International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 03 | March – 2024 DOI: 10.55041/ISJEM01352
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

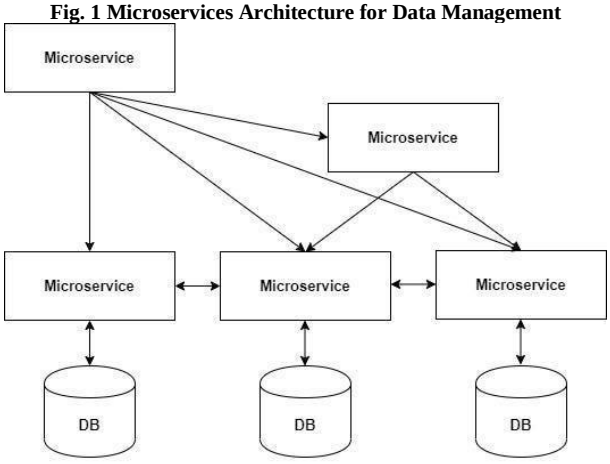
Scaling Data Infrastructure for High-Volume Manufacturing: Challenges and Solutions in Big Data Engineering

Tarun Parmar
(Independent Researcher)
Austin, TX
ptarun@ieee.org

***Abstract*—Scaling data infrastructure for high-volume manufacturing presents significant challenges owing to the rapid growth, diversity, and complexity of the data generated by modern production processes. This review explores the key challenges and solutions in big-data engineering to enable efficient, scalable, and reliable data management in manufacturing environments. The primary challenges include handling the volume, velocity, and variety of data; ensuring real-time processing and analysis; managing data storage and retrieval at scale; and maintaining data quality and consistency. To address these challenges, various big data engineering solutions have been discussed, including distributed computing frameworks, cloud-based storage and computing resources, data lakes, data governance and metadata management, stream processing technologies, machine learning, and AI for predictive analytics. This review also examines the role of data architecture and infrastructure in building scalable systems, highlighting the importance of microservices, containerization, orchestration, NoSQL databases, and data security and privacy. Performance optimization techniques, such as query optimization, data partitioning, sharding, caching, and data compression, have been explored to ensure efficient operation of large-scale data systems. The review includes case studies of successful implementations and discusses emerging trends, such as edge computing, and the growing importance of data interoperability and standardization. Future research directions were identified, emphasizing the need for ongoing development in this field to meet the ever-growing demand for high-volume manufacturing.**

***Keywords*—high-volume manufacturing, big data, data infrastructure, distributed computing frameworks, cloud computing, data lakes, microservices, containerization, orchestration, query optimization, data partitioning, sharding, caching, data compression**

Fig. 1 Microservices Architecture for Data Management



I. INTRODUCTION

Scaling data infrastructure for high-volume manufacturing involves navigating a complex landscape of challenges and solutions within the realm of big-data engineering. As manufacturers increasingly rely on data-driven decision making to enhance operational efficiency, they encounter significant hurdles, including rapid data growth, diverse data types, system performance limitations, cost control, and the need for skilled personnel. The ability to effectively manage and analyze vast datasets is crucial as it directly impacts production quality, maintenance practices, and overall competitiveness in the industry. Notably, the explosive growth of data generated by sensors, devices, and production processes poses critical storage and accessibility challenges [1]. Traditional

storage systems often struggle to cope with the volume and diversity of data, necessitating the adoption of advanced solutions, such as cloud computing and scalable data architectures.

Integrating data from multiple sources and formats poses difficulties for ensuring consistency and accuracy [2]. Real-time data processing and analysis are crucial for timely decision-making but require sophisticated systems to handle the high velocity of incoming information. Storing and retrieving large datasets efficiently becomes complex as data volumes grow exponentially. Maintaining the data quality and consistency across distributed systems is also a significant challenge.

