So far in the Scala courses...

Focused on:

- **Basics of Functional Programming**. Slowly building up on fundamentals.
- **Parallelism**. Experience with underlying execution in shared memory parallelism.

So far in the Scala courses...

Focused on:

- **Basics of Functional Programming**. Slowly building up on fundamentals.
- **Parallelism**. Experience with underlying execution in shared memory parallelism.

This course:

Not a machine learning or data science course!

- This is a course about distributed data parallelism in Spark.
- Extending familiar functional abstractions like functional lists over large clusters.
- Context: analyzing large data sets.

Why Scala? Why Spark?

Why Scala? Why Spark?

Normally:

Data science and analytics is done *"in the small"*, in R/Python/MATLAB, etc

Why Scala? Why Spark?

Normally:

Data science and analytics is done *"in the small"*, in R/Python/MATLAB, etc

If your dataset ever gets too large to fit into memory,

these languages/frameworks won't allow you to scale. You've to reimplement everything in some other language or system.

Why Scala? Why Spark?

Normally:

Data science and analytics is done *"in the small"*, in R/Python/MATLAB, etc

If your dataset ever gets too large to fit into memory,

these languages/frameworks won't allow you to scale. You've to reimplement everything in some other language or system.

Oh yeah, there's also the massive shift in industry to data-oriented decision making too!

...and many applications are "data science in the large".

Why Scala? Why Spark?

By using a language like Scala, it's easier to scale your small problem to the large with Spark, whose API is almost 1-to-1 with Scala's collections.

That is, by working in Scala, in a functional style, you can quickly scale your problem from one node to tens, hundreds, or even thousands by leveraging Spark, successful and performant large-scale data processing framework which looks a and feels a lot like Scala Collections!

Why Spark?

Spark is...

- **More expressive**. APIs modeled after Scala collections. Look like functional lists! Richer, more composable operations possible than in MapReduce.

Why Spark?

Spark is...

- **More expressive**. APIs modeled after Scala collections. Look like functional lists! Richer, more composable operations possible than in MapReduce.
- **Performant**. Not only performant in terms of running time... But also in terms of developer productivity! Interactive!

Why Spark?

Spark is...

- **More expressive**. APIs modeled after Scala collections. Look like functional lists! Richer, more composable operations possible than in MapReduce.
- **Performant**. Not only performant in terms of running time... But also in terms of developer productivity! Interactive!
- **Good for data science**. Not just because of performance, but because it enables *iteration*, which is required by most algorithms in a data scientist's toolbox.

Also good to know...

Spark and Scala skills are in extremely high demand!

In this course you'll learn...

- **Extending data parallel paradigm to the distributed case, using Spark.**
- **Spark's programming model**
- **Distributing computation, and cluster topology in Spark**
- **How to improve performance; data locality, how to avoid recomputation and shuffles in Spark.**
- **Relational operations with DataFrames and Datasets**

Prerequisites

Builds on the material taught in the previous Scala courses.

- **Principles of Functional Programming in Scala.**
- **Functional Program Design in Scala**
- **Parallel Programming (in Scala)**

Or at minimum, some familiarity with Scala.

Books, Resources

Many excellent books released in the past year or two!

Learning Spark (2015), written by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia

Books, Resources

Many excellent books released in the past year or two!

