# Performance Improvement of Information Systems with File-based Data Storage

***Matthias Pohl, Paul Langisch***
*German Aerospace Center (DLR) - Institute of Data Science*
*Jena, Germany* *matthias.pohl, paul.langisch@dlr.de*

***Christian Haertel, Daniel Staegemann, Klaus Turowski***
*Otto von Guericke University*
*Magdeburg, Germany* *christian.haertel, daniel.staegemann, klaus.turowski@ovgu.de*

**Abstract**

Modern information systems, such as enterprise resource planning (ERP) systems, are vital for real-time decision-making, providing instant access to essential data. The rise of file-based storage engines provides valuable alternatives in information system architecture, especially for resource-constrained organizations. This study conducts comprehensive performance benchmarks across diverse database architectures, columnar file-based, in-memory, and traditional RDBMS, using standardized ERP workloads that simulate transaction processing, analytical reporting, and mixed operations with varying data volumes and concurrency levels.

**Keywords:** Information Systems, Database Benchmark, Duckdb, Performance, Data Storage

## 1. Introduction

In today's fast-paced business environment, modern information systems, such as enterprise resource planning (ERP) systems, are vital for enabling real-time decision-making, providing instant access to essential data. Key operations such as inventory checks and financial transactions are time-sensitive and cannot afford any delays [5]. As businesses evolve, they accumulate vast amounts of data, which in turn increases operational complexity. Modern ERP systems integrate information from source events, providing a comprehensive view on organizational activities [3], [8]. Furthermore, information systems support many users engaged in read and write operations simultaneously. To maintain performance under this heavy load, fast and efficient storage solutions are essential [2], [11]. Organizations are increasingly integrating real-time analytics into their information systems. This requires fast data access to support complex queries, enabling leaders to make informed decisions rapidly [4], [14]. Moreover, information systems manage high volumes of transactions that must be executed quickly and reliably, which is crucial for maintaining competitive advantage and operational efficiency in a fast-paced market [6], [12]. The rise of in-memory databases has transformed data storage in information systems. By utilizing RAM, these databases deliver performance improvements of 100 to 1,000 times for many operations. This technology eliminates traditional disk I/O limitations, enhancing both transactional and analytical workloads for smoother application performance. In-memory databases eliminate the need for separate OLTP and OLAP systems, allowing both processes to run on a single dataset, which simplifies management and enhances consistency. Advanced compression techniques help mitigate the costs of RAM storage while efficiently handling larger datasets [6], [11, 12]. While in-memory databases offer advantages, they also have significant drawbacks. RAM is considerably more expensive per gigabyte compared to SSDs and HDDs, leading to higher storage costs. Additionally, data in RAM is at risk of loss during power outages, requiring persistence solutions that can affect performance. The physical limitations of RAM can also restrict database size compared to disk-based systems, and specialized servers

with large RAM requirements raise setup costs. Transitioning to in-memory databases often requires substantial redesigns of existing applications, complicating the process. Finally, many in-memory databases come with higher licensing fees than traditional databases due to their advanced features. The rise of file-based storage engines provides valuable alternatives in information system architecture, especially for resource-constrained organizations [1], [10]. Small and medium enterprises (SMEs) with limited budgets can leverage solutions (e.g., DuckDB) to transform regular data files into efficient database repositories with minimal overhead [13]. These lightweight engines simplify database management by eliminating traditional administration needs while still offering analytical capabilities comparable to dedicated systems. This approach allows for the implementation of a functional database layer on existing file repositories, balancing performance and ease of use, and making advanced data processing accessible to SMEs.

Therefore, the paper at hand addresses the following research question. **RQ:** *To what extent can file-based database engines deliver the performance, scalability, and reliability requirements of modern enterprise information systems, particularly in ERP deployments with mixed transactional and analytical workloads?* This study conducts comprehensive performance benchmarks across diverse database architectures, columnar file-based, in-memory, and traditional RDBMS, using standardized ERP workloads that simulate transaction processing, analytical reporting, and mixed operations with varying data volumes and concurrency levels [9]. The measurements focus on throughput, latency, resource utilization, and scalability characteristics under controlled conditions, with particular attention to how each architecture handles the transition between OLTP and OLAP operations typical in modern ERP environments. The experiments are designed to provide actionable insights for database selection in enterprise information systems.

