Renan Souza, Ph.D.

✉ contact@renansouza.org renan-souza in renansouza1
ORCID 0000-0002-1794-808X RenanSouza.org
x9t36ewAAAAJ Citations: 816, h-index: 15, i10-index: 23

Summary

Tech lead, senior software engineer, and research scientist of intelligent data and AI platforms to accelerate scientific discovery. With 15+ years at IBM, ORNL, SLAC National Accelerator Laboratory, and the Federal University of Rio de Janeiro (UFRJ), I foster user-centric system design by keeping experts in the development loop, rapidly translating abstract requirements from domains such as Energy, Chemistry, Biology, and Climate into production-grade systems that are easier to operate, maintain, and scale. My research focuses on highly scalable, low-latency, observable, provenance- and metadata-first architectures that facilitate comprehensive data analysis across heterogeneous infrastructure, bridging edge instruments, cloud clusters, and leadership-class supercomputers, as well as data integration across SQL and NoSQL databases, knowledge graphs, messaging, streaming systems, and parallel file systems. My current focus includes AI and machine learning, LLM-driven workflows, and agentic systems. I authored 50+ papers, received best thesis and paper awards, held 10+ United States Patent and Trademark Office (USPTO) patents, and reviewed for major venues including IEEE Transactions on Parallel and Distributed Systems (TPDS), IEEE Big Data, IEEE eScience, Future Generation Computer Systems (FGCS), the Very Large Databases (VLDB) Journal, and ACM/IEEE Supercomputing.

Areas of Expertise

AI/ML, LLM-driven, and Agentic workflows ◆ Edge-Cloud-HPC computing ◆ Provenance-driven data analysis, lineage, and observability ◆ Scalable data engineering (SQL, NoSQL, KGs, Streaming, Parallel Data Processing)

Education

Federal University of Rio de Janeiro, Brazil - Rio de Janeiro, Brazil

Ph.D. in Computer Science | Sep 2015 — Dec 2019 | COPPE/Data and Knowledge Engineering
Thesis: Supporting User Steering in Large-scale Workflows with Provenance Data
Supervisor: Marta Mattoso and Patrick Valduriez.

M.Sc. in Computer Science | Jan 2013 — Jul 2015 | COPPE/Data and Knowledge Engineering
Thesis: Controlling the Parallel Execution of Workflows Relying on a Distributed Database
Supervisor: Marta Mattoso

B.Sc. in Computer Science | Jan 2009 — Dec 2012
Thesis: Linked Open Data Publication Strategies: An Application in Network Performance Data
Supervisor: Maria Luiza Machado Campos

International experience:
Visiting Ph.D. Student - Inria/Univ. Montpellier, France (Ph.D.)
Computer Science exchange student - Missouri State University, U.S. (B.Sc.)

Experience

Oak Ridge National Laboratory — **Oct 2022 — Present**
Staff Scientist & Sr. Software Engineer, HPC Workflows, Data & AI — **Knoxville, USA**

- Leading R&D on workflow provenance and observability for AI-driven science, focusing on transparency, reliability, and reproducibility in end-to-end workflows.
- Designing and developing provenance models and open source systems (e.g., Flowcept) to connect user intent, agent decisions, workflow executions, and downstream results in unified traces.
- Validated and applied these methods through high-profile projects in additive manufacturing, electron microscopy, and advanced biological analysis across Edge-Cloud-HPC environments.
- Published and presented results in HPC and eScience venues, and drove community engagement through tutorials and reference architectures.

IBM Research **Apr 2015 — Oct 2022**
Staff Scientist & Sr. Software Engineer, Cloud, Data & AI **Rio de Janeiro, Brazil**

- Led applied R&D on hybrid cloud and HPC data platforms for AI systems, advancing scalable architectures on Kubernetes and OpenShift for distributed, enterprise-grade workloads.
- Developed and validated knowledge graph-centric approaches for large-scale data integration, lineage, and governance across heterogeneous and distributed data stores and AI pipelines.
- Partnered closely with internal global teams and major external clients, particularly in the Energy sector, to translate research into deployable systems adopted in production.
- Produced sustained research impact through peer-reviewed publications and 10+ USPTO patents spanning provenance, polystores, AI lifecycle management, and hybrid cloud systems.

SLAC National Accelerator Laboratory **May 2013 — Dec 2014**
Research Software Engineering Intern **Menlo Park, USA**

- Applied semantic web and scalable data management methods to publish structured measurement data for broad community use.

Federal University of Rio de Janeiro **Jan 2010 — Sep 2014**
Software Engineer (Intern → Engineer) **Rio de Janeiro, Brazil**

- Led applied research on semantic web and linked open data systems, translating ontology-based models into production platforms for public-sector information access in user-facing systems.
- Developed data warehousing approaches for integrating structured and unstructured data to support big data analytics, reporting, and information discovery across heterogeneous sources.

Petrobras **May 2007 — May 2008**
IT Intern **Rio de Janeiro, Brazil**

- Early industry experience in software development and user support.

Technical Knowledge

Programming Languages: Python, Java, C, C++, C#, Shell, NodeJS, Scala, Lua
Data Science & ML: PyTorch, MLFlow, Airflow, Pandas, Polars, Jupyter, Matplotlib, Plotly
Agentic AI: MCP, LangChain, CrewAI, Streamlit, Chainlit, RAG, LLM-based orchestration
Big Data, Streaming, and Messaging: Spark, Dask, Parsl, Kafka, Redis, RabbitMQ
Databases & Data Lakes: PostgreSQL, MySQL Cluster, MongoDB, Elasticsearch, HBase, Hive, Redis, LMDB; Object Storages, Polystores, Data lakes, Data warehouses, and Data Lakehouses
Knowledge Graphs: AllegroGraph, Jena, Virtuoso, RDF, SPARQL, OWL
Parallel & Distributed Programming: MPI, OpenMP, CUDA, PubSub
Cloud, HPC, DevOps: Kubernetes, OpenShift; Slurm, LSF; Nvidia/AMD GPU Profiling; Prometheus, Grafana

Selected Projects

Orchestrated Platform for Autonomous Laboratories (OPAL)
OPAL (FAMOUS) advances autonomous science across multiple laboratories using AI agents, robotics, and automation, enabling HPC-scale, human-in-the-loop discovery workflows.