To address these challenges, big-data engineering offers several solutions. Distributed computing frameworks such as Apache Hadoop and Spark enable the processing of large datasets across computer clusters. Cloud-based solutions provide scalable storage and computing resources, allowing manufacturers to adapt to changing data requirements [3] [4]. Data lakes offer centralized repositories for storing diverse data types, thus facilitating easier integration and analysis [5]. Implementing robust data governance and metadata management practices ensures data quality and consistency across infrastructure.

Stream-processing technologies, such as Apache Kafka and Flink, enable real-time data handling, which is crucial for monitoring production processes [6]. Machine learning and AI algorithms can be deployed for predictive analytics, helping to optimize manufacturing operations and predict maintenance requirements. These technologies can process large amounts of historical and real-time data to generate actionable insights.

The architecture of scalable data systems often employs microservices to construct flexible and modular data pipelines [7]. Containerization and orchestration technologies, such as Docker and Kubernetes, facilitate the easy scaling and management of data processing components. NoSQL databases are frequently used to handle unstructured data common in manufacturing environments [8]. Ensuring data security and privacy is paramount, particularly when scaling systems across multiple locations or to the cloud.

Optimizing the performance of large-scale data systems is crucial for their efficient operation. Techniques such as query optimization, data partitioning, and sharding help improve the database performance. Caching mechanisms can significantly reduce data access times for frequently used information [9]. Data compression techniques help manage storage requirements, whereas continuous monitoring and tuning ensures optimal system performance.

The successful implementation of scalable data infrastructure in manufacturing often involves a combination of these technologies and approaches. Case studies of leading manufacturers demonstrate how these solutions can lead to improved productivity, reduced downtime, and better-quality control. These examples provide valuable insights into the best practices and potential pitfalls in scaling data infrastructure.

Emerging trends in data infrastructure for manufacturing include edge computing, which brings data processing closer to the source and reduces latency and bandwidth usage [10]. The growing importance of data interoperability and standardization is driving efforts to create common data models and protocols across industries. These developments promise to further enhance the capabilities of the data infrastructure in supporting high-volume manufacturing.

In conclusion, scaling data infrastructure for high-volume manufacturing requires addressing complex challenges through innovative big-data engineering solutions. By implementing robust, scalable, and flexible data systems, manufacturers can harness the power of data to drive operational efficiency, quality, and innovation. As technology continues to evolve, ongoing research and development in this field will be crucial for meeting the ever-growing demands of modern manufacturing environments.

II. Challenges in Scaling Data Infrastructure

A. Volume, velocity and variety of data generated in high-volume manufacturing

Scaling data infrastructure, particularly in high-volume manufacturing environments, presents several critical challenges that organizations must navigate to maintain efficiency and performance. These challenges can be

broadly categorized as volume, velocity, and the variety of data generated in high-volume manufacturing [11] [12]. Addressing these challenges requires a multifaceted approach combining cutting-edge technologies, strategic planning, and skilled personnel. Organizations must invest in scalable cloud infrastructure, implement advanced data compression and archiving techniques, and leverage distributed computing frameworks to handle increasing data volumes. Additionally, developing a comprehensive data governance strategy and fostering a data-driven culture within an organization are crucial steps in overcoming the complexities of scaling data infrastructure in high-volume manufacturing environments.

High-volume manufacturing generates enormous volume, velocity, and variety of data, presenting significant challenges for data infrastructure scaling. The sheer quantity of data produced by sensors, machines, and processes can quickly overwhelm traditional data-management systems. Real-time data streams from production lines, quality control systems, and supply chain operations contribute to the high velocity of information flow. Additionally, the variety of data types, including structured, semi-structured, and unstructured data from diverse sources, further complicate data handling and analysis.

B. Data Integration from multiple sources and formats

Data integration from multiple sources and formats poses a major challenge in scaling data infrastructure. Manufacturing environments often rely on a combination of legacy systems, modern IoT devices, and third-party applications, each of which generates data in different formats [11]. Harmonizing these disparate data sources requires sophisticated ETL (Extract, Transform, Load) processes and data mapping techniques [13]. Ensuring data consistency and maintaining data lineages across various systems becomes increasingly complex as the number of data sources increases.

