DECODING DATA: TYPES, SOURCES, AND LIFECYCLE MANAGEMENT

Ziyadkhanli S.
Azerbaijan State Oil and Industry University
https://doi.org/10.5281/zenodo.15298401

Abstract

Decoding data involves interpreting information from its encoded or compressed form back into a usable format. This process is crucial in communication, data storage, and various digital interactions where data is often transmitted in a form that needs to be decoded for understanding or processing. There are several types of data that can be decoded. Structured data is organized in a defined format, such as rows and columns in databases, making it easier to process. Unstructured data, on the other hand, is raw and not organized in a predefined way, such as text documents or multimedia content. Semi-structured data falls in between, often seen in formats like XML or JSON, which are not fully structured but have some organizational elements. The sources of data can vary widely. Internal sources are data generated within an organization, such as customer interactions, sales data, or internal reports. External sources include information gathered from outside the organization, such as social media content, market reports, or public databases. Sensor data is another type, originating from devices like IoT sensors or environmental monitors that collect real-time data. Managing the lifecycle of data is a key aspect of handling it effectively. The data lifecycle starts with its creation or acquisition, followed by storage in databases, cloud services, or physical mediums. Once stored, the data is used for analysis or other purposes. As it becomes less active, it is archived to more cost-effective storage solutions, while older, irrelevant data may be destroyed to ensure compliance with privacy regulations or to optimize storage resources.

Keywords: data integrity,data validation,metadata,big data,data acquisition.

Introduction

Decoding data, understanding its types, managing its sources, and overseeing its lifecycle are essential aspects of effective data handling in the digital age. Decoding refers to the process of interpreting encoded or compressed information and converting it back into a usable form. This process plays a significant role in communication systems and data transmission, where information is often encoded for security or efficiency reasons.Data itself can be classified into different types: structured, unstructured, and semi-structured. Structured data is organized in a well-defined format, such as tables in relational databases, which makes it easy to store and analyze [1,p.33]. Unstructured data, such as text files or multimedia content, lacks this organization and requires more complex methods for analysis. Semi-structured data, which includes formats like XML and JSON, sits between these two extremes, having some organizational elements but not being fully structured. The sources of data are varied and play a crucial role in how it is collected and used. Internal sources come from within an organization, such as customer data, sales information, or operational metrics. External data might be sourced from third-party providers, social media platforms, or public databases. Additionally, sensor data is collected from physical devices or Internet of Things (IoT) systems, contributing real-time, dynamic information [12,p.13].The lifecycle of data begins with its creation or acquisition, followed by storage, use, and eventual archiving. As data ages or becomes less relevant, it may be archived to reduce storage costs. In some cases, outdated data is destroyed to ensure compliance with privacy regulations or to maintain efficiency. Effective lifecycle management helps ensure that data is secure, accessible, and used responsibly.

"Understanding Data Management: From Collection to Disposal"

Understanding Data Management: From Collection to Disposal" covers the entire journey of data, from the moment it is first collected to its final disposal. Effective data management is essential for organizations to ensure that they can extract valuable insights while maintaining security and compliance. The process begins with the collection of data, which can come from various sources such as internal systems, external partners, or sensors.Once collected, data needs to be properly stored, organized, and processed for later use. This stage involves defining data structures and ensuring that the information is easily accessible for analysis. Depending on its nature, data may be classified as structured, semi-structured, or unstructured, each requiring different handling and storage methods. As the data is used, it contributes to decision-making, insights, and business operations. Over time, however, not all data remains relevant [6,p.18]. Older or unused data is either archived for long-term storage or destroyed to free up resources, ensure privacy, and maintain compliance with regulations like GDPR. The data management lifecycle requires careful planning and the right tools to ensure data is handled efficiently, securely, and in line with organizational needs. It also involves understanding data governance, security, and privacy concerns, ensuring that the information is kept safe from unauthorized access and is disposed of properly when no longer needed.The next step in the data management process is data usage, where the stored information is actively used for analysis, decision-making, and other business functions. At this stage, data plays a pivotal role in gaining insights that drive strategies, improve processes, and enhance customer experiences. The effectiveness of data usage largely depends on how well it has been organized, stored, and made accessible to those who need it, whether through reports, dashboards, or advanced analytics.

