https://doi.org/10.58245/ipsi.tir.2502.08

Current Aspects of the Data Ecosystems

Borkovcova, Monika; Kvet, Marek; and Dokoupil, Tomas

Abstract: ***Current trends and future developments in the field of data storage, processing, analysis, output, and visualization include various possible directions and future developments of these systems. This paper discusses the current state of data ecosystems and possible future developments based on an analysis of publications and studies from the last time. Commonly known trends point to approaches that enable more efficient data management, faster data analysis and, of course, the integration of different types of data coming from different sources. Additionally, the study evaluates the ETL process performance, identifying key metrics, optimizations, and bottlenecks in data processing workflows. The benefits and development of generative models are also changing the way we work with data, with implications for improved visualizations and general process automation. Thus, in the area of analysis, it can be assumed there will be an even greater emphasis on interactive and intuitive reports that make it as easy as possible to understand complex data structures. In terms of the use of system resources, the use of Cloud services is the basis for the comprehensive use of scalability and flexibility at lower infrastructure costs. Another area that the paper discusses is the current growing demand for real-time data analysis due to the increasing need for quick decisions based on data evaluation.***

Index Terms: ***cloud computing, data analysis, data management, data visualization, real-time processing***

1. INTRODUCTION

Data ecosystems are complex environments in which different actors, technologies and data interact to create value. These ecosystems play a key role in the modern digital economy, where data serves as a strategic resource for innovation, decision-making and process improvement. The following text summarizes key insights from five selected articles, which were chosen as introductory because they address different aspects of data ecosystems from different perspectives, ranging from their governance and sustainability to their application in specific industries.

A data ecosystem reflecting environmental data analysis involves a variety of interconnected systems, tools, and processes that allow for the collection, processing, analysis, and dissemination of environmental data. The goal of such an ecosystem is to provide insights that help in environmental decision-making, resource management, and policy development. Key components of a data ecosystem in environmental data analysis are data collection (sensors, satellite imagery, manual observations), data storage, and data processing (preprocessing, real-time processing, geospatial temporal analysis [1, 2]. The data analysis itself points to statistical analysis, machine learning, and modeling. Data visualization techniques cover dashboards, maps, graphs and charts. The Erasmus+ project EverGreen (Including Everyone in Green Data Analysis) deals with climate change monitoring, biodiversity, and sustainability [3]. The data used for the evaluation study reflects airplane monitoring to handle flight efficiency by limiting impacts on the environment.

One of the key topics is the governance and optimization of policies in data ecosystems, as shown in the study "Towards a policy tuning method for data ecosystems". The authors focus on how the right rules and policies can impact the performance of data ecosystems and align them with business objectives. They propose methods for optimizing these policies to enable more effective use of data and support the achievement of organizations' strategic goals. This approach highlights the importance of data flow management and its impact on overall system efficiency. [4]

Feedback loops in open data ecosystems form other important aspects, as described in the paper "Feedback Loops in Open Data Ecosystems". Open data, which is often provided by public institutions, plays a crucial role in increasing transparency and fostering innovation. The study analyses how feedback between data providers and data users affects the quality and usability of this data. This approach promotes

Manuscript received February 15, 2025. This scientific paper was supported by the international Erasmus+ project 2022-1-SK01-KA220-HED-000089149 “Including EVERyone in GREEN Data Analysis (EVERGREEN)” funded by the European Union.

collaboration between different actors and shows how open data can contribute to the development of innovative solutions. [5]

A specific application of data ecosystems is their use in industrial production, as shown in the article "Conceptualising Data Ecosystems for Industrial Food Production". The authors focus on the food industry, where data can be used to create new business models, improve production efficiency and promote sustainability. The study highlights the importance of integrating data from different sources and collaboration between actors within the ecosystem, enabling better decision-making and process optimization. [6]

Sustainability and governance of data ecosystems is another key topic, as shown in the article "Sustainability and Governance of Data Ecosystems". The authors discuss technical and organisational approaches that can contribute to the creation of sustainable data ecosystems. The study highlights that good data governance and technical decisions are essential for the long-term sustainability of these systems. Sustainability is understood here not only in terms of environment but also in terms of economic and social impact. [7]

The Fifth article, "Shaping the Future of Data Ecosystem Research-What Is Still Missing?", offers a broader view of the current state of data ecosystem research. The authors conduct a bibliometric analysis that identifies key trends and research gaps. The study shows that although research on data ecosystems is rapidly evolving, there are still areas that require deeper investigation, such as issues of standardization, interoperability, and ethics in data use. [8]

The above studies show that data ecosystems are dynamic and complex environments that require careful management and collaboration between different actors. Key success factors are the right policy settings, feedback loops, sustainability and the application of data in specific sectors. At the same time, however, research in this area still has room for further development, especially on issues of standardization and ethical considerations. Modernization of Data Ecosystems thus presents not only a challenge but also an opportunity for organizations that want to harness the potential of data to achieve their strategic goals.

