DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

DECLARATION

I hereby declare that the project entitled **"K-MEANS CLUSTERING ALGORITHM ON MAP REDUCE ARCHITECTURE"** submitted by me in the partial fulfillment of the requirements for the award of the degree of Master of Technology (Software Engineering) of Delhi Technological University is record of my own work carried under the supervision and guidance of **Dr. Kapil Sharma.**

To the best of my knowledge this project has not been submitted to Delhi Technological University or any other University or Institute for the award of any degree.

ROHIT NEGI
[2K13/SWE/19]

ACKNOWLEDGEMENT

In the sense of great pleasure and satisfaction I present this project entitled **"K-MEANS CLUSTERING ALGORITHM ON MAP REDUCE ARCHITECTURE"**.

The completion of this project is no doubt a product of invaluable support and contribution of number of people.

I would like to express my sincere thanks to my guide **Dr. Kapil Sharma** (Associate Professor, Department of Computer Science and Engineering) for his continuous help and valuable suggestions and also providing encouraging environment, without which my project and its documentation would not have been possible.

The completion of any task is not only the reward to the person activity involved in accomplishing it, but also the person involved in inspiring and guiding. I am grateful to my friends and family for their constant motivation and comments that has helped me to complete this report.

ROHIT NEGI

[2K13/SWE/19]

ABSTRACT

The rapid development of the Internet and its impact on every aspect of life has resulted in the size of the data to increase from GB level to TB even PB level. This has brought about new technologies such as Hadoop for efficient storage and analysis the data. Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Cluster is a collection of data members having similar characteristics. The process of establishing a relation or deriving information from raw data by performing some operations on the data set like clustering is known as data mining. Data collected in practical scenarios is more often than not completely random and unstructured. Hence, there is always a need for analysis of unstructured data sets to derive meaningful information. This is where unsupervised algorithms come in to picture to process unstructured or even semi structured data sets by resultant.

K-Means Clustering is one such technique used to provide a structure to unstructured data so that valuable information can be extracted. This paper discusses the implementation of the K-Means Clustering Algorithm over a distributed environment using ApacheTM Hadoop. The key to the implementation of the K-Means Algorithm is the design of the Mapper and Reducer routines which has been discussed in the later part of the paper. The steps involved in the execution of the K-Means Algorithm has also been described in this paper based on a small scale implementation of the K-Means Clustering Algorithm on an experimental setup to serve as a guide for practical implementations.

TABLE OF CONTENTS

APPENDIX A

LIST OF FIGURES

APPENDIX B

LIST OF TABLES

REFERENCES

1. Armour, F., Kaisler, S., & Espinosa, A. (2015). Introduction to Big Data Analytics: Concepts, Methods, Techniques and Applications. *48th Hawaii International Conference on System Sciences* (p. 886). IEEE.

2. Bing, L., & Chan, K. C. (2014). A Paralleled Big Data Algorithm with MapReduce Framework for Mining Twitter Data. *IEEE Computer Society* (pp. 121-128). IEEE.

3. Dashti, H. T., Simas, T., Ribeiro, R. A., Assadi, A., & Moitinho, A. (2010). MK-means - Modified K-means clustering algorithm. IEEE.

4. Demchenko, Y., Laat, C. d., & Membrey, P. (2014). Defining Architecture Components of the Big Data Ecosystem. *IEEE* (pp. 104-112). IEEE.

5. Elagib, S. B., Najeeb, A. R., Hashim, A. H., & Olanrewaju, R. F. (2014). Big Data Analysis Solutions Using Map Reduce Framework. *5th International Conference on Computer & Communication Engineering* (pp. 127-130). IEEE.

6. Frank, A., Kaisler, S., & Espinosa, A. (2015). Introduction to Big Data Analytics: Concepts, Methods, Techniques and Applications. *48th Hawaii International Conference on System Sciences* (p. 886). IEEE.

7. Gohil, P., Garg, D., & Panchal, B. (2014). A Performance Analysis of MapReduce Applications on Big Data in Cloud based Hadoop. *ICICES.* S.A.Engineering College, Chennai, Tamil Nadu, India: IEEE.

