AN INNOVATION DEVELOPMENT OF RELIABLE REDUNDANCY OF DATA BACKUP IN BIG DATA SERVERS USING RAID ARRAYS

Nirmal Adhikari[1], G. Ramesh[2] and V. Aravindarajan[3]
[1]*School of Built Environment Engineering and Computing, Leeds Beckett University, United Kingdom*
[2,3]*Department of Information Technology, KLN College of Engineering, India*

Abstract

Backup is an important process in the life of any IT infrastructure. It is a rescue parachute in case of an unexpected disaster. At the same time, the backup is used to create a kind of historical archive of the company's business activities over a certain period of its life. Working without insulation is like living outdoors - the weather can turn bad at any time, and there's nowhere to hide. But don't lose important data and spend huge sums of money on it. Backup can restore both complete backup and individual files and folders. The backup can create an exact copy of the HDD, including the operating system, boot logs and other information. It can copy and restore an entire hard drive and individual partitions. In this paper, a reliable redundancy of data backup system was proposed in big data servers using RAID arrays. Using the Backup Disaster Recovery utility you can create a custom bootable recovery disk on which you can boot the operating system in case of failure and restore the system, settings and all data. It can copy ODBC compatible databases and also has special plug-in to accurately copy databases like DB2, Oracle, MS SQL, MySQL etc. Another hobby besides writing articles is promoting free software. To organize the backup system more effectively, you need to develop a real strategy for storing and retrieving information.

Keywords:

Backup, Disaster, Restores, HDD, Booting, Hard Drive, RAID Arrays

1. INTRODUCTION

. Backups cannot be avoided even if you have a clustered architecture. A failover cluster actually maintains the functionality of the services entrusted to it if one of the servers fails. In case of the above problems like virus attack or data corruption due to bad “human factor”, no cluster saves [1]. The only thing that can serve as an inferior backup substitute for disaster recovery is a mirrored backup server with continuous data replication from the primary server to the backup. In this case, if the main server fails, its tasks will be taken over by the backup, and you will not need to transfer the data [2]. But such a system is very expensive and takes time to organize. Don't forget about the need for consistent replication. The importance of distinguishing backups from redundant backup systems should also be evaluated when developing a data backup plan, whether it concerns company or home computers [3]. Ask yourself why you are making duplicates. If we talk about backup, we mean protecting data in case of accidental (intentional) action [4]. Redundant redundancy makes it possible to save data, including backups, in case of equipment failure. There are many low-cost devices on the market today that provide reliable redundancy using RAID arrays or the cloud [5]. It is recommended to use both types of data backups simultaneously. In order to avoid unnecessary material costs when organizing a backup and, if possible, not to go beyond the backup window, several backup technologies have been developed that are used depending on the specific situation [6]. Full backup is the main and basic method of creating backups in which a selected sequence of data is completely copied. It is the most complete and reliable backup method, although it is also the most expensive one [7]. If it is necessary to store multiple copies of data, the total stored volume will increase in proportion to their number. To prevent such waste, compression methods are used, as well as a combination of this method with other types of backups: incremental or differential [8]. And, of course, a full backup is essential when you need to prepare a backup for a quick restore of a new computer [9].

Unlike a full backup, in this case, not all data (files, partitions, etc.) is copied, but only what has changed since the last backup. Various methods can be used to determine when a copy was made, for example, on computers running Windows operating systems, the relative file attribute (archive bit) is used, which is set when the file is modified and restored by backup [10]. On other systems, the date the file was modified may be used. Clearly, a plan using this type of backup is incomplete if you don't perform a full backup every now and then [11]. When doing a full system restore, you must restore from the last copy created by a full backup, and then roll over the data from the incremental copies one by one in the order of their creation. In the case of creating archival copies, it is important to reduce the amount consumed on storage devices. This will also allow you to reduce the processing time of backup jobs, which is especially important when you have to work on a 24x7 busy schedule or pump out large amounts of information. One thing to keep in mind is incremental replication. Step by step recovery also provides deleted files needed for recovery period [12].

