REBUILDING OF DATA IN A CLOUD COMPUTING: A SURVEY

Rabit Ul Islam[1], Jyoti Arora[2],
[1]*M.Tech Student, Desh Bhagat University,Mandi Gobindgarh*
[2] *Assistant Professor, Desh Bhagat University,Mandi Gobindgarh*
(E-mail: arr.raabit@gmail.com)

***Abstract*—**in cloud computing, data generated in electronic form are large in amount. To maintain this data efficiently, there is a necessity of data recovery services. Cloud computing provides accessing of any kind of services dynamically over Internet on demand basis. One of the most significant service that is being provided is storage as a service. Cloud customer can store any amount of data into cloud storage results to huge amount of data at the datacenter. The data may get deleted by man-made disaster (either CSP or customer itself without their knowledge) or by natural disasters (either earth quakes or volcanoes) from the datacenters. Nowadays, data has been generated in large quantity that requires the data recovery services or techniques. Therefore there is a requirement for designing an efficient data recovery technique to recover the lost data. Many researchers have proposed different data recovery techniques but they lack in efficiency and reliability. This paper provides an extensive survey on rebuilding of data and research in the cloud environments. We present different taxonomy of data recovery mechanisms, main challenges and proposed solutions. We also describe the cloud-based data recovery platforms and identify open issues related to data recovery.

***Keywords*—** *Cloud Computing; IAAS; PAAS; SAAS.*

I. INTRODUCTION

Cloud computing becomes more popular in large-scale computing day by day due to its ability to share globally distributed resources. Users can access to cloud-based services through Internet around the world. The biggest IT companies are developing their data centers in the five continents to support different cloud services. The total value of the global cloud computing services market revenues is expected to reach about $241 billion by the end of 2020 (Reid et al., 2011). Rapid development in cloud computing is motivating more industries to use variety of cloud services (Arean, 2013), for instance near to 61% of UK businesses are relying on some kinds of cloud services (White paper, 2013). However, many security challenges have been raised, such as risk management, trust and recovery.

Cloud Service Models Software as a Service (SaaS): SaaS is a collection of application and software; it allows the clients to subscribe the software instead of purchasing it. Software application is presented as service to the customer based on their demand. Twitter, Facebook, whats app provides Software as a service.

Platform as a Service (PaaS): This model provides platform as a service. This provides clients to develop his own application using the tools and programming languages. This service is hosted in cloud and accessed by clients using internet. Google App engine, Amazon AWS provides the platform as service.

Infrastructure as a Service (IaaS): This model provides the shared resource services. It provides the computing infrastructure like storage, virtual machine, network connection, bandwidth, IP address. IaaS is complete package for computing. Amazon, Go Grid provides the infrastructure as the service to the user [1].

A. Cloud Deployment Models Public

Cloud: A public cloud is available to any user with an internet facility, is less secure than the private cloud because it can be accesses by general public.

Private Cloud: Private cloud is available to a specific organization so that the user who belongs to that organization can have access the data. It is more secure than the cloud because of its private nature.

Hybrid Cloud: The hybrid cloud is basically combination of no less than two clouds such as combination of private, community or public cloud.

Community Cloud: Community cloud allows the resources and system to be accessible by number of associated organization. Data storage is one of the most significant services provided by cloud computing technology. But, recovering the lost data is one of the challenging issue in cloud computing paradigm. A brief overview of data recovery in cloud computing is discussed below.

B. Data Recovery in Cloud Computing

Data stored at the datacenter is increasing day by day it leads into huge amount of data storage in cloud and results into issues such as data loss, data breach etc. There is a need of an efficient technique if the data get destroyed or deleted by mistake to recover the data from any backup server. In business continuity if the system crashed or any type of natural or human made disaster occurred then there is chance of data loss and it may also cause the financial loss. By using some of the data recovery techniques the original data can be recovered. But, the existing recovery techniques are not efficient and reliable hence, to recover the lost original data a technique is needed to meet efficiency and reliability.

II. RELATED WORK

This section presents summary of some of the data backup and recovery techniques in cloud computing.

In paper [2], author has proposed Hierarchical attribute based user classification algorithm to prevent the access of information from les privileged user. To do so author proposed a delegation approach. Further to provide the physical control of data to data owner author has divided the data into three categories such as Privacy Not Required (PNR), Privacy Required with Trusted Provider (PRTP) and Privacy Required with Non-Trusted Provider (PRNTP).

In paper [3], author has proposed the DR-Cloud model which is fault tolerant multi cloud storage, it makes use of DR XOR codes which provides data redundancy and uses minimum repair traffic during data transmission. DR-Cloud acts as interface between user application and multi cloud server.

