Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

The impact of data recovery criteria, data backup schedule and data backup prosses on the efficiency of data recovery management in data centers

Maen T. Alrashdan[a*], Mutaz Abdel Wahed[a], Emran Aljarrah[a], Mohammad Tubishat[b], Malek Alzaqebah[c,d] and Nader Aljawarneh[e]

[a]*Department of Computer Networks and Cybersecurity, Jadara University, Jordan*
[b]*College of Technological Innovation, Zayed University, Jordan*
[c]*Department of Mathematics, College of Science, Imam Abdulrahman Bin Faisal University, Saudi Arabia*
[d]*Basic and Applied Scientific Research Center, Imam Abdulrahman Bin Faisal University, Saudi Arabia*
[e]*Department of Business Administration, Jadara University, Jordan*

C H R O N I C L E

Article history:
Received: January 6, 2024
Received in revised format: February 20, 2024
Accepted: May 2, 2024
Available online: May 2, 2024

Keywords:
Data recovery
Data backup
Data center
Network security
Reliability

A B S T R A C T

A large-scale cloud data center must have a low failure incidence rate and great service dependability and availability. However, due to several issues, such as hardware and software malfunctions that regularly cause task and job failure, large-scale cloud data centers still have high failure rates. These mistakes can have a substantial impact on cloud service dependability and need a large resource allocation to recover from failures. Therefore, it is important to have an efficient management of data recovery to protect organizations data from loss. This paper aims to study some factors that may improve the management of data recovery by using quantitative research design as a methodology. The results of hypothesis testing give strong evidence supporting the positive and significant correlations between the proposed hypotheses and the efficiency of data management recovery. This study finds that the presence of a data center in an organization demands the development of a solid plan for the most effective usage of a software program to handle data recovery.

1. Introduction

Data storage has been acknowledged as one of the primary issues with information technology during the past few decades. Distributed storage has replaced server-attached storage due to the advantages of network-based apps. Considerable work has been done in the field of distributed storage security since data security is the cornerstone of information security. But the field of study on cloud computing security is quite young (Voorsluys et al., 2011).

The expanded development and integration of TCP/IP-based computer technology—which includes quick microprocessors, massive memory, quick networks, and stable system architecture—is known as cloud computing. Without the established connectivity protocols and complex data center technology assembly procedures, cloud computing would not exist. IBM and Google announced their cloud computing partnership in October 2007. After it, the phrase "cloud computing" gained popularity. A workable conceptual framework for cloud services includes, among other things, Salesforce's CRM, Amazon Elastic Compute Cloud (EC2), and Google App Engine. Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) are the three main categories into which cloud computing services are split (Gong et al., 2010).

* Corresponding author.
E-mail address dr.maen@jadara.edu.jo (M. T. Alrashdan)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.5.004

The public began to have access to cloud storage services in the 2000s, enabling users to store files and programs on distant servers that belong to third parties. These remote servers essentially serve as "global storage," offering electronically stored data that can be accessed and stored at any time, from any location, using any kind of technology. When they sign up, users frequently have to accept non-negotiable terms of service. open an account with a company that offers cloud storage. The terms under which a user may and may not use the service, as well as the provider's usage and access to information, are typically specified in the service agreement, which also manages the user's relationship with the service (Johnson, 2017). Natural disasters like earthquakes may cause the data to be erased from the data centers, or man-made disasters like CSPs or customers deleting data without the customers' knowledge (Challagidad et al., 2017).

2. Data Centers Storage System's Tiering

Using tiering techniques, data is stored on SSD and HDD in the appropriate hot and cold categories. There are no duplicate copies of the data on HDD and the data stored in the SSD tier is permanent rather than transient like in SSD caching. Observing that the tiered SSD tiering approach is the subject of some research, Two categories can be distinguished amongst storage architectures: which are layered structures that are device-controlled and host-controlled. Within the host-managed tiered structure, the system controls the data motions host by utilizing the data from the host side, while the device Data movements are regulated via a controlled tiered framework. Examining the characteristics of the recently received data and the past information (Li, 2014).

