A systematic literature review of the data replication techniques in the cloud environments

Bahareh Alami Milani, Nima Jafari Navimipour*

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

*Corresponding Author:

Name : Nima Jafari Navimipour (Ph.D.)

Tel. : +989144021694

Fax. : +984134203292

Email : jafari@iaut.ac.ir

Abstract

Cloud computing has various challenges, one of them is using copied data. Data replication is an important technique for distributed mass data management. The aim of the general idea of data replication is placing replications at different places, while there are several replications of a specific file at different points. Replication is one of the most broadly studied phenomena in the distributed environments 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. Applications and architecture of distributed computing have changed drastically during last decade and so has replication protocols. Different replication protocols may be suitable for different applications. However, despite the importance of this issue, in a cloud environment as a distributed environment, this issue has not been investigated so far systematically. The data replication in the cloud environment falls into two categories of static and dynamic methods. In the static patterns, a number of created replicas is constant and fixed from the beginning. The number is either determined by the user from the beginning or the cloud environment determines such number. However, in the dynamic algorithm and considering its environment, the number is determined based on user's access algorithm. The objective of this paper is to review the data replication techniques in these two main groups systematically as well as a discussing the main features of each group.

Keywords: Cloud computing, data replication, static, dynamic, systematic literature review.

1. Introduction

Nowadays, wide developments on IT-based systems with the recent advances in cloud computing (Asghari & Navimipour, 2016; Ashouraie & Jafari Navimipour, 2015; Chiregi & Navimipour, 2016; B. A. Milani & N. J. Navimipour, 2016; Nima Jafari Navimipour, Rahmani, Navin, & Hosseinzadeh, 2015), grid computing (Nima Jafari Navimipour & Khanli, 2008; Nima Jafari Navimipour, Rahmani, Navin, & Hosseinzadeh, 2014; Souri & Navimipour, 2014), expert cloud(Ashouraie & Jafari Navimipour, 2015; Nima Jafari Navimipour, Rahmani, et al., 2015), peer to peer computing(Chiregi & Navimipour, 2016) and wireless networks (Jafari & Es-Hagi, 2011; Nima Jafari Navimipour, 2011; Nima Jafari Navimipour & Rahmani, 2009; Nima Jafari Navimipour, Shabestari, & Samaei, 2012) has been created. These technologies facilitate data access and resource sharing (A. S. Milani & N. J. Navimipour, 2016).

Cloud computing as a network-based infrastructure (Nima Jafari Navimipour & Soltani, 2016; Zareie & Navimipour, 2016) provides computing resources such as operating systems, storage, networks, hardware, databases, and even entire software applications to users as on-demand fashion (Buyya, Yeo, & Venugopal, 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 & Ranjan, 2010). Cloud services are classified into some categories such as SaaS (Software as a Service) (Almorsy, Grundy, & Ibrahim, 2014; Buxmann, Hess, & Lehmann, 2008; Choudhary, 2007; Lin, Fu, & Zhu, 2009; Zeng & Veeravalli, 2014), IaaS (Infrastructures as a Service) (Bhardwaj, Jain, & Jain, 2010; Iosup, Prodan, & Epema, 2014; Khajeh-Hosseini, Greenwood, & Sommerville, 2010; Lin et al., 2009; Nathani, Chaudhary, & Somani, 2012; Wang, Liang, & Li, 2013; Zeng & Veeravalli, 2014), PaaS (Platforms as a Service) (Dinesha & Agrawal, 2012; Eludiora et al., 2011; Lin et al., 2009; Mell & Grance, 2009; Miller & Lei, 2009; Sellami, Yangui, Mohamed, & Tata, 2013; Zeginis et al., 2013; Zeng & Veeravalli, 2014) and EaaS (Expert as a Service) (Ashouraie, Jafari Navimipour, Ramage, & Wong, 2015; N Jafari Navimipour & Milani, 2015; Nima Jafari Navimipour, 2015; Nima Jafari Navimipour & Khezr, 2015; Nima Jafari Navimipour, Navin, Rahmani, & Hosseinzadeh, 2015; 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, Zhao, & Chen, 2014). So, data replication is generally used to manage a great deal of data (Wolfson, Jajodia, & Huang, 1997) by creating identical copies of data (files, databases, etc.) in geographically distributed sites, which are called replicas (Lamehamedi & Szymanski, 2007; Meroufel & Belalem, 2013). The advantage of data replication is speeding up data access, reducing access latency and increasing data availability (Berl et al., 2010; Long, Zhao, & Chen, 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, Zhang, Huang, Shi, & Cao, 2012). Replication is one of the most broadly studied phenomena in the distributed environments (Goel & 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 & Khayyambashi, 2013; Goel & Buyya, 2006).

