Journal of Cleaner Production 371 (2022) 133633

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Exploring the sustainability challenges facing digitalization and internet data centers

Dlzar Al Kez[a,b,*], Aoife M. Foley[c,d], David Laverty[a], Dylan Furszyfer Del Rio[c], Benjamin Sovacool[e,f]

[a] *School of Electronics, Electrical Engineering and Computer Science, Queen's University of Belfast, Belfast, United Kingdom*
[b] *Electrical Engineering Department, University of Sulaimani, Sulaymaniyah, Iraq*
[c] *School of Mechanical and Aerospace Engineering, Queen's University of Belfast, Belfast, United Kingdom*
[d] *Department of Civil, Structural and Environmental Engineering, Trinity College, Dublin, Ireland*
[e] *Science Policy Research Unit, University of Sussex, Brighton, United Kingdom*
[f] *Department of Business Technology and Development, Aarhus University, Denmark*

ARTICLE INFO

Handling Editor: Prof. Jiri Jaromir Klemeš

Keywords:
Data center emissions
Dark data
Data storage
Environmental footprints
Power consumption

ABSTRACT

Internet data centers have received significant scientific, public, and media attention due to the challenges associated with their greenhouse gas, water, and land footprint. This resource greedy data services sector continues to rapidly grow driven by data storage, data mining, and file sharing activities by a wide range of end-users. A fundamentally important question then arises; what impact does data storage have on the environment and is it sustainable? Water is used extensively in data centers, both directly for liquid cooling and indirectly to generate electricity. Data centers house a huge number of servers, which consume a vast amount of energy to respond to information requests and store files and large amounts of resulting data. Here we examine the environmental footprint of global data storage utilizing extensive datasets from the latest global electricity generation mix to throw light on this data sustainability issue. The analysis also provides a broad perspective of carbon, water, and land footprints due to worldwide data storage to through some light on the real impact of data centers globally. The findings indicate that if not properly handled, the annual global carbon, water and land footprints resulting from storing dark data might approach 5.26 million tons, 41.65 Gigaliters, and 59.45 square kilometers, respectively.

1. Introduction

The development of information and communications technologies (ICT), such as the Internet of Things (IoT) sensors, cloud computing services, big data analytics, and the introduction of new smart devices and software applications, have generated exponential growth in the volumes of data generated over the past two decades (Corallo et al., 2021). The ICT industry is fast evolving with the emergence of cloud computing, the expansion of 5G networks, artificial intelligence, and big data leading to the creation of huge amounts of data. Worldwide Internet Protocol (IP) traffic and Internet data grew more than 10-fold between 2010 and 2018, while global data center storage capacity increased by a factor of 25 (Shehabi et al., 2016).

Modern ICT technologies make it very easy to generate large volumes of data, and because storage is quite cheap, there is a tendency to keep that data regardless of whether it has a point (i.e., synchrophasors and smart meters). Thus, companies are expected to have a high capacity for gathering and managing large amounts of data, both technologically and in terms of the skills and capabilities required from employees (Abdulsalam et al., 2019). Organizations recognize the significance of data and are investing extensively in data management as they move closer to data-driven business models. They continuously generate an overwhelming flow of data, during routine activities, from various sources (e.g., enterprise systems, machines, sensors, controllers, and demand-side digitalization). These data come in multiple formats (i.e., dark data, redundant, and critical) and are referred to herein as big data, which includes a wide range of information streams, log files, master data, and manually entered operator data (Nagorny et al., 2017). The term "dark data" refers to unstructured and inert content, which is fundamentally opposed to critical structured data. However, redundant

* Corresponding author. School of Electronics, Electrical Engineering and Computer Science, Queen's University of Belfast, Belfast, United Kingdom.
E-mail address: dalkez01@qub.ac.uk (D. Al Kez).

https://doi.org/10.1016/j.jclepro.2022.133633
Received 16 June 2022; Received in revised form 28 July 2022; Accepted 11 August 2022
Available online 16 August 2022

data is semi-structured information with a high risk of becoming dark (Imdad et al., 2020).

Currently, data centers are considered one of the fastest-growing electricity consumers (Jones, 2018). According to the International Energy Agency, they consume around 1% of global electric power generation, which is about 205 TWh (IEA, 2020), with computing power accounting for 43% of this figure, power provision systems for another 11% (Dayarathna et al., 2016), storage drives for about 11% excluding their share of cooling and other infrastructure energy consumption, networks for 3%, and cooling represent about 32% which highly benefits from access to natural resources for cooling (Shehabi et al., 2016). Data centers, as heavy energy consumers, are at the core of discussions over energy efficiency and carbon dioxide (CO_2) emissions. However, reliable information on the extent of energy consumption, and underlying emissions, behind data center infrastructure, remains fragmented, difficult to acquire, and even more difficult to authenticate. This is because the rapid expansion and growth of data centers make reliable and recent data difficult to compile, and also because energy efficiency gains are outpaced (Koot and Wijnhoven, 2021).

The digital revolution is coincident with global warming, and the increasing need to mitigate emissions and environmental stress. Along with the effects of climate change, many freshwaters and other natural systems are losing their ability to maintain ecological functions while also being forced to meet increasing agricultural and industrial demands. Depending on how data centers are powered, they indirectly consume a significant amount of water for electricity generation (Shehabi et al., 2016), and some use direct water to cool servers and storage drives. For example, considering water deployed in utility dispatched power stations, Microsoft used 3.96 Gigalitres (GL) of water in 2020, up from 1.91 GL in 2017 (Microsoft, 2020), while Google utilized 21.5 GL in 2021, up from 11.62 GL in 2017 (Google, 2021). While some IT companies, such as Google and Amazon, have made significant strides toward reducing their environmental impact by investing in renewable energy (Amazon, 2022) and enhanced data storage (Jones, 2022), others are still trailing behind in the transition to green energy and storage optimization. Furthermore, it is currently impossible to precisely calculate emissions associated with data storage but the entire ICT sector is estimated to account for about 1.4% of global CO_2 emissions due to the large amounts of energy they consume that are often carbon-intensive (Cunliff, 2020).

The motivation for this analysis is that a portion of the actual physical center environmental footprint is associated with worldwide data storage. In the United States, for example, power consumption due to data center data storage was estimated to be at 14 TWh in 2020 resulting in almost 6.5 million metric tons (MT) of CO_2 emissions (Backup Works Storage Solutions, 2020). It is worth mentioning that dark data accounts for 54% of worldwide data storage, and the storage power required to hold and process dark data is estimated to emit 5.8 MT of CO_2 (Veritas, 2015). Nonetheless, one core challenge remains in the data center sectors and that is their commitment to achieving net-zero carbon emissions as part of their social and corporate responsibilities. It should also be noted that they are not going to magic up some technologies to undo all the damage they are doing between now and when they invent this as yet unheard of net-zero device. Until now, much of the attention of the tech industry and large data center operators have been on the transition to cleaner, renewable energy sources. While this is an important part of developing a more sustainable company, it ignores one of the most significant characteristics of a green data center including data storage optimization and waste reduction.

Renewable energy is a long term plan and not the only way to achieve sustainability, and it comes with its own set of challenges (Al Kez et al., 2022). There is a high potential to accelerate progress toward sustainability by eliminating resource waste and ensuring that investments in data storage infrastructure yield maximum value (Vries and Stoll, 2021). Minimizing dark data storage and employing modern tape systems, for example, can help to speed up progress toward sustainability by lowering energy use and CO_2 emissions (Cooke et al., 2021). However, the problem is before data centers commit their data to tape, they have to admit to themselves that the data is useless, and they will never want to look at it again, meaning that the most sustainable solution of all is to delete it. An important question arises: what impact does power consumption due to data storage have on the environment and is it sustainable? In this context, this research uses a fundamental proxy variable-based footprint assessment method to radically determine CO_2, water, and land footprints associated with data storage. The main contributions of this research are the following:

- Provide an estimation of carbon, water, and land footprints due to data storage using datasets for the most recent global electricity generation mix.
- Differentiate between normal data, critical data, abandoned data, and dark data.
- Determine the environmental footprint deviation for twelve data center dominant countries for global parameters.

The rest of the paper is organized as follows: Section 2 presents an overview of global ICT environmental footprints and the research gap within the literature. The data hierarchy of need and classification of data types are given in Section 3. This is followed by the proposed methodology to assess the environmental footprint of data storage. Analysis and results are illustrated in Section 5. Discussions and recommendations are provided in Section 6. Finally, Section 7 concludes the work with some remarkable future directions.

2. Literature review

ICT, like all sectors, confronts obstacles in its efforts to reduce carbon emissions. It also allows other industries to become more energy efficient. Our previous analysis examined the possible system benefits of integrating data centers with variable renewable energy technologies to support grid services (Al Kez et al., 2020) while facilitating secure integration of higher levels of intermittent renewable systems (Al Kez et al., 2021). Our results demonstrated that instead of simply being a power load, data center businesses might ease the transition towards renewable electricity by utilizing the potential for demand response to match data center demand with times when high renewable power is generated. As this reduces the environmental footprint associated with total ICT energy usage, this industry can solve energy and environmental challenges. Despite these capabilities, ICT industries have not yet been actively deployed to provide grid flexibilities and thus still have a considerable environmental impact (Bloomberg, 2021).