C. Challenges of real-time data processing and analysis

Real-time data processing and analysis present significant hurdles in high-volume manufacturing. The need for immediate insights to support rapid decision-making and process optimization requires low-latency data processing capabilities. Streaming analytics platforms must be able to handle massive data influxes while performing complex computations and applying machine learning algorithms in real-time [14] [15]. Balancing the trade-offs between processing speed, accuracy, and resource utilization is crucial for maintaining efficient operations.

D. Complexities of data storage and retrieval at scale

Data storage and retrieval at this scale introduce complexities in terms of infrastructure design and management. As data volumes grow exponentially, traditional relational databases may struggle to provide the required performance and scalability. Distributed storage systems and NoSQL databases offer potential solutions but come with their own challenges in terms of data consistency, fault tolerance, and query optimization [16]. The implementation of effective data partitioning, replication, and caching strategies is essential to ensure fast and reliable data access across geographically distributed manufacturing sites. However, NoSQL databases lack standardized data modeling methods, which can present challenges when handling database relationships.

E. Ensuring data quality and consistency

Ensuring data quality and consistency in large-scale manufacturing data infrastructure is a persistent challenge[17]. The high volume and velocity of the data increase the likelihood of errors, inconsistencies, and duplications. The implementation of robust data validation, cleansing, and deduplication processes is critical for maintaining data integrity. Additionally, enforcing data governance policies and maintaining data catalogs across the organization become more complex as the scale of data operations increases. Balancing the need for data quality with the demands for real-time processing and analysis requires sophisticated data-management strategies and continuous monitoring.

III. BIG DATA ENGINEERING SOLUTIONS

Addressing the multifaceted challenges of data integration necessitates innovative solutions. Distributed computing frameworks, such as Apache Hadoop and Spark, offer powerful tools for unifying data from diverse sources, showcasing how technology can directly address these integration hurdles.

International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 03 | March – 2024 DOI: 10.55041/ISJEM01352
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

A. Distributed computing frameworks for processing large datasets

Distributed computing frameworks play a crucial role in efficiently processing large datasets efficiently [18]. Apache Hadoop and Apache Spark are two prominent examples that enable parallel processing across computer clusters. These frameworks distribute data and computational tasks across multiple nodes, allowing for the faster processing of massive datasets. Hadoop's MapReduce paradigm and Spark's in-memory processing capabilities provided robust solutions for batch processing and iterative algorithms, respectively.

B. Cloud-based solutions for scalable storage and computing resources

Cloud-based solutions offer scalable storage and computing resources, which are essential for big-data engineering. Cloud platforms, such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure, provide a range of services tailored for big data processing. These include managed Hadoop and Spark clusters, object storage systems, and server-less computing options. Cloud solutions allow organizations to scale their infrastructure dynamically, reducing the need for large upfront investments in hardware and maintenance.

C. Role of Data lakes in centralizing and managing data

Data lakes have emerged as central repositories for storing diverse types of data in raw format [19] [20]. Unlike traditional data warehouses, data lakes can accommodate structured, semi-structured, or unstructured data. This flexibility allows organizations to store data from various sources without the need for an immediate schema definition. Data lakes facilitate data discovery, analytics, and machine learning by providing a single truth source for all data assets. Unlike the 'schema-on-write' approach of traditional databases that requires defining data structures before storage, data lakes employ a 'schema-on-read' methodology, allowing structured, semi-structured, and unstructured data [21].

D. Data governance and metadata management

Data governance and metadata management are critical components of successful big-data initiatives. Appropriate governance ensures data quality, security, and compliance with regulations. Metadata management provides context and lineage information for data assets, making it easier to understand and utilize the data effectively [21]. Implementing robust data-cataloging tools and establishing clear data ownership and access policies are essential steps in this process.

