Raul Castro Fernandez

Office 32-G930, Stata Center
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology (MIT)
Cambridge, MA, USA
raulcf@csail.mit.edu

31/Oct/2018

SUMMARY

In my research, I build high-performance and scalable systems to discover, prepare and process data.

PUBLICATIONS

- Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Sam Madden, Michael Stonebraker, **Aurum: A Data Discovery System**, 34th IEEE International Conference on Data Engineering, (**ICDE**), Paris, France, 2018
- Raul Castro Fernandez, Essam Mansour, Abdulhakim Qahtan, Ahmed Elmagarmid, Ihab Ilyas, Sam Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang, **Seeping Semantics: Linking Datasets using Word Embeddings for Data Discovery**, 34th IEEE International Conference on Data Engineering, (**ICDE**), Paris, France, 2018
- Raul Castro Fernandez, William Culhane, Pijika Watcharapichat, Matthias Weidlich, Victoria Lopez Morales, Peter Pietzuch, **Meta-Dataflows: Efficient Exploratory Dataflow Jobs**, ACM International Conference on Management of Data (**SIGMOD**), Houston, (TX), 2018
- Andrew Ilyas, Joana M. F. da Trindade, Raul Castro Fernandez, Sam Madden, **Extracting Syntactical Patterns from Databases**, 34th IEEE International Conference on Data Engineering (**ICDE**), Paris, France, 2018
- Mourad Ouzzani, Nan Tang, Ahmed Elmagarmid, Raul Castro Fernandez, Abdulhakim A. Qahtan, **FAHES: A Robust Disguised Missing Values Detector**. 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (**KDD**), London, 2018
- Badrish Chandramouli, Raul Castro Fernandez, Jonathan Goldstein, Ahmed Eldawy, Abdul Quamar, **Quill: Efficient, Transferable, and rich analytics at scale**, 43rd International Conference on Very Large DataBases (**VLDB**), Munich, Germany, 2017
- Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed Elmagarmid, Ihab Ilyas, Samuel Madden, Mourad Ouzzani, Nan Tang, **The Data Civilizer System**, 8th Biennial Conference on Innovative Data Systems Research (**CIDR**), Asilomar (CA), 2017
- Pijika Watcharapichat, Victoria Lopez Morales, Raul Castro Fernandez, Peter Pietzuch, **Ako: Decentralized Deep Learning With Partial Gradient Descent**, ACM Symposium on Cloud Computing (**SOCC**), Santa Clara, (CA), 2016
- Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, Nan Tang, **Detecting Data Errors: Where are we and what needs to be done?**, 42nd International Conference on Very Large DataBases (**VLDB**), New Delhi, India, 2016
- Alexandros Koliousis, Matthias Weidlich, Raul Castro Fernandez, Paolo Costa, Alexander Wolf, Peter Pietzuch, **SABER: Window-Based Hybrid Stream Processing for Heterogeneous Architectures**, ACM International Conference on Management of Data (**SIGMOD**), San Francisco, (CA), 2016

- Raul Castro Fernandez, Peter Pietzuch, Jay Kreps, Neha Narkhede, Jun Rao, Joel Koshy, Dong Lin, Chris Riccomini, Guozhang Wang, **"Liquid: Unifying Nearline and Offline Big Data Integration"**, 7th Biennial Conference on Innovative Data Systems Research (**CIDR**), Asilomar (CA), 2015
- Raul Castro Fernandez, Matteo Migliavacca, Evangelia Kalyvianaki and Peter Pietzuch, **"Making State Explicit for Imperative Big Data Processing"**, USENIX Annual Technical Conference (**USENIX ATC**), Philadelphia (PA), 2014
- Raul Castro Fernandez, Matteo Migliavacca, Evangelia Kalyvianaki and Peter Pietzuch, **"Integrating Scale Out and Fault Tolerance in Stream Processing using Operator State Management"**, ACM International Conference on Management of Data (**SIGMOD**), New York (NY), 2013
- M. Garcia-Valls and Raul Castro Fernandez, "Timely service composition in service-based distributed real-time systems", International Conference on Embedded Software and Systems (ICESS), Liverpool, 2012
- M. Garcia-Valls, Raul Castro Fernandez and Iria Estevez Ayres, "iLAND Deliverable D3.4: Deterministic System-Level Reconfiguration Strategies and Composition Algorithms for Bounded-Time and Bounded-Error", 2012
- Raul Castro Fernandez and Javier Carbo. "APoDDS: A DDS-Based Approach to Promote Multi-Agent Systems in Distributed Environments." Distributed Computing and Artificial Intelligence, Springer Berlin/Heidelberg, 2010