8. Guoli, L., LangFang, H., Tingting, W., Yanping, L., Limei, Y., & Jinqiao, G. (2013). The Improved Research on K-Means Clustering Algorithm in Initial Values. *International Conference on Mechatronic Sciences, Electric Engineering and Computer* (pp. 2124-2127). IEEE.

9. *http://www.coreservlets.com/hadoop-tutorial/*. (n.d.). Retrieved 2015, from coreservlets.com: http://www.coreservlets.com

10. https://hadoop.apache.org/. (n.d.). Retrieved 2015, from apache: http://www.apache.org

11. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An Efficient k-Means Clustering Algorithm:Analysis and Implementation. *IEEE TRANSACTIONS* (pp. 881-892). IEEE.

12. Li, H. G., Wu, G. Q., Hu, X. G., Zhang, J., Li, L., & Wu, X. (2011). K-Means Clustering with Bagging and MapReduce. *Proceedings of the 44th Hawaii International Conference on System Sciences* (pp. 1-8). IEEE.

13. Liao, Q., Yang, F., & Zhao, J. (2013). An Improved parallel K-means Clustering Algorithm with MapReduce. *Proceedings of ICCT* (pp. 764-768). IEEE.

14. Maitrey, S., & Jha, C. K. (2015). Handling Big Data Efficiently by using Map Reduce Technique. *IEEE International Conference on Computational Intelligence & Communication Technology* (pp. 703-708). IEEE.

15. Manikandan, S. G., & Ravi, S. (2014). Big Data Analysis using Apache Hadoop. *IEEE.* IEEE.

16. Marozzo, F., Talia, D., & Trunfio, P. (2011). Framework for Managing MapReduce Applications in Dynamic Distributed Environments. *19th International Euromicro Conference on Parallel, Distributed and Network-Based Processing* (pp. 149-158). IEEE.

17. Na, S., yong, G., & Xumin, L. (2010). Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm. *Third International Symposium on Intelligent Information Technology and Security Informatics* (pp. 63-67). IEEE.

18. Nandimath, J., Patil, A., Banerjee, E., Kakade, P., & Vaidya, S. (2014). Big Data Analysis using Apache Hadoop. *IEEE IRI.* IEEE.

19. Ren, S., & Fan, A. (2011). K-means Clustering Algorithm Based On Coefficient Of Variation. *4th International Congress on Image and Signal Processing* (pp. 2076-2079). IEEE.

20. Verma, J. P., Patel, B., & Patel, A. (2015). Big Data Analysis: Recommendation System with Hadoop Framework. *IEEE International Conference on Computational Intelligence & Communication Technology* (pp. 92-97). IEEE.

21. Wang, J., & Su, X. (2011). An improved K-Means clustering algorithm. (pp. 44-46). IEEE.

22. Wang, S. (2013). Improved K-means Clustering Algorithm Based on the Optimized Initial Centriods. *3rd International Conference on Computer Science and Network Technology* (pp. 450-453). Dalian, China: IEEE.

23. Xu, M., & Franti, P. (2004). A HEURISTIC K-MEANS CLUSTERING ALGORITHM BY KERNEL PCA. *International Conference on Image Processing* (pp. 3503-3506). IEEE.

24. Yang, J., & Li, X. (2013). MapReduce based Method for Big Data Semantic Clustering. *IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2814-2819). IEEE.

25. Yu, H. T., Cheng, X., Jia, M., & Jiang, Q. (2013). Optimized K-Means Clustering Algorithm based on Artificial Fish Swarm. *International Conference on Mechatronic Sciences, Electric Engineering and Computer* (pp. 1783-1787). IEEE.

26. Zhao, W., Ma, H., & He, Q. (2009). Parallel K-Means Clustering Based on MapReduce. *Springer-Verlag Berlin Heidelberg* , 674-679.

27. Wang, J., Yuan, D., & Jiang, M. (2012). Parallel K-PSO Based on MapReduce. *IEEE*, (pp. 1203-1208).