2. *Related Work*

In today's digital environment, the exponential growth of data creation has made big data analysis a key tool for innovation and strategic decision-making. [1, 2, 3, 9]

Big data has traditionally been characterized by three key attributes, known as the 3Vs: Volume, Velocity, and Variety. [2, 10]

These dimensions reflect the challenges of managing large datasets from multiple sources processed at high speed. For example, social media platforms generate huge volumes of user-generated content (Volume), real-time transaction records require fast processing (Velocity), and data from IoT devices or multimedia files emphasize the variety of data formats (Variety).

In some frameworks, the characteristics of Big Data extend beyond the 3Vs to include Veracity and Value, forming the 5Vs model. [10]

Veracity refers to the reliability and quality of data and addresses the problems posed by noisy, incomplete, or inconsistent information. Ensuring veracity is crucial for accurate analysis and decision-making, especially in areas such as healthcare or financial services. Value, on the other hand, highlights the potential insights and business benefits of analyzing large volumes of data, making it essential to turn raw data into actionable information. [11]

Big data comes from a variety of sources, including social networks, Internet of Things (IoT) devices, online transactions, and scientific research. Social networks play an important role in the generation of big data through the interactions of users [12], who share opinions, ideas, and perspectives on various topics daily. The modern world allows individuals, companies, and government institutions not only to connect but also to create content in different formats, such as text, images, or videos. This vast flow of information forms what is called "Big Social Data", which can be used to gain insights through machine learning and social data analytics. With these data, organizations can improve their data-driven decision-making process by using techniques such as text analytics, which is one of the most common methods of social media analytics.

However, effective use of big data depends on robust data management frameworks such as ETL and the emerging Data Lakehouse architecture to ensure scalability, reliability, and ease of access. [13]

ETL is a fundamental process for integrating data in modern systems and preparing them for analysis. This process consists of three main

phases: extract, transform and load. The goal is to integrate data from disparate sources, improve its quality, and ensure that it is properly stored in data warehouses or lakes. In the first phase, extraction, data is extracted from various sources, including databases, CSV or XML files. In the transformation phase, data are cleaned, normalized, and enriched with additional information, including removing errors, filling in missing values, and standardizing formats. The final step is to upload the data to the target repository, where it is optimized for fast query and analysis operations. [14]

Automating ETL processes brings significant savings in time and resources. Modern tools enable real-time data integration, automated cleaning, and advanced data quality monitoring. [15]

Current trends also include the use of machine learning and artificial intelligence for anomaly detection and predictive data cleaning, reducing human intervention, and increasing the accuracy of data processes. [16]

Data warehouses are the foundation of modern data architecture and serve as centralized repositories where data from various sources are collected, organized, and stored for analysis and reporting purposes. [17]

Traditionally, data warehouses support the extraction, transformation, and loading (ETL) process in which data are extracted from various operational systems, transformed to ensure quality and consistency, and loaded into the warehouse. Their schemas, typically optimized for query performance, use approaches such as stars or snowflakes, making them ideal for business intelligence (BI) tools and analytical applications. As data ecosystems have grown in complexity, data warehouses have evolved. [18]

Two important approaches to data warehouse architecture have been defined by Ralph Kimball and Bill Inmon, each offering a different methodology. Kimball's approach focuses on bottom-up design, where the emphasis is on creating data warehouses tailored to specific business processes. These data marts, structured in a star schema, are later integrated into a complete data warehouse. This design emphasizes simplicity, speed of implementation, and direct alignment with business needs. In contrast, Inmon advocates a top-down approach that envisions the data warehouse as a centralized enterprise-wide repository designed with a normalized structure (3NF). This architecture promotes scalability and consistency and ensures that all enterprise functions derive insights from a single source of truth.

Data lakes are a central element of modern big data architectures and represent a natural evolution of data warehousing systems that address big data requirements characterized by their volume, variety, and velocity. Alfredo [16] to data lakes as the evolution of data warehouses. Unlike traditional data warehouses, data lakes are designed to handle raw, unstructured, and partially structured data in addition to structured data sets. This flexibility allows them to serve as repositories that satisfy a variety of analytical needs. [19]

Notable is the emergence of hybrid Data Lakehouse (LH) architecture, which is sometimes presented as the "next step". These combine the structured and managed nature of data warehouses with the flexibility and scalability of data lakes. This hybrid model addresses the limitations of traditional data warehouses, particularly in handling unstructured and semi-structured data types generated from different sources. Powered by advanced analytics and machine learning, LH exemplifies the shift towards unified, universal data platforms that bridge the gap between the management of structured and unstructured data. [13, 20]

