

## TABLE OF CONTENT

FOREWORD.....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES .....	vi
LIST OF TABLES .....	viii
LIST OF LISTINGS .....	x
ABSTRACT .....	xi
CHAPTER I INTRODUCTION .....	1
1.1 Research Background .....	1
1.2 Research Problem .....	3
1.3 Research Scope .....	3
1.4 Research Objectives .....	3
1.5 Research Benefits .....	4
CHAPTER II LITERATURE REVIEW .....	5
CHAPTER III THEORETICAL BASIS .....	21
3.1 Big Data.....	21
3.2 Cluster computer.....	22
3.2.1 Cluster computer Definition.....	23
3.2.2 Cluster Computer Management.....	26
3.3 Parallel programming.....	27
3.4 Cluster Management Tool.....	28
3.4.1 Apache Spark.....	28
3.4.2 Spark Scheduling Algorithm.....	32
3.5 Benchmark Tool.....	34
3.5.1 Spark-Bench.....	34
3.5.2 Spark-Bench Workload.....	35
CHAPTER IV RESEARCH METHODOLOGY.....	36
4.1 Research Description.....	36

4.1.1 Computer Cluster Design.....	37
4.2 Scenario Evaluation.....	36
4.2.1 Scenario Evaluation 1.....	38
4.2.2 Scenario Evaluation 2.....	40
4.2.3 Scenario Evaluation 3.....	44
4.3 Tools and Software.....	46
4.3.1 Tools.....	46
4.3.2 Software.....	46
4.4 Evaluation Metrics.....	47
4.4.1 Total Runtime .....	47
4.4.2 Execution Time .....	47
4.5 Pearson Correlation.....	48
CHAPTER V IMPLEMENTATION.....	49
5.1 Cluster computer deployment.....	49
5.1.1 Cluster computer installation.....	49
5.1.2 Installation of Apache Spark.....	51
5.1.3 Installation of Spark-Bench.....	53
5.1.4 Spark-Bench workload deployment.....	54
5.2 Scenario Evaluations.....	55
5.2.1 Evaluation of scenario 1.....	55
5.2.2 Evaluation of scenario 2.....	57
5.2.3 Evaluation of scenario 3.....	59
5.2.4 Data Processing.....	61
CHAPTER VI RESULTS AND DISCUSSION.....	64
6.1 Results and Discussions Scenario 1.....	64
6.2 Results and Discussions Scenario 2.....	68
6.3 Results and Discussions of Fair and FIFO Scenario 3.....	75
CHAPTER VII CONCLUSIONS.....	80
REFERENCES .....	81

## LIST OF FIGURES

Figure 2.1 Result of <i>Apache Spark</i> and <i>Hadoop MapReduce</i> , evaluate by using <i>Logistic Regression</i> , <i>K-Means</i> , and Java (Saouabi and Ezzati, 2017).....	9
Figure 2.2 Result of <i>Apache Spark</i> and <i>Hadoop MapReduce</i> , evaluate by using <i>Logistic Regression</i> , <i>K-Means</i> , Java and <i>Scala</i> (Saouabi and Ezzati, 2017).....	9
Figure 2.3 results of <i>FIFOBRO</i> , <i>SPSABOR</i> , and <i>OSPSA</i> with scan (Jianchao et al., 2016).....	10
Figure 2.4 results of <i>FIFOBRO</i> , <i>SPSABOR</i> , and <i>OSPSA</i> with Aggregation evaluation (Jianchao et al., 2016).....	10
Figure 2.5 results of <i>FIFOBRO</i> , <i>SPSABOR</i> , and <i>OSPSA</i> with Join evaluation (Jianchao et al., 2016).....	11
Figure 2.6 Results of Response Time Statistics in Boxplot Methods (Chen and Wang, 2015).....	12
Figure 2.7 <i>Spark</i> Analysis Server Throughput Results (Chen and Wang,2015).....	12
Figure 2.8 Results of Failure Rate (Jianchao et al.,2016).....	13
Figure 2.9 Results of 100,000 Query Requests at the Concurrent Level of 10 (Jianchao et al.,2016).....	13
Figure 2.10 Pseudocode of job scheduler (Hadjar and Jedidi, 2019).....	14
Figure 2.11 Results of the evaluation of Task scheduler performance (Hadjar and Jedidi, 2019).....	15
Figure 2.12 Finding speedup (Cheng et al., 2019).....	17
Figure 2.13 The completion time and CPU utilization (Cheng et al., 2019).....	18
Figure 2.14 Results of Reduce job completion time, reservation time and increase CPU utilization (Cheng et al., 2019).....	19
Figure 3.1 architecture of cluster computer (Buyya, 1999) .....	24
Figure 3.2 brief of High Performance Computing (Vugt, 2014) .....	25
Figure 3.3 general view of load balancing clusters (Vugt, 2014) .....	25
Figure 3.5 Overview of master-slave node (Liu, Ding, and Xu, 2010).....	27

