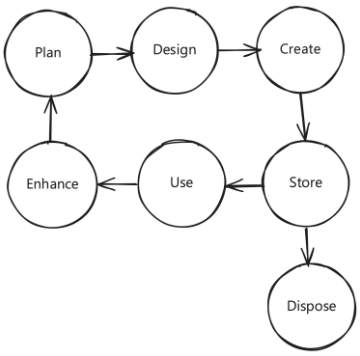
Coalitie Duurzame Digitalisering

Sustainable data management: Implementing a Sustainable Data Strategy Across the Data Lifecycle

EXECUTIVE SUMMARY

The increasing digitization of our world has led to an exponential growth in data creation, storage, and processing. With this growth comes a significant environmental impact due to the energy consumption and cooling requirements of data centers, as well as the raw materials required for storage solutions. Sustainable data management has therefore emerged as a critical strategy for organizations to address the environmental footprint of their data lifecycle.

Sustainable data management is not just a technical or operational issue; it is a strategic imperative at the executive level as well as a cultural mind shift.



This paper uses the DMBoK phases to address the sustainability considerations in all the phases of your data lifecycle, from planning, creation until the data is destroyed. We address the following:

Planning

- Create awareness in your organisation about the sustainability impact of data storage and usage.
- Plan to avoid copies of data in your organisation and use one single version of the truth as a central source of information
- Plan to use the best technology for processing data.

Design:

- Use design patterns for optimised data management:
  - Compress your data
  - Avoiding data duplication in your design
  - Design to use a tiered storage solution
  - Consider the usage of a zero-copy cloning if virtualisation or access to the source cannot be used
- Make sure that you design / include the metadata
- Include data disposal in your design (use cases)

Create:

- If you can create data in a structured way (XML, JSON) instead of unstructured that is preferred

Store:

- Implement the best mechanism to store and retrieve data. E.g. RDMS
- Select the different storage tiers and implement tiered storage
- Store the data in a compressed format (data at rest)
- The place where you store your data and the products that you use should be evaluated from the perspective of sustainability.

Use:

- Compress the data while it is moved to be processed and when processing tool place
- Be laser focused and only retrieve those data that you need, both in number of documents and rows as well as on the content.
- Right level of utilisation: when and where do you need it?
- Use green coding practices to process data
- Use optimised search algorithms for data processing

Enhance:

- If you update your data, make sure that you delete the old version
- If you must keep history, store the delta's only
- Apply retention policies, both for the new as well as the old version
- Aggregate data if you do no longer need the details

Dispose Data:

- Set the retention policies on the data
- Run the process to clean your data
- Comply and report on your data retention standards

We do understand that any of those measures is situational, depends on your context and is not complete. The purpose of this "checklist" is to achieve sustainable data management by implementing a sustainable data strategy across the lifecycle of your data.

INTRODUCTION

In this article, we focus on the ***environmental sustainability*** aspects within the lifecycle of data.

Even though it may look obvious that companies should have a data strategy, many companies don't. They do realize that data are their critical assets and a data strategy that includes a strategy to manage the lifecycle of their data is important. But creating a good data strategy across all departments is a cumbersome task. Legislation like GDPR and the EU Data Act is definitely driving the importance of establishing a data lifecycle to actively manage data. Next to the enormous potential of (Gen)AI that is also fully relying on high quality data.

Why is this important and why now?

Why is Sustainable Data Management important. The reason it is so relevant and important is multifaceted:

1. **Regulatory Compliance and Reporting**: Legislation like the GDPR and the EU Data Act has heightened the importance of data lifecycle management. Organizations must actively manage their data to comply with these regulations and to report on their Environmental, Social, and Governance (ESG) commitments.

2. **Cost Efficiency:** By adopting sustainable data management practices, companies can significantly reduce costs associated with data storage, processing, and transport. This is not only beneficial for the environment but also for the bottom line.

3. **Reputation and Competitive Advantage**: Companies that demonstrate a commitment to sustainability can enhance their brand reputation and gain a competitive edge in the market. Customers and stakeholders are increasingly valuing environmental responsibility.

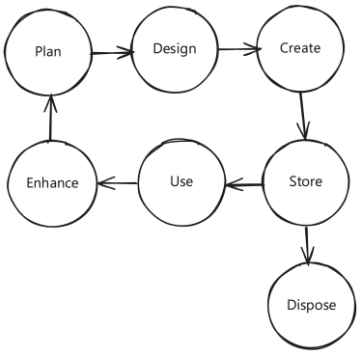
As the volume of data continues to grow, the energy, cooling (carbon emissions) and raw materials required to store and process this data will also increase. According to NetApp, the quantity of global GHG emissions from data centers (2.5%) are greater than that of the global airline industry [11]. They also state in their paper on the Waste Index that most of the data that is stored never will be used (>67%) [11]. This contributes to unnecessary energy consumption and carbon emissions.

*With **environmental sustainability** we mean the usage of energy (carbon emission) and the aspect of circularity, the reduction of waste.*

In this paper we do address another reason to have your data lifecycle management in place, namely ***environmental sustainability***. Data storage, processing and transport consumes a lot of energy, and the devices handling this do create a lot of waste. There are many reasons to establish a data strategy, including data lifecycle management. This paper is there to assist with the "environmental" aspects of your data lifecycle. And when you put your strategy in place, don't forget that you need to report it in your yearly ESG report.