E. Stream processing technologies for real-time data handling

Stream processing has emerged as a crucial technology for handling real-time data in several domains. It enables the continuous processing of unbounded data streams, making it ideal for applications requiring timely decision making [22]. Stream-processing technologies enable real-time data handling, which is crucial for applications that require immediate insights [23]. Apache Kafka, Apache Flink, and Apache Storm are popular frameworks for processing continuous data streams. These technologies allow organizations to ingest, process, and analyze data in motion, enabling use cases, such as real-time fraud detection, IoT sensor data processing, and live recommendations.

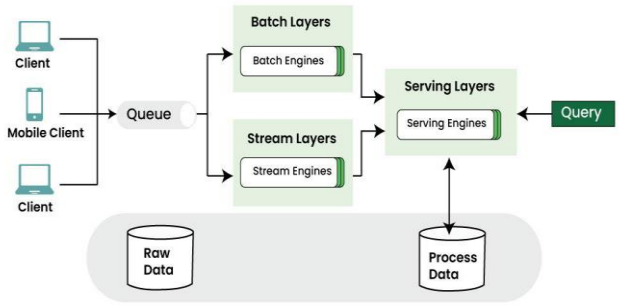


Fig. 2 Lambda Architecture for Batch and Stream Processing

F. Machine learning and AI for predictive analytics

Machine learning and AI implementation for predictive analytics are key aspects of modern big-data solutions. Frameworks like TensorFlow, PyTorch, and scikit-learn facilitate the development and deployment of machine learning models at scale. These tools, combined with big data processing capabilities, enable organizations to build sophisticated predictive models that can analyze historical data and make accurate forecasts. In healthcare, these technologies have been applied to predict patient outcomes, improve diagnoses, and enhance treatment strategies [24] [25]. Implementing MLOps ensures the smooth integration of machine learning models into

production environments, enabling continuous improvement and monitoring of model performance.

IV. Data Architecture and Infrastructure

Data architecture and infrastructure play crucial roles in building scalable and efficient data systems. Various architectural patterns have emerged to address the challenges of handling large volumes of data and the complex processing requirements. A detailed feature taxonomy for comparing and evaluating distributed database platforms has been created to address the challenges in designing massively scalable and highly available big data systems [26]. For instance, the Lambda architecture combines batch and stream processing to provide a comprehensive and real-time data analysis [Fig. 2]. Another popular pattern is the Kappa architecture, which simplifies the lambda approach using a single-stream processing engine for both real-time and historical data processing. These patterns enable organizations to design systems that can handle massive data influxes while maintaining performance and reliability.

Microservices have revolutionized the construction of data pipelines, offering increased flexibility and modularity. By breaking down complex data processing tasks into smaller, independent services, the microservice architecture allows for easier maintenance, scalability, and deployment of individual components. This approach enables teams to develop and update specific parts of the data pipeline without affecting the entire system, thereby resulting in faster iterations and improved overall system resilience. Additionally, microservices facilitate the use of different technologies and programming languages for different components, allowing organizations to leverage the most suitable tools for each specific task.

Containerization and orchestration have become essential elements for building scalable data systems. Containers provide a lightweight and portable environment for running applications and services, thereby ensuring consistency across different development and production environments [27]. For example, Docker has become a popular containerization platform, allowing developers to package applications with all their dependencies and deploy them seamlessly across various infrastructure setups. Orchestration tools, such as Kubernetes, complement containerization by automating the deployment, scaling, and management of containerized applications. This combination enables organizations to efficiently scale their data infrastructure up or down based on demand, optimize resource utilization, and ensure the high availability of services.