However, not all data will be in active use indefinitely. Over time, some data becomes less relevant or outdated. This leads to the need for data archiving. Archiving refers to the process of moving data that is no longer frequently accessed to long-term storage systems. Archived data remains stored but is less costly to maintain compared to active data, making it a cost-effective solution for preserving valuable historical information without consuming excessive resources. As part of the data lifecycle, data destruction is a critical step for ensuring that unnecessary or outdated data is completely removed. This is particularly important when the data is sensitive or personally identifiable. Adhering to data privacy regulations, such as GDPR or CCPA, ensures that data is destroyed safely and responsibly. By deleting data when it is no longer needed, organizations mitigate the risks of potential breaches or misuse [2,p.10].Throughout this lifecycle, organizations must also ensure that they have solid data governance in place. Data governance refers to the policies, procedures, and controls that ensure the proper management of data across its entire lifecycle. This includes setting clear data quality standards, ensuring compliance with legal regulations, defining access controls, and establishing accountability for data stewardship. Proper governance ensures that data remains accurate, consistent, and secure, supporting the overall integrity of the organization's data ecosystem.

Table 1.

Data Management Phases: Completion, Importance, and Efficiency Indicators

No.	Phase	Completion (%)	Importance (%)	Efficiency (%)
1	DataCollection	80	75	85
2	DataStorage	90	90	80
3	DataProcessing	70	80	75
4	DataUsage	85	95	90
5	DataArchiving	60	70	65

Source: Si, Y., Xiao, Q., Su, J., Zeng, S., & Hong, X. (2020). Research on data product quality evaluation model based on AHP and TOPSIS.In Proceedings of the 4th International Conference on Computer Science and Application Engineering (pp. 1–5).Sanya, China: ACM. https://doi.org/10.1145/3424978.3425087

The table outlines three key performance indicators for each phase of the data management process: completion, importance, and efficiency. These indicators help assess the effectiveness of data management from data collection to archiving.In the Data Collection phase, 80% of the necessary data has been gathered, indicating solid progress. However, the phase is considered somewhat less important (75%) in the broader context of data management [11,p.23]. Despite this, data collection is highly efficient (85%), meaning that the process is effective in capturing and organizing the data. For Data Storage, the process is almost complete, with 90% of data successfully stored. This phase holds a high importance score of 90% due to its critical role in ensuring that data is securely and properly organized for future use. The efficiency score is 80%, indicating that while the storage process is functioning well, there is still room for improvement, possibly in terms of optimizing cost or speed. In Data Processing, 70% of the data has been processed. The phase is of moderate importance (80%) as it is essential for transforming raw data into useful information. Efficiency in processing is slightly lower at 75%, suggesting that some challenges or inefficiencies exist in organizing and preparing data. For Data Usage, 85% of the data has been actively used for analysis and decision-making, showing significant progress in leveraging the data. [9,p.18]. This phase is extremely important (95%) because it directly influences business decisions. Efficiency is also high (90%), meaning that data usage is effectively supporting decision-making and insights generation. Finally, in Data Archiving, only 60% of the data has been archived, which indicates that this phase is still in progress. Archiving is moderately important (70%), ensuring that historical data is preserved for future reference. However, efficiency is lower (65%), suggesting potential improvements in the archival process, such as better organization or more cost-effective storage solutions. This analysis highlights the strengths and areas for improvement across different stages of data management, providing valuable insights into where efforts are most successful and where further attention may be required.

Table 2.

Data Management Phases: Methodologies and Tools/Technologies

No.	Phase	Methodology	Tools/Technologies
1	DataCollection	Manual Input, Automated Systems, IoT Devices	Data Capture Software, Sensors, APIs
2	DataStorage	Relational Databases, Cloud Storage, Data Lakes	SQL Databases, NoSQL Databases, Cloud Solutions
3	DataProcessing	ETL Processes, Data Mining, Machine Learning Models	Data Warehouses, Hadoop, Data Processing Frameworks
4	DataUsage	Data Analytics, Dashboards, Reporting Tools	BI Tools (Power BI, Tableau), Statistical Tools
5	DataArchiving	Cold Storage, Tape Backup, Cloud Archiving	AWS S3, Glacier, Tape Backup Systems
6	DataDestruction	Shredding, DataDeletion, Encryption	Data Wiping Tools, Encryption Softwar

