Spark\_R

Bo

2019/10/18

# Mastering Spark with R

The Complete Guide to Large-Scale Analysis and Modeling Javier Luraschi, Kevin Kuo & Edgar Ruiz Foreword by Matei Zaharia

Apache Spark is a distributed computing platform built on extensibility: Spark’s APIs make it easy to combine input from many data sources and process it using diverse programming languages and algorithms to build a data application. R is one of the most powerful languages for data science and statistics, so it makes a lot of sense to connect R to Spark. Fortunately, R’s rich language features enable simple APIs for calling Spark from R that look similar to running R on local data sources. With a bit of background about both systems, you will be able to invoke massive computations in Spark or run your R code in parallel from the comfort of your favorite R program‐ ming environment. This book explores using Spark from R in detail, focusing on the sparklyr package that enables support for dplyr and other packages known to the R community. It covers all of the main use cases in detail, ranging from querying data using the Spark engine to exploratory data analysis, machine learning, parallel execution of R code, and streaming. It also has a self-contained introduction to running Spark and moni‐ toring job execution. The authors are exactly the right people to write about this topic—Javier, Kevin, and Edgar have been involved in sparklyr development since the project started. I was excited to see how well they’ve assembled this clear and focused guide about using Spark with R. I hope that you enjoy this book and use it to scale up your R workloads and connect them to the capabilities of the broader Spark ecosystem. And because all of the infra‐ structure here is open source, don’t hesitate to give the developers feedback about making these tools better. — Matei Zaharia Assistant Professor at Stanford University, Chief Technologist at Databricks, and original creator of Apache Spark