NoSQL databases have gained prominence in handling unstructured and semi-structured data, which are increasingly common in modern data ecosystems [28]. Unlike traditional relational databases, NoSQL databases offer flexible schema designs and horizontal scalability, making them well-suited for handling diverse data types and large-scale data processing. Document-oriented databases such as MongoDB, key-value stores such as Redis, and wide-column stores such as Cassandra provide different approaches to efficiently store and query unstructured data. These databases can accommodate the evolving nature of data structures and scale horizontally to handle growing data volumes, making them invaluable for organizations that deal with complex and varied data sources.

As data systems scale, ensuring data security and privacy becomes increasingly critical. Implementing robust security measures is essential to protect sensitive information from unauthorized access, breaches, and data leaks [29]. This includes employing encryption techniques for data at rest and in transit, implementing strong authentication and access control mechanisms, and regularly auditing and monitoring data-access patterns. Additionally, organizations must adhere to data privacy regulations, such as GDPR and CCPA, which require careful management of personal data and user consent. Implementing data governance frameworks and privacy-by-design principles ensures that security and privacy considerations are integrated into the data architecture from the ground-up, fostering trust and compliance in scaled data systems.

V. Performance Optimization

Query performance optimization in large-scale databases is crucial for efficient data retrieval and processing. Several techniques can be employed to enhance the query execution speed. Indexing is a fundamental strategy that creates data structures to quickly locate specific records, thereby reducing the need for full-table scans. Query plan optimization involves analyzing

and restructuring queries to minimize resource usage and execution time [30]. Materialized views can precompute and store complex query results, allowing for faster access to frequently requested data. Additionally, denormalization techniques can be applied judiciously to reduce the need for complex joins, thereby improving query performance.

Data partitioning and sharding are essential for managing large-scale databases. Partitioning involves dividing a large table into smaller, more manageable segments based on specific criteria such as date ranges or categories. This approach allows parallel processing and faster query execution on relevant partitions. Sharding extends this concept by distributing data across multiple servers or nodes, thereby enabling horizontal scaling and improving the performance [31]. Sharding strategies can be based on various factors, including a hash-based distribution, range-based partitioning, or a combination of methods. Proper implementation of these strategies can significantly enhance the query performance and overall system scalability.

Caching mechanisms play a vital role in improving data-access speeds. By storing frequently accessed data in memory or in faster storage tiers, caching reduces the need to repeatedly fetch information from slower primary storage. Various caching strategies can be employed, such as page, query result, and object caching. Distributed caching systems can further enhance performance by allowing multiple nodes to share cached data [32]. Implementing intelligent cache invalidation and update mechanisms ensures data consistency, while maintaining the performance benefits of caching.

Data compression is an effective technique for reducing storage requirements in large-scale databases. Compression algorithms can significantly decrease the amount of physical storage required, leading to cost savings and potentially improving I/O performance. Studies have shown that data-reduction techniques can decrease the volume of data transferred and stored by as much as 80% in some cases, resulting in substantial savings in storage and networking costs [33]. Column-based compression techniques are particularly effective for analytical workloads, because they can achieve high compression ratios for similar data types. However, it is important to balance the benefits of compression with the computational overhead of decompression during data retrieval. Adaptive compression techniques can optimize this trade-off by applying different compression methods based on the data characteristics and access patterns.

Monitoring and tuning system performance is critical for maintaining optimal database operations. Implementing robust monitoring tools allows administrators to track key performance indicators such as query execution times, resource utilization, and system bottlenecks. Regular performance audits can identify areas for improvement and guide optimization efforts [34]. Automated tuning mechanisms, such as self-tuning databases, can continuously adjust the system parameters based on workload patterns and resource availability. Additionally, proactive capacity planning and regular hardware upgrades ensure that the database infrastructure can be scaled to meet growing demands while maintaining performance standards.