American Science Cloud (AmSC) **2025 — Present**
AmSC is a core pillar of the DOE Genesis Mission, delivering a secure, federated platform for AI-driven science across national laboratories. At ORNL, work within the Intelligent Interface team focuses on shaping agentic AI workflows that integrate data, compute, and facilities for scalable, reusable, mission-aligned discovery.

Advanced Manufacturing into Leadership-class Supercomputers via AI Agents **2025 — Present**
A core AmSC use case demonstrating agentic AI integration between advanced manufacturing facilities and leadership-class supercomputers. I provide technical leadership in defining the end-to-end architecture and translating the scientific vision into an operational platform, including multi-agent communication, provenance-aware infrastructure, and dynamic steering across facilities.

Flowcept **2023 — Present**

Flowcept is a provenance platform that captures runtime data with low overhead and links tasks, lineage, telemetry, and AI-agent interactions into end-to-end traces for accountability and reproducibility. I created and lead the platform, which underpins multiple DOE initiatives, such as OPAL (BER/ASCR) and broader Autonomous Science (ASCR-ACT), and research work on provenance for agentic workflows.

ProvLake **2018 — 2023**

ProvLake is a knowledge-graph-driven data lineage and management platform for hybrid data lakes spanning SQL databases, NoSQL stores, cloud object storage, and HPC file systems. It captures and integrates fine-grained data relationships across distributed workflows and services into a unified, semantic-rich provenance knowledge graph, enabling cross-store querying, explainability, and governance of complex AI and scientific pipelines. I created and led ProvLake as a foundational platform adopted across multiple IBM Research programs, supporting large-scale industry and internally funded research initiatives.

Selected Programmatic Leadership, Contributions, and Artifacts

OPAL - Orchestrated Platform for Autonomous Laboratories (FAMOUS) **2025 — Present**

- **Role:** Technical leadership for agentic AI and cross-facility workflow architecture
- **Scope and contributions:** Technical leader designing, implementing, and deploying the Agentic AI platform, translating expert intent into interactive HPC execution and multimodal analytics, using Flowcept for agentic provenance while integrating ORNL's APPL experimental workflows with OLCF Frontier in collaboration with biologists, AI researchers, and software engineers.
- **Validation and impact:** Demonstrated and reviewed with ORNL and ANL senior leadership in biological systems; endorsed by DOE leadership; publicly highlighted by the DOE Undersecretary; used within OPAL and informing Genesis-aligned autonomous laboratory efforts.

American Science Cloud (AmSC) **2025 — Present**

- **Role:** Technical contributor shaping agentic AI workflow frameworks
- **Scope and contributions:** Contributing to the early design of agentic AI workflow patterns to integrate data, compute, facilities, and AI agents across national laboratories, working with multidisciplinary teams of domain scientists, computational scientists, and engineers as the platform evolves.
- **Validation and impact:** Contributions inform ongoing architectural discussions within AmSC and related Genesis-aligned efforts (e.g., ModCon, FAMOUS).

Additive Manufacturing Agentic Workflow - MDF / AmSC Use Case **2025 — Present**

- **Role:** Technical leadership for agentic AI and cross-facility workflow architecture
- **Scope and contributions:** Acted as technical leader driving architectural convergence and resolving system design deadlocks within a multidisciplinary team of 10+ contributors, delivering an agentic cross-facility workflow connecting advanced manufacturing at MDF with the OLCF ACE Testbed. Established Flowcept-enabled provenance for agentic safety, transparency, and end-to-end workflow traceability, enabling agent-to-agent and agent-to-compute coordination with dynamic steering.
- **Validation and impact:** Established as a reusable architectural reference within AmSC and adopted as a core use case, positioning ORNL as a leader across DOE labs in cross-facility agentic workflows. Recognized by senior DOE leadership as a breakthrough example of AI agents leveraging leadership-class computing, with early publications at Supercomputing workshops and eScience. Supported AmSC infrastructure deployment and DOE booth demonstrations at Supercomputing’25, ensuring successful presentation to external audiences.

ModCon - Transformational AI Models Consortium **2025 — Present**

- **Role:** Technical collaborator supporting Genesis-aligned platform direction
- **Scope and contributions:** Collaborating with assigned ModCon team members by sharing architectural patterns, lessons learned, and implementation guidance from agentic, cross-facility workflow and provenance work, contributing to early design discussions within multidisciplinary teams.
- **Validation and impact:** Contributions inform ongoing technical discussions with senior ANL researchers and early prototypes.

INTERSECT LDRD: Multi-workflow Orchestration and Integrated Data Analysis Across Facilities **2024 — 2025**

- **Role:** Principal Investigator

- **Scope and contributions:** Led a program on multi-workflow orchestration and data analysis across facilities, coordinating teams of 10+ scientists, computer scientists, and engineers across NCCS and CSM divisions to orchestrate AI workflows between ORNL MDF to OLCF ACE Testbed and ORNL CNMS to OLCF Summit; and established Flowcept as a provenance foundation through partnerships across ORNL and ANL.
- **Validation and impact:** Managed $1M+ in funding; delivered cross-facility workflow demonstrations presented to senior DOE labs personnel, published in peer-reviewed venues; produced reusable orchestration and provenance artifacts (open source software and system architectures) that informed subsequent DOE efforts.

