Renan Souza

contact@renansouza.org • RenanSouza.org • renansouza1
ID: x9t36ewAAAAJ • renan-souza • Generated on May 17, 2020

Bio

Renan Francisco Santos Souza holds a Ph.D. (2019) and an M.Sc. (2015) in Computer Science from COPPE/Federal University of Rio de Janeiro (UFRJ), and a B.Sc. in Computer Science from UFRJ (2009-2013). Since 2015, he works at IBM Research Brazil, where he is a Research Scientist in the Industrial Cloud Technologies group. He has been working both as a software engineer and a researcher in several projects since 2010 and has been actively publishing scientific papers in refereed international conferences and journals since 2014. During his B.Sc., he spent a school year at Missouri State University and did a summer internship at Stanford University in the SLAC National Laboratory. During his Ph.D., he was a visiting researcher at Inria/Univ. Montpellier in France in 2019. In 2017, he won the best M.Sc. thesis award from SBBD, the main conference on data management in Latin America. He researches large-scale data science and engineering techniques for the support of Artificial Intelligence systems.

Research Interests

Large-scale Data Science and 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 and Patrick Valduriez
  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
- M.Sc. in Computer Science, Federal Univ. of Rio de Janeiro, Brazil — Jan 2013 – Jul 2015
  Supervised by Marta Mattoso
  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
  Title: Linked Open Data Publication Strategies: An Application in Network Performance Data (in pt)
- Technical Degree in Information Systems, Lemos de Castro — Jan 2005 – Dec 2007

Experience

- IBM Research, Research Scientist — Aug 2019 – present
- IBM Research, Research Engineer — Sep 2015 – Aug 2019
- IBM Research, Software Engineer Intern — Apr 2015 – Sep 2015
- Stanford Univeristy, SLAC, Research Collaborator — Aug 2013 – Dec 2014
- Stanford Univeristy, SLAC, Research Intern — May 2013 – Aug 2013
- CAPGov - Government Technologies, Leading Software Engineer — Dec 2013 – Sep 2014

- CAPGov - Government Technologies, Software Engineer — Sep 2013 – Jan 2014
- CAPGov - Government Technologies, Software Engineer Intern — Jan 2011 – Jun 2012
- Federal Univ. of Rio de Janeiro, Software Engineer Intern — Jan 2010 – Jul 2011

Selected Publications

For complete list, visit: RenanSouza.org/publications

[1] R. Souza, L. 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, "Provenance data in the machine learning lifecycle in computational science and engineering," in *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, pp. 1–10. DOI: 10.1109/WORKS49585.2019.00006. [Online]. Available: https://arxiv.org/pdf/1910.04223.

[2] 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://hal-lirmm.ccsd.cnrs.fr/lirmm-02265932.

[3] 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://hal-lirmm.ccsd.cnrs.fr/lirmm-02127456.

[4] R. Souza, V. Silva, A. L. G. A. Coutinho, P. Valduriez, M. Mattoso, "Data reduction in scientific workflows using provenance monitoring and user steering," *Future Generation Computer Systems*, vol. online, pp. 1–34, 2017, ISSN: 0167-739X. DOI: 10.1016/j.future.2017.11.028. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01679967/document.

Technical Knowledge

- **Languagess** : 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, 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 DataFrames, Jupyter, Numpy, SciPy, Tensorflow, and 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, HPC Clusters
- **DevOps** : Containers, Kubernetes, OpenShift, CI/CD Pipelines, GitHub, Travis, Jenkins
- **Domain-specific knowledge** : Techniques to analyze large amounts of geological data (SEG-Y, Well logs), OpendTect
- **Web Development** : Python Flask/UWSGI, Java EE: JSP, JSF, JPA, Hibernate, Tomcat/JBoss

Languages

- **English** - Full proficiency
  - Missouri State University, U.S. Jun 2012 – Aug 2012 (150h)
    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

Grants & Awards

- SBBD Best M.Sc. Thesis Award 2017
- Honored Mention at SBBD 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

