VIEW POINT

THE ROLE OF AI IN SUSTAINABLE DATA MANAGEMENT

Abstract

The exponential growth of digital data presents a significant environmental crisis, burdening the planet with energy consumption, resource depletion, and electronic waste. This document details the lifecycle footprint of data, highlighting the environmental costs associated with each stage from creation to disposal, particularly within the life science industry. Traditional data management's focus on performance and cost has exacerbated these issues. However, Artificial Intelligence (AI) offers transformative solutions. By optimizing data capture, storage, processing, transmission, and disposal, AI can substantially reduce environmental impact. Case studies within this document illustrate both the challenges and the potential of AI-driven strategies to achieve sustainable data management practices.

Infosys®
Navigate your next

Understanding Data's Environmental Impact: Lifecycle Footprint

The environmental impact of data within the life science industry is significant and pervasive, spanning its entire data lifecycle from creation to disposal. This includes considerable energy consumption, depletion of resources for specialized hardware, and the generation of electronic waste (e-waste).

Data Creation & Capture

- Generating vast life science datasets (genomics, medical imaging, clinical trial data) requires energy-intensive specialized equipment, such as high-throughput sequencers used in genomics, contributing to significant energy consumption.
- Manufacturing these specialized equipment resources, adding to the overall environmental burden even before data is processed.

Data Storage

- Storing massive, often long-term life science datasets (patient records, genomic libraries, images) requires energy-hungry data centers with significant power, cooling, and water demands.
- The constant need for high-capacity storage drives resource depletion during manufacturing and generates significant e-waste when hardware is replaced.

Data Processing

- Analyzing large, complex life science data (genomic analysis, drug discovery simulations, diagnostic AI) demands substantial computational power (often HPC), leading to high energy use and associated carbon emissions.
- **Training a single complex AI model, increasingly used in life sciences, can have a carbon footprint comparable to a car's lifetime emissions.**

Data Transmission

- Transferring large life science files (e.g., genomic datasets, high-resolution images) between institutions or to the cloud consumes considerable network energy.
- The underlying internet and network infrastructure supporting this data transfer contribute significantly to global electricity consumption (estimated at 3-5%).

Data Disposal

- Inefficient data disposal practices lead to prolonged storage of massive datasets (e.g., genomic libraries), driving up energy consumption and increasing the demand for storage resources. Inefficient data management practices that result in unnecessary data retention contribute significantly to the overall environmental footprint of data storage due to increased energy demands.

Impact and Limitations of Traditional Data Management

Traditional data management practices, with their linear approach, performance/cost prioritization, and lack of holistic data lifecycle thinking, contribute significantly to environmental challenges like increased energy consumption, carbon emissions, water depletion, and e-waste generation. These limitations are particularly impactful in the life science industry, which generates and manages vast amounts of complex and sensitive data.

Category	Issue	Description	Key Details/Consequence
Quantifying Environmental Impact of LS Data Management	Emissions from Life Science Data Management.	The contribution of life science data generation, storage, processing, and transmission to greenhouse gas (GHG) emissions..	Life science's energy-intensive computing for processing and storing data such as genomic sequences (e.g., from high-throughput sequencing), drug discovery simulations, medical images (e.g., MRI, CT scans), and large datasets contributes significantly
	Carbon Intensity of Data	The carbon footprint associated with each unit of data.	**Metric:** Grams to tens of grams of CO2e per gigabyte. **Life Science Impact:** Large datasets from sequencing, medical imaging, and patient records increase this intensity.
	Data Center Water Consumption	The large amounts of water used by data centers for cooling.	**Large data centers:** Consume millions of gallons daily - **Hyperscale centers:** Consume 10+ million gallons daily. **Life Science Impact:** Data centers storing genomic data, research data, and patient information require significant cooling.
	E-waste Generation from Data Management Infrastructure.	The increasing volume of electronic waste generated by the infrastructure used for life science data management.	**Projection:** Global e-waste to exceed 82 million metric tons by 2030, with a significant portion stemming from the Life science Industry.
Limitations of Traditional Data Management	Unsustainable Linear Model	Traditional data management's "take-make-dispose" approach.	**Outcome:** Resource depletion and increased e-waste. **Life Science Impact:** Frequent upgrades of research equipment and IT infrastructure in life sciences exacerbate e-waste.
	Prioritization of Performance/Cost in selecting computing resources Computing resources (servers, cloud instances, etc.).	Emphasis on performance and cost-efficiency over environmental sustainability.	**Consequence:** Ignores energy efficiency and environmental concerns. **Life Science Impact:** The need for high-speed computing in life sciences often leads to energy-intensive solutions.
	Lack of Holistic Lifecycle Thinking	Fragmented data management practices that disregard the full environmental impact.	**Result:** Obscures the true environmental cost of data. **Life Science Impact:** Life science organizations may overlook the environmental impact of data throughout its lifecycle in research and development.
	Limited Automation and Optimization	Inefficient manual practices and insufficient use of intelligent, dynamic resource optimization.	**Implication:** Hinders sustainability efforts. **Life Science Impact:** Manual data management in life sciences can lead to inefficiencies and higher energy consumption.
	Insufficient Governance and Measurement	Inadequate impact tracking and reporting mechanisms.	**Effect:** Limits accountability and impedes progress in achieving sustainability - **Principle:** Unmeasured impact remains unmanaged and unimproved. **Life Science Impact:** Lack of standardized metrics for environmental impact in life science data management.

