PAPER NAME

M22CS005_Thesis-Devansh.pdf

AUTHOR

Devansh Kaushik

WORD COUNT

3222 Words

CHARACTER COUNT

18146 Characters

PAGE COUNT

15 Pages

FILE SIZE

603.8KB

SUBMISSION DATE

Dec 6, 2023 10:26 AM GMT+5:30

REPORT DATE

Dec 6, 2023 10:27 AM GMT+5:30

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 15% Internet database
- 12% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

Project Summarisation and Robust Backup Utility for MIS

A Project Report Submitted by

Devansh Kaushik

in partial fulfillment of the requirements for the award of the degree of

Master of Technology

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology, Jodhpur

Computer Science & Engineering

December, 2023

Declaration

I hereby declare that the work presented in this Project Report titled Project Summarisation and Robust Backup Utility for MIS submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of Master of Technology, is a bonafide record of the research work carried out under the supervision of Dr. Sumit Kalra. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Signature

Devansh Kaushik

M22CS005

Certificate

This is to certify that the Project Report titled Project Summarisation and Robust Backup Utility for MIS, submitted by Devansh Kaushik(M22CS005) to the Indian Institute of Technology Jodhpur for the award of the degree of Master of Technology, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature

Dr. Sumit Kalra

Acknowledgements

I would like to thank my supervisor Dr. Sumit Kalra to give me the opportunity to work under his guidance for my MTP. I would thank him for guiding me throughout the project and providing me valuable insights.

I would also to thank Department of Computer Science and Engineering for providing with me these resources to work on this required. I have received great exposure from this project. He provided various resources which helped me during my research project. I want to express my gratitude Mr. Anurag Purohit with whom we had numerous discussions as the project progressed. Finally, I would thank my parents who provided constant support and motivation throughout the project. ...

Abstract

With the evolving development of responsive and secure software systems, the necessity for government agencies to update and shift from document mode to software applications has increased. Acknowledging this view, CWDB (Cotton Wool Developemnt Board), a sub-department of Ministry of Textiles, has been collaboraring with IIT Jodhpur in designing and deploying a software architecture for the MIS (Management Information System), alongside their landing wbesite. This project focuses on project summarisation and proposing a backup architecture for the MIS.

CWDB has 4 schemes with sub-components under which IAs (Implementing Agencies) and propose projects and funds for the region under their jurisdiction. The Board would require constant progress reports and proofs regarding expenditures and beneficiaries related to the project. This would also prevent misuse of funds and resources and direct them to relevant authorities and beneficiaries. The project summarization module will focus on sub-components of all schemes and keep track of all resources and expenditures. This basically means IAs have to fill quarterly quotations and reports of their respective component, which will cover all significant data required to prevent misuses and show progress.

The backup module is supposed to take scheduled backups of MIS data in regular intervals, focus on the design architecture and distributed aspects for faster yet secure data retrieval and updates. The current architecture involves an NIC Server hosted at Jaipur in India, with a server being an older version whose details we will delve into in the later part of this thesis. The architecture will also attempt to ensure least data loss, fault tolerant and available data storage.

Contents

List of Figures

List of Tables

Project Summarisation and Robust Backup Utility for MIS

1 Introduction and background

In today's rapid technological advancements, incorporating modern Information Systems has become essential for boosting the efficiency and transparency of governmental operations. This thesis focuses on implementing a Management Information System (MIS) customized for the CWDB (Cotton Wool Development Board) under the Ministry of Textiles, Government of India. The main goal is to move beyond traditional document-based methods, introducing a digital transformation for project proposal submission, scheme approval, and fund releases. This bold initiative takes the form of a web application designed to revolutionize the ministry's workflow, simplifying bureaucratic processes and enabling better-informed decision-making.

The MIS portal consists of 7 modules, out of which module 2, 3, 4, 6 have been completed:

Module 1: User and Role Management Module

Module 2: Project Proposal Submission Module

Module 3: Project Approval Management Module

Module 4: Project Progress Report Management Module

Module 5: Project Fund Release Management Module

Module 6: Summary Report Generation Module

Module 7: Backup Module

This thesis will focus on the project reports in Module 4 for summarisation, and Module 7 for backup utility. We meticulously delve into two pivotal modules of the overarching project. Project summarisation will involve data collection from all Implementing agencies (IAs). Also, the number of viruses and malwares developed keeps increasing at a terrifying rate. According to Symantec's 2018 Internet Security Threat Report [2], 92% of new malware have been found in the year. And according to Microsoft Defense Report 2022 [3], there has been a 74% increase in password attacks in a year. Hence, A backup module will be proposed for resilient, fault-tolerant and secure backup of data on regular intervals, which would is divided into 3 regions: Un-Safe, Middle and Safe regions, depending on the risk imposed by infecting malware.

