Addressing Challenges in Data Science: Scale, Skill Sets and Complexity

Joseph Bradley
Databricks, Inc.
joseph@databricks.com

ABSTRACT

Data science in modern applications is pushing the limits of tools and organizations. The scale of data, the breadth of required skill sets, and the complexity of workflows all cause organizations to stumble when developing data-powered applications and moving them to production. This talk will discuss these challenges and Databricks' efforts to overcome them within open source software projects like Apache Spark and MLflow.

Apache Spark has simplified large-scale ETL and analytics, and its Project Hydrogen helps to bridge the gap between Spark and ML tools such as TensorFlow and Horovod. MLflow, an open source platform for managing ML lifecycles, facilitates experimentation, reproducibility and deployment. We will present insights from our collaborations on these projects, as well as our perspective at Databricks in facilitating data science for a wide variety of organizations and applications.

BIOGRAPHY

Joseph Bradley is an Apache Spark PMC member and a Machine Learning Software Engineer at Databricks. At the intersection of Big Data and Machine Learning, his team works on scaling ML, integrating Big Data and ML tools, and simplifying the management of environments and workflows for Data Science. Major projects include Apache Spark MLlib and GraphFrames.

Previously, Joseph was a postdoc at UC Berkeley after receiving his Ph.D. in Machine Learning from Carnegie Mellon. His research interests included optimization techniques for sparse regression, systems for peer grading and rating, and probabilistic graphical models.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

KDD '19, August 4–8, 2019, Anchorage, AK, USA.
© 2019 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6201-6/19/08.
DOI: https://doi.org/10.1145/3292500.3340408