Co-funded by the European Union

INCREASING THE VISIBILITY AND KNOWLEDGE BASE ON AIR QUALITY WITH BIG DATA AND ITS IMPACT ON THE POPULATION HEALTH

Erasmus+ KA220 Cooperation Partnerships

A Actual Look at Artificial Intelligence Tools Used in Big Data Processing

Project number	2023-1-SK01-KA220-HED-000158470
Document type	Research report

Author(s) / Institution	**Vladimir Vietoris** Slovak University of Agriculture in Nitra, SPVTS Slovakia

Version	1.0

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 European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

1 INTRODUCTION

The convergence of artificial intelligence (AI) and big data has revolutionized how organizations process, analyze, and derive value from massive datasets. This chapter explores the key AI tools and frameworks that have become instrumental in big data processing across various industries. As data volumes continue to grow exponentially, traditional data processing methods have proven insufficient to handle the velocity, variety, and volume challenges presented by modern datasets. AI tools have emerged as critical solutions that enable organizations to extract meaningful insights efficiently and on a scale.

2 BACKGROUND

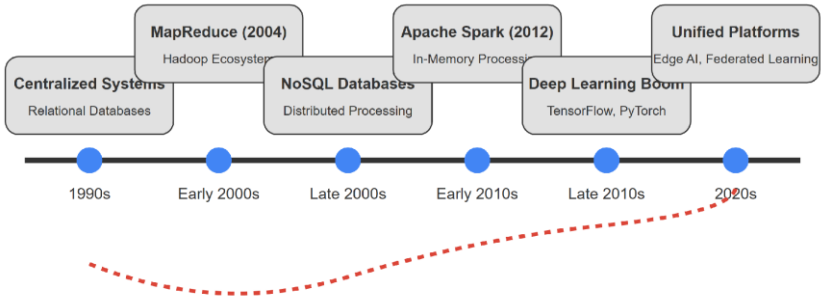
2.1 Historical Overview of Big Data Processing

The evolution of big data processing represents one of the most significant technological transformations of the past three decades. In the early 1990s, data processing was primarily conducted on centralized mainframes and relational database management systems with limited scalability (Gray, 1992). The internet boom of the late 1990s generated unprecedented volumes of data that traditional systems could not handle, leading to what Laney (2001) would later characterize as the "3Vs" challenge: volume, velocity, and variety. This data explosion catalyzed innovations like Google's MapReduce paper in 2004 (Dean & Ghemawat, 2004) and the subsequent development of Hadoop by Yahoo! in 2006 (White, 2015). These technologies revolutionized data processing by introducing distributed computing on commodity hardware clusters, allowing organizations to process petabytes of data cost-effectively, as documented by Shvachko et al. (2010) in their analysis of the Hadoop Distributed File System. The early 2010s saw the rise of in-memory processing frameworks like Apache Spark (Zaharia et al., 2012), which dramatically improved processing speeds for iterative algorithms essential to machine learning applications. Simultaneously, specialized NoSQL databases emerged to address specific data types and workloads (Cattell, 2011). Cloud computing's maturation in the mid-2010s further transformed the landscape, offering elastic, managed big data services that reduced operational complexity (Hashem et al., 2015). Real-time stream processing frameworks like Apache Kafka (Kreps et al., 2011) and Apache Flink (Carbone et al., 2015) enabled organizations to process data as it was generated rather than in batches. The late 2010s brought significant advancement in AI integration with big data systems, including automated machine learning platforms (Feurer et al., 2015) and specialized hardware accelerators like GPUs and TPUs that dramatically reduced the time required to train complex models on massive datasets (Jouppi et al., 2017). Most recently, the trend has shifted toward unified data platforms that integrate storage, processing, analytics, and AI capabilities within coherent architectures (Armbrust et al., 2021), while edge computing pushes data processing capabilities closer to data sources, reducing latency and bandwidth requirements (Shi et al., 2016). This historical progression reflects a continuous response to increasing data volumes, velocity, and variety, driven by both technological innovation and evolving business

requirements for deriving timely insights from complex data ecosystems (Davenport & Bean, 2018).

