Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Informatics

Bachelor Thesis

Data Backup and Archiving

Mykola Bas

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Mykola Bas

Informatics

Thesis title

Data backup and archiving

Objectives of thesis

The goal of this thesis is to develop a data backup and archiving strategy for a business project. This strategy needs to ensure the integrity, availability, and security of the contained data. The thesis aims to specify the requirements of the business about data backup and archiving, evaluate different backup and archiving solutions and propose a proper backup and archiving plan.

Methodology

The thesis methodology will include:

1. A literature review of the matter

2. Interviews with a key stakeholder (project manager)

3. Analysis of existing solutions

4. Proposal of suggested solution

The literature review will be focused on the best practices for data backup and archiving and available solutions on the market.

The interviews with the project manager will help to identify the requirements of the business.

The analysis of existing backup and archiving solutions will help evaluate the different options available and identify the most suitable solution for the business.

The proposal of suggested solution will suggest the most suitable backup and archiving solution within particular restrictions.

The proposed extent of the thesis

35-45p.

Keywords

data backup, data archiving, data backup strategy, enterprise backup strategy, backup and archiving solutions

Recommended information sources

Backup Strategy Best Practices Organizations Should Follow. VM Backup Solution for VMware, Hyper-V, EC2 | NAKIVO. [Online] NAKIVO, 15 August 2022. [Cited: 01 May 2023.] https://www.nakivo.com/blog/how-to-build-a-backup-strategy-full-guide/.

Coronel, Carlos, Morris, Steven and Rob, Peter. Database Systems: Design, Implementation, and Management, Ninth Edition. Boston : Cengage Learning, 2010. ISBN: 978-0-538-74884-1

Implementation of Data Backup and Synchronization Based on Identity Column Real Time Data Warehouse. Adnyana, I Gede and Jendra Sulastra, I Made Dwi. 1, Denpasar, Indonesia : Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 2020, Vol. 11. ISSN 2541-5832

Little, David B. and Chapa, David A. Implementing Backup and Recovery: The Readiness Guide for the Enterprise. Indianapolis : Wiley Publishing, Inc., 2003. ISBN: 0471227145.

Preston, W. Curtis. Backup & Recovery. Inexpensive Backup Solutions for Open Systems. Sebastopol, CA : O'Reilly, 2006. ISBN: 978-0-596-10246-3.

Expected date of thesis defence

2023/24 SS – PEF

The Bachelor Thesis Supervisor

Ing. Martin Havránek, Ph.D.

Supervising department

Department of Information Technologies

Electronic approval: 4. 7. 2023

doc. Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 3. 11. 2023

doc. Ing. Tomáš Šubrt, Ph.D.

Dean

Prague on 10. 03. 2024

Declaration

I declare that I have worked on my bachelor thesis titled "Data backup and Archiving" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 15.03.2024 ____________________________

Acknowledgement

I would like to thank Ing. Martin Havránek, Ph.D. and all other persons, for their advice and support during my work on this thesis.

Data Backup and Archiving

Abstract

The aim of the thesis is to develop a data backup and archiving strategy for the business project, for improving data integrity and retain its accessibility in case of corruption or loss. The methodology includes a literature review, interviews with the project manager, analysis of the current solution, and the suggestions for improvement.

The findings of the thesis recommend a hybrid approach for the backup, detailed and fixed scheduling of the backups (for both full and incremental backups), automation of the processes that are done manually. Also, the implementation of Veeam Backup & Replication Community Edition software is covering all of the current system's flaws.

As the conclusion of the work, this research found that the current solution has some significant weaknesses, such as manual operations usage, lack of data encryption, and the offsite storage, which is not used. Fixing the gaps stated above not only protects the data against the disasters or data losses, but also ensures the business will operate without disruptions.

Keywords: data backup, data archiving, data backup strategy, enterprise backup strategy, backup and archiving solutions

Zálohování a archivace dat

Abstrakt

Cílem práce je navrhnout strategii zálohování a archivace dat pro obchodní projekt, pro zlepšení integrity dat a zachování jejich dostupnosti v případě poškození nebo ztráty. Metodika zahrnuje rešerši literatury, rozhovory s projektovým manažerem, analýzu aktuálního řešení a návrhy na zlepšení.

Závěry práce doporučují hybridní přístup k zálohování, podrobné a pevné plánování záloh (pro plné i přírůstkové zálohy), automatizaci procesů, které jsou prováděny ručně. Implementace softwaru Veeam Backup & Replication Community Edition také pokrývá všechny nedostatky současného systému.

V závěru práce tento výzkum zjistil, že současné řešení má některé významné slabiny, jako je používání ručních operací, chybějící šifrování dat a úložiště mimo pracoviště, které se nepoužívá. Oprava výše uvedených nedostatků nejen chrání data před katastrofami nebo ztrátami dat, ale také zajišťuje, že podnikání bude fungovat bez přerušení.

Klíčová slova: zálohování dat, archivace dat, strategie zálohování dat, podniková strategie zálohování, řešení zálohování a archivace.

Table of content

List of figures

List of tables

1 Introduction

Information plays a significant, if not crucial role in the modern business world. It is thread through all parts of the enterprise and it's very hard to imagine a 21st century business without the information systems and information technologies implemented. So, if information is so essential, it is critical for businesses to maintain their integrity and speed of recovery, if any unpredicted situation (disaster) happens.

Also, there is a major point that recovering from data corruption/loss is much more time and money consuming to than constantly maintaining a backup solution, which restores the lacking information parts and keeps business up and running (1).

In order to make this point more specific, there are six major reasons for implementing a backup:

1. *Hardware may eventually fail*. In the last few decades, most businesses shifted to computer information systems. So, if something unpredicted happened to the data-storing hardware, the data is also lost.
2. *Human error*. Regardless of how professional the employees may be, they are human beings and prone to make mistakes. In some sophisticated procedures it is possible to miss an important step or accidentally delete a file, which is thought to be unnecessary. Also, it may even happen to spill a cup of tea on a laptop with crucial information.
3. *Nature and other disasters*. Sometimes, some unpredicted nature events may occur. Tornadoes, extreme heat, or freeze may destroy or severely corrupt the data. Apart from that, no one is 100% secure from accidental fire or flood in the working premises.
4. *Company market competition*. In a modern world of information technology, losing or disclosing the customers' data will severely damage the reputation and trust to the business. And, consequently, reputation loss leads to loss of money by a decrease in new customers incoming and old customers' retention.
5. *Data Theft*. Some employees may be extremely dissatisfied with the working conditions, so they may decide to commit a "revenge" by stealing or deleting the data.
6. *Future-proof solutions.* If an enterprise stores the data within its facilities only, there is not only a risk of data loss, but also, makes the work with data inflexible.

On the other hand, having the off-site or cloud backup solutions allow employees to work with data remotely and in a more fluid manner.

After a decision to implement a backup is made, the business needs to start planning backup architecture (2 p. 18). The planning starts from the fundamental questions, which help to build the proper backup solution, not missing any important points and not overspending the company's budget.

