Benchmarking Study – Checkpoint 2

Identifying Use Case, Sample Data Sources, and Database Systems Study

Pooja Bendre, Ezhilarasu R, and Ritesh Kumar

2024WI_MS_DSP_420-DL_SEC61_SEC62: Database Systems

Benchmarking Study – Term Checkpoint 2

Abid Ali and Jaya R

February 26, 2024

Abstract:

In the rapidly evolving landscape of financial markets, efficient management and analysis of large datasets have become crucial. This study focuses on benchmarking various database systems—PostgreSQL, MySQL, Neo4J, and others—for analyzing daily equity Futures and Options (F&O) data from India's largest stock exchange. Employing Python and SQL, the research evaluates these databases on performance, efficiency, and resource utilization metrics. The dataset includes two decades of trading data from the National Stock Exchange of India, providing a rich source for analysis. The study aims to offer insights into the optimal database technology for financial market data analysis, aiding database administrators and financial analysts in making informed decisions. By comparing database and table creation times, query performance, and data handling efficiency, this research contributes to the financial technology field by elucidating each system's strengths and weaknesses in managing high-volume, dynamic financial data. Through a structured benchmark study, involving setup, execution of tests, and analysis, this project addresses the challenges of data volume, complexity, and fair comparison, aiming to provide actionable recommendations for technology selection in financial data analysis.

Introduction:

The finance sector's landscape is undergoing a transformative shift, propelled by the explosion of data generated in stock exchanges worldwide. At the heart of this transformation is the ability to analyze and manage vast quantities of financial data, a task that has become increasingly complex and critical. Daily equity Futures and Options (F&O) data from India's largest stock exchange, for instance, encapsulates a wealth of information crucial for market analysis, risk management, and decision-making processes. This underscores the

indispensable role of database systems, which act as the foundational infrastructure for storing, retrieving, and processing this colossal data efficiently.

However, the choice of database technology is not straightforward. The market offers a wide spectrum of database systems, from traditional relational databases like PostgreSQL and MySQL, renowned for their reliability and SQL support, to NoSQL options such as Neo4J, celebrated for their scalability and flexibility in handling unstructured data. Each of these systems presents unique strengths and limitations, especially when dealing with the intricacies of financial market data, characterized by its volume, velocity, and variety.

This research is motivated by the pressing need to navigate this complex landscape, aiming to identify the most suited database technology for handling financial data. By benchmarking a selection of database systems against key performance indicators—such as query performance, data handling efficiency, and resource utilization—we seek to provide a comprehensive analysis that aids database administrators and financial analysts in making informed decisions. This not only enhances the efficiency of financial data management but also contributes significantly to the broader domain of financial technology, driving innovation and optimization in financial market analysis.

Literature Review:

The field of financial data analysis has garnered significant interest, sparking numerous studies on enhancing database technologies for sophisticated datasets. A notable work, "A Performance Comparison of SQL and NoSQL Databases (2013)" by Le et al., investigated the efficiency of SQL versus NoSQL databases. Silva and Almeida highlighted SQL's robustness in managing Big Data across distributed, scalable systems, emphasizing its capacity to handle datasets characterized by large volume, high velocity, and diverse variety.

Additionally, an article by Matsiaka titled "PostgreSQL vs MySQL: A Detailed Comparison for Database Selection" for MarketSplash provided an in-depth analysis of these two databases. Shah et al., in their study "Performance Study of Time Series Databases," identified marked differences in data injection and query execution times between real and synthetic datasets. Furthermore, Dey et al.'s work, "Predictive Analytics with Structured and Unstructured Data – A Deep Learning Approach," introduced a deep learning framework for predictive analytics that leverages both structured and unstructured data.

These contributions highlight a movement towards a range of database technologies, each with distinct advantages for managing financial data. Our research builds on these insights, aiming to conduct a thorough benchmarking study to clarify the most suitable database system for analyzing financial market data.

Methods:

Our research methodology encompasses a structured approach to evaluate and compare various database systems for financial data analysis. Initially, we have selected a diverse set of database technologies including traditional relational databases like PostgreSQL and MySQL, alongside NoSQL options such as Neo4J. The criteria for selection were based on their popularity, unique features, and relevance to financial data management.

The core of our study involves the use of a comprehensive dataset from India's largest stock exchange, covering two decades of Futures and Options (F&O) daily trading data. This dataset is instrumental for benchmarking the databases on aspects such as data ingestion speed, query performance, scalability, and resource utilization under varying loads.

Our research employs Python for scripting and data manipulation, leveraging libraries such as pandas for dataset handling and SQLAlchemy for database interaction. SQL queries are used

for direct database operations, ensuring a consistent evaluation framework across all database systems.

We have designed a series of benchmark tests that include database and table creation times, data import efficiency, and the execution speed of various queries ranging from simple lookups to complex aggregations. Additionally, we assess the databases' capability to handle concurrent access and scalability challenges, simulating real-world application scenarios.

The outcome of these tests will be analyzed to identify performance trends, strengths, and weaknesses of each database system, providing a holistic view of their suitability for financial market data analysis.

Results:

From our comprehensive benchmarking study on various database systems using financial market data, we anticipate uncovering valuable insights into the performance, efficiency, and scalability of each system under real-world data analysis scenarios. Specifically, we expect to learn:

1. Performance Metrics: Which database systems offer the fastest query response times for both simple and complex queries, and how they manage data ingestion processes efficiently. We aim to quantify the speed and reliability of PostgreSQL, MySQL, and Neo4J when handling large datasets typical of financial markets.
2. Scalability and Concurrency: How each database technology scales with increasing data volumes and concurrent access requests. This will shed light on their capacity to support growing financial data analysis needs and the resilience of each system under high load conditions.
3. Resource Utilization: The comparative analysis of CPU, memory, and storage requirements for each database system will provide insights into their operational

efficiency. Understanding the resource footprint is crucial for evaluating the cost-effectiveness and sustainability of deploying these technologies in a financial data analysis context.

