DATA DEDUPLICATION:
IMPORTANCE, BENEFITS AND CRITERIA FOR SELECTION

RAIN.TECH
Custom Cloud & Managed IT

CONTENTS

WHAT IS DATA CLEANSING?

Experts estimate that humans produce 2.5 quintillion bytes of data every day.[1] That includes a lot of duplicate, inaccurate and outdated information. Data cleansing is the process of identifying and rectifying such corrupt or flawed data from a data set, table or database. It helps you substitute, alter or delete dirty data.

Data cleansing includes five elements — data standardization, data validation, data deduplication, data analysis and quality check.

Data Standardization

Most businesses use data from multiple sources such as data warehouses, cloud storage and databases. But data from distinct sources may not be in a consistent format, which could lead to trouble down the line. This is where data standardization helps. It's the process of converting data into a consistent format.

Data Deduplication

Data deduplication refers to the process of eliminating duplicate data in a data set by deleting an additional copy of a file and leaving just a single copy to be stored.

Data Normalization

This is the process of organizing data within a database. It involves making data tables and identifying relationships between those tables based on the rules designed to reduce data redundancy and improve data integrity.

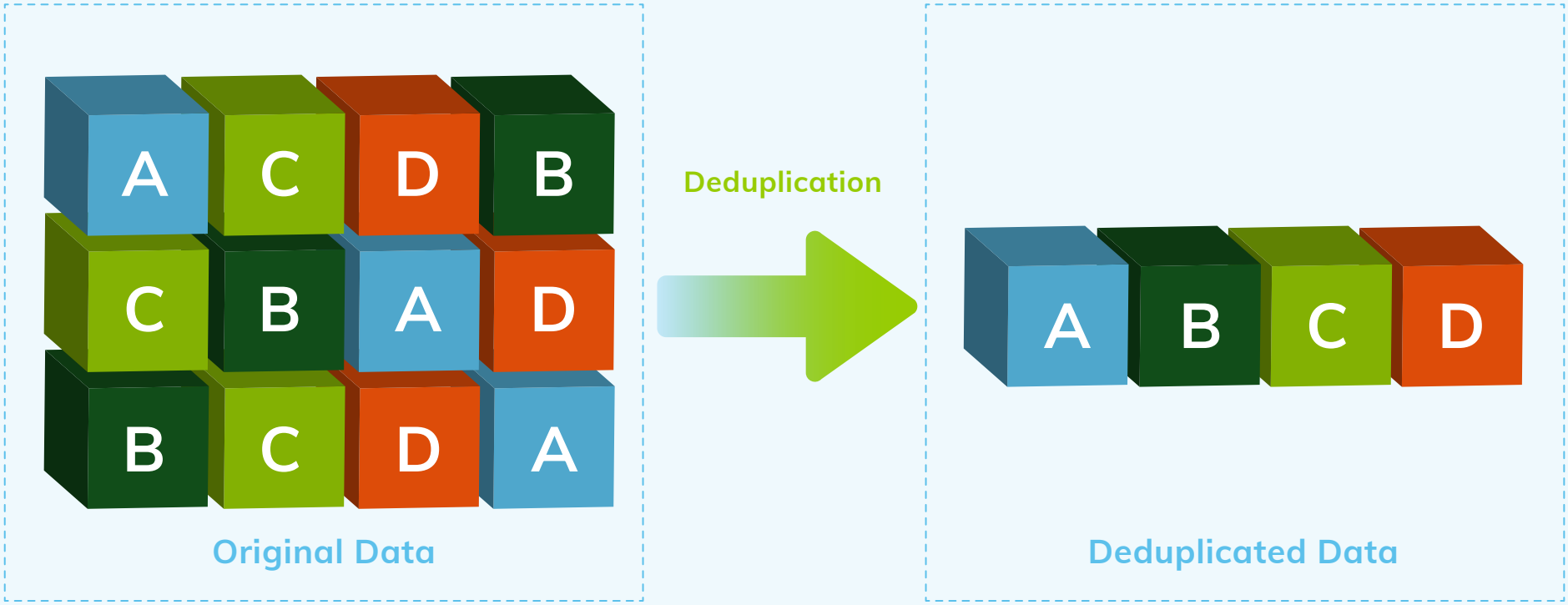
Quality Check

Businesses need good quality data to make the right decisions because if dirty data influences decision making, it can lead to costly mistakes. Therefore, quality checks are essential for every business dealing with considerable amounts of data.

