Making Sense of Big Data with the Berkeley Data Analytics Stack

Michael Franklin
UC Berkeley
franklin@cs.berkeley.edu

ABSTRACT
The Berkeley AMPLab is creating a new approach to data analytics. Launching in early 2011, the lab aims to seamlessly integrate the three main resources available for making sense of data at scale: Algorithms (machine learning and statistical techniques), Machines (in the form of scalable clusters and elastic cloud computing), and People (both individually as analysts and in crowds). 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 four 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 features 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 advanced analytics and to make people a full-fledged resource for making sense of 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 our newest efforts, including some or all of: the GraphX graph processing system, the Velox and MLBase machine learning platforms, and the SampleClean framework for hybrid human/computer data cleaning. 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.

Categories and Subject Descriptors
E.0 [General]

Keyword
Big Data

WSDM'15, February 2–6, 2015, Shanghai, China.
ACM 978-1-4503-3317-7/15/02.
http://dx.doi.org/10.1145/2684822.2685326

BIO

Michael Franklin is the Thomas M. Siebel Professor of Computer Science and Chair of the Computer Science Division at the University of California, Berkeley. He has over 30 years of experience in the database, data analytics, and data management fields as a researcher, lab director, faculty member, entrepreneur, and software developer. Prof. Franklin is also the Director of the Algorithms, Machines, and People Laboratory (AMPLab) at UC Berkeley. The AMPLab currently works with 23 industrial sponsors including founding sponsors Amazon Web Services, Google, and SAP, and received a National Science Foundation CISE "Expeditions in Computing" award, announced as part of the White House Big Data research initiative in March 2012. AMPLab is well-known for creating a number of popular systems in the Open Source Big Data ecosystem including Spark, Mesos, Shark, GraphX and MLlib, all parts of the Berkeley Data Analytics Stack (BDAS). Prof. Franklin is also a co-PI and Executive Committee member for the Berkeley Institute for Data Science, part of a multi-campus initiative to advance Data Science Environments. He is an ACM Fellow, a two-time winner of the ACM SIGMOD "Test of Time" award, and recipient of the outstanding Advisor Award from the Computer Science Graduate Student Association at Berkeley..