Devarajan et al. (2020) create and put into use HCompress, a hierarchical data compression library that can enhance the performance of the application by combining data compression with multi-tiered storage in a harmonious way. We have created a new compression selection technique that makes it easier to match compression libraries to tiered storage in the best possible way. Our assessment indicates that HCompress, in comparison to other cutting-edge tiered storage solutions, can increase scientific applications' performance by a factor of seven. Through simulations, observations, and analysis, large volumes of data are read and written by modern scientific programs. These programs execute I/O operations for most of their runtime. Fast node-local and shared storage resources are two features of HPC storage systems that can lift applications out of this constraint. Additionally, a number of middleware libraries—like Hermes, for example—are suggested to transfer data transparently between these tiers. Another method for reducing data production and enhancing I/O performance is data reduction. When combined, these two technologies have advantages over one another. Different compression algorithms can be used based on the features of the various tiers to increase the effectiveness of data compression, and additional capacity can help the multi-tiered hierarchy function better (Devarajan et al., 2020). Data center storage services make judgments all the time on things like block allocation, prefetching, and cache admission. Heuristics based on statistical features, such as temporal locality or common file sizes, usually guide these judgments. Application-level data such as can help make decisions of higher quality. The database operation to which a request is related. Although application cues (e.g., explicit prefetches) can be used to exploit such capabilities, this method is labor-intensive and should only be used for the most optimized workloads (Zhou & Maas, 2021).

3. Data Reliability in Data Centers

For large Internet companies such as Facebook, dependable network infrastructure is essential, both within and between data centers. Meza et al. (2018) conducted a study that uses operational data from Facebook's production infrastructure to analyze data center network dependability over time. Over seven years and eighteen months, thousands of intra-data center network events as well as inter-data center accidents are included in the analysis. The analysis demonstrates how various network architectures and device types impact network dependability. As software systems get more complicated and distributed, network infrastructure can become a significant limiting factor. Capable of managing large-scale distributed software systems with dependability. The study aims to increase the stability of large-scale data center networks and systems. A study that used trees as a basis for hierarchical modeling offered a way to assess the reliability and availability of data centers (Nguyen et al., 2019). Three layers make up the hierarchical model: 1) a fault tree, which represents the subsystems' architecture; 2) reliability graphs at the top, which represent the topology of the system network; and 3) stochastic reward nets, which accurately represent the relationships and behaviors of the subsystem components. Two popular data center networks' three-tier and fat-tree topologies are carefully modeled and investigated. The research examined many case studies to determine how networking and administration affected cloud computing facilities. Moreover, it does several thorough analyses on the dependability and accessibility of the system models. The researcher's conclusions show that data centers' availability and dependability may be increased by optimizing node placement inside networks by employing appropriate networking.

Couto et al. (2016) looked at the benefits of several data center layouts while taking survivability and dependability into account. Three distinct data center architectures—fat-tree, BCube, and DCell—were examined by the researchers. They also made a comparison between these topologies and the conventional three-layer data center topology. The study is not restricted to any particular hardware, traffic patterns, or network protocols in order to guarantee universality. For every topology, they provide closed-form computations of the Mean Time To Failure. They can determine the best topology for every failure situation thanks to the results. According to the investigation, DCell is the most robust to switch failures, while BCube is more resilient to link failures than other topologies. They proved that for both kinds of failures, all other topologies perform better

than the three-layer structure. They examined how BCube and DCell's dependability is affected by the quantity of network interfaces on each server.

4. Data Loss

A data center network is an essential component of many current information technology applications. Data centers hold files containing vital information, but their lifespan is limited. Natural catastrophes and man-made damage to a data center will result in the loss of all information saved in the files. Therefore, duplicating essential files in other data centers is required to enhance the lifetime of these files in a data center network (Ma & Fan, 2021). Excellent service reliability, availability, and a low failure incidence probability are essential for a large-scale cloud data center. Even yet, there are still many reasons why large, contemporary cloud data centers fail, such as broken hardware and software, which often results in task and job failures. These kinds of errors can drastically reduce the dependability of cloud services and are very expensive to repair. To avoid unanticipated waste, it is therefore essential to correctly predict activities or work failures in advance (Gao et al., 2020). Liu and Kuhn (2010) divided the loss of data into two types; Leakage occurs when sensitive data is no longer within the organization's control, resulting in a loss of confidentiality. Hacked customer databases are a major cause of data loss, with the most typical outcome being identity theft. Disappearance or damage to a proper data copy, resulting in a loss of integrity or availability for the company.