- *Why?* Why protect the data from a disaster or an unpredictable event? What will the business lose if it loses the data? What types of data does the business have and how "expensive" each type of data is?
- *What?* What data will be backed up? Is it the backup of the entire enterprise's infrastructure or only some distinct drives and filesystems should be backed up?
- *When?* When is the best moment for making a backup?
- *Where?* Where to store the backups to keep them as safe as possible?
- *Who?* Who is going to keep the backup infrastructure up and running? Who is going to be responsible for the software and hardware?
- *How?* How is the enterprise going to implement backing up? Will it be accomplished with the RAID, mirroring, or the off-site storage?

.

2 Objectives and Methodology

2.1 Objectives

The goal of this thesis is to develop a data backup and archiving strategy for a business project. This strategy needs to ensure the integrity, availability, and security of the contained data. The thesis aims to specify the requirements of the business about data backup and archiving, evaluate different backup and archiving solutions and propose a proper backup and archiving plan.

2.2 Methodology

The thesis methodology will include:

1. A literature review of the matter
2. Interviews with a key stakeholder (project manager)
3. Analysis of existing solutions
4. Proposal of suggested solution

The literature review will be focused on the best practices for data backup and archiving and available solutions on the market.

The interviews with the project manager will help to identify the requirements of the business.

The analysis of existing backup and archiving solutions will help evaluate the different options available and identify the most suitable solution for the business.

The proposal of suggested solution will suggest the most suitable backup and archiving solution within particular restrictions.

3 Literature Review

3.1 What is backup and why do enterprises need it?

Backup is a copy of data, stored in another place (another drive, storage or in the cloud) and used for recovery if the original data is lost or, for some reason, corrupted. Also, backup may be used for retrieving the earlier versions of data if needed.

3.2 Basic data backup definitions

To dive deeper into the details of how to choose the best backup plan, at first, there is a need to clarify the basic backup definitions, which are type, location, scheduling and keeping duration.

3.2.1 Backup types

There are main types of a backup, which are following:

- *Full backup*. It means that all of the files are copied to the recovery storage, regardless of whether they have been changed or not since previous full backup. It consumes relatively more storage space and time, and therefore, it is done less frequently (once a week or once a month).
- *Incremental backup*. When this backup type is carried out, it stores only the changes in files since the last backup (no matter, was it full or another incremental backup). It is the most space- and time-saving backup type. However, to restore data from an incremental backup, the system needs the last full backup and other incremental backups in-between. Apart from that, the incremental backups may be tagged with a level (from 1 to 9) for understanding, in which order do they go. For example, level 1 incremental backup keeps the changes since the last full backup. Level 2 incremental backup keeps the changes since the last level 1 incremental backup or the lower-level backup, and so on. Full backup is usually marked as level 0.
- *Differential backup*. This type of backup stores files modified since the last full backup. It ignores other differential backups made in between. The differential backup consumes more storage, than incremental backups, but it needs only the last full backup to restore the data (no backups in between needed).

- *Mirroring (passive backup).* A mirror backup is an exact copy of selected folders and files from the source. On the physical level mirroring is usually implemented via a RAID 1 disk array, where the data is written to two (or more) hard drives simultaneously. Mirror backups are known for their speed, copying files to separate storage without compression. However, there are some significant drawbacks, like a significant risk is data deletion (data deleted on the source will also delete the data on other disks), password protection is not possible, and tracking different file versions is challenging for such backup type.
- *Clustering (distributed system).* In the context of backup and data archiving distributed systems store data on multiple storage nodes or servers working together in different places. This helps to increase data availability, reliability, and future growth. Data is replicated across the nodes, which ensures redundancy and high availability, minimizes downtime in case of a disaster, and provides constant access to the data. Clusters may offer load balancing, distributing workloads efficiently, scalability (for future expansion of data), data integrity checks, automatic recovery, and fault tolerance. The drawbacks are: complexity, cost, single points of failure (for example, management component or shared storage), scalability limits, and management complexity.

Also, it is worth noting that all these types of backup work in the best way if properly combined. For example, it is technically possible to always conduct a full backup. However, it will consume senselessly much more storage space and time for the backup process.

3.2.2 Backup location

Backup location describes where the data will be physically stored, whether it will be stored in the premises of a business, or somewhere else. There are three location categories: local, offsite and cloud based.

- Local. Local backups are stored closely to the source data (usually, in the same building). The data under question can be stored on external storage, tape drive or Network-attached storage (NAS).

  Pros: full control over the storage, backup speed.

Cons: vulnerable to any local disasters, like the source data itself (fire, flood, hardware failure, etc.) (3)

- *Cloud-based*. This type of location uses distributed datacentres for storing and managing the backup infrastructure.

  Pros: More accessible from the distant places (only internet connection required); may be cheaper, than offsite; easier to scale the infrastructure, if needed; more tools for backup process automation.

  Cons: It is dependent on the Internet connection; More difficult to change the cloud backup provider if backups are large (4).

Location plays a crucial role in data integrity and costs incurred. Also, as with the types of backup, you need to combine your backup locations to achieve the best integrity/cost proportion.

3.2.3 Scheduling

This category is responsible for when and how often the business creates a particular of backup (full/incremental/differential). It is very important to plan the schedule carefully to avoid unnecessary storage consumption and ensure that downtime for business operations will be minimal (2 pp. 27-33).

Also, the time of the backup is very important, because it should not interfere with the normal functioning of the information systems. For instance, it is advised to back files up at night, because the load on network and storage is usually the lowest at that time. There might be an exception to this when the business is a 24/7 online store, which needs to operate consistently throughout any time. In this case time needs to be chosen at the lowest load level. Several examples presented in the tables below (Table 1 and Table 2).

Table 1: Example of weekly full and daily differential backup schedule

Day of week	Backup start time	Backup type/level
Sunday	01:00	Full/0
Monday	01:00	Differential/1
Tuesday	01:00	Differential/1
Wednesday	01:00	Differential/1
Thursday	01:00	Differential/1
Friday	01:00	Differential/1
Saturday	01:00	Differential/1

Source: (2).

Table 2: Example of weekly full and daily incremental backup schedule

Day of week	Backup start time	Backup type/level
Sunday	01:00	Full/0
Monday	01:00	Incremental/1
Tuesday	01:00	Incremental/2
Wednesday	01:00	Incremental/3
Thursday	01:00	Incremental/4
Friday	01:00	Incremental/5
Saturday	01:00	Incremental/6

Source: (2).

Schedule and timings of a backup may differ. It depends on the amount of data backed up, frequency of data usage and the time when the system is the least loaded.

3.3 Software solutions for backup

To make a comparison table of the backup software, which will be used later, the following preliminary criteria were chosen to leave the most appropriate software solutions:

- Multiplatform support: the software should be compatible with both Windows operating systems and Linux distributions, as it would keep the same solution if the OS changed.
- Backup schedule support: To ensure the backups will be done without the need for a responsible person to perform the backups manually.

The comparison table consists of cost, ease of use, features and security.