It is clear that such a solution would be cost-effective only in the case of critical services with high requirements for fault tolerance and minimum recovery time. As a rule, such programs are used in very large companies with a lot of products and cash flow [13]. This plan is an incomplete replacement for backup, because anyway, if the data is damaged by a computer virus, ineffective user actions or incorrect operation of the application, the data and software on both servers can be affected. And, of course, no redundant backup system solves the problem of maintaining a data archive over a period of time [14]. The way to solve these problems described above suggests itself: to postpone the start of the process of creating copies for an idle period, the mutual influence of the backup and other working systems will be less. This period is called the “safety window”.

2. RELATED WORKS

The LAN-free backup method does not solve the backup window problem. Also, this method creates an additional load on the client servers, entrusting them with additional functions of the

servers to copy the data being backed up. Some applications allow online backups, which are implemented in many transactional applications and specialized backup software options such as Open File Copiers [1]. However, the use of such technologies does not reduce the load on production servers, which can increase the time to solve basic tasks above the acceptable threshold with large amounts of data (terabytes and so on) [2]. For disaster recovery and archival storage, it is necessary to determine the list of copied resources, the time of execution of tasks and where, how and for how long the backups will be stored [3].

With small amounts of data and a more complex IT infrastructure, you can try to combine these two tasks together, for example, make a daily full copy of all disk partitions and databases. But it is better to distinguish between the two ends and choose the right way for each of them. Accordingly, even if there are universal solutions such as the same Acronis True Image package or the ntbackup program, a separate tool is used for each task [5]. When defining the goals and objectives of backup, as well as solutions for implementation, it is clear that it is necessary to proceed from the needs of the business [6]. There are various strategies to implement disaster recovery. In some cases, bare metal requires a direct computer recovery. For example, this can be done using Acronis True Image combined with Universal Restore. In this case, the server configuration can be returned to service in a very short time [8]. For example, it is quite possible to raise a partition with a 20 GB operating system from a backup in eight minutes (if the backup copy is on a 1 GB / s network).

In another case, it is more appropriate to “return” settings to a newly installed system, such as copying the configuration files from the / etc folder on UNIX-like systems (in Windows, this is roughly equivalent to copying and restoring the system. level) [10]. Of course, with this approach, the server will not be activated before the operating system is installed and the necessary settings will be reset, which will take more time. But in any case, what constitutes disaster recovery evolves from the needs and resource constraints of the business [12].

3. PROPOSED MODEL

All data backup systems can be divided into three types according to the replication method: file-by-file, volume, or application-level replication. The block backup system (English image-level or volume-level backup) works directly with the media, ignores the file structure and completely protects all contents - operating system, working data, settings, etc. The advantage of this type of backup is its high speed. However, when performing regular copy operations, it is necessary to pause the operation of the applications to keep the copy consistent. When performing backup operations in file-level or file-based backup, the file system is used. In this case, recovering some specific files is a relatively straightforward task. In general, backup operations take a long time, additional loading of the operating system occurs, and there is a problem with accessing open files. The RAID arrays (RAID 0) management was shown in the Fig.1.

Backups can also be performed on an application-level backup. Copy and restore operations are performed using a specially provided API (Application Programming Interface) in the redundant application. A backup is a collection of files and other objects, defined by an application that represents the state of the application at a particular point in time. The RAID arrays (RAID 5) management was shown in the Fig.2.

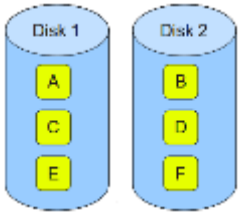

Fig.1. RAID arrays (Blocks striped, No mirror and No parity)

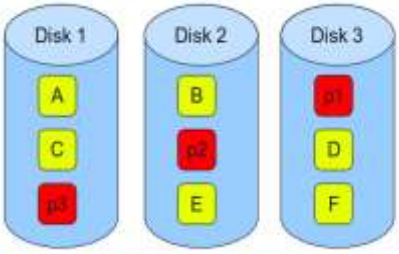

Fig.2. RAID 5 arrays (Blocks striped, distributed parity)

With this method of backup, there may be compatibility issues between different versions of applications and backup systems that implement the corresponding interface. In the process of organizing a backup, two main tasks are set:

- Restoring the infrastructure in case of failures (disaster recovery) and maintaining a data archive to provide access to information from the past.
- A classic example of a disaster recovery backup is an image of a server's system partition created by Acronis True Image.