5. Data recovery management

Data integrity and recovery management are particularly crucial in cloud computing since data is everywhere. Backup recovery and security provide significant challenges. It is necessary to create an effective and dependable data storage system (Gokulakrishnan & Gnanasekar, 2020a). A study focuses on and proposes a novel approach for data recovery and data management to provide high-level scalability and high order dependability in cloud-based systems. The study offers a methodology for segmenting data and creating tokens for data split-up that include the cloud address or cloud storage locations via the tailing approach. Thus, the missing section of any problematic node is easily detected within a narrow range of bounds, and the data backup from the nearby nodes. Another proposal proposes a new technique for backup and recovery management, resulting in a more reliable and scalable cloud infrastructure (Gokulakrishnan & Gnanasekar, 2020b). The methodology presents a way for data segmentation. Tokens are generated for data splitting by attaching cloud storage locations or addresses. A problematic cloud server without a location segment is identified within a narrower broadcasting limit and backup is retrieved from secure surrounding sources.

6. Data Backup Criteria

Backup copy attributes are organized into a group for administrative purposes. Multiple backup copy groups can be supported in the same management class, each with distinct backup criteria. For example, one group may be used for daily backups, while another may specify quarterly backups (Kaczmarski & Pease, 2003). The criteria for selecting a backup service are that the supplier must have redundant sources of vital supplies to ensure optimal dependability. For an online service to be trustworthy, it must have two key characteristics: competent and redundant sources of supply. This requires the employment of high-quality equipment, media, and supply services, as well as varying levels of redundancy. Second, it must have strong environmental and operational controls in place (Garnett, 2008).

7. Data Backup Schedule

Such backups could generate a lot of traffic, which could put a lot of strain on the communications network underneath. Van de Ven et al. (2014) conducted a study that looked at the trade-off between regular backups and a decline in network peak demand. The study addresses the problem of shifting backup traffic, while staying within user connectivity constraints, from peak to off-peak hours. A distributed protocol, which is based on a set of probabilities that indicate the likelihood of a user starting a backup within a specific hour, is used to organize backups. A backup plan is essential for guaranteeing the right degree of data security and reducing system downtime in a system with multiple data sources. A framework for automatically creating an MDP instance from system specifications and data protection criteria is proposed in a study that examines the use of Markov Decision Process (MDP) to guide the scheduling of data backup activities (Xia et al., 2014). In order to reduce system downtime while satisfying requirements, the solution optimizes the timetable. Effective data security is extremely challenging due to the daily data generated by large company enterprises and the backup system's amazing capacity increase. Backup administrators have typically been left to handle this challenge since they oversee creating complex static backup plans that guarantee backup and recovery objectives are met. Nevertheless, backup solutions fall short of their goals because static backup schedules can't adjust to the changing backup environment. Because static backup schedules are ineffective, developing a dynamic backup system scheduling technique is essential (Qin et al., 2018).

8. Data and system recovery process

Research found that by implementing a data and system recovery procedure, additional attributes whose data has not been lost can be used to retrieve missing data using the trained model. Experiments demonstrate that an approach with numerous

characteristics is efficient and can enhance data recovery accuracy when compared to other techniques (Cheng et al., 2021). Machine learning is an effective auto-learning approach for obtaining the underlying rules automatically. The paper presented an intelligent recovery technique for massive data on the Internet of Things using Multi-Attribute Assistance and Extremely Randomized Trees (MAET).

