

# Table of Contents

1. Abstract.....	I
2. Introduction.....	1
3. Motivation.....	2
4. RelatedWork.....	3
5. Background.....	4
5.1 Apache Hadoop.....	4
5.2 Hbase.....	6
5.3 Pig.....	6
5.4 Hadoop Yarn.....	6
5.5 MapReduce.....	9
5.6 HDFS Architecture.....	15
5.7 Hadoop Cluster Architecture.....	16
5.8 Experimental Setup.....	16
6. MapReduce Programming Model.....	17
7. K-Means Based Distributed Document Clustering.....	18
8. Experiments and Results.....	19
9. Conclusion and Future Work.....	28
10. References.....	29