Chapter 29
Data Intensive Cloud Computing:
Issues and Challenges

Jayalakshmi D. S.
M. S. Ramaiah Institute of Technology, India

R. Srinivasan
S. R. M. University, India

K. G. Srinivasa
M. S. Ramaiah Institute of Technology, India

ABSTRACT

Processing Big Data is a huge challenge for today's technology. There is a need to find, apply and analyze new ways of computing to make use of the Big Data so as to derive business and scientific value from it. Cloud computing with its promise of seemingly infinite computing resources is seen as the solution to this problem. Data Intensive computing on cloud builds upon the already mature parallel and distributed computing technologies such HPC, grid and cluster computing. However, handling Big Data in the cloud presents its own challenges. In this chapter, we analyze issues specific to data intensive cloud computing and provides a study on available solutions in programming models, data distribution and replication, resource provisioning and scheduling with reference to data intensive applications in cloud. Future directions for further research enabling data intensive cloud applications in cloud environment are identified.

INTRODUCTION

Massive amounts of data are being generated in scientific, business, social network, healthcare, and government domains. The "Big Data" so generated is typically characterized by the three Vs: Volume, Variety, and Velocity. Big data comes in large volumes, from a large number of domains, and in different formats. Data can be in structured, semi-structured or unstructured format, though most of the Big Data

DOI: 10.4018/978-1-4666-9840-6.ch029

is unstructured; the data sets might also grow in size rapidly. There are many opportunities to utilize and analyze the Big Data to derive value for business, scientific and user-experience applications. These applications need to process data in the range of many terabytes or petabytes and are called data intensive applications. Consequently, computing systems which are capable of storing, and manipulating massive amounts of data are required; also required are related software systems and algorithms to analyze the big data so as to derive useful information and knowledge in a timely manner.

In this chapter we present the characteristics of data intensive applications in general and discuss the requirements of data intensive computing systems. Further, we identify the challenges and research issues in implementing data intensive computing systems in cloud computing environment. Later in this chapter, we also present a study on programming models, data distribution and replication, resource provisioning and scheduling with reference to data intensive applications in cloud.

Data Intensive Computing Systems

Data Intensive Computing is defined as "a class of parallel computing applications which use a data parallel approach to processing large volumes of data" (”Data Intensive Computing”, 2012). They devote most of their processing time to I/O and manipulation of data rather than computation (Middleton, 2010). According to the National Science Foundation, data intensive computing requires a "fundamentally different set of principles' ' to other computing approaches. There are several important common characteristics of data intensive computing systems that distinguish them from other forms of computing(Middleton, 2010).

- Data and applications or algorithms are co-located so that data movement is minimized to achieve high performance in data intensive computing
- Programming models that express the high level operations on data such as data flows are used, and the runtime system transparently controls the scheduling, execution, load balancing, communications and movement of computation and data across the distributed computing cluster.
- They provide reliability, availability and fault tolerance.
- They are linearly scalable to handle large volumes of data.

Challenges and Research Issues for Data Intensive Computing Systems

Parallel processing using data-parallel approach is widely accepted as the way to architect data intensive applications. Many different system architectures such as parallel and distributed relational database management systems have been implemented for data intensive applications and big data analytics. However these assume that data is in structured form whereas most of the big data is in unstructured or semi-structured form.

Typical data intensive applications include scientific applications handling large amounts of geo-distributed data for which grid architectures have been used extensively. Hence, loosely coupled distributed systems with message passing are preferred over typical, tightly-coupled HPC systems. The challenge is to architect and implement applications that can scale to handle voluminous and geo-distributed data in different forms in a reliable manner, and in real time in some applications.

Cloud computing systems with their promise of seemingly infinite, elastic resources lend themselves to these requirements and hence data-intensive cloud applications are the focus of current research.