Hypothesis

Based on other researchers' studies and previous related works, there are some factors that play a role in having an efficient data recovery management as shown in figure 1. The paper proposes three hypotheses:

H_1: *There is a positive and significant role in data backup criteria in improving the efficiency of data backup management in data centers.*

H_2: *There is a positive and significant role in data backup schedules in improving the efficiency of data backup management in data centers.*

H_3: *There is a positive and significant role in the data and system recovery process in improving the efficiency of data backup management in data centers.*

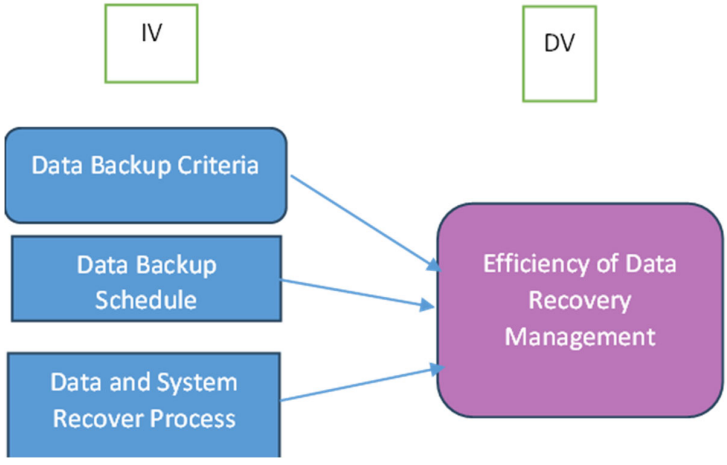

Fig. 1. Conceptual Framework

9. Methodology

This study employs a quantitative research design to assess the influence of data backup criteria, scheduling practices, and data recovery processes on the efficiency of data recovery management within organizations. Drawing from a sample of 363 participants primarily sourced from information technology departments in large organizations, the study utilizes a structured questionnaire as the primary data collection instrument. This questionnaire was developed based on the literature on data backup, scheduling, and recovery processes. Upon distribution, participants are provided with clear instructions and a designated timeframe to complete the questionnaire. However, inferential statistics, such as correlation and regression analyses, are utilized to explore the effect of data backup criteria, scheduling, and recovery processes on the efficiency of data recovery management.

10. Results

10.1 Validity

This study aims to evaluate how applying data backup criteria, scheduling and data recovery process in data centers affects the data recovery management recovery efficiency in the organizations. However Smart PLs4 was used to investigate the efficient data recovery management factors in data centers. The first to examine outer loading values of each indicator as shown in Fig. 2 below. For several research variables, the outer loading values of each indicator are seen to exceed 0.7, indicating strong correlations between the indicators and their corresponding latent constructs. However, it is interesting that several indicators are still indicating outer loading levels below 0.7. The prerequisites for convergent validity are said to be satisfied by outer loading values between 0.5 and 0.6, according to Mulyono et al. (2020). However, the outer loading values of all the variable indicators in this study are more than 0.5, indicating that they may be used in research and that further investigation is necessary to determine their relationship to their latent components.