Teaching and Supervisions

Teaching:

- Databases Laboratory, graduate, UFRJ 2017
  Teacher assistant to Prof. Marta Mattoso
- Logics for Computer Science, undergraduate, UFRJ 2012–2013
  Teacher assistant to Prof. Mario Benevides

Supervisions of final dissertations

- Pedro Paiva Miranda, undergraduate, UFRJ, Co-supervision with Prof. Marta Mattoso 2015
  Thesis title: *A Mechanism for Fault Tolerance in Parallel Executions of Workflows supported by a Database*
- Pedro Paiva Miranda, undergraduate, UFRJ, Co-supervision with Prof. Marta Mattoso 2015
  Thesis title: *Publication of Workflows Provenance Data in the Semantic Web*

Talks and Events Participation

- **Open Subsurface Data Universe Development Workshop** in Houston, TX 2020
- **IEEE/ACM Supercomputing (SC)** in Denver, CO 2019
  Workflows in Support of Large-scale Science (WORKS)
  - Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering, Oral presentation
- **SciDISC Workhsop** in Rio de Janeiro, Brazil 2019
  - Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering, Oral presentation
- **Open Subsurface Data Universe F2F Meeting** in Houston, TX 2019
- **IEEE International Conference on e-Science** in San Diego, CA 2019
  - Efficient Runtime Capture of Multiworkflow Data using Provenance, Oral presentation
- **INRIA Talks** in Montpellier, France 2019
  - Providing Online Data Analytical Support for Humans in the Loop of Computational Science and Engineering Applications, Oral presentation
- **IBM Regional Technical Exchange** in Rio de Janeiro, Brazil 2019
- **Provenance Week** in London, UK 2018
  International Provenance and Annotation Workshop (IPAW)

- Provenance of Dynamic Adaptations in User-steered Dataflows, Oral presentation
- Capturing Provenance for Runtime Data Analysis in Computational Science and Engineering Applications, Poster presentation

Computational Reproducibility Workshop
- Provenance of Dynamic Adaptations in User-steered Dataflows, Oral presentation

- **International Conference on Very Large Databases (VLDB)** in Rio de Janeiro, Brazil 2018
Latin American Data Science Workshop
  - Tracking Hyperparameter Tuning in Deep Learning Training, Oral presentation
- **SBC Brazilian Syposium on Databases (SBBD)** in Rio de Janeiro, Brazil 2018
- **SBC Brazilian Syposium on Databases (SBBD)** in Uberlandia, Brazil 2017
  - Spark Scalability Analysis in a Scientific Workflow, Oral presentation
  - Controlling the Parallel Execution of Workflows Relying on a Distributed Database, Oral presentation
- **Federal University of Uberlandia, Brazil** in Uberlandia, Brazil 2017
  - Kubernetes, Invited talk
- **Hacker at the Smart City Cloud Hackathon OpenStack Rio** in Rio de Janeiro, Brazil 2017
- **Computer Science Week at UFRJ** in Rio de Janeiro, Brazil 2017
  - Kubernetes, Oral presentation
- **SBC Brazilian Conference on Artificial Intelligence (BRACIS)** in Recife, Brazil 2017
  - Graph Analytics with Spark, Tutorial , link
- **IEEE/ACM Supercomputing (SC)** in Salt Lake City, UT 2016
Workflows in Support of Large-scale Science (WORKS)
  - Online Input Data Reduction in Scientific Workflows, Oral presentation
- **ASE BigData/SocialCom/CyberSecurity** in Stanford University, Menlo Park, CA 2014
  - Revise name, poster presentation

All Publications and Patents

Journal Articles

[J1] 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://hal-lirmm.ccsd.cnrs.fr/lirmm-02127456.

[J2] V. Silva, L. Neves, R. Souza, A. L. G. A. Coutinho, D. Oliveira, M. Mattoso, "Adding domain data to code profiling tools to debug workflow parallel execution," *Future Generation Computer Systems*, pp. 624–643, 2018, ISSN: 0167-739X. DOI: 10.1016/j.future.2018.05.078. [Online]. Available: https://doi.org/10.1016/j.future.2018.05.078.