The rest of this thesis is organised as follows. Section 2 introduces existing works and related studies for backup module. Section 3 explains the methodology of project summarization and proposed backup architecture.

2 Literature survey

This section briefly discusses existing works on backup architectures. There are different Various MNCs have developed NAS systems which provide backup systems and enabling prevention from malware such as Viruses, Worm, Trojans-horse, Rootkit, Backdoor, Botnet, Spyware, Ransomware etc. Corporations like Synology [4], QNAP [5] and et al. offer NAS-based backup systems for Ransomware threats, however, they are unable to disconnect the backup system automatically. They depend on the end user to, if possible, disconnect the system manually as soon as possible to prevent virus from entering the system. But, this brings us back to the point of creating or putting up antivirus, which can't detect and remove all viruses. [6]

In [7], Dong Jin et al. proposed a container-based backup system where the backup from both local and remote systems was taken in a secure docker container. However, the docker container doesn't detect any virus, thus has no mechanism to deal with virus if the backed up data itself contains malware. Thus, the thesis will inspire their architecture from a backup system proposed by Myungjoon Shon et al.[8], take ensure critical data reliability at low cost [9], and hold encrypted data after lossless compression [10].

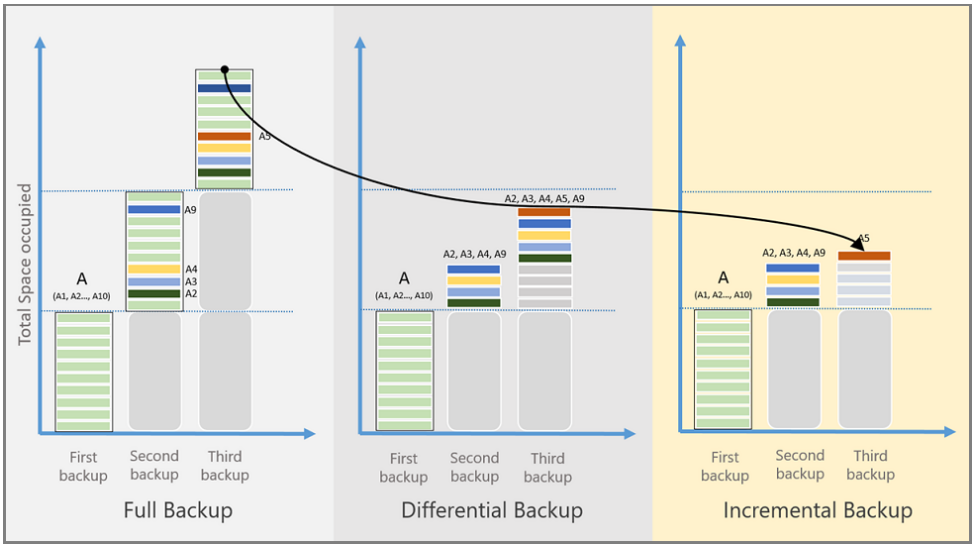

Figure 2.1: Source: Types of backups [1]

A full backup captures all files on a selected drive, including system, application, and user data. However, it is time-consuming. Incremental backups are faster but involve a more complex recovery process. For instance, if a file is updated after a Saturday full backup and an issue occurs on Tuesday, recovery would require access to the Monday night backup.

Differential backups, like incremental, only capture updated files since the last full backup. They differ in that they do not clear the archive bit. Consequently, a file updated after a full backup will be archived with each differential backup until the next full backup clears the archive bit [8]. Hence our focus will be on developing a differential backup-based architecture, with automated cut-off from the maintain to prevent malware from entering the backup storage.

3 Problem definition and Objective

Historically, conventional approaches, involving manual procedures, restricted communication channels, and paper-centric systems, have influenced the dissemination of information and the delivery of services. Although effective in their respective periods, these methods now encounter a range of challenges that impede their capacity to address the evolving requirements of the contemporary era.