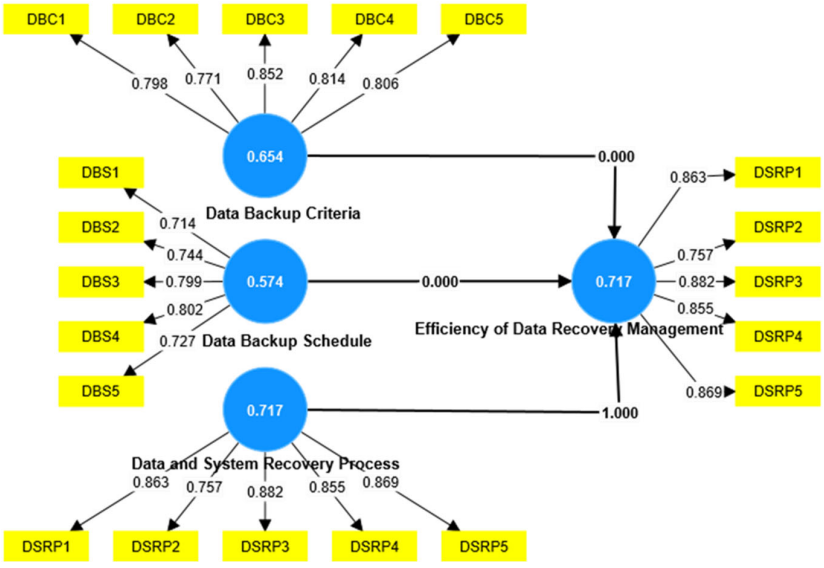

Fig. 2. Examination of Validity

10.2 AVE and Reliability

Table 1 employs three commonly used measures to assess variable dependability: Cronbach's Alpha, composite reliability, and AVE. These metrics evaluate the convergent validity, internal consistency, and overall dependability of the constructs in SEM-PLS analysis. The reliability and trustworthiness of the constructs in Table 1 for the SEM-PLS investigation are shown by their high Cronbach's Alpha coefficients, composite reliability values better than 0.70, and extracted average variance values. These measures ensure the desired conceptions' correctness, consistency, and reliability.

Table 1
Reliability and AVE

	Cronbach's alpha	Composite reliability (rho_a)	Composite reliability (rho_c)	Average variance extracted (AVE)
Data Backup Criteria	0.867	0.868	0.904	0.654
Data Backup Schedule	0.815	0.820	0.871	0.574
Data and System Recovery Process	0.900	0.904	0.927	0.717
Efficiency of Data Recovery Management	0.900	0.904	0.927	0.717

The results of the average variance extracted (AVE) assessment and reliability testing are shown in Table (1) for the four research variables: data recovery management efficiency, data and system recovery process, data and backup criteria, and data backup schedule. The internal consistency of the variable's items is measured by Cronbach's alpha values, which vary from 0.815 to 0.900 and denote good to outstanding reliability. The measurement model's composite dependability, as determined by both rho_a and rho_c, is strong, with values ranging from 0.868 to 0.904, above the suggested criterion of 0.7 for all variables. Additionally, the acceptable convergent validity is shown by the AVE values, which exceed the minimal criterion of 0.5 and indicate the fraction of variance captured by the latent variables relative to measurement error. These values range from 0.574 to 0.717. All of these results point to a strong internal consistency, reliability, and convergent validity of the measurement model, demonstrating the validity and reliability of the variables in measuring the various constructs.

10.3 Hypothesis Testing

In statistical analysis, a variety of indicators are used to evaluate hypotheses. Examples of these indicators include original value sample estimates (O), p-values (P), and t-statistics. They provide their perspectives on the nature and relevance of the link between the factors. The numerical estimate obtained from the sample data is represented by the original value sample estimate (O). It is considered positive if the correlation between the variables is around +1, and negative if it is around -1. T-statistics (T) are also used to determine the significance of a connection. At a 95% confidence level, a t-statistics value greater than 1.96 indicates a significant relationship between the variables. P-values (Ps) are an important statistic for determining significance. a p-value less than the desired significance level. If the p-value is less than the chosen significance threshold, which is commonly less than 0.05, then the variables have a statistically significant relationship. Figure 3 depicts the outcomes of the hypothesis testing.