Table 3: Comparison table of software solution for backuping

Solution	Pricing	Ease of Use, pts.	Features
Acronis Cyber Protect	From 1 856,00 Kč, free tryout period available	9.1/10	• Integrated Backup and Cybersecurity • Advanced Security • Forensic Backup
NinjaOne	n/a, free tryout period available	9.3/10	• Incremental Block-Level Backup • Secure Restore Options

			• Versatile Backups and Recovery
Veeam Backup & Replication	Free (Community Edition), additional functions	9.1/10	• Flexible Data Protection • Ransomware-Proof Backups • Advanced Security
Veritas Backup Exec	n/a, free tryout period available	8.2/10	• Broad Environment Support • Efficiency and Management

Source: (5).

3.4 Backup hardware

Choosing the appropriate backup hardware and media ensures that the data will be stored safely and may be retrieved as fast as possible in case of unpredictable situations. Also, the business needs to consider the cost and limitations of a particular media, as some of them may not be suitable for smaller organizations.

3.4.1 Storage media

Hard Disk Drives (HDDs): Traditional HDDs have relatively affordable price per gigabyte and offer the storage of a relatively high capacity. They are usually used for local backups and may be easily disconnected for offline storage.

Solid State Drives (SSDs): SSDs are faster and more durable than HDDs, but they are more expensive in terms of price per gigabyte. They are suitable for backups that require quick access to the data.

Network-Attached Storage (NAS) Devices: These devices often use multiple HDDs or SSDs in RAID configurations to provide redundancy and data replication.

Tape Cartridges: Tape cartridges are a long-term storage solution, providing high capacity and data integrity. They are suitable for archiving and regulatory compliance. However, they require a special tape interface, which may cost substantial money. Also, reading and writing speed is relatively low.

Optical Media (e.g., DVDs, Blu-rays): Optical discs are portable and durable for archival purposes, but they have limited storage capacity compared to other options.

Cloud Storage: Cloud storage services like Amazon S3, Google Cloud Storage, or Dropbox offer offsite backup solutions. Data is stored in data centers maintained by the service providers, providing redundancy and accessibility from anywhere with an internet connection.

3.4.2 RAID configurations

RAID stands for "redundant array of independent drives". RAID allows to operate several drives as one storage unit and helps to restore data if one (or more) of drives fails. There are many RAID configurations and combinations, but the most commonly used are RAID 0 (stripping), RAID 1 (mirroring), RAID 5, and RAID 10 (6).

RAID 0 (striping): requires two or more drives, and it reads/writes information across all drives simultaneously, so the read/write speed increased, compared to a single drive. However, this configuration has no redundancy, so, if any of the drives fail, all data will be lost.

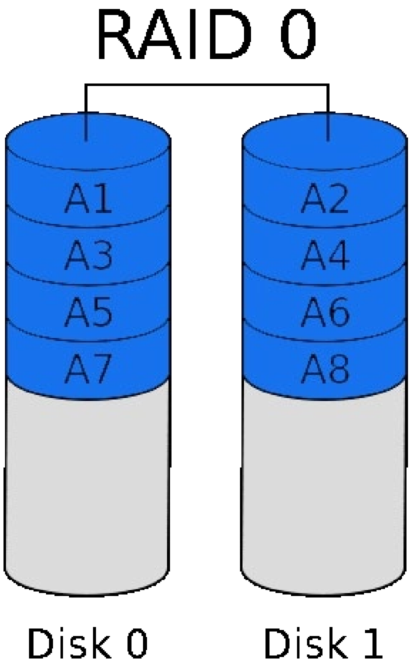

Figure 1: Visual representation of RAID0 data distribution, Source: (18).

RAID 1 (mirroring): reads/writes data across two or more hard drives. It provides data redundancy by duplicating information on each drive. If one of the drives fails, data will still be accessible using the mirrored drive, ensuring data integrity, and minimizing downtime of the system. However, it uses twice or more drives (depending on configuration) to achieve the same storage space compared to a single drive. RAID 1 primarily serves as a data backup and fault tolerance solution for mission critical data, such as accounting.

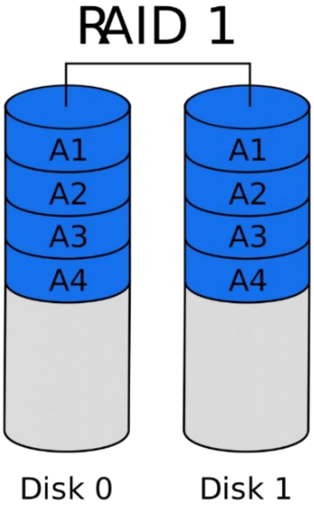

Figure 2: Visual representation of RAID1 data distribution, Source: (18).

RAID 5 (disk stripping with parity): typically requires a minimum of three drives to function. RAID 5 uses striping like RAID 0 for faster read/write speed but also includes distributed parity information across the drives. Parity information used for restoring the data from a failed drive. With three drives, the data is stripped across two drives and parity information on the third one. This allows for data recovery in case one drive fails. However, RAID 5 can also be implemented with more drives for increased storage capacity and performance. Also, RAID 5 configuration is suitable for NAS.

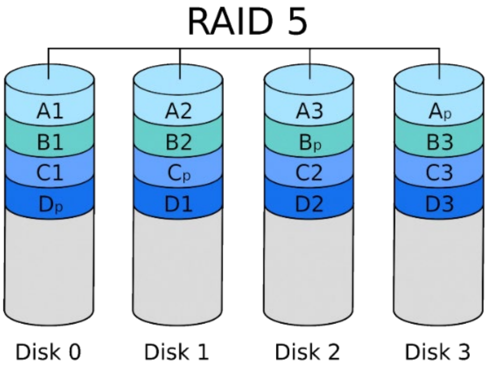

Figure 3: Visual representation of RAID5 data distribution, Source: (18).

RAID 10 (also known as RAID 1+0): combines the RAID 1 (mirroring) and RAID 0 (striping). It requires at least four drives to function. This configuration provides both data redundancy and higher read/write speed. Data is mirrored across pairs of drives, and then these pairs are striped for speed. RAID 10 offers high fault tolerance, as it can withstand the failure of one or more drives in each mirrored pair. So, this RAID configuration is suitable for tasks where both read/write speed and redundancy are required.

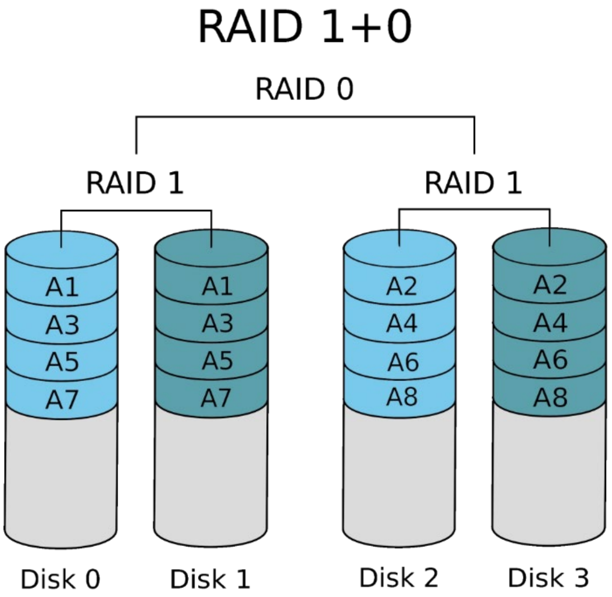

Figure 4: Visual representation of RAID10 (1+0) data distribution, Source: (18).

