



Lightning+ Benchmark Whitepaper

Real-Time Performance for AI Agentic Workloads

Overview

This document presents the results of a performance benchmark conducted on the Lightning+ in-memory data server. Designed for real-time AI agent retrieval, Lightning+ was evaluated under a demanding workload of 1,000,008 structured data queries, each consisting of a single, complex SQL query. The query randomized the target SYMBOL and selected a dynamic date range of up to 1 year. Despite this variability, Lightning+ demonstrated the ability to deliver low-latency, high-throughput responses. The entire benchmark was completed in under 20 minutes.

Data Source

The benchmark queries were executed against a preloaded in-memory dataframe containing S&P 500 daily stock data spanning from January 2000 onward.

Rows: 2,234,876 Format: Arrow

Tables Engine: Embedded DuckDB

Example Query

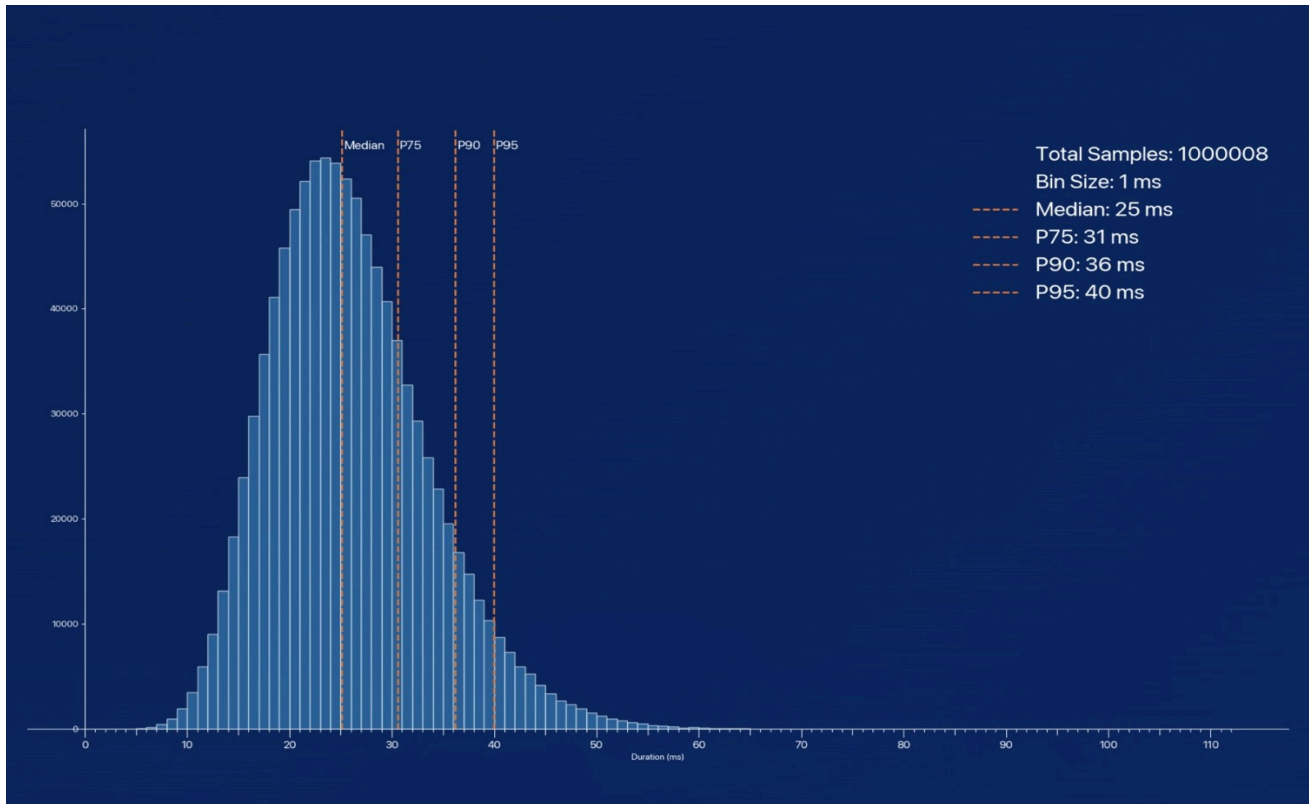
```
SELECT
    SYMBOL,
    DATE,
    CLOSE,
    100 * (CLOSE - FIRST_VALUE(CLOSE) OVER w) / FIRST_VALUE(CLOSE) OVER w AS RETURN_PCT
FROM sp500_daily
WHERE
    SYMBOL = ?
    AND DATE BETWEEN ? AND ?
WINDOW w AS (
    PARTITION BY SYMBOL
    ORDER BY DATE
    ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW
);
```

This query randomly selects a SYMBOL and generates a dynamic 1-year date range to calculate return percent.



Response Time Distribution

Histogram of Lightning+ Query Latencies



This histogram visualizes the latency distribution of 1,000,008 AI agentic queries processed by Lightning+. Over 95% of responses complete in under 40 milliseconds, with a sharp median at 25 ms.

Key Latency Metrics:

- Median (P50): 25 ms
- P75: 31 ms
- P90: 36 ms
- P95: 40 ms
- Maximum Latency Observed: 80 ms

Benchmark Duration: ~19 minutes

Sustained Throughput: ~52,600 responses per minute

These results validate Lightning+ as a consistent, real-time retrieval engine for structured data.



©2025 DADOS Technology. All rights reserved

Why This Benchmark Matters

Modern AI agents — from copilots to planners — depend on fast, accurate access to structured data. Yet most data infrastructure was built for dashboards, not decisions.

When retrieval is slow or stale, agents:

- Hallucinate
- Drop context
- Fail to act in real time

Lightning+ changes the game.

Built for real-time AI workloads, it eliminates ingestion delays and cloud latency by keeping data local, columnar, and query-ready.

Key Takeaways from the Benchmark

- 1,000,008 complex queries completed in under 20 minutes
- Median response time: 25 ms
- 95% of queries completed in < 40 ms
- Sustained throughput: 52,600+ queries per minute
- Max latency: under 100 ms — even under load

Why It Matters for Production AI

- Instant access to live, structured data — no prep pipelines required
- Consistent performance keeps agents responsive and grounded
- Edge-optimized — no egress, no cloud dependency, full control

Lightning+ is not just fast — it's ready for agent-scale AI.

- Learn more or request early access: <https://dadotech.com>



Appendix: Benchmark Configuration

Run Metadata		Aggregated Query Stats	
Host: Mac.attlocal.net		Avg Time: 25.95 ms	
Elapsed Time: 1083.21 sec		Min Time: 2.22 ms	
CPU Cores: 12		Max Time: 80 ms	
Timestamp: 20250529_050626		Median (P50): 25.11 ms	
Machine: Mac Studio 2023		Std Dev: 7.76 ms	
Chip: Apple M2 Ultra		P75: 30.56 ms	
Memory: 128 GB		P90: 36.21 ms	
OS: Sequoia		P95: 39.97 ms	
Requests/Thread: 41,667		Test Time: 1082.91 sec	
Threads: 24		Throughput: 923.4 qps	
Total Requests: 1,000,008			