We do not want to invent another framework to describe the environmental sustainability aspects, therefor as a starting point we use the DMBoK (Data Management Body of Knowledge)[1] data lifecycle management phases.

We discuss the data lifecycle stages in summary from the perspective of sustainability. More details about the phases can be found in publications about DMBoK, for now a simple summary:



- **Plan:** The sustainability aspects that are relevant and should be considered in the project.
- **Design** (& Enable)**:** The sustainable design decisions that should be made based on the sustainability strategy.
- **Create**/Obtain**:** Define the most sustainable way to create or obtain data from other sources.
- **Store**/Maintain**:** Define the most sustainable way to store the data.
- **Enhance:** Define the most sustainable way to update and transform the data.
- **Dispose of:** Make sure that data will be disposed in a sustainable way.

THINK!

Both plan and design phase are part of the solution definition and therefore these are combined in this point of view. The other phases will focus on the actual implementation. The first two phases are about **Think!** The other five focus on: **Do!**

PLAN

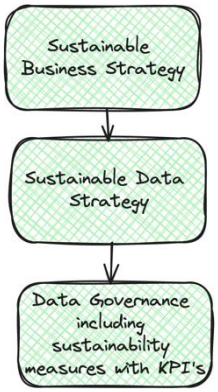
The *plan phase* defines the business objectives and requirements to direct the activities across the data lifecycle. Within the *design phase* those objectives and requirements are "translated" to principles and decisions.

SUSTAINABLE DATA MANAGEMENT

Sustainability by Design

Our overarching-principle is '*Sustainability by Design'*. How would that translate into sustainable data requirements?

To scope this viewpoint the assumption is that a sustainable data management is in place. Even though some of the principles and decisions that are described can be applied bottom up, it is preferable to have a sustainable data management capability in place that is enabled and governed via overall data management practice within the organization.



A sustainable business strategy leads to a sustainable data strategy which result in a sustainable data governance and framework that will be used to calculate and measure the value of sustainable data management. For this viewpoint, the framework with metrics is assumed to be there and you have a way to govern it.

The annual electricity report from the International Energy Agency of 2024[5] says data centers consumed 460TWh in 2022. This figure could rise to 1,050TWh by 2026 in a worst-case scenario due to developments like AI. This increase is equivalent to adding the entire power consumption of a country like Germany.

AWARENESS

The demand for compute is still growing at an enormous pace. The first thing to do is to ensure that people are aware of the energy consumption involved in data *processing* and *storage*. The energy consumption of data processing and storage should be added as an important requirement within IT projects. It is important that the deployed solutions are sustainable from an environmental perspective which go hand in hand with the economic perspective in many cases. But also, the *creation of the solution itself* requires attention. The development of an application, compilations, (automated) testing do cost energy. Especially when training (Gen)AI models, this could be an energy intensive process.

According to OpenAI researchers, since 2012, the amount of computing power required to train cutting-edge AI models has doubled every 3.4 months. The energy cost by training GPT-4 equivalents of driving a gasoline car for nearly 18 million miles or the equivalents of powering more than 1300 homes for one year.[8, 9]

In this viewpoint we address principles and decisions that could be made during the whole data lifecycle to limit the power consumption of the implemented data solutions from an end-to-end perspective.

A good way to create awareness are examples. We address several of those in our viewpoint. In many cases when talking about numbers, different sources have different numbers which could be argued about. However, the

message is clear: We must implement a sustainable data strategy across the data life cycle to keep the environment impact in control.

Single Version of the Truth

There is a tendency to copy data into your own domain. From that moment you have full control over it, a nice feeling. However, is it really necessary? It is important to limit the number of copies of data. This limits the total amount of storage that is required and lowers the risk of data quality issues due to inconsistencies. There are multiple possible measures to limit the need for data duplication. The following measures are the most common ones:

- *Normalized databases design* techniques that organizes data in a manner that reduces redundancy of data
- *Data virtualization technologies* that enable data sharing without the need to create copies
- *Centralized data platform* like a data warehouse that create a single view on the data within an organization. This platform reduces also the need to store history within the operational systems which results in more efficient systems.

Service oriented architecture (SOA) or a micro services-based architecture promotes the reuse of services across different applications, which can help reduce data duplication.

Normalized database design

Normalized databases are efficient for data storage and store data once. The counter site is that sometimes complex queries are required to collect the necessary data. In the paragraph *Use optimized storage mechanism* in the Store phase different mechanisms are discussed.

Example: For the actual sales results, you might access a service that aggregates the results at the source location and return actual results, while a reference table of zip-codes may be one that you copy locally. It is used often, and it changes rarely.

Data Virtualization & Processing at the source

Current data virtualization technologies are providing reliable access to remote data. The advantage is also that the data you access is actual. You must balance between the number of times the data is accessed, the amount of data that is accessed and transported, the frequency that the copy must be refreshed, etc.

Centralized Data Platform

Data warehouse, data lake, data mesh or data fabric what is the most sustainable? Except for a data mesh, the data platform architectures have in common that they are based on a centralized structure.