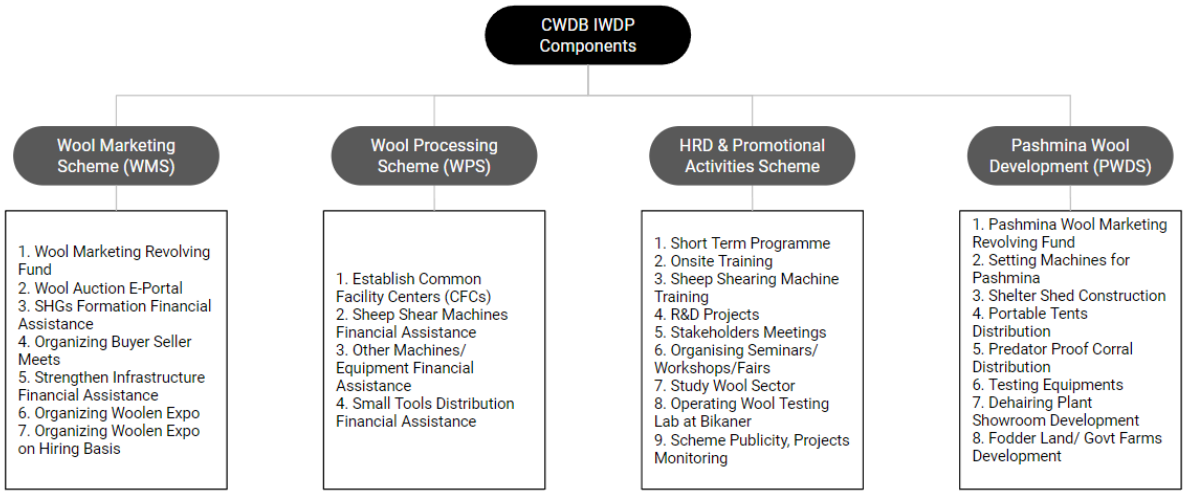

Figure 3.1: CWDB IWDP Components

3.1 Project Summarisation Module

- **Data Collection:** Data must be collected from all beneficiaries, in order to prevent wrong data added.
- **Scheme Component-wise reports:** With CWDB's 4 ongoing schemes and 28 sub-components, each components will have tailored-made progress reports to be summarized.
- **Data Analytics Dataflow:** Meaningful parameters need to be sent to dashboard side to display descriptive statistics - budget allotted, spent, number of beneficiaries, categories benefitted, state-wise distribution etc.

Project proposals must be done by relevant authiorities or Implementing agencies only, and budget/resources expenditure must be tracked with quarterly submission of these progress reports.

3.2 Backup Module

- **Design**: A differential distributed architecture divided into 3 regions, with fault-tolerant models to check malware infections, before transferring files to offsite backup (Safe Region).
- **Evaluation**: Certain software quality parameters must be met for the architecture to be resilient, reliable and secure.

4 Methodology

Firstly, we will discuss the data collected on the basis of the schemes and previously kept hard documented data.

4.1 Data Collection

The data collected and required can be classified into 2 types - beneficiaries data and expenditure data. IAs should give all beneficiaries contact and identity information, which can be later cross-checked for verification if they have received said resources by the respective IA. Payment-proofs must be submitted in pdf format. Expenditure data involves sharing allotted budget and component-wise budget spent for the quarter.

4.2 Dataflow & Storage

The MIS system is being developed on Django. Since all components have a different data requirement, there would be atleast 28 data models to be in which data can be submitted. The progress report would be reminded for all users every quarter, if not submitted, the dept would be notified. Since beneficiaries could be in hundreds, MS Excel sheets would be shared with specific columns. Invalid entries would be prevented using added data validation from Excel. Each sheet can have maximum 1000 entries.

Since all beneficiaries and expense Excel sheets would have hundreds of entries, combining them everytime to create views for the end user would be expensive. Hence, parameters such as number of beneficiaries, category, address/state, budget allotted and spent would have a separate data model and stored database, to reduce computation.