The global carbon footprint of Internet use ranges from 28 to 63 g (g) CO_2 equivalent per Gigabyte (GB), whereas the water and land footprints are 0.1–35 L (L)/GB and 0.7–20 cm^2/GB, respectively (Obringer et al., 2021). In 2015, a water footprint of data centers up to 205 L/GB was reported by (Ristic et al., 2015). This includes footprints associated with both transmission and data storage globally in a data center. However, research by (Obringer et al., 2021) identified that this number has reduced almost by 150% during five years to around 35 L/GB based on the global energy mix in 2018. Data storage alone has been reported to have 5 L/GB water footprints (Obringer et al., 2021). The significant reduction in the environmental footprints of data centers was justified and attributed to advancements and efficiency improvements in servers, storage devices, network switches, and data center infrastructure (Siddik et al., 2021). Another study by (Belkhir and Elmeligi, 2018) reported the differences between annual global CO_2 emissions from ICT and data center industries over the last ten years, as shown in Table 1. In contrast to what was previously claimed, the data in the table highlights a considerable increase in the amount of CO_2 emissions associated with both the ICT and data centers in recent years. However, a report by the Internet data center corporation indicated that CO_2 emissions in 2020 would be far lower, at 230 MT, than those in Table 1 (Cooke et al.,

Table 1
Annual worldwide ICT, data centers, and communication network carbon footprint.

Global carbon footprints in (MT CO_2e)											
Years	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Data centers	159	179	200	224	251	281	315	352	395	442	495
Communication network	138	152	167	179	192	205	218	230	243	256	269
ICT sector	610	700	800	850	900	950	1000	1050	1150	1210	1300
ICT in percentage of global footprint (%)	1.85	2.1	2.27	2.3	2.5	2.73	2.83	3	3.23	3.45	3.7

2021).

In fact, such considerable discrepancies in environmental footprints lie in the scope and the underlying assumptions used in previous research. This is because experts make assumptions and build models to work with energy consumption figures due to business competitive concerns, lack of knowledge, technical challenges, or lack of transparency. Global figures are also extremely subject to geographical region considerations. Researchers looked at data center energy usage and performance from a variety of perspectives, including computer power, cooling, and network-related issues. Many businesses continue to use PUE based assessment methodologies. However, assessing a data center's long-term sustainability necessitates looking at a variety of factors, some of which are difficult to quantify. As an example, studies by (Aslan et al., 2018) and (Malmodin and Lunden, 2018) have estimated the CO_2 footprint of data storage, transmission, and consumption, while the footprint associated with storing dark data is not considered. It is also to be noted that the CO_2 footprint alone cannot provide a complete picture of the environmental impact of Internet use (Ristic et al., 2019). Despite the importance of the environmental challenges, only a few studies consider the complete picture of CO_2, water, and land footprints of data centers (Obringer et al., 2021). However, the environmental footprints associated with dark data storage are once again ignored.

The challenges associated with data management (i.e., the process of acquiring and storing data, as well as preparing and retrieving it for analysis) have been reported previously in terms of data governance, data and information sharing, operational costs, data ownership, privacy, and security (Sivarajah et al., 2017). However, now another consequence has been brought to light; according to Veritas (2015), the energy required to store dark data contributes significantly to the carbon footprints of the data center. Furthermore, there is also a need to revise earlier energy estimates regularly, given technological and efficiency advancements in the Internet industry, as well as shifting energy supply portfolios around the world. Thus, the main objective of this research is to fill this gap through a rough estimation of three major environmental footprints from a broader perspective (i.e., CO_2 footprint, water footprint, and land footprint) associated with different data storage types (i.e., dark data, redundant data, and critical data).

3. Data hierarchy of needs

The data science hierarchy of needs includes data collection, transport, and storage, data exploration and manipulation, aggregation and labeling, learning and optimization, and artificial intelligence and deep learning. The data science pyramid is inspired by Maslow's hierarchy of needs, and it allows data to pass through many phases to locate useful data to act on, as illustrated in Fig. 1 (Renze, 2019).

First, to adopt Maslow's concept of working from the ground up, companies must begin with data collecting. It is essential to know what type of data they require among the big data. For every data-driven company, this is a vital step at the base of the hierarchy and the most basic requirement. It establishes the foundation for the company's higher and better objectives (Rotem-Gal-Oz, 2015). Basic data gathering and operations start with documenting transactions, reporting faults, and digitizing analog data. To build a strong dataset, firms must look at the data coming in from sensors and the methods important user interactions are being documented before advancing to the next pyramid level.

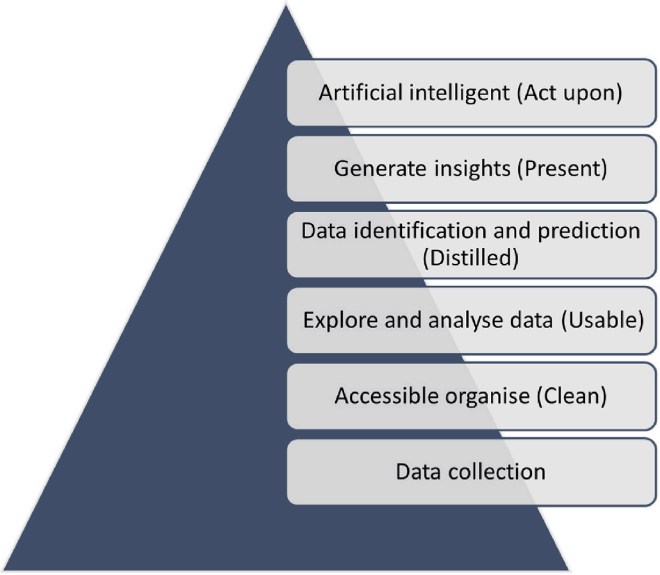
Artificial intelligent (Act upon)
Generate insights (Present)
Data identification and prediction (Distilled)
Explore and analyse data (Usable)
Accessible organise (Clean)
Data collection

Fig. 1. Data hierarchy of needs.

The data must then be relocated to a secure and easily accessible location to help the researchers to find the information they need. A corporation should now be able to move on to securing information flow, structure, and storage, as well as tracking how data moves through the system, with accurate, reliable, and complete data collection. Data sets are often chaotic; therefore, data scientists must find a means to verify that the data they are collecting is properly formatted and ready to be analyzed (Hashem et al., 2015). The data must be structured and converted into a format that can be analyzed. This begins with fundamental data-organization tasks such as data transformation, cleaning, and storage.

The big data appears to be chaotic, but it can be used to uncover hidden insights through exploration. The exploration and data analysis via anomaly identification and data cleaning is the next step in the data science hierarchy of needs pyramid to make the data usable. This is a crucial first step toward a more robust data organization at the following levels. If the outcomes are not up to standard, it might be time to revisit the foundation and refocus on gathering methods. It generally begins with basic data analysis tools, such as reports and dashboards.

Firms must use data to develop insights that drive business choices once it has been acquired, stored, converted, and analyzed. Descriptive, diagnostic, predictive, and prescriptive data analytics are four forms of data analytics that can aid in the development of insights. This typically entails incorporating increasingly complex types of data analysis into their data-science pipelines, such as predictive analytics, prescriptive analytics, and machine learning. This step helps understand what is occurring in the company and why it is happening. In fact, there are lots of software and hardware technologies in use in data centers that predict whether data is likely to be accessed or not and move data up and now a hierarchy depending on if it needs to be "snappy" or if the end user is unlikely to notice if it takes a while to be retrieved (Zhu et al., 2019).

This is generally obfuscated by the customers unless they want to get into higher tier plans and specific service level agreements (SLA).

Finally, to close the loop and eliminate the human from the process, data-science operations must be automated to act upon the data. This approach employs artificial intelligence, deep learning, and reinforcement learning to reduce human involvement costs while increasing revenue. Here the worst form of dark data might be that at the bottom which simply exists and does not even get any further processing. The best type of dark data might be that which was once acted upon but now is of no relevance ever again.

3.1. *Classification of big data*

Big data is defined as data with greater diversity, emerging in larger volumes, and with higher velocity, which is also referred to as the three Vs (Oracle, 2022). The details of these three aspects are as follow: 1) The numerous different sorts of data that are available are referred to as variety. Traditional data formats were well-structured and fit into a relational database with ease. With the rise of big data, new unstructured data kinds have emerged. To derive meaning and support metadata, unstructured and semistructured data types including text, audio, and video require further preprocessing. 2) The rate at which data is received and (perhaps) acted on is referred to as velocity. In most cases, data is streamed directly into memory rather than being written to a disc. 3) It is important to consider the amount of data available. Data centers have to process a lot of low-density, unstructured data with big data. This can be unvalued data like social media data feeds, clickstreams on a website or mobile app, or sensor-enabled equipment. This might be tens of gigabytes of data for certain firms. Over the last few years, two more Vs have appeared: value and veracity which refer to the process of rapidly generating and identifying enormous hidden values from vast datasets of various forms (Zhang and Yang, 2021).

Classification of big data into several types is essential to analyzing their features due to the availability of a vast amount of data in the cloud. They are typically categorized based on four different factors including, source of data, the content of data, data store, and data staging and processing, as shown in Fig. 2 (Hashem et al., 2015). Each of these categories has certain unique characteristics and levels of complexity, but the data content is the main focus of this study. Examples of data sources are Internet data, sensing, and any repositories of transnational information that range from unstructured to highly structured and store its content in a variety of formats.

The massive hadron collider (LHC) is probably the most famous big data example. The LHC program initiated two general-purpose experiments, ATLAS and CMS, to search for the Higgs boson. One of the most difficult aspects of these experiments is managing and analyzing a large volume of data from the High Energy Physics (HEP) detectors. Firms are confronted with a concrete case of Big Data in the LHC era: the LHC produces 40 million collisions of protons every second, or around 15 trillion collisions per year (Innovation News Network, 2021). Every collision produces one Mbyte of data, or 2000 Tbytes (TB), 2 Petabytes (PB) of data each year for the ATLAS detector alone. Furthermore, a comparable amount of simulated data derived from various theoretical models is required for comparison with the ’real data’ gained from tests. After that, in order to obtain physics results, all of these data must be run through analytic algorithms. Finally, although there are hundreds of good big data initiatives in science and industry, advertising and medical research remain hot topics (Au-Yong-Oliveira et al., 2021).