IBM Research - Knowledge-Centric Systems (Internal Strategic Programs) **2020 — 2022**

- **Role:** Technical leadership for platform architecture, integration, and reuse
- **Scope and contributions:** Led platform integration for the AI Workbench, enabling knowledge-centric scientific discovery across global IBM Research teams. Led provenance and knowledge management for context-aware platform reconfiguration, supporting hybrid cloud-HPC execution in interactive notebooks, and established ProvLake as a core shared platform adopted across IBM Research labs.
- **Validation and impact:** Peer-reviewed publications, executive-level demonstrations, cross-lab adoption, and a portfolio of patents supporting hybrid cloud-HPC platforms.

IBM Research - Oil & Gas AI Client Programs (Galp, Shell, ExxonMobil) **2018 — 2020**

- **Role:** Technical leadership for data lakes and provenance platforms
- **Scope and contributions:** Led the design and delivery of reusable data lake and provenance platforms supporting multiple concurrent AI programs for subsurface exploration, coordinating with 20+ researchers and engineers across IBM Research and industry-funded engagements. This work initiated the ProvLake research and development effort, which later evolved into a shared provenance platform adopted across programs and IBM Research labs worldwide.
- **Validation and impact:** Multi-year client adoption, executive-level demonstrations, media coverage, productization of platform components, peer-reviewed publications in scientific and industrial venues, and granted patents.

IBM Research - Conversational AI Platform (Pre-LLM) **2016 — 2018**

- **Role:** Technical leadership for scalability, deployment, and operations
- **Scope and contributions:** Led architectural design, DevOps, and scalability for a conversational AI platform coordinating user interaction with multiple bots in a same chat, supporting multiple concurrent users accessing the platform, coordinating engineers and researchers to achieve high availability and efficiency on cloud infrastructure.
- **Validation and impact:** Industry client demonstrations, peer-reviewed publications in AI venues, a highly cited (100+) patent, and successful scaling to thousands of concurrent users.

Programs listed reflect sustained platform leadership, cross-team coordination, and reusable system delivery across DOE national laboratory and industry contexts.

Selected Publications

Complete list of publications and patents is in the end of this document.

- **R. Souza**, T. J. Skluzacek, S. R. Wilkinson, M. Ziatdinov, and R. F. da Silva, "Towards Lightweight Data Integration using Multi-workflow Provenance and Data Observability," in IEEE International Conference on e-Science, 2023, doi: 10.1109/e-Science58273.2023.10254822.
- **R. Souza**, L. G. Azevedo, V. Lourenço, E. Soares, R. Thiago, et al., "Workflow Provenance in the Lifecycle of Scientific Machine Learning," Concurrency and Computation: Practice and Experience, 2021, doi: 10.1002/cpe.6544.
- **R. Souza**, A. Gueroudji, S. DeWitt, D. Rosendo, T. Ghosal, et al., "PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows," in 2025 IEEE International Conference on eScience (eScience), 2025, doi: 10.1109/eScience65000.2025.00093.
- **R. Souza**, S. Caino-Lores, M. Coletti, T. J. Skluzacek, A. Costan, et al., "Workflow Provenance in the Computing Continuum for Responsible, Trustworthy, and Energy-Efficient AI," in IEEE International Conference on e-Science, 2024, doi: https://doi.org/10.1109/e-Science62913.2024.10678731.

Selected Institutional Recognition

- ORNL performance award for top performers (2025)
- IBM Patent Plateaus for high-impact software innovations (8+ USPTO patents) (2020, 2021)

Selected Academic Recognition

- 2025, Distinguished Paper Award, WORKS @ IEEE/ACM Supercomputing — The (R)evolution of Scientific Workflows in the Agentic AI Era: Towards Autonomous Science
- 2021, Runner-up (2nd Place) - Best Ph.D. Thesis — User Steering Support in Large-scale Workflows
- 2017, Honorable Mention - Best Paper — Spark Scalability Analysis in a Scientific Workflow
- 2015, Best M.Sc. Thesis Award — Parallel Execution of Workflows driven by Distributed Database Techniques

Media Highlights

Galp-IBM AI Platform for Seismic Interpretation link
Led workflows management for AI and knowledge-engineering for an AI-assisted seismic interpretation platform developed at IBM Research and deployed in production with Galp. The system integrated machine learning, visual analytics, and domain knowledge to accelerate geological decision-making on large-scale seismic data.

AI4Seismic: AI Platform for Geological Discovery link
Highlighted by SIAM News, this presentation introduced AI4Seismic, an end-to-end AI platform for seismic analysis that captures expert geological knowledge and integrates ML, provenance, and reproducible workflows to accelerate energy-critical geological discovery.

OPAL: AI-Assisted Biological Discovery with Frontier link1 link2 link3
ORNL Communications highlighted the OPAL project's Agentic AI platform as an early, high-impact result of integrating laboratory instruments with the Frontier exascale supercomputer to enable autonomous, human-in-the-loop biological discovery under the DOE Genesis Mission.

DOE Genesis Mission Platform Demonstration link
Work referenced in U.S. Congressional testimony by DOE undersecretary for science highlighting early Genesis Mission milestones, including AI-enabled workflows that autonomously coordinate experiments, HPC execution, and analysis across national laboratories.

Grants and Fellowships

CAPES International Science Grant (2012-2013)
Competitive national fellowship supporting international research exchange, enabling academic placement at Missouri State University and research internship at SLAC National Accelerator Laboratory.

CAPES Master's Scholarship (2013-2015)
Nationally funded graduate research scholarship.

