Chapter 4
Parallel Data Reduction Techniques for Big Datasets

Ahmet Artu Yıldırım
Utah State University, USA

Cem Özdoğan
Çankaya University, Turkey

Dan Watson
Utah State University, USA

ABSTRACT

Data reduction is perhaps the most critical component in retrieving information from big data (i.e., petascale-sized data) in many data-mining processes. The central issue of these data reduction techniques is to save time and bandwidth in enabling the user to deal with larger datasets even in minimal resource environments, such as in desktop or small cluster systems. In this chapter, the authors examine the motivations behind why these reduction techniques are important in the analysis of big datasets. Then they present several basic reduction techniques in detail, stressing the advantages and disadvantages of each. The authors also consider signal processing techniques for mining big data by the use of discrete wavelet transformation and server-side data reduction techniques. Lastly, they include a general discussion on parallel algorithms for data reduction, with special emphasis given to parallel wavelet-based multi-resolution data reduction techniques on distributed memory systems using MPI and shared memory architectures on GPUs along with a demonstration of the improvement of performance and scalability for one case study.

DOI: 10.4018/978-1-4666-4699-5.ch004

INTRODUCTION

With the advent of information technologies, we live in the age of data – data that needs to be processed and analyzed efficiently to extract useful information for innovation and decision-making in corporate and scientific research labs. While the term of 'big data' is relative and subjective and varies over time, a good working definition is the following:

- **Big Data:** Data that takes an excessive amount of time/space to store, transmit, and process using available resources.

One remedy in dealing with big data might be to adopt a distributed computing model to utilize its aggregate memory and scalable computational power. Unfortunately, distributed computing approaches such as grid computing and cloud computing are not without their disadvantages (e.g., network latency, communication overhead, and high-energy consumption). An "in-box" solution would alleviate many of these problems, and GPUs (Graphical Processing Units) offer perhaps the most attractive alternative. However, as a cooperative processor, GPUs are often limited in terms of the diversity of operations that can be performed simultaneously and often suffer as a result of their limited global memory as well as memory bus congestion between the motherboard and the graphics card. Parallel applications as an emerging computing paradigm in dealing with big datasets have the potential to substantially increase performance with these hybrid models, because hybrid models exploit both advantages of distributed memory models and shared memory models.

A major benefit of data reduction techniques is to save time and bandwidth by enabling the user to deal with larger datasets within minimal resources available at hand. The key point of this process is to reduce the data without making it statistically indistinguishable from the original data, or at least to preserve the characteristics of the original dataset in the reduced representation at a desired level of accuracy. Because of the huge amounts of data involved, data reduction processes become the critical element of the data mining process on the quest to retrieve meaningful information from those datasets. Reducing big data also remains a challenging task that the straightforward approach working well for small data, but might end up with impractical computation times for big data. Hence, the phase of software and architecture design together is crucial in the process of developing data reduction algorithm for processing big data.

In this chapter, we will examine the motivations behind why these reduction techniques are important in the analysis of big datasets by focusing on a variety of parallel computing models ranging from shared-memory parallelism to message-passing parallelism. We will show the benefit of distributed memory system in terms of memory space to process big data because of the system's aggregate memory. However, although many of today's computing systems have many processing elements, we still lack data reduction applications that benefit from multi-core technology. Special emphasis in this chapter will be given to parallel clustering algorithms on distributed memory systems using the MPI library as well as shared memory systems on graphics processing units (GPUs) using CUDA (Compute Unified Device Architecture developed by NVIDIA).

GENERAL REDUCTION TECHNIQUES

Significant CPU time is often wasted because of the unnecessary processing of redundant and non-representative data in big datasets. Substantial speedup can often be achieved through the elimination of these types of data. Furthermore, once non-representative data is removed from