Spark in Action (2017), written by Petar Zecevic and Marko Bonaci

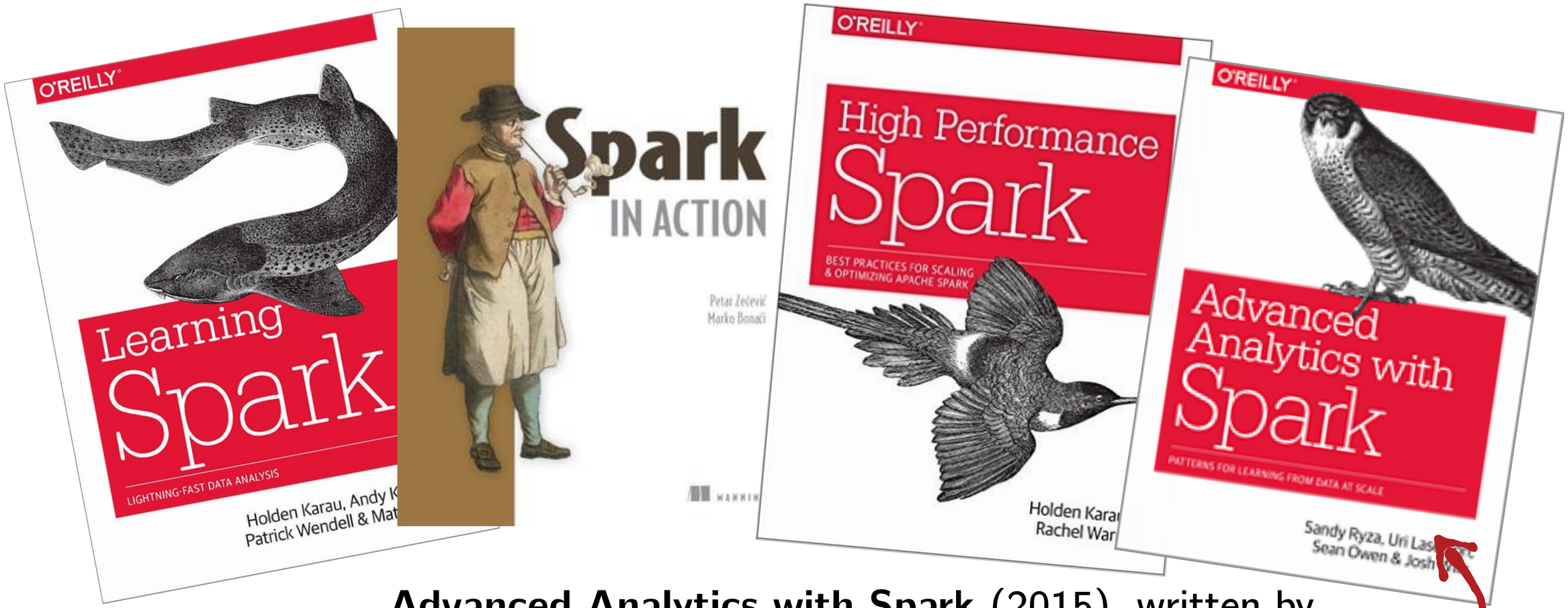
Books, Resources

Many excellent books released in the past year or two!

High Performance Spark (in progress), written by Holden Karau and Rachel Warren

Books, Resources

Many excellent books released in the past year or two!



Advanced Analytics with Spark (2015), written by Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills

Books, Resources

GitBook WE ARE HIRING! Pricing Explore About Blog Sign In Sign Up

jaceklaskowski > Mastering Apache Spark 2

Updated an hour ago

Mastering Apache Spark 2, by Jacek Laskowski

ABOUT 138 DISCUSSIONS 0 CHANGE REQUESTS

★ Star 682 Subscribe 308

Download PDF Read

Mastering Apache Spark 2

Welcome to Mastering Apache Spark 2 (aka #SparkLikePro)!

I'm Jacek Laskowski, an **independent consultant** who is passionate about **Apache Spark**, Apache Kafka, Scala, sbt (with some flavour of Apache Mesos, Hadoop YARN, and DC/OS). I lead Warsaw Scala Enthusiasts and Warsaw Spark meetups in Warsaw, Poland.

Tools

As in all other Scala courses...

- IDE of your choice
- sbt
- Databricks Community Edition *(optional)*

  Free hosted in-browser Spark notebook. Spark "cluster" managed by Databricks so you don't have to worry about it. 6GB of memory for you to experiment with.

Assignments

Like all other Scala courses, this course comes with autograders!

Course features 3 auto-graded assignments that require you to do analyses on real-life datasets.

Let’s jump in!