3.2. *Data storage types*

It is projected that the amount of Global Datasphere (i.e., the summation of all data created, captured, or replicated) to grow 5-fold from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025 (Reinsel et al., 2018). Despite the development of new data storage technologies, data volumes are roughly doubling every two years. Organizations are still struggling to keep up with their data and find effective storage solutions. Cloud computing has opened up even more options for big data. Developers may easily spin up ad hoc clusters to test a fraction of data in the cloud, which provides genuinely elastic scalability. With its ability to present huge volumes of data in a way that makes analytics rapid and thorough, graph databases are also becoming more essential. However, simply storing the data is insufficient. To be valuable, data must be used, and this is dependent on curation. It takes a lot of effort to get clean structured data or data that is relevant to the customer and arranged in a way that allows for useful analysis. This large volume of data has to be classified according to the type of data produced and its information. The big data content is mainly classified as critical, redundant, and dark data, as shown in Fig. 3.

3.2.1. *Critical data*

Critical structured business data is required information used to run a

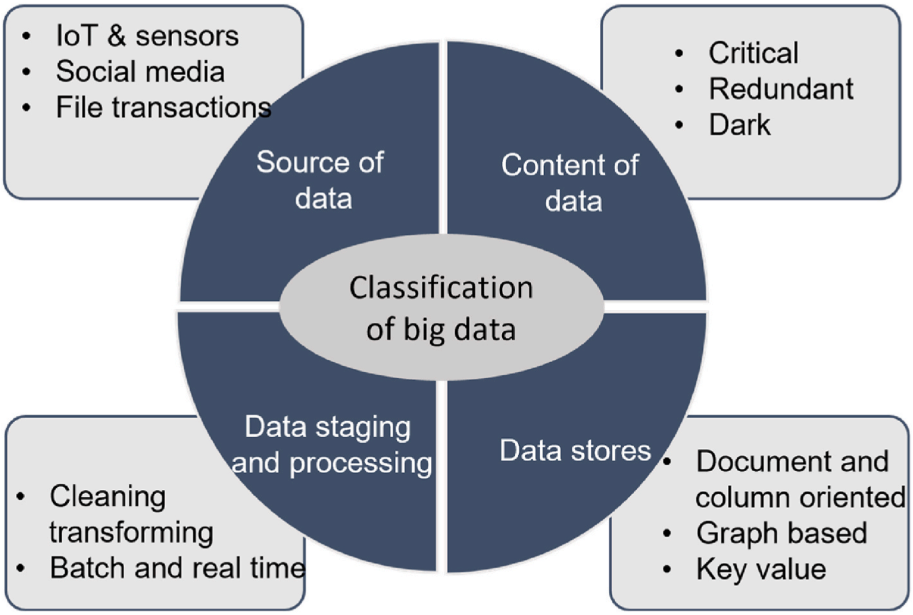

Fig. 2. The classifications of big data.

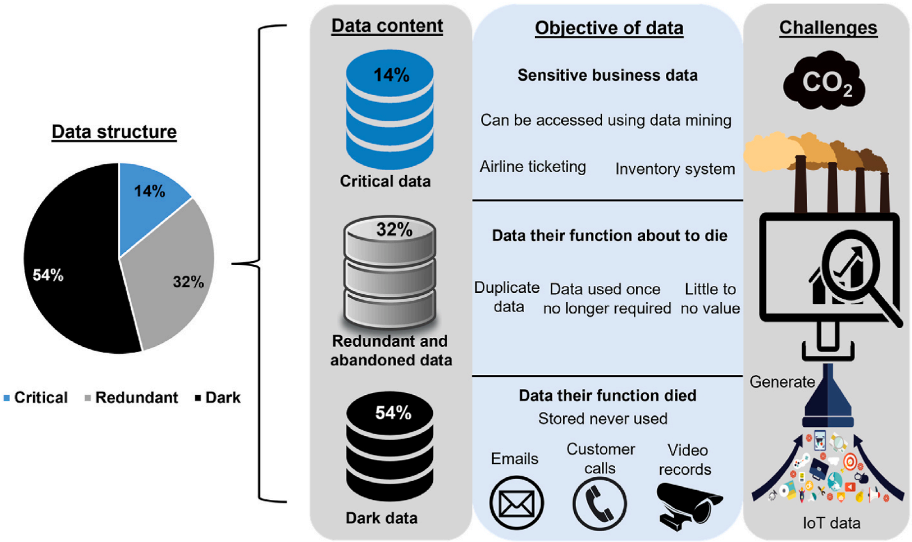

Fig. 3. General content format of big data classifications.

firm, guaranteeing that objectives are accomplished and allowing it to develop every year. The universe of big data is mainly concerned with critical data or data that is readily accessible to the user. The content of this type of data can be managed and used to enhance record management, maintain uniformity in filing, and ease access, storage, protection, and retention. Structured data are frequently maintained using structured query language, a programming language designed for handling and querying data in relational database management systems (Hashem et al., 2015). A study demonstrated by (Imdad et al., 2020) that organizations utilize less than 50% of their structured critical data for business analytics and decision making, while the rest of their structured data has the potential to become dark as well.

3.2.2. *Redundant data*

Redundant data is semi-structured data that may no longer be used or is irrelevant to the companies' data requirements. Semi-structured data is information that does not adhere to a traditional database structure. Structured data that is not arranged in relational database models is considered semi-structured data (Hashem et al., 2015). Using a fixed file format to capture semi-structured data for analysis is not the same as using a fixed file format to capture semi-structured data. As a result, acquiring semi-structured data necessitates the application of complicated rules that dynamically determine the next step after the data has been captured. Organizations store large volumes of data in semi-structured formats (i.e., CSV files, relational, and databases), and publish data on the Web in other semi-structured formats (i.e., XML, JSON, etc.). This type of data needs mapping languages and engines to modify, integrate, and feed data into knowledge graphs (Ryen et al., 2022).

3.2.3. *Dark data*

Finally, *dark* and abandoned data storage contains information that has been gathered or stored but not used to generate insights for decision making (e.g., data derived by synchrophasors). That said, dark data is generally defined as the scarcity of information that an organization develops and uses only once before being hidden among a massive and disorganized collection of other content assets (Goodwin, 2019). Dark data is untapped, buried, or undigested data for businesses since it has little value potential. Big data information assets are acquired, analyzed, and stored in normal business operations, particularly concerning digital transformation efforts, but are largely impractical for other uses. Business relationship management and analytics are also two examples. All this information will be aggregated over time as frequently retained by businesses for regulatory reasons, and it then lies inactive in storage hardware archives.

Unstructured data is also becoming more prevalent as a result of customer behavior in converting text, pictures, or music into a digital format for computer processing, social networking, search engine inquiries, and real-time streaming (Goodwin, 2019). This could also be due to the other tools users are used to, such as Instagram, TikTok, Facebook, and YouTube, which all pull content from a huge online sea instead of placing it within a structured hierarchy. Furthermore, on the university computers, there are hundreds of unorganized student files, and many of his items may contain enormous amounts of unstructured data. All of these assist in creating a large amount of digital data that needs to be stored and may not be accessed later. Although the terms data and information are frequently used interchangeably, data transforms into information when seen in context or analyzed to provide insights.

With the emergence of IoT, every industrial equipment (i.e., smartphones and smartwatches) utilized today can have data-gathering chips installed into it and broadcast all data via the Internet. The IoT has the potential to create hidden information in logs, metadata, text fields and documents, video, audio, and photographs. About 90% of generated data by IoT devices is never evaluated (Gimpel and Alter, 2021), and up to 60% of that data loses value within milliseconds of generation (Corallo et al., 2021). Storing such a high percentage of concealed data may be difficult to analyze as dark data far outnumbers the amount of visible data. While visible data may be easily accessed in databases, dark data requires a more sophisticated extraction process before being actively utilized. Storing and safeguarding such data usually comes at a higher cost and, in some cases, a higher risk than the data itself (Schembera and Duran, 2020). This is because part of this data might become more valuable and a target of theft and malware attacks (i.e., ransomware) (Goodwin, 2019).

Fig. 4 displays how anything transferred over the Internet has the potential to become unstructured dark data. Dark data is the most important subset of unstructured, accounting for 90% of all data (Gimpel and Alter, 2021). Yet less than 1% of it is ever accessed again (Imdad et al., 2020) for business analytics and decision making. Another analysis conducted by (Veritas, 2015) reveals that an average of 54% of stored data by worldwide enterprises is classified as dark since individuals in charge of it are unaware of its content and usefulness. According to a study conducted at the Stuttgart high-performance computing center, more than 49% of their user accounts are idle and classed as dark (Schembera and Duran, 2020). The annual CO_2

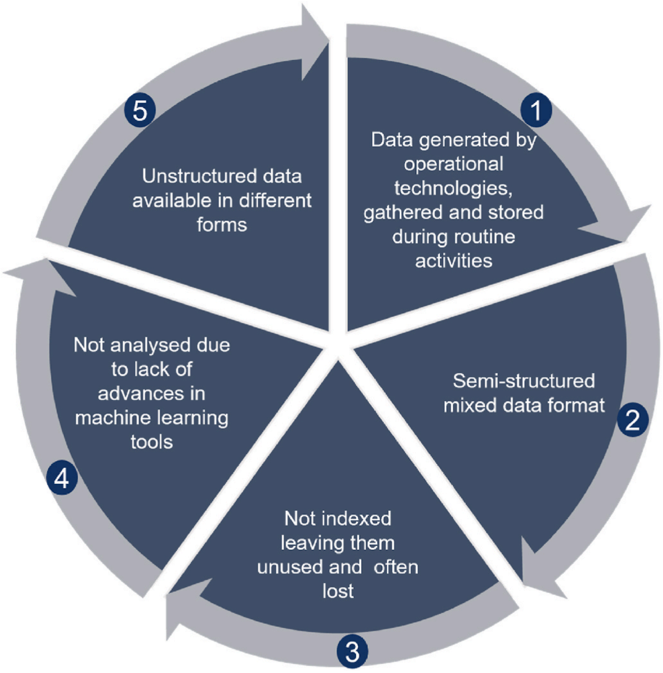
1
Data generated by operational technologies, gathered and stored during routine activities
2
Semi-structured mixed data format
3
Not indexed leaving them unused and often lost
4
Not analysed due to lack of advances in machine learning tools
5
Unstructured data available in different forms

Fig. 4. Process of dark unstructured data creation.

emissions from preserving this type of data are estimated to be around 5.8 MT (Veritas, 2015). To put things in perspective, an analysis by (Bayern, 2020) illustrates that 91 ZB of dark data would exist in the next five years, which is more than four times the quantity currently held. A large portion of this needs to be stored requiring significant storage space unless data users and organizations adjust their activities.