The data mesh is based on the domain driven software development principles and introduces the concept of data products that are domain specific. A clear definition of the data products is very important to ensure an efficient data mesh set up without a lot of data duplication. The architecture embraces the use of domain specific data definitions. The definition of a customer can be different from a marketing perspective than from a finance perspective. The same customers will also be available in multiple products. This is not really different than in the more centralized architectures like a data warehouse. A data warehouse within the most organizations consists of a backroom (historical integration layer) and a front room (end user layer with

the data marts). The centralized platforms strive to a single version view of the enterprise with universal definitions but at the same time copy the data to use it for multiple purposes, like to create end user data sets based on specific business rules with specific optimizations to increase query performance.

Centralizing data for BI and Analytics is a sustainable practice, however it comes with the cost of good

All data platform architectures have the risk of a lot of waste due to bad management. It is very important to have strong monitoring in place (incl. FinOps, utilization and end user usage). This to ensure that all the stored data has a purpose and is processed in an efficient way. We don't see big differences from a sustainability perspective regarding to the centralized data platforms. We advise to look at the specific situation to choose for the architecture with the best fit.

Centralizing the data and storing it in a central repository to be accessed there is a good practice. At the beginning of this point of view we discussed the difficulty of a data strategy. These concepts require good data governance, not an easy thing to do either.

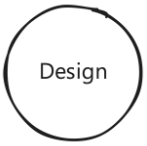
Master data management is a concept that centralizes the storage and access of your master data. The driver is most of the times to make sure that there is only one source of truth, but it definitely has a sustainability advantage.

Fit for Purpose data processing

Ensure that there is no waste in the processing of data. The processed datasets should be as small as possible, and the data should be of good quality.

Use database queries for instance that only select specific data (select specific columns and datasets). This limits the required resources for processing and network transport.

The quality of the data is very important to deliver data products that deliver value. If the data quality (complete, accurate and consistent) is to low data products will not be used which leads to waste. Low data quality will also lead to the need of additional computing resources to check and correct the data (if possible).



DESIGN PHASE

Within the *design phase* objectives and requirements are “translated” to principles and decisions.

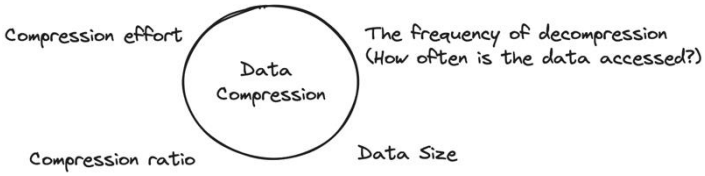
USE THE SERVICE WITH THE BEST FIT FOR THE IMPLEMENTED USE CASE.

In the design process we need to consider the best approach for storing and accessing the data. In many cases it will be a trade-off between different aspects. For example, security may imply additional measures that cost energy or performance will require latest processors which have impact on circularity. In architectural terms these are called opposing factors. To make a trade-off between these factors, each factor can be assigned a weight. This way you can weight opposing factors. These factors also include architectural principles and business requirements. Where architectural principles are hardly negotiable, resolving the requirement usually have some negotiation space. An architectural decision is the typical tool to support the argumentation and documentation of a decision. The *implementation* of these decisions take place during the subsequent phases. The following Storage Efficiency aspects should be considered:

- Data compression
- Data deduplication
- Storage tiering
- Data storage and retrieval services
- Zero copy cloning

Aspects that should be balanced are depicted in the following pictures:

DATA COMPRESSION



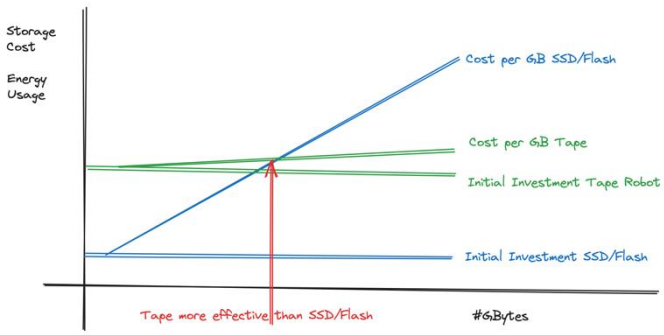
Data compression reduces the amount of physical storage required to store your data, but it requires CPU cycles to perform the compression and decompression. Dependent on the type of data, the compression could be significant or almost irrelevant. A text file can be significantly compressed while a .PDF file is already compressed. If you compress all your data and 95% is already compressed, you burn unnecessary CPU cycles.

DATA DEDUPLICATION

We have the intend to copy data instead of using pointers or links. We would like to bring the data under our control. Think about an email with an attachment. Why not a pointer to a shared repository. This is also a good practice from a GDPR perspective to ensure that data can be deleted of required. Within the boundaries of a company that is usually very well possible. It becomes a security issue when giving access to external users.

The best solution is to resolve it on an application level, but if that fails, there is always deduplication on storage level.

Storage Tiering



Storage tiering is traditionally based on TCO. Fast storage like flash and SSD is relatively expensive compared to spinning disks. More smaller disks are faster, more expensive and cost more energy as well as materials. The perspective of the storage tiering concept should no longer be based on TCO but energy consumption and material usage. This means that spinning disk should no longer be used. In order to achieve acceptable cost, modern tape technology is one of the alternatives as second tier storage. And if that is used as second tier storage, it may as well be used for (air-gapped) backups. More details on storage tiering are discussed in the chapter about data storing.