AI-Powered Solutions for Sustainable Data Management

AI offers a transformative suite of solutions to address the environmental challenges posed by the life science data lifecycle. By leveraging intelligent automation, predictive capabilities, and optimization algorithms, AI can significantly reduce energy consumption, optimize resource utilization, and minimize waste across various stages of managing complex data, from genomic sequences and clinical trial data to drug discovery information.

Smarter Data Ingestion: Harvesting Only What We Need

AI optimizes life science data capture (e.g., genomic sequences, microscopy images) to reduce energy and resource strain. It employs **edge computing** for initial processing, **machine learning** to recognize valuable patterns, and **predictive analytics** to forecast data needs.

Leaner Data Transfer: Minimizing Network Energy Consumption

AI reduces energy consumption from data transmission in life science crucial for collaborations involving large-scale data sharing between research institutions or pharmaceutical companies by enabling edge processing and federated learning, which minimize data transfer. This is achieved using **distributed computing, machine learning** for optimized routing and compression, and **edge computing** platforms.

Intelligent Storage: Right Data, Right Place, Right Time

AI optimizes life science data storage by automating tiering based on access patterns (e.g., frequently accessed research data vs. archived clinical trial data) and using deduplication and compression to minimize storage needs and energy consumption. This is achieved using **data classification algorithms, machine learning** for access pattern prediction, and **AI-optimized data compression**

Beyond E-waste: AI for Reclaiming the Data Assets of Life Science

AI improves data management practices for end-of-life lab equipment in the life science industry by facilitating intelligent data archiving and potential anonymization for secondary use in research or training. This involves **machine learning** for data categorization and tagging, **natural language processing** for extracting relevant information, and secure **data anonymization techniques** to maximize data value while adhering to privacy regulations.

Autonomous Data Centers: Self-Optimizing for Efficiency

AI minimizes data center energy consumption critical for life science organizations relying on high-performance computing for tasks like drug discovery and genomic analysis by autonomously monitoring and optimizing operations such as cooling, workload management, and resource allocation. This is achieved using **machine learning** for anomaly detection, **AI-driven control systems, real-time data analytics, and automation.**

AI-Powered ESG: Quantifying and Projecting Sustainability

AI streamlines ESG (Environmental, Social, and Governance) reporting for life science companies by automating data collection, analyzing sustainability trends, and predicting environmental impacts to inform sustainability strategies. This is achieved using **data analytics, machine learning** for trend analysis, and **natural language processing** for report generation.

Case Study: Infosys Implements AI-Driven Sustainable Data Management for Global Pharmaceutical Company

Overview:

A global pharmaceutical company sought to enhance its data management practices with a focus on environmental sustainability. The company faced challenges in managing its vast data landscape while minimizing its environmental impact. To address these issues, the pharmaceutical company partnered with Infosys to implement AI-driven solutions that promote sustainable data management practices.

Challenges:

The primary challenges faced by the pharmaceutical company included:

1. High energy consumption and environmental footprint due to extensive data operations.
2. Inefficient data management processes leading to increased resource usage and waste.
3. Difficulty in maintaining data quality and compliance with environmental regulations.
4. Fragmented data systems that hindered efficient data handling and sustainability efforts.

Solution:

Infosys proposed a comprehensive AI-driven solution to optimize the pharmaceutical company's data management practices while minimizing environmental impact. The key components of the solution were:

1. AI-Powered Data Integration:

Infosys implemented AI algorithms to integrate data from various sources, ensuring seamless data flow and reducing redundancy. This integration facilitated real-time data access and improved data consistency.