The relationship between AI and big data processing has evolved significantly over the past decade. Early applications focused primarily on basic statistical analysis and rule-based systems with limited scalability (Chen et al., 2014). Today's AI tools for big data processing incorporate sophisticated machine learning algorithms, neural networks, and natural language processing capabilities that can handle petabytes of structured and unstructured data. This evolution has been driven by advancements in computational power, distributed computing frameworks, and algorithmic innovations that have made AI-powered data processing both feasible and cost-effective for organizations of various sizes (LeCun et al., 2015).

Figure 1: Evolution of Big Data Processing



Source: Author´s own processing

2.2 Machine Learning Frameworks for Big Data

Machine learning frameworks have become the backbone of AI-driven big data processing. TensorFlow, developed by Google, has emerged as one of the most widely adopted frameworks for implementing machine learning models at scale. Its distributed computing capabilities allow it to process massive datasets across clusters of machines, making it particularly suitable for big data applications (Abadi et al., 2016). Similarly, Apache Spark's MLlib provides a comprehensive suite of machine learning algorithms optimized for distributed processing environments. The integration of Spark with Hadoop ecosystems has made it a popular choice for organizations seeking to implement machine learning capabilities within their existing big data infrastructure (Meng et al., 2016).

PyTorch, developed by Facebook's AI Research lab, has gained significant traction for its dynamic computational graph and intuitive Python interface. While initially focused on research applications, PyTorch has increasingly been deployed in production environments

for big data processing tasks, particularly those involving deep learning and natural language processing (Paszke et al., 2019). These frameworks share a common goal: enabling the efficient implementation of complex machine learning models capable of extracting patterns and insights from vast datasets that would be impossible to analyze manually.

Table 1: Popular Machine Learning Frameworks for Big Data

Framework	Description	Key Strengths
Apache Spark ML	Integrated ML library within Spark ecosystem	Scalability, speed
TensorFlow	Deep learning and neural networks	Versatility, performance
PyTorch	Dynamic computational graph, deep learning models	Ease of use, flexibility
Apache Mahout	Distributed linear algebra and collaborative filtering	Scalability, parallel computing
Scikit-learn	Robust algorithms for predictive analytics	User-friendly, extensive APIs

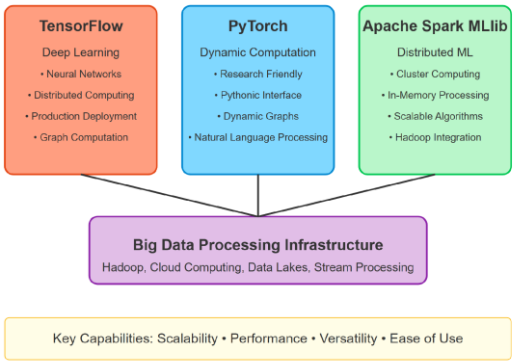
Source: Author´s own processing

Deep Learning in Big Data Analytics

Deep learning has transformed the landscape of big data analytics by enabling the processing and analysis of complex, unstructured data types such as images, audio, and text at unprecedented scales. Convolutional Neural Networks (CNNs) have revolutionized image processing within big data pipelines, enabling automated classification and feature extraction from visual datasets containing millions of images (Krizhevsky et al., 2012). This capability has found applications in diverse fields ranging from medical imaging analysis to satellite imagery processing for environmental monitoring.

Recurrent Neural Networks (RNNs) and their variants, particularly Long Short-Term Memory (LSTM) networks, have become essential tools for processing sequential data in big data environments. These architectures have proven especially valuable for analyzing time-series data, natural language, and other sequential information that constitutes a significant portion of today's big data landscape (Hochreiter & Schmidhuber, 1997). The ability of deep learning models to automatically extract hierarchical features from raw data has eliminated much of the manual feature engineering previously required in big data processing pipelines, significantly accelerating the time-to-insight for organizations dealing with complex datasets.

Figure 2: Machine Learning Frameworks for Big Data



Source: Author´s own processing

Deep Learning in Big Data: Analyzing Air Pollution and Outdoor Air Quality

Air pollution represents one of the most significant environmental health risks globally, with the World Health Organization (WHO) attributing approximately 7 million premature deaths annually to exposure to polluted air (WHO, 2021). Monitoring, analyzing, and predicting air quality has become increasingly critical for public health policy, urban planning, and environmental management. The convergence of big data analytics and deep learning has revolutionized our capacity to understand air pollution dynamics at unprecedented temporal and spatial resolutions.