VI. Future Trends

The future of data infrastructure scaling in high-volume manufacturing is poised for significant transformation driven by several emerging technologies. 5G networks have been set to revolutionize data transmission from factory floors, offering faster and more reliable connectivity. This enhanced communication capability enables the real-time monitoring and control of manufacturing processes with unprecedented precision. Quantum computing has the potential to dramatically increase data-processing capabilities, allowing for complex simulations and optimizations that were previously unfeasible. Blockchain technology is expected to play a crucial role in enhancing data security and traceability throughout supply chains, thereby ensuring the integrity and authenticity of manufacturing data.

Advanced AI and machine-learning models are becoming increasingly sophisticated, enabling more nuanced and predictive data analytics. These technologies will allow manufacturers to extract deeper insights from their data, leading to improved decision making and process optimization. In addition, the deployment of next-generation sensors and IoT devices will generate richer and more diverse datasets, providing a more comprehensive

view of manufacturing operations and enabling more granular control over processes.

Edge computing is emerging as a game changer in manufacturing data infrastructure. Edge computing offers several advantages by processing data closer to its source. It significantly reduces latency, enabling real-time decision making on factory floors, which is crucial for time-sensitive operations. The improved reliability of edge computing achieved through local data processing ensures that critical operations can continue even in the event of network disruptions. Furthermore, edge computing optimizes bandwidth usage by filtering and aggregating data at the source, thereby reducing the strain on network resources. This approach enhances data privacy and security by limiting the transmission of sensitive information. Most importantly, edge computing is a key enabler for autonomous systems and robotics in manufacturing, providing the local processing power required for these advanced technologies to operate effectively.

The growing importance of data interoperability and standards cannot be overstated in the context of scaling the data infrastructure for high-volume manufacturing. There has been a concerted effort across industries to develop common data models, facilitating easier integration and analysis of data from diverse sources. The adoption of standardized communication protocols, such as the OPC UA and MQTT, is gaining momentum, ensuring seamless communication between different systems and devices. Efforts are underway to create unified metadata schemas that will greatly simplify data integration processes and enable more efficient data management across the manufacturing ecosystem.

There is also a strong push for open data formats to facilitate information sharing both within and between organizations. Openness is crucial for fostering innovation and collaboration in the manufacturing sector. Industry bodies are increasingly collaborating to establish data exchange standards, recognizing the need for a unified approach to data management in an increasingly interconnected manufacturing landscape. The implementation of digital twins, which require standardized data representations to create accurate virtual models of physical assets and processes, further underscores the importance of data interoperability and standards in modern manufacturing.

These trends collectively indicate an ongoing evolution in data infrastructure to meet the growing demand for high-volume manufacturing. As these technologies mature and become more widely adopted, they enable manufacturers to achieve unprecedented levels of efficiency, flexibility, and innovation in their operations.

VII. CONCLUSION

Before The key challenges in high-volume manufacturing include managing large-scale data collection, ensuring data quality, integrating diverse sources, and implementing real-time analytics. Solutions involve the adoption of scalable cloud platforms, robust data governance, advanced analytics, and standardized data models. A scalable data infrastructure enables seamless data integration, provides computational power, supports real-time monitoring, enables predictive maintenance, and allows flexible resource scaling. Future research areas include advanced AI for predictive analytics, edge computing, enhanced cybersecurity, industry-specific data standards, and the integration of emerging technologies such as digital twins and augmented reality.

REFERENCES

[1] L.-M. Ang and K. P. Seng, “Big Sensor Data Applications in Urban Environments,” *Big Data Research*, vol. 4, pp. 1–12, Mar. 2016, doi: 10.1016/j.bdr.2015.12.003.

[2] M. Y. Santos *et al.*, “A Big Data Analytics Architecture for Industry 4.0,” springer, 2017, pp. 175–184. doi: 10.1007/978-3-319-56538-5_19.

[3] N. Ferry, P. Kalweit, G. Terrazas, S. Ratchev, D. Weinelt, and A. Solberg, “Towards a big data platform for managing machine generated data in the cloud,” Jul. 2017, pp. 263–270. doi: 10.1109/indin.2017.8104782.

