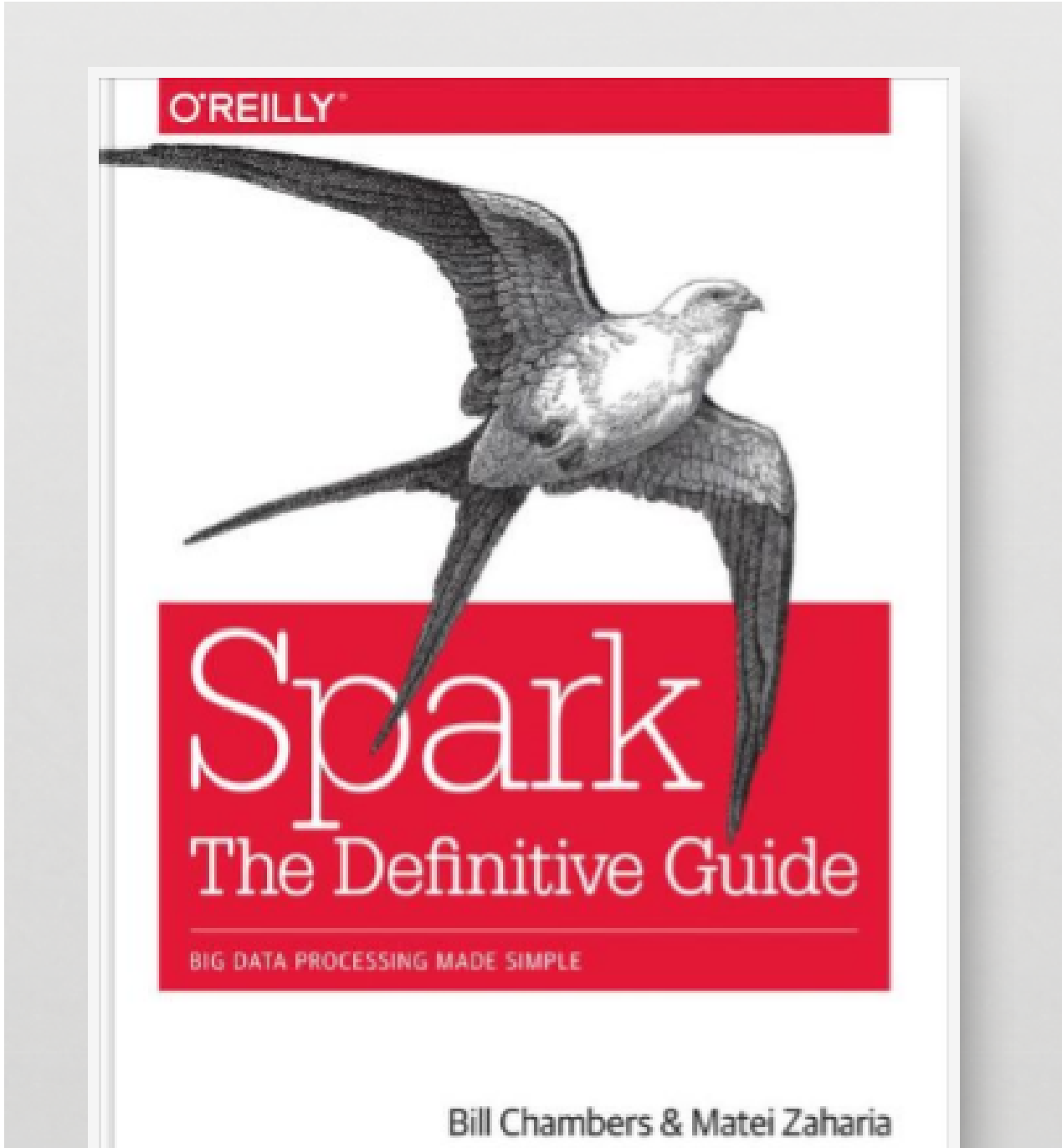
Spark the Definitive Guide 2nd Edition

Chapter 03

A Tour of Spark's Toolset

A Tour of Spark's Toolset

Text Book


O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia

Objectives and Outcomes

- Take a tour of Spark's toolset
- Understand how to run production Spark applications
- Understand type-safe APIs for structured data
- Understand Structured Streaming and Machine Learning
- Understand SparkR and Resilient Distributed DataSets

Review

So far we have:

- learned about core architecture of Spark
  - learned about executors
  - learned about partitions
  - learned about drivers
- learned about datatypes
  - DataFrames
  - APIs
- learned about transformations
- learned about actions
- learned how to put it together from the Spark CLI