Another part of this topic is spatial data which refers to information about places and their characteristics. Spatial data has a wide range of applications in many fields, such as urban planning, environmental protection, and logistics. Due to its richness and the possibility to use it for decision-making, working with such data has attracted a lot of interest in the past. Initial approaches included the extension of relational databases with spatial data or the creation of spatial data warehouses, which enabled more efficient querying and management of these data. [21, 22]

However, with the advent of Big Data, traditional methods began to reach their limits and new concepts such as data lakes and later the LakeHouse architecture began to emerge. The LakeHouse architecture combines the flexibility of data lakes with the robustness and reliability of data warehouses, enabling efficient handling of large volumes of spatial data. Moreover, thanks to the cloud, it is possible to use almost unlimited computing power, which significantly speeds up analytical queries over spatial databases and opens new possibilities, which dramatically reduces response time when working with queries. These innovations are key to coping with the increasing demands for spatial data analysis and processing in modern

applications. [22, 23]

3. ETL Process – Performance Evaluation Study

The ETL (Extract, Transform, Load) process is a critical component in data warehousing and data analysis. It facilitates the integration of data from multiple sources, ensuring its quality and usability for business intelligence and decision-making. This study evaluates the performance of an ETL process based on key performance metrics, bottlenecks, and optimization strategies. The efficiency of the process is delimited by the execution time, data accuracy of the results, as well as resource utilization. Besides, it is necessary to identify limitations and component bottlenecks. Thanks to that, optimization techniques and recommendations can be provided.

Fig. 1 shows the flow of the ETL process consisting of three phases loading the data into the data warehouse. There are many factors influencing the performance of the ETL process:

- Volume and velocity – the size of the dataset strongly impacts the performance of the extraction and consecutive processing. It is necessary to identify relevant data to be handled.
- Type of processing - real-time data or streaming commonly needs different strategies compared to batch processing.
- Data quality – additional validation can cause significant overhead.
- Hardware and available resources - CPU, memory, disk I/O, and network bandwidth are critical for the data movement speed.
- Data model and database layer optimization – indexes, partitioning, SQL query tuning, partitioning, and data distribution.

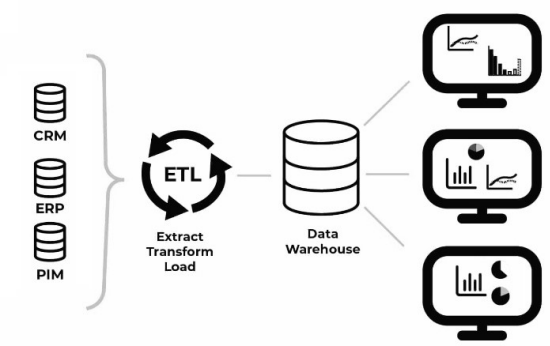

Figure 1. ETL process

Fig. 1 shows the basic components of the ETL process. The data source can be a relational database (like Oracle, MySQL, Postgres), or cloud storage (like Oracle Object Storage, Amazon S3, Google Cloud storage). Among that, data can be provided by various interfaces or Rest APIs.

Next, the transformation process includes data cleansing (handling undefined values, duplicate tuples, etc.), data aggregation, normalization, and conversion. A typical example is related to the date and time processing and various input formats [23].

The loading process can take a data warehouse, data lake, or other architectures for the analytical-oriented storage.

For the performance evaluation study, the following characteristics were used:

- Extraction time
- Transformation time
- Loading time
- CPU utilization
- Memory utilization
- I/O performance
- Data accuracy
- Throughput

The data originated from aircraft monitoring over Europe in the period 2017-2020. It took the whole process from starting, taxi, departure, and flight itself up to landing and parking. To optimize the flight, it is necessary to identify optimal and real route.

To handle that, three scenarios were used:

- Scenario 1 (S1) - Conventional performance test - running the ETL process under normal conditions to establish benchmarks.
- Scenario 2 (S2) - High volume test - processing large datasets to evaluate scalability.
- Scenario 3 (S3) - Testing concurrency - running multiple ETL processes simultaneously to test system concurrency.

Results are present in Tab. 1.

Table 1. Evaluated model description

	S1	S2	S3
Extraction time (min)	12	27	17
Transformation time (min)	18	44	25
Load time (min)	10	35	14
CPU utilization	65	90	78

Memory utilization	6	15	12
Accuracy	99.7	99.3	99.6

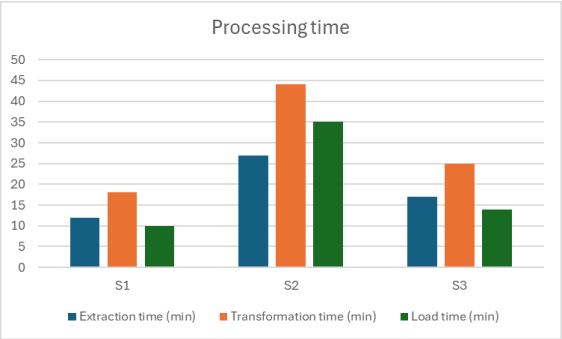

Figure 2. Processing time results

The results show, that the main aspect of the ETL process impacting the performance is just the transformation process of the data from various sources. With an increasing number of data, the problem can be even more and more demanding.
To highlight that, several bottlenecks can be identified – slow extraction speed, high memory utilization or database write speed. When dealing with the last aspect approaching data loading, it is necessary to find a proper balance between the data retrieval performance and writing operations. It is related to the physical infrastructure, data layer, and model, but mostly indexes.

