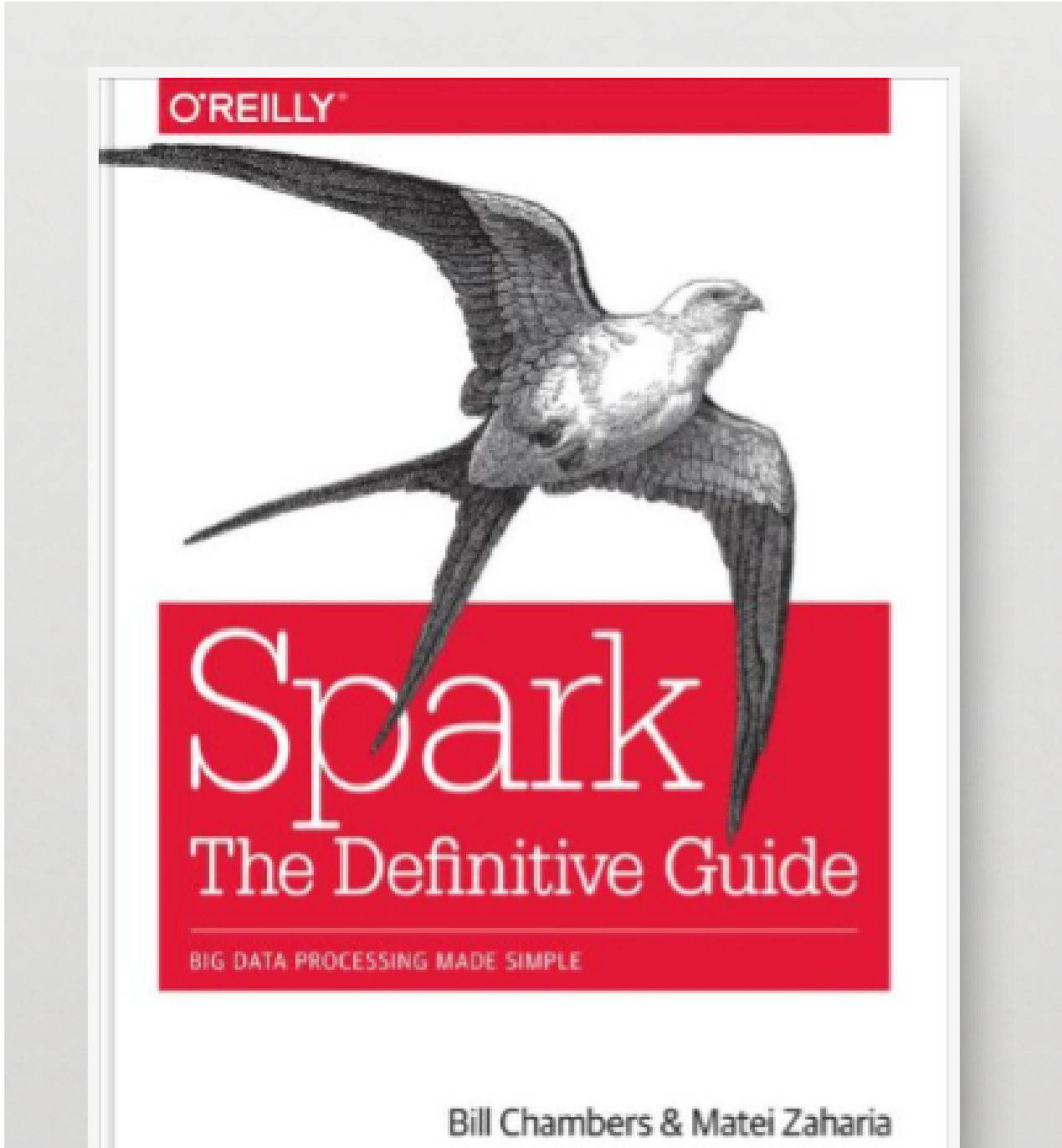
Spark the Definitive Guide 2nd Edition

Chapter 04

Structured API Overview

Structured API Overview

Text Book


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

Objectives and Outcomes

- Introduced to Spark's Structured APIs, DataSets, DataFrames, and SQL Views
- Learn how Spark transforms into a physical execution plan on a cluster

Review

So far:

- We learned about Spark's programming model
- We learned how to run production code
- We were introduced to type-safe data structures in Spark
- We were introduced to Structured Streaming on Spark
- We were introduced to Machine Learning on Spark
- We were introduced to 3rd party Spark packages

API Overview 66

- Three datatypes in Spark:
  - DataFrames
    - Can you define this term?
  - Datasets
    - Can you define this term?
  - SQL Tables and Views
    - Can you define these terms?
- With these data types we can manipulate dispirate types of data
  - Unstructured log files
  - Semi-structured CSV files
  - Structured Parquet files

Structured API concepts

- These concepts refer to both *batch* and *streaming*
  - Code should easily switch between the two
  - We will cover Streaming later in the course, Chapter 20

Structured Collections

- Spark has two notions of structured collections:
  - Datasets and DataFrames
- Each are distributed table-like collections with well defined rows and columns
  - Each row must have the same number of columns
  - Both are **immutable**
  - Both allow for lazily evaluated plans that are only deployed when an **action** is called

Schemas

- A **schema** defines the column names and data types of the column
  - Schemas can be defined manually or inferred
  - Schema on Read
- All of Spark actions take place in the internal Spark language called Catalyst
  - We don't write in this language but the JVM allows us to write in higher level languages that convert to Catalyst

DataFrames vs Datasets

- DataFrames have types of a sort
  - These are maintained by Spark internally
  - Schema only checked at *runtime*
- Datasets are typed DataFrames
  - Only available in Scala and Java
  - Enforce type at compile time
  - P. 54

Conclusion

- We were introduced to Spark's Structured APIs, DataSets, DataFrames, and SQL Views
- We learned how Spark transforms into a physical execution plan on a cluster