

# Renan Souza

✉ [contact@renansouza.org](mailto:contact@renansouza.org) • 🌐 [RenanSouza.org](https://RenanSouza.org) • in [renansouza1](https://www.linkedin.com/in/renansouza1)  
🔑 ID: x9t36ewAAAAJ • 🌐 [renan-souza](https://github.com/renan-souza) • Updated on June 26, 2025.

## Bio

---

Renan Francisco Santos Souza holds a Ph.D., M.Sc., and B.Sc. in Computer Science (2009–2019) from the Federal University of Rio de Janeiro (UFRJ). Since 2022, he has been a staff research scientist at the Oak Ridge National Laboratory. From 2015 to 2022, he worked as a research scientist and software engineer at IBM Research. During his Ph.D., he was a visiting researcher at Inria, France. During his B.Sc., he spent one year at Missouri State University and interned at the SLAC National Laboratory at Stanford University. He has been active as a software engineer, researcher, and technical lead on multiple projects since 2010. He was awarded with the best M.Sc. thesis and an honorable mention for best Ph.D. thesis from SBBD, the leading data science and engineering conference in Latin America. His research and development focus on large-scale data management and AI to support the next generation of Edge–Cloud–HPC workflows.

## Research Interests

---

Large-scale Data Science and Data Engineering • Edge–Cloud–HPC Workflows • Provenance Data • Big Data Analytics • Machine Learning Systems • Agentic AI

## Education

---

- **Ph.D. in Computer Science**, Federal Univ. of Rio de Janeiro, Brazil Sep 2015 – Dec 2019  
Supervised by Marta Mattoso (COPPE/UFRJ) and Patrick Valduriez (Inria).  
Title: Supporting User Steering in Large-scale Workflows with Provenance Data
- Visiting Ph.D. Student, Inria/Univ. Montpellier, France Jan 2019 – Mar 2019  
Supervised by Patrick Valduriez (Inria).
- **M.Sc. in Computer Science**, Federal Univ. of Rio de Janeiro, Brazil Jan 2013 – Jul 2015  
Supervised by Marta Mattoso (COPPE/UFRJ).  
Title: Controlling the Parallel Execution of Workflows Relying on a Distributed Database
- Computer Science exchange student, Missouri State University, U.S. Jun 2011 – Jun 2012
- **B.Sc. in Computer Science**, Federal Univ. of Rio de Janeiro, Brazil Jan 2009 – Dec 2012  
Supervised by Maria Luiza Machado Campos (DCC/UFRJ).  
Title: Linked Open Data Publication Strategies: An Application in Network Performance Data
- **Technical Degree in Information Systems**, Lemos de Castro Jan 2005 – Dec 2007

## Experience

---

- **Oak Ridge National Laboratory** Oct 2022 – Present  
**Research Scientist, HPC Workflows, Data & AI** Knoxville, United States  
As Principal Investigator and lead software developer, he leads the design and development of large-scale data and AI systems within the Workflows and Ecosystem Services group at ORNL, focusing on AI-driven observability, data integration, workflow provenance, and LLM-based agentic workflows to accelerate scientific discovery in Edge–Cloud–HPC environments.

- **IBM Research**  
**Research Scientist and Software Eng., Cloud, Data & AI**

As a Staff Research Scientist (2021–2022), he was a lead researcher and developer on projects in large-scale data science and engineering to support AI systems in hybrid cloud and cluster environments with highly distributed and heterogeneous workloads for clients across energy, finance, physics, and cheminformatics domains. Although primarily focused on backend development, data engineering, cloud, and DevOps, he collaborated closely with front-end developers and HCI researchers to design domain-specific applications. As a Research Software Engineer (2015–2021), he worked with and led R&D projects on large-scale data integration for AI systems in the energy sector, targeting cluster and cloud platforms. He also led the Cloud DevOps team responsible for developing and deploying conversational AI systems.

As a Software Engineering Intern (2015), he designed and implemented big data and machine learning solutions for real-time analysis of streaming social data.

**Apr 2015 – Oct 2022**  
**Rio de Janeiro, Brazil**
- **SLAC National Accelerator Laboratory, Stanford Univ.**  
**Research Software Engineering intern**

Led the development of a cloud platform utilizing semantic web, big data, and data warehousing techniques. This platform is designed to store, retrieve, visualize, and publish structured data about internet performance worldwide, enabling understanding of global Internet quality.

**May 2013 – Dec 2014**  
**Menlo Park, United States**
- **COPPE-UFRJ**  
**Software Engineer**

As a Lead Software Engineer (2013–2014), he led the development of a system that facilitated access of information about public services offered by the Brazilian Federal Government. Also lead the development of a platform to publish linked open data from the Brazilian Federal Register on the semantic web, applying agile practices, ontology-based data modeling, and natural language processing.

As a Full-Stack Software Engineering Intern (2011–2013), he worked on the R&D of various web systems.

**Dec 2011 – Sep 2014**  
**Rio de Janeiro, Brazil**
- **Federal Univ. of Rio de Janeiro**  
**Software Engineering intern**

Developed a system that integrated data warehouse environments with both structured and unstructured data, enabling the generation of more intelligent and flexible information reports.