During the evaluation study, several optimization strategies were identified:

- Parallel processing – multi-thread environment accessing data for the extraction and transformation concurrently.
- Loading data incrementally – efficient identification of the new data tuples to reduce handled data amount during the transformation batch.
- Database schema optimization using advanced techniques for indexing, partitioning, and data distribution with emphasis on the physical infrastructure.
- Data compression - apply efficient storage formats to reduce I/O overhead.
- Load balancing – using multiple processing nodes and distributing workload uniformly across them. An associated feature is safety and reliability; once a node goes down, the processing is migrated to another available one.

This performance evaluation highlights key areas where the ETL process can be optimized for efficiency and scalability. By implementing the recommended optimizations, you can enhance data processing speed, reduce resource consumption, and improve overall data pipeline reliability.

4. *Attributes Monitored*

From the above publications and studies, the monitored parameters can be defined, where different attributes are monitored in different types of data ecosystems depending on their purpose, structure and involved actors. In general, however, certain key attributes are common across most data ecosystems. These attributes can be divided into several categories:

1. **Data quality**
   - Accuracy: How accurate and reliable the data is.
   - Completeness: Whether the data is complete and contains all necessary information.
   - Consistency: Whether the data are consistent across different sources and systems.
   - Update: How often the data are updated and whether they are up to date?
   - Relevance: Whether the data is relevant to the needs of the users and the purpose of the ecosystem.

2. **Data flows and uses**
   - Accessibility: Who has access to the data and how easily the data can be retrieved?
   - Interoperability: The ability of data systems to work together and share data.
   - Usage tracking: How the data is used and by whom.
   - Data sharing: Mechanisms and rules for sharing data between actors in the ecosystem.

3. **Data security and protection**
   - Data integrity: Ensuring that data has not been tampered with.
   - Privacy: Ensuring that personal data is protected following legislation (e.g. GDPR).
   - Encryption: Use of encryption methods to protect data during transmission and storage.
   - Threat detection: Monitoring cyber threats and vulnerabilities.

4. **Ecosystem performance**
   - Data processing speed: How fast data is processed and analyzed.
   - Availability: Whether systems and data are available in real-time.
   - Reliability: How often system failures or errors occur.

- Scalability: The ability of the system to handle increasing volumes of data and users.

5. **Economic and business attributes**
   - Data Value: The economic value of data to organizations and users.
   - Cost of data management: The costs associated with storing, processing, and managing data.
   - Data Monetization: How are data monetized (e.g., data sales, analytics)?

6. **Sustainability and governance**
   - Sustainability: How the ecosystem contributes to long-term sustainability (e.g. environmental impacts).
   - Rules and policies: Monitoring compliance with ecosystem rules and policies.
   - Actor involvement: Level of cooperation and involvement of different actors in the ecosystem.

7. **Feedback and development**
   - User feedback: How users evaluate the quality and usability of the data.
   - Innovation: The ability of the ecosystem to adapt to new technologies and trends.
   - Process improvement: Monitoring process efficiency and identifying areas for improvement.

These attributes are key to the effective functioning of data ecosystems and their ability to create value. Monitoring these aspects allows organizations to better manage data, and ensure data quality and security while fostering collaboration between the various actors within the ecosystem.

5. *Fulfilment of Monitored Attributes and Their Use in Data Ecosystems*

Data quality is a fundamental pillar of modern data ecosystems, where organizations are faced with challenges around data accuracy, completeness, and consistency. The main technical challenges are inconsistent input data formats, data transmission failures, and conflicts in concurrent data access. Ensuring high data quality is complicated by the increasing volume of data and the diversity of data sources.

Data flows and its use is currently facing challenges in ensuring seamless availability and effective interoperability between systems. The technical challenges mainly include optimizing performance with a high number of concurrent accesses, addressing API incompatibilities, and managing large amounts of logs to track data usage. Effective management of access rights and optimization of the network infrastructure also play a key role.

Data security and protection requires a comprehensive approach including ensuring data integrity, protecting privacy, and implementing advanced encryption mechanisms. The technical challenges relate mainly to the high computing power requirements for encryption, the complexity of managing encryption keys and certificates, and the difficulty of detecting security threats in real-time.