For example, consider a university with hundreds of modules and thousands of student coursework assignments produced each year. Previously, these were printed on paper, marked, and then either returned to the student or discarded or not for decades forever. When these became electronic, they were marked and most likely removed for a short period of time. This was most likely done through a custom web server at the institution. Then the storage went "cloud," and hard disk drive (HDD) prices are so cheap that no one ever deletes anything anymore. Coursework assignments from a decade ago may still exist and that is just the cloud copies. There are extra copies on local computers, email inboxes, and USB sticks. This means writing once, reading once, and never erasing the data. In contrast to the 1990s, institutions had to erase things because HDDs were so expensive and easy to fill up.

3.3. Data storage power consumption

Research by (Goodwin, 2019) demonstrated that worldwide enterprise data storage is rising at a 27% compound annual growth rate (CAGR). As indicated in Fig. 5, global data storage will reach over 11 ZB per year by 2025, up from 2.6 ZB in 2018, and will double every two to three years (Cooke et al., 2021) and (Reinsel et al., 2018). This indicates that the quantity of information stored in data centers is predicted to surge by 27% every year until 2025, while energy usage grew by 31% between 2017 and 2020. The majority of this data is eventually stored on traditional disc, cloud, or tape, necessitating enormous storage and management systems. Traditional storage solutions keep data on local physical discs at the client's main location, which is characterized by fast, manual security set up by the user, and can be recovered without accessibility issues (Benadjila et al., 2022). Data is mainly stored by users on disk-based technology, which is also used for data management and integration into the software. On the other hand, the cloud affords large storage capacity for users and access to data through separate geographical locations (Jalil et al., 2022). Additionally, it leverages the network to store the data on a service provider-owned remote server. With an internet connection, storage is simply available and more effective and simple to set up than traditional ones in terms of data security possibilities.

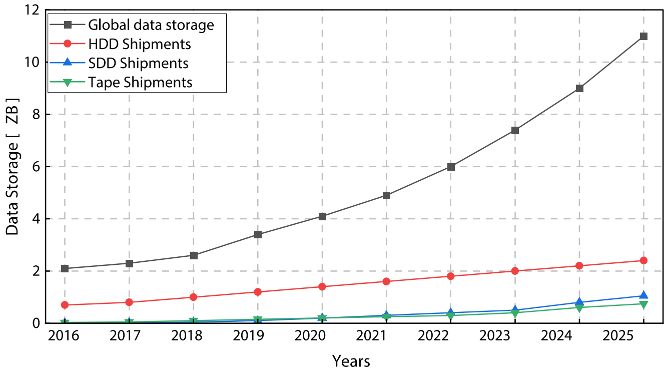

Fig. 5. Worldwide cumulative data storage requirements.

In a typical data center, a ratio of 40–50% of traditional HDD and solid-state drive (SSD) is usually available to store data (Shehabi et al., 2018). Currently, SSDs and HDDs serve very distinct purposes, since SDDs are generally used on "computing" servers and HDDs on "storage" servers. Each year, the energy usage of SSD drives decreases by roughly 2.3%, while that of HDD drives decreases by 5.3% (Shehabi et al., 2018). In 2016, SSDs began with an average of 6.0 Watt (W) per TB drive, whereas HDDs began with an average of 8.1 W per TB drive (Shehabi et al., 2018). Annually, the number of SSDs increases by 9.5%, while HDDs decline by 2.9%. A study by (Zhang and Yang, 2021) reported 27.8 kWh of power consumption per TB of data for information technology (IT) storage devices, whereas data centers utilized 46.33 kWh per TB of data per year, resulting in approximately 35 kg of CO_2 emissions per TB of data per year.

Data storage demand is expected to rise from 118.93, 235.63, and 309.14 Exabytes (EB) in 2016 to 368.47, 5023.40, and 24,840.67 EB in 2030 for traditional, cloud, and hyperscale data centers, respectively (Koot and Wijnhoven, 2021). Due to the deployment of more energy efficient SSD devices, this sudden increase in storage demand is anticipated to have a negligible impact on escalating storage device electricity demand in the near future. In contrast, a study by (Koot and Wijnhoven, 2021) claimed storage-related power consumption is expected to fall from 18.33 TWh in 2016 to 15.23 TWh in 2030, justified by the fact that demand will continue to be offset by ongoing efficiency improvements. Furthermore, hyperscale data centers are projected to consume the majority of storage-related energy due to global workload allocations. In 2030, hyperscale data centers are expected to utilize 12.72 TWh, compared to 0.13 and 2.39 TWh for traditional and cloud data centers, respectively (Koot and Wijnhoven, 2021). This indicates that data storage still accounts for a large fraction of total energy consumption with disk drives being the primary source of storage energy use.

Storing a massive amount of dark data on this type of storage is currently wasting a significant amount of energy, mostly powered by non-renewable resources, to run storage devices and associated storage management systems and thus increase CO_2 emissions. This, in addition to the heat produced as a by-product of production, traffic, and storage, necessitates cooling. Storage drives power demand could increase to 19% of overall data center energy consumption if storage cooling share infrastructure is taken into account (Backup Works Storage Solutions, 2020). When compared to general compute loads, write operations in data center storage systems can be somewhat energy intensive due to the redundant array of inexpensive disks (RAID) setups that require the calculation of parity checks and other calculations. The energy consumption once written to disk is, of course, difficult to determine. It depends on whether the end user pays for the drives to be online and spinning up, or whether they allow the dark data to be placed in a glacier-like environment where it can be retrieved but with high latency (because the disks are shut down).

Thus, for large-scale data centers and cloud providers, finding ways to use electricity more effectively is critical to make business and reducing bills. Given the size of these facilities, seemingly minor modifications can have a significant impact on the cost of environmental issues and carbon footprint. Although cleaner energy sources are an essential area of focus for green data centers, reducing waste and optimizing resources can also play a big part in the progress of the green data center. Combining improved energy efficiency with greener energy sources yields the best results in terms of cost savings and carbon emissions reduction. Due to low-cost storage media, minimum power consumption, and low cooling requirements, today's highly improved magnetic tape has witnessed a resurgence in the marketplace in recent years (Cooke et al., 2021). The tape storage does not use any electricity when not in use (although, so does an HDD when you power it off) and has low embodied energy per tape. However, the main cons of the tape storage medium are extremely slow if the tape is in the drive, if the tape needs to be loaded by a robot, and if the tape requires to be loaded by a human, maybe from a distant warehouse.

3.4. *Contextualizing digitalization through a sustainability lens*

Maintaining large amounts of data or transferring them over a network or the cloud can use a lot of energy and result in a significant environmental footprint. However, storage and processing requirements can be reduced by eliminating data storage and flows that are no longer required. Progressive firms who are committed to improving their sustainability are, according to an analysis conducted by (Goodwin, 2019), projected to reassess their current data storage techniques and explore using more modern magnetic tape as a viable option for achieving long-term success. For regulatory and governance reasons, analytics, and other use cases, businesses may choose to keep dark unstructured data on nearline media or the "active archive" tier of storage. This information does not require the use of disk drives, which require continuous power and cooling to operate. This means that while the data is not in use, it is still available for productive usage and consumes minimum processing power and environmental resources.

Given the emphasis on sustainability and the massive quantities of data storage devices needed to store the growing amounts of data in the coming years, organizations have an opportunity to decarbonize, improve sustainability, and lower the costs by migrating less frequently visited data from HDD-based storage to sophisticated tape storage systems. As a hypothetical example, research by (Johns, 2021) looked at the impact of storing 100 PB of data for ten years on the storage media. An active archive that migrated 60% of the HDD resident data to tape storage decreased CO_2 emissions by 57% and electronic waste by 48% when compared to HDD-based storage, as illustrated in Fig. 6. However, if all the data is presumed to be dark and moved to tape, CO_2 emissions and e-waste are decreased by 95% and 80%, respectively. It is also worth noting that this research only looked at emissions from storage media, not the IT infrastructure that supports them.