Scientific Community Service

Chair and Editor

- Frontiers in High Performance Computing - Editorial Board
- IEEE International Conference on e-Science (eScience'23) - Session Chair
- Brazilian Symposium on Databases (SBBD'20) - Session Chair

Technical Program Committee

- International Workshop on AI Principles in Science Communication (AISC'25)
- Int. Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'25,26)
- Workflows in Distributed Environments (WiDE'24)
- IEEE/ACM Supercomputing (SC'24)
- IEEE International Conference on e-Science (eScience'23)

- Workflows in Support of Large-Scale Science (WORKS at IEEE/ACM Supercomputing'20, 21, 23, 24, 25)
- Brazilian Workshop on Database and Artificial Intelligence Integration
- Brazilian Symposium on Databases (SBBD'20, 23, 24, 25, 26)
- Brazilian e-Science (BreSci'26)
- Innovation Summit on Information Systems (at SBSI'19,20)

Journal Reviewer

- IEEE Transactions on Parallel and Distributed Systems
- Future Generation Computer Systems
- Concurrency Computation Practice and Experience
- Journal of Parallel and Distributed Computing
- The Very Large Databases (VLDB) Journal
- IEEE Transactions on Big Data
- Journal of Cloud Computing
- Computer Physics Communications
- Discover Data
- Frontiers in High Performance Computing

Teaching, Supervisions, and Mentoring

Courses

- Databases Laboratory (UFRJ, 2017). Teacher assistant to Prof. Marta Mattoso
- Semantic Web (UFRJ, 2013). Teacher assistant to Prof. Maria Luiza Machado Campos
- Logics for Computer Science (UFRJ, 2012-2013). Teacher assistant to Prof. Mario Benevides
- Metadata Management (UFRJ, 2011). Teacher assistant to Prof. Adriana Vivacqua

Academic Supervisions

- Pedro Paiva Miranda, *A Mechanism for Fault Tolerance in Parallel Executions of Workflows supported by a Database* (undergraduate, 2015)
- Rachel de Castro, *Publication of Workflow Provenance Data in the Semantic Web* (undergraduate, 2015)

Internship Supervisions

- Timothy Poteet — ORNL
- Luke Christenson — ORNL
- Rennan Gaio — IBM Research
- Lucca Martins — IBM Research
- Marcelo Costalonga — IBM Research
- Additional interns advised (names available upon request)

Early-Career Guidance and Onboarding Support

- Tyler Skluzacek — ORNL
- Daniel Rosendo — ORNL

Badges and Certifications

- Machine Learning Specialist Professional (2022). Exploratory Data Analysis, Regression, Classification, Deep Learning, Reinforcement Learning, Unsupervised Learning, Time Series and Survival Analysis, AI Ethics and Explainability
- Trustworthy AI and AI Ethics (2022). Trustworthy AI foundations and applied AI ethics
- Enterprise Design Thinking Practitioner (2022)
- LinkedIn Skill Assessment: Python, MySQL, Linux, T-SQL, NoSQL

Languages

- English: Full professional proficiency
- Portuguese: Native
- Spanish: Reading fluent; speaking/listening intermediate

Invited Talks and Conference Presentations

- **Agentic AI for User Facilities** (2026) | Lawrence Berkeley National Laboratory
- Oral presentation: Provenance Data in Agentic Workflows
- **Workflows Community Initiative** (2025) | Remote
- Oral presentation: For Provenance and with Provenance: The Role of Provenance Data in Agentic Workflows
- **IEEE/ACM Supercomputing (SC)** (2025) | St. Louis, U.S.A.
- Oral presentation: LLM Agents for Interactive Workflow Provenance: Reference Architecture and Evaluation Methodology
- **IEEE International Conference on e-Science** (2025) | Chicago, U.S.A
- Tutorial: Large-scale Workflow Provenance Data Management in the AI Lifecycle using Flowcept
- Oral presentation: PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows
- **IEEE International Conference on e-Science** (2024) | Osaka, Japan (Virtual)
- Oral presentation: Workflow Provenance in the Computing Continuum for Responsible, Trustworthy, and Energy-Efficient AI
- **IEEE/ACM Supercomputing (SC)** (2024) | Atlanta, GA
- Oral presentation: Integrating Evolutionary Algorithms with Distributed Deep Learning for Optimizing Hyperparameters on HPC Systems
- **IEEE/ACM Supercomputing (SC)** (2023) | Denver, CO
- **IEEE International Conference on e-Science** (2023) | Limassol, Cyprus
- Oral presentation: Towards Lightweight Data Integration using Multi-workflow Provenance and Data Observability
- **Brazilian Symposium on Databases (SBBD)** (2021) | Rio de Janeiro, RJ (virtual)
- Oral presentation: User Steering Support in Large-Scale Workflows
- **Federal Fluminense University (UFF) Computer Science Seminars** (2021) | Rio de Janeiro, RJ (virtual)
- Invited talk (Portuguese): A Knowledge-centric Approach to Support Large-scale AI Systems
- **SIAM Conference on Computational Science and Engineering** (2021) | Forth Worth, TX (virtual)
- Invited talk, Highlighted by the SIAM press: AI4Seismic: An AI-Driven Platform to Accelerate Geological Discoveries
- Oral presentation: Workflow Provenance in the Lifecycle of Scientific Machine Learning
- **ACM International Conference on Management of Data (SIGMOD)** (2020) | Portland, OR (virtual)
- **Brazilian Symposium on Databases (SBBD)** (2020) | Rio (virtual)
- **High-Performance Data Science workshop** (2020) | Rio (virtual)
- **Computational Science and Engineering Seminar at COPPE/UFRJ** (2020) | Rio (virtual)
- Invited talk: Workflow Provenance in the Lifecycle of Scientific Machine Learning
- **Open Subsurface Data Universe Development Workshop** (2020) | Houston, TX
- **Open Subsurface Data Universe Development Workshop** (2019) | Houston, TX
- **IEEE/ACM Supercomputing (SC)** (2019) | Denver, CO
- **Scientific Data Analysis using Data-intensive Scalable Computing Workshop** (2019) | Rio de Janeiro, Brazil
- Invited talk: Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering
- **Open Subsurface Data Universe F2F Meeting** (2019) | Houston, TX
- **IEEE International Conference on e-Science** (2019) | San Diego, CA
- Oral presentation: Efficient Runtime Capture of Multiworkflow Data using Provenance
- **Inria Talks** (2019) | Montpellier, France
- Invited talk: Providing Online Data Analytical Support for Humans in the Loop of Computational Science and Engineering Applications
- **IBM Regional Technical Exchange** (2019) | Rio de Janeiro, Brazil