[J3] M. G. Bayser, P. Cavalin, R. Souza, A. Braz, H. Candello, C. Pinhanez, J.-P. Briot, "A hybrid architecture for multi-party conversational systems," *arXiv preprint Computation and Language (cs.CL)*, pp. 1–40, 2017. DOI: arXiv:1705.01214. [Online]. Available: https://arxiv.org/abs/1705.01214.

[J4] R. Souza, V. Silva, A. L. G. A. Coutinho, P. Valduriez, M. Mattoso, "Data reduction in scientific workflows using provenance monitoring and user steering," *Future Generation Computer Systems*, vol. online, pp. 1–34, 2017, ISSN: 0167-739X. DOI: 10.1016/j.future.2017.11.028. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01679967/document.

Conference and Workshop Papers

[C1] L. Azevedo, R. Souza, E. Soares, M. Moreno, "Modern federated databases: An overview," in *International Conference on Enterprise Information Systems (ICEIS)*, 2020.

[C2] L. Azevedo, R. Souza, R. Thiago, E. Soares, M. Moreno, "Experiencing provlake to manage the data lineage of ai workflows," in *Meeting in Innovation in Information Systems (EISI) in Brazilian Symposium in Information Systems (SBSI)*, 2020.

[C3] R. Souza, J. Camata, M. Mattoso, A. Coutinho, "Runtime steering of parallel cfd simulations," in *International Conference on Parallel Computational Fluid Dynamics*, 2020.

[C4] R. Souza, A. Codas, J. A. Nogueira Junior, M. P. Quinones, L. Azevedo, R. Thiago, E. Soares, M. Cardoso, L. Martins, "Supporting the training of physics informed neural networks for seismic inversion using provenance," in *American Association of Petroleum Geologists Annual Convention and Exhibition (AAPG)*, 2020.

[C5] R. Thiago, R. Souza, L. Azevedo, E. Soares, R. Santos, W. Santos, M. De Bayser, M. Cardoso, M. Moreno, R. Cerqueira, "Managing data lineage of O&G machine learning models: The sweet spot for shale use case," in *European Association of Geoscientists and Engineers (EAGE) Digitalization Conference and Exhibition*, 2020.

[C6] R. Souza, L. 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, "Provenance data in the machine learning lifecycle in computational science and engineering," in *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, pp. 1–10. DOI: 10.1109/WORKS49585.2019.00006. [Online]. Available: https://arxiv.org/pdf/1910.04223.

[C7] 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://hal-lirmm.ccsd.cnrs.fr/lirmm-02265932.

[C8] R. Souza, E. V. Brazil, L. Azevedo, D. Ferreira, E. Soares, R. Thiago, M. Nery, V. Torres, R. Cerqueira, "Managing data traceability in the data lifecycle for deep learning applied to seismic data," in *American Association of Petroleum Geologists Annual Convention and Exhibition (AAPG)*, 2019. [Online]. Available: www.searchanddiscovery.com/abstracts/html/2019/ace2019/abstracts/1718.html.

[C9] M. G. Bayser, C. Pinhanez, H. Candello, M. Affonso, M. P. Vasconcelos, M. A. Guerra, P. Cavalin, R. Souza, "Ravel: A mas orchestration platform for human-chatbots conversations," in *The 6th International Workshop on Engineering Multi-Agent Systems (EMAS@AAMAS 2018), Stockholm, Sweden*, 2018.

[C10] V. Silva, R. Souza, J. Camata, D. Oliveira, P. Valduriez, A. L. G. A. Coutinho, M. Mattoso, "Capturing provenance for runtime data analysis in computational science and engineering applications," in *Provenance and Annotation of Data and Processes*, ser. Lecture Notes in Computer Science (LNCS), Springer International Publishing, 2018, pp. 183–187, ISBN: 978-3-319-98379-0. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-98379-0_15.

