FOREWORD to "Data Algorithms with Spark"

When I started the Apache Spark project a decade ago, one of my main goals was to make it easier for a wide range of users to implement parallel algorithms. New algorithms acting on large-scale data are having a profound impact in all areas of computing, and I wanted to help developers implement new algorithms and reason about their performance without having to build a distributed system from scratch.

I am therefore very excited to see this new book by Dr. Mahmoud Parsian on data algorithms with Spark. Dr. Parsian has extensive research and practical experience with large-scale data-parallel algorithms, including developing new algorithms for bioinformatics as the lead of Illumina's big data team. In this book, he introduces Spark through its Python API, PySpark, and shows how to implement a wide range of useful algorithms efficiently using Spark's distributed computing primitives. He also explains the workings of the underlying Spark engine and how to optimize your algorithms through techniques such as controlling data partitioning. This book will be a great resource for both readers looking to implement existing algorithms in a scalable fashion and readers who are developing new, custom algorithms using Spark.

I am also thrilled that Dr. Parsian has included working code examples for all the algorithms he discusses, using real-world problems where possible. These will serve as a great starting point for readers who want to implement similar computations. Whether you intend to use these algorithms directly or build your own, custom algorithms using Spark, I hope that you enjoy this book as an introduction to the open-source engine, its inner workings, and the modern parallel algorithms that are having such a broad impact across computing.

Matei Zaharia
Assistant Professor of Computer Science, Stanford
Chief Technologist, Databricks
Original Creator of Apache Spark