High Performance Clustering to Improve Data Mining Techniques

Heba Mohammed Fadhil a,∗

a Department of Information and Communication, Al-Khwarizmi College of Engineering, University of Baghdad, Baghdad, Iraq

∗ Corresponding Author: fadhheba@gmail.com

Abstract—An increasing amount of data is approaching into existence each day and power to process this data calls for cluster computers which are a set of a computer that operates together, only to be viewed as a single computer, to process and compute data. People are pretty dependent on computers and their computational power more and more every day. As a consequence, those outcomes need to be accurate and be calculated fast. Parallel computing is a more rapid way to process instructions by breaking the large task into smaller tasks using an organized effort to process data simultaneously. Data clustering has been confirmed to be a favorable data mining technique. Recently, there have been various attempts for clustering data mining algorithms. For that reason, Cluster computing is used to create redundancy in a computer network to ensure that it will regularly be to be had and that it’s going to not fail.

Keywords—Data clustering; Data mining; Parallel Programming; High Performance.

RESEARCHERS utilize computers to run simulations of natural phenomena or analyze data about their work. As these programs grow in complexity and accuracy so does the computer resources required to process them. To develop systems that can run these programs effectively, researchers create computing systems which are consisting of multiple processors and machines working together. These systems are called distributed systems. Given a task these processors within the system will each take a part of the problem and process their sections simultaneously [1]. The field, that is involved in designing and improving these distributed computing systems, is called High-Performance Computing (HPC). Even with the computing power of the supercomputers produced by aggregating multiple processors and machines, the ambitions and program complexity will still be increasing throughout the scientific and business communities. A currently growing example of this trend is Big Data. Big Data is the name that is given to incredibly large amounts of stored information on databases and clouds. Researchers utilize big data to discover trends or patterns to help guide their businesses’ decisions [2].

Most of the association rule mining algorithms suffer from the time-consuming elaboration on finding all candidates that fit the subjective conditions. We believe that the most effective way is to develop parallel algorithms to promote performance. However, prior parallel architectures and algorithms suffer from overhead in inter-site communications or requiring a huge number of spaces to maintain the local support counts of a large number of candidate’s sets [3].

Data mining algorithms are essential tools to extract information from the increasing number of large datasets, also called Big Data. However, these algorithms demand huge amounts of computing power to achieve reliable results. Although conventional High-Performance Computing (HPC) platforms can deliver such performance, they are commonly expensive and power-hungry [4].

Beowulf cluster is simply a collection of identical computer hardware-based systems, networked together and running some kind of parallel processing software that allows each node in the cluster to share data and computation [5]. Typically, the parallel programming software is MPI (Message Passing Interface), which utilizes TCP/IP along with some libraries to allow programmers to create parallel programs that can split one task into parts appropriate to run on multiple machines simultaneously [6].

An important distinguishing feature is that only one computer— the head node—is communicating with the
outside network. A Beowulf cluster is dedicated only to jobs assigned through its head node.

High-performance computing (HPC) on the other hand, is the use of supercomputers and parallel processing techniques for solving complex computational problems. HPC technology focuses on developing parallel processing algorithms and systems by incorporating both administration and parallel computational techniques.

Clusters are usually deployed to improve performance and availability over that of a single computer, while typically being much more cost-effective than single computers of comparable speed or availability. Computer clusters have emerged as a result of the convergence of several computing trends which is including the availability of low-cost microprocessors, high-speed networks, and software for high-performance distributed computing [7].

With Best Regards,
Heba M. Fadhil, Editor
ICSES Interdisciplinary Transactions on Cloud Computing, IoT, and Big Data
21-3-2020

REFERENCES