This analysis explores how deep learning techniques, when applied to big data from diverse sources, are transforming air quality monitoring and analysis. Modern air quality management generates massive datasets from heterogeneous sources, including ground-based sensor networks, satellite imagery, meteorological stations, traffic monitoring systems, and citizen science initiatives. Deep learning approaches excel at extracting meaningful patterns from these complex, high-dimensional datasets, enabling more accurate predictions, source attribution, and intervention planning than traditional statistical methods (Wen et al., 2022).

Other case studies across the sectors

We try to present several real-world case studies that illustrate how artificial intelligence (AI) tools and big data technologies are implemented across various sectors. The goal is to demonstrate the practical benefits and challenges that organizations face when integrating these technologies into their operational processes.

Healthcare and Medical Diagnostics

In healthcare, AI and big data play a crucial role in image analysis, disease prediction, and personalized medicine. Convolutional Neural Networks (CNNs) are commonly used for the automated analysis of medical images, enabling faster and more accurate diagnoses of conditions such as tumors and degenerative diseases (Krizhevsky et al., 2012). Case studies have shown that the implementation of these technologies in hospitals can significantly enhance diagnostic quality and, consequently, patient treatment. Furthermore, the analysis of large datasets collected from various sources, including wearable devices, contributes to better monitoring of population health, supporting preventive measures (WHO, 2021).

Financial Sector and Fraud Detection

In finance, AI tools are key to detecting fraud and managing risk. Machine learning is used to identify unusual transaction patterns in real time, allowing for rapid responses to potentially fraudulent activities. Case studies from the banking sector indicate that the use of distributed streaming technologies – such as Apache Kafka combined with real-time data processing (Carbone et al., 2015; Kreps et al., 2011) – significantly reduces the risk of financial losses and enhances the security of financial transactions.

Industry and Smart Manufacturing

Efficiency is a key factor in the industrial sector. AI tools help optimize manufacturing processes through predictive maintenance, quality monitoring, and real-time production line management. The use of machine learning and deep learning frameworks such as TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019) enables more accurate predictions of machine failures and minimizes downtime. Case studies from smart factories demonstrate that integrating these technologies not only boosts productivity but also leads to significant cost savings.

Transportation and Urban Mobility

In the transportation sector and urban infrastructure, AI and big data tools enable the creation of intelligent systems for traffic management, congestion prediction, and public transportation optimization. The use of edge computing – which processes data directly at its source (Shi et al., 2016) – in conjunction with real-time analytics improves system response times and contributes to more efficient urban network management. Case studies from smart

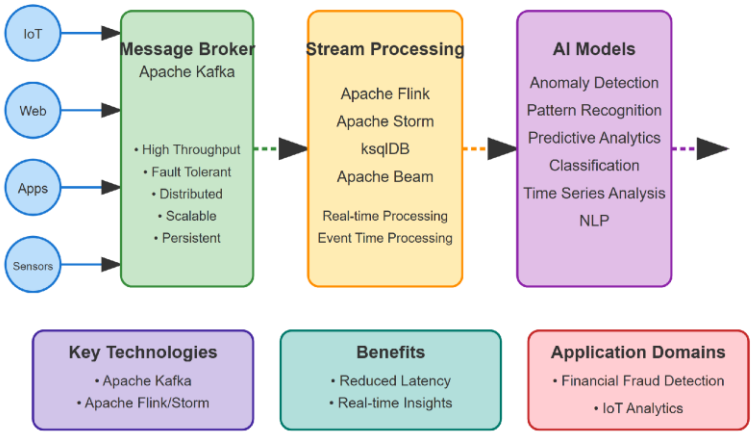
city projects have shown that such systems can significantly reduce greenhouse gas emissions and enhance the quality of urban life.

Real-time Analytics and Stream Processing with AI

The need to derive insights from data in real-time has driven the development of AI tools specifically designed for stream processing. Apache Kafka, combined with stream processing frameworks like Apache Flink and Apache Storm, has created powerful platforms for implementing real-time AI analytics on continuously generated data streams (Kreps et al., 2011). These technologies enable organizations to apply machine learning models to data as they are created, rather than processing it in batches, significantly reducing the latency between data generation and insight extraction.

