Conquering Big Data with Spark

Prof. Ion Stoica

UC Berkeley, USA

Abstract

Today, big and small organizations alike collect huge amounts of data, and they do so with one goal in mind: extract "value" through sophisticated exploratory analysis, and use it as the basis to make decisions as varied as personalized treatment and ad targeting. To address this challenge, we have developed Berkeley Data Analytics Stack (BDAS), an open source data analytics stack for big data processing.

In this talk I'll focus on the execution engine in BDAS: Apache Spark. Apache Spark is a cluster computing engine that is optimized for in-memory processing, and unifies support for a variety of workloads, including batch, streaming, and iterative computations. Spark is now the most active big data project in the open source community, and is already being used by over one thousand organizations.

Biography

Ion Stoica is a Professor in the EECS Department at University of California at Berkeley. He received his PhD from Carnegie Mellon University, and his B.S. from Polytechnic Institute Bucharest. He does research on cloud computing and networked computer systems. Past work includes the Dynamic Packet State (DPS), Chord DHT, Internet Indirection Infrastructure (i3), declarative networks, replay-debugging, and multi-layer tracing in distributed systems. His current research focuses on resource management and scheduling for data centers, cluster computing frameworks, and network architectures. He is an ACM Fellow and has received numerous awards, including the SIGCOMM Test of Time Award (2011), and the ACM doctoral dissertation award (2001). In 2006, he co-founded Conviva, a startup to commercialize technologies for large scale video distribution, and in 2013, he co-founded Databricks a startup to commercialize technologies for Big Data processing.