Data storage and retrieval services

There are different ways of storing and retrieving data. Think about the following: flat files, relational databases, hierarchical databases, (No)SQL databases. It requires a study on its own to decide what the most sustainable solution is for a particular workload. Probably the best measure to consider is performance. The more performant it is, the more sustainable. Of course, you must baseline the infrastructure and you need to consider the overhead. You also have to consider if you do many “write” activities or mainly “read”. When you compare a SQL database running 64 cores with a MongoDB running .5 core, that is probably not a good comparison. There are many aspects to consider. Some of those we will discuss more extensively in the section about data usage. More about the comparisons of different workloads vs services can be found in the section on data storage.

Zero copy cloning technologies

Zero copy cloning is a technology that is mostly used in the cloud data warehouse databases space. There are also storage technologies that have a cloning option based on this technology. A zero-copy clone creates a replica of the data by only copying the metadata, while still referencing the same physical tables or files. This enables the creation of clones in just a few seconds by only copying metadata. The data itself will only change when a write operation is performed on the clone. In this case the copied metadata of the copied data will be repointed to the impacted data by the write operation. This technology reduces the need for storage in case data cloning is required to support DTAP acceptance environments for instance.

Be clean

Example: The Dutch tax legislation [6] requires that administrative records need to be stored for 7 years. Records of real estate and rights to real estate must be kept for 10

Ensure clear and sustainable data retention policies (external and internal) that are actively communicated to ensure data disposal as much as possible. This saves storage and limits the need for processing capacity. We should only store data that has value for future use or that should be stored due to external regulations. A strong data classification scheme connected to the data assets is an important prerequisite for this. All data should be labeled and stored in a data catalog. This catalog is a prerequisite to enable effective data management.

According to the European General Data Protection Regulation (GDPR) for instance, personal data should only be collected for specified, explicit, and legitimate purposes. Customer data may be stored for as long as necessary to fulfill the purposes for which it was collected. Once the original purpose has been fulfilled, data should be deleted.

Example: An insurer filed a claim with the insurance company. The insurer sends a dashcam recording as evidence. The dashcam recording must contain metadata that relates to the claim. The moment the claim is disposed also the corresponding data, like this dashcam recording must be disposed.

ENSURING THE AVAILABILITY OF METADATA TO ALLOW DISPOSAL OF THE DATA

To keep your environment clean it is necessary that the right metadata is available on the data to classify it for the retention process. The metadata that must be made available to have the document destroyed depends on your disposal process. It could be very simple by adding a deletion data or it could be more complex where a combination of creation data, last updated, type of document is required. In situation where compliance becomes important, the implementation of a records management solution is required. Since records management relates to privacy and not to environmental sustainability, we do not discuss this further.

DO!

Now we have defined a comprehensive set of measures (principles) to run a sustainable data project. But what does this mean for the implementation? In the following phases a list of practices has been described.

CREATE

The create step is the first step of the realization. Based on our plan and our design we will start working with the data that can come from different sources. There are different types of data sources:

65% of the data is used for once or less [7]

1. You create data from nothing. Newly entered information, for example a new request for applying a Visa.
2. You create data from existing data (enriching).
3. You create data from non-digital information. You have a form, use OCR to read the form and digitize it.
4. You use existing data. This could be by creating a copy or provide access.

There are a couple of principles that should be applied:

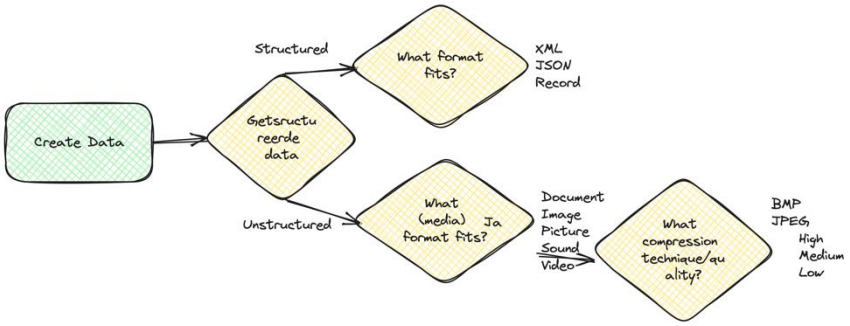
USE STRUCTURED DATA OVER UNSTRUCTURED.

Example: You made a video recording of an interview and used AI to create a transcript of this recording. Is there a need to keep the recording or would the transcript or even a summary be sufficient?

Using structured data is for more efficient than unstructured data. If you capture data, do that in a structured way and store it in a structured way. That could be a database record or an XML file. You may be able to convert unstructured into structured data. In many cases we are hesitant to delete the original data. Delete unstructured data if there is no value to keep it. For example, a form that is used to file a claim. There is also the tendency to add photos or video as evidence. Is this evidence used or stored “just in case that”. Is it stored somewhere else? Maybe if you have seen the evidence, approve the claim and subsequently you can remove the photos and videos.

A refinement of this principle is the creation of data in a preferred format.

For example, JSON, XML or records in database can be used to store structured data. From a storage perspective XML would probably be more efficient than JSON.



For unstructured data the first question you should consider is: what type of data is required? Picture enough or video better? Followed by the question What is the most efficient format to store it? Bitmap or is JPEG sufficient?

