The Berkeley Data Analytics Stack: Present and Future

Mike Franklin
UC Berkeley, USA

Abstract

The Berkeley AMPLab was founded on the idea that the challenges of emerging Big Data applications requires a new approach to analytics systems. Launching in early 2011, the project set out to rethink the traditional analytics stack, breaking down technical and intellectual barriers that had arisen during decades of evolutionary development. The vision of the lab is to seamlessly integrate the three main resources available for making sense of data at scale: Algorithms (such as machine learning and statistical techniques), Machines (in the form of scalable clusters and elastic cloud computing), and People (both individually as analysts and en masse, as with crowdsourced human computation). To pursue this goal, we assembled a research team with diverse interests across computer science, forged relationships with domain experts on campus and elsewhere, and obtained the support of leading industry partners and major government sponsors. The lab is realizing its ideas through the development of a freely-available Open Source software stack called BDAS: the Berkeley Data Analytics Stack. In the nearly three years the lab has been in operation, we've released major components of BDAS. Several of these components have gained significant traction in industry and elsewhere: the Mesos cluster resource manager, the Spark in-memory computation framework, and the Shark query processing system. BDAS shows up prominently in many industry discussions of the future of the Big Data analytics ecosystem - a rare degree of impact for an ongoing academic project. Given this initial success, the lab is continuing on its research path, moving "up the stack" to better integrate and support deep machine learning and to make people a full-fledged resource for making sense of Big Data.

In this talk, I'll first outline the motivation and insights behind our research approach and describe how we have organized to address the cross-disciplinary nature of Big Data challenges. I will then describe the current state of BDAS with an emphasis on the key components listed above and will address our current efforts on machine learning scalability and ease of use, and hybrid human/computer processing. Finally I will present our current views of how all the pieces will fit together to form a system that can adaptively bring the right resources to bear on a given data-driven question to meet time, cost and quality requirements throughout the analytics lifecycle.

Biography

Michael Franklin is the Thomas M. Siebel Professor of Computer Science at UC Berkeley, where he also serves as Director of the Algorithms, Machines and People Lab (AMPLab). The Berkeley AMPLab is a collaboration of over 60 researchers supported by Founding Sponsors Amazon Web Services, Google, and SAP, along with 17 other leading companies, the Darpa XData program, and an NSF Expeditons in Computing award. The latter was announced as part of the Obama Administration's Big Data research initiative in 2012. His research interests include large-scale data management and analytics, data integration, and hybrid human/computer data processing systems. He was founder and CTO of Truvisio, a real-time data analytics company acquired by Cisco Systems in 2012. He is an ACM Fellow and two-time winner of the ACM SIGMOD Test of Time Award (2013 and 2004). He also recently received the Best Paper awards at ICDE 2013 and NSDI 2012, a "Best of VLDB 2012" selection, Best Demo awards at SIGMOD 2012 and VLDB 2011 and the Outstanding Advisor Award from the Computer Science Graduate Student Association at Berkeley. He is a committee member on the U.S. National Academy of Sciences study on Analysis of Massive Data and a Transportation Research Board committee on long-
term data stewardship. Prof. Franklin received his Ph.D. in Computer Science from the University of Wisconsin-Madison in 1993.