MANUSCRIPTS AND WORK IN PROGRESS

- Raul Castro Fernandez, Sam Madden, **The Fabric of Data: Learning a Relational Embedding for Database Exploration**, *Work in Progress*
- Raul Castro Fernandez, Sam Madden, **The Fabric of Data: Filling Values by Posing the Right Questions with an OpenQA System**, *Work in Progress*
- Raul Castro Fernandez, Sam Madden, **Termite: A System for Tunneling Through Heterogeneous Data**, *Work in Progress*
- Raul Castro Fernandez, Jisoo Min, Demitri Nava, Sam Madden, **Lazo: A Cardinality-Based Method for Coupled Estimation of Jaccard Similarity and Containment**, *Under Submission*
- Mohammad Mahdavi, Ziawash Abedjan, Raul Castro Fernandez, Sam Madden, Nan Tan, Mourad Ouzzani, Michael Stonebraker, **Raha: A Configuration-Free Error Detection System**, *Under Submission*

BOOK CHAPTERS

- Mourad Ouzzani, Nan Tang, Raul Castro Fernandez, **Data Civilizer: How to make your data great again.** Chapter contribution in *Making Databases Work: The Works of Michael Stonebraker*, ACM Books Turing Series, Michael L. Brodie (editor), in preparation for publication in 2019.
- Raul Castro Fernandez, **Aurum: A Story about Research Taste**. Chapter contribution in *Making Databases Work: The Works of Michael Stonebraker*, ACM Books Turing Series, Michael L. Brodie (editor), in preparation for publication in 2019.

WORKSHOP, DEMO, OTHER

- Essam Mansour, Dong Deng, Raul Castro Fernandez, Abdulhakim Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab Ilyas, Sam Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang, **Building Data Civilizer Pipelines with an Advanced Workflow Engine**, 34th IEEE International Conference on Data Engineering, (**ICDE**), Paris, France, 2018
- Raul Castro Fernandez, Dong Deng, Essam Mansour, Abdulhakim Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab Ilyas, Sam Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang, **A Demo of the Data Civilizer System**, ACM International Conference on Management of Data (**SIGMOD**), Chicago, (IL), 2017
- Raul Castro Fernandez, Ziawasch Abedjan, Samuel Madden, Michael Stonebraker, **Towards Large-Scale Data Discovery**, 3rd International Workshop on Exploratory Search in Databases and the Web, (**ExploreDB@SIGMOD**), San Francisco, (CA), 2016
- Alexandros Koliousis, Matthias Weidlich, Raul Castro Fernandez, Alexander L. Wolf, Paolo Costa, and Peter Pietzuch, **The SABER System for Window-Based Hybrid Stream Processing with GPGPUs**, 10th ACM International Conference on Distributed and Event-Based Systems (**DEBS**), Irvine, CA, 2016
- Raul Castro Fernandez, Panagiotis Garefalakis, Peter Pietzuch, **Java2SDG: Stateful Big Data Processing for the Masses (Demo paper)**, 32nd IEEE International Conference on Data Engineering (**ICDE**), Helsinki, Finland, 2016
- Raul Castro Fernandez, Matthias Weidlich, Peter Pietzuch and Avigdor Gal, Grand Challenge: Scalable Stateful Stream Processing for Smart Grids, ACM Conference on Distributed Event Based Systems (DEBS), Mumbai, India, 2014
- Raul Castro Fernandez, Peter Pietzuch, "Towards Low-Latency and In-Memory Large Scale Data Processing", ACM International Conference on Distributed Event-Based Systems (DEBS), Arlington (TX), 2013

PROFESSIONAL EXPERIENCE

Massachusetts Institute of Technology (MIT) October 2015 – Ongoing
Postdoctoral Associate

At MIT I work on data discovery with Samuel Madden and Michael Stonebraker. With large amounts of heterogeneous data spread across many databases, data lakes and the cloud, data scientists spend more time finding relevant data for the task at hand than analyzing it. In this work, we are building systems to help data analysts find data.