STORE

You created the data in the previous phase. You took important decisions to do it in the most effective way. As soon as the data is created, you probably want to keep it in a persistent way. In the store and maintain phase of data lifecycle management, the primary objective is to securely store and manage data to preserve its integrity and ensure accessibility over time. This necessitates the implementation of robust storage solutions and protocols to mitigate the risks of data loss, corruption, or unauthorized access. However, it's important to note that traditional storage practices can often consume significant amounts of (unnecessary) energy. Therefore, adopting a sustainable approach to storage is essential to minimize environmental impact while maintaining data security and availability.

USE OPTIMIZED STORAGE MECHANISM

Optimizing storage mechanisms such as SQL databases or NoSQL databases is a critical aspect of promoting sustainability within the data lifecycle. These optimized storage solutions can efficiently manage and organize data, thereby reducing storage costs, and energy consumption. SQL databases, known for their structured data model and strong consistency guarantees, are well-suited for applications requiring complex queries and transactional integrity. By carefully designing database schemas with normalization techniques and indexing strategies, organizations can optimize SQL databases to minimize storage overhead and enhance query performance, leading to reduced resource usage and improved energy efficiency. On the other hand, NoSQL databases offer flexibility and scalability for handling large volumes of unstructured or semi-structured data, supporting distributed architectures and horizontal scaling.

In addition to traditional SQL and NoSQL databases, other storage mechanisms and their uses are listed below. By selecting the appropriate storage mechanism based on the nature of the data and the application requirements, organizations can optimize resource utilization, enhance performance, and minimize storage footprint.

Database	Data Complexity	Scalability	Flexibility	Query Needs
Relational (SQL) Databases	Structured data	Scalable	Limited flexibility caused by fixed schemas	Complex queries, ACID transactions
NoSQL Databases	Varies	Highly scalable	Flexible schema	Varied query languages
Object-Oriented Databases	Complex data	Scalable	Flexible data structures	Complex queries
Key-Value Stores	Simple data	Highly scalable	Limited flexibility	Fast data retrieval, simple
Document-Oriented Stores	Semi-structured data	Scalable	Flexible schema	Complex queries
Graph Databases	Complex relationships	Scalable	Flexible schema	Complex relationship queries

Example: Imagine using MongoDB to store financial data where you need to track customer accounts and their transactions. If transactions between accounts occur, MongoDB doesn't natively support complex joins and multi-document ACID transactions, leading to higher energy consumption for maintaining data integrity.

It's also essential to be aware of non-sustainable storage practices that can undermine these efforts. Examples of non-sustainable storage practices include:

1. Using NoSQL databases for highly relational data: NoSQL databases are optimized for handling unstructured or semi-structured data and may not be suitable for highly relational data with complex transactional requirements.
2. Using key-value stores for complex queries: Key-value stores are optimized for simple read and write operations based on key lookups. Using key-value stores for complex queries requiring joins or aggregations can lead to inefficient data access and reduced performance compared to using more suitable storage mechanisms such as relational databases.

STORAGE TIERING

Storage tiering is a crucial tactic for promoting sustainability within the data lifecycle. By efficiently categorizing data into different tiers based on their usage patterns and importance, organizations can optimize storage resources, reduce costs, and minimize energy consumption. Tiered storage assigns various categories of data to different types of storage media, such as high-speed drives, lower-speed drives, MAID (Massive Array of Idle Disks), or tape, depending on their access frequency and performance requirements. Lower performance storage options typically consume less electricity, making them more energy-efficient and cost-effective for storing less critical or infrequently accessed data. High-speed drives are reserved for mission-critical applications that demand instantaneous response times. Additionally, storage tiers can incorporate varying levels of data protection to ensure data integrity and availability. Automated storage tiering (AST) further enhances efficiency by dynamically moving data between different disk types and backup levels based on capacity, performance, and cost considerations. This dynamic approach to storage management helps optimize resource utilization while meeting the organization's operational needs.

Data Category	Description	Best Practices
Hot Data	Frequently accessed, mission-critical data	- Store on high-speed drives for instant access - Implement RAID for redundancy and data protection
Warm Data	Moderately accessed, important data	- Store on lower-speed drives for cost-effectiveness - Implement RAID for data protection
Cold Data	Infrequently accessed, less critical data	- Store on low-speed drives, MAID, or tape for cost-efficiency and energy savings - Implement data compression for further storage optimization
Archive Data	Data that needs to be retained for compliance or long-term purposes	- Store on tape or cloud archive for cost-effectiveness and long-term retention

DATA COMPRESSION AT REST

Implementing data compression [2] at rest can also enhancing sustainability in storage phase. This technique involves reducing the storage footprint of data by encoding it in a more compact form, thereby optimizing storage space and reducing resource consumption. It's important to note that data should be compressed based on its usage frequency, as it shouldn't remain unchanged all the time.

Compression can be done at several levels, for example:

Storage. Low level compression. The advantage is that compression can be built into hardware which ensures performance with a low footprint. Compression and decompression happen every time the data is stored and retrieved but it is fast.

Application. At the application level, compression can be applied selectively based on specific use cases. Applications can compress data before writing it to disk or sending it over the network and decompress it when needed for processing. This approach offers flexibility, allowing developers to choose the most efficient compression algorithms based on the data type and usage pattern. Application-level compression can also be combined with other optimizations such as caching and deduplication to further enhance performance and reduce resource consumption.