[C11] R. Souza and M. Mattoso, "Provenance of dynamic adaptations in user-steered dataflows," in *Provenance and Annotation of Data and Processes*, ser. Lecture Notes in Computer Science (LNCS), Springer International Publishing, 2018, pp. 16–29, ISBN: 978-3-319-98379-0. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-98379-0_2.

[C12] R. Souza, L. Neves, L. Azeredo, R. Luiz, E. Tady, P. Cavalin, M. Mattoso, "Towards a human-in-the-loop library for tracking hyperparameter tuning in deep learning development," in *Latin American Data Science (LaDaS) workshop co-located with the Very Large Database (VLDB) conference*, Rio de Janeiro, Brazil, 2018, pp. 84–87. [Online]. Available: http://ceur-ws.org/Vol-2170/paper12.pdf.

[C13] R. Souza, V. Silva, J. Camata, A. Coutinho, P. Valduriez, M. Mattoso, "Tracking of online parameter fine-tuning in scientific workflows," in *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)*, Denver, CO, 2017. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620974.

[C14] R. Souza, V. Silva, P. Miranda, A. A. B. Lima, P. Valduriez, M. Mattoso, "Spark scalability analysis in a scientific workflow," in *Simpósio Brasileiro de Banco de Dados (SBBD)*, 2017, pp. 288–293. [Online]. Available: http://sbbd.org.br/2017/wp-content/uploads/sites/3/2018/02/p288-293.pdf.

[C15] T. Barbosa, R. Souza, S. Cruz, M. Campos, R. L. Cottrell, "Applying data warehousing and big data techniques to analyze internet performance," SLAC National Accelerator Lab., Menlo Park, CA (United States), Tech. Rep., 2016.

[C16] P. Cavalin, F. Figueiredo, M. Bayser, L. Moyano, H. Candello, A. Appel, R. Souza, "Building a question-answering corpus using social media and news articles," in *International Conference on Computational Processing of the Portuguese Language*, 2016, pp. 353–358.

[C17] V. Silva, L. Neves, R. Souza, A. Coutinho, D. D. Oliveira, M. Mattoso, "Integrating domain-data steering with code-profiling tools to debug data-intensive workflows," in *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)*, Salt Lake City, USA, 2016.

[C18] R. Souza, V. Silva, A. Coutinho, P. Valduriez, M. Mattoso, "Online input data reduction in scientific workflows," in *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, pp. 1–10. [Online]. Available: https://hal.archives-ouvertes.fr/lirmm-01400538.

[C19] R. Castro, R. Souza, V. Silva, K. Ocaña, D. Oliveira, M. Mattoso, "Uma abordagem para publicação de dados de proveniência de workflows científicos na web semântica," in *Simpósio Brasileiro de Banco de Dados (SBBD)*, 2015.

[C20] R. Souza, V. Silva, D. Oliveira, P. Valduriez, A. A. B. Lima, M. Mattoso, "Parallel execution of workflows driven by a distributed database management system," in *ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, Salt Lake City, USA, 2015, pp. 1–3. [Online]. Available: http://sc15.supercomputing.org/sites/all/themes/SC15images/tech_poster/tech_poster_pages/post284.html.

[C21] R. Souza, L. Cottrell, B. White, M. L. Campos, M. Mattoso, "Linked open data publication strategies: Application in networking performance measurement data," in *ASE BigData/SocialCom/CyberSecurity, Stanford, CA*, 2014.

Patents

[P1] A. Braz, P. R. Cavalin, F. Figueiredo, M. G. De Bayser, R. Souza, *System and method for managing artificial conversational entities enhanced by social knowledge*, Granted, US Patent Application 15/265,615, 2018.

[P2] M. G. De Bayser, A. Braz, P. R. Cavalin, F. Figueiredo, R. Souza, *Creating coordinated multi-chatbots using natural dialogues by means of knowledge base*, Granted, US Patent Application 15/217,660, 2018.

[P3] A. P. Appel, A. Gama Leal, R. Souza, *Predicting user question in question and answer system*, Granted, US Patent Application 15/171,055, 2017.