## 2. Comparison of Information System Architectures

Following an enterprise architecture (e.g., TOGAF [7]), the structure of an Information System can be designed by a 4-layer architecture. The Business Architecture encompasses the entire structure of an organization, characterized by its core business processes, the roles of its personnel, and the overall organizational framework. The Application Architecture is defined by the array of services and functionalities that the organization offers. The Data Architecture serves as a blueprint for managing the organizational data assets, including the types of data collected, the models that describe their relationships, and the information flows that support decision-making processes. The Technology Architecture outlines the IT infrastructure that supports the business functions. It comprises the hardware, software, and networking components in use, detailing how these technologies interact and are deployed to meet organizational requirements. In developing Information Systems, design characteristics can vary greatly depending on specific system requirements and objectives. Various realizations of reference models for business processes significantly influence both business and application architectures, shaping how components interact. The technology architecture is primarily determined by the database systems used, highlighting the need for careful selection. For example, open-source systems like Odoo typically rely on open databases such as PostgreSQL for flexibility and cost-efficiency. On the other hand, proprietary solutions like SAP S/4 utilize specialized databases like SAP HANA, offering advanced performance and integrated functionalities suited for enterprises. The research proposes a transformative approach to technology architecture by integrating modern, lightweight database engines (e.g., DuckDB) specifically designed for particular file formats. This method benefits SMEs by allowing them to use common tools like spreadsheets (e.g., Excel) as a backend. This not only simplifies data management but also enables the efficient conversion of spreadsheet data into high-performance files for rapid query processing. The performance benchmarks will assess the strengths and limitations of these database technologies to

determine their viability as alternatives in Information Systems architecture, potentially reshaping data handling practices for organizations.

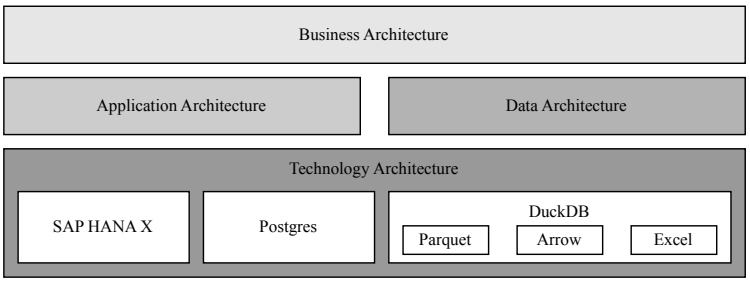


**Fig. 1.** Variation of Database Systems in the Technology Architecture of an Information System.

## 3. Experimental Performance Benchmarks

The performance benchmarks will follow TPC settings and include modifications for Online Transaction Processing (OLTP) that meet ERP system requirements [8]. This involves creating complex queries, ranging from simple aggregations of key performance indicators like total sales and average inventory, to advanced joins that analyze customer behavior, sales trends, and inventory turnover. A simulation environment will allow for the concurrent execution of OLTP and Online Analytical Processing (OLAP) operations, mirroring real-world scenarios where transactional data interacts with analytical reporting. Extensive testing will assess different read-to-write operation ratios, ensuring performance benchmarks are met and the system can handle real-time business demands.

**Testing Parameters:** The variation of different testing parameter allows for a comprehensive understanding of system performance. Utilizing datasets that span from a baseline of 100GB to a substantial upper limit of 10TB will assess the behavior of a system as it scales. To evaluate system performance under load, realistic user interactions will be simulated by 10 and 1,000 concurrent queries. In general, query patterns are defined by the TPC benchmarking, and the complexity of the data model is constructed by a realistic dataset of a SAP ERP system.