Nevertheless, to the best of our knowledge, despite the importance of data replication mechanisms in cloud environments, there is not any detailed and comprehensive systematic review of these mechanisms. Therefore, the purpose of this paper is to survey the existing techniques, compares the differences between mentioned mechanisms and outlines the types of challenges that could be addressed. We divided most of the introduced data replication algorithms into two main categories, static and dynamic. To the best of our knowledge, this survey represents the first attempt to systematically examine data replication with a specific focus on cloud computing. Briefly, the contributions of this paper are as follows:

- Providing background information about related concepts regarding this study.
- Describing how the systematic literature review was conducted.
- Discussing the findings and how they map to prior research.

The rest of this paper is structured as follows. The background information is provided in the next section. Section 3 discusses how the systematic literature review conducted. Section 4 presents explanation about data replication. At last, Section 5 comes up with the conclusion of this paper.

2.1 Background information

In this section, we describe the background related to data replication and advantages and disadvantaged of data replication in cloud computing environment, and introduce the key concepts that using in this study. First, we introduce the data replication. Then, we review the mechanisms of data replication in a cloud environment and give a brief overview of existing review studies on replication.

Replication is one of the most widely studied phenomena in a distributed environment. It is a strategy in which multiple copies of some data are stored at multiple sites. The reason for such a widespread interest is due to following facts: high availability, high performance, and high reliability. By storing the data at more than one node, if a data node fails, a system can operate using replicated data, thus increasing availability and fault tolerance. At the same time, as the data is stored at multiple nodes, the request can find the data close to the site where the request originated, thus increasing the performance of the system. But the benefits of replication, of course, do not come without overheads of creating, maintaining and updating the replicas. If the application has read-only nature, replication can greatly improve the performance. But, if the application needs to process update requests, the benefits of replication can be neutralized to some extent by the overhead of maintaining consistency among multiple replicas.

Replication has been an area of interest for many years in World Wide Web (Qiu, Padmanabhan, & Voelker, 2001), peer-to-peer networks (Aazami, Ghandeharizadeh, & Helmi, 2004; Nima Jafari Navimipour & Milani, 2014), ad-hoc and sensor networking (Intanagonwiwat, Govindan, & Estrin, 2000; Tang, Gupta, & Das, 2008), and mesh networks (Jin & Wang, 2005). Replication is a strategy that creates multiple copies of some data and stored them at multiple sites (Goel & Buyya, 2006). It 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, Fauzi, Sidek, Zin, & Beg, 2010). More recently, the emergence of large-scale distributed systems such as Grid (Dabrowski, 2009; Nima Jafari Navimipour et al., 2014; Navin, Navimipour, Rahmani, & Hosseinzadeh, 2014; Souri & Navimipour, 2014) and cloud (Ashouraie et al., 2015; Bonvin, Papaioannou, & Aberer, 2009; Jafari Navimipour, Masoud Rahmani, Habibizad Navin, & Hosseinzadeh, 2014; N Jafari Navimipour & Milani, 2015; Talia, Trunfio, & Marozzo, 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. In a static replication strategy, the host node and the number of replicas are predetermined and well-defined (Ghemawat, Gobioff, & Leung, 2003; Rahman, Barker, & Alhajj, 2006; Shvachko, Kuang, Radia, & Chansler, 2010). Whereas, dynamic strategies automatically create and remove replicas based on the changes in user access pattern, storage capacity and bandwidth (Chang & Chang, 2008; Doğan, 2009; Lei, Vrbsky, & Hong, 2008; Li, Yang, & Yuan, 2011; Wei, Veeravalli, Gong, Zeng, & Feng, 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 distributed (Doğan, 2009; Ghemawat et al., 2003; Shvachko et al., 2010; Wei et al., 2010) and centralized algorithms (Chang & Chang, 2008; Lei et al., 2008; Rahman et al., 2006; Sun, Chang, Gao, Jin, & Wang, 2012).

