCH METHODS.....	32
3.1 Research Overview.....	32
3.2 Materials and Equipment.....	36
3.2.1 Data Preparation.....	36
3.2.2 MySQL Database.....	36
3.2.3 Apache Hadoop.....	36
3.2.4 Cloudera Hadoop QuickStarts VM.....	36
3.2.5 Google Cloud Platform.....	37
3.2.6 Git Version Control System.....	37
3.3 Analytical Method.....	38
3.3.1 Data Gathering.....	38
3.3.2 Machine Learning Algorithm to Develop an Accurate Recommender System.....	38
3.3.3 Performance Test Comparison Between Centralized and Distributed Cluster Recommender System with a Big Data.....	41
CHAPTER 4 – RESULTS AND DISCUSSIONS.....	43
4.1 Data Gathering.....	43
4.2 Interview Result.....	44
4.2 Recommendation System Design.....	46
4.4 Database System Design.....	47
4.5 Interview Data Conversion.....	50
4.6 Centralized Job Recommender System Design.....	53

4.7 Distributed Cluster Job Recommender System Design.....	56
4.7.1 Cloudera Hadoop System Environment Setup.....	56
4.7.2 Google Cloud Platform Hadoop System Environment Setup.....	58
4.7.3 Job Recommender Program Architecture.....	59
4.8 Accuracy Analysis on Similarity Measure Algorithm.....	63
4.8.1 Analysis of the Cosine Similarity-based Job Recommendation Output....	65
4.8.2 Analysis of the Jaccard Similarity-based Job Recommendation Output..	67
4.8.3 Analysis of the Euclidean Distance-based Job Recommendation Output.	68
4.9 Using Cosine Similarity with Big Amount of Data.....	69
4.9.1 Centralized Recommender System.....	71
4.9.2 Distributed Cluster Recommender System.....	73
4.9.3 Analysis between Centralized Recommender System and Distributed Recommender System.....	75
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS.....	79
5.1 Conclusions.....	79
5.2 Future Works.....	79
GLOSSARY.....	80
REFERENCES.....	81
CURRICULUM VITAE.....	83



SWISS GERMAN UNIVERSITY