Ecosystem performance is a critical factor that affects processing speed, service availability and overall system reliability. The main technical challenges are optimizing I/O operations, addressing bottlenecks in data processing, ensuring efficient load balancing, and managing distributed systems as they scale.

The economic and business attributes focus on maximizing the value of data and optimizing the cost of managing the data ecosystem. Technical challenges primarily include the difficulty of measuring actual data usage, predicting needed capacity, and implementing effective data monetization models.

Sustainability and management of data ecosystems require a balance between system performance and its environmental impact. The technical challenges lie in optimizing power usage, implementing effective compliance monitoring systems, and ensuring effective collaboration between the various system actors.

Feedback and development are key to the continuous improvement of the data ecosystem. Technical challenges mainly include the complexity of implementing systems for collecting and analyzing feedback, ensuring compatibility when implementing innovations, and effectively testing changes in the production environment.

The following figures show the most frequently observed combinations of properties (attributes) in modern data ecosystems according to their focus. In general terms, attributes can be understood as factors or criteria that contribute to the overall evaluation of a system, project, or solution variant. Attributes typically together form a whole (100%), with each attribute having a defined weight that reflects its relative importance in a given context. These weights can be predetermined based on organizational priorities, project requirements, or expected outcomes. For example, if there is a strong emphasis on environmental sustainability, the attribute

"environmental impact" may have a higher weight than other attributes. Conversely, in other cases where the priority is, for instance, the economic aspect, the relevant attribute will be dominant.

The overall value of the attributes is then used to evaluate the variants, whereby their mutual ratio allows for a more objective and systematic comparison of individual solutions. This approach also supports transparency in decision-making.

Figure 3. Colour differentiation of the attributes

This first variant may be used in situations, where organizations, working with sensitive data, such as in the healthcare or financial sectors or situations where personal data protection (GDPR and other regulatory standards) is paramount. These projects focus on long-term data quality and trustworthiness, regardless of immediate economic returns.

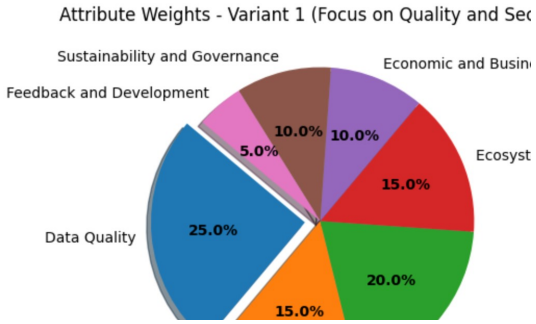

Figure 4. Focus on Quality and Security

The second variant is designed for tech companies where speed to market, performance, and innovation are key. For organizations focusing on market growth, business expansion, and flexibility or for situations where data is not highly sensitive, the main goal is to optimize overall performance.

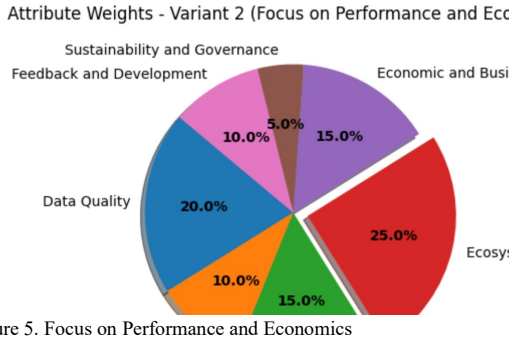

Figure 5. Focus on Performance and Economics

Variant 3 focuses on organizations that emphasize minimizing environmental impacts and long-term system management. This option is ideal for companies wishing to implement practices aimed at sustainability and efficiency. Projects oriented towards ecological and systemic sustainability can benefit from this configuration due to its focus on stability and process transparency. Businesses aiming to gain a competitive advantage by adopting environmentally friendly and long-term sustainable strategies will profit from a robust approach focused on responsible operational management, resource optimization, and building trust among partners and customers. This variant can also help organizations reduce costs associated with inefficiencies or environmental fines, thereby supporting both financial and environmental prosperity.

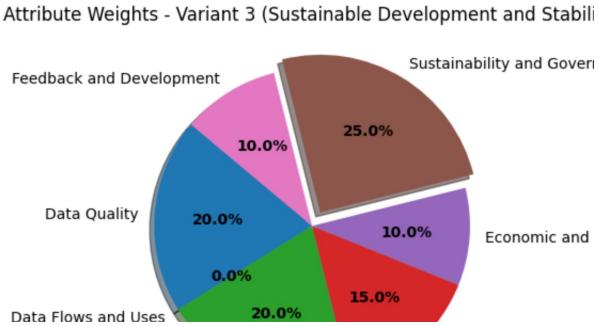

Figure 6. Focus on Sustainable Development and Stability

Variant 4 is suitable for start-ups or companies in a phase of rapid growth, where prioritizing iterative development and innovation is crucial. This configuration is also ideal for projects requiring flexibility, experimentation, and quick adaptation to market demands. Organizations seeking new approaches to system improvement can leverage this variant to support creative solutions and implement adaptive methods in a dynamic environment.