Technology and Location Consciousness

Technology differs, some products are more sustainable than others. Datacenter location differ. How green is your datacenter? How sustainable your (public) cloud provider? That information is available and there are big differences per locations even from the same provider. That information does not include the difference in transport energy if you retrieve it from the location around the corner or from the location at the other side of the world.

Local green data centers offer another viable option that can sometimes be more cost-effective and sustainable, especially in terms of energy transport. By moving less frequently accessed data (cold data) to the cloud or local green data center, organizations can reduce the energy consumption of their on-premises data centers, and this may have some sustainability benefits:

Efficient Use of Resources: Both public cloud providers and local green data centers often operate using energy-efficient infrastructure and renewable energy sources, reducing the overall energy consumption associated with data storage.

Green Data Centers: Many public cloud providers and local green data centers are committed to sustainability and operate green data centers that minimize carbon emissions. By leveraging these facilities, organizations can significantly reduce their overall carbon footprint. However, this should not relive you from your data sustainability responsibilities!

Dynamic Scaling: Both cloud storage and local green data centers offer scalable solutions that adjust based on demand. This flexibility ensures that resources are used efficiently, optimizing energy consumption, and reducing waste.

Improved Data Management: Public cloud services offer advanced technologies like automated tiering, which moves cold data to more energy-efficient storage tiers. This reduces the energy requirements for managing large datasets and enhances overall sustainability [12].

Some of the best practices that we discussed like compression, tiering and deduplication are easily available.

USE

The data is created, and it is stored. 2/3rd of the stored data is never used [3]. That's a shame. So, if you store it, you better use it. How to do that in a sustainable way?

The fifth phase of the lifecycle is where data is actively utilized for various purposes. During this phase, organizations leverage collected data to derive insights, make informed decisions, and drive business processes. Data is accessed, analyzed, and manipulated to extract valuable information and support operations across different departments. This phase involves tasks such as data querying, reporting, visualization, and (AI) modeling to extract actionable intelligence from the data. Additionally, data may be shared among stakeholders and integrated into various applications to facilitate decision-making and enhance productivity. Following are the sustainable practices to minimize environmental impact for the data utilization phase and ensure long-term sustainability in data operations.

DATA COMPRESSION IN MOTION

Data compression in motion refers to the practice of compressing data while it is being transmitted or used, presenting a proactive approach to enhancing sustainability in use stage. Unlike static compression strategies like data compression in rest applied during storage, data compression in motion operates dynamically, compressing data in real-time as it is being transmitted or accessed. This proactive strategy not only optimizes energy consumption in storage hardware by reducing the amount of physical space required but also contributes to significant reductions in data transmission time and communication bandwidth usage. By compressing data during its active usage, organizations can minimize their environmental footprint while simultaneously improving operational efficiency.

However, implementing data compression in motion has trade-offs. One significant trade-off is the increased use of computing resources needed for real-time compression algorithms. These algorithms require computational power, which can raise energy consumption in data processing. Also, real-time compression might introduce overhead and latency, affecting system performance and user experience. To address these challenges, organizations must carefully consider the trade-offs involved in data compression, weighing the environmental benefits against the additional computational resources required.

ONLY PROCESS THE DATA THAT IS REQUIRED TO EXECUTE THE PROCESS

Another green strategy in data management is to only process the data that is required to execute the process efficiently. This approach involves carefully assessing the data needs for each specific task or operation and processing only the relevant data, thereby minimizing unnecessary data processing and reducing energy consumption. By implementing data filtering and optimization techniques, organizations can avoid processing excessive volumes of data that may not contribute to the desired outcomes. This strategy not only conserves computational resources but also helps streamline data workflows, leading to improved efficiency and reduced

environmental impact. Additionally, by focusing on processing only essential data, organizations can enhance data security and privacy by minimizing the exposure of sensitive information.

If your data is required for a report that will be published by the end of the month, it may be more efficient to wait until the end of the month to process it, together with the other data that will be published in that report.

ENSURE THAT THE DATA IS PROCESSED ON THE RIGHT LEVEL OF UTILIZATION

Following the strategy of only processing the data required to execute processes efficiently, another green approach in data lifecycle management involves ensuring that the data is processed at the right level of utilization. While the former strategy emphasizes minimizing unnecessary data processing, this strategy focuses on *optimizing the utilization level of the processed* data. This entails evaluating the workload demands and resource availability to determine the appropriate level of data processing. By striking a balance between processing efficiency and resource utilization, organizations can optimize energy consumption and maximize operational efficiency. Implementing dynamic workload management systems and adaptive processing techniques enables organizations to adjust data processing levels based on real-time demand, thereby optimizing resource utilization and minimizing energy wastage.

ENSURE THAT THE DATA IS PROCESSED VIA GREEN CODE

Ensure that the data is processed via green code technologies. Green code refers to software development practices that prioritize energy efficiency and environmental sustainability. By leveraging green code technologies, organizations can minimize the energy consumption associated with data processing operations.

Structural changes, such as optimizing energy use in multi-core processor-based applications and leveraging green IT infrastructure like virtual machines and containers, are crucial for enhancing energy efficiency. Additionally, adopting microservices architecture and cloud-based DevOps practices can further optimize energy usage by improving resource utilization and reducing network energy consumption.