- **Provenance Week** (2018) | London, UK
- **International Conference on Very Large Databases (VLDB)** (2018) | Rio de Janeiro, Brazil
- **Brazilian Syposium on Databases (SBBD)** (2018) | Rio de Janeiro, Brazil
- **Brazilian Syposium on Databases (SBBD)** (2017) | Uberlandia, Brazil
- Oral presentation: Spark Scalability Analysis in a Scientific Workflow
- Oral presentation: Controlling the Parallel Execution of Workflows Relying on a Distributed Database
- **Federal University of Uberlandia, Brazil** (2017) | Uberlandia, Brazil
- Invited talk: Kubernetes
- **Smart City Cloud Hackathon OpenStack Rio** (2017) | Rio de Janeiro, Brazil
- **Computer Science Week at UFRJ** (2017) | Rio de Janeiro, Brazil
- Oral presentation: Kubernetes
- **Brazilian Conference on Artificial Intelligence (BRACIS)** (2017) | Recife, Brazil
- Tutorial: Graph Analytics with Spark
- **IEEE/ACM Supercomputing (SC)** (2016) | Salt Lake City, UT
- **ASE BigData/SocialCom/CyberSecurity** (2014) | Stanford University, Menlo Park, CA
- poster presentation: Linked open data publication strategies: Application in networking performance measurement data

All Publications and Patents

Journal Articles

[J1] M. Dorier, A. Gueroudji, V. Hayot-Sasson, H. Nguyen, S. Ockerman, **R. Souza**, T. Bicer, H. Pan, P. Carns, K. Chard, and others, "Toward a Persistent Event-Streaming System for High-Performance Computing Applications," *Frontiers in High Performance Computing*, 2025. DOI: 10.3389/fhpcp.2025.1638203

[J2] **R. Souza**, V. Silva, A.A.B. Lima, D. Oliveira, P. Valduriez, and M. Mattoso, "Distributed In-memory Data Management for Workflow Executions," *PeerJ Computer Science*, 2021. DOI: 10.7717/peerj-cs.527

[J3] **R. Souza**, L.G. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E.V. Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira, and M.A.S. Netto, "Workflow Provenance in the Lifecycle of Scientific Machine Learning," *Concurrency and Computation: Practice and Experience*, 2021.

[J4] L.G. Azevedo, **R. Souza**, R. Brandão, V.N. Lourenço, M. Costalonga, M.d. Machado, M. Moreno, and R. Cerqueira, "Adding Hyperknowledge-enabled data lineage to a machine learning workflow management system for oil and gas," *First Break*, 2020. DOI: 10.3997/1365-2397.fb2020055

[J5] **R. Souza**, V. Silva, J.J. Camata, A.L.G.A. Coutinho, P. Valduriez, and M. Mattoso, "Keeping Track of User Steering Actions in Dynamic Workflows," *Future Generation Computer Systems*, 2019. DOI: 10.1016/j.future.2019.05.011

[J6] V. Silva, L. Neves, **R. Souza**, A.L.G.A. Coutinho, D.d. Oliveira, and M. Mattoso, "Adding Domain Data to Code Profiling Tools to Debug Workflow Parallel Execution," *Future Generation Computer Systems*, 2018. DOI: 10.1016/j.future.2018.05.078

[J7] **R. Souza**, V. Silva, A.L.G.A. Coutinho, P. Valduriez, and M. Mattoso, "Data Reduction in Scientific Workflows Using Provenance Monitoring and User Steering," *Future Generation Computer Systems*, 2017. DOI: 10.1016/j.future.2017.11.028

Conference and Workshop Papers

[C1] **R. Souza**, A. Gueroudji, S. DeWitt, D. Rosendo, T. Ghosal, R. Ross, P. Balaprakash, and R.F.d. Silva, "PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows," *2025 IEEE International Conference on eScience (eScience)*, 2025. DOI: 10.1109/eScience65000.2025.00093

[C2] **R. Souza**, T. Poteet, B. Etz, D. Rosendo, A. Gueroudji, W. Shin, P. Balaprakash, and R.F.d. Silva, "LLM Agents for Interactive Workflow Provenance: Reference Architecture and Evaluation Methodology," *Workflows in Support of Large-Scale Science (WORKS) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2025. DOI: 10.1145/3731599.3767582

[C3] W. Shin, **R. Souza**, D. Rosendo, F. Suter, F. Wang, P. Balaprakash, and R.F.d. Silva, "The (R) evolution of Scientific Workflows in the Agentic AI Era: Towards Autonomous Science," *Workflows in Support of Large-Scale Science (WORKS) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2025.

[C4] D. Rosendo, S. DeWitt, **R. Souza**, P. Austria, T. Ghosal, M. McDonnell, R. Miller, T.J. Skluzacek, J. Haley, B. Turcksin, and others, "AI Agents for Enabling Autonomous Experiments at ORNL’s HPC and Manufacturing User Facilities," *Extreme-Scale Experiment-in-the-Loop Computing (XLOOP) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2025. DOI: 10.1145/3731599.3767592

[C5] A. Gueroudji, T. Mallick, **R. Souza**, R.F.d. Silva, R. Ross, M. Dorier, P. Carns, K. Chard, and I. Foster, "ControlA: Agentic Workflow Control Mechanisms for Reliable Science," *IEEE International Conference on e-Science*, 2025.

