Optimizing Apache Spark

Introductions & Agenda

Introductions & Agenda

The Course

- This course will focus on some of the most significant performance problems associated with developing Spark applications
- We will learn what those problems are
- We will learn how to identify those problems in existing code
- And we will look at options for mitigating those problems

Introductions & Agenda

The 5 Most Common Performance Problems (The 5 Ss)

The "5 Ss" refers to the five most common performance problems that every developer needs to be aware of: Spill, Skew, Shuffle, Storage, and Serialization. By developing a solid understanding of these problems, every developer is better equipped to diagnose and fix various performance problems.

Introductions & Agenda

The Spark UI Simulator

- https://www.databricks.training/spark-ui-simulator
- Preran notebooks
- A full capture of the notebook, cluster and history server's state
- Experiments are tailored to specific topic
- Experiments are 100% reproducible by students
- Always available for future reference