A
Dissertation
On

Data Clustering using Black Hole Algorithm using MapReduce on Hadoop Framework

Submitted in Partial Fulfillment of the Requirement
For the Award of the Degree of

Master of Technology

in

Computer Science and Engineering

by

Prinsi Sharma
University Roll No.:- 2K14/CSE/12

Under the Esteemed Guidance of

Dr. Kapil Sharma
Associate Professor, Computer Science and Engineering Department, DTU

2014-2016

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

DELHI TECHNOLOGICAL UNIVERSITY

DELHI – 110042, INDIA

ABSTRACT

The major drawback of conventional data clustering algorithms is that they are inefficient for analyzing large-scale datasets as most of them are tailored for a centralized system, that means if the size of input dataset exceeds the size of storage or memory of such a system, it would make the job of clustering much more difficult. To solve this problem, an efficient clustering algorithm, called Black Hole using MapReduce on Hadoop framework is proposed to ascend the strength of the black hole algorithm and the MapReduce programming model of Hadoop to accelerate the clustering speed by virtue of both software and hardware.

By using MapReduce, the algorithm will then divide a large dataset into a number of small data sets and cluster these smaller data sets in parallel. Moreover, it inherits the characteristics of the black hole algorithm, meaning that no parameters are to be set manually; thus, the implementation is easy. To evaluate the performance of the proposed algorithm, several datasets are used with different numbers of nodes. Experimental results show that the proposed algorithm can provide a significant speedup as the number of nodes increases.

Index Terms—Black hole algorithm, clustering, Hadoop, and MapReduce.

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Dr. Kapil Sharma for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided to me without which the project could not have been a success.

I am also grateful to Dr. O.P Verma, HOD, Computer Science and Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents, friends and seniors for constantly encouraging me during the completion of work.

Prinsi Sharma
University Roll no: 2K14/CSE/12
M.Tech (Computer Science and Engineering)
Department of Computer Engineering
Delhi Technological University
Delhi – 110042

CERTIFICATE

This is to certify that the dissertation entitled "**Data Clustering using Black-Hole algorithm on MapReduce Framework**" is a bonafide record of work done by **Prinsi Sharma, Roll No.-2K14/CSE/12** at **Delhi Technological University**for the partial fulfillment of the requirement for the degree of **Master of Technology** in **Computer Science and Engineering**. This project was carried out under my supervision and has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma to the best of my knowledge and belief.

Date: _____

(Dr. Kapil Sharma)
Associate Professor and Project Guide
Department of Computer Science and Engineering
Delhi Technological University

TABLE OF CONTENTS

Abstract **ii**

Acknowledgement **iii**

Certificate **iv**

List of Figures & Tables **vii**

Chapter 1

Introduction **1**

1.1. Motivation of the work 2

1.2. Aim of the thesis 2

1.3. Thesis Organization 3

Chapter 2

Literature Survey **4**

2.1. Meta-Heuristic Algorithms 4

2.2. Types of meta-heuristic algorithms 4

Chapter 3

Research Methodology **8**

3.1. Apache HADOOP 8

3.1.1 Key Features of Hadoop 8

3.1.2 Hadoop Cluster 9

3.1.3 Why Hadoop 10

3.1.4 HDFS 11

3.1.5 Files and Blocks 14

3.1.6 HDFS File Read and Write 15

3.1.7 Difference between GFS & HDFS 17

3.1.8 Disadvantages of HDFS 18

3.1.9 MapReduce 19

3.1.10 Inefficiencies of Hadoop 21
3.2. Data Clustering 22
3.2.1 Types of clustering methods 22
3.2.2 Black-hole phenomenon and algorithm 24
3.3. Test Datasets 28

Chapter 4
Proposed Work **29**
4.1. Pseudo code of Black Hole Mapreduce Algorithm 30
4.2 Flowchart of Black Hole Mapreduce Algorithm 31

Chapter 5
Simulation Result & Analysis **33**
5.1. Simulation Setup 33
5.2. Performance Evaluation- No: of nodes Vs time plots 33
5.3. Performance Evaluation- No: of iterations Vs Fitness plots 37

Chapter 6
Conclusion and Future Work **39**

Reference **40**

LIST OF FIGURES

Figure 1 Hadoop Cluster [23] 10
Figure 3 Processes of HDFS 14
Figure 4 Replication of Blocks with replication factor 3 [21] 14
Figure 5 HDFS Write [21] 16
Figure 6 HDFS Read [21] 17
Figure 7 MapReduce [23] 19
Figure 8 Map Task Execution [24] 20
Figure 9 Reduce Task Execution [24] 21
Figure 10 Black Hole with its event horizon 25
Figure 11 Flowchart of Black Hole algorithm 27
Figure 12 Flowchart for the α-black hole algorithm 32
Figure 13 Nodes Vs time plot for iris dataset 34
Figure 14 Nodes Vs time plot for glass dataset 35
Figure 15 Nodes Vs time plot for glass dataset 35
Figure 16 Nodes Vs time plot for magic dataset 36
Figure 17 Nodes Vs time plot for poker hand dataset 36
Figure 18 Iteration Vs Fitness plot for iris dataset 37
Figure 19 Iteration Vs Fitness plot for glass dataset 37
Figure 20 Iteration Vs Fitness plot for wine dataset 38
Figure 21 Iterations Vs Fitness plot for magic dataset 38
Table -4.1 Characteristics of test datasets. 28
Table -5.1 Configuration of systems used 33
Table 5.2 Sum of intra-cluster distances for various datasets 34