The paper [4] presents the novel technique to recover the data. It solves all existing problems with data recovery by automatically compressing and decompressing the data before the backup of the data. Dual backup system was used. The dual system provides the high reliability and better bandwidth utilization of data storage.

In paper [5], author has proposed the Advanced Encryption Standard (AES) and Seed Block algorithm (SBA) method to perform the smart remote data backup in cloud computing environment. The proposed technique uses the AES and seed block algorithm. If the data gets deleted by mistake then we can get it from the remote server. This method takes less time to recovery the data and solves the time related issues. Thus the method provides an efficient security mechanism for the data stored in the cloud environment.

The paper [7] presents the cloud mirroring technique. It uses the mirroring algorithm. The method provides the high availability, integrity of the data, recovery of the data and minimizes the data loss. This method can be applied to any kind of the cloud. Cost to recover the data is also less.

In paper [8], author proposed the data backup and recovery technique. This technique provides the data protection from the service failure and also decreases the cost of solution. By using this technique the process of migration becomes simple and also removes the cloud vendor dependency. They proposed an effective data backup technique to recover the data from the server in case of data loss. For every business it is essential to back up the data to avoid the data loss.

The paper [9] presents a method which includes business service procedure (BSP) and disaster recovery procedure (DRP) with an assistance of cloud environment in order to avoid disaster recovery problems. The work employs priority based technique towards data recovery. The proposed approach ensures that it can provide security to entire organization datasets, which may contain log, account files. It also ensures that it can minimize the time required to get better organization data within small amount of time.

In paper [10], review has been done on distribution of data in cloud environment by construction of privacy preserving techniques and RBD. In order to perform smart RBD the system employs encryption and compression methods. The paper aims to preserve user privacy. System verified that it can overcome time related issues and are also solved by encryption and compress techniques.

In paper [11], the author implements the PRS algorithm for distributed disaster recovery system. The system On DDR provides the data security through 1+1+N distributed architecture in case of multi-node damage. To improve the system performance it uses the RS erasure coding and it also helps to reduce the storage resource consumption caused by data redundancy.

In the paper [12] discussed the tools to study the disaster management. Now-a-days cloud computing technology is increasing day by day; the huge amount of data is stored on cloud. There is a chance of data loss and disaster. There is necessary to study the tools to manage the disaster in cloud environment. This paper aims to analyzing the various types of disaster and recovery techniques.

The paper [13] presented the Secure Erasure Coding (SEC) technique. This technique helps to retrieve the data from remote server in the absence of network connection and helps to recovery the data even if the data deleted from the server or cloud get destroyed. This method does not uses any kind of encryption techniques but also provides the security. It uses the less time to recovery the data.

From literature survey we found different techniques to recover the data. Each technique has its own advantages and disadvantages. During the data recovery process there present some issues. These issues are discussed below.

Issues Identified Data Storage: All the enterprises stores there large amount of data in the cloud. For providing the security to data the computing is distributed but storage is centralized. Therefore single point failure and data loss are critical challenge to store the data in cloud.

Data security: User stores their huge data in the cloud. The stored data may be confidential and sensitive data. Providing the security to these data is important [6].

Lack of redundancy: If the cloud gets destroyed due to any reason then secondary site gets activated in order to provide the data to user when primary storage fails to provide the data. Dependency: Customer doesn't have control on their system and data. Backup service is provided to overcome this drawback.

III. DATA RECOVERY CHALLENGES

A. Dependency:

One of the disadvantages of cloud services is that customers do not have control of the system and their data. Data backup is on premises of service providers as well. This issue makes dependency on CSPs for customers (such as organizations) and also loss of data because of disaster will be a concern for customers. Dependency (Javaraiah, 2011) also creates another challenge which is the selection of a trusted service provider.

B. Cost

It is obvious that one of the main factors to choose cloud as a DR Service is its lower price. So, cloud service providers always seek cheaper ways to provide recovery mechanisms by minimizing different types of cost. The yearly cost of DR systems can be divided in three categories (Alhazmi and Malaiya, 2012): • Initializing cost: amortized annual cost • Ongoing cost: storage cost, data transfer cost and processing cost • Cost of potential disaster: Cost of recovered disasters and also cost of unrecoverable disasters.

C. Failure Detection

Failure detection time strongly effects on the system downtime, so it is critical to detect and report a failure as soon as possible for a fast and correct DR. On the other hand, in multiple backup sites there is a major question: How to distinguish between network failure and service disruption.

D. Security

As mentioned before, DR can be created by nature or can be human-made. Cyber-terrorism attack is one of human-made disasters which can be accomplished for many reasons. In this case, protection and recovery of important data will be a main goal in DR plans beside of system restoration.