3.4.3 Planning of capacity and future scalability

Planning the capacity and future scalability of backup hardware is very important, because your backup solution needs to accommodate the current and future data needs while keeping the costs low. The steps to consider are the following:

Current Data Assessment: The process starts from estimating how much data should be backed up.

Data Growth Rate: Then, the data growth rate over the next few years is assessed. There are important factors to consider like new projects, increased storage demand for the current projects, and regulatory requirements for data retention.

Retention Policies: It should be decided how long the company needs to retain backups. Compliance and business requirements may help to find specific retention periods.

Backup Frequency: How often is it required to perform backups (for example, daily, weekly or in real time). Frequent backups require more storage.

Redundancy: The options like RAID configurations or backup copies redundancy are considered to protect the system against hardware failures.

Future Scalability: The backup hardware to consider needs to be easily expanded via adding more storage devices or upgrading the existing hardware.

Storage Needs: based on current data and expected growth the storage needs are calculated. The calculations include redundancy and additional storage overhead for efficient data recovery.

Cloud Integration: Cloud storage may also be a good possibility for backup. Many backup solutions may offer hybrid options that use both local and cloud storage.

Backup Hardware Selection: Deciding to choose the appropriate backup hardware, such as NAS devices, tape libraries, or cloud-based solutions. The selection is made based on the budget and company requirements.

Regular Assessments: Occasionally, hardware and capacity planning needs to be reassessed in order to keep up with the changes in data size growth.

Budget: To ensure the company can afford scalable solutions when needed, the budget is planned.

Data Deduplication and Compression: Data deduplication and compression are the things to think of in order to optimize or free up the storage.

Backup Software: Making sure that the backup software in use supports the hardware and scalability requirements.

3.5 Backup Strategies

3.5.1 “3-2-1” backup rule

3-2-1 backup rule is the concept of keeping the data, which was first formulated by US photographer Peter Krogh. In a nutshell, this rule has three parts:

- There should be *three* copies of data under the backup. That is one primary copy, and two backup copies are kept. The more copies are created, the less probability that all of them will fail simultaneously.
- Having *two* different storage media. Storing copies on the same type of storage increases the risk of failure because of media wear-out. To make this part work, the data and backups should be stored on the different devices, such as NAS or external hard drives.

- *One* copy of data should be stored offsite. This part will make sure that even if some local disaster happens and destroys both production and local backup copies, the offsite one will be safe (7).

3.5.2 Grandfather-father-son backup scheme

Grandfather-Father-Son (GFS) Backup Strategy: This is a frequently used method of performing backup rotations in storage media. It includes multiple backup cycles, typically three or more, and offers a combination of various backup frequencies. These frequencies may include daily, weekly, monthly, quarterly, half-yearly, and/or annual backups. Moreover, the GFS strategy is versatile, accommodating full, incremental, or differential backup methods. A key feature of this approach is the dual-location storage of backups, that enhances data security and recoverability (8) (9).

Backup Management in the GFS Scheme: In the Grandfather-Father-Son backup strategy, some backups are periodically removed for the purposes of safekeeping or disaster recovery. This approach implements the ‘first in, first out' principle, so, the oldest backups are deleted first, to free up the storage for new backups. This systematic deletion makes sure that storage space is used more efficiently and maintains a cycle of backups for optimal data security and recovery.

An example of GFS implementation may be the following:

- **Grandfather Component:** Involves a monthly full backup, executed once each month. An example could be the last day of the month. The backup is stored off-site or in a cloud-based storage solution, ensuring that data is safe and is out of the primary location.
- **Father Component:** Consists of a weekly full backup. A standard practice is to perform this backup on a specific day each week, such as every Monday, however, the day can be chosen based on the enterprise’s needs. These backups are typically stored locally.
- **Son Component:** Includes daily backups, which could either be incremental or differential, depending on the enterprise’s preference. These backups are performed daily, possibly even more frequently, and can also be saved in a local storage system.

Pros of GFS Backup: The Grandfather-Father-Son (GFS) backup method makes lots of backup copies. This means if a disaster or data corruption happens, like a computer crash or a broken disk, it’s less likely to lose the data. It stores the backups in different places, which makes it easier to get the files or system back safely.

Cons of GFS Backup: This method takes a lot of time and needs two kinds of places to store your backups. It gets even more complicated if you're doing incremental backups (where you only save changes since the last backup), setting up different rules for how long to keep backups, and using the offsite storage. Moreover, if the data is restored from an incremental backup, all backups linked need to be in perfect condition.

3.6 Data Recovery and Disaster Recovery

3.6.1 Data recovery methods and procedures

Data Recovery in Enterprises: Data recovery is restoring data that was lost, corrupted, or inaccessible for any reason. This process usually relies on backups stored in a different location, ensuring minimal data loss and business discontinuity. Data recovery is essential for businesses to prevent significant data loss and involves having a robust backup and restore plan as part of a larger disaster recovery strategy. It encompasses various methods like software for undeleting files, services for physical data recovery from damaged disks, and restoring data from cloud-based backups, with a particular focus on enterprise needs (10).

Backup Plan for Enterprises: To make safe against data loss from human errors, disasters, or cyberattacks, the companies should have a comprehensive backup and recovery solution. It should cover all crucial data, including that on servers (both on-premises and cloud-hosted), storage area networks, and endpoint devices. The backup plan should include not only files but also applications, databases, system data, and runtimes. Regular backups of these elements ensure that the business can quickly recover if data loss happens.

Recovery Point Objective and Recovery Time Objective: Determining the Recovery Point Objective (RPO) and Recovery Time Objective (RTO) is crucial in a backup strategy. RPO refers to the maximum tolerable age of backup data, while RTO is the longest allowable downtime of the system. These objectives vary based on the data's importance, the business sector, and other factors and, consequently, affect backup frequency, methods, and locations. A reliable data recovery service provider should offer a Service Level Agreement (SLA) that outlines achievable RPOs, RTOs, security measures, and compliance with industry regulations.

Cloud Data Recovery Solutions: Cloud backup and recovery solutions gain popularity for their ability to handle large data volumes and for reducing infrastructure and administrative costs. These solutions store data copies offsite, accessible via a subscription model, usually based on storage or bandwidth usage. Cloud storage is scalable, offering flexibility and resource optimization, and ensures data accessibility whenever needed.

Secure Data Recovery: The security of data backups consists of two main factors: encryption of backup files both in transit and at the storage site, and decent physical security, user authentication, and access controls at the storage location. Compliance with industry-specific data privacy and security regulations is very important, especially for protecting intellectual property and sensitive information. Ensuring the complete destruction of unnecessary backups is also a key aspect of secure data recovery.