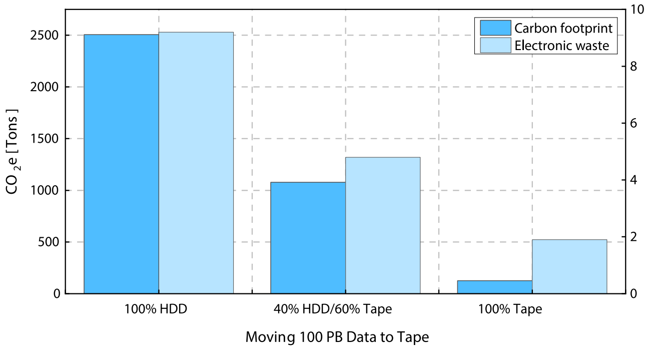

Fig. 6. Ten years 100 PB of data transferring to tape storage.

When deciding on data storage infrastructure, the long-term impact on energy usage should be taken into account. International Data Corporation (IDC) conducted in-depth research and constructed a scenario to better understand the impact that tape storage could have on CO_2 emissions if more data is shifted to tape storage. Shifting to tape can result in a large, observable change in energy resource consumption when considering overall resource usage over the life term of data storage (Goodwin, 2019). The reductions in energy costs, as well as the reduction in CO_2 emissions, are compelling reasons to consider expanding the usage of tape storage. Data migration has an immediate positive impact, resulting in lower electricity consumption. The annual CO_2 reduction by 2030 is 43.7% if an increasing proportion of data is considered to be archival, with 80 percent of archived data kept on enterprise storage systems and 57% of replicated data transferred to tape (Cooke et al., 2021). This indicates that between 2019 and 2030, 664 MT of CO_2 emissions might be avoided. This quantity is equal to the annual greenhouse gas emissions from 144 million passenger automobiles or the annual energy consumption of 80 million houses.

Furthermore, the growing demand for data storage and server cooling has prompted some global cloud infrastructure to be migrated to temperate climate regions such as Sweden, Iceland, and Ireland, which rely on natural cooling to reduce heat instead of using high-powered cooling equipment all the time (Vonderau, 2019). Data migration refers to the process of transferring data from one place, one format, or one application to another as a result of introducing new systems or locations for the data. The process involves data profiling, data cleaning, data validation, and the ongoing data quality assurance process in the destination system. Data migrations are widely deployed nowadays as businesses switch from on-premises infrastructure and applications to cloud-based storage and apps in an effort to optimize or transform their business (Maniah et al., 2022).

Migrating a large volume of data between different geographical locations, for the purpose of environmental footprint reduction, sounds simple in theory, but the approach does not always work as many countries legislate that their citizens' data must be stored domestically (Ali and Osmanaj, 2020). Furthermore, due to data gravity, data migration has been perceived as a challenge and a risk. Even though the aspect of data gravity has been around for a while, the dilemma is becoming more noticeable as data migrates to cloud infrastructures (Laurent et al., 2020). Data gravity is a metaphor that illustrates how data attracts other data, how data is integrated into a business, and how data becomes more tailored over time. Gartner advises "disentangling" data and applications as a way to combat data gravity and shift apps and data to more beneficial locations (Gartner, 2017). Finally, transferring unstructured data is one of the most difficult challenges that businesses will face when migrating their data to the cloud. The challenge with migrating unstructured data is that most public cloud providers do not prioritize unstructured data migration, instead focusing on the accessibility and scalability of critical data.

4. Methodology to assess environmental footprint of data storage

This subsection presents the methodology deployed to identify global environmental footprints associated with different types of data storage. Some of the earliest projections of the world's CO_2 emissions (The Climate Group, 2018) and energy consumption (Bordage, 2019) were based on imprecise, vague, and out-of-date data and lacked the transparency necessary to be relied upon. Malmodin and Lunden deployed the publicly available datasets from the ICT industries to estimate the carbon footprint of each device connected to a data center (Malmodin and Lundén, 2018). The analysis is based on a broad dataset that incorporates primary and secondary data for operational energy usage and life cycle CO_2 for the covered sub-sectors. Belkhir and

Elmeligi were built on this while estimating the annual lifecycle footprint for each desktop, including data centers and networking equipment, using the following quantities for each component of the ICT industry: production energy, component lifecycle, and annual energy consumption (Belkhir and Elmeligi, 2018). However, the main issue with these studies is that they assume that most electric energy is generated by fossil fuels, which is not the case in most renewable-dominant countries such as Brazil. This means that the electricity mix is an important consideration in addressing the actual carbon footprint of the ICT sectors, as most data center industries are still powered by nonrenewable resources (Obringer et al., 2021). This is because the power generation mix varies depending on the country's strategy for approaching sustainability goals. Thus, the main objective of this study is to estimate the environmental footprint of data storage in some data center-dominant countries by using the global average electricity mix and emission factors from various sources of electricity production.

As shown in Fig. 7, two main steps are used to determine the environmental impact of the worldwide electric power generation mix, followed by calculating total storage energy usage. The first step elaborates on the analysis conducted by (Obringer et al., 2021), specifically looking at the global power generation mix in (kWh) from various conventional and renewable energy resources using the most recent global data for 2020. The breakdown of the electricity mix by country (Ritchie and Roser, 2021), renewable power generation by technology (Bahar and Bojek, 2020), and accumulated data (Ember, 2021). These data are then utilized to determine carbon, water, and land footprints for each source of power generation (Obringer et al., 2021), (Ristic et al., 2019), and the total land use (Fritsche et al., 2017). It is worth noting that, in this phase, the overall environmental footprints of electric power sources are calculated using a range of truncated log-normal distribution (i.e., min, median, and max) values to bound the footprint of each component between the minimum and maximum values reported in the literature.

In step II, global power generation data is used to determine the required energy consumption for data centers to keep a gigabyte of data storage alive in kWh/GB. Although research conducted by (Koot and Wijnhoven, 2021) estimated the worldwide data storage power consumption to be 18 TWh, the study did not define what is deemed within the scope of the estimates provided and the real storage energy requirements may be underestimated. For example, given that storage devices account for 11% (Shehabi et al., 2016) of total data center power consumption (i.e., 205 TWh) excluding their share of cooling infrastructure, the total energy usage of storage drives might be as high as 22.55 TWh. Therefore, different from (Obringer et al., 2021), in this analysis, the required energy in kWh/GB values is calculated by dividing 22.55 TWh by the total data storage capacity of 4.9 ZB, yielding 0.0046 kWh for each GB data storage on a disc drive (Cooke et al., 2021). It is worth noting that a tape storage system with the same capacity uses only 0.0006 kWh, meaning an 87% reduction in energy use (Backup Works Storage Solutions, 2020). The metric kWh/GB is then used to calculate the water and land footprints L/GB and cm^2/GB, respectively, for various power generation types using a variety of water intensity L/kWh metrics and m^2 of land coverage. This means that the scope of the analysis begins with a broad perspective of environmental footprint due to worldwide data storage capacity and progresses to the effect of accumulated data storage on the environment within each country in subsequent analysis. The supplementary file contains the complete computation techniques deployed to determine the figures in this research.

The analysis is first performed assuming that all the data is stored on disk drives, considering power consumption determined in step II. In the second scenario, 80% of the data to be stored is designated as archival, with 43% of this data to be stored on business storage drives and 57% of replicate data to be migrated to tape storage (Vries and Stoll, 2021). Finally, the share of energy consumption due to dark, redundant, and crucial data storage is calculated from data available in (Veritas, 2015) and (Veritas, 2020), based on the premise that each data type represents 54%, 32%, and 14% of the current global data storage, respectively. The carbon, water, and land footprint for each data type are computed based on the percentage of energy consumption to store these types of data including the avoided footprint due to shifting to tape storage. The same methodology is implemented to compute the energy required for data storage as well as the environmental footprints in some dominant countries based on the electric power generation mix of each country. Percentage values are used to determine the environmental impact of each country concerning global footprint values, which will be shown in the next section.

5. Analysis and results

Fig. 8 depicts the global environmental footprints of storing various data types. As demonstrated, the water footprint in the minimum and median scenarios is relatively small for all data storage classifications, which are less than 0.05 L/GB. These relatively tiny footprints are rather big compared to the massive amounts of multi-gigabyte data connected with Internet use. However, this figure has increased to nearly 2.44 L/GB during the maximum scenario. When this number is further broken down, it can be seen that almost 54% of the total water footprints are due to dark data storage while 32% considered for redundant data which is around 0.78 L/GB followed by critical data storage of 0.34 L/GB which accounted for 14% of the total maximum water footprints. These results indicate that energy consumption of worldwide data storage

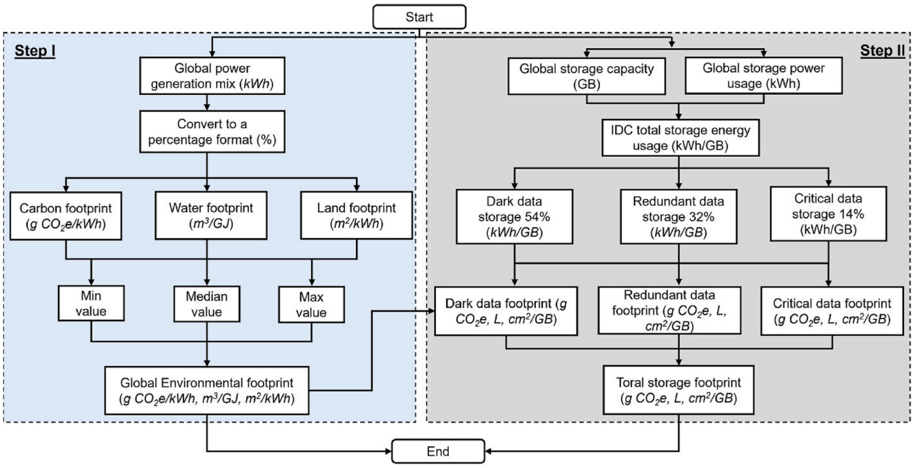
Start
Step I
Global power generation mix (kWh)
Convert to a percentage format (%)
Carbon footprint (g CO2e/kWh)
Water footprint (m3/GJ)
Land footprint (m2/kWh)
Min value
Median value
Max value
Global Environmental footprint (g CO2e/kWh, m3/GJ, m2/kWh)
Step II
Global storage capacity (GB)
Global storage power usage (kWh)
IDC total storage energy usage (kWh/GB)
Dark data storage 54% (kWh/GB)
Redundant data storage 32% (kWh/GB)
Critical data storage 14% (kWh/GB)
Dark data footprint (g CO2e, L, cm2/GB)
Redundant data footprint (g CO2e, L, cm2/GB)
Critical data footprint (g CO2e, L, cm2/GB)
Toral storage footprint (g CO2e, L, cm2/GB)
End

Fig. 7. Methodology deployed to compute environmental footprint of data center storage devices.