4. Suitability for Financial Data Analysis: By examining the specific features and limitations of each database system, such as support for complex queries, transaction integrity, and data model flexibility, we expect to identify which technologies are best suited for different types of financial data analysis tasks.
5. Trends and Innovations: Emerging trends in database technology and their implications for financial data analysis. This includes the potential of graph databases to uncover complex relationships in market data and the use of NoSQL databases for unstructured data analytics.

Ultimately, this study represents a modest effort to assist database administrators, financial analysts, and technologists in navigating the complex landscape of database technologies. By providing a detailed comparison of various database systems' performance, efficiency, and scalability when managing financial market data, we hope to offer practical insights that can aid in selecting the most suitable database solutions. This endeavor seeks to enhance the operational efficiency and effectiveness of financial data analysis, contributing to the broader goal of supporting professionals in adapting to the fast-paced changes within the financial sector.

Conclusion:

The conclusions from our forthcoming benchmarking study on database systems, focusing on their application in financial market data analysis, will emphasize the critical significance of selecting the appropriate database technology to align with specific financial data analysis needs. This decision extends beyond mere technical considerations, influencing the

efficiency, scalability, and overall success of financial operations significantly. Our anticipated findings are expected to unveil:

1. Performance Diversity: We anticipate discovering notable differences in performance among various database systems, with some excelling in query speed while others stand out in data ingestion efficiency. This expected outcome will highlight the importance of a discerning approach to database selection, where the unique requirements of financial data tasks are carefully aligned with the strengths of each database system.
2. Scalability and Concurrency Considerations: With the continuous growth in financial data volumes and the increasing need for real-time analysis, we foresee that the scalability and concurrency capabilities of database technologies will be critically evaluated. Our study aims to identify which databases can effectively manage large datasets and handle simultaneous queries without a drop in performance.
3. Resource Efficiency: The operational costs and resource utilization associated with different database systems will be a significant focus, guiding organizations towards optimizing their technology investments. We expect to shed light on database solutions that offer an optimal balance between performance and resource efficiency, facilitating more sustainable and cost-effective technology decisions.
4. Technological Appropriateness: The appropriateness of a database system is anticipated to vary depending on the specific use cases, be it managing high frequency trading data, analyzing complex financial networks, or conducting large-scale historical data analyses.

In conclusion, our study is set to provide foundational insights into the intricate relationship between database technologies and financial data analysis. By elucidating the comparative strengths and limitations of various database systems, this study could equip financial sector professionals with the knowledge to make informed, strategic technology choices, enhancing their analytical capabilities and operational effectiveness in a forward-looking manner.

References:

1. Data: NSE India Futures & Options Daily (2000-20) https://www.kaggle.com/datasets/tanay001/nseindia-futures-options-daily/data
2. NSE Historical Reports https://www.nseindia.com/resources/historical-reports-capital-market-daily-monthly-archives
3. Practical SQL – Second Edition, Debarros A., https://nostarch.com/practical-sql-2nd-edition
4. Data Engineering with Python, Crickard P., https://www.packtpub.com/en-fi/product/data-engineering-with-python-9781839214189?type=ebook
5. Efficient MySQL Performance, Nichter D., https://www.oreilly.com/library/view/efficient-mysql-performance/9781098105082/#:~:text=Daniel%20Nichter%20shows%20you%20how,the%20most%20important%20MySQL%20metrics
6. Building Knowledge Graphs: A Practitioner's Guide, Barrasa J. and Webber J, https://neo4j.com/knowledge-graphs-practitioners-guide/?utm_source=google&utm_medium=PaidSearch&utm_campaign=GDB&utm_content=AMS-X-SEM-Category-Expansion-Evergreen-Search&utm_term=&gad_source=1&gclid=CjwKCAiAiP2tBhBXEiwACslfnlvKh8bgC8n93lU8rhc_BJk5IR-eFhNCW5BNieF3L8AWfPS8rEagZBoC6ugQAvD_BwE
7. A performance comparison of SQL and NoSQL databases. IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing - Proceedings. 15-19. 10.1109/PACRIM.2013.6625441 – Li, Yishan & Manoharan, Sathiamoorthy. (2013). https://www.researchgate.net/publication/261079289_A_performance_comparison_of_SQL_and_NoSQL_databases

8. SQL: From Traditional Databases to Big Data – Silva, Yasin & Almeida, Isadora & Queiroz, Michell. (2016). https://www.researchgate.net/publication/311488672_SQL_From_Traditional_Databases_to_Big_Data
9. PostgreSQL Vs MySQL: A Detailed Comparison For Database Selection - Matsiaka, MarketSplash. https://marketsplash.com/tutorials/mysql/postgresql-vs-mysql/
10. Performance Study of Time Series Databases – Shah, Bonil & Jat, P. & Sashidhar, Kalyan. (2022). https://www.researchgate.net/publication/363128579_Performance_Study_of_Time_Series_Databases
11. Predictive Analytics with Structured and Unstructured Data – A Deep Learning Approach (2017) – Lipika Dey, Hardik Meisheri and Ishan Verma. https://www.comp.hkbu.edu.hk/~iib/2017/Dec/article5/iib_vol18no2_article5.pdf