3.6.2 Disaster recovery planning and implementation

Disaster Recovery (DR) is a policy and procedures which will be applied if a disaster (natural or human factor) when it is needed to recover the systems in question and the business up and running. Key aspects of DR include:

Risk Assessment: Identifying the most important systems and functions, and the impact if they are corrupted or lost.

DR Plan Development: Creating a plan that includes recovery strategies for different types of disasters (e.g., cyber-attacks, natural disasters).

Data Center Redundancy: Establishing backup data centers or cloud-based solutions to retain data availability.

Regular Testing and Updates: Regularly testing the DR plan to make sure its effectiveness and updating it to reflect changes in the business environment and technology.

Employee Training: making sure that staff is trained on DR procedures and their roles if a disaster happens.

3.7 Basic data archiving definitions

3.7.1 What is the difference between archiving and backup?

Backup and archiving are both a copy of source data, but they have different purposes. So, the key differences are purpose, location, speed, searchability and duration (11).

From the purpose point, a backup is for making operating restoration in a certain point in time, while archives are used to keep information for a long period and data within is not required on regular basis. Also, archives are made for reference purposes, so they are used to compare some data from the previous periods and is not used for restoration (11).

In terms of location the backups mean that the source information is kept, and the copy is made. However, when data is archived, the source information is being deleted in order to free the space up (11).

Regarding the speed, it is crucial for backups to restore corrupted data as quickly as possible, while archives don’t need that level of speed as data in archives primarily used for reference or legal purposes (11).

3.7.2 Data Lifecycle Management

Data Lifecycle Management (or DLM) refers to the policies, processes, and tools that are used to manage data within its lifecycle, from creation and initial storage to the time it becomes obsolete and is deleted. This process involves identifying, classifying, archiving, and destroying data in a systematic and controlled manner. DLM ensures that the data is kept secure and accessible if required, while also complying with legal and regulatory requirements. Also, this process makes sure the obsolete data is destroyed when it is not needed for both operational and legal purposes.

3.7.3 Archiving methods

Data archiving can be done through physical media (like tapes or HDDs), cloud storage, or hybrid systems combining both approaches. The choice of archiving method depends on factors like data volume, retrieval needs, security requirements, and cost. Cloud archiving is increasingly popular due to its scalability, accessibility, and cost-effectiveness.

3.7.4 Security concerns

There are four main concerns that needs to be covered:

- Data Integrity: Ensuring the data is not modified during archiving process.
- Access Control: Restricting access to archived data to authorized personnel only.
- Encryption: Protecting data from unauthorized access, especially during transfer stage.
- Compliance: Adhering to industry and legal standards for data protection and privacy.

3.7.5 Cost questions

The cost of keeping the archive running is based upon these three categories:

- Storage Costs: Costs vary based on the medium (physical vs. cloud) and the size of data archived.
- Maintenance Costs: Includes expenses related to data management, security, and technology updates.
- Cost-Benefit Analysis: Organizations must consider the cost of archiving against the risks and costs of data loss or non-compliance.

3.7.6 When is it ok to delete data from an archive?

This question consists of four following key points:

- Deletion Criteria: Data can be deleted when it is no longer legally and operationally needed to be stored.
- Regulatory Compliance: Ensuring deletion complies with legal data retention requirements.
- Data Review Process: Regularly reviewing archived data to determine its relevance and necessity.
- Secure Deletion: Making sure that once deleted, data cannot be recovered or accessed to prevent any security breaches.

3.8 Software solutions for archiving

Software solutions for archiving play an important role in data management for both Windows and Linux operating systems. They offer a range of functionalities to efficiently store, retrieve, and manage archived data.

Windows Archiving Solutions:

- WinRAR: Known for its robust compression and encryption capabilities; supports various formats including RAR, ZIP, and others. Also, it offers a graphical interface and integration with Windows context menu for ease of use.
- 7-Zip: Free to use and modify, making it a popular choice for cost-conscious businesses. 7-Zip supports the most important archive formats like 7z, ZIP, TAR, RAR, and more. This tool is also known for its high compression ratio.
- Acronis True Image: Not just a sole archiving tool but also a full-functional backup solution. This program offers cloud storage options for archiving, enhancing data accessibility and disaster recovery.

Linux Archiving Solutions:

- tar (Tape Archive): A standard Linux tool for combining multiple files into a single archive file, commonly used in combination with compression tools like gzip. It also supports various compression methods and is highly scriptable for automated tasks.
- gzip/gunzip: Often used in conjunction with tar for creating compressed archives. Also known for its efficient compression algorithm and speed.
- rsync: More than just an archiving tool, also used for data backups and synchronization. Ideal for creating backups over the network because of its delta-transfer algorithm, which sends only changed parts of files.

Cross-Platform Solutions:

- Veritas Enterprise Vault: Offers archiving solutions for large businesses. Provides features like data deduplication, retention management, and eDiscovery.

- Zmanda: Offers comprehensive backup and archiving solutions. Integrates with cloud storage services for off-site archiving and disaster recovery.
- Duplicati: Suitable for smaller businesses or individual use because it’s free and open source. Supports various cloud storage providers, along with encryption and incremental backups.

3.9 Archiving hardware

The archiving hardware used to store the data for extended periods of time, mostly, more than several years. The most frequently used archiving hardware includes tape drives, HDDs, Optical Media (DVDs, Blu-Rays), SSDs, Network Attached Storages (NAS), and cloud-based solutions.

- Magnetic Tape Drives: One of the oldest and most reliable forms of data storage. Magnetic tape drives store data on magnetic tape cartridges. It has the advantages of high capacity, low cost per gigabyte, long shelf life (up to 30 years), and low energy consumption since they don't require power when they are not in use. The drawbacks are slow read/write speed (and it’s even more reduced if data accessed randomly). This solution is ideal for very large volumes of data that don't require frequent access, such as archival of historical data.
- Hard Disk Drives (HDDs): Traditional spinning disk drives used for storing digital data. It has Higher access speeds compared to tapes, good for data that requires semi-regular access. However, it has a higher cost per gigabyte compared to tapes and has a shorter lifespan. Suitable for active archives where data needs to be accessed more frequently.
- Solid-State Drives (SSDs): Storage devices using flash memory, faster than traditional HDDs. The pros are high-speed data access, lower power consumption, and no moving parts, which reduces the risk of mechanical failures. Cons are higher cost per gigabyte and potential for data degradation over very long periods of inactivity. Ideal for high-performance archival needs where speed is a priority.
- Optical Media: Includes CDs, DVDs, and Blu-ray discs. The pluses are longevity (can last several decades), portability, and immunity to

electromagnetic interference. However, it is limited by lower capacity compared to other media and slower access speeds. Best suited for small to medium-sized data archives and for data that requires physical distribution.