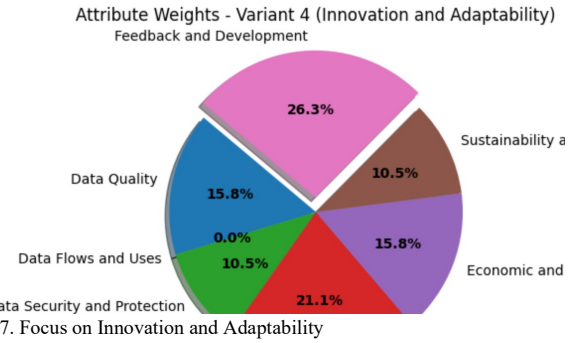

Figure 7. Focus on Innovation and Adaptability

6. *DATA ECOSYSTEMS AND THE NEAR FUTURE*

Modern data ecosystems are becoming increasingly complex and offer a range of opportunities for efficient data collection,

processing, and use. One of the key trends is the growing emphasis on data quality, which includes data accuracy, timeliness and completeness. Organizations need to implement mechanisms for automatic data cleansing and validation to ensure a high level of confidence in all analytical outputs. It is also necessary to monitor progress on data governance, which includes structured processes and rules for managing data across the entire lifecycle, including sharing, storage, and disposal.

Another important aspect to watch is the integration of data from different sources. Increasing volumes of data, coming from both internal systems and IoT sensors or external data marketplaces, are putting pressure on the use of modern data orchestration tools and ETL (Extract, Transform, Load) processes. Standardization of data formats and interoperability between different systems is becoming a necessity, allowing organizations to more effectively connect and leverage data from different platforms.

Predictive and prescriptive modelling capabilities are growing rapidly in analytics and artificial intelligence. Keeping up with innovations in machine learning and neural networks is becoming essential as modern data ecosystems already routinely incorporate AI-based tools. Organizations must prepare for increasing demands on computing power and storage, especially if they want to massively leverage AI for the real-time processing of large volumes of data.

Data security is another key issue that needs to be constantly monitored. With the increasing number of cyber-attacks, organizations are forced to invest in sophisticated technologies such as AI threat detection, multi-factor authentication and zero-trust architectures. Soon, we can expect even stricter compliance requirements for privacy legislation such as GDPR or CCPA, which will force companies to improve their mechanisms for managing personal data.

Cloud technologies also play a significant role. The move to hybrid or fully cloud-based data infrastructure models offers greater flexibility and scalability but also brings new cost and performance management requirements. Organizations will need to invest in tools to optimize cloud expenditure and follow trends in edge computing, which allows data to be processed closer to where it is generated.

Within data visualization and accessibility, there is an increasing emphasis on self-service BI solutions that allow users to easily interpret complex analytics without the need for advanced technical knowledge. Keep an eye on advances in natural language processing (NLP), which enables querying and interacting with data in natural language. This trend will make it easier for even less tech-savvy employees to extract valuable insights from analyses.

With sustainability comes new challenges and opportunities. Data centers are one of the big energy consumers, so organizations are starting to look for greener alternatives, such as using renewable energy and optimizing the energy consumption of computing infrastructure. The issue of data waste, unused or redundant data that puts unnecessary strain on systems, is also coming to the fore.

Soon, we can expect higher demands for automation and autonomous systems. Data ecosystems will need to be able to react to environmental conditions in real-time without the need for manual intervention. An example is the autonomous decision-making of AI-driven systems that will be used in manufacturing, logistics or healthcare.

Another key aspect is the question of ethics when working with data. With the increasing use of AI and massive data analytics, there is a need to define ethical principles for working with data to ensure fair and transparent decision-making. Organizations will need to invest in training employees and putting in place regulatory mechanisms to prevent misuse of data.

Finally, it is important to stress that successful data ecosystems of the future will be built on ambient collaboration and innovation. Multi-organizational collaboration and data sharing will enable a better understanding of markets, customer behavior and global challenges such as climate change or health crises. The key challenge will be not only to manage the technical challenges but also to set rules that promote transparency and trust across the data world.

7. CONCLUSIONS

This paper discusses data ecosystems, their current state and possible future developments. The paper analyzes recent publications and studies and discusses trends in data storage, processing, analysis, output, and visualization. The authors focus on approaches to enable more efficient data management, faster analysis, and integration of different types of data from different sources. The paper also mentions the importance

of generative models, interactive visualizations and cloud services in the context of data ecosystems.