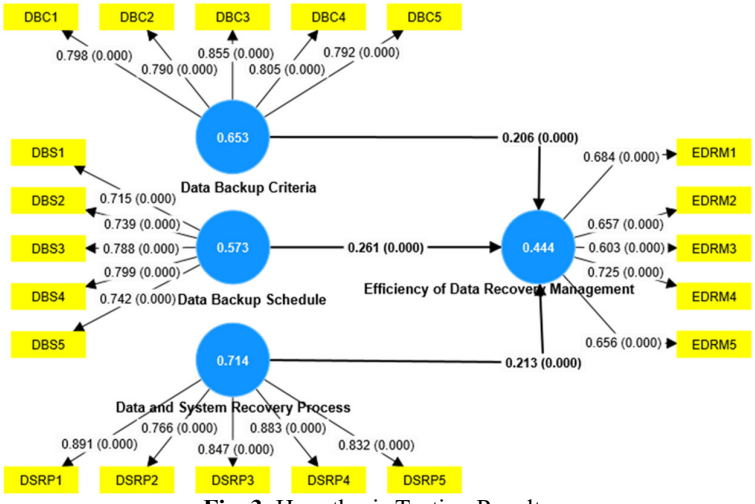

Fig. 3. Hypothesis Testing Results

Fig. 3 depicts the process of testing hypotheses, which includes evaluating the research hypotheses. The previously mentioned route coefficients give critical information for this testing. Table 3 shows the findings of hypothesis testing for direct impacts. This table offers a complete summary of the variables' relationships and enables for hypothesis testing. By evaluating the figures in Table 3, researchers may decide whether to accept or reject their study hypothesis based on the observed direct effects between variables. Tables are a helpful tool for analyzing the results of hypothesis testing.

Table 2
Results of Hypothesis Testing

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values	
Data Backup Criteria → Efficiency of Data Recovery Management	0.206	0.209	0.052	3.931	0.000	Supported
Data Backup Schedule → Efficiency of Data Recovery Management	0.261	0.264	0.060	4.323	0.000	Supported
Data and System Recovery Process → Efficiency of Data Recovery Management	0.213	0.214	0.058	3.639	0.000	Supported

The results of the hypothesis test are displayed in the above table, which also provides information on the effects of several parameters on the effectiveness of data recovery. But for H1, which suggests that Data Backup Criteria have a positive and important influence in improving the effectiveness of data recovery management, the findings show a strong correlation with a coefficient of 0.206, which is much higher than the threshold for statistical significance. This implies that companies which use good data backup standards should see an increase in the effectiveness of their data recovery procedures.

Similarly, a coefficient of 0.261 strongly supports H2, which postulates a positive and significant impact of Data Backup Schedule on Data Recovery Management Efficiency. This suggests that creating and following structured backup schedules can significantly improve recovery efficiency in data centers. Additionally, with a value of 0.213, H3, which suggests that the Data and System Recovery Process has a positive and important effect in enhancing the efficiency of data recovery management, obtains strong validation. This emphasizes how crucial proper data and system recovery procedures are to enabling timely and successful data restoration in the case of interruptions or breakdowns. Taken together, these results highlight how crucial it is to use strategic methods for scheduling, data backup, and recovery procedures in order to maximize the effectiveness of data recovery management in corporate data centers.

11. Discussion

The results of the hypothesis testing offer strong evidence in favor of the significant and positive associations suggested by hypotheses H1, H2, and H3. First off, the data backup criteria and the effectiveness of data recovery management in data centers have a strong and positive link, which is supported by the study and hypothesis H1. This emphasizes how crucial it is to have clear criteria for data backup operations in order to enable more efficient recovery processes. Likewise, hypothesis H2 is confirmed, showing a strong and positive correlation between the effectiveness of data recovery management and the scheduling of data backups. This demonstrates how important it is to have regular backup plans in place if you want to improve data centers' overall capacity for recovery. Furthermore, data recovery management efficiency and data and system recovery procedures have a positive and substantial association, as supported by the empirical evidence for hypothesis H3. This

emphasizes how crucial reliable recovery procedures and systems are to effective and timely data restoration. As a result, this paper's entire hypothesis is accepted.

12. Conclusion

This research focused on the impact of assigning data recovery criteria, having data backup schedule, and implementing data backup process on the efficiency of the data recovery management in organizations data centers. The results showed the importance of enhancing the data recovery management by having a clear classification of the data and the processes and a priority in data recovery. Also, the data backup schedule plays an important role in monitoring the transferred and stored data. Finally, it is recommended to train employees in organizations in the optimal use of the processes of data backup software applications.

References

Challagidad, P. S., Dalawai, A. S., & Birje, M. N. (2017). Efficient and reliable data recovery technique in cloud computing. *Internet of Things and Cloud Computing*, *5*(1), 13-18.