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 & Singh, 2015). In the static patterns, a number of created replicas is constant and fixed from the beginning. The number is either determined by the user from the beginning or the cloud environment determines such number.

Dynamic strategies for data replication in cloud environments automatically create and delete the replicas according to changes in user access pattern, storage capacity and bandwidth (Chang & Chang, 2008; Doğan, 2009; Lei et al., 2008; Li et al., 2011; Wei et al., 2010). They make intelligent choices about the location of data depending upon the information of the current environment. But, it has some drawbacks such as difficulty to collect runtime information of all the data nodes in a complex cloud infrastructure and maintaining the data file consistency (Long et al., 2014). Dynamic data replication strategies include some phases: analyzing and modeling the relationship between the number of replicas and system availability; recognizing the popular data and triggering a replication operation when the popularity data passes a dynamic threshold; evaluating a suitable number of copies to meet a reasonable system byte effective rate requirement and insertion replicas among data nodes in a balanced way; and designing the dynamic data replication algorithm in a cloud.

3. Related work

Many types of research have been done in the field of cloud computing and general challenges including data replication, scheduling, resource discovery and etc. However, there is a little comprehensive research about data replication in cloud computing has been done yet. In this section, we describe to some papers that there are in the field of data replication in cloud computing.

One of the survey of data replication is about dynamic replication mechanisms in the grid have proposed by Amjad, Sher, and Daud (2012), this paper has classified dynamic replication strategies for a data grid environment. All replication techniques address some attributes like fault tolerance, scalability, improved bandwidth consumption, performance, storage consumption, data access time etc. In this paper, different issues involved in data replication are identified and different replication techniques are studied to find out which attributes are addressed in a given technique and which are ignored. It can be seen that different strategies have presented their own terms for the evaluation of their proposed methods. Most of the techniques included in this survey have used simulation to evaluate and test the algorithms. The paper also includes some discussion about future work in this direction by identifying some open research problems. However, their data replication survey was in grid computing.

Tarek Hamrouni, Slimani, and Charrada (2015) have proposed a critical survey of data grid replication strategies based on data mining techniques. The main objective of this paper consists in

the study of how the data mining techniques can be applied to access historical data of data grids and how do they infer file correlations knowledge and use them to enhance replication strategies performance. Also, a new guideline to data mining application in the context of data grid replication strategies. In this paper, have suggested this guideline would facilitate further research works in this promising area and to give hints to other works to be done in the area of data mining and data grid replication. From this survey, it can be seen that the number of the proposed replication strategies based on data mining techniques is limited and so there is still a lot of work to be done in the field of data replication based on data mining. However, this survey was limited to data mining techniques and their data replication survey was in the field of the grid.

Another survey is a survey of dynamic replication and replica selection strategies based on data mining techniques in data grids that have proposed by T Hamrouni, Slimani, and Charrada (2016). This paper has focused particularly on how extracted knowledge enables enhancing data replication and replica selection strategies which are important data management techniques commonly used in data grids. Indeed, relevant knowledge such as file access patterns, file correlations, user or job access behavior, prediction of future behavior or network performance, and so on, can be efficiently discovered. These findings are then used to enhance both data replication and replica selection strategies. Various works in this respect are then discussed along with their merits and demerits. In addition, they have proposed a new guideline to data mining application in the context of data replication and replica selection strategies. However, this survey was limited data mining techniques and their data replication survey was in the field of the grid.