4.3 Proposed Backup Architecture

As mentioned before, the architecture has 3 regions:

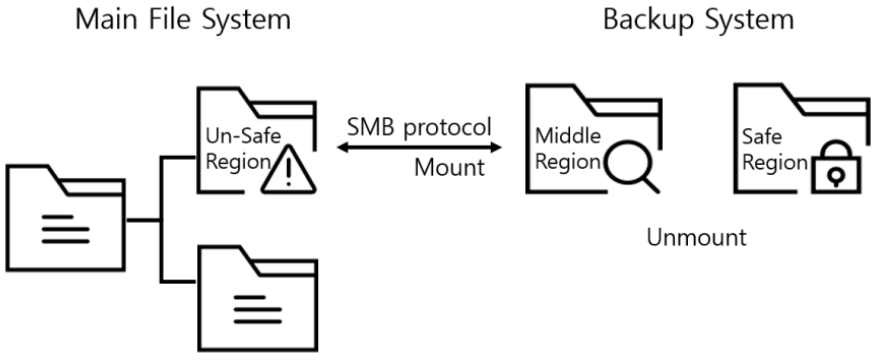

Figure 4.1: Backup System Architecture

- **Main File System:** It is a common file system, exposed to malware and viruses. Data would be transferred to the middle regions after mounting through SMB (Server Message Block) protocol

[11]. Files can be transferred to the local or remote backup systems. The unsafe region will be unmounted when checking for malware in the local or remote backup system.

- **Middle Region:** Connects safe and unsafe region, mounts a connection and transfers all the files to safe region if files are scanned to have no malware after running the detection system.
- **Safe Region:** aka offsite backup, isn't connected to the internet. If no malware is found in the middle region, safe region can connect over the network and transfer files.

4.4 Backup Procedure

Normal backup of data would be done weekly when the network congestion or data interaction is supposedly the least, for assumption, let it be Saturday. Then the backup would be kept on the un-safe file system until next normal backup, i.e., next Saturday, with archive bit set to 1. So if Monday's data is required on Tuesday, it can be retrieved from the un-safe system. On Saturday, all files be transferred to middle regions with archive bit set to 0 using SMB protocol. At this moment, the safe region is disconnected and isolated from the middle regions. After backing up files, the middle regions are unmounted from the main file system, and the files are scanned for malware.

The middle regions will be remounted to the main file system rather than being mounted to the safe area if files inside them are infected, since there can be errors in the malware detection and the backup method should be attempted once more. Files can be compressed before transferring to safe region, thus optimizing disk space. [10]

Even though SSDs are generally used to store data which has high-density and low-cost flash memory, its poor reliability is one ofn the major concerns. Virtual Tape Libraries (VTLs) do have rapid backup and recovery [12], restoring data from VTL can take a long time, and they often rely on cloud-based tapes which can be more difficult to manage compared to physical tapes. To address this, a **critical data backup** design can be used to reduce cost and increase reliability. The fundamental concept involves storing two duplicates of crucial data initially in high-speed memory to fully leverage its performance and durability. Subsequently, one of the duplicates will be transferred to the slower memory within the stripe to mitigate the write amplification resulting from distinct access granularity between the two memory types.

5 Theoretical/Numerical/Experimental findings

Currently, the static website is being hosted as a VM by NIC, Govt of India, in Delhi. After setting up remoter connection with the remote server, the current system is a 64-bit Windows Server 2012 R2 Datacenter, Intel Xeon CPU E5-2640, implying the processors are pretty good, but the OS hasn't been updated since 2012. The server has an average 10% CPU usage and 87% memory usage, since the system has only 2GB primary memory. Hence, a shift to new VM is required to deploy MIS + the dynamic website as well. An additional VM with local server hardware located in the dept can be set up, with the additional VM being the middle region and the local department server as the safe region. Since the NIC hosted VM is a distributed system with 180 servers, a malware attack will hit this system and the additional VM, thereby isolating local server from the malware.

To analyse if the system holds the standard software architecture quality attributes[13]:

- **Availability:** The use of a common file system, exposed to malware and viruses, may pose availability risks. However, the design mitigates this by transferring data to the middle regions after mounting through SMB protocol. This helps in isolating the unsafe region during malware checks, ensuring that the main file system is available for regular operations. The safe region is already isolated, hence ensure safe region is free from malware.

- **Scalability:** The architecture allows for scalability in terms of storage by using the main file system, middle region, and safe region. Each component can be scaled independently based on the organization's requirements.

- **Security:** The middle region acts as a buffer between the safe and unsafe regions, ensuring that only malware-free files are transferred to the safe region. The use of SMB protocol and unmounting during malware checks adds a layer of security.

- **Performance:** The backup procedure is designed to optimize performance. Weekly backups are scheduled during periods of low network congestion. The use of SMB protocol and the transfer of files with archive bits set to 0 contribute to efficient data transfer. The use of high-speed memory for critical data initially, followed by the transfer to slower memory, helps in leveraging the performance of high-speed memory while still ensuring reliability.