Data Analysis

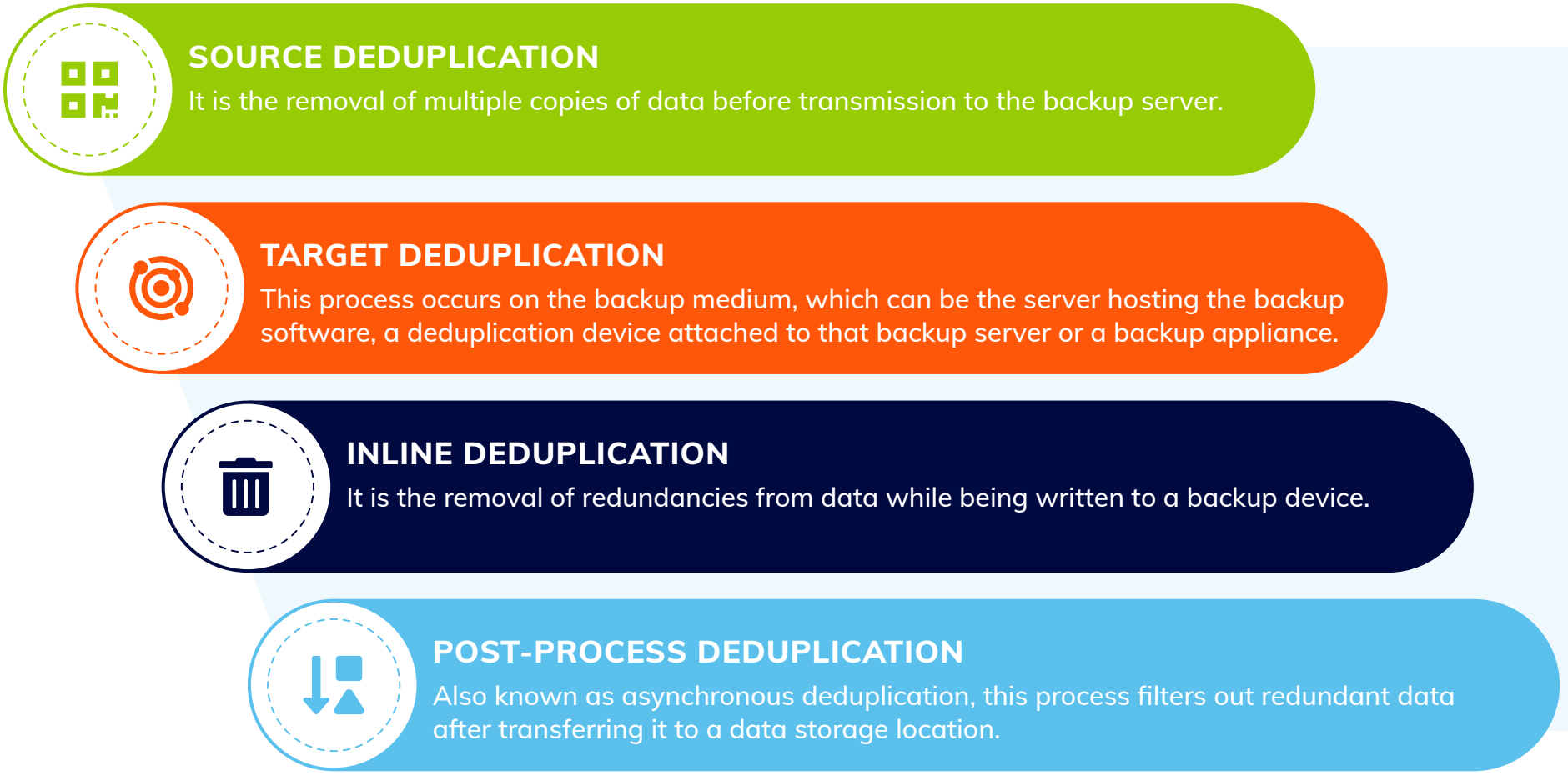
Data analysis is the process of analyzing data using logical and analytical reasoning to get valuable insights. The derived information helps make sensible decisions.

In data deduplication, data gets divided into several blocks that are compared with each other. Each block is assigned a unique hash code. If the hash code of one block matches the hash code of another, it is considered a duplicate copy and gets deleted. This ensures that only a unique copy of the data is stored. Deduplication can detect redundant copies of data across data types, directories, servers and locations.



COMMON TYPES OF DATA DEDUPLICATION

Some of the popular deduplication techniques are:



WHY IS DATA DEDUPLICATION IMPORTANT?

The storage capacity for most small and medium-sized businesses (SMBs) is often limited, but the amount of data generated, transferred and stored steadily grows over time. When this contains duplicate copies of files, it can create a serious storage issue.

The process of data deduplication helps tackle this issue by:

- **Reducing storage space requirements by storing only a single copy of a file**
- **Minimizing the network load since less data is transferred, thus leaving more bandwidth for other tasks**

In other words, by deleting duplicate copies, deduplication optimizes storage capacity, boosts on-appliance retention and minimizes the landing zone space necessary for backups and the number of bytes transferred between endpoints.