**Performance Metrics:** The performances will be measured with a specific set of metrics. First, the number of transactions or queries processed per second by a system will be obtained, indicating the ability of a higher throughput. Further, latency measures the time delay experienced in processing requests, expressed in response time percentiles. As the load increases, whether through additional data or an increase in the number of queries, the performance of the system will be assessed to measure its scalability. In general, the aspect of resource efficiency evaluates the utilization of system resources, including CPU, memory, storage, and network bandwidth.

**Testing Infrastructure:** Typical server setups range from 16 to 64 cores and 64 to 512 GB of RAM, catering to a wide range of workloads, from light applications to intensive processing. Utilizing cloud platforms like AWS, Azure, and GCP enables organizations to access scalable resources, supporting flexible computing that adapts to demand through services. Different network latencies impact performance. LAN provides the lowest latency, whereas WAN connections, especially those spanning long distances, can introduce delays that affect the user experience. Bandwidth limitations influence data transfer rates and application performance. Thus, the benchmarks will be conducted on both on-premise server configurations and cloud-based environments.

## 4. Conclusion

The benchmark study would establish clear performance thresholds where file-based database solutions remain competitive with traditional systems, demonstrating their efficiency for analytical workloads while identifying limitations under heavy concurrent write operations. We expect to document significant advantages in resource efficiency, with file-based solutions requiring substantially lower memory, CPU and storage resources compared to traditional databases for similar workloads. Finally, the research yield recommended architectural patterns for incorporating file-based engines within ERP implementations, particularly hybrid approaches leveraging their analytical strengths while compensating for transaction processing limitations.

## References

[1] Abadi, D., Boncz, P., Harizopoulos, S., Idreos, S., Madden, S., et al.: The design and implementation of modern column-oriented database systems. Foundations and Trends in Databases 5(3), pp. 197–280 (2013)

[2] Ahmad, Tahir and Waheed, Mehwish: Cloud computing adoption issues and applications in developing countries : a qualitative approach. The International Arab Journal of E-Technology 4(2), pp. 84–93 (2015)

[3] Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS quarterly pp. 1165–1188 (2012)

[4] Davenport, T.H.: Analytics 3.0. Harvard business review 91(12), pp. 64–72 (2013)

[5] Elragal, A., Haddara, M.: The future of erp systems: look backward before moving forward. Procedia Technology 5, pp. 21–30 (2012)

[6] Färber, F., May, N., Lehner, W., Große, P., Müller, I., Rauhe, H., Dees, J.: The sap hana database–an architecture overview. IEEE Data Eng. Bull. 35(1), pp. 28–33 (2012)

[7] Harrison, R.: TOGAF® 9 Foundation Study Guide. Van Haren (2016)

[8] Nambiar, R., Wakou, N., Carman, F., Majdalany, M.: Transaction processing performance council (tpc): State of the council. TPCTC'10, Springer-Verlag, Berlin, Heidelberg (2010)

[9] Osterthun, A., Pohl, M.: Foxbench: Benchmark for n-dimensional array file formats in data analytics environments. In: Datenbanksysteme für Business, Technologie und Web (BTW 2025). LNI, vol. P-361, pp. 545–564. Gesellschaft für Informatik e.V. (2025)

[10] Özcan, F., Tian, Y., Tözün, P.: Hybrid transactional/analytical processing: A survey. In: Proceedings of the 2017 ACM International Conference on Management of Data. pp. 1771–1775 (2017)

[11] Plattner, H.: A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases. Springer Publishing Company, Incorporated (2013)

[12] Plattner, H., Zeier, A.: In-Memory Data Management: Technology and Applications. Springer Publishing (2016)

[13] Raasveldt, M., Mühleisen, H.: Duckdb: an embeddable analytical database. In: Proceedings of the 2019 International Conference on Management of Data. p. 1981–1984. SIGMOD '19, ACM, New York, NY, USA (2019)

[14] Wixom, B., Watson, H.: The bi-based organization. International Journal of Business Intelligence Research (IJBIR) 1(1), pp. 13–28 (2010)