An example of an archive is a monthly download of databases from 1C, recorded on cassettes and then stored in a specially designated place. Composition of information to be copied an archived copy usually contains only user and business data for a specific period of time. A disaster recovery copy contains, in addition to this data, computer images or copies of the operating system and application software systems and other information required for recovery. Sometimes these tasks can be combined. The big data disk array in RAID 5 was shown in Fig.3

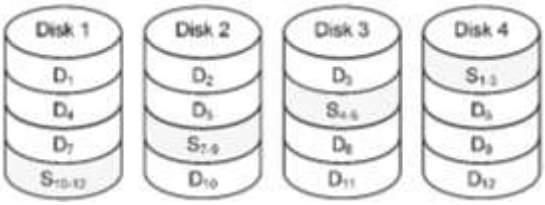

Fig.3. Big data disk array in RAID 5

An annual collection of monthly full “snapshots” of a file server and changes made during the week. A real image is suitable as a tool for creating such a backup. The most important thing is to clearly understand why the reservation is being made. Let me give you an example: A critical SQL Server crashed due to a disk array failure. We have the correct hardware in stock, so the only solution is software and data recovery. During the entire service life of the server, only the databases are regularly backed up, not taking into account the need to restore the server with all settings,

including the software of the DBMS. Shadow copy provides creation of snapshots of the file system, and changes to the original do not affect them in any way. Using this function, it is possible to create multiple blind copies of a file over a period of time, as well as on-the-fly backup copies of files that are open for writing. Volume Copy Shadow Service is responsible for the task of shadow copy.

System status provided the System Level Copy creates backups of critical components of Windows operating systems. It allows you to restore a previously installed system after destruction. When copying the system state, the registry, boot and other files important to the system are saved, including restoring the active directory, certificate service database, COM + class registry database, SYSVOL directory. On UNIX-based operating systems, the indirect analog of system state copying stores the contents of the / etc, /usr / local / etc directories, and other files necessary to restore the system state. The hardware redundancy systems are the introduction of some redundancy into the hardware so that it continues to function if a component suddenly fails. A good example in this case is a RAID (redundant array of independent disks). In the event of a disk failure, information loss can be avoided and a safe replacement can be made, data can be protected by a specific arrangement of disk arrays.

4. RESULTS AND DISCUSSION

The proposed reliable redundancy of data backup (RRDB) was compared with the existing High reliability provision (HRP), classified enhancement model (CEM), Distributed File Systems (DFS) and Antivirus Security Systems (ASS)

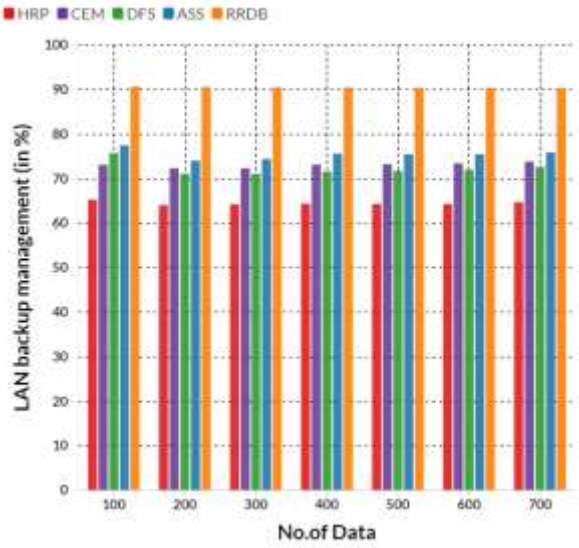

Fig.4. Comparison of LAN backup management

LAN backup management: the advent of Storage Area Networks (SANs), a dedicated backup network was used to reduce backup traffic on the core network, as well as a tiered system with multiple replica servers. Dedicating a copy server and locating it on the network “closer” to production servers that process large volumes of information allows you to localize backup traffic between the copy server and production servers and reduce the load on the shared LAN. The comparison of LAN backup management was shown in the Fig.4.