Spaho, Barolli, and Xhafa (2014) have proposed a survey about data replication strategies in P2P systems it conducts a theoretical survey of replication strategies in P2P systems. Also, it described different strategies and discussed their advantages and disadvantages and Replica Placement Strategies in unstructured P2P networks and they classified replica placement strategies by using two criteria: techniques related to site selection and techniques related with replica distribution In order to increase availability and reliability, data replication techniques are considered commonplace in P2P computing systems. However, their data replication survey was in P2P systems.

Another survey is proposed by Malik et al. (2016); they have studied data replication and management, two instrumental technologies that are widely used to manage massive quantities of data on cloud services. These techniques are required to assure strict QoS on data operations (search, upload, download, replicate, and the like). A comprehensive survey of techniques along with the (a) advantages, (b) disadvantages, (c) assumptions, and (d) SLA-based performance metric topographies are explored in this paper. The techniques are compared and analyzed based on the abovementioned features. They also analyze the working of numerous data replication techniques and how data-intensive applications are deployed in the cloud. The knowledge provided in the paper can be further exploited to design and model new mechanisms or approaches in the cloud. Furthermore, the analysis of each approach, issue, and suitability to support and operate in certain environments led us to the identification of the following open research issues. However, this survey did not provide a systematic literature review.

B. A. Milani and N. J. Navimipour (2016) have proposed a review of data replication mechanisms in a cloud environment, the comprehensive and detailed study and survey of the state of art techniques and mechanisms in this field are provided. Also, they 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. However, this survey did not provide a systematic literature review.

It is important to point out that none of these surveys present a pure systematic literature-based review of the existing data replication techniques with a discussion on their categorization, future challenges that data replication could have in a cloud environment. We formalized some questions about data replication in cloud computing in the next section. We select papers about data replication in cloud computing to answer these questions and we can explore data replication mechanisms and the challenge of them.

4. Research method

An SLR is a research method (Kitchenham, 2004) originating from the field of medicine which provides a repeatable research method and should supply sufficient detail to be replicated by other researchers(Nima Jafari Navimipour & Charband, 2016). In terms of lead to detailed answer within necessity of cloud computing, we developed four research questions to address the key concerns of data replication in cloud computing. In the next section, we formalize these questions. At the moment, there is no systematic literature review (SLR) on the data replication in a cloud environment. This section provides a systematic literature review on state-of-the-art approaches and techniques in addressing data replication in cloud computing(Jula, Sundararajan, & Othman, 2014). An SLR was chosen as a research method because the study is more about trying to understand a problem than trying to find a solution to it. Also, there was already existing literature that could be synthesized. An SLR is a research method originating from the field of medicine(Kupiainen, Mäntylä, & Itkonen, 2015). It requires a comprehensive and unbiased coverage of searched literature. To maximize the coverage of our searched literature, we started by identifying some of the most used alternative words/concepts and synonyms in the research questions.

4.1. Research questions

The present research aims at collecting and investigating all of the credible and effective studies that have examined data replication in the cloud environment and study feature, challenge and relevant issues in data replication in the cloud. More specifically, the extraction of important features and methods of papers in cloud data replication will be considered, and their characteristics will be described. To achieve the above-mentioned goals and identify the methods that have been selected by researchers for their studies and result assessment methods, case studies are covered for which new methods are proposed and datasets and benchmarks are used. The following research questions (RQs) are raised.

- RQ1: What is the importance of data replication in cloud computing? The purpose of this question is to show the number of published papers about data replication in the cloud.

- RQ2: How many the number of data replication techniques is? The purpose of this question is to show existing method of data replication in a cloud environment.
- RQ3: What is the problem of data replication in cloud computing? The purpose of this question is to identifying challenge of data replication and the role of replication in the cloud.

We search the papers about data replication to answer these questions.

4.2. Search query

The goal of the search process is the identification of the journal articles that investigated the data replication mechanisms and it focuses on the factor of research that affect their acceptance. The search process for the review uses online scientific databases. We defined a query string by selecting the most appropriate keywords. The research questions are used to construct a search string in the databases. By adding synonyms and alternative spellings for the question elements, the following search string was defined: (cloud replication) OR (cloud data replication) OR (data replication in the cloud).
The search was conducted from 2010 to 2015.