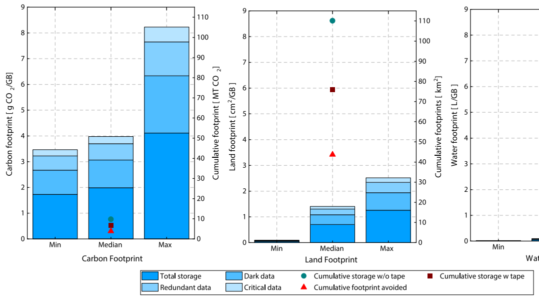

Fig. 8. Minimum, median, and maximum CO2, water, and land footprints due to different data storage categories including cumulative and avoided footprints with tape storage.

would require about 77.13 GL of water in the median scenario. The water footprint attributable to dark data storage, redundant data, and critical structured data each account for 41.65, 24.68, and 10.79 GL, respectively, in the figure.

Fig. 8 also reveals that in the minimum scenario, the land footprint due to data storage presents only a small amount of less than 0.045 cm^2/GB. However, this figure has increased significantly to 0.7 cm^2/GB and 1.25 cm^2/GB during the median and maximum footprint scenarios. Meaning that in the median scenario total data storage results in an annual median land footprint of about 110.1 square kilometers (km^2), with dark data accounting for 59.45 km^2 followed by 35.23 and 15.41 km^2 due to redundant and critical data storage, respectively.

Differently, the total carbon footprint due to data storage is significantly higher than water and land footprints. As shown, the carbon footprint due to data storage reaches nearly 1.73 g CO$_2$/GB in the minimum scenario while this figure has raised to approximately 1.98 g

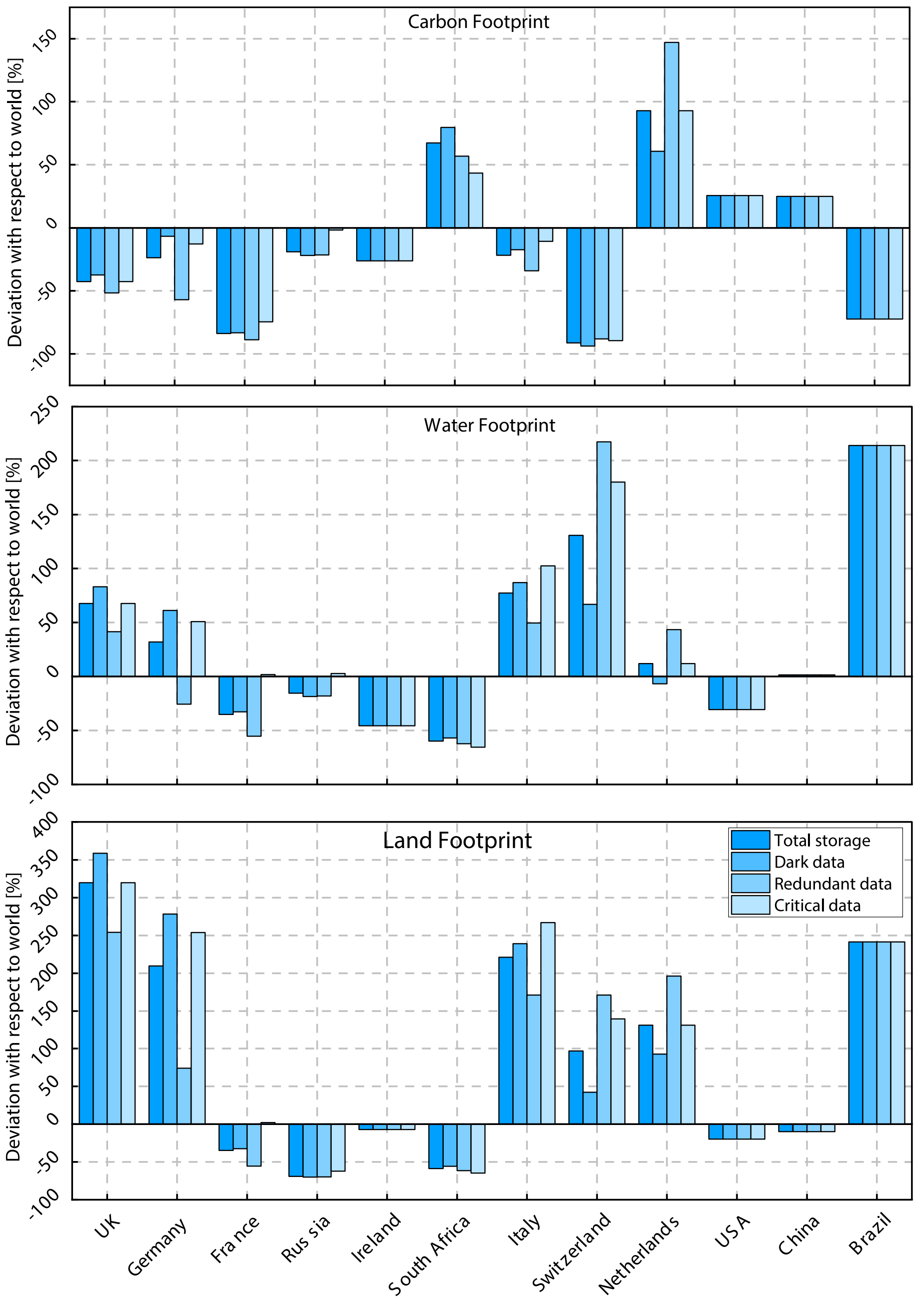

Fig. 9. Carbon, water, and land footprint deviations for various countries concerning the world median.

CO_2/GB and 4.11 g CO_2/GB during both the median and the maximum scenarios, respectively. These values are used to determine that worldwide data storage results in annual CO_2 emissions of 8.49, 9.74, and 20.15 MT during the minimum, median, and maximum scenarios, respectively. Breaking down these numbers, it is possible to identify that dark data alone is responsible for 5.26 MT of CO_2 emission during the median scenario. Furthermore, redundant and critical data storage could result in 3.12 and 1.36 MT of CO_2 emissions, respectively. It is worth noting that these figures are comparable to those found in a study conducted in 2015 (Veritas, 2015).

Finally, as can be seen from the right side of Fig. 8, expanding the use of tape storage can positively affect sustainability goals and the impact can be recognized immediately with a significant reduction in CO_2 from 9.74 MT to 6.7 MT, water footprint from 77.13 GL to 53 GL, and land footprint from 110.1 km^2 to 75 km^2. This enhancement is due to the reduced electricity consumption while migrating data to tape storage devices. It is also worth noting that these figures exclude the impact of power consumption from cooling and infrastructure supporting storage data, implying that the impact of data migration to tape storage could be much larger when power consumption from cooling systems is factored in. Given the massive amounts of data on the horizon, businesses should assess whether they need all of it "live" and available, or whether some, if not all, can be stored as an archive in a tape environment.

The analysis further examines how individual countries' footprints compare to the total global footprints due to various data storage categories. The top panel of Fig. 9 depicts variances in carbon footprint in a few data center dominant countries compared to overall world calculations, while the middle and bottom graphs show comparable calculations for water and land footprints. As shown, in terms of the environmental footprints of an average unit of energy needed for storing dark data, some countries perform better than others due to differences in the energy mix in 2020. As an example, the carbon footprint of total data storage in the median scenario in the UK is roughly 50% lower than the global median, while the water footprint and land footprints are nearly 50% and 300% higher than the global median, respectively. The carbon footprint has declined marginally compared with those obtained by (Obringer et al., 2021) for the year 2018, but the water and land footprints have increased slightly. This is because the total electricity generation in the UK fell by 3.6% between 2019 and 2020 (Statistics, 2021), while demand for data center usage has increased by 20% due to the COVID19 pandemic's complete lockdown (Kang et al., 2020). The share of electricity produced by coal, gas, and nuclear fell by 0.3%, 5%, and 1.3%, respectively. This enabled higher power generation from renewables, which climbed from 36.9% to 43.1%.

Due to the low level of renewable energy in the Netherlands (i.e., 25%), the median carbon footprint was nearly 100% higher than the global median. Similarly, the land footprint was substantially higher than the world median, while the water footprint remains close to the world median. In Ireland, Switzerland, and Brazil, on the other hand, the share of renewable power generation accounted for 40.7%, 65.1%, and 84.3% in 2020, respectively. Therefore, it is obvious that the carbon footprint of these countries is significantly lower than the world median values. Instead, the water footprint of storing data in Brazil is 214% higher than the global median. The country obtained more than 64% of its electricity from hydropower plants, resulting in a higher water footprint and lower carbon footprint compared to other countries. Indeed, hydropower plants use more water than conventional power plants during the electricity generation process. Furthermore, the water from a hydro station is basically clean and can be used downstream, which is not the same as foul water from an industrial process. Clearly, these differences emphasize the influence of diverse energy mixes on the overall footprints of the data center as the higher percentage of renewable power generation leads to smaller environmental footprints.