2. Automated Data Governance:

The solution utilized AI and Machine Learning (ML) models to automate data governance tasks, such as data classification, tagging, and cleansing. This automation reduced manual efforts and enhanced data quality while minimizing resource usage.

3. Predictive Analytics for Resource Optimization:

Infosys deployed predictive analytics models to monitor and optimize resource consumption, including energy usage. These models provided actionable insights to reduce the carbon footprint of data operations.

4. Sustainable Data Center Management:

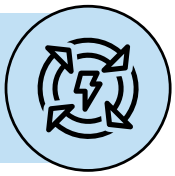
The solution included AI-driven strategies to optimize data center operations, such as dynamic power management and cooling systems. This ensured efficient energy use and reduced environmental impact.

Results:

The implementation of the AI-driven environmentally sustainable data management solution by Infosys resulted in significant improvements for the pharmaceutical company:

1. Reduced Carbon Footprint:

The AI-powered resource optimization and sustainable data center management practices led to a substantial **reduction in energy consumption and carbon emissions.**



2. Enhanced Data Quality:

The automated data governance processes ensured high data quality and consistency across the organization, **supporting efficient and sustainable data management.**

3. Cost Savings:

Optimization of resource usage and reduced energy consumption led to significant cost savings, lowering operational expenses.

4. Improved Compliance:

Predictive analytics models helped the organization maintain **compliance with environmental regulations, reducing the risk of non-compliance.**

Conclusion

The environmental burden of data, fueled by rising energy demands and the mounting challenge of e-waste, demands a fundamental shift away from traditional, unsustainable data management practices, particularly within the data-intensive life science sector. While conventional methods struggle to address the full lifecycle impact of data, Artificial Intelligence offers transformative solutions tailored to the industry's needs.

By optimizing processes across the data lifecycle from streamlining data capture in genomics and medical imaging, to enhancing data storage and transmission, and enabling more sustainable data disposal of lab equipment. AI empowers life science organizations to significantly reduce energy consumption, optimize resource utilization, and minimize waste generation.

Embracing these AI-powered sustainable strategies is not merely an option, but a crucial imperative for responsible environmental stewardship and for fostering a sustainable future for life science innovation.

Reference

1. https://deepmind.google/discover/blog/safety-first-ai-for-autonomous-data-centre-cooling-and-industrial-control/
2. https://engineering.fb.com/2024/09/10/data-center-engineering/simulator-based-reinforcement-learning-for-data-center-cooling-optimization/
3. https://download.schneider-electric.com/files?p_File_Name=SPD_ASTE-6Z5L2Z_EN.pdf&p_enDocType=White+Paper&p_File_Type=application/pdf
4. https://www.komprise.com/
5. https://www.vertiv.com/en-us/about/news-and-insights/articles/white-papers/sensor-networks-the-foundation-for-smarter-data-center-optimization/
6. https://www.ibm.com/topics/predictive-maintenance
7. https://xcelerator.siemens.com/global/en/industries/data-centers/use-cases/use-ai-to-optimize-cooling-infrastructure.html
8. https://www.researchgate.net/publication/344019968_Federated_Learning_for_Edge_AI_Applications

About the Authors

Shanmugam Lakshmanan

Shanmugam Lakshmanan is a Senior Principal in Life Sciences (LS) & Consumer Healthcare (CH), specializing in Responsible AI Enablement, Gen AI Industrialization, Data Management, and Analytics within the LS and CH industries.

Santosh Bhat

Santosh Bhat is a Senior Consultant in Business Consulting, bringing expertise in solution and design, data governance, and quality design/implementation. He is also DCAM Certified from the EDM Council.

Yashvardhan Chamoli

Yashvardhan Chamoli is a Consultant in Business Consulting, specializing in Business Analytics with a focus on Strategy and Marketing. He brings consulting expertise within the US Life Sciences market.

Muskan Bansal

Muskan Bansal is a Consultant in Business Consulting, with experience in Ab Initio development, data warehousing, and analyzing business requirements to support and develop ETL projects.

Shriyani Roy

Shriyani Roy is an Analyst in Business Consulting with experience in pharmaceutical market research and data analysis. She specializes in analyzing market trends and strategic insights.

For more information, contact askus@infosys.com

Infosys®
Navigate your next