Spark Overview

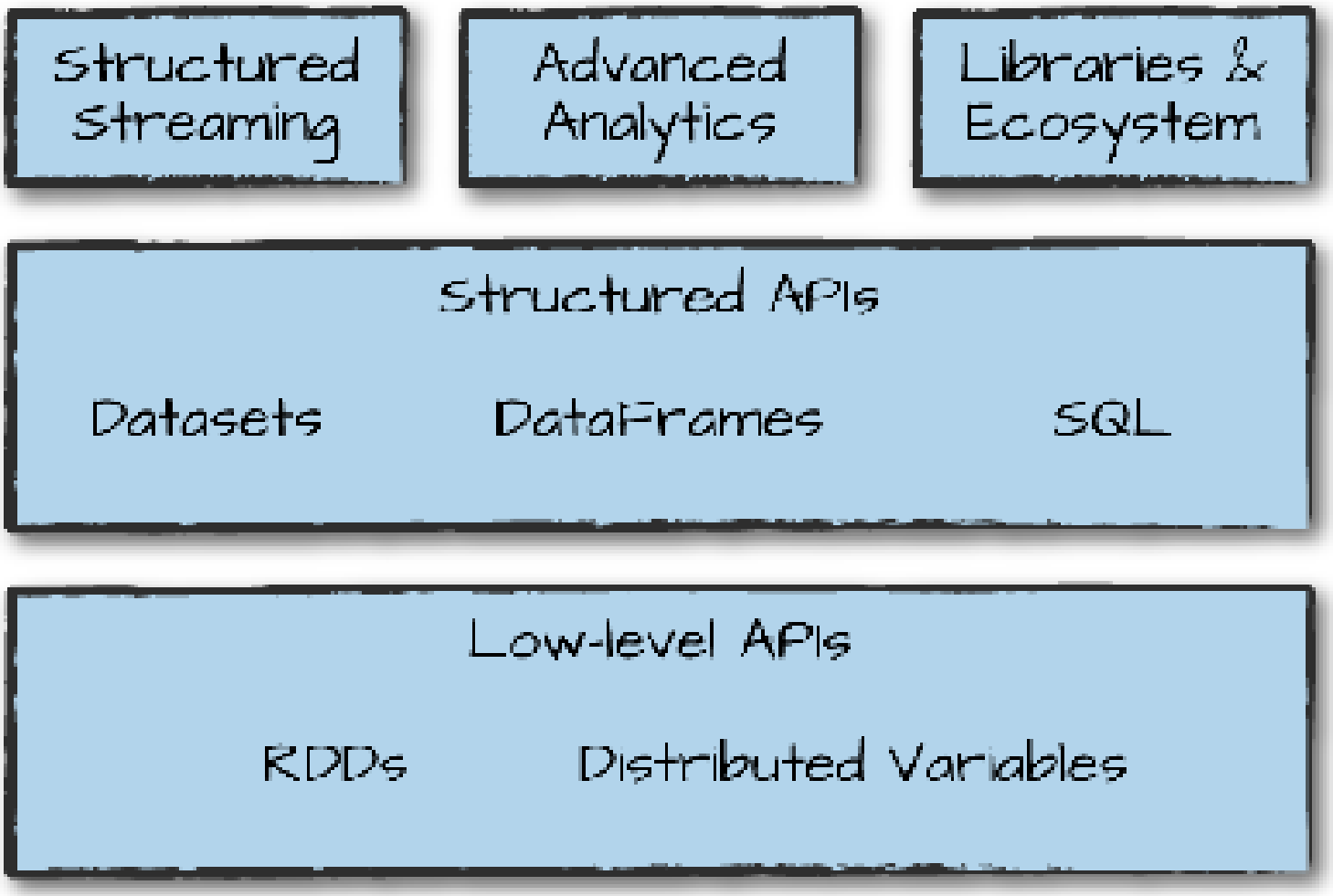


Figure 1-1. Spark's toolkit

Running Production Applications

- `spark-submit`
  - Different from the interactive shell commands we saw in chapter 02
  - `spark-submit` does one thing: send your code to a cluster for execution
  - Application will run until finished or reports and error
- Types of **cluster managers include**:
  - local system (as threads)
  - Mesos
  - YARN

Sample Code

- `spark-submit --class org.apache.spark.examples.SparkPi --master local ./examples/jar/spark-examples_2.11-2.4.4.jar 10`
  - The file name was changes since we are using version 2.4.4 not 2.2.0

Conclusion

- Spark is great