A closer inspection of these graphs reveals that the share of the environmental footprint due to various data storage options fluctuates significantly concerning the world median scenario. For instance, South Africa's carbon footprint resulting from overall data storage is 67% greater than the worldwide median, while its dark data storage deviance is 79.6% higher. This is followed by redundant and critical data storage, both of which are 56% and 43% greater than the global median, respectively. Dark, redundant, and critical data storage, however, have smaller water and land footprints than the global median, with −56.9%, −62.3%, and −65.6%, respectively. Turning now to the UK, Germany, and Italy, it is obvious that dark data storage has significantly larger water and land footprints than the worldwide median when compared to redundant and critical data footprints. These differences highlight the impact of the different shares of dark, redundant, and critical data storage in each country on the overall storage footprint of data centers. When the UK and Ireland are compared, the results demonstrate that the UK has a significantly larger land footprint than Ireland. This is mainly because the UK produced more than 46 TWh of its electricity from hydro and bioenergy in 2020, which are among the main sources of electricity that result in large land footprints (Holmatov et al., 2019). However, during the same year, these sources provided less than 2 TWh of Ireland's electricity.

Dark data, for example, is expected to account for about 59% of total data storage in the UK, while redundant and crucial data account for 27% and 14%, respectively (Veritas, 2015). Meanwhile, France has one of the highest percentages of clean and identifiable business critical and redundant data storage, accounting for 22% and 22% of total data storage, while dark data is only 56% (Veritas, 2015). Finally, smaller countries like Switzerland fare better in dark data, accounting for only 39% of overall data storage, followed by 44% redundant data and 17% vital data (Veritas, 2015). Comparing the disparities between data storage footprints not only illustrates the trade-off between different sources of power generation but also depicts the importance of evaluating many environmental footprints for storing each data type simultaneously. This is different from the conventional practice of focusing solely on the carbon footprint of the complete data center.

6. Discussions and recommendations

Global warming is a serious challenge. Governments are applying new, more expensive reporting requirements and regulations, while consumers prefer more robust government policies that preserve the environment. To address sustainability, companies are introducing a number of programs and efforts to better understand and manage the impact of their activities on the environment throughout their lifecycles. Firms can choose from a variety of different alternative projects to tackle and conform to these sustainability issues, which will vary depending on the sector and business. The perfect sustainability program, on the other hand, will drastically reduce carbon emissions, minimize a product's environmental impact, save costs, and be simple to implement. One of these solutions is to eliminate dark data and shift regularly visited archive data from the disc to tape storage, especially for organizations with a considerable amount of stored data. To take advantage of this possibility, two core recommendations arise from our analysis concerning the environmental impacts associated with holding dark data and Internet sustainability more generally.

First, businesses and data centers firms can start to make a major difference by simply taking control of data storage, assessing the storage rules, and ensuring they are not retaining data that is no longer needed, as per Fig. 10. In fact, as far as the data center is concerned, the more data their customers store the better that is for business. However, this is a terrible thing as far as the climate is concerned, yet that is the business model they operate on. Eliminating data centers' unstructured data improves regulatory compliance and lowers costs and helps mitigate emissions and safeguard the environment. Data centers must begin to improve their data management policies, utilize the correct technologies to identify which data adds value, and eliminate dark data from their data centers to avoid emissions spiraling out of control and avoid digital waste. Filtering dark data and removing unnecessary information should

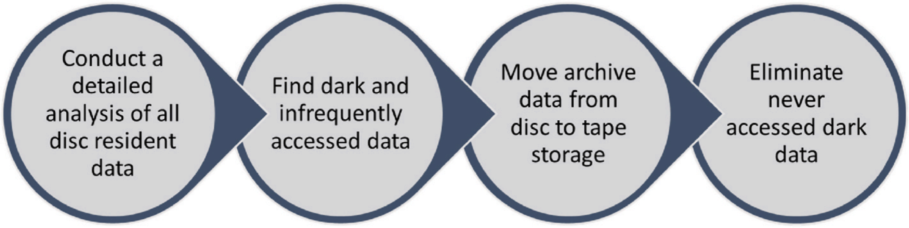
Conduct a detailed analysis of all disc resident data
Find dark and infrequently accessed data
Move archive data from disc to tape storage
Eliminate never accessed dark data

Fig. 10. Data assessment to minimize dark data storage.

become a moral responsibility for organizations worldwide. Although artificial intelligence is there for data cleaning, the problem needs more radical solutions such as notifying the public, the government, and the industry about the environmental concern of storing massive redundant and dark data.

There are, however, ways to utilize less while rendering data centers more sustainable and, as a result, more competitive in the long run. But the question is how to archive data and why even we store never accessed again data? Perhaps international standards or national approaches are needed to protect and archive sensitive and important data on modern tape storage devices. Adopting tape storage early enough is prudent if a business needs to save redundant data for a long period of time. It is evident that if enough businesses around the world adopted this policy, the environmental impact would be significant. Considering the prevailing events and increased focus on sustainability now is the best time to review IT data storage policies and migrate infrequently viewed data to contemporary tape storage or eliminate them entirely. IT executives should also work with other business stakeholders to create data quality programs that are efficient and sustainable while achieving intended outcomes.

Data centers are not really responsible for any of this though, it is their customers who need to improve their data management policies. Thus, the second critical step in avoiding an irreversible path to an unsustainable digital world is for society to recognize the power of collective action in reducing the environmental impact of holding dark data. People's widespread adoption of environmentally responsible online conduct is vital for preventing climate change and maintaining long-term sustainability. Making Internet users aware of the costs of online acts and the advantages of making tiny behavioral changes through information campaigns, behavioral nudges, and other means is crucial to fostering sustainable digital behavior. As cloud storage is so cheap and widely available, consumers play a role in storing thousands of films and digital photographs that will never be viewed. In tandem, people should be aware of emails, instant messages, documents, presentations, and spreadsheet that will never be read and lose track of what have been saved along the way. Small steps like removing emails and unneeded content on cloud-based storage services, unsubscribing from email lists, and turning off videos during online meetings can help lessen the environmental impact of Internet use. A tiny action like taking a photo with a smartphone and posting it on social media generates two types of dark data: the post and the image itself and the metadata around it. Essentially, the metadata is basically trivial compared to the volume of data in the image itself. It is worth stating, however, that the metadata is essentially insignificant in comparison to the abundance of data in the image itself. Presently people are being encouraged to upload and share stuff constantly so that it creates data for advertising algorithms to target them with, and also feeds into other machine learning projects, etc. However, for the sake of the environment, businesses and people worldwide must manage their data daily in order to avoid creating dark data in the first place or deleting it.

7. Conclusions

The data center sector consumes an increasing amount of land, electricity, water, and raw materials while also producing an increasing amount of waste. In summary, despite great improvements in data center efficiency, this sector is still responsible for a large amount of CO_2 emissions and contributes significantly to global water use and land footprints. This research showed that, if not adequately managed, the worldwide CO_2 emissions resulting from the storing of dark data may exceed 5.26 MT per year. This is followed by water and land footprints each with 41.65 GL, and 59.45 km^2, respectively. This study also examined the footprint deviations for various data center dominant countries concerning total global parameters. It showed that for a country like the United Kingdom when 43% of its electricity is produced from renewables in 2020, carbon footprint deviation might be modest compared to the entire global footprint, yet water and land footprints are considerable. Finally, as we increase reliance on data centers the problem with their environmental footprints will be more and more exposed, so it is critical that the prioritization process sieve out redundant and dark data in a sensitive manner so that we archive data sustainable for future generations. Even though data center owners and operators have made a significant effort to decarbonize their power consumption, they continue to rely heavily on fossil fuel power plants to keep data servers up, running, and cooling. Furthermore, power consumption is the problem, and the users are constantly encouraged to consume through a variety of IoT applications. Thus, future analysis should continue to examine the environmental impact due to powering different user applications. Future research should also assess the environmental impact of data storage on digital media against hard copies of printed material data archives.

CRediT authorship contribution statement

Dlzar Al Kez: Investigation, Writing – original draft, Visualization. **Aoife M. Foley:** and, Supervision, Conceptualization, Investigation, Writing – review & editing, Funding acquisition. **David Laverty:** Supervision, Conceptualization, Investigation, Writing – review & editing, Funding acquisition. **Dylan Furszyfer Del Rio:** and, Conceptualization, Writing – review & editing. **Benjamin Sovacool:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is part of the Collaborative REsearch of Decentralization, ElectrificatioN, Communications and Economics (CREDENCE) project, which is funded by a US-Ireland Department for the Economy (DfE), Science Foundation Ireland (SFI) and a US National Science Foundation

(NSF) award under the Research and Development Partnership Program (Centre to Centre) award (grant number USI 110). The views and opinions expressed in this paper do not necessarily reflect those of the DfE, SFI, and NSF.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2022.133633.

References

Abdulsalam, Y., Shailendra, S., Shamim, H.M., Ghulam, M., 2019. IoT big data analytics for smart homes with fog and cloud computing. Future Generat. Comput. Syst. 91, 563–573.

Al Kez, D., et al., 2021. Potential of data centers for fast frequency response services in synchronously isolated power systems. Renew. Sustain. Energy Rev. 151, 111547.

Al Kez, D., Foley, A.M., Morrow, J.D., 2022. Analysis of fast frequency response allocations in power systems with high system non-synchronous Penetrations. IEEE Trans. Ind. Appl. 58 (3), 3087–3101.

Al Kez, D., Foley, A.M., S.M., M., Morro, J.D., 2020. Manipulation of static and dynamic data center power responses to support grid operations. IEEE Access 8, 182078–182091.