LAN-free backup management: With the advent of the SAN, backup traffic could no longer be transferred over the LAN, but from servers directly to storage devices (usually tape libraries) attached to the SAN. This method is called “LAN-free backup”. When using this method, the server-client acts as a server to copy the backed-up data to storage devices accessible via the SAN, among other tasks. In this case, the backup management server delegates the task of executing the backup schedule by providing control actions via LAN (via TCP/IP protocol) and controlling the execution of tasks by copy servers. Thus, the task of reducing the traffic of backup data on the LAN is solved. The comparison of LAN-free backup management was shown in the Fig.5.

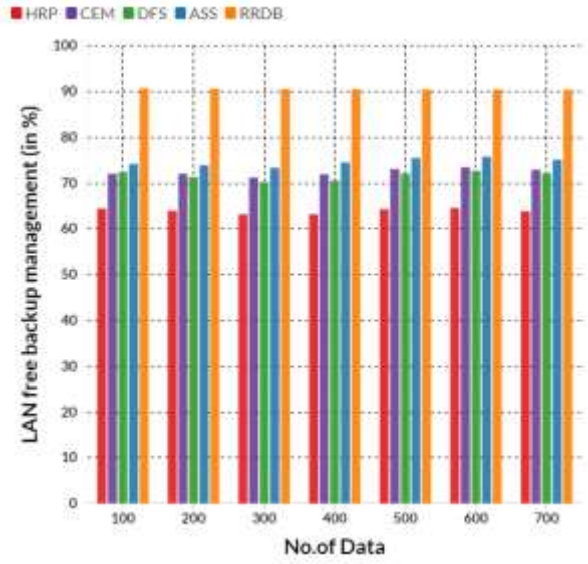

Fig.5. Comparison of LAN-free backup management

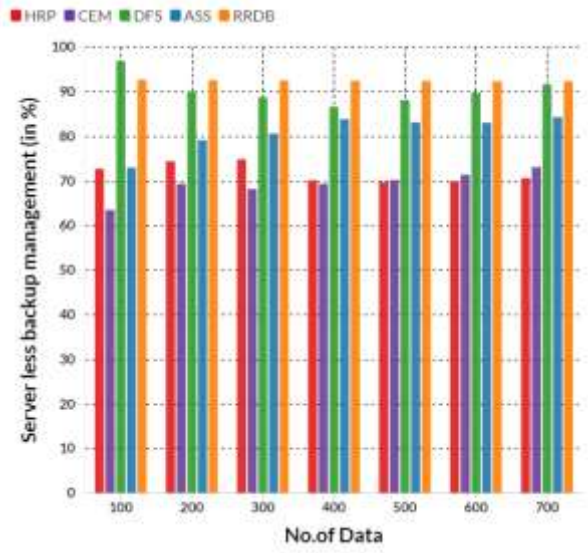

Fig.6. Comparison of Server less backup management

Server less backup management: An ideal backup plan is when the client server data is backed up to a storage device over a SAN by some third-party device (called a “data mover”), without using the client server's computing resources and without interrupting its operation. This backup method is called “Server less Backup”. The role of “data mover” can be performed by a

dedicated server connected to the same disk array as the production server, or by a special device - a router. The comparison of Server less backup management was shown in the Fig.6.

Continuous Data protection: Continuous Data Protection (CDP), as defined by SNIA, is a technique for continuously tracking data changes, storing them in a repository independent of the original data, and allowing retrieval at any time in the past. CDP systems can be implemented at the block, file, or application level and can provide fine granularity of object retrieval down to a single write operation at any time. The comparison of Continuous Data protection was shown in the Fig.7