Efficient data processing techniques, such as list comprehension and parallelization, contribute significantly to energy efficiency in data processing. List comprehension, for example, allows for the creation of lists without using traditional loops, reducing computational overhead and energy consumption. Parallelization leverages multiple cores in modern computers to divide tasks efficiently, maximizing computing power and energy efficiency.

Cultural changes are equally essential in driving the adoption of green coding practices. Empowering management and employees to embrace sustainability initiatives fosters a culture of environmental responsibility and encourages innovation in energy-efficient data processing techniques. By implementing green coding practices, organizations can not only reduce energy costs but also accelerate progress toward sustainability goals and achieve higher earnings.

Beyond energy savings, green coding offers additional benefits. It reduces energy costs, crucial in today's dynamic energy market, ensuring both

environmental and business sustainability. Moreover, it accelerates progress towards sustainability goals, essential for organizations aiming for net-zero emissions. CEOs implementing green coding often report higher operating margins, according to a CEO study [4]. Additionally, it fosters better development discipline, simplifying infrastructures and saving time for software engineers.

OPTIMIZED SEARCH ALGORITHM

Implementing an optimized search algorithm is also a green tactic in the “use” phase. This approach focuses on enhancing the efficiency of search operations to minimize computational resources and energy consumption while maximizing search performance. Optimized search algorithms prioritize techniques that streamline the search process, reducing the time and energy required to locate relevant information within datasets.

One key aspect of optimized search algorithms is the utilization of efficient data structures and indexing methods. By organizing data in a structured and indexed format, search operations can be performed more quickly and with reduced computational overhead. Techniques such as binary search trees, hash tables, and inverted indexes enable rapid retrieval of data based on specific search criteria, leading to improved efficiency and reduced energy consumption.

Moreover, optimized search algorithms leverage advanced search techniques such as parallel processing, distributed computing, and caching mechanisms to further enhance performance and reduce energy usage. Parallel processing allows search tasks to be divided among multiple processors or cores, accelerating search operations and minimizing idle time. Distributed computing distributes search tasks across multiple nodes or servers, enabling parallel execution of search queries and optimizing resource utilization.

Additionally, caching mechanisms store frequently accessed data or search results in memory, reducing the need to recompute or retrieve data from disk during subsequent search operations. By caching relevant data, optimized search algorithms can significantly reduce energy consumption associated with disk access and data retrieval, further enhancing efficiency and sustainability.

ENHANCE

Once the data is created, stored and being used, you may want to add or modify your data. You received review comments on a document you created. You have your original document, the documents with comments and you create a third document that applies the comments. Because you would like to demonstrate that you have applied the changes to the original version, you keep all three documents. This is called data lineage. But is it really necessary to keep them all?

Enhancing data feels like buying new shoes. You do not throw away your old ones, who knows, the new ones may not be as comfortable as the old ones, so keep them, but for how long?

THROW AWAY OLD VERSIONS

When a new version of a document (or any form of data) is created, the most sustainable approach is to delete the previous version (overwrite). But on many occasions, there is a lot of hesitance to do that, who knows… Sooner or later you may run into this situation that you have to rollback some changes, but how often did that occur and in that case did you remember which version contained that original piece of text?

In case of structured data, for example transaction data there is often a lot of interest in history, for example to analyze trends. Worst case this result in storing each modified transaction again and again. On top of that, you need to add a date to the transaction and the query must contain an additional field that select the latest transaction.

The principle is clear: **Delete old versions**. The following measures only apply when the old version cannot be deleted (overwritten).

GIT is an excellent example that stores only delta's and makes good use of that.

ONLY STORE DELTA'S

When you do not want to throw old versions away, make sure only the deltas are stored. You can do that both for structured data as well as for unstructured data.

APPLY RETENTION POLICIES FOR OLD DATA

Usually there is a different retention need for older versions of data. The latest version is important, older versions may be deleted after some time. Therefor include a retention policy for older data as well.

AGGREGATE DATA

There could be a lot of value in historical data. You may be interested in your electricity usage of three years ago. However, you are most likely not interested in the usage of your electricity on a specific hour of a specific day three years ago. You may want to look at the usage of a specific month. Therefor you could aggregate the hourly and daily usage and aggregate that to monthly usage and probably if it is more than three years, you are only interested in the yearly number.

DISPOSE PHASE

The ‘Dispose’ phase of the data life cycle is a critical juncture that goes beyond simply deleting obsolete information. This phase ensures the secure and compliant removal of data that is no longer needed or should be removed based on legal requirements. Less data means a lower power consumption and reduced need for resources like disks. Storing large volumes of unnecessary or redundant data requires significant energy to maintain servers and data centers. By cleaning data, organizations can reduce the amount of storage needed, thereby decreasing energy consumption. Next to the consumed power for storage, processing and managing large datasets requires substantial computational resources. Reducing the volume of data through cleaning minimizes the energy required for these operations.

In the ‘Design’ and ‘Plan’ phase we have already defined the data cleaning and archiving strategy regarding to new developments. This paragraph will cover the governance on the disposal of data and the activities that you should perform to dispose your data.