Ali, O., Osmanaj, V., 2020. The role of government regulations in the adoption of cloud computing: a case study of local government. Comput. Law Secur. Rep. 36, 105396.

Amazon, 2022. Amazon extends position as world's largest corporate buyer of renewable energy [Online] Available at: https://www.aboutamazon.com/news/sustainability/amazon-extends-position-as-worlds-largest-corporate-buyer-of-renewable-energy. (Accessed 28 July 2022).

Aslan, J., Mayers, K., Koomey, J.G., France, C., 2018. Electricity intensity of internet data transmission: untangling the estimates. J. Ind. Ecol. 22 (4), 785–798.

Au-Yong-Oliveira, M., et al., 2021. The potential of big data research in HealthCare for medical doctors' learning. J. Med. Syst. 45 (13).

Backup Works Storage Solutions, 2020. Reducing data center energy consumption and carbon emissions with modern tape storage [Online] Available at: https://www.backupworks.com/tape-storage-reduce-energy-consumption-and-carbon-emissions.aspx. (Accessed 19 December 2021).

Bahar, H., Bojek, P., 2020. Renewable power [Online] Available at: https://www.iea.org/reports/renewable-power. (Accessed 25 September 2021).

Bayern, M., 2020. 6.4M tons of CO2 will pollute the atmosphere in 2020 due to dark data [Online] Available at: https://www.techrepublic.com/article/6-4m-tons-of-co2-will-pollute-the-atmosphere-in-2020-due-to-dark-data/. (Accessed 3 October 2021).

Belkhir, L., Elmeligi, A., 2018. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J. Clean. Prod. 177, 448–463.

Benadjila, R., Khati, L., Vergnaud, D., 2022. Secure storage—confidentiality and authentication. Computer Science Review 44, 100465.

Bloomberg, 2021. Data Centers and Decarbonization, Unlocking Flexibility in Europ's Data Centers. BloombergNEF, New York, United States.

Bordage, F., 2019. The Environmental Footprint of the Digital World. Green IT, Corenc, France.

Cooke, J., Goodwin, P., Nadkarni, A., Sheppard, E., 2021. Accelerating Green Datacenter Progress with Sustainable Storage Strategies. IDC Corporate, Massachusetts, United States.

Corallo, A., et al., 2021. Understanding and defining dark data for the manufacturing industry. IEEE Trans. Eng. Manag. 1–13.

Cunliff, C., 2020. Beyond the Energy Techlash: the Real Climate Impacts of Information Technology. Information Technology and Innovation Foundation, USA.

Dayarathna, M., Wen, Y., Fan, R., 2016. Data center energy consumption modeling: a survey. IEEE Communications Surveys & Tutorials 18 (1), 732–794.

Ember, 2021. Renewables overtook fossil fuels [Online] Available at: https://ember-climate.org/european-electricity-transition/. (Accessed 26 September 2021). Accessed.

Fritsche, U.R., et al., 2017. Energy and Land Use. IRENA, Abu Dhabi.

Gartner, 2017. Liberate applications for migration by disentangling data [Online] Available at: https://www.gartner.com/en/documents/3840464. (Accessed 26 July 2022).

Gimpel, G., Alter, A., 2021. Benefit from the internet of things right now by accessing dark data. IT Professional 23 (2), 45–49.

Goodwin, P., 2019. Tape and Cloud: Solving Storage Problems in the Zettabyte Era of Data. IDC Corporate, Massachusetts, United States.

Google, 2021. Environmental Report. Google, California, United States.

Hashem, I.A.T., et al., 2015. The rise of "big data" on cloud computing: review and open research issues. Inf. Syst. 47, 98–115.

Holmatov, B., Hoekstra, A.Y., Krol, M., 2019. Land, water and carbon footprints of circular bioenergy production systems. Renew. Sustain. Energy Rev. 111, 224–235.

IEA, 2020. Data Centres and Data Transmission Networks. International Energy Agency, Paris.

Imdad, M., et al., 2020. Dark data: opportunities and challenges. Int. Res. J. Comput. Sci. Technol. 1 (1).

Innovation News Network, 2021. The big data challenge at the large hadron collider [Online] Available at: https://www.innovationnewsnetwork.com/big-data-challenge-large-hadron-collider/11359/. (Accessed 7 June 2022).

Jalil, B.A., Hasan, T.M., Mahmood, S.G., Abed, H.N., 2022. A secure and efficient public auditing system of cloud storage based on BLS signature and automatic blocker protocol. J. King Saud Univ. 34, 4008–4021.

Johns, B., 2021. Improving Information Technology Sustainability with Modern Tape Storage. Consulting LLC, Tucson, Arizona.

Jones, E., 2022. Google Cloud vs AWS in 2022 (Comparing the Giants) [Online] Available at: https://kinsta.com/blog/google-cloud-vs-aws/. (Accessed 28 July 2022). Accessed.

Jones, N., 2018. How to stop data centres from gobbling up the world's electricity. Nature 561, 163–166.

Kang, C., Alba, D., Satariano, A., 2020. The New York times [Online] Available at: https://www.nytimes.com/2020/03/26/business/coronavirus-internet-traffic-speed.html. (Accessed 29 September 2021).

Koot, M., Wijnhoven, F., 2021. Usage impact on data center electricity needs: a system dynamic forecasting model. Appl. Energy 291, 116798.

Laurent, A., Libourel, T., Madera, C., Miralles, A., 2020. The gravity principle in data lakes. In: Laurent, A., Laurent, D., Madera, C. (Eds.), Data Lakes. John Wiley & Sons, Inc, Hoboken, pp. 187–199.

Malmodin, J., Lundén, D., 2018. The energy and carbon footprint of the global ICT and E&M sectors 2010–2015. Sustainability 10 (9), 3027.

Malmodin, J., Lunden, D., 2018. The energy and carbon footprint of the global ICT and E&M sectors 2010–2015. Sustainability 10, 3027.

Maniah, Soewito, B., Gaol, F.L., Abdurachman, E., 2022. A systematic literature Review: risk analysis in cloud migration. J. King Saud Univ. – Comput. 34, 3111–3120.

Microsoft, 2020. Environmental Sustainability Report (Washington, United States: Microsoft).

Nagorny, K., Lima-Monteiro, P., Barata, J., Colombo, A.W., 2017. Big data analysis in smart manufacturing: a review. Int. J. Commun. Netw. Syst. Sci. 10 (3).

Obringer, R., et al., 2021. The overlooked environmental footprint of increasing Internet use. Resour. Conserv. Recycl. 167, 105389.

Oracle, 2022. What is big data? [Online] Available at: https://www.oracle.com/uk/big-data/what-is-big-data/. (Accessed 5 June 2022).

Reinsel, D., Gantz, J., Rydning, J., 2018. The Digitization of the World, from Edge to Core. IDC Anlayze the Future, USA.

Renze, M., 2019. The data science hierarchy of needs [Online] Available at: https://matthewrenze.com/articles/the-data-science-hierarchy-of-needs/. (Accessed 8 June 2022).

Ristic, B., Kaveh, M., Zen, M., 2015. The Water Footprint of Data Centers, Sustainability, p. 11260, 1128.

Ristic, B., Mahlooji, M., Gaudard, L., Madani, K., 2019. The relative aggregate footprint of electricity generation technologies in the European Union (EU): a system of systems approach. Resour. Conserv. Recycl. 143, 282–290.

Ritchie, H., Roser, M., 2021. *Electricity Mix.* [Online] Available at: https://ourworldindata.org/electricity-mix. (Accessed 25 September 2021).

Rotem-Gal-Oz, A., 2015. Data's hierarchy of needs [Online] Available at: https://arnon.me/2015/06/data-hierarchy-of-needs/. (Accessed 8 June 2022).

Ryen, V., Soylu, A., Roman, D., 2022. Building semantic knowledge graphs from (Semi-) Structured data: a review. Future Internet 14 (5), 129.

Schembera, B., Duran, J.M., 2020. Dark Data as the New Challenge for Big Data Science, vol. 33. Philosophy & Technology, pp. 93–115.

Shehabi, A., Smith, S.J., Masanet, E., Koomey, J., 2018. Data center growth in the United States: decoupling the demand for services from electricity use. Environ. Res. 13 (12), 124030.

Shehabi, A., et al., 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, USA.

Siddik, M.A.B., Shehabi, A., Marston, L., 2021. The environmental footprint of data centers in the United States. Environ. Res. Lett. 16 (6), 064017.

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286.

Statistics, N., 2021. UK Energy in Brief. Department for Business, Energy & and Industrial Strategy, London.

The Climate Group, 2018. SMART 2020: Enabling the Low Carbon Economy in the Information Age. Global eSustainability Initiative, Brussels, Belgium (GeSI).

Veritas, T., 2015. The Databerg See what Other Do Not Identify the Value, Risk and Cost of Your Data. Veritas Databerg Research, California.

Veritas, T., 2020. The Benefits of Reducing Dark Data. Veritas Databerg Research, California.

Vonderau, A., 2019. Storing Data Infrastructuring the Air: Thermocultures of the Cloud, vol. 18. The Nature of Data Centers, pp. 1–12.

Vries, A., Stoll, C., 2021. Bitcoin's growing e-waste problem. Resour. Conserv. Recycl. 175, 105901.

Zhang, Q., Yang, S., 2021. Evaluating the sustainability of big data centers using the analytic network process and fuzzy TOPSIS. Environ. Sci. Pollut. Control Ser. 28, 17913–17927.

Zhu, L., et al., 2019. Big data analytics in intelligent transportation systems: a survey. IEEE Trans. Intell. Transport. Syst. 20 (1), 383–398.