- Network Attached Storage (NAS): Storage devices connected to a network, allowing multiple users and client devices to retrieve data from a centralized location. It's easy to set up, it is a scalable solution, and provides shared access to files. Useful for small to medium businesses for centralized data storage and sharing.
- Cloud Storage Solutions: Remote storage accessible over the internet. It can be public, private, or hybrid. Advantages: Scalability, accessibility from any place with Internet connection, and potential cost savings on hardware and maintenance. Disadvantages: Requires reliable internet connectivity and ongoing subscription costs. Use Cases: Suitable for organizations of all sizes, particularly those with distributed teams and remote work policies.

3.10 Archiving strategies

Archiving strategies is the systematic approach to store, manage, and preserve data for long-term use. In order to develop an archiving strategy the company should consider the following steps:

- Defining Archiving Policies and Procedures: Establishing clear policies for what data should be archived, its importance, legal requirements, and frequency of access, and creating standard operating procedures for how data is archived, retrieved, and eventually destroyed.
- Data Classification and Categorization: Determining which data is critical for business operations, legal compliance, or historical value. Classifying data into categories for easier management and retrieval. Categories could be based on the data type, department, or sensitivity level.
- Implementing Data Lifecycle Management: Managing data through its entire lifecycle, from creation and active use to archiving and deletion. Using software tools to automate the transition of data from active to archived status based on predefined rules.

- Selecting Appropriate Archiving Solutions: Choosing the right combination of hardware (like tapes, HDDs, SSDs) and software solutions based on data volume, retrieval needs, and budget. Also, deciding between cloud-based solutions for flexibility and scalability, or on-premises solutions for greater control and security.
- Regular Data Audits and Reviews: Conducting regular audits to ensure the archived data is still relevant, accessible, and in good condition. Data Purging: Identifying and securely deleting data that is no longer needed or has surpassed its required retention period.
- Ensuring Data Security and Compliance: Implementing strong encryption, access controls, and other security measures to protect archived data and ensuring the archiving strategy complies with relevant legal and industry-specific data retention and protection regulations.
- Disaster Recovery and Business Continuity Integration: Maintaining backup copies of the archived data in separate physical or cloud locations and integrating archiving with the organization’s disaster recovery and business continuity plans.
- Cost Management and Efficiency: Using cost-effective storage solutions for long-term archiving, such as tape drives for rarely accessed data and regularly reviewing and optimizing storage usage to balance costs and performance.

4 Practical Part

4.1 Current State Assessment

4.1.1 Introduction

This part of the thesis starts from the practical side of data backup and archiving. Because a Non-Disclosure Agreement (NDA) was signed by me, some details have been made general or anonymous.

Before starting to formulate the survey, there is a series of the abstract questions was made. This will help to estimate the business needs and will make the interview more consistent and relevant to the existing issues. The questions are:

- Is the current backup solution sufficient for the company?
- Is there a need for the data archiving?
- What is the data lifecycle for the matter?
- Are the recovery time objectives defined?
- Are there any specific constraints or resources available for being utilized for the backup process?

To get the data for further steps, a series of questions was prepared and grouped together for conducting an interview with a responsible person (product owner). This helps to organize the research and make sense of the information that is gathered. This information will give an understanding of the existing backup solution and the business requirements for the solution to propose.

4.1.2 Questions for the key stakeholder

To assess the current state of the backup process the following groups of questions were asked:

1. Current solution satisfaction:
    a. Is the existing solution fulfilling the expectations and needs of the business?
2. Nature of Data (gives the understanding of what data to back up and its importance):
    a. What types of data does the company want to backup?

    b. How critical is this data for company functioning?
    c. How is sensitive or confidential information under the backup?

3. Data Lifecycle (what is the procedure of data creation, editing and disposal):
    a. Description of the lifecycle of the data within the organization?
4. Data Volume and Growth (how big is the data size and its potential growth rate):
    a. How much data does the company currently manage, and how is it expected to grow in the future?
    b. Are there seasonal or cyclical patterns in data usage?
5. Data Access and Retrieval (frequency and timings of data retrieval):
    a. How frequently is data accessed and retrieved?
    b. Are there specific time-sensitive requirements for data retrieval?
6. Regulatory and Compliance Requirements (any government regulations):
    a. Are there any industry-specific regulations or compliance standards that the company needs to adhere to regarding data storage and retention?
7. Recovery Objectives (how fast the system should be recovered in case of failure):
    a. What are the recovery time objectives (RTO) for this type of data?
8. Backup Storage and Infrastructure (what hardware and software does the company have at the moment):
    a. What storage infrastructure is currently in place for backups?
    b. Is there a mix of on-premises and offsite/cloud-based storage solutions?
9. Data Retention Policies (what are the organization's policies of keeping the data):
    a. Are there specific policies in place regarding the retention of data?
    b. How is data classified in terms of its retention requirements?
10. Backup Testing and Validation (checking whether the backups are eligible to be used in case of a disaster or data corruption):
    a. How often are backup and recovery procedures tested?
    b. Is there a regular schedule for validating the integrity of backup data?
11. Security Measures:
    a. What security measures are in place to protect backup data?
    b. How is encryption applied to ensure the confidentiality of backed-up information?
12. Budget and Resource Allocation:
    a. How is the budget allocated for backup and archiving solutions?

   b. Are there any resource constraints or considerations that impact the implementation of backup strategies?

13. Specific Constraints and Requirements:
    a. What may be other constraints, not mentioned above?
    b. Requirements for the backup

4.1.3 Interview results

1. Current solution satisfaction:
   a. Is the existing solution fulfilling the expectations and needs of the business?

      No, the project manager would like to improve the existing backup solution.
2. Nature of Data:
   a. What types of data does the company want to backup?

      Images for training neural networks.

   b. How critical is this data for company functioning?

      Very critical, because it can severely impact the main products' development.

   c. How is sensitive or confidential information under the backup?

      This data is very sensitive, because it contains additional metadata created by the company for internal use.
3. Data Lifecycle:
   a. Description of the lifecycle of the data within the organization.

      1) Unprocessed images are uploaded to the database.

      2) The images are gradually being processed and metadata added.

      3) (optionally) The processed images and their metadata are being corrected.

      4) Some particular images are being used for neural networks training.

      5) Unwanted/unusable images are deleted.
4. Data Volume and Growth:
   a. How much data does the company currently manage, and how is it expected to grow in the future?

      Now, it has approximately 100GB of data, expected to grow to 160-200GB in the next one-two years.

   b. Are there seasonal or cyclical patterns in data usage?

      Usually, data is accumulated, but is not deleted, as it might be useful in other projects. So, the most common pattern of this data usage is the following:

      1) Images uploaded.

2) Images processed and metadata added.

3) Data consumed for neural networks training.

5. Data Access and Retrieval:
   a. How frequently is data accessed and retrieved?

      Some parts of the dataset may be accessed two-three times a day.

   b. Are there specific time-sensitive requirements for data retrieval?