[C6] **R. Souza**, S. Caino-Lores, M. Coletti, T.J. Skluzacek, A. Costan, F. Suter, M. Mattoso, and R.F.d. Silva, "Workflow Provenance in the Computing Continuum for Responsible, Trustworthy, and Energy-Efficient AI," *IEEE International Conference on e-Science*, 2024. DOI: https://doi.org/10.1109/e-Science62913.2024.10678731

[C7] T.J. Skluzacek, **R. Souza**, M. Coletti, F. Suter, and R.F.d. Silva, "Towards Cross-Facility Workflows Orchestration through Distributed Automation," *Practice and Experience in Advanced Research Computing (PEARC 24)*, 2024. DOI: 10.1145/3626203.3670606

[C8] R.F.d. Silva, W. Shin, F. Suter, A. Gainaru, **R. Souza**, D. Dietz, and S. Jha, "Eco-Driven AI-HPC: Optimizing Energy Efficiency in Distributed Scientific Workflows," *Energy-Efficient Computing for Science Workshop*, 2024.

[C9] R.F.d. Silva, K. Maheshwari, T. Skluzacek, **R. Souza**, and S. Wilkinson, "Advancing Computational Earth Sciences: Innovations and Challenges in Scientific HPC Workflows," *European Geosciences Union (EGU)*, 2024.

[C10] M. Coletti, **R. Souza**, T.J. Skluzacek, F. Suter, and R.F.d. Silva, "Integrating Evolutionary Algorithms with Distributed Deep Learning for Optimizing Hyperparameters on HPC System," *Workflows in Support of Large-Scale Science (WORKS) workshop co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2024.

[C11] L.G. Azevedo, **R. Souza**, E. Soares, R.M. Thiago, J.C.C. Tesolin, A.C.C.M. Oliveira, and M.F. Moreno, "HKPoly: A Polystore Architecture to Support Data Linkage and Queries on Distributed and Heterogeneous Data," *Proceedings of the 20th Brazilian Symposium on Information Systems (SBSI)*, 2024. DOI: 10.1145/3658271.3658322

[C12] **R. Souza**, T.J. Skluzacek, S.R. Wilkinson, M. Ziatdinov, and R.F.d. Silva, "Towards Lightweight Data Integration using Multi-workflow Provenance and Data Observability," *IEEE International Conference on e-Science*, 2023. DOI: 10.1109/e-Science58273.2023.10254822

[C13] D. Rosendo, M. Mattoso, A. Costan, **R. Souza**, D. Pina, P. Valduriez, and G. Antoniu, "ProvLight: Efficient Workflow Provenance Capture on the Edge-to-Cloud Continuum," *IEEE International Conference on Cluster Computing*, 2023. DOI: 10.1109/CLUSTER52292.2023.00026

[C14] **R. Souza**, "User Steering Support in Large-scale Workflows," *PhD Thesis Contest: Brazilian Symposium on Databases (SBBD)*, 2021.

[C15] E. Soares, **R. Souza**, R. Thiago, M. Machado, and L. Azevedo, "A Recommender for Choosing Data Systems based on Application Profiling and Benchmarking," *Brazilian Symposium on Databases (SBBD)*, 2021.

[C16] R.L. Cunha, L.V. Real, **R. Souza**, B. Silva, and M.A. Netto, "Context-aware Execution Migration Tool for Data Science Jupyter Notebooks on Hybrid Clouds," *IEEE International Conference on e-Science*, 2021. DOI: 10.1109/eScience51609.2021.00013

[C17] L. Azevedo, **R. Souza**, E. Soares, R. Thiago, A. Oliveira, and M. Moreno, "Supporting Polystore Queries using Provenance in a Hyperknowledge Graph," *International Semantic Web Conference (ISWC)*, 2021.

[C18] **R. Souza**, J. Camata, M. Mattoso, and A. Coutinho, "Runtime Steering of Parallel CFD Simulations," *International Conference on Parallel Computational Fluid Dynamics*, 2020.

[C19] R. Thiago, **R. Souza**, L. Azevedo, E. Soares, R. Santos, W. Santos, M.D. Bayser, M. Cardoso, M. Moreno, and R. Cerqueira, "Managing Data Lineage of O\&G Machine Learning Models: The Sweet Spot for Shale Use Case," *European Association of Geoscientists and Engineers (EAGE) Digitalization Conference and Exhibition*, 2020. DOI: 10.3997/2214-4609.202032075

[C20] **R. Souza**, A. Codas, J.A.N. Junior, M.P. Quinones, L. Azevedo, R. Thiago, E. Soares, M. Cardoso, and L. Martins, "Supporting the Training of Physics Informed Neural Networks for Seismic Inversion Using Provenance," *American Association of Petroleum Geologists Annual Convention and Exhibition (AAPG)*, 2020.

[C21] R. Brandão, V. Lourenço, M. Machado, L. Azevedo, M. Cardoso, **R. Souza**, G. Lima, R. Cerqueira, and M. Moreno, "A Knowledge-Based Approach for Structuring Cyclic Workflows," *International Semantic Web Conference (ISWC)*, 2020.

[C22] R. Brandão, V. Lourenço, M. Machado, L. Azevedo, M. Cardoso, **R. Souza**, G. Lima, R. Cerqueira, and M. Moreno, "Cycle Orchestrator: A Knowledge-Based Approach for Structuring Cyclic ML Pipelines in the O\&G Industry," *International Semantic Web Conference (ISWC)*, 2020.

[C23] L. Azevedo, **R. Souza**, E. Soares, and M. Moreno, "Modern Federated Databases: an Overview," *International Conference on Enterprise Information Systems (ICEIS)*, 2020.