The integration of complex event processing capabilities with AI has further enhanced real-time analytics in big data environments. Technologies like Confluent's ksqlDB and Apache Beam provide declarative interfaces for expressing complex analytics logic that can be applied to streaming data using AI models (Akidau et al., 2015). These tools have found applications in diverse domains including financial fraud detection, network security monitoring, and IoT analytics, where the ability to quickly identify patterns and anomalies in high-velocity data streams provides significant competitive advantages.

Figure 3: Real-time Analytics and Stream Processing with AI



Source: Author´s own processing

Edge Computing and AI for Distributed Data Processing

The proliferation of IoT devices and mobile technologies has created new challenges and opportunities for big data processing. Edge computing, combined with optimized AI models designed to run on resource-constrained devices, has emerged as a powerful approach for processing data closer to its source (Shi et al., 2016). Frameworks like TensorFlow Lite and PyTorch Mobile enable the deployment of compact machine learning models on edge devices, reducing the need to transmit raw data to centralized processing facilities.

This distributed approach to AI-powered data processing offers several advantages in big data contexts, including reduced latency, decreased bandwidth requirements, and enhanced privacy through local data processing (Li et al., 2018). The development of federated learning techniques, where models are trained across multiple edge devices without centralizing the data, represents a particularly promising direction for applying AI to distributed big data environments while respecting privacy constraints. These approaches are increasingly being adopted in applications ranging from smart cities to healthcare monitoring systems where data is generated across widely distributed networks of sensors and devices.

AI-Powered Data Preparation and Integration

Data preparation and integration represent critical challenges in big data environments, often consuming up to 80% of data scientists' time (Lohr, 2014). AI tools have emerged to automate and streamline these processes. Machine learning-based data quality tools can automatically identify and correct inconsistencies, duplications, and errors in large datasets at scales that would be impossible to address manually. These tools employ techniques ranging from simple rule-based systems to sophisticated anomaly detection algorithms to maintain data integrity across diverse data sources.

AI-powered data integration platforms like Tamr and Trifacta leverage machine learning to automate the mapping and transformation of data from disparate sources into unified schemas (Stonebraker et al., 2013). These platforms can learn from user feedback to continuously improve their integration logic, making them particularly valuable in dynamic big data environments where new data sources are frequently added. By reducing the manual effort required for data preparation and integration, these AI tools enable organizations to focus more resources on extracting insights from their data rather than preparing it for analysis.

Ethical Considerations and Governance in AI-Driven Big Data

The application of AI to big data processing raises important ethical considerations that organizations must address. Issues of bias, transparency, and fairness become amplified when AI models are applied to massive datasets that may contain inherent biases or unrepresentative samples (O'Neil, 2016). AI governance frameworks have emerged to guide the responsible deployment of AI in big data environments, establishing processes for model

validation, monitoring, and explanation that help ensure algorithmic decisions are fair and accountable.

Tools for algorithmic transparency and explainability, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), have become essential components of responsible AI deployments in big data contexts (Ribeiro et al., 2016). These tools help data scientists and stakeholders understand how AI models are making decisions when applied to complex datasets, facilitating the identification and mitigation of potential biases. As regulatory frameworks like the EU's General Data Protection Regulation (GDPR) increasingly emphasize the "right to explanation" for algorithmic decisions, these explainability tools will continue to grow in importance for organizations leveraging AI for big data processing.

3 CONCLUSION

The integration of AI tools with big data processing capabilities has transformed how organizations extract value from their data assets. From distributed machine learning frameworks to automated data preparation tools, AI technologies have addressed many of the challenges associated with the volume, velocity, and variety of modern datasets. As these technologies continue to evolve, several trends are likely to shape the future landscape of AI in big data processing.

The development of more efficient algorithms and hardware architectures specifically optimized for AI workloads will further enhance the scalability of AI applications in big data environments. Quantum computing, though still in its early stages, holds particular promise for accelerating certain classes of AI algorithms when applied to massive datasets (Biamonte et al., 2017). Similarly, the continued refinement of automated machine learning platforms will further democratize access to AI capabilities, enabling a broader range of organizations to leverage these tools for big data processing.

