Renan Souza

contact@renansouza.org • RenanSouza.org • renansouza1
ID: x9t36ewAAAAJ • renan-souza • Updated on April 14, 2022

Bio

Renan Francisco Santos Souza holds a Ph.D., M.Sc., and B.Sc (2009–2019) in Computer Science from the Federal University of Rio de Janeiro (UFRJ). Since 2015, he has been at IBM Research, where he is a Research Scientist in the Intelligent Cloud Technologies group. He has been working both as a software engineer and a researcher on several projects since 2010. During his B.Sc., he spent a year at Missouri State University and was an intern at Stanford University in the SLAC National Laboratory. During his Ph.D., he was a visiting researcher at Inria, France. He received the best M.Sc. thesis and an honored mention for the best Ph.D. thesis awards from SBBD, the main conference on data science in Latin America. He researches large-scale data management techniques to support the evolution of Artificial Intelligence systems in the cloud.

Research Interests

Large-scale Data Science and Data Engineering • Parallel Workflows • Data Provenance • Big Data Analytics • High Performance Computing in Clusters and Clouds • Machine Learning

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

- **IBM Research** **Apr 2015 – present**
  Research Scientist, Cloud & AI Data Management **Rio de Janeiro, Brazil**
  As a Research Scientist (since 2021), he leads R&D projects in large-scale data science and data engineering to support Artificial Intelligence systems running on hybrid cloud and cluster environments with highly distributed and heterogeneous applications, data, and users. He develops software and do applied research to solve problems in different industries, such as energy, financial, and cheminformatics.
  As a Research Software Engineer (2015–2021), he participated in several R&D projects with clients in the energy field by developing techniques and systems for large-scale data integration of AI systems running on clusters and clouds. He also led the Cloud DevOps team to develop conversational AI systems.
  As a Software Engineering intern (2015), he designed and implemented big data and machine learning solutions to analyze streaming social data.

- **SLAC National Accelerator Laboratory, Stanford Univ.** **May 2013 – Dec 2014**
  Research Software Engineering intern **Menlo Park, CA**
  Developed a project to build a cloud platform that uses semantic web, big data, and data warehousing techniques to store, retrieve, visualize, and publish structured data about internet performance worldwide, enabling a rich understanding of information about the Internet quality around the world.
- **CAPGov COPPE/UFRJ** **Dec 2011 – Sep 2014**
  Software Engineer **Rio de Janeiro, Brazil**
  As a software engineer (2013–2014), he led the development of a system that helped the Brazilian population to have easy access to information about public services provided by the Federal Government. He also participated in the development of a system to publish linked open data of the Brazilian Federal Register ("Diário Oficial da União") on the semantic web using agile methodology, ontology data modeling, and natural language processing.
  As a Software Engineering intern (2011–2013), he participated in several R&D web systems for the Brazilian Federal Government.
- **Federal Univ. of Rio de Janeiro** **Jan 2010 – Jul 2011**
  Software Engineering intern **Rio de Janeiro, Brazil**
  Developed a system to integrate data warehouse environments with structured and unstructured data to enable more intelligent and flexible information reports.
- **Petrobras** **May 2007 – May 2008**
  IT Intern **Rio de Janeiro, Brazil**
  Helped to implement features and provided maintenance for web systems to support Petrobras employees.

Technical Knowledge

- **Languages**: Python, Java, C, C++, Shell scripting, NodeJS, Scala, Lua
- **Relational DBMS**: PostgreSQL/PostGIS, DB2, DashDB, MySQL, MySQL Cluster, MS SQL Server
- **NoSQL DBMS**: MongoDB, AllegroGraph, Jena, Blazegraph, Virtuoso, Sesame, Cloudant, CouchBase, Redis, Impala, Elasticsearch, HBase, Hive, Apache Ignite
- **Heterogeneous Data Management**: Data Integration, Multi-database Queries, Polystores, Foreign Data Wrappers
- **Big Data Frameworks**: Apache Spark: RDD, DataFrames, Streaming, MLib, GraphX, GraphFrames; Hadoop Ecosystem
- **Message Brokers**: Kafka, RabbitMQ
- **Data Science/ML Technologies**: Pandas, Jupyter Notebooks, Numpy, Matplotlib, Tensorflow, Scikit-Learn, Keras, PyTorch
- **Big Data 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**: MPI, OpenMP, CUDA, Data-centric distributed and parallel programming
- **Cloud and Cluster computing**: VMs, Dockers, Kubernetes, OpenShift, HPC Clusters
- **DevOps**: Containers, Kubernetes, OpenShift, CI/CD Pipelines, GitHub, Travis, Jenkins
- **Web Development**: Python Flask/UWSGI, Java EE, Tomcat/JBoss, Spring Boot

Selected Publications

For complete list, visit: RenanSouza.org/publications

[1] **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/.

[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. [Online]. Available: https://doi.org/10.1002/cpe.6544.

[3] **R. Souza**, L. Azevedo, R. Thiago, E. Soares, M. Nery, M. Netto, E. V. Brazil, R. Cerqueira, P. Valduriez, M. Mattoso, "Efficient runtime capture of multiworkflow data using provenance," in *IEEE International Conference on e-Science (eScience)*, 2019, pp. 1–10. DOI: 10.1109/eScience.2019.00047. [Online]. Available: https://doi.org/10.1109/eScience.2019.00047.

[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
- SBBD Honored Mention for the Best Ph.D. Thesis Award 2021
- 1st IBM Patent Plateau (>4 patents submitted to USPTO) 2020
- SBBD Best M.Sc. Thesis Award 2017
- SBBD 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
- **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