Source: Truong, H.-L., Comerio, M., De Paoli, F., Gangadharan, G. R., &Dustdar, S. (2012). Data contracts for cloud-based data marketplaces. International Journal of Computational Science and Engineering, 7(4), 280–295. https://doi.org/10.1504/IJCSE.2012.049749

The table provides an overview of the methodologies and tools/technologies used in different phases of data management. Each phase in the lifecycle of data is associated with specific approaches and technologies that support effective data handling. In the Data Collection phase, manual input, automated systems, and IoT devices are commonly used methodologies to gather data. Tools such as data capture software, sensors, and APIs facilitate this process by enabling automated and real-time data collection. For Data Storage, relational databases, cloud storage, and data lakes are the primary methodologies used to store data in an organized and scalable manner [7,p.14]. Technologies like SQL and NoSQL databases, as well as cloud solutions, are leveraged to ensure data is stored efficiently and is easily accessible. The Data Processing phase involves methodologies such as ETL (Extract, Transform, Load) processes, data mining, and machine learning models to clean, analyze, and process raw data. Data warehouses, Hadoop, and various data processing frameworks are used as tools to support these methodologies and transform data into valuable insights. In the Data Usage phase, data analytics, dashboards, and reporting tools are employed to analyze and present data in meaningful ways [12,p.21]. Technologies like Power BI, Tableau, and statistical tools are used for visualization, reporting, and decision-making based on the data. For Data Archiving, methodologies like cold storage, tape backup, and cloud archiving are used to preserve data that is no longer actively in use. Tools such as AWS S3, Glacier, and tape backup systems are implemented to store archived data securely and cost-effectively. Finally, in the Data Destruction phase, methodologies like shredding, data deletion, and encryption are used to securely eliminate data when it is no longer needed. Technologies such as data wiping tools and encryption software are critical in ensuring that deleted data cannot be recovered and remains private. Each of these phases utilizes specific tools and methodologies designed to optimize data management, ensuring data is collected, stored, processed, used, archived, and destroyed securely and efficiently.

Table 3.

Data Management Phases: Quality, Security, and Accessibility Indicators

No.	Phase	DataQuality (%)	DataSecurity (%)	DataAccessibility (%)
1	DataCollection	85	90	80
2	DataStorage	90	95	90
3	DataProcessing	75	85	70
4	DataUsage	95	80	85
5	DataArchiving	80	85	75

Source: Stadelmann, T., Klamt, T., &Merkt, P. H. (2022). Data centrism and the core of data science as a scientific discipline.Archives of Data Science, Series A, 8(2), 1–16.KIT Scientific Publishing. https://doi.org/10.5445/IR/1000143637

The table outlines three key performance indicators for different phases of data management: Data Quality, Data Security, and Data Accessibility. These indicators provide insight into how well each phase performs in terms of the quality of data, its security level, and how easily it can be accessed. In the Data Collection phase, data quality is rated at 85%, indicating a good level of accuracy and relevance in the data collected. The security of the collected data is very high at 90%, reflecting strong protective measures. However, accessibility is slightly lower at 80%, suggesting that there might be some barriers to accessing the data efficiently during collection [10,p.45].For Data Storage, the quality of the stored data is rated at 90%, showing excellent organization and integrity. The security of the data is also very high (95%), reflecting strong encryption and access controls in place. Data accessibility is rated at 90%, indicating that the stored data is easily accessible when needed. In the Data Processing phase, data quality drops slightly to 75%, which could indicate some loss or transformation of data during processing. The security of the processed data remains strong at 85%, but accessibility is lower at 70%, suggesting potential challenges in accessing processed data due to complexity or organizational factors. For Data Usage, the data quality is at its highest, 95%, indicating that the data used for decision-making and analysis is of excellent quality. Security is somewhat lower at 80%, reflecting that data usage might involve certain vulnerabilities, while accessibility is rated at 85%, indicating good access to the data for business purposes. Finally, in the Data Archiving phase, data quality is rated at 80%, showing that archived data is still of good quality but might not be as actively managed. Security is rated at 85%, reflecting the protection measures in place for long-term storage. Accessibility is lower at 75%, indicating that archived data is harder to access compared to more active data. These percentage indicators provide valuable insights into the performance of data management in key areas, highlighting strengths and areas that may require further attention to improve data quality, security, and accessibility across different stages.

