Chapter 71
Synchronizing Execution of Big Data in Distributed and Parallelized Environments

Gueyoung Jung
Xerox Research Center Webster, USA

Tridib Mukherjee
Xerox Research Center India, India

ABSTRACT

In the modern information era, the amount of data has exploded. Current trends further indicate exponential growth of data in the future. This prevalent humungous amount of data—referred to as big data—has given rise to the problem of finding the "needle in the haystack" (i.e., extracting meaningful information from big data). Many researchers and practitioners are focusing on big data analytics to address the problem. One of the major issues in this regard is the computation requirement of big data analytics. In recent years, the proliferation of many loosely coupled distributed computing infrastructures (e.g., modern public, private, and hybrid clouds, high performance computing clusters, and grids) have enabled high computing capability to be offered for large-scale computation. This has allowed the execution of the big data analytics to gather pace in recent years across organizations and enterprises. However, even with the high computing capability, it is a big challenge to efficiently extract valuable information from vast astronomical data. Hence, we require unforeseen scalability of performance to deal with the execution of big data analytics. A big question in this regard is how to maximally leverage the high computing capabilities from the aforementioned loosely coupled distributed infrastructure to ensure fast and accurate execution of big data analytics. In this regard, this chapter focuses on synchronous parallelization of big data analytics over a distributed system environment to optimize performance.

DOI: 10.4018/978-1-4666-9840-6.ch071

INTRODUCTION

Dealing with the execution of big data analytics is more than just a buzzword or a trend. The data is being rapidly generated from many different sources such as sensors, social media, click-stream, log files, and mobile devices. Recently, collected data can exceed hundreds of terabytes and moreover, they are continuously generated from the sources. Such big data represents data sets that can no longer be easily analyzed with traditional data management methods and infrastructures (Jacobs, 2009; White, 2009; Kusnetzky, 2010). In order to promptly derive insight from big data, enterprises have to deploy big data analytics into an extraordinarily scalable delivery platform and infrastructure. The advent of on-demand use of vast computing infrastructure (e.g., clouds and computing grids) has been enabling enterprises to analyze such big data using with low resource usage cost.

A major challenge in this regard is figuring out how to effectively use the vast computing resources to maximize the performance of big data analytics. Using loosely coupled distributed systems (e.g., clusters in a data center or across data centers; public cloud with the internal clusters as hybrid cloud formation) is often better choice to parallelize the execution of big data analytics compared to using local centralized resources. Big data can be distributed over a set of loosely-coupled computing nodes. In each node, big data analytics can be performed on the portion of the data transferred to the node. This paradigm can be more flexible and has obvious cost benefits (Rozsnyai, 2011; Chen, 2011). It enables enterprises to maximally utilize their own computing resources and effectively utilize external computing resources that are further optimized for the big data processing.

However, contrary to common intuition, there is an inherent tradeoff between the level of parallelism and performance of big data analytics. This tradeoff is primarily caused by the significant delay for big data to get transferred to computing nodes. For example, when a big data analytics is run on a pool of inter-connected computing nodes in hybrid cloud (i.e., the mix of private and public clouds), it is often experienced that an extended period of data transfer delay is comparable or even higher than the time required to data computation itself. Additionally, the heterogeneity of computing nodes on computation time and data transfer delay can make the tradeoff issue being further complicated. The data transfer delay mostly depends on the location and network overhead of each computing node. A fast transfer of data chunks to a relatively slow computing node can cause data overflow, whereas a slow transfer of data chunks to a relatively fast computing node can lead to underflow causing the computing node to be idle (hence, leading to low resource utilization of the computing node).

This chapter focuses on optimally parallelizing big data analytics over such distributed heterogeneous computing nodes. Specifically, this chapter will discuss how to improve the advantage of parallelization by considering the time overlap *across computing nodes* as well as *between data transfer delay and data computation time* in each computing node. It should be noted here that the data transfer delay may be reduced by using data compression techniques (Plattner, 2009; Seibold, 2012). However, even with such reduction, overlapping the data transfer delay with the execution can reap benefits in the overall turnaround of the big data analytics. Ideally, the parallel execution should be designed in such a way that the execution of big data analytics at each computing node, including such data transfer and data computation, completes at near same time with other computing nodes.

This chapter will 1) discuss the performance issue of big data analytics in loosely-coupled distributed systems; 2) describe some solution approaches to address the issue; and 3) introduce a case study to demonstrate the effectiveness of the solution approaches using a real world big data analytics application on hybrid cloud environments. Readers of this chapter will have a clear picture of the importance of the