[4] G. Terrazas, N. Ferry, and S. Ratchev, “A cloud-based framework for shop floor big data management and elastic computing analytics,” *Computers in Industry*, vol. 109, pp. 204–214, May 2019, doi: 10.1016/j.compind.2019.03.005.

[5] W. Yu, T. Dillon, F. Mostafa, W. Rahayu, and Y. Liu, "A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 183–192, May 2019, doi: 10.1109/tii.2019.2915846.

[6] K. Peddireddy, "Streamlining Enterprise Data Processing, Reporting and Realtime Alerting using Apache Kafka," May 2023, pp. 1–4. doi: 10.1109/isdfs58141.2023.10131800.

[7] R. Laigner, M. Kalinowski, Y. Liu, M. A. V. Salles, and Y. Zhou, "Data management in microservices," *Proceedings of the VLDB Endowment*, vol. 14, no. 13, pp. 3348–3361, Sep. 2021, doi: 10.14778/3484224.3484232.

[8] S. Sicari, A. Rizzardi, and A. Coen-Porisini, "Security&privacy issues and challenges in NoSQL databases," *Computer Networks*, vol. 206, p. 108828, Feb. 2022, doi: 10.1016/j.comnet.2022.108828.

[9] T. Le, M. Gerla, and Y. Lu, "Social caching and content retrieval in Disruption Tolerant Networks (DTNs)," Feb. 2015. doi: 10.1109/iccnc.2015.7069467.

[10] X. Li, M. Xia, M. Imran, J. Wan, H.-N. Dai, and A. Celesti, "A Hybrid Computing Solution and Resource Scheduling Strategy for Edge Computing in Smart Manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4225–4234, Jul. 2019, doi: 10.1109/tii.2019.2899679.

[11] P. Fang, J. Yang, L. Zheng, R. Y. Zhong, and Y. Jiang, "Data analytics-enable production visibility for Cyber-Physical Production Systems," *Journal of Manufacturing Systems*, vol. 57, pp. 242–253, Oct. 2020, doi: 10.1016/j.jmsy.2020.09.002.

[12] G. A. Lakshen, V. Janev, and S. Vranes, "Big data and quality: A literature review," Nov. 2016, vol. 4, pp. 1–4. doi: 10.1109/telfor.2016.7818902.

[13] S. Kavuri and S. Narne, "Improving Performance of Data Extracts Using Window-Based Refresh Strategies," *International Journal of Scientific Research in Science, Engineering and Technology*, pp. 359–377, Sep. 2021, doi: 10.32628/ijsrset2310631.

[14] J. Fu, K. Wang, and J. Sun, "SPARK – A Big Data Processing Platform for Machine Learning," Dec. 2016. doi: 10.1109/iciicii.2016.0023.

[15] D. García-Gil, S. García, S. Ramírez-Gallego, and F. Herrera, "A comparison on scalability for batch big data processing on Apache Spark and Apache Flink," *Big Data Analytics*, vol. 2, no. 1, Mar. 2017, doi: 10.1186/s41044-016-0020-2.

[16] S. Hamouda and Z. Zainol, "Document-Oriented Data Schema for Relational Database Migration to NoSQL," Aug. 2017, vol. 9, pp. 43–50. doi: 10.1109/innovate-data.2017.13.

[17] M. A. Serhani, A. Nujum, H. T. El Kassabi, and I. Taleb, "An Hybrid Approach to Quality Evaluation across Big Data Value Chain," Jun. 2016, pp. 418–425. doi: 10.1109/bigdatacongress.2016.65.

[18] S. Dolev, I. Singer, S. Sharma, P. Florissi, and E. Gudes, "A Survey on Geographically Distributed Big-Data Processing Using MapReduce," *IEEE Transactions on Big Data*, vol. 5, no. 1, pp. 60–80, Mar. 2019, doi: 10.1109/tbdata.2017.2723473.