[C24] L. Azevedo, **R. Souza**, R. Thiago, E. Soares, and M. Moreno, "Experiencing ProvLake to Manage the Data Lineage of AI Workflows," *Innovation Summit on Information Systems (EISI) in Brazilian Symposium in Information Systems (SBSI)*, 2020.

[C25] **R. Souza**, L. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E.V. Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira, and M.A.S. Netto, "Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering," *Workflows in Support of Large-Scale Science (WORKS) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2019. DOI: 10.1109/WORKS49585.2019.00006

[C26] **R. Souza**, E.V. Brazil, L. Azevedo, R. Ferreira, D. Chevitarese, E. Soares, R. Thiago, M. Nery, V. Torres, and R. Cerqueira, "Managing Data Traceability in the Data Lifecycle for Deep Learning Applied to Seismic Data," *American Association of Petroleum Geologists Annual Convention and Exhibition (AAPG)*, 2019.

[C27] **R. Souza**, L. Azevedo, R. Thiago, E. Soares, M. Nery, M. Netto, E.V. Brazil, R. Cerqueira, P. Valduriez, and M. Mattoso, "Efficient Runtime Capture of Multiworkflow Data Using Provenance," *IEEE International Conference on e-Science*, 2019. DOI: 10.1109/eScience.2019.00047

[C28] P. Valduriez, M. Mattoso, R. Akbarinia, H. Borges, J. Camata, A.L.G.A. Coutinho, D. Gaspar, N. Lemus, J. Liu, H. Lustosa, F. Masseglia, F.N.D. Silva, V. Silva, **R. Souza**, K. Ocaña, E. Ogasawara, D. Oliveira, E. Pacitti, F. Porto, and D. Shasha, "Scientific Data Analysis Using Data-Intensive Scalable Computing: the SciDISC Project," *LADaS: Latin America Data Science Workshop*, 2018.

[C29] **R. Souza**, L. Neves, L. Azeredo, R. Luiz, E. Tady, P. Cavalin, and M. Mattoso, "Towards a human-in-the-loop library for tracking hyperparameter tuning in deep learning development," *Latin American Data Science (LaDaS) workshop co-located with the Very Large Database (VLDB) conference*, 2018.

[C30] M.G.d. Bayser, C. Pinhanez, H. Candello, M. Affonso, M.P. Vasconcelos, M.A. Guerra, P. Cavalin, and **R. Souza**, "Ravel: A MAS orchestration platform for Human-Chatbots Conversations," *International Workshop on Engineering Multi-Agent Systems (EMAS@AAMAS 2018)*, 2018.

[C31] **R. Souza** and M. Mattoso, "Provenance of Dynamic Adaptations in User-Steered Dataflows," *Provenance and Annotation of Data and Processes - International Provenance and Annotation Workshop (IPAW)*, 2018. DOI: 10.1007/978-3-319-98379-0_2

[C32] V. Silva, **R. Souza**, J. Camata, D.d. Oliveira, P. Valduriez, A.L.G.A. Coutinho, and M. Mattoso, "Capturing Provenance for Runtime Data Analysis in Computational Science and Engineering Applications," *Provenance and Annotation of Data and Processes - International Provenance and Annotation Workshop (IPAW)*, 2018. DOI: 10.1007/978-3-319-98379-0_15

[C33] **R. Souza**, "Parallel Execution of Workflows driven by Distributed Database Techniques," *MSc Thesis Contest: Brazilian Symposium on Databases (SBBD)*, 2017.

[C34] **R. Souza**, V. Silva, J. Camata, A. Coutinho, P. Valduriez, and M. Mattoso, "Tracking of online parameter fine-tuning in scientific workflows," *Workflows in Support of Large-Scale Science (WORKS) workshop co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2017.

[C35] **R. Souza**, V. Silva, P. Miranda, A.A.B. Lima, P. Valduriez, and M. Mattoso, "Spark Scalability Analysis in a Scientific Workflow," *Brazilian Symposium on Databases (SBBD)*, 2017.

[C36] P. Cavalin, F. Figueiredo, M.d. Bayser, L. Moyano, H. Candello, A. Appel, and **R. Souza**, "Building a question-answering corpus using social media and news articles," *International Conference on Computational Processing of the Portuguese Language*, 2016.

[C37] J.J. Camata, J.M. Cela, D. Costa, A.L.G.A. Coutinho, D. Fernández-Galisteo, **R. Souza**, and others, "Applying future Exascale HPC methodologies in the energy sector," *Russian Supercomputing days*, 2016.

[C38] J. Camata, J.M. Cela, D. Costa, A.L.G.A. Coutinho, D. Fernández-Galisteo, C. Jimenez, V. Kourdioumov, M. Mattoso, R. Mayo-García, T. Miras, J.A. Moríñigo, J. Navarro, P.O.A. Navaux, D.D. Oliveira, M. Rodríguez-Pascual, V. Silva, **R. Souza**, and P. Valduriez, "Enhancing Energy Production with Exascale HPC Methods," *CARLA: Latin American High Performance Computing Conference*, 2016. DOI: 10.1007/978-3-319-57972-6_17

[C39] **R. Souza**, V. Silva, A. Coutinho, P. Valduriez, and M. Mattoso, "Online Input Data Reduction in Scientific Workflows," *Workflows in Support of Large-Scale Science (WORKS) workshop co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2016.

[C40] V. Silva, L. Neves, **R. Souza**, A. Coutinho, D.D. Oliveira, and M. Mattoso, "Integrating Domain-data Steering with Code-profiling Tools to Debug Data-intensive Workflows," *Workflows in Support of Large-Scale Science (WORKS) workshop co-located with the ACM/IEEE International Conference for High Performance Computing, Networking,* *Storage, and Analysis (SC)*, 2016.