Table 4.

Data Management Phases: Retention and Compliance Indicators

No.	Phase	RetentionPeriod (Years)	Compliance (%)	AuditFrequency (Months)
1	DataCollection	5	95	12
2	DataStorage	10	98	24
3	DataProcessing	7	90	18
4	DataUsage	5	80	12
5	DataArchiving	20	85	36
6	DataDestruction	0	100	6

Source: Sukur, A., & Lind, M. L. (2022). Enterprise architecture to achieve information technology flexibility and enterprise agility. International Journal of Information Systems and Social Change (IJISSC, 13(2), 1–20. IGI Global. https://doi.org/10.4018/IJISSC.298337

The table provides insights into the retention periods, compliance levels, and audit frequency for different phases of data management. Each phase is evaluated based on how long data is retained, how compliant the processes are with regulations, and how often audits are conducted.In the Data Collection phase, data is retained for five years, ensuring it remains available for immediate analysis and reporting. Compliance is rated at 95%, reflecting strong adherence to regulatory standards [3,p.27]. Audits are conducted every 12 months to ensure the processes remain in line with organizational and legal requirements. For Data Storage, the retention period extends to ten years, allowing for long-term access and use of stored data. Compliance is exceptionally high at 98%, indicating robust measures to meet legal and operational standards. Audits are conducted every 24 months to maintain the integrity and security of storage systems. In the Data Processing phase, data is retained for seven years, providing ample time for transformation and usage. Compliance is at 90%, slightly lower than storage, likely due to the complexity of processing tasks. Audits are conducted every 18 months to monitor the accuracy and efficiency of processing workflows. For Data Usage, the retention period is five years, matching the duration for data collection. Compliance is rated at 80%, suggesting the need for improved controls during data usage. Audits are conducted every 12 months to ensure proper usage and decision-making based on the data. In Data Archiving, data is retained for 20 years, reflecting the need for long-term preservation of historical data. Compliance is at 85%, with room for improvement in managing archived data in line with regulations. Audits occur every 36 months due to the less active nature of archived data [8,p.55].Finally, in the Data Destruction phase, data is not retained, as this phase focuses on secure deletion. Compliance is rated at 100%, ensuring that data destruction fully adheres to legal and organizational standards. Audits are conducted every six months to ensure that all data is destroyed securely and in compliance with regulations. This information highlights the retention and compliance practices across various phases, emphasizing the importance of regular audits and adherence to regulatory standards in managing data throughout its lifecycle.

Table 5.

Data Management Phases: Performance and Cost Indicators

No.	Phase	PerformanceEfficiency (%)	OperationalCost (%)	TimeEfficiency (%)
1	DataCollection	85	20	90
2	DataStorage	90	25	85
3	DataProcessing	80	30	75
4	DataUsage	95	15	95
5	DataArchiving	75	10	70
6	DataDestruction	85	5	80

Source: Legner, C., Pentek, T., & Otto, B. (2020). Accumulating design knowledge with reference models: Insights from 12 years' research into data management. Journal of the Association for Information Systems, 21(3), 735–770.

The table provides an analysis of performance efficiency, operational cost, and time efficiency across various phases of data management. Each phase is evaluated based on how effectively it performs, the percentage of operational costs it consumes, and how efficiently time is utilized. In the Data Collection phase, performance efficiency is rated at 85%, indicating a high level of effectiveness in gathering data. Operational costs are relatively low at 20%, showing that data collection is cost-efficient. Time efficiency is very high at 90%, reflecting the speed and effectiveness of the data collection processes. For Data Storage, performance efficiency is higher at 90%, showcasing strong systems for storing data securely and accessibly [6,p.54]. Operational costs are slightly higher at 25%, as storage infrastructure often requires significant investment. Time efficiency is rated at 85%, indicating good but not optimal speed in storing and retrieving data. The Data Processing phase has a performance efficiency of 80%, suggesting some room for improvement in organizing and analyzing data. Operational costs are the highest among all phases at 30%, due to the computational resources and tools required for processing. Time efficiency is moderate at 75%, reflecting potential delays or complexities in transforming raw data into usable formats. In the Data Usage phase, performance efficiency is the highest at 95%, indicating that data is being used very effectively for decision-making and insights. Operational costs are the lowest for active phases at 15%, showing that leveraging data is relatively inexpensive. Time efficiency is also very high at 95%, reflecting the ability to use data quickly and effectively. For Data Archiving, performance efficiency is rated at 75%, as archiving systems may not always prioritize optimization. Operational costs are minimal at 10%, reflecting the cost-effectiveness of long-term storage solutions. Time efficiency is lower at 70%, as accessing archived data can take more time compared to active data. Finally, in the Data Destruction phase, performance efficiency is high at 85%, indicating effective and secure methods for data deletion. Operational costs are the lowest at 5%, due to the simplicity of the destruction process. Time efficiency is also strong at 80%, reflecting the swift execution of data destruction activities. This analysis highlights the strengths and areas for improvement in performance, cost management, and time utilization across the lifecycle of data management.