Microsoft Research July 2015 – September 2015
Research Intern

At Microsoft I worked with Badrish Chandramouli and Jonathan Goldstein on designing a new distributed data processing system that follows an innovative design adapted to modern clouds. We implemented a prototype and evaluated it in the high-performance Azure cloud. The system was later deployed internally at Microsoft.

LinkedIn June 2014 – August 2014
Software Engineering Intern

At LinkedIn I worked in the Distributed Data Systems group, with Neha Narkhede, Jay Kreps, and Jun Rao from the Apache Kafka team, and Chris Riccomini from Apache Samza. Kafka is a highly available

messaging system, and Samza, built on top, is a stream processing system. Together, these two technologies are central to the backend infrastructure of LinkedIn, handling large amounts of data every day.

My work there focused on providing *exactly once semantics* to both the Kafka and Samza layers. To do this, I collaborated in the design of a transactional mechanism that I implemented while collaborating with a team of engineers. The ultimate goal of this project was to improve the messaging guarantees of these systems in order to enable new use cases both within and outside the company.

Ecana Sistemas de Informacion SL July 2011 – January 2013
Founder and CTO

- Designed the core technology, a data acquisition and processing engine on top of Wireless Sensor Networks to compute key performance indicators for customers by means of a dashboard that shows interactive visualizations
- Designed and planned technology projects with special emphasis on the design and implementation of MVP (Minimum Viable Products) for startups
- European research grant proposal (FP7) writing and planning

University Carlos III of Madrid February 2010 – April 2011
Research assistant

- iLAND European project (ARTEMIS-JU FP7), http://www.iland-artemis.org/
- Worked with communication middleware technologies (e.g. DDS, ICE, CORBA) and communication technologies (e.g. WS* or REST)
- Developed a time-deterministic service-oriented architecture middleware
- Developed core services of the iLAND middleware, such as service composition algorithms and reconfiguration processes
- Produced deliverables and research papers

EDUCATION

PhD in Computer Science
Imperial College London
Supervisor: Dr. Peter Pietzuch
Funding: CASE PhD Award from BAE Systems October 2011 – May 2015

My PhD research was broadly motivated by two fundamental challenges in data management:

First, we need natural programming models that more people—not only skilled engineers—know how to use. For example, there is great value in bridging the programming skills gap between domain-specific scientists—that are used to mainstream programming languages—and the platform-specific programming models of current Big Data systems, which are designed to facilitate parallelism and achieve fault-tolerance.

Second, it is crucial to use resources efficiently, as data grows faster than compute power and new applications keep imposing new and challenging requirements on data analysis. This requires new data management techniques and systems to implement them efficiently.

There is, however, a tension between these two challenges: common programming models hinder performance because they make it more difficult to extract data- and pipeline-parallelism than in the purpose-built, constrained programming models of current systems.

To address both challenges, I have proposed a new data-parallel processing model that, unlike current stateless dataflow graph models that represent only data and computation, contains explicit state in the

dataflow. This stateful processing model: (i) allows the use of state in the programming model, e.g. as found in Matlab, Java or C++ (**USENIX ATC14**) and; (ii) includes a set of state management techniques that allows the system to dynamically partition, checkpoint, backup and recover state (**SIGMOD13**). This facilitates high-performance and low-latency processing by exploiting the data- and pipeline-parallelism in distributed clusters.