[19] R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data Lakes: A Survey of Functions and Systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12571–12590, Dec. 2023, doi: 10.1109/tkde.2023.3270101.

[20] A. M. Olawoyin, A. Cuzzocrea, and C. K. Leung, "Open Data Lake to Support Machine Learning on Arctic Big Data," Dec. 2021, vol. 8, pp. 5215–5224. doi: 10.1109/bigdata52589.2021.9671453.

[21] F. Ravat and Y. Zhao, "Metadata Management for Data Lakes," springer, 2019, pp. 37–44. doi: 10.1007/978-3-030-30278-8_5.

[22] D. K. Lal and U. Suman, "Towards comparison of real time stream processing engines," Dec. 2019, pp. 1–5. doi: 10.1109/cict48419.2019.9066123.

[23] W. Hummer, B. Satzger, and S. Dustdar, "Elastic stream processing in the Cloud," *WIREs Data Mining and Knowledge Discovery*, vol. 3, no. 5, pp. 333–345, Aug. 2013, doi: 10.1002/widm.1100.

[24] N. Liu, Z. Zhang, A. F. Wah Ho, and M. E. H. Ong, "Artificial intelligence in emergency medicine," *Journal of Emergency and Critical Care Medicine*,

vol. 2, p. 82, Oct. 2018, doi: 10.21037/jeccm.2018.10.08.

[25] H. Wang, M. A. Ahmed, Z. Yang, Q. Zu, and J. Chen, "Application of Artificial Intelligence in Acute Coronary Syndrome: A Brief Literature Review.," *Advances in Therapy*, vol. 38, no. 10, pp. 5078–5086, Sep. 2021, doi: 10.1007/s12325-021-01908-2.

[26] I. Gorton, J. Klein, and A. Nurgaliev, "Architecture Knowledge for Evaluating Scalable Databases," May 2015, vol. 3, pp. 95–104. doi: 10.1109/wicsa.2015.26.

[27] Z. Zhong, M. A. Rodriguez, C. Xu, M. Xu, and R. Buyya, "Machine Learning-based Orchestration of Containers: A Taxonomy and Future Directions," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–35, Jan. 2022, doi: 10.1145/3510415.

[28] J. Bhogal and I. Choksi, "Handling Big Data Using NoSQL," Mar. 2015, pp. 393–398. doi: 10.1109/waina.2015.19.

[29] S. Chenthara, H. Wang, K. Ahmed, and F. Whittaker, "Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing," *IEEE Access*, vol. 7, pp. 74361–74382, Jan. 2019, doi: 10.1109/access.2019.2919982.

[30] J. Arnold, B. Glavic, and I. Raicu, "A High-Performance Distributed Relational Database System for Scalable OLAP Processing," May 2019, vol. 11, pp. 738–748. doi: 10.1109/ipdps.2019.00083.

[31] P. M. Dhulavvagol, V. H. Bhajantri, and S. G. Totad, "Performance Analysis of Distributed Processing System using Shard Selection Techniques on Elasticsearch," *Procedia Computer Science*, vol. 167, pp. 1626–1635, Jan. 2020, doi: 10.1016/j.procs.2020.03.373.

[32] S. Sasaki, Y. Oyama, O. Tatebe, K. Takahashi, and R. Matsumiya, "RDMA-Based Cooperative Caching for a Distributed File System," Dec. 2015. doi: 10.1109/icpads.2015.51.

[33] Y. Alforov, A. Novikova, J. Kunkel, T. Ludwig, and M. Kuhn, "Towards Green Scientific Data Compression Through High-Level I/O Interfaces," Sep. 2018, vol. 2, pp. 209–216. doi: 10.1109/cahpc.2018.8645921.

[34] S. Huang, Z. Li, X. Zhang, B. Cui, Y. Tu, and Y. Qin, "Survey on performance optimization for database systems," *Science China Information Sciences*, vol. 66, no. 2, Jan. 2023, doi: 10.1007/s11432-021-3578-6.