E. Replication Latency

DR mechanisms rely on replication technique to make backups. Current replication techniques are classified into two categories: synchronous and asynchronous (Ji et al., 2003). However, both of them have some benefits and some flaws. Synchronized replication, guarantees very good RPO and RTO, but it is expensive and also can effect on system performance because of large overhead. This issue is more serious in multi-tier web applications, because it can significantly increase Round Trip Time (RRR) between primary and backup site. On the other hand, a backup model adopted with a sync replication is cheaper and also system suffers low overhead, but the quality of DR Service will be decreased. Therefore, trading off between cost, performance of the system and also replication latency is an undeniable challenge in cloud disaster solutions.

F. Data Storage

Business database storage is one of the problems of enterprises which can be solved by cloud services. By increasing of cloud usage in business and market, enterprises need to storage huge amount of data on cloud-based storages. Instead of conventional data storage devices, cloud storage service can save money and is also more flexible. The architecture of a cloud storage system includes four layers: physical storage, infrastructure management, and application interface and access layer. In order to satisfy applications and also to guarantee the security of data, computing has to be distributed but storage has to be centralized. Therefore, storage single point of failure and data loss are critical challenges to store data in cloud service providers (Pokharel et al., 2010).

G. Lack of Redundancy

When a disaster happens, primary site becomes unavailable and secondary site has to be activated. In this case, there is no ability to sync or a sync replication in a backup site but data and system states only can be stored locally. It is a serious threat to the system. This issue is temporary and will be removed after recovery of the primary site. However, to achieve the best DR solutions, especially in high availability services (such as business data storage), it is better to consider all risky situations.

IV. DATA RECOVERY SOLUTIONS

DR Solutions In this section, we will discuss some DR solutions which have been proposed to overcome the problems and challenges in cloud-based DR.

A. Local Backup

A solution for dependency problem has been proposed in (Javaraiah, 2011). A Linux box can be deployed on the side of customers to make control of data and to get backup of both data and even complete application. Local storage can be updated through a secured channel. By this technique, migration between cloud service providers and also migration between public to private, and private to public is possible. In the event of a disaster, local backup can provide the services that were served by the service provider.

B. Geographical Redundancy and Backup (GRB)

Although geographical redundancy can be used in traditional model, but it is expensive and unaffordable. In (Pokharel et al., 2010), two cloud zones have a replication of each other. If one zone becomes down, then another zone will be on and provide the services. There is a module that monitors the zones to detect disaster. Primary zone has an active load balancer to request extra resources or even released unused resources. Second zone also has a passive load balancer. Another research (Khan and Tahboub, 2011) has been proposed a method to select optimal locations for multiple backup. The number of places is decided based on the nature of application and priority of services. Distance and bandwidth are two factors to choose the best sites in this method. However, this work neglects some critical factors such as the capacity of mirror sites and the number of node sources which can be hosted in each location.

C. Inter-Private Cloud Storage (IPCS)

This approach was proposed for cloud data storage (Jian-hua and Nan, 2011). According to Storage Networking Industry Association (SNIA), at least three backup locations are necessary for business data storage. Users' data should be stored in three different geographical locations: Servers, Local backup server (LBS) and remote backup server (RBS). The private clouds are established for any enterprises consist some servers and an LBS; and also an inter-private cloud storage is created in a public cloud consists the RBSs to be shared between public clouds. This model gives communication ability to backup locations in order to increase data integration.

D. Resource Management

Heterogeneous clouds consist many different hardware and software such as hybrid storage and diverse disks. In cloud-based enterprises, entire business data are stored in the cloud storage. So, data protection, safety and recovery are critical in these environments. Data in danger is the data which has been processed at the primary host but has not taken place in the backup host yet. So, in the case of disaster, it is necessary to use enhanced technology for data recovery in storage clouds. There are three solutions for data recovery proposed in (Patil et al., 2012):

- Using fastest disk technology in the event of a disaster for replication of data in danger.
- Changing dirty page threshold: The percentage of dirty pages in RAM which have to be waited for flushing to disk might be reduced (Rudolph, 1990).
- Prediction and replacement of risky devices: Some important factors such as power consumption, heat dissipation, carbon credit utilization and importance of data (stored on each disk) can be calculated in a specific period of time. By these factors, a mathematical equation will be formed to make a replace priority list.