Duplicate data is something every business deals with. If neglected, this data accumulates over time and reduces valuable storage space, leading to wastage of resources. What's more? Excessive amounts of duplicate data can cause poor data quality and inaccurate analytics. Therefore, SMBs should embrace data deduplication as a best practice.

But don't confuse data deduplication with data compression.

Although some online forums use them interchangeably, their meanings are different. While deduplication removes duplicate data without compromising vital information, data compression restructures data and reduces data size. Also, compression works with a file or a set of files while data deduplication operates with data at the block level and deletes redundant blocks.

BENEFITS OF DATA DEDUPLICATION

If you don't have to restore extra copies/versions of a file, recovery times after a data loss incident can be much faster.

Faster Recovery

Would you pay for three LinkedIn premium accounts just for yourself? Probably not. The same concept applies to data deduplication. Having a business-class backup solution in place is good. However, if employees are saving multiple versions of the same file all over your network, it will likely cost you a lot of extra storage just to house what is essentially the same information.

Saves on Storage Costs

Supports Productivity

Organizations collectively lose thousands of productive hours per year simply because employees cannot quickly locate the information needed to carry out tasks.

Minimizes Version Control Issues

If one file is individually downloaded and saved across multiple endpoints, in the end, you will run into version control issues in which the asset's accuracy may come into question. This is a serious problem, especially if pricing data or other sensitive information is involved. This could, in turn, affect your business reputation and company revenue.

Enables Collaboration

Housing information centrally via collaboration tools like SharePoint or Google Docs, and equipping employees to provide inputs using these tools, can eliminate the challenge of accurately aggregating inputs on the back end.

5 DATA DEDUPLICATION BEST PRACTICES

1 IDENTIFY THE BEST-SUITED DEDUPLICATION TYPE

Although different deduplication techniques remove duplicate files by identifying patterns within chunks of data, they each perform differently. While selecting one that suits your business, consider factors like cost and storage requirements. Always opt for a deduplication type that makes sense for your business. Avoid selecting one solely based on what your competitor is doing or what an online forum suggests. If needed, consult an expert.

2 SORT FILES BY DATA TYPE

Deduplication might not be very effective with some media files such as MP4 and JPEG, so remember to sort the data types that you handle. Otherwise, deduplication efficiency might be affected, and you could end up being disappointed with the results.

3 DO NOT FOCUS ON REDUCTION RATES

If someone promises that they can help reduce your data size by 50%, 80%, etc., do not blindly accept it. Actual reduction rates will depend on the type of backup, type of data and frequency of change in the data. Therefore, make sure your expectations are based on facts.

4 DECIDE DEDUPLICATION LOCATIONS

You need not deploy a deduplication solution on every storage media since it will not be cost-effective. In most cases, only secondary locations like backup, where cost is a concern, need deduplication. Plus, deployment of deduplication in primary storage like data centers can affect storage performance.

5 CONSIDER ALL EXPENSES

To avoid being surprised by hidden costs at a later stage, you must consider the full range of expenses needed for deduplication, i.e., remember to consider factors such as maintenance and management costs along with the cost of physical storage.

CRITERIA FOR SELECTING A DATA DEDUPLICATION SOLUTION

Consider the following criteria when evaluating a deduplication solution:

1 Compatibility

A good deduplication solution must be non-disruptive.
It must:

- Integrate with a business' existing backup environment
- Have flexible implementation options to provide coverage across multiple sites — on-premise, remote and the cloud.

2 Scalability

Scalability is a vital aspect that you must consider for a deduplication solution since the amount of data handled will increase when the business grows. The best deduplication solution must be flexible enough for initial implementation and future growth. Upgrades must be available according to your increasing data volumes.

3 Restore Performance

Some solutions are excellent at deduplicating data but might take considerable time to restore an extensive database. Select a solution with reasonable restore speeds.

4 Availability

Since a deduplication solution merges a lot of data in one place, the probability of data loss is high. But access to deduplicated data is vital and must not be susceptible to failure. Hence, select a deduplication solution with features such as mirroring to safeguard against local storage failure and replication to protect against disaster.

5 Efficiency

Efficiency must be a prime consideration while selecting a solution to make the purchase economically justifiable. For a deduplication solution to be truly cost-effective, it must have robust duplicate detection capabilities.

TRAINING/PROCESS DOCUMENTATION IS VITAL FOR DEDUPLICATION

For the successful implementation of data deduplication, you must provide employees with the necessary training and process documentation. Employees must know the best practices for data hygiene and avoiding dirty data.

Regular and comprehensive training aims at improving the awareness, skills and knowledge of employees. This will have a positive influence on the way employees do their job. Eventually, it will contribute in a big way to your business' growth and success.