[C41] R. Castro, **R. Souza**, V. Silva, K. Ocaña, D. Oliveira, and M. Mattoso, "Uma Abordagem para Publicação de Dados de Proveniência de Workflows Científicos na Web Semântica," *Brazilian Symposium on Databases (SBBD)*, 2015.

[C42] T. Barbosa, **R. Souza**, S. Cruz, M. Campos, and R.L. Cottrell, "Applying data warehousing and big data techniques to analyze internet performance," *International Conference on Internet Applications, Protocols, and Services*, 2015.

[C43] **R. Souza**, V. Silva, D. Oliveira, P. Valduriez, A.A.B. Lima, and M. Mattoso, "Parallel Execution of Workflows Driven by a Distributed Database Management System," *ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2015.

[C44] **R. Souza**, L. Cottrell, B. White, M.L. Campos, and M. Mattoso, "Linked open data publication strategies: Application in networking performance measurement data," *ASE BigData/SocialCom/CyberSecurity, Stanford, CA*, 2014.

Technical Reports

[R1] B. Etz, S. Oral, R.F.D. Silva, R. Adamson, A. Alnajjar, T. Beck, A. Barker, M. Brim, P. Bryant, **R. Souza**, and others, "OLCF's Advanced Computing Ecosystem (ACE): FY25 Update for Ongoing Efforts," **, 2025. DOI: 10.2172/3006499

[R2] D. Bard, K. Chard, S.d. Witt, I.T. Foster, C. Goble, W. Godoy, J. Gustafsson, U. Haus, S. Hudson, L. Los, **R. Souza**, and others, "Workflows Community Summit 2024: Future Trends and Challenges in Scientific Workflows," *Distributed, Parallel, and Cluster Computing (cs.DC)*, 2024. DOI: https://doi.org/10.48550/arXiv.2410.14943

[R3] R.F.d. Silva, R.M. Badia, V. Bala, D. Bard, P. Bremer, I. Buckley, S. Caino-Lores, K. Chard, C. Goble, S. Jha, ..., **R. Souza**, and e. al., "Workflows Community Summit 2022: A Roadmap Revolution," *arXiv preprint Distributed, Parallel, and Cluster Computing (cs.DC)*, 2023. DOI: 10.48550/arXiv.2304.00019

[R4] L.G. Azevedo, **R. Souza**, E.F.d.S. Soares, R.M. Thiago, J.C.C. Tesolin, A.C. Oliveira, and M.F. Moreno, "A Polystore Architecture Using Knowledge Graphs to Support Queries on Heterogeneous Data Stores," *arXiv preprint Databases (cs.DB)*, 2023. DOI: 10.48550/arXiv.2308.03584

[R5] R.F.d. Silva, H. Casanova, K. Chard, ..., **R. Souza**, and e. al., "Workflows Community Summit: Advancing the State-of-the-art of Scientific Workflows Management Systems Research and Development," *arXiv preprint Distributed, Parallel, and Cluster Computing (cs.DC)*, 2021.

[R6] M.G.d. Bayser, P. Cavalin, **R. Souza**, A. Braz, H. Candello, C. Pinhanez, and J. Briot, "A Hybrid Architecture for Multi-party Conversational Systems," *arXiv preprint Computation and Language (cs.CL)*, 2017.

Patents

[P1] M.A.S. Netto, L.C.V. Real, B. Silva, and **R. Souza**, "Shortened narrative instruction generator for software code change," *US Patent App. 17/819,025*, 2024.

[P2] L.C.V. Real, R.L.D.F. Cunha, **R. Souza**, and M.A.S. Netto, "Data transformation for acceleration of context migration in interactive computing notebooks," *US Patent App. 17/683,279*, 2023.

[P3] M.A.S. Netto, B. Silva, R.L.D.F. Cunha, **R. Souza**, and L.C.V. Real, "Remotely healing crashed processes," *US Patent App. 17/480,087*, 2023.

[P4] L.C.V. Real, M.A.S. Netto, R.L.D.F. Cunha, **R. Souza**, and A. Braz, "Program context migration," *US Patent App. 17/216,817*, 2022.

[P5] L.C.V. Real, R.L.D.F. Cunha, M.N.d. Santos, and **R. Souza**, "Asset identification for collaborative projects in software development," *Granted, US Patent App. 17/118,646*, 2022.

[P6] A.P. Appel, C.R.L.D. Freitas, **R. Souza**, C.R.D.A. Mendes, A. Vital, N. Dos, S. Marcelo, M.A.S. Netto, P.B. Avegliano, and C. Villas, "Model Document Creation in Source Code Development Environments using Semantic-aware Detectable Action Impacts," *US Patent App. 17/353,731*, 2022.

[P7] **R. Souza**, R. Mozart, F.R.D. Silva, A. Vital, and V.T.d. Silva, "Metadata-based scientific data characterization driven by a knowledge database at scale," *Granted, US Patent App. 16/527,546*, 2021.

[P8] L.C.V. Real, M.N.d. Santos, and **R. Souza**, "Continuous storage of data in a system with limited storage capacity," *Granted, US Patent App. 16/678,375*, 2021.

[P9] M.G.D. Bayser, A. Braz, P.R. Cavalin, F. Figueiredo, and **R. Souza**, "Creating coordinated multi-chatbots using natural dialogues by means of knowledge base," *Granted, US Patent Application 15/217,660*, 2018.

[P10] A. Braz, P.R. Cavalin, F. Figueiredo, M.G.D. Bayser, and **R. Souza**, "System and method for managing artificial conversational entities enhanced by social knowledge," *Granted, US Patent Application 15/265,615*, 2018.

[P11] A.P. Appel, A.G. Leal, and **R. Souza**, "Predicting user question in question and answer system," *Granted, US Patent Application 15/171,055*, 2017.