Master of Science, Telematics engineering
University Carlos III of Madrid October 2010 – July 2011
MSc Thesis - *Bounded time composition algorithm for SOA-oriented middleware*

Course, Entrepreneurship
EOI Escuela Organizacion Industrial February 2011

Master of Science, Computer Science and Technology (AI and Distributed Systems major)
University Carlos III of Madrid October 2009 – November 2010
MSc Thesis - *APoDDS: A DDS-Based Approach to Promote Multi-Agent Systems in Distr. Environments*

Bachelor of Science, Telematics Engineering
University Carlos III of Madrid September 2005 – September 2009
Bachelor Thesis - *Development of video-surveillance software for distributed embedded systems with ICE*

TEACHING AND MENTORING

Teaching activities:

- 6.830/6.814 Database Systems. This is the undergraduate and graduate class on databases at MIT, taken by around 70 students. Co-teaching with professor Tim Kraska. Fall 2018
- Guest Lecture: "Data Parallel Processing Architectures". Invited lecture in the database systems class taught by professor Samuel Madden at MIT. Fall 2016

Mentoring while doing my Postdoc at MIT (UROP stands for Undergraduate Research Opportunities Program and it is a program at MIT in which undergraduate students collaborate with faculty and postdocs on research projects):

- Andrew Ilyas (UROP). Title: *Automating metadata discovery.*
- Gina Yuan (UROP). Title: *Annotation Framework for data discovery*
- Gina Yuan (UROP). Title: *Improving the performance of data profiling*
- Famien Koko (UROP). Title: *Distributed Master-Worker Data Profiling*
- Demitri Nava (UROP). Title: *Comparison of One-Permutation Hash Sketches*
- Jisoo Min (UROP). Title: *Improving quality of Jaccard Similarity estimation with sketches*
- Kevin Fang (UROP). Title: *On building easy-to-use discovery systems*

Mentoring while doing my PhD at Imperial College:

- Ian O'Neill (Undegraduate project). Title: *Merging Streaming and Historical Data.*
- Martin Rouaux (MSc project). Title: *Framework for Dynamic Scaling in Master-Slave Stream Processing Systems.*
- Andrei Antonescu (MSc project). Title: *Cluster resource management in a multi-tenant scenario with low-latency and batch jobs.*

- Irina-Elena Veliche (MSc project). Title: *Scaling Natural Language Processing for High-Throughput Question Answering.*
- Group Project (Undergraduate group project). Title: *HDFS integration in the SEEP platform.*

INDEPENDENT PROJECTS

contxt.in July 2012 – July 2014

contxt aggregates, summarizes and presents conveniently to its users news coming from multiple sources. In today's information rich world, users are suffering of information overload: too much repetitive and overlapping information which leaves little time to think. contxt aims to ameliorate this problem. It uses a pipeline that scrapes content (with publisher permission) from the news pages. Then it cleans it, it formats it into a format amenable to the backend and then it uses NLP techniques to find out linkages between different news sources. Finally, it chooses among a few different visualizations to present the information to users through a mobile app.

HONORS AND AWARDS

- SIGMOD 2013 Travel Award.
- CASE PhD award from EPSRC/BAE Systems.
- Best academic record of students finishing their engineering studies on telecommunications engineering: telematics (3-year) between 1 January and 30 October 2009 (Year 2008-2009).
- Two times awardee of MEC (Spanish government) grants during bachelor (Covering full tuition costs).
- Chosen student representative for MSc in Computer Science and Technology (Year 2009-2010).

SERVICE

- Program Committee SIGMOD'19
- Reviewer for VLDBJ
- Reviewer for TODS (Transactions on Database Systems)
- Reviewer for TKDE (Transactions on Knowledge and Data Engineering)
- Reviewer for O'Reilly
- Eurosys 2012 (Shadow PC)

REFERENCES

- Samuel Madden, (madden@csail.mit.edu), CSAIL, MIT
- Peter Pietzuch, (prp@doc.ic.ac.uk), Department of Computing, Imperial College London
- Michael Stonebraker, (stonebraker@csail.mit.edu), CSAIL, MIT
- Badrish Chandramouli, (badrishc@microsoft.com), Microsoft Research
- Tim Kraska, (kraska@csail.mit.edu), CSAIL, MIT