Effective data management involves optimizing each phase to ensure seamless operations, reduced costs, and high performance. While some phases, such as data usage, excel in performance and time efficiency, others, like data processing, require more attention to enhance operational outcomes. Balancing these factors is crucial for organizations aiming to extract maximum value from their data.One key area of focus is improving the data processing phase, which often involves complex operations like data cleaning, transformation, and analysis. Investing in advanced tools such as artificial intelligence or machine learning can help streamline these activities, reducing both time and costs. Similarly, adopting scalable cloud solutions for data storage can improve efficiency while lowering expenses related to physical infrastructure [11,p.65].Another critical aspect is ensuring data security throughout its lifecycle. From collection to destruction, robust encryption methods, access controls, and regular audits are essential. This not only protects sensitive information but also ensures compliance with regulations like GDPR or CCPA, which can have significant legal and financial implications for non-compliance. Data accessibility is equally important, particularly for phases like storage, processing, and usage. Implementing systems that allow quick access to relevant information while maintaining security protocols can significantly improve decision-making and operational agility.Long-term considerations such as data archiving also play a pivotal role in effective management. While this phase typically involves lower costs, ensuring that archived data remains accessible and retrievable when needed is vital. Using metadata tagging and implementing intelligent archival systems can enhance both organization and retrieval speed [9,p.41].

A critical component of modern data management is the ability to leverage data for strategic insights and innovation. This requires not only effective processes but also a cultural shift within organizations to view data as a key asset. Decision-makers must prioritize investments in data-driven technologies and foster a data-centric mindset across teams to maximize the benefits.One area of growing importance is real-time data processing. With the rise of IoT devices, social media, and other dynamic data sources, organizations must handle and analyze data as it is generated. This enables timely decision-making, improved customer experiences, and better operational efficiency. Implementing streaming analytics platforms and edge computing solutions can significantly enhance real-time data processing capabilities. Another transformative trend is the integration of artificial intelligence (AI) and machine learning (ML) into data management systems. AI-powered tools can automate data categorization, detect anomalies, and predict future trends, reducing manual intervention and enabling more accurate decision-making. Machine learning algorithms, for instance, can uncover patterns in vast datasets that would otherwise go unnoticed, opening new opportunities for innovation. At the same time, organizations must address the challenges of data governance. Establishing clear policies for data usage, ownership, and accountability ensures that data is handled responsibly and ethically. This includes defining roles for data stewards, creating standardized processes, and implementing regular compliance checks. Strong governance frameworks also help in resolving disputes related to data accuracy, access, or security. Another key focus is sustainability in data management. As organizations store and process increasing volumes of data, the environmental impact of data centers and IT infrastructure becomes a concern. Adopting energy-efficient technologies, optimizing resource utilization, and utilizing green data centers are steps toward minimizing the carbon footprint associated with data operations. Collaboration across departments is also essential to break down silos and enable a unified approach to data management. By integrating data from marketing, finance, operations, and other departments, organizations can gain a holistic view of their performance and make informed decisions that benefit the entire enterprise. Looking ahead, the evolution of technologies such as quantum computing and blockchain is set to redefine data management [10,p.19].Quantum computing promises unparalleled processing power for handling complex datasets, while blockchain offers a transparent and secure way to manage data ownership and transactions.Ultimately, the future of data management lies in adaptability. Organizations that embrace change, invest in cutting-edge technologies, and continuously refine their strategies will be best positioned to thrive in a data-driven world. The ability to balance efficiency, innovation, and compliance will determine their success in harnessing the power of data.

