Journal of Network and Computer Applications 64 (2016) 229–238

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

A comprehensive review of the data replication techniques in the cloud environments: Major trends and future directions

Bahareh Alami Milani [1], Nima Jafari Navimipour *

Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form
9 January 2016
Accepted 11 February 2016
Available online 23 February 2016

Keywords:
Cloud computing
Replication
Big data
Static
Dynamic

ABSTRACT

Nowadays, in various scientific domains, large data sets are becoming an important part of shared resources. Such huge mass of data is usually stored in cloud data centers. Therefore, data replication which is generally used to manage large volumes of data in a distributed manner speeds up data access, reduces access latency and increases data availability. However, despite the importance of the data replication techniques and mechanisms in cloud environments, there has not been a comprehensive study about reviewing and analyzing its important techniques systematically. Therefore, in this paper, the comprehensive and detailed study and survey of the state of art techniques and mechanisms in this field are provided. Also, we discuss the data replication mechanisms in the cloud systems and categorize them into two main groups including static and dynamic mechanisms. Static mechanisms of data replication determine the location of replication nodes during the design phase while dynamic ones select replication nodes at the run time. Furthermore, the taxonomy and comparison of the reviewed mechanisms are presented and their main features are highlighted. Finally, the related open issues and some hints to solve the challenges are mapped out. The review indicates that some dynamic approaches allow their associated replication strategies to be adjusted at run time according to changes in user behavior and network topology. Also, they are applicable for a service-oriented environment where the number and location of the users who intend to access data often have to be determined in a highly dynamic fashion.

Contents

* Corresponding author. Tel.: +98 9144021694.
E-mail address: jafari@iaut.ac.ir (N. Jafari Navimipour).
[1] Tel.: +98 9144126036.

http://dx.doi.org/10.1016/j.jnca.2016.02.005
1084-8045/

1. Introduction

Cloud computing is a network-based infrastructure where information technology (IT) and computing resources such as operating systems, storage, networks, hardware, databases, and even entire software applications are delivered to users as on-demand services (Buyya et al., 2008). Cloud computing does not consider a lot of new technologies, however, it saves the cost and increases the scalability to manage IT services (Buyya and Ranjan, 2010). The provided in cloud computing are grouped into 4 categories, including Software as a Service (SaaS) (Almorsy et al., 2014; Buxmann et al., 2008; Choudhary, 2007; Lin et al., 2009; Zeng and Veeravalli, 2014), Infrastructures as a Service (IaaS) (Bhardwaj et al., 2010; Iosup et al., 2014; Khajeh-Hosseini et al., 2010; Lin et al., 2009; Nathani et al., 2012; Wang et al., 2013; Zeng and Veeravalli, 2014), Platforms as a Service (PaaS) (Dinesha and Agrawal, 2012; Eludiora et al., 2011; Lin et al., 2009; Mell and Grance, 2009; Miller and Lei, 2009; Sellami et al., 2013; Zeginis et al., 2013; Zeng and Veeravalli, 2014) and Expert as a Service (EaaS) (Ashouraie et al., 2015; Nima Jafari Navimipour and Milani, 2015; Nima Jafari Navimipour, 2015; Nima Jafari Navimipour et al., 2015a, 2015b; Oussalah et al., 2014).

On the other hand, currently, in different scientific disciplines, an enormous amount of data is an important and vital part of shared resources. The mass of data is measured in terabytes and sometime in petabytes in many fields. Such enormous mass of data is typically kept in the cloud data centers (Long et al., 2014). So, data replication is generally used to manage a great deal of data (Wolfson et al., 1997) by creating identical copies of data (files, databases, etc.) in geographically distributed sites, which are called replicas (Lamehamedi and Szymanski, 2007; Meroufel and Belalem, 2013). The advantage of data replication is speeding up data access, reducing access latency and increasing data availability (Berl et al., 2010; Long et al., 2013). A general method is using multiple replicas which are distributed in geographically-dispersed clouds to increase the response time to users. It is important to guarantee replica's availability and data integrity features; i.e., the same as the original data without any interfering and corruption. Remote data ownership checking is an effective method to prove the replica's availability and integrity (He et al., 2012). Replication is one of the most broadly studied phenomena in the distributed environments (Goel and Buyya, 2006) in which multiple copies of some data are stored at multiple sites where overheads of creating, maintaining and updating the replicas are important and challenging issues (Dayyani and Khayyambashi, 2013; Goel and Buyya, 2006).