E. Secure-Distributed Data Backup (SDDB)

An innovative technique has been presented in (Ueno et al., 2010) to protect data in the event of disaster. The data protection technique has six stages:

- First data encryption: Data has to be encrypted after receiving into a data center.
- Spatial scrambling: By a spatial scrambling algorithm, the order of data files is changed.
- Fragmentation, duplication: Data files are divided into some fragments and these fragments are duplicated in terms of service level agreement.
- Second encryption: Fragments are encrypted again with a different key.
- Shuffling & Distribution: In the last stage, fragments are distributed using a shuffling method into unused memory capacities.
- Transferring Metadata to backup server: Metadata including encryption keys, shuffling, fragmentation and distribute information is sent to a supervisory server. If a disaster happens, the supervisory server will gather all information from distributed devices and performs decryption (2nd), sort & merge, inverse spatial scrambling and decryption (1st), respectively.

F. Pipelined Replication

This replication technique (Wood et al., 2011) aims to gain both the performance of a sync replication and the consistency of sync replication. In sync replication, processing cannot continue until replication is completely finished at the backup site. Whereas, in a sync replication, after storing data in the local storage the process can be started. The result can be replied to the client, and then the writes are replicated to the backup site in an epoch. Pipelined replication performs replication and process in parallel as in the following scenario.

- Scenario of usage: The client sends a request to the web server. The web server processes the requests, then sends data to the local database in the primary data center. At this stage, the writes are flushed in the remote backup site, and the process operation can be performed in parallel. However, the reply to the client can be committed only after receiving the Ack from the backup site. Therefore, Pipeline replication facilitates replication procedure, and also guarantees the writes protection.

G. Scale Up/Down

Sometimes, performing functions with high priority can decrease money loss or even increase the revenue in the event of a disaster. Priority of services can be defined by some different features such as service level agreement, and the amount of revenue and urgent needs. After a natural disaster occurs in an area, cloud service providers are faced with flooding service requests. In this case, service providers have to manage their existent users' services and also handle new user requests. Service providers must satisfy existent users and should serve to new customers as much as possible. In (Nakajima et al., 2013), a management engine has been introduced for carrier networks. In case of a large scale natural disaster (like earthquakes), this system uses a DR scenario by scaling up resources for the high-priority services (e.g., voice communication) and scaling down allocated resources to low-priority service (e.g., video on-demand).

H. Dual-Role Operation

For increasing utilization of recourses, (Aghdaie and Tamir, 2003) introduces a simple technique. As shown in Figure 4, in this technique each host can operate as the primary host for some applications and can also be the backup host for some other applications. In this architecture, clients send their requests to the backup host first, then the backup host transmits those requests to primary host. After processing, primary host sends a log to the backup and finally reply to the clients. When a failure happens, the primary host becomes unavailable, and backup host has to handle the requests of the failed host. However, this technique cannot guarantee a good service restoration by itself, because backup site must share the resources between its own requests and redirected requests.

V. CONCLUSION

In this paper, we have provided an in depth analysis of the state of the art for DR in cloud computing. First, we briefly introduced cloud computing, including background, properties, advantages and challenges. Then, we discussed the details of cloud-based disaster recovery and compared it with traditional approaches. In addition, we also derived the main challenges of DR mechanisms and proposed solutions to overcome them. Furthermore, the main DR platforms are discussed, followed by open issues and future direction in the field of cloud-based DR mechanisms.

REFERENCES

[1] P. S. Challagidad, M. N. Birje, "Hierarchical Attribute-based Access Control with Delegation Approach in Cloud", Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353 2017 4th International Conference on "Computing for Sustainable Global Development", 01st - 03rd March, 2017.

[2] Greeshma Radhakrishnan, Chenni Kumaran, "DR – Cloud: Multi-Cloud Based Disaster Recovery Service", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, March 2016.

[3] Megha Rani Raigonda, Tahseen Fatima, "A Cloud Based Automatic Recovery and Backup System with Video Compression", International Journal of Engineering and Computer Science, ISSN: 2319-7242, Vol. 5, Issue 09, and September 2016.

[4] Tanay Kulkarni, Sumit Memane, "Intelligent Cloud Security Back-Up System", International Journal of Technical Research and Applications, Vol. 3, Issue 2, Mar-Apr 2015.

[5] Shilpi U. Vishwakarma and Praveen D. Soni, "Cloud Mirroring: A Technique of Data Recovery", International Journal of Current Engineering and Technology, Vol. 5, No. 2, March 2015.

[6] PS. Vijayabaskaran, "Efficient Backing up Data for Migrating Cloud to Cloud", International Journal of Computer Science and Information Technologies, Vol. 6, 2015.

[7] Atesh Kumar, Saurabh Mishra, "Priority with Adoptive Data Migration in Case of Disaster using Cloud Computing use style", International