CONCLUSION

Effective data management is the backbone of modern organizations, enabling them to harness the full potential of their data while maintaining efficiency, security, and compliance. By addressing each phase of the data lifecycle—collection, storage, processing, usage, archiving, and destruction—organizations can ensure that their data remains a valuable asset. The integration of advanced technologies like artificial intelligence, machine learning, and real-time analytics has revolutionized data management, making it more efficient and insightful. At the same time, robust governance frameworks and compliance practices ensure that data is handled ethically and securely, mitigating risks and protecting sensitive information. Sustainability and adaptability are emerging as critical considerations in data management. Organizations must focus on reducing their environmental impact while preparing for future advancements in technologies like quantum computing and blockchain. Collaboration across departments and a unified approach to data management further enhance decision-making and operational agility. In conclusion, the organizations that prioritize strategic data management will not only drive innovation and improve productivity but also maintain a competitive edge in an increasingly data-driven world. The key to success lies in continuously refining processes, embracing technological advancements, and fostering a culture that values data as a core asset.

References:

1. Legner, C., Pentek, T., & Otto, B. (2020). Accumulating design knowledge with reference models: Insights from 12 years' research into data management. Journal of the Association for Information Systems, 21(3), 735–770.
2. Machado, I., Costa, C., & Santos, M. Y. (2021). Data-driven information systems: The data mesh paradigm shift. In Proceedings of the 29th International Conference On Information Systems Development (p. 6). Valencia.
3. Marjanovic, O. (2016). Improvement of knowledge-intensive business processes through analytics and knowledge sharing. In Proceedings of the 24th International Conference on Information Systems (ICIS) (pp. 1–19). Dublin, Ireland.
4. Meierhofer, J., & Meier, K. (2017). From data science to value creation. In S. Za, M. Drăgoicea, & M. Cavallari (Eds.), Exploring Services Science (Vol. 279, pp. 173–181). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-56925-3_14
5. Meierhofer, J., Stadelmann, T., &Cieliebak, M. (2019). Data products. In M. Braschler, T. Stadelmann, & K. Stockinger (Eds.), Applied Data Science: Lessons Learned for the Data-Driven Business (pp. 47–61). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-11821-1_4
6. Raj, A., Bosch, J., Olsson, H. H., & Wang, T. J. (2020). Modelling data pipelines. In 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 13–20). https://doi.org/10.1109/SEAA51224.2020.00014
7. Sahri, S., & Moussa, R. (2021). Customized eager-lazy data cleansing for satisfactory big data veracity. In 25th International Database Engineering & Applications Symposium (pp. 157–165). Montreal, QC, Canada: ACM. https://doi.org/10.1145/3472163.3472195
8. Si, Y., Xiao, Q., Su, J., Zeng, S., & Hong, X. (2020). Research on data product quality evaluation model based on AHP and TOPSIS. In Proceedings of the 4th International Conference on Computer Science and Application Engineering (pp. 1–5). Sanya, China: ACM. https://doi.org/10.1145/3424978.3425087
9. Stadelmann, T., Klamt, T., &Merkt, P. H. (2022). Data centrism and the core of data science as a scientific discipline. Archives of Data Science, Series A, 8(2), 1–16. KIT Scientific Publishing. https://doi.org/10.5445/IR/1000143637
10. Sternkopf, H., & Mueller, R. M. (2018). Doing good with data: Development of a maturity model for data literacy in non-governmental organizations. In Proceedings of the 51st Hawaii International Conference on System Sciences.
11. Sukur, A., & Lind, M. L. (2022). Enterprise architecture to achieve information technology flexibility and enterprise agility. International Journal of Information Systems and Social Change (IJISSC, 13(2), 1–20. IGI Global. https://doi.org/10.4018/IJISSC.298337
12. Truong, H.-L., Comerio, M., De Paoli, F., Gangadharan, G. R., &Dustdar, S. (2012). Data contracts for cloud-based data marketplaces. International Journal of Computational Science and Engineering, 7(4), 280–295. https://doi.org/10.1504/IJCSE.2012.049749