      The main data access and modification timeframes are the following:

- *Monday – 9:00 – 17:00.*
- *Tuesday – 9:00 – 17:00.*
- *Wednesday – 9:00 – 17:00.*
- *Thursday – 9:00 – 17:00.*
- *Friday – 9:00 – 17:00.*

There are also some data access and modifications happen out of this timeframe, but they are not very "heavy" to take them under consideration.

6. Regulatory and Compliance Requirements:
   a. Are there any industry-specific regulations or compliance standards that the company needs to adhere to regarding data storage and retention?

      No such regulations.

7. Recovery Objectives:
   a. What are the recovery time objectives (RTO) for this type of data?

      The data should be restored within 24 hours of disaster occurrence not to severely interrupt the normal business operations.

8. Backup Storage and Infrastructure:
   a. What storage infrastructure is currently in place for backups?

      Currently, there is a server in the office premises which contains both production and backups. The backup is performed once a month manually by a data engineer and stored for six months. The capacity dedicated to the dataset limited to 2TB.

   b. Is there a mix of on-premises and offsite/cloud-based storage solutions?

      No.

9. Data Retention Policies:
   a. Are there specific policies in place regarding the retention of data?

The backups are kept for six months and then being deleted.

b. How is data classified in terms of its retention requirements?

No specific classifications.

10. Backup Testing and Validation:

a. How often are backup and recovery procedures tested?

Approximately once every year.

b. Is there a regular schedule for validating the integrity of backup data?

No.

11. Security Measures:

a. What security measures are in place to protect backup data?

The dataset is accessible only from particular machines and IP addresses.

b. How is encryption applied to ensure the confidentiality of backed-up information?

No data encryption is applied to the dataset.

12. Budget and Resource Allocation:

a. How is the budget allocated for backup and archiving solutions?

The company is ready to allocate 200EUR per month for the backup purposes.

b. Are there any resource constraints or considerations that impact the implementation of backup strategies?

- *The company owns in-premises and an offsite storages, which may also be used for storing the backups. The offsite storage was also limited to 2TBs.*
- *The machine responsible for the operation of the dataset runs on Windows 10 operating system.*
- *Cloud solutions must not be used due to security concerns.*
- *Archiving idea was rejected, because the dataset is constantly growing, and the data is deleted very rarely.*

5 Results and Discussion

5.1 Proposed Backup System

5.1.1 Overview of Proposed Changes

Considering the available options and limitations the following changes proposed:

- Hybrid Backup (fulfils the questions 12b, Budget and Resource Allocation and 8b, Backup Storage and Infrastructure): Because cloud solutions are not an option, a hybrid approach using both on-premises and offsite storage is recommended. This approach utilizes the company's existing storage infrastructure and enhances data protection and recovery capabilities.
- Backup Strategy (fulfils the question 5, Data Access and Retrieval): Given the nature of the data, the following backup schedule and types are proposed:
    - Full backup each beginning of month (e.g. January 1st, February 1st, etc.).
    - Incremental backup two times a week (is relatively frequent and not so storage consuming).
    - All of the backups to conduct at 23:00, as this time is out of the peak of data access and modifications for any day of the week.
- Encryption and Security Implementation (fulfils the question 11b, Security Measures): Implement encryption for backup data to protect sensitive information. This can be done using software-based encryption tools compatible with Windows 10 OS.
- Automated Backup Schedule (fulfils the question 8a, Backup Storage and Infrastructure): Transition from manual monthly backups to automated backups. This could be set to occur weekly or bi-weekly to ensure more recent data is always backed up.
- Data Lifecycle Management (fulfils the question 3, Data Lifecycle): Align the backup strategy with the lifecycle of the data. Regular backups will ensure all stages from unprocessed images to processed data are secured.
- Data Volume Management (fulfils the question 4a, Data Volume and Growth): Given the growth projection, ensure the backup system can scale,

even if current storage seems enough. Regularly monitor storage usage to plan for future expansion needs.

- Implement Veeam Backup & Replication Community Edition (fulfils the question 12a Budget and Resource Allocation): This tool covers all of the flaws of the existing backup plan, such as lack of encryption, manual backups and not utilizing available offsite storage. Furthermore, this tool is free, so it does not require any additional investments, unless additional functionality is required.

5.2 Implementation Plan

5.2.1 Detailed plan for implementing the proposed changes.

1. Infrastructure Setup: Configure the existing on-premises and offsite storage for the hybrid backup system.
2. Backup Software Configuration: Install and configure Veeam Backup & Replication Community Edition.
3. Security Implementation: Implement encryption solutions for the backup data.
4. Automation Setup: Configure the backup software for automated scheduling.

5.2.2 Backup infrastructure

The proposed backup infrastructure consists of two main parts:

- In-premises server machine, which is used for production and the backups. The dedicated capacity for the project is 2TBs.
- Offsite server machine, which was not used for the backups, but it is also possible to dedicate 2TBs of storage for the backup purposes as well.

5.2.3 Data flow

In the solution proposed for the project, the data flow focuses on the collection of the images used for training the neural networks. Also, it is worth mentioning that this collection of the images constantly grows and the data within is rarely deleted and modified after the metadata has been added and images have been processed properly.

As the first step of the data flow, the unprocessed images are added to the dataset. The images that are collected don't have any metadata or corrections added. These images are used as a basis for the next step, which is processing and adding the metadata.

After the images have been collected and inserted into the dataset, the appropriate metadata is added, and the corrections are made (if required). This step is extremely important because it adds the information about the tags, categories, descriptive attributes, etc. to proceed with the training processes.

Once the images are processed and filled with appropriate metadata, they are being used for neural networks training. This step is the actual use of the dataset, where the tagged and processed images are inserted into the machine learning algorithms for the further development of the company's products.

The backup is made independently from the data flow of the production, so, both raw images and the images which are tagged and processed are the subject of the backup. Also, when the backup is performed, the data is encrypted and saved (mirrored) into both in-premises server and an offsite server machines.

5.2.4 Disaster recovery plan

There are three most possible disaster scenarios are considered:

I. Production data is corrupted or destroyed.

II. In-premises backup is corrupted or destroyed, but the offsite backup and production are in the normal state.

III. Offsite backup is corrupted or destroyed, but the in-premises backup and production are in the normal state.

For the scenario "I" the recovery plan is the following:

1. Attempt to perform recovery from the in-premises server machine, as it is the fastest way to resume the normal workflow.
2. If the in-premises server machine backups fail as well, perform a recovery from the offsite server.

For the scenario "II" the recovery plan is the following:

1. Copy the backup data from the offsite server to the in-premises server.

For the scenario “III” the recovery plan is the following:

1. Copy the backup data from the in-premises server to the offsite server machine.

5.2.5 Disaster recovery testing

For the disaster recovery testing the simulation tests were chosen, as they are modelling the actual disaster, but they don’t interrupt the normal functioning of the production dataset.

Regarding the schedule of the disaster recovery testing, it was decided to perform them once a year, at the end of the month the testing is planned on. It is decided so for not interrupting the full backup processes, which have been proposed to perform each start of the month.

