S_CODA: An Emerging Pipeline for Evaluating Distributed Database Performance to Support Operational Data Analytics

Nicholas M. Synovic and George K. Thiruvathukal
Department of Computer Science, Loyola University Chicago
Chicago, IL, USA
{nsynovic,gkt}@luc.edu

Shilpika, Silvio Rizzi, Doug Waldron, and Michael E. Papka
Leadership Computing Facility, Argonne National Laboratory
Lemont, IL, USA
{shilpika,srizzi,dwaldron,papka}@anl.gov

Abstract

As high-performance computing (HPC) systems scale toward the exascale era, operational data analytics (ODA) play an increasingly central role in managing system security, health, scheduling, and scientific productivity. Supercomputing facilities continuously generate massive volumes of logs and system metrics. To make actionable insights, distributed database management systems (DBMSs) are often employed, but their behavior under realistic production HPC workloads remains underexplored. This poster presents S_CODA (Supercomputing Operational Data Analytics), an emerging benchmarking pipeline designed to evaluate distributed DBMS solutions—including relational, document, time-series databases and lakehouse solutions—using real and synthetic HPC environment logs. By working alongside our business intelligence colleagues to systematically model and implement common ODA workflows, S_CODA enables data-driven comparisons of competing DBMS platforms and identifies trade-offs in ingestion, querying, and concurrent access. We present our methodology, preliminary benchmarks, and lessons learned from applying S_CODA to multiple DBMS platforms at Argonne National Laboratory.

CCS Concepts

• **Information systems** → **Database performance evaluation**; *Parallel and distributed DBMSs*; Data stream management; • **Computer systems organization** → *High-performance computing*.

Keywords

high-performance computing, operational data analytics, distributed databases, benchmarking, exascale systems

ACM Reference Format:
Nicholas M. Synovic and, George K. Thiruvathukal, and Shilpika, Silvio Rizzi, Doug Waldron, and Michael E. Papka. 2025. S_CODA: An Emerging Pipeline for Evaluating Distributed Database Performance to Support Operational Data Analytics. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Supercomputing centers produce terabytes of operational data daily, spanning scheduler logs, hardware telemetry, and user activity [6, 7]. These data streams are critical for predictive maintenance, failure analysis, and planning for future systems. Anecdotally, we heard from our business intelligence team colleague that our current DBMS hosting systems consume substantial resources. As we transition to exascale systems, managing such volumes requires robust distributed DBMS infrastructures [9]. Prior work has explored ODA at HPC facilities, workflows, and systems [2–4, 8, 9]. However, limited prior work has systematically evaluated DBMS performance in the HPC operations context [8]. S_CODA addresses this gap by providing a reproducible benchmarking pipeline for operational data analytics workflows.

2 Goals

The primary goals of S_CODA are threefold: (1) benchmark relational, document, and time-series distributed databases and lakehouse systems [1] on realistic ODA workloads; (2) investigate performance trade-offs under concurrent ingress and egress patterns characteristic of HPC environments; and (3) enable stakeholders—including facility operators and researchers—to make informed decisions regarding database adoption and optimization for exascale readiness.

3 Methodology

Our methodology combines both real world and synthetic datasets to replicate HPC operational conditions. We leverage publicly available environment logs [5] and generate synthetic environment data following of supercomputing log schemas. DBMS platforms are deployed in containerized environments orchestrated by Docker Compose, ensuring portability and reproducibility. We design benchmarking workflows to capture:

- Sequential and batch ingestion patterns.
- Concurrent read/write (ingress/egress) scenarios.
- Data aggregation and query-intensive workloads.

Performance metrics include throughput, latency, and resource utilization.

4 Preliminary Results

Our preliminary benchmarks span three relational, two document, and two time-series DBMSs, and two lakehouse systems. We find that document databases generally outperform relational databases on both sequential and batch ingestion tasks. Time-series databases

offer efficient storage and retrieval for structured logs (e.g. temperature sensor data) but show limitations under high concurrency. Lakehouse solutions provide promising integration of analytic and transactional capabilities but require further optimization for HPC-scale logging.

5 Discussion

These findings highlight both opportunities and challenges in applying distributed DBMS technologies to HPC ODA workloads. No single solution emerges as universally superior: relational databases excel in transactional consistency, document stores provide flexible schema handling, time-series systems enable temporal analysis, and lakehouses integrate analytics. S_CODA 's pipeline approach reveals where trade-offs occur, empowering practitioners to tailor DBMS selection to workflow priorities. Future extensions will incorporate in-flight compression operations, machine learning-based anomaly detection, and real-time visualization as part of the benchmark suite.