4.3. Selection of sources

The search string was limited by searching only for journal articles and conference papers as they obtain validated results. The query strings were applied titles, abstract and body of studies, the search was conducted 2010-2015 using the online scientific database. We subsequently classified and analyzed these publishers in order to extract relevant results. Google scholar was adapted as our data source. Online databases included Elsevier, IEEE, Springer, ACM, and DOAJ. The online database was shown in table 1.

Table1: Online database

Online database	URL
Elsevier	http://www.Sciencedirect.com
IEEE	http://www.ieeexplore.ieee.org
Springer	http://link.springer.com
ACM	http://dl.acm.org
DOAJ	http://doaj.org

4.4. Study selection

We selected papers according to a search query on online databases. First, we select articles that their subjects are about computer science and information technology and other articles ignored. Then, we checked the title of the papers and removed the irrelevant title. Then, we ignored books and removed the paper was written in not English language and studied abstract and conclusions of selected papers and ignored papers that seemed to be unused.

In summary, our study selection process includes 4 stages:

1. Exclusion based on the subject area (Computer Science and information technology).
2. Exclusion based on the title.
3. Exclusion based on removing unusable articles (We removed books and articles not written in the English language).
4. Exclusion based on the content of abstract and conclusions.

Fig 1 shown this process.

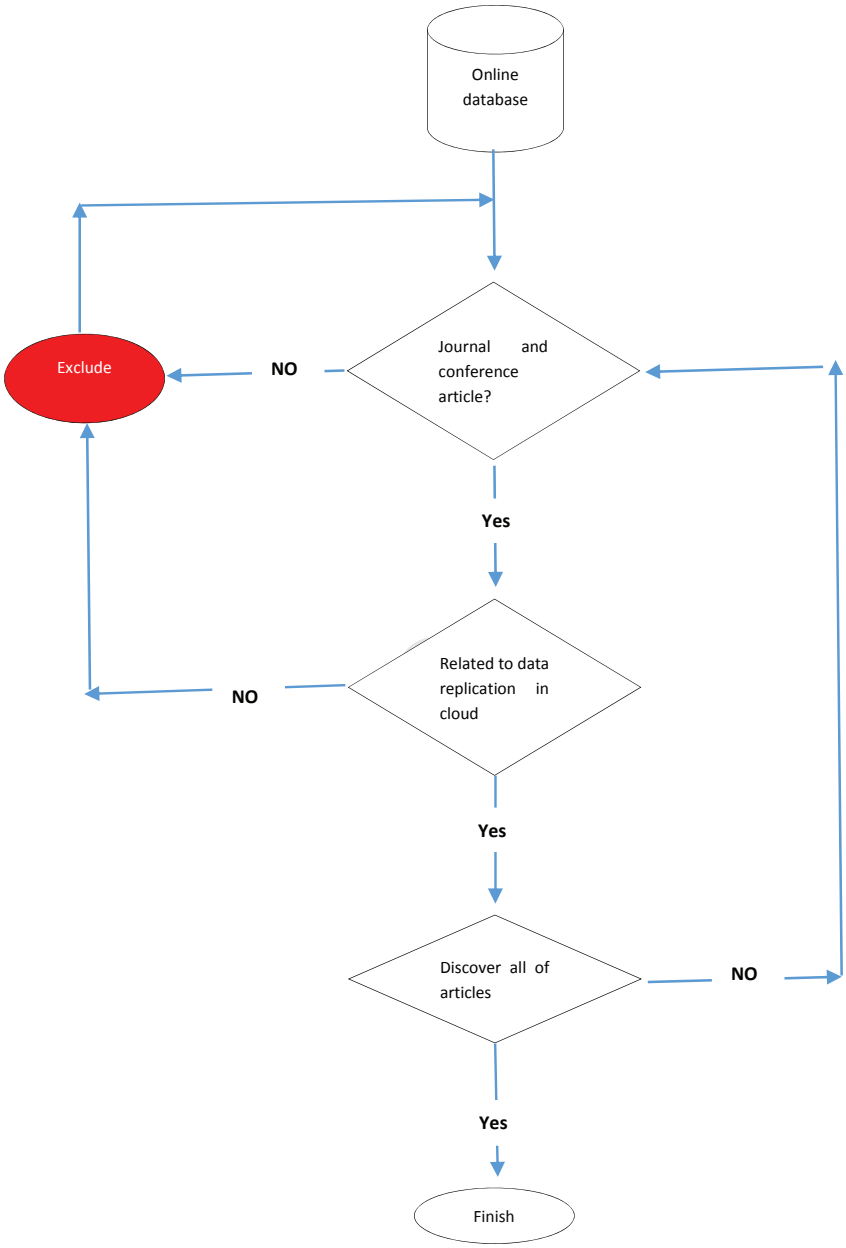

Fig1. Study process

4.5 Results

This section summarized the results of the systematic literature review. Only 41 papers have qualified, the distribution of articles by year of publication is shown in Fig. 2. It shows the distribution of the articles by publishers are related to data replication mechanisms in cloud environments from 2010 to 2015 among 6 publishers, where 59% of the total article of journals belong to IEEE. To further investigate the foundation journal of articles 20% of the literature is related to the Springer, 11% of the literature are related to Elsevier, 8% of the literature are related to DOAJ and 2% of the literature are related to ACM. Due to our first formalization question (RQ1), it distinctly outlines the importance of data replication and necessity of new and improved data replication mechanisms along with the rise in the utilization of cloud computing. Fig. 3 shows the distribution of the articles in all investigated categories including Elsevier, Springer, Emerald, IEEE, ACM, DOAJ and ACM.