Modern data ecosystems are undergoing a dynamic evolution characterized by increasing complexity and new challenges in data management, processing, and use. Several key findings emerge from the analysis.

First, data quality and data governance are becoming critical success factors. Organizations must implement sophisticated mechanisms for automatic data validation and cleansing, with an emphasis on accuracy, timeliness, and completeness of information.

Second, integrating disparate data sources requires advanced data orchestration tools and ETL processes. The paper specifically evaluates the performance of ETL processes, identifying key bottlenecks such as transformation times and resource consumption, while proposing optimization strategies for improving data workflows. Standardization of data formats and interoperability of systems are essential for effective use of data across platforms. The evaluation study of this paper focuses on the ETL process performance.

Third, the growing importance of artificial intelligence and predictive analytics places increased demands for computing power and storage capacity. Organizations must be prepared to implement advanced AI tools for real-time data processing.

Last but not least, the importance of data security and compliance with regulatory requirements must be emphasized. As the number of cyber threats increases, the need to implement robust security mechanisms and ensure compliance with regulatory requirements increases.

The future of data ecosystems will be characterized by greater emphasis on automation, ethical use of data and cross-industry collaboration. Successful organizations will be those that can effectively combine technological innovation with a responsible approach to data and its use. The future of data ecosystems is also linked to the further increase in data volume, which will require even more sophisticated methods for data processing and analysis. A key aspect will be increasing the speed and efficiency of data processing thanks to advances in machine learning, artificial intelligence, and cloud technologies.

There will also be a focus on interoperability and integration of different data sources, allowing for more comprehensive and accurate data insights. These developments will support both the wider use of the Internet of Things (IoT) and the deployment of real-time data analytics technologies.

Data security and protection will also remain a key challenge. As the amount of sensitive information grows, advanced security methods will need to be developed to prevent data breaches and ensure user and business confidence.

Overall, therefore, data ecosystems will continue to evolve towards greater interconnectivity, automation, and the ability to provide valuable insights for rapid and informed decision-making. This evolution will offer new business opportunities and streamline operations across many industries, but will also require increased attention to the ethical and security aspects of data processing.

ACKNOWLEDGEMENT

This scientific study was financially supported by the international Erasmus+ project: Project number: 2022-1-SK01-KA220-HED-000089149, Project title: Including EVERYone in GREEN Data Analysis (EVERGREEN) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Slovak Academic Association for International Cooperation (SAAIC). Neither the European Union nor SAAIC can be held responsible for them.

REFERENCES

[1] M. Kvet, S. Toth and E. Krsak E., "Concept of temporal data retrieval: Undefined value management", Concurrency Computation Practice and Experience. 2020, https://doi.org/10.1002/cpe.5399

[2] M. Kvet, "Developing Robust Date and Time-Oriented Applications in Oracle Cloud: A comprehensive guide to efficient date and time management in Oracle Cloud", Packt Publishing, 2023, ISBN: 978-1804611869

[3] Erasmus+ project EverGreen, Accessed: 15.1.2025, Available: https://evergreen.uniza.sk/

[4] M. Salnitri, E. Ramalli and B. Pernici, "Towards a policy tuning method for data ecosystems," *2024 IEEE International Conference on Web Services (ICWS)*,

Shenzhen, China, 2024, pp. 55-64, https://doi.org/10.1109/ICWS62655.2024.00024

[5] D. Rudmark and M. Andersson, "Feedback Loops in Open Data Ecosystems," in *IEEE Software*, vol. 39, no. 1, pp. 43-47, Jan.-Feb. 2022, https://doi.org/10.1109/MS.2021.3116874

[6] C. Rix, H. Stein, Q. Chen, J. Frank and W. Maass, "Conceptualizing Data Ecosystems for Industrial Food Production," *2021 IEEE 23rd Conference on Business Informatics (CBI)*, Bolzano, Italy, 2021, pp. 201-210, https://doi.org/10.1109/CBI52690.2021.00031

[7] E. Ramalli and B. Pernici, "Sustainability and Governance of Data Ecosystems," *2023 IEEE International Conference on Web Services (ICWS)*, Chicago, IL, USA, 2023, pp. 740-745, https://doi.org/10.1109/ICWS60048.2023.00100

[8] T. Brée, E. Karger and F. Ahlemann, "Shaping the Future of Data Ecosystem Research—What Is Still Missing?," in IEEE Access, vol. 12, pp. 103162-103175, 2024, https://doi.org/10.1109/ACCESS.2024.3432969

[9] Dhaouadi, A.; Bousselmi, K.; Gammoudi, M.M.; Monnet, S.; Hammoudi, S. "Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons". Data 2022, 7, 113. https://doi.org/10.3390/data7080113

[10] Favaretto M, De Clercq E, Schneble CO, Elger BS (2020) What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. PLOS ONE 15(2): e0228987. https://doi.org/10.1371/journal.pone.0228987