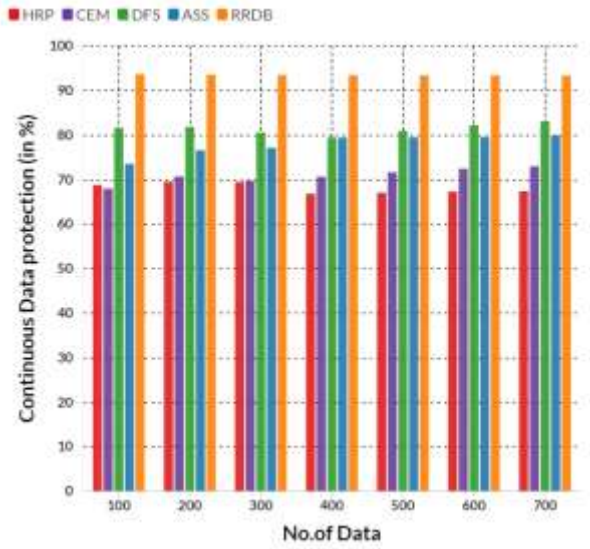

Fig.7. Comparison of Continuous Data protection

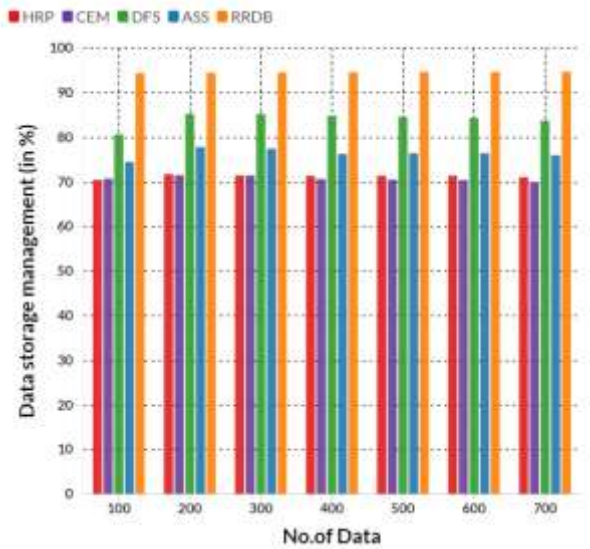

Fig.8. Comparison of Data storage management

Rapid Data Access management: Speed of access to long-term archive is not critical in most cases. Usually the need to "raise data for a period" arises during document reconciliation, rollback, etc., i.e. not in emergency mode. Another thing is disaster recovery, when the necessary data and performance of services must be restored quickly. In this case, the speed of access to the backup is very important. The comparison of Rapid Data Access management was shown in the Fig.9

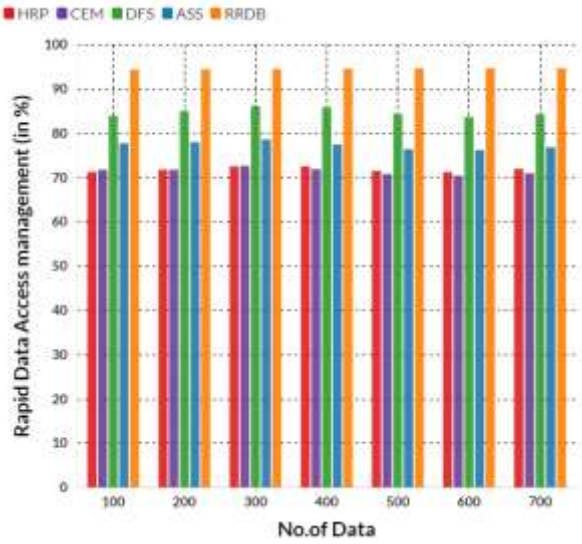

Fig.9. Comparison of Rapid Data Access management

Data storage management: For archival copies, it is much longer. In some cases, this is regulated not only by business requirements but also by law. Disaster recovery copies have a relatively small amount. Typically, one or two (with increased reliability requirements) backups for disaster recovery are created at maximum intervals of one or two days, after which they are overwritten by new ones. In particularly critical cases, a disaster recovery backup can be updated more frequently, for example, every few hours. The comparison of Data storage management was shown in the Fig.8.

Performing a backup causes unnecessary server overload. This is especially true for disk subsystems and network connections. In some cases, when the replication process has enough priority, it may cause some services to become unavailable. In addition, copying data at the time of making changes is associated with significant difficulties. Of course, there are technical means to avoid problems while maintaining data integrity in this case, but if possible, it is better to avoid such copying on the fly.