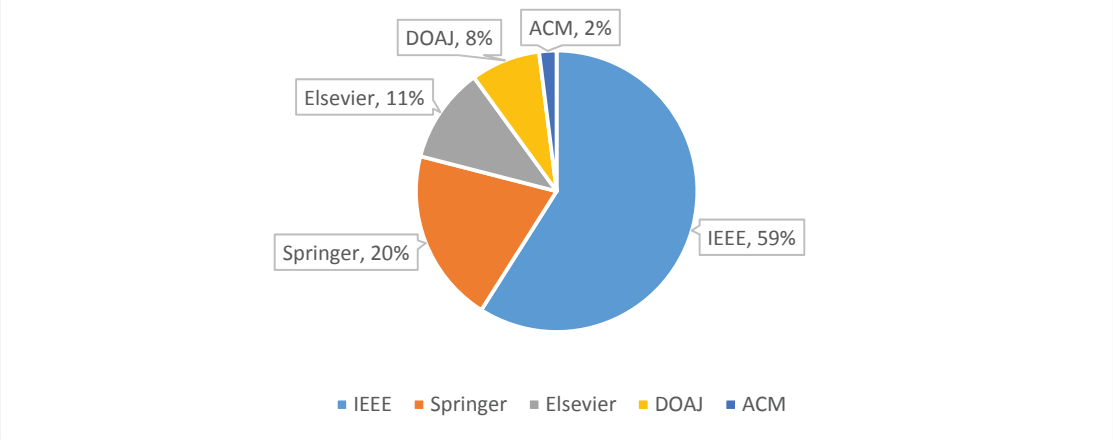

Fig2. Distribution of paper by publisher

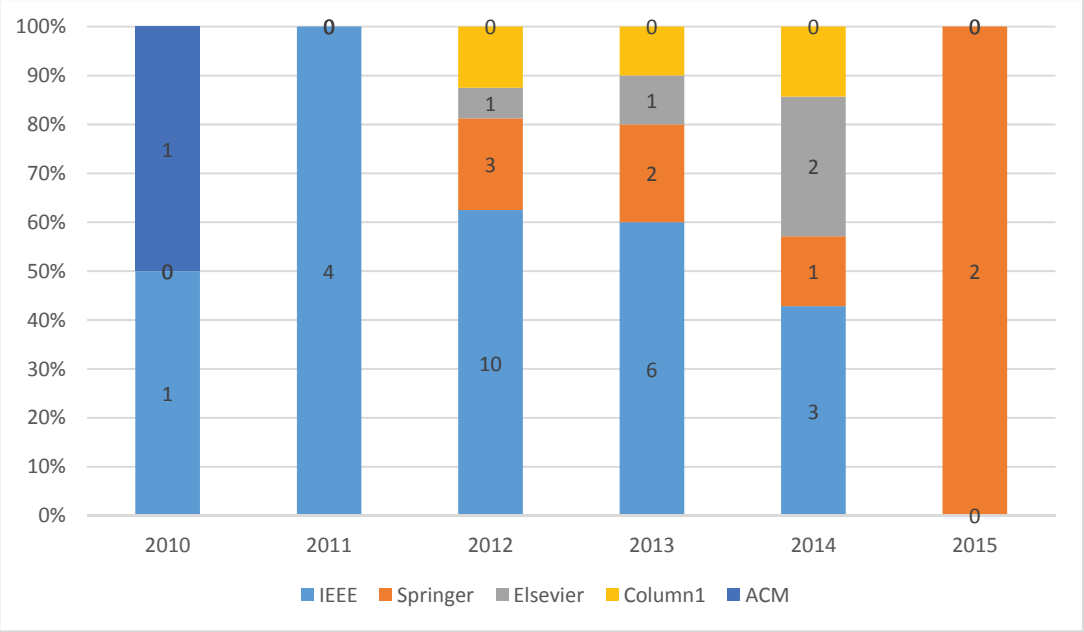

Fig3. Number of paper between 2010-2015

5. Data replication strategies

In this section, we reviewed selected studies to answer RQ2 and RQ3. Data replication is an important technique for distributed mass data management. The aim of the general idea of data replication is placing replications at different places, while there are several replications of a specific file at different points. The data replication in the cloud environment falls into two categories of static and dynamic methods. In the static patterns, a number of created replicas is constant and fixed from the beginning. The number is either determined by the user from the beginning or the cloud environment determines such number. However, in the dynamic algorithm and considering its environment, the number is determined based on user's access algorithm. Static approaches determine the locations of replication nodes during the design phase while dynamic ones select replication nodes at a run time. Some dynamic approaches even allow their associated replication strategies to be adjusted at run time according to changes in user behavior and network topology. Dynamic replication is generally more appropriate 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. In the static replication strategy, the number of replicas and their locations is initially set in advance. Instead, dynamic replication strategy dynamically creates and deletes replicas according to changing environment load conditions. There has been an interesting number of works for data replication in the Cloud computing. Where most of them compared and analyzed in this paper.