Close to 30% of employees believe they do not get adequate training from their employer while about half believe that their company could have benefitted greatly if there was proper training.[2]

Training is usually carried out at three levels:

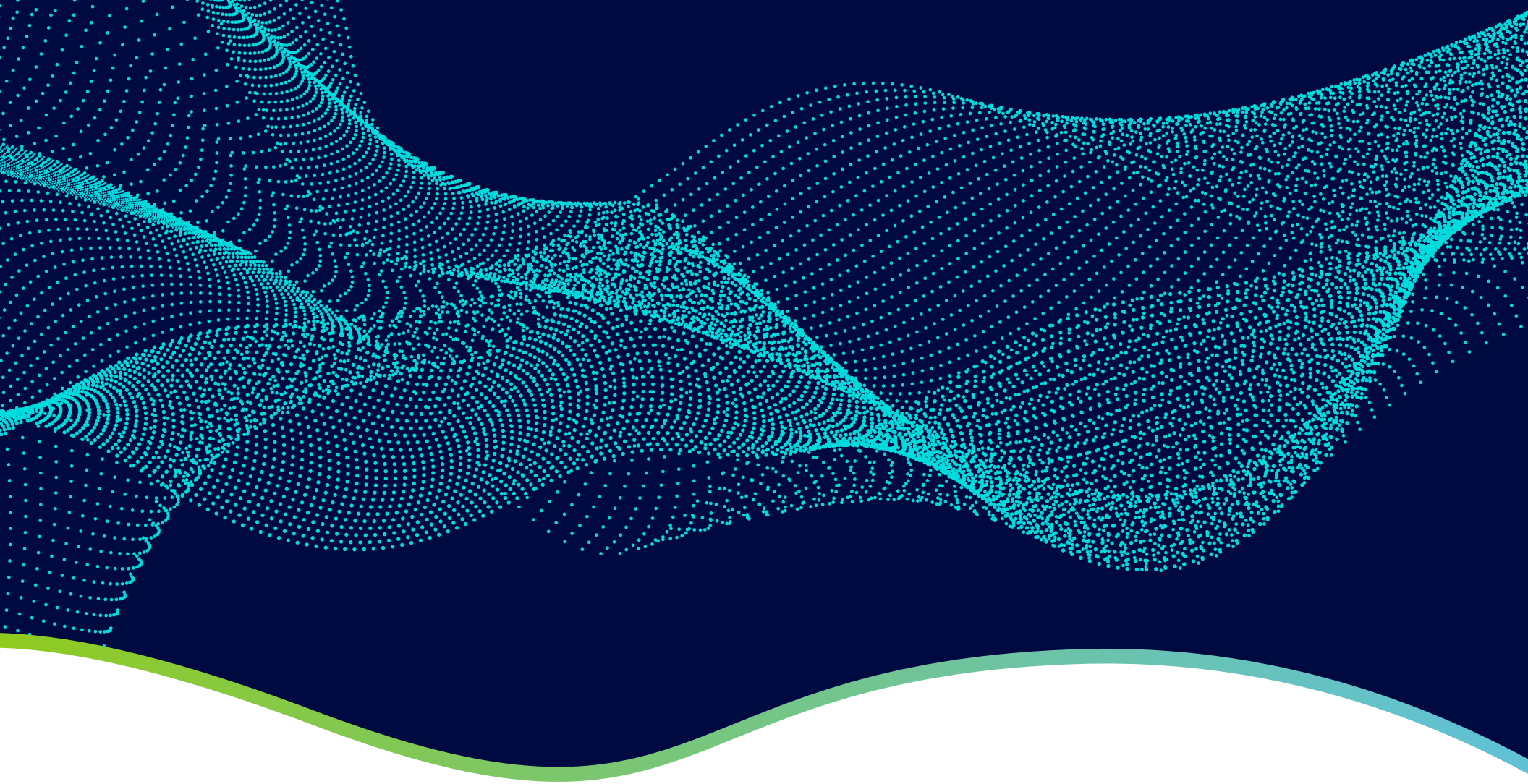
1. Train a new employee when they join the organization to familiarize them with deduplication in general as well as other do's and don'ts.
2. All current employees must get refresher sessions to ensure they do not forget the foundational lessons.
3. Provide training to update employees about every change to the deduplication solution and policies.

A deduplication process document is internal, ongoing documentation that focuses on how to implement the process safely and successfully. A business comprises multiple interrelated processes, which if not documented, can affect daily operations.

PARTNER FOR SUCCESS

Data deduplication is essential to reducing the volume of data transferred across a network and saving a substantial amount of money in terms of storage costs and backup speed. Get started on your path to data deduplication with an experienced partner like us. Knowing that the process is in safe hands not only gives you peace of mind but also frees you up to focus on growing your business.

Get in touch with us to learn how we help organizations enact the best data deduplication policies that save storage and enable faster recovery after a data-loss incident.



Sources:

1. Techjury.net
2. Workplace Knowledge and Productivity Report