5. CONCLUSION

If a hard drive fails, the same RAID array protects data from destruction. But it does not save from data damage by computer virus or ineffective user actions. It does not save the RAID even if the file system crashes as a result of an unauthorized reboot. If a disk fails, data is restored due to a redundancy mechanism, specifically, using stored checksums. At the same time, there is a significant decrease in performance, the server freezes, control is practically lost. The system administrator simply restarts the server with a cold restart. As a result of this direct overload, file system errors occur. The best thing to expect in this case is a long period of disk verification to restore the integrity of the file system. In the worst case, you have to say goodbye to the file system and where, how and when you can restore data and server performance.

REFERENCES

[1] R. Nachiappan and K.M. Matawie, "Cloud Storage Reliability for Big Data Applications: A State of the Art Survey", *Journal of Network and Computer Applications*, Vol. 97, pp. 35-47, 2017.

[2] C. Liu, Y. Gu and D. Wang, "R-ADMAD: High Reliability Provision for Large-Scale De-Duplication Archival Storage Systems", *Proceedings of International Conference on Supercomputing*, pp. 370-379, 2009.

[3] H. Huang and S. Zhou, "Classified Enhancement Model for Big Data Storage Reliability based on Boolean Satisfiability Problem", *Cluster Computing*, Vol. 23, No. 2, pp. 483-492, 2020.

[4] J.Y. Lee and S.Y. Noh, "Performance Evaluations of Distributed File Systems for Scientific Big Data in FUSE Environment", *Electronics*, Vol. 10, No. 12, pp. 1471-1478, 2021.

[5] Z. Qiao and B. Settlemyer, "Developing Cost-Effective Data Rescue Schemes to Tackle Disk Failures in Data Centers", *Proceedings of International Conference on Big Data*, pp. 194-208, 2018.

[6] G. Ramesh, J. Logeshwaran and V. Aravindarajan, "The Performance Evolution of Antivirus Security Systems in Ultra dense Cloud Server Using Intelligent Deep Learning", *BOHR International Journal of Computational Intelligence and Communication Network*, Vol. 1, No. 1, pp. 15-19, 2022.

[7] M. Grawinkel, L. Nagel and A. Brinkmann, "Lonestar Raid: Massive Array of Offline Disks for Archival Systems", *ACM Transactions on Storage*, Vol. 12, No. 1, pp. 1-29, 2016.

[8] C. Liu and Y.W. Leung, "R-Memcached: A Reliable In-Memory Cache System for Big Key-Value Stores", *Proceedings of International Conference on Big Data Computing and Communications*, pp. 243-256, 2015.

[9] J. Logeshwaran, "AICSA - An Artificial Intelligence Cyber Security Algorithm for Cooperative P2P File Sharing in Social Networks", *ICTACT Journal on Data Science and Machine Learning*, Vol. 3, No. 1, pp. 251–253, 2021.

[10] X.C. Chai, W.Q. Wang and Y. Li, "Research on a Distributed Processing Model based on Kafka for Large-Scale Seismic Waveform Data", *IEEE Access*, Vol. 8, pp. 39971-39981, 2020.

[11] A. Thomasian and M. Blaum, "Higher Reliability Redundant Disk Arrays: Organization, Operation, and Coding", *ACM Transactions on Storage*, Vol. 5, No. 3, pp. 1-59, 2009.

[12] Z. Qiao and B. Settlemyer, "Enabling Proactive Data Protection in ZFS To Build Reliable Big Data Storage Systems", *Proceedings of International Conference on Big Data*, pp. 1-13, 2021.

[13] J. Logeshwaran, "The Control and Communication Management for Ultra Dense Cloud System using Fast Fourier Algorithm", *ICTACT Journal on Data Science and Machine Learning*, Vol. 3, No. 2, pp. 281-284, 2022.

[14] C.H. Wu and P. Hsu, "Cost-Effective and Reliable Cloud Storage for Big Data", *Proceedings of the ASE Big Data and Social Informatics*, pp. 1-6, 2015.