6 Conclusions and Next Steps

S_CODA provides a foundation for evaluating distributed DBMS performance under realistic HPC workloads. By combining real and synthetic logs, containerized deployments, and workload-driven benchmarks, S_CODA advances understanding of the strengths and weaknesses of different database paradigms in supercomputing environments. Ongoing work will broaden the range of DBMS platforms and configurations benchmarked, integrate streaming data workflows-including online machine learning, in-flight data compression and decompression, and visualizations-and engage stakeholders across the HPC community to validate and refine the pipeline. Ultimately, S_CODA aims to guide both database vendors and HPC practitioners toward architectures capable of sustaining the demands of exascale operational data analytics.

Acknowledgments

This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy Office of Science user facility at Argonne National Laboratory. It was supported by the U.S. DOE Office of Science, Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357.

References

[1] Ben Lorica, Michael Armbrust, Reynold Xin, Matei Zaharia, and Ali Ghodsi. 2020. *What Is a Lakehouse?* https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

[2] Ben Lenard, Eric Pershey, Zachary Nault, and Alexander Rasin. 2023. An Approach for Efficient Processing of Machine Operational Data. In *Database and Expert Systems Applications: 34th International Conference, DEXA 2023, Penang, Malaysia, August 28–30, 2023, Proceedings, Part I* (Berlin, Heidelberg, 2023-08-28). Springer-Verlag, 129–146. doi:10.1007/978-3-031-39847-6_9

[3] Alessio Netti, Micha Müller, Carla Guillen, Michael Ott, Daniele Tafani, Gence Ozer, and Martin Schulz. 2020. DCDB Wintermute: Enabling Online and Holistic Operational Data Analytics on HPC Systems. In *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing* (New York, NY, USA, 2020-06-23) *(HPDC '20)*. Association for Computing Machinery, 101–112. doi:10.1145/3369583.3392674

[4] Michael Ott, Woong Shin, Norman Bourassa, Torsten Wilde, Stefan Ceballos, Melissa Romanus, and Natalie Bates. 2020. Global Experiences with HPC Operational Data Measurement, Collection and Analysis. In *2020 IEEE International Conference on Cluster Computing (CLUSTER)* (2020-09). 499–508. doi:10.1109/CLUSTER49012.2020.00071 ISSN: 2168-9253.

[5] Tapasya Patki, Barry Rountree, Torsten Wilde, Andrea Bartolini, Stephanie Brink, Esa Heiskanen, Sachin Idgunji, Matthias Maiterth, James Rogers, Ermal Rrapaj, Ralf Schneider, Woong Shin, Kathleen Shoga, Christian Simmendinger, Nicholas J Wright, and Zhengji Zhao. 2025. LAST/Power-Provisioning-Dataset at main · LLNL/LAST. https://github.com/LLNL/LAST/tree/main/Power-Provisioning-Dataset

[6] Sean Peisert. 2017. Security in high-performance computing environments. 60, 9 (2017), 72–80. doi:10.1145/3096742

[7] Sean Peisert, Thomas E. Potok, and Todd Jones. 2015. ASCR Cybersecurity for Scientific Computing Integrity - Research Pathways and Ideas Workshop. (2015). doi:10.2172/1236181

[8] Eric Pershey, Ben Lenard, Brian Toonen, Peter Upton, and Alexander Rasin. 2025. A Big Data Approach for Efficient Processing of Machine Operational Data. In *Proceedings of the 37th International Conference on Scalable Scientific Data Management* (New York, NY, USA, 2025-06-22) *(SSDBM '25)*. Association for Computing Machinery, 1–6. doi:10.1145/3733723.3733729

[9] Woong Shin, Tim Osborne, Ahmad Maroof Karimi, Rachel Palumbo, Alex May, Corwin Lester, Jesse Hines, Naw Safrin Sattar, Leah Huk, Scott Simmerman, Wesley Brewer, Jeffrey Miller, Ryan Adamson, Olga Kuchar, Ryan Prout, Feiyi Wang, Scott Atchley, and Sarp Oral. 2024. Navigating Exascale Operational Data Analytics: From Inundation to Insight. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2024-11). 1795–1804. doi:10.1109/SCW63240.2024.00226