

TABLE OF CONTENT

APPROVAL PAGE	II
STATEMENT	III
PREFACE.....	VI
TABLE OF CONTENT	VIII
LIST OF TABLE.....	XI
TABLE OF FIGURES	XII
ABSTRACT.....	XIV
CHAPTER I INTRODUCTION	1
1.1 BACKGROUND PROBLEM	1
1.2 RESEARCH PROBLEM	2
1.3 RESEARCH SCOPE	2
1.4 RESEARCH OBJECTIVE	3
1.5 RESEARCH METHODOLOGY	3
1.6 THESIS ORGANIZATION	3
CHAPTER II LITERATURE REVIEW	5
CHAPTER III THEORETICAL BASES	10
3.1 DATA	10
3.2 METADATA	11
3.2.1 <i>Metadata Models</i>	11
3.2.2 <i>Metadata Standards</i>	12
3.3 APACHE TIKA.....	13
3.3.1 <i>Tika Architecture</i>	13
3.3.2 <i>Metadata in Apache Tika</i>	14
3.3.3 <i>Full Text Extraction</i>	15
3.4 ELASTICSEARCH	16
3.4.1 <i>Data Architecture</i>	17
3.4.2 <i>System Architecture</i>	18
3.4.3 <i>Function</i>	20
3.5 MAP REDUCE	21
3.5.1 <i>Programming Model</i>	21



3.6	HADOOP	23
3.6.1	<i>Hadoop MapReduce Class</i>	24
3.6.2	<i>HDFS Block Size</i>	25
3.6.3	<i>WordCount</i>	25
3.6.4	<i>WordMean</i>	26
3.6.5	<i>WordMedian</i>	26
3.6.6	<i>Word Standard Deviation</i>	26
CHAPTER IV ANALYSIS AND SYSTEM DESIGN		27
4.1	RESEARCH STEPS	27
4.2	ANALYSIS	29
4.2.1	<i>Data Analysis</i>	29
4.2.2	<i>Functional Requirement Analysis</i>	32
4.3	RESEARCH DESIGN.....	33
4.3.1	<i>Workflow Design</i>	33
4.3.2	<i>Extraction Process Design</i>	35
4.3.3	<i>Indexing Process Design</i>	37
4.3.4	<i>Testing Process Design</i>	38
4.3.5	<i>Testing Design</i>	38
CHAPTER V IMPLEMENTATION		55
5.1	RESEARCH ENVIRONMENT.....	55
5.1.1	<i>Hardware Specification</i>	55
5.1.2	<i>Software Specification</i>	55
5.2	IMPLEMENTATION.....	55
5.2.1	<i>Extraction Process</i>	55
5.2.2	<i>Indexing Process</i>	59
5.2.3	<i>Hadoop</i>	70
5.2.4	<i>Testing Process</i>	76
CHAPTER VI EXPERIMENT AND DISCUSSION.....		79
6.1	QUANTITY TEST RESULT	79
6.1.1	<i>WordCount</i>	81
6.1.2	<i>WordMean</i>	84



6.1.3	<i>WordMedian</i>	87
6.1.4	<i>Word Standard Deviation</i>	90
6.2	MAPPER AND REDUCER CONFIGURATION TEST RESULT	93
6.2.1	<i>WordCount</i>	93
6.2.2	<i>WordMean</i>	99
6.2.3	<i>WordMedian</i>	105
6.2.4	<i>WordStandardDeviation</i>	111
CHAPTER VII CONCLUSION AND FUTURE WORKS.....		118
7.1	CONCLUSION	118
7.2	RECOMMENDATION	119
REFERENCES.....		120



LIST OF TABLE

Table II-1 Literature Review.....	9
Table IV-1 Data Research Description.....	29
Table IV-2 PDF Metadata	30
Table IV-3 Data Quantity Test Design	39
Table IV-4 Mapper and Reducer Configuration Test Design.....	45
Table V-1 Hadoop Configuration Table	71
Table VI-1 Indexing Process Time.....	80
Table VI-2 Time Table Wordcount Quantity Test	81
Table VI-3 Wordcount JSON and csv Bytes Size Read and Written Table	82
Table VI-4 Time Table Wordmean Quantity Test	84
Table VI-5 Wordmean JSON and csv Bytes Size Read and Written Table.....	85
Table VI-6 Time Table Wordmedian Quantity Test	87
Table VI-7 Wordmedian JSON and csv Bytes Size Read and Written Table.....	88
Table VI-8 Time Table Word Standard Deviation Quantity Test	90
Table VI-9 Word Standard Deviation JSON and csv Bytes Size Read and Written Table	91
Table VI-10 WordCount Total Time Matrix.....	93
Table VI-11 WordCount Map Time Matrix	95
Table VI-12 WordCount Reduce Time Matrix.....	97
Table VI-13 WordMean Total Time Matrix	99
Table VI-14 Word Mean Map Time Matrix	101
Table VI-15 Word Mean Reduce Time Matrix	103
Table VI-16 WordMedian Total Time Matrix	105
Table VI-17 WordMedian Map Time Matrix	107
Table VI-18 WordMedian Reduce Time Matrix	109
Table VI-19 Word Standard Deviation Total Time Matrix	111
Table VI-20 Word Standard Deviation Map Time Matrix	113
Table VI-21 Word Standard Deviation Reduce Time Matrix.....	115

TABLE OF FIGURES

Figure II.1 log events collecting and indexed stored (Bai, 2013)	5
Figure II.2 Searching Flow (Bai, 2013)	6
Figure III.1 Classes of Metadata Models (Manning Tika in Action)	12
Figure III.2 Apache Tika Framework Architecture System (Manning Tika in Action)	14
Figure III.3 Metadata Extraction Class Apache Tika (Manning Tika in Action)	15
Figure III.4 Apache Tika PDF Extraction	15
Figure III.5 Elasticsearch and RDMS Data Architecture Comparison	18
Figure III.6 Elasticsearch System Architecture	20
Figure III.7 Map and Reduce Function	22
Figure III.8 Map and Reduce Phase (Rehan and Gandgodkar, 2015)	22
Figure III.9 Map and Reduce Performance (Lin and Dyer, 2010)	23
Figure III.10 Hadoop Modul	24
Figure III.11 WordCount Process	26
Figure IV.1 General Research Flow	28
Figure IV.2 Raw Data Structure	29
Figure IV.3 JSON Data Structure	30
Figure IV.4 Pre-processing Data Illustration	32
Figure IV.5 Detailed Research Workflow	34
Figure IV.6 First Extraction Process	36
Figure IV.7 Indexing process flowchart	37
Figure IV.8 Testing Process Flowchart	38
Figure V.1 PDF file before extraction	56
Figure V.2 First extraction result	57
Figure V.3 Data cleaning code	58
Figure V.4 JSON after cleaning	59
Figure V.5 Github page of Elasticsearch-head plugin	60
Figure V.6 Elasticsearch status on terminal	61
Figure V.7 Elasticsearch info on localhost:9200	62
Figure V.8 Elasticsearch-head plugin	63
Figure V.9 Elasticsearch Info	64
Figure V.10 Indices Stats	65
Figure V.11 Cluster Health	66
Figure V.12 Index Creation	67
Figure V.13 Elasticsearch Bulk Indexing Process Script	67
Figure V.14 Indexing Process	68
Figure V.15 Indexing Process result	68
Figure V.16 Index Structured Query	69
Figure V.17 Es2csv Command	69
Figure V.18 The Exportation Process	69
Figure V.19 Indexing and Exporting Script	69
Figure V.20 Indexing and Exporting Process	70