5.2.6 Price of suggested solution

The total price of the solution is considered as free of charge, due to the software solution suggested is free and that the company owns the offsite server, which is planned to use for the backup as well. However, there might be some additional hidden costs incurred, such as:

- Payments for additional software functionality (the free version of Veeam Backup & Replication Community Edition covers all the issues as of now and will cover some reasonable expansions, but if the company needs will increase beyond that, the additional functionality should be purchased separately).
- Increased cost of the offsite server (because more data will be stored on the server and, consequently, more bandwidth will be required, the server maintenance cost may be higher).

5.2.7 Timeline and milestones for the implementation.

Considering the fact that it is not possible to exclude the dataset from the production, and it needs to constantly continue being available for data adding, refining and consuming, the following implementation schedule proposed:

- Month 1-2: Infrastructure setup and software installation.
- Month 3: Testing and final adjustments.
- Month 4: Go-live with the new backup system.

6 Conclusion

In this thesis, the research of the backup solution for the business project was carried out and recommendations for improving the current backup solution were proposed. The analysis of an existing backup solution revealed its flaws, the main of which were the lack of backup process automatization, absence of the encryption of the existing backups, and no offsite backup utilization.

The key implications of this work have shown some interesting findings, among which that the project has been heavily reliant on the manual "copy-paste" approach, which in turn, is a potential failure point. Also, despite the data under backup is critical for the business and relatively expensive to fetch, the backups were not encrypted.

Limitations. This research was limited to one particular business project, which has some relatively unique limitations, because of the data value and sensitivity. Another thing, worth mentioning is the fact that it is relatively hard to make the maximally objective comparison of the backup software solutions, as there are many different solutions from simple OS out-of-the-box utilities to sophisticated solutions, which may automate most of the workloads.

Research Process. The research started from literature and other sources review, which helped to understand the general matter and get the information about the best practices to work with. After that, the interview was conducted to get the practical information from the project, and also, receive the business requirements. Then, the analysis of the existing solution has shown the space for improvements and potential fixes. The process wraps up by proposing the solution for making the backup process better under the given requirements and limitations.

Concluding Thoughts. The work has shown that the backup and archiving process is not only storing the data somewhere else, but also includes some sophisticated processes and details, good to be aware of. Furthermore, when the theory about the matter meets the real cases, with its limitations, requirements, and scarce available resources, the solution becomes not quite obvious. Apart from that, poor backup solution may lead to business ineffectiveness, such as redundant manual work, which may be the weak point.

7 References

1. **Dave, Henry.** 6 Reasons You Must Have Data Backup and Recovery For Your Company. *Veritas.* [Online] Veritas, 28 June 2022. [Cited: 20 August 2023.] https://www.veritas.com/blogs/6-reasons-you-must-have-data-backup-and-recovery-for-your-company.

2. **Curtis-Preston.** *Backup and Recovery.* Sebastopol, CA : O'Reilly Media, 2007. pp. 27-33. ISBN: 978-0-596-10246-3.

3. **Helen.** Onsite vs Offsite Backup Reviews, Strategies, and Best Practice. *MiniTool Software.* [Online] 5 November 2021. [Cited: 16 October 2023.] https://www.minitool.com/news/offsite-onsite-backup.html.

4. **Yasar, Posey.** What is Off-site Backup? *TechTarget.* [Online] TechTarget, 01 August 2023. [Cited: 13 October 2023.] https://www.acronis.com/en-us/blog/posts/cloud-vs-local-backup/.

5. **G2.** *Compare Acronis Cyber Protect Cloud vs. NinjaOne (formerly NinjaRMM) vs. Veeam Data Platform vs. Backup Exec | G2.* [Online] G2. [Cited: 04 October 2023.] https://www.g2.com/compare/acronis-cyber-protect-cloud-vs-ninjaone-formerly-ninjarmm-vs-veeam-data-platform-vs-veritas-backup-exec.

6. What is RAID? RAID 0, 1, 5, & 10 Made Easy. *HelloTech.* [Online] HelloTech, 14 January 2020. [Cited: 04 October 2023.]

7. Team, NAKIVO. 3-2-1 Backup Rule: A Guide to Efficient Data Protection. *#1 Backup and Recovery Solution | NAKIVO.* [Online] NAKIVO, 01 June 2023. [Cited: 30 11 2023.] https://www.nakivo.com/blog/3-2-1-backup-rule-efficient-data-protection-strategy/.

8. Ivy. Grandfather-Father-Son Backup Scheme - Safest Backup Strategy. *AOMEI Backupper.* [Online] AOMEI Backupper, 09 November 2022. [Cited: 02 October 2023.] https://www.ubackup.com/articles/grandfather-father-son-backup-5740.html.

9. Andrew, Abwoga. The Grandfather-Father-Son (GFS) Backup Strategy. *Backup Ninja.* [Online] Backup Ninja, 03 December 2020. [Cited: 04 October 2023.] https://backup.ninja/news/grandfatherfatherson-gfs-backup-strategy.

10. Data Recovery Explained. *IBM.* [Online] IBM. [Cited: 01 October 2023.] https://www.ibm.com/topics/data-recovery.

11. Technologies, Waterford. The Key Difference Between Data Backup and Data Archiving. *Waterford Technologies.* [Online] Waterford Technologies, 20 June 2021.

[Cited: 24 October 2023.] https://waterfordtechnologies.com/difference-data-backup-archiving/.

12. Preston, W. Curtis. *Backup & Recovery. Inexpensive Backup Solutions for Open Systems.* Sebastopol, CA : O'Reilly, 2006. ISBN: 978-0-596-10246-3.

13. Preston. Backup vs. archive: Why it's important to know the difference. *Network World | From the data center to the edge.* [Online] Network World, 09 July 2018. [Cited: 26 October 2023.] https://www.networkworld.com/article/965986/backup-vs-archive-why-its-important-to-know-the-difference.html.

14. Little, David B. and Chapa, David A. ***Im**plementing Backup and Recovery: The Readiness Guide for the Enterprise.* Indianapolis : Wiley Publishing, Inc., 2003. ISBN: 0471227145.

15. Coronel, Carlos, Morris, Steven and Rob, Peter. ***D**atabase Systems: Design, Implementation, and Management, Ninth Edition.* Boston : Cengage Learning, 2010. ISBN: 978-0-538-74884-1.

16. *Implementation of Data Backup and Synchronization Based on Identity Column Real Time Data Warehouse.* Adnyana, I Gede and Jendra Sulastra, I Made Dwi. 1, Denpasar, Indonesia : Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 2020, Vol. 11. ISSN 2541-5832.

17. Backup Strategy Best Practices Organizations Should Follow. *VM Backup Solution for VMware, Hyper-V, EC2 | NAKIVO.* [Online] NAKIVO, 15 August 2022. [Cited: 01 May 2023.] https://www.nakivo.com/blog/how-to-build-a-backup-strategy-full-guide/.

18. Selecting Your RAID: Breaking Down Popular Levels. *Protek-IT.* [Online] Protek-IT, 08 June 2020. [Cited: 06 October 2023.] https://teamprotek-it.com/blog/most-popular-raid-levels-explained/**.**