KDD 2016 Tutorials

Introduction to Spark 2.0

Michael Armbrust / Databricks

Doug Bateman / Databricks

Reynold Xin / Databricks

Matei Zaharia / Databricks

Tutorials

Originally started as an academic research project at UC Berkeley, Apache Spark is one of the most popular open source projects for big data analytics. Over 1000 volunteers have contributed code to the project; it is supported by virtually every commercial vendor; many universities are now offering courses on Spark.

Spark has evolved significantly since the 2010 research paper: its foundational APIs are becoming more relational and structural with the introduction of the Catalyst relational optimizer, and its execution engine is developing quickly to adopt the latest research advances in database systems such as whole-stage code generation.

This tutorial is designed for academic researchers (graduate students, faculty members, and industrial researchers) interested in a brief hands-on overview of Spark. This tutorial covers the core APIs for using Spark 2.0, including DataFrames, Datasets, SQL, streaming and machine learning pipelines. Each topic includes slide and lecture content along with hands-on use of a Spark cluster through a web-based notebook environment. In addition, we will dive into the engine internals to discuss architectural design choices and their implications in practice. We will guide the audience to ``hack'' Spark by extending its query optimizer to speed up distributed join execution.