By comparing the mechanisms, a specific mechanism to provide all mentioned issues will become a challenging problem and are interesting lines for future research and work. Also, consistency is a problem of replication, maintaining data integrity and consistency in a replicated environment is of prime importance. High precision applications may require strict consistency of the updates made by transactions. In these papers solve this problem by using a lazy update. The lazy update method is used to separate the processes of data replica updates and data accesses, which can improve the throughput of data accesses and reduces response time. Another challenge of replication is downtime during new replica creation, if strict data consistency is to be maintained, performance is severely affected if a new replica is to be created. As sites will not be able to fulfill request due to consistency requirements. Maintenance overhead is another problem in this scope if the files are replicated at more than one sites, it occupies storage space and it has to be administered. Thus, there are overheads in storing multiple files. Last challenge that mentioned in these papers is lower write performance, Performance of write operations can be dramatically lower in applications requiring high updates in replicated environment because the transaction may need to update multiple copies.

6. Conclusion

This paper presented a systematic review of data replication. In this paper, we have reviewed the past and the state of the art mechanisms in the field of cloud data replication. Furthermore, we introduced a taxonomy of the reviewed cloud data replication mechanisms. Data replication in cloud environment increases the availability of data, Increased performance, and enhanced reliability by storing the data at more than one site, if a data site fails, a system can operate using replicated data, thus increasing availability and fault tolerance. At the same time, as the data is stored at multiple sites, the request can find the data close to the site where the request originated, thus increasing the performance of the system.

Regrences