**Jan 2010 – Jul 2011**  
**Rio de Janeiro, Brazil**
- **Petrobras**  
**IT Intern**

Implemented features and provided ongoing maintenance for web systems supporting Petrobras employees.

**May 2007 – May 2008**  
**Rio de Janeiro, Brazil**

## Technical Knowledge

---

- **Languages:** Python, Java, C, C++, Shell scripting, NodeJS, Scala, Lua
- **Data Science/ML Technologies:** Pandas, Polars, Jupyter Notebooks, Numpy, Matplotlib, Seaborn, Plotly, Tensorflow, ScikitLearn, Keras, PyTorch, MLFlow, Airflow, Grafana
- **Agentic AI and LLMs:** MCP Agents, Crew AI (Multi-agent Framework), LangChain, Streamlit for AI Agents, RAG, Prompt Engineering Techniques
- **Big Data & Parallel Processing Frameworks:** Dask; Apache Spark: RDD, DataFrames, Streaming, MLib, GraphX, GraphFrames; Hadoop Ecosystem
- **Cloud and Cluster computing:** VMs, Dockers, Kubernetes, OpenShift, HPC (Slurm, LSF, PBS)
- **DevOps:** Containers, Kubernetes, OpenShift, CI/CD Pipelines, GitHub, GitHub Actions, Travis, Jenkins
- **Message Queueing Systems:** Kafka, RabbitMQ, Redis
- **GPU Programming and Profiling:** NVIDIA and AMD Python APIs for GPU performance analysis
- **Relational DBMS:** PostgreSQL/PostGIS, DB2, SQLite, MySQL, MySQL Cluster, MS SQL Server
- **NoSQL DBMS:** MongoDB, AllegroGraph, Jena, Blazegraph, Virtuoso, Sesame, Cloudant, CouchBase, Redis, Impala, Elasticsearch, HBase, Hive, Apache Ignite, LMDB
- **Heterogeneous Data Management:** Data Integration, Multi-database Queries, Polystores
- **Cluster Deployment:** YARN, Mesos, Standalone deployment
- **Business Intelligence:** MS SQL Server BI developer studio, Pentaho Solutions, Talend;
- **Semantic Web Tools/Languages:** OWL, RDF, SPARQL, Protege

- **Distributed and Concurrent Programming:** PubSub, MPI, OpenMP, CUDA, Data-centric distributed and parallel programming
- **Web Development:** Python Flask/UWSGI, Java EE, Tomcat/JBoss, Spring Boot

## Selected Publications

For complete list, visit: [RenanSouza.org/publications](https://RenanSouza.org/publications)

- [1] **R. Souza**, T. J. Skluzacek, S. R. Wilkinson, M. Ziatdinov, R. F. 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. [Online]. Available: <https://doi.org/10.1109/e-Science58273.2023.10254822>.
- [2] **R. Souza**, L. G. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E. Vital Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira, M. A. S. Netto, "Workflow provenance in the lifecycle of scientific machine learning," *Concurrency and Computation: Practice and Experience*, vol. e6544, pp. 1–21, 2021. DOI: 10.1002/cpe.6544. [Online]. Available: <https://doi.org/10.1002/cpe.6544>.
- [3] **R. Souza**, V. Silva, A. A. B. Lima, D. Oliveira, P. Valduriez, M. Mattoso, "Distributed in-memory data management for workflow executions," *PeerJ Computer Science*, vol. 7, pp. 1–30, 2021. DOI: 10.7717/peerj-cs.527. [Online]. Available: <https://peerj.com/articles/cs-527/>.
- [4] **R. Souza**, V. Silva, J. J. Camata, A. L. G. A. Coutinho, P. Valduriez, M. Mattoso, "Keeping track of user steering actions in dynamic workflows," *Future Generation Computer Systems*, vol. 99, pp. 624–643, 2019, ISSN: 0167-739X. DOI: 10.1016/j.future.2019.05.011. [Online]. Available: <https://doi.org/10.1016/j.future.2019.05.011>.

## Grants and Awards

- 2nd IBM Patent Plateau (8+ patents submitted to USPTO) 2021
- SBBB Honored Mention for the Best Ph.D. Thesis Award 2021
- 1st IBM Patent Plateau (4+ patents submitted to USPTO) 2020
- SBBB Best M.Sc. Thesis Award 2017
- SBBB Honored Mention on the paper  
*Spark Scalability Analysis in a Scientific Workflow* 2017
- CAPES M.Sc. Grant 2013 – 2014
- Brazil Science Mobility Grant - Missouri State University 2012 – 2013
- Scientific Initiation Grant - Federal Univ. of Rio de Janeiro 2010

## Badges and Certifications

- **Machine Learning Specialist Professional** Course duration: 73h — 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** Course duration: 3.5h — 2022
- **Enterprise Design Thinking Practitioner** 2022
- **LinkedIn Skill Assessment: Python, MySQL, Linux, T-SQL, NoSQL**

## Languages

- **English** - Full proficiency
  - Missouri State University, U.S. Duration: 150h — Jun 2012 – Aug 2012  
Scientific English for Graduate Students
  - Cultura Inglesa (English Culture), Rio de Janeiro, Brazil 2001 – 2009
- **Portuguese** - Native
- **Spanish** - Fluent reading, intermediate speaking and understanding, limited writing