THE DATA RETENTION POLICY

Ensure that sustainability is addresses in the corporate data retention policy as a viewpoint next to the legal requirements. All the data that has no business value anymore should be disposed as a generic guideline. The value is in meeting regulatory compliance, required input for the execution of business processes and decision making. An important prerequisite for this is that the data retention policies within the enterprise are clear, translated in practical guidelines that are actively communicated throughout the organization.

An example of a practical guideline is a policy that Teams sites are deleted when they are not used for 6 months.

You could join the yearly world digital cleanup day or organize your own cleanup initiative.

CLEANUP

Get clean via cleaning initiatives to ensure that all the data that has no business value anymore is disposed. A perfect moment in time could be in the closing phase of your project. The closing phase is a very important phase, but unfortunately only the best project manager honor this. In this phase you finalize documentation, harvest reusable assets, clean your code, and why not delete unnecessary data…

Maybe you have a moment in the week that you do some personal administration. Why not take a look at your download folder and delete everything that is left there.

Plan cleanup events to clean personal and team data… You could join the word digital cleanup day with your organization or organize your own cleanup event. It is fun to cleanup your mess in an interactive way as a group.

After you have done the painting of your house you are not ready painting, you need to clean the brushes, correct some small mistakes and try to remove the painting from your hands (and sometimes more than hands only)

DATA RETENTION STANDARDS AND TOOLING

People, Process and Technology is a triangle that need to be balanced to achieve results. Many tools(technology) provide the capability of setting

Have courage to implement retention policies unconditionally!

retention policies. Make sure that you have policies (process) in place and develop a culture of data cleaning. Many tools, specifically on content platforms do provide content retention capabilities. Ensure that these data retention policies are implemented and automated if possible. This includes a mature version management practice but also courage. How much courage do you have? We will end with the principle: Have courage to implement the policies unconditionally.

Final Thoughts

We discussed several measures for reducing energy consumption that you can consider during the lifecycle of your data. Of course, energy is only one aspect that you have to consider. It should be balanced against privacy, security, ethics, performance and many other non-functional requirements.

For some measures you may consider the application of AI to support you with the application of the measures. For example:

Data selection - AI models can be used to determine the most relevant data to collect and reduces the collection of unnecessary data.

Storage tiering - AI can dynamically manage data across different storage tiers based on usage patterns.

Data governance - AI can help enforce sustainable data governance practices by automating policy enforcement to be compliant with sustainability standards.

There is much to gain, and remember, today every bit of data is at least multiplied by a factor: High availability, Backup, Copy in a data warehouse, a copy of your data that is again made high available etc. Imagine what you can save!

REFERENCES

[1] DAMA-DMBoK. DAMA-DMBoK. Available at: https://www.dama.org/

[2] TechTarget. Data Compression. Available at: https://www.techtarget.com/searchstorage/definition/compression

[3] Market Logic Software. Wasted Info: Why So Much Enterprise Data Goes Unused. Available at: https://marketlogicsoftware.com/blog/wasted-info-why-so-much-enterprise-data-goes-unused/

[4] IBM. 2022 CEO Study. Available at: https://www.ibm.com/thought-leadership/institute-business-value/en-us/c-suite-study/ceo

[5] International Energy Agency (IEA). Electricity 2024: Report on the Analysis and Forecast to 2026. Available at: https://www.iea.org/reports/electricity-2024

[6] Belastingdienst. Dutch Tax Office: Administration Requirements. Available at: https://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/btw/administratie_bijhouden/administratie_bewaren/

[7] Digital Decarbonization. Source. Available at: https://digitaldecarb.org

[8] Substack. The Carbon Impact of Large Language Models. Available at: https://tinyml.substack.com/p/the-carbon-impact-of-large-language

[9] Scientific Computing. The True Cost of AI Innovation. Available at: https://www.scientific-computing.com/analysis-opinion/true-cost-ai-innovation

[10] NetApp. People, Profit, and Planet: Sustainability. Available at: https://www.netapp.com/blog/people-profit-planet-sustainability/

[11] NetApp. Data Waste Index Research Analysis. Available at: https://www.netapp.com/media/83312-data_waste_research_report_final_for_submission.pdf

[12] NetApp. 8 Scenarios for Using NetApp Cloud Tiering. Available at: https://bluexp.netapp.com/blog/8-reasons-to-use-cloud-tiering-in-your-data-center

ACCOUNTABILITY

This point of view is developed under the umbrella of the NCDD. It is a result of the Data Lifecycle Management stream within the working group Architecture. The authors are from different organizations and companies. The sole objective of this point of view is that companies become more aware about the sustainability measures that could be taken. Any reference to a product or service is only used as example, reference to any of our companies' products or services is a coincidence.

We really hope this point of view will trigger the awareness in your organization and will result in sustainable measures.

The authors:

- **Eloise Zhang**, MSc Computer Science UVA/VU | Sustainable & Secure AI. *LinkedIn: https://www.linkedin.com/in/eloise-zhang-82b6ab27b/*
- **Pepijn van der Veen**, Enterprise Architect Data. *LinkedIn: https://www.linkedin.com/in/pepijn-van-der-veen-7023444/*
- **Ronald Meijer**, IT Architect. *LinkedIn: https://www.linkedin.com/in/ronald-meijer-0a715b1/*
- **Gerrit Ouderkerken**, Manager District Sales. *LinkedIn: https://www.linkedin.com/in/gerritouderkerken/*