Big Data Processing in Cloud Computing Environments

A. Noraziah 1, 2, Mohammed Adam Ibrahim Fakherldin3, Khalid Adam1, Mazlina Abdul Majid1

1Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, 26300 Kuantan, Pahang, Malaysia
2IBM Centre of Excellence, University Malaysia Pahang 26300 Kuantan, Pahang, Malaysia
3Faculty of Computer Science and Information Systems, Jazan University, P.O Box 114-, Saudi Arabia.

While technologies to build and run big data projects have started to mature and proliferate over the last couple of years, exploiting all potentials of big data is still at a relatively early stage. In fact, big data is term refer to huge data sets, have high velocity, high volume and high variety and complex structure with the difficulties of management, analyzing, storing and processing. Due to characteristic of big data it becomes very difficult to Management, analysis, Storage, Transport and processing the data using the existing traditional techniques. This paper introduces several big data processing technics from system and application aspects. First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data. Following, we present the cloud computing for big data and related work. Furthermore, we also discuss, big data moving to the cloud. Finally, we present the conclusion and future work.

Keywords: Big Data, Cloud Computing, Data Management; Distributed Computing.

1. INTRODUCTION

Nowadays, organizations are focusing to shift their data and infrastructure in the form of private or public clouds due to their increasing computational needs and explosion of data from exabyte to zettabyte. In order to handle the data growth, companies are looking towards new tools and technologies to process and visualize their data to make timely and better valued decisions. Massive data is being generated by many organizations in structured and unstructured form like online transactions, emails, photos, videos through social network sites, web pages, server logs etc. [1]. This rapid growth of colossal data is involving organizations into big data problems. Big data of sizes exabyte to zettabytes do not fit into the current database architectures and invites new challenges beyond the processing power of conventional software tools and hardware systems to manage, store and visualize the data in acceptable time delay. Challenges include processing power, storage space, searching, sharing, analyzing and visualizing such data. For example, Facebook manages 40 billion photos of users, Wallmart process more than one million customer transactions per hour which are being saved in the database that are estimated to contain approximately 2.5 petabytes of data [2]. There are more than 38000 Google searches per second and 2.2 billion email users sending 144 billion emails per day [3]. Present big data needs cannot be suitably handled with conventional relational databases, desktop analysis and statistical packages. There is need of immense parallel processing of software running on hundreds and the outstands of machines concurrently. However, big data pose four unique dimensions of problems described as four vs, namely, volume, velocity, variety and veracity [1], [4], [5]. volume refers to amount of data increasing exponentially day by