As AI and big data technologies continue to converge, the boundary between data infrastructure and AI systems will increasingly blur. The emergence of unified platforms that seamlessly integrate data storage, processing, and AI capabilities will create new possibilities for end-to-end data analytics pipelines that can adapt and learn as they process information. This evolution will require organizations to develop new competencies and governance structures that span traditional boundaries between data management and artificial intelligence disciplines.

The convergence of AI and big data processing continues to redefine organizational capabilities, driven by innovations in generative AI, federated learning, and edge computing. Looking ahead, advancements in self-supervised learning (Chen et al., 2021) will reduce reliance on labeled datasets, enabling AI systems to extract insights from unstructured data at unprecedented scales. Meanwhile, quantum machine learning (Biamonte et al., 2017)

holds potential to revolutionize optimization tasks, such as feature selection in high-dimensional datasets, though practical applications remain nascent. Organizations must also navigate evolving ethical landscapes. The proliferation of generative AI necessitates frameworks to audit synthetic data for bias, while federated learning demands robust governance to ensure equitable participation in collaborative models. As unified platforms blur the lines between data infrastructure and AI, interdisciplinary competencies—spanning data ethics, quantum computing, and decentralized systems—will become critical for sustaining innovation. In the next decade, AI-driven big data processing will likely transition from reactive analytics to proactive intelligence, where systems not only interpret data but also anticipate trends and autonomously optimize workflows.

By embracing emerging technologies while addressing their societal implications, organizations can unlock transformative value in an increasingly data-driven world.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation,* pp. 265-283.

Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., ... & Whittle, S. (2015). The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12), 1792-1803.

Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hover, H., Drabas, A., Griffith, A., Ionescu, D., Łuszczak, A., Michalska, M., Nakandala, S., Nykodym, D., Oguntebi, T., Park, S., Shah, K., Wollman, M., Xin, R., Zaharia, M., Zang, R., Zhang, Y., Delimitrou, C. & Kozyrakis, C. (2021). Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores. *Proceedings of the VLDB Endowment*, 13(12), 3411-3424.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549(7671), 195-202.

Brown, T. et al. (2020). *Language Models are Few-Shot Learners*. NeurIPS.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S. & Tzoumas, K. (2015). Apache Flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 36(4), 28-38.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), 12-27.

Chen, M., Mao, S. & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, 19(2), 171-209.

Chen, T. et al. (2021). Emerging Paradigms in Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Davenport, T.H. & Bean, R. (2018). Big data and AI executive survey 2019. *NewVantage Partners LLC*.

Dean, J. & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI)* (pp. 137-150).

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M. & Hutter, F. (2015). Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems,* pp. 2962-2970.

Gray, J. (1992). *Benchmark handbook for database and transaction processing systems*. Massachusetts, USA: Morgan Kaufmann.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. & Khan, S.U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, 47, 98-115.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Jouppi, N.P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A. et al. (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture,* pp. 1-12.

Kreps, J., Narkhede, N. & Rao, J. (2011). Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB,* pp. 1-7.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems,* pp. 1097-1105.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *META Group Research Note*, 6(70), 1.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Li, H., Ota, K. & Dong, M. (2018). Learning IoT in edge: Deep learning for the Internet of Things with edge computing. *IEEE Network*, 32(1), 96-101.

Lohr, S. (2014). For big-data scientists, 'janitor work' is key hurdle to insights. *New York Times*, 17.

McMahan, B. et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *AISTATS*.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Talwalkar, A. (2016). MLlib: Machine learning in Apache Spark. *The Journal of Machine Learning Research*, 17(1), 1235-1241.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Portland, USA: Broadway Books.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems,* pp. 8026-8037.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* pp. 1135-1144.

Shi, W., Cao, J., Zhang, Q., Li, Y. & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637-646.

Shvachko, K., Kuang, H., Radia, S. & Chansler, R. (2010). The Hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST),* pp. 1-10. IEEE.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B., ... & Xu, S. (2013). Data curation at scale: The data tamer system. In *CIDR,* pp. 1-10.

White, T. (2015). *Hadoop: The definitive guide* (4th ed.). California, USA: O'Reilly Media.

WHO. (2021). *Air quality and health report*.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M.J., Shenker, S. & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI),* pp. 15-28.