# Chapter 39
# Big Data Processing on Cloud Computing Using Hadoop Mapreduce and Apache Spark

**Yassir Samadi**
*Mohammed V University, Morocco*

**Mostapha Zbakh**
*Mohammed V University, Morocco*

**Amine Haouari**
*Mohammed V University, Morocco*

## ABSTRACT

*Size of the data used by enterprises has been growing at exponential rates since last few years; handling such huge data from various sources is a challenge for Businesses. In addition, Big Data becomes one of the major areas of research for Cloud Service providers due to a large amount of data produced every day, and the inefficiency of traditional algorithms and technologies to handle these large amounts of data. In order to resolve the aforementioned problems and to meet the increasing demand for high-speed and data-intensive computing, several solutions have been developed by researches and developers. Among these solutions, there are Cloud Computing tools such as Hadoop MapReduce and Apache Spark, which work on the principles of parallel computing. This chapter focuses on how big data processing challenges can be handled by using Cloud Computing frameworks and the importance of using Cloud Computing by businesses*

## INTRODUCTION

Cloud Computing and Big Data induce a major transformation in the digital use by all economic sectors companies. Related issues link the activity and job creation within the digital actors, and enable user companies to generate competitiveness gains. Nowadays, the enterprises and organizations are producing

DOI: 10.4018/978-1-7998-5339-8.ch039

and storing data on large scale every day and the rate is dynamic by nature, mainly in the web and online social networks applications, such as Facebook, Twitter, and YouTube, to name a few. The quantitative explosion of digital data has forced researchers and developers to find new ways of seeing and analyzing the world. This is to discover new orders of magnitude concerning acquisition, searching, sharing, storage, analysis and presentation of the data. The main concern of Big Data (Gandomi & Haider, 2015) is storing a tremendous amount of information on a numerical basis that becomes difficult to process with conventional database management tools. Big data is not just data, it is also a set of technologies, architecture, tools and procedures allowing an organization to quickly capture, process and analyze large quantities of heterogeneous data, and extract relevant information at an affordable cost. The main challenges of data-intensive computing are analyzing and processing exponentially growing data volumes for different purposes in a minimum delay. Also, new algorithms which can scale to search and process massive amounts of data should be developed. Several solutions are available to deal with the requirements of Big Data. Among the proposed solutions, there are Cloud Computing tools such as Hadoop MapReduce and Apache Spark.

Hadoop Mapreduce is a framework that has mainly been used to store and analyze a large amount of data. Hadoop was designed for batch processing providing scalability and fault tolerance but not fast performance (Apache Hadoop, 2017). It enables applications to run in thousands of nodes with petabytes of data. Hadoop Mapreduce responds to the large amount of data by splitting up the data elements and assigns each element in a given cluster node for analysis. It follows a similar strategy for computing by breaking jobs into a number of smaller tasks that will be executed in nodes of the cluster. However, Hadoop's performance is not suitable for real-time applications (SAP Business By Design, 2017) because it writes and reads data from and to an external storage system, e.g., a distributed file system. This generates additional overheads due to data replication and input/output operations on a physical disk, which can increase the application's execution time. To solve this problem, Matei Zaharia has proposed a new framework called Spark (Zaharia, Chowdhury, Michael, & Shenker, 2010). Spark minimizes these data transfers from and to disk by using effectively the main memory and performing in-memory computations. Also, Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming.

Cloud Computing affirms the ability to scale computing resources as needed without a large upfront investment in infrastructure and with affordable cost. Therefore, Cloud Computing facilitates movement towards Big Data, linked to the need for greater computing capacity and storage of data flow from the increased use of new digital technologies. Consequently, Companies should continue to manage an exponential increase in the volume of generated data (structured, semi- structured or unstructured) and analyze as soon as possible to try to extract value. Cloud Computing and Big Data represent a rapidly developing field, providing many opportunities for value creation.

This chapter focuses first on integration of Big Data frameworks on Cloud Computing environment and the reason that enterprises should migrate their applications to the cloud. The chapter is organized as follows: Section II gives an overview of Cloud Computing and its architecture. Section III provides comprehensive review of Big Data and its classification. Section IV describes the relationship between Cloud Computing and Big Data. In section V, the authors focus on the current researches targeting the issues and challenges of Big Data storage and management for analytics. Section VI outlines the two Cloud computational frameworks Hadoop Mapreduce and Apache Spark. Section VII discusses the main advantages and benefits of Big Data processing in Cloud Computing for business. Section VIII