- **Reliability:** Storing duplicates of crucial data in high-speed memory and transferring one duplicate to slower memory enhances reliability by providing redundancy and mitigating write amplification.

6 Summary and Future plan of work

In this thesis, we collected and extracted meaningful data from 28 components from all schemes in the form of progress reports. Adding data validation in Microsoft Excel prevents invalid data entries. beneficiaries entries can be cross-checked with contact information while expenditure data can be referenced with payment proofs. The proposed backup system proposes a reliable and secure system for backing up data, within a focus on satisying all architecture quality attributes and implementing critical backup design for faster and persistent storage in the safe region with compressed files. Below future work of the project:

- The proposed backup system/architecture appears to have a balanced approach to address availability, scalability, security, performance, and reliability. However, the effectiveness of the system would also depend on the implementation details, the specific tools and technologies used, and the organization's specific requirements and environment.
- Regular testing and updates to the backup strategy are crucial to ensuring its ongoing effectiveness and alignment with evolving security threats.
- Detection systems cannot detected every malware, thus, a certain guarantee need to be calculated and ensured by the detection system.
- The security of the current system, both software and hardware anomalies, need to be checked and solved. The new dynamic website is going to be deployed on Drupal, while the MIS is deployed on Django. Hence, code malfunctions, package updates, architecture flaws are yet to be detected.
- Since the MIS System is currently being developed, further code optimizations can be done after its development.

References

[1] A. Mallick and et al., "Azure backup architecture and components," in *https://learn.microsoft.com/en-us/azure/backup/backup-architecture*, 2023.

[2] "Symantec internet security threat report," in *https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-executive-summary-en.pdf*, 2018.

[3] "Microsoft digital defense report," in *https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5bUvv?culture=uscountry=us*, 2023.

[4] Synology, "Protect yourself against encryption-based ransomware — synology inc." in *https://www.synology.com/en-global/solution/ransomware*, 2020.

[5] QNAP, "Mitigate the threat of ransomware with qnap nas," in *https://www.qnap.com/solution/ransomware/en/*, 2020.

[6] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[7] Y. Jin, M. Tomoishi, S. Matsuura, and Y. Kitaguchi, "A secure containerbased backup mechanism to survive destructive ransomware attacks," in *International Conference on Computing, Networking and Communications (ICNC), 1-6*, 2018.

[8] M. Shon, H. Kim, K. Park, J. W. Park, K. Won, and J. Hong, "A robust and secure backup system for protecting malware," in *https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-executive-summary-en.pdf*, 2018.

[9] L. Luo, D. Yu, Y. Lv, and L. Shiuchi, "Critical data backup with hybrid flash-based consumer devices," in *ACM Transactions on Architecture and Code Optimization*, 2023.

[10] X. Li and J. Chen, "Innovative architecture of college sports online training data based on cloud backup of remote data center," in *4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2022.

[11] V. Javaraiah, "Backup for cloud and disaster recovery for consumers and smbs," in *Fifth IEEE International Conference on Advanced Telecommunication Systems and Networks (ANTS)*, 2011.

[12] Y. Zhao and N. Lu, "Research and implementation of data storage backup," in *IEEE International Conference on Energy Internet*, 2018.

[13] "Reasoning about software quality attributes," in *https://insights.sei.cmu.edu/library/reasoning-about-software-quality-attributes/*, Software Engineering Institute, CNY, 2018.

18% Overall Similarity

Top sources found in the following databases:

- 15% Internet database
- Crossref database
- 7% Submitted Works database
- 12% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	researchgate.net Internet	4%
2	Myungjoon Shon, Heejin Kim, Kicheol Park, Juw Won Park, Kwanghee ... Crossref	3%
3	Indian Institute of Technology Jodhpur on 2021-01-18 Submitted works	3%
4	Indian Institute of Technology Jodhpur on 2021-01-19 Submitted works	2%
5	iitj on 2023-11-24 Submitted works	1%
6	Longfei Luo, Dingcui Yu, Yina Lv, Liang Shi. "Critical Data Backup with ... Crossref	1%
7	v1.overleaf.com Internet	<1%
8	Xiaofei Li, Juan Chen. "Innovative Architecture of College Sports Onlin... Crossref	<1%

9	arxiv.org Internet	<1%
10	link.springer.com Internet	<1%
11	nukib.cz Internet	<1%
12	coursehero.com Internet	<1%
13	learn.microsoft.com Internet	<1%
14	diva-portal.se Internet	<1%