Figure 3.6 Spark architecture (Spark .Apache.org, 2019).....	29
Figure 3.7 Architecture of FIFO (White, 2015).....	33
Figure 3.8 Architecture of Fair Scheduler (White, 2015).....	33
Figure 4.1 Computer Cluster Design.....	37
Figure 4.2 Flowchart of First Scenario.....	38
Figure 4.3 Flowchart of First Scenario.....	39
Figure 4.4 Flowchart of FIFO Second Scenario.....	41
Figure 4.5 Flowchart of Fair Second Scenario.....	42
Figure 4.6 Flowchart of Third Scenario.....	44
Figure 6.1 Comparison of K-Means First Scenario.....	64
Figure 6.2 Comparison of Linear Regression First Scenario.....	65
Figure 6.3 Comparison of Sleep First Scenario.....	65
Figure 6.4 Comparison of SparkPI First Scenario.....	66
Figure 6.5 Comparison of SQL First Scenario.....	66
Figure 6.6 Comparison Averages of First Scenario.....	67
Figure 6.7 Comparison averages of spark.memory.fraction.....	70
Figure 6.8 Comparison averages of spark.files.openCostInBytes.....	70
Figure 6.9 Comparison averages of spark.memory.storageFraction.....	71
Figure 6.10 Comparison averages of spark.files.maxPartitionBytes.....	71
Figure 6.11 Comparison averages of spark.reducer.maxSizeInFlight.....	72
Figure 6.12 Comparison averages of spark.files.useFetchCache.....	72
Figure 6.13 Comparison averages of spark.storage.memoryMapThreshold.....	73
Figure 6.14 Comparison averages of spark.shuffle.file.buffer.....	73
Figure 6.15 Comparison of K-Means with A Combination of Parameter Configurations.....	77
Figure 6.16 Comparison of Linear Regression with A Combination of Parameter Configurations.....	78
Figure 6.17 Comparison of Sleep with A Combination of Parameter Configurations.....	78

Figure 6.18 Comparison of SparkPI with A Combination of Parameter

Configurations.....79

Figure 6.19 Comparison of SQL with A Combination of Parameter

Configurations.....79

## LIST OF TABLES

Table 2.1 Result of Apache Spark and Hadoop MapReduce, evaluate by using Logistic Regression, K-Means, and Java (Saouabi and Ezzati,2017).....	6
Table 2.2 Result of Apache Spark and Hadoop MapReduce, evaluate by using Logistic Regression, K-Means, and Java (Saouabi and Ezzati, 2017) .....	7
Table 2.3 Results for K-Means using Spark (Gopalani and Arora, 2015).....	8
Table 2.4 Results for K-Means using Map Reduce (Gopalani and Arora, 2015).....	8
Table 2.5 Specifications of homogeneous cluster (Hadjar and Jedidi, 2019).....	15
Table 2.6 Specifications of homogeneous cluster (Samadi, Zbakh and Tadonki, 2016).....	16
Table 2.7 Summary of literature reviews.....	21
Table 3.1 Apache Spark Parameter Configurations .....	31
Table 4.1 Unit of Parameters.....	37
Table 4.2 Observation of the first scenario.....	39
Table 4.3 Observation of Second Scenario.....	43
Table 4.4 Example of FIFO scheduling Table with Combinations of Parameter Configurations.....	44
Table 4.5 Example of Fair scheduling Table with Combinations of Parameter Configurations.....	45
Table 4.6 Interpretation of the correlation coefficient (Schmidt And Osesbold, 2017).....	47
Table 5.1 General Configuration of Apache Spark.....	51
Table 5.2 Configuration of Fair Scheduler.....	51
Table 5.3 FIFO Property with Configurations and Value.....	52
Table 5.4 Fair Property with Configurations and Value.....	52
Table 6.1 Observation of Fair First Scenario.....	66
Table 6.2 Observation of FIFO First Scenario.....	67
Table 6.3 Results of Parameter Configurations Fair Scheduling Second Scenario.....	69
Table 6.4 Results of Parameter Configurations FIFO Scheduling Second Scenario.....	69



Table 6.5 Fair scheduling Table with Combinations of Parameter Configurations.....	76
Table 6.6 Results of Parameter Configurations Between FIFO Scheduling Second Scenario.....	77

## LIST OF LISTINGS

Listing 5.1 Passwordless SSH.....	49
Listing 5.2 Configuration of spark-env.sh.....	49
Listing 5.3 Configuration of Fair Scheduling.....	50
Listing 5.4 Configuration of spark-bench-env.sh.....	52
Listing 5.5 SparkPI configuration.....	53
Listing 5.6 K-Means configuration.....	53
Listing 5.7 Sleep configuration.....	53
Listing 5.8 Linear Regression configuration.....	54
Listing 5.9 SQL configuration.....	54
Listing 5.10 Automation Code of First Scenario.....	56
Listing 5.11 Running the Spark-Bench.....	57
Listing 5.12 Parameter Configurations.....	57
Listing 5.14 Automation Code of Second Scenario.....	58
Listing 5.14 Automation Code of Scenario 3.....	60
Listing 5.15 Fair Parameter Configurations.....	61
Listing 5.16 FIFO Parameter Configurations.....	61
Listing 5.17 Benchmark Data Processing.....	62