[11] M. Dave and J. Kamal, "Identifying big data dimensions and structure," 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 2017, pp. 163-168, https://doi.org/10.1109/ISPCC.2017.8269669

[12] M. S. Rahman and H. Reza, "A Systematic Review Towards Big Data Analytics in Social Media," in Big Data Mining and Analytics, vol. 5, no. 3, pp. 228-244, September 2022, https://doi.org/10.26599/BDMA.2022.9020009

[13] D. Oreščanin and T. Hlupić, "Data Lakehouse - a Novel Step in Analytics Architecture," 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2021, pp. 1242-1246, https://doi.org/10.23919/MIPRO52101.2021.9597091

[14] L. Zineb and F. Rachid, "ETL Technologies for Big Data: A Comparative Study," 2023 IEEE International Conference on Advances in Data-Driven Analytics And Intelligent Systems (ADACIS), Marrakesh, Morocco, 2023, pp. 1-6, https://doi.org/10.1109/ADACIS59737.2023.10424341

[15] Gorka Zarate, María Jose Lopez Osa, Ana I. Torre-Bastida, Urtza Iturraspe, Jordi Arjona, Benjamín Navarro, and Antoni Gimeno. 2024. Evolution of Extract-Transform-Load (ETL) processes towards data product pipelines. In Proceedings of the 4th Eclipse Security, AI, Architecture and Modelling Conference on Data Space (eSAAM '24). Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/3685651.3686662

[16] Kartick Chandra Mondal, Neepa Biswas, and Swati Saha. 2020. Role of Machine Learning in ETL Automation. In Proceedings of the 21st International Conference on Distributed Computing and Networking (ICDCN '20). Association for Computing Machinery, New York, NY, USA, Article 57, 1–6. https://doi.org/10.1145/3369740.3372778

[17] M. Kvet and J. Papan, "Enhancing Analytical Select Statements Using Reference Aliases," in IEEE Access, vol. 12, pp. 27311-27330, 2024, https://doi.org/10.1109/ACCESS.2024.3366455

[18] H. Akid, G. Frey, M. B. Ayed and N. Lachiche, "Performance of NoSQL Graph Implementations of Star vs. Snowflake Schemas," in IEEE Access, vol. 10, pp. 48603-48614, 2022, https://doi.org/10.1109/ACCESS.2022.3171256

[19] A. Cuzzocrea, "Big Data Lakes: Models, Frameworks, and Techniques," *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju Island, Korea (South), 2021, pp. 1-4, https://doi.org/10.1109/BigComp51126.2021.00010

[20] Ahmed A. Harby and Farhana Zulkernine. 2025. Data Lakehouse: A survey and experimental study. Inf. Syst. 127, C (Jan 2025). https://doi.org/10.1016/j.is.2024.102460

[21] Kovacs-Györi, A.; Ristea, A.; Havas, C.; Mehaffy, M.; Hochmair, H.H.; Resch, B.; Juhasz, L.; Lehner, A.; Ramasubramanian, L.; Blaschke, T. Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning. *ISPRS Int. J. Geo-Inf.* 2020, *9*, 752. https://doi.org/10.3390/ijgi9120752

[22] S. Ait Errami, K. Ait El Kadi, H. Hajji, and H. Badir, "Spatial big data architecture: From Data Warehouses and Data Lakes to the LakeHouse," Journal of Parallel and Distributed Computing, vol. 176, pp. 70–79, Jun. 2023, https://doi.org/10.1016/j.jpdc.2023.02.007

[23] Rodrigo Costa Mateus, Thiago Luís Siqueira, Valéria Cesário Times, Ricardo Rodrigues Ciferri, and Cristina Dutra Aguiar Ciferri. 2016. Spatial data warehouses and spatial OLAP come towards the cloud: design and performance. Distrib. Parallel Databases 34, 3 (September 2016), 425–461. https://doi.org/10.1007/s10619-015-7176-z

Monika Borkovcova is an assistant professor at the University of Pardubice, Faculty of Electrical Engineering, and Informatics, Studentská 95, 532 10 Pardubice, Czech Republic. Her teaching and research activities are related to the database systems and data analytics. Email: monika.borkovcova@upce.cz, ORCID 0000-0001-7820-2241.

Marek Kvet is an associate professor at the University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 01026 Žilina, Slovakia. Modeling and optimization are the key areas of his research by focusing on the ambulance system optimization. Email: marek.kvet@fri.uniza.sk, ORCID 0000-0001-5851-1530.

Tomas Dokoupil is a lecturer and Ph.D. student at the University of Pardubice, Faculty of Electrical Engineering and Informatics, Studentská 95, 532 10 Pardubice, Czech Republic. His research focuses on data layer optimization and data analytics within various architectures.