Cheng, H., Shi, Y., Wu, L., Guo, Y., & Xiong, N. (2021). An intelligent scheme for big data recovery in Internet of Things based on multi-attribute assistance and extremely randomized trees. *Information Sciences*, *557*, 66-83.

Couto, R. D. S., Secci, S., Campista, M. E. M., & Costa, L. H. M. K. (2016). Reliability and survivability analysis of data center network topologies. *Journal of Network and Systems Management*, *24*, 346-392.

Devarajan, H., Kougkas, A., Logan, L., & Sun, X. H. (2020, May). Hcompress: Hierarchical data compression for multi-tiered storage environments. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (pp. 557-566). IEEE.

Gao, J., Wang, H., & Shen, H. (2020). Task failure prediction in cloud data centers using deep learning. *IEEE transactions on services computing*, *15*(3), 1411-1422.

Garnett, B. S. (2008). Better safe than sorry: criteria for choosing a data backup service. *Orthodontic Products*, *15*(3), 106-109.

Gokulakrishnan, S., & Gnanasekar, J. M. (2020a). Data integrity and recovery management under peer to peer convoluted fault recognition cloud systems. *Journal of Computational and Theoretical Nanoscience*, *17*(5), 2147-2150.

Gokulakrishnan, S., & Gnanasekar, J. M. (2020b). Data integrity and recovery management in cloud systems. In *2020 Fourth International Conference on Inventive Systems and Control (ICISC)* (pp. 645-648). IEEE.

Gong, C., Liu, J., Zhang, Q., Chen, H., & Gong, Z. (2010, September). The characteristics of cloud computing. In *2010 39th International Conference on Parallel Processing Workshops* (pp. 275-279). IEEE.

Johnson, E. (2017). Lost in the cloud: Cloud storage, privacy, and suggestions for protecting users' data. *Stanford Law Review*, *69*, 867.

Kaczmarski, M., Jiang, T., & Pease, D. A. (2003). Beyond backup toward storage management. *IBM Systems Journal*, *42*(2), 322-337.

Li, Z. (2014). *GreenDM: A versatile tiering hybrid drive for the trade-off evaluation of performance, energy, and endurance* (Doctoral dissertation, State University of New York at Stony Brook).

Liu, S., & Kuhn, R. (2010). Data loss prevention. *IT professional*, *12*(2), 10-13.

Ma, F. Q., & Fan, R. N. (2021). Markov Processes in Data Center Networks. *IEEE Access*, *9*, 42216-42225.

Meza, J., Xu, T., Veeraraghavan, K., & Mutlu, O. (2018, October). A large scale study of data center network reliability. In *Proceedings of the Internet Measurement Conference 2018* (pp. 393-407).

Mulyono, H., Hadian, A., Purba, N., & Pramono, R. (2020). Effect of service quality toward student satisfaction and loyalty in higher education. *The Journal of Asian Finance, Economics and Business (JAFEB), 7*(10), 929-938.

Nguyen, T. A., Min, D., Choi, E., & Tran, T. D. (2019). Reliability and availability evaluation for cloud data center networks using hierarchical models. *IEEE Access*, *7*, 9273-9313.

Qin, Y., Hoffmann, B., & Lilja, D. J. (2018, November). Hyperprotect: Enhancing the performance of a dynamic backup system using intelligent scheduling. In *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)* (pp. 1-8). IEEE.

van de Ven, P. M., Zhang, B., & Schörgendorfer, A. (2014, April). Distributed backup scheduling: Modeling and optimization. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications* (pp. 1644-1652). IEEE.

Voorsluys, W., Broberg, J., & Buyya, R. (2011). Introduction to cloud computing. *Cloud computing: Principles and paradigms*, 1-41.

Xia, R., Machida, F., & Trivedi, K. (2014, June). A Markov decision process approach for optimal data backup scheduling. In *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks* (pp. 660-665). IEEE.

Zhou, G., & Maas, M. (2021). Learning on distributed traces for data center storage systems. *Proceedings of Machine Learning and Systems*, *3*, 350-364.