Since data replication is coming to play an increasingly important role in the cloud, the purpose of this paper is to survey the existing techniques and to outline the types of significant challenges and issues that can be addressed in the cloud replication domain. To the best of our knowledge, this survey paper is a first attempt to comprehensively and systematically examine the data replication problem with a specific focus on the cloud. The contributions of this paper are as follows:

- Providing the basic concepts and terminologies which are used in the field of data replication.
- Discussing the data replication mechanisms in the cloud systems and categorizing them into two main groups including static and dynamic mechanisms.
- Presenting the taxonomy and comparison of the reviewed mechanisms and highlighting their features.
- Mapping out the related open issues and some hints to solve the existing problems.

The rest of this paper is structured as follows. Section 2 discusses the data replication approaches in a cloud environment and classifies them. Section 3 presents the taxonomy and comparison of the reviewed mechanisms. Section 4 maps out some open issues. At last, Section 5 comes up with the conclusion of this paper.

2. Data replication mechanisms

Replication has been an area of interest for many years in World Wide Web (Qiu et al., 2001), peer-to-peer networks (Aazami et al., 2004; Nima Jafari Navimipour and Milani, 2014), ad-hoc and sensor networking (Intanagonwiwat et al., 2000; Tang et al., 2008), and mesh networks (Jin and Wang, 2005). Replication is a strategy that creates multiple copies of some data and stored them at multiple sites (Goel and Buyya, 2006). Data replication is a technique which is used in the cloud to decrease the user waiting time, to increase data availability and to minimize cloud system bandwidth consumption utilizing different replicas of the same service (Ahmad et al., 2010). More recently, the emergence of large-scale distributed systems such as Grid (Dabrowski, 2009; Nima Jafari Navimipour et al., 2014; Navin et al., 2014; Souri and Navimipour, 2014) and cloud (Ashouraie et al., 2015; Bonvin et al., 2009; Jafari Navimipour et al., 2014; Nima Jafari Navimipour and Milani, 2015; Talia et al., 2016) has made data replication becoming a research hot spot once again. In data clouds, enormous scientific data and complex scientific applications require different replication algorithms, which have attracted more attention recently. Data replication techniques can be classified into two main groups including static and dynamic replication mechanisms that are shown in Fig. 1. The number of replicas and the host node is predetermined and well-defined in the static replication strategies (Ghemawat et al., 2003; Rahman et al., 2006; Shvachko et al., 2010). Whereas, dynamic strategies automatically create and remove replicas based on the changes in user access pattern, storage capacity and bandwidth (Chang and Chang, 2008; Doğan, 2009; Lei et al., 2008; Li et al., 2011; Wei et al., 2010). It makes intelligent choices about the location of data depending upon the information of the current situation. But, it has some drawbacks such as difficulty to collect runtime information of all the data nodes in a complex cloud infrastructure and hard to maintain consistency of data file (Long et al., 2014). Static and dynamic replication algorithms can be further classified into groups as distributed (Doğan, 2009; Ghemawat et al., 2003; Shvachko et al., 2010; Wei et al., 2010) and centralized algorithms (Chang and Chang, 2008; Lei et al., 2008; Rahman et al., 2006; Sun et al., 2012).

2.1. Static mechanisms

In this section, the static mechanisms of data replication and their basic properties are described. Then, eight most popular static mechanisms of data replication are discussed. Finally, these mechanisms are compared and summarized in Section 2.1.3.

2.1.1. Overview of the static strategies

Static replication strategies follow deterministic policies, therefore, the number of replicas and the host node is well-defined and predetermined (Long et al., 2014). Also, these strategies are simple to implement but it is not often used because it does not adapt according to the environment (Gill and Singh, 2015). In the next sub-section, some applicable and popular static data replication mechanisms in the cloud environments are reviewed and discussed.