Scala for Apache Spark

Level: Advanced
Duration: undefined hours

We are very happy to announce that Prof Darren Wilkinson is running a series of four courses for data science and statistics with Scala.

Course 3 will be dedicated to understanding Apache Spark, the distributed Big Data analytics platform for Scala. Spark's Resilient Distributed Dataset (RDD) will be compared to the parallel collections examined in Course 1, and it will be shown how it can be used not only for the processing of very large data sets, but also for the parallel and distributed analysis of large or otherwise computationally-intensive models. We will see how Spark can be used both interactively and as a Scala library, producing compiled Spark applications for submission to a Spark cluster. We will also cover the use of Spark's DataFrame for more convenient processing of tabular data.

Course Outline

Course 3 will be dedicated to understanding Apache Spark, the distributed Big Data analytics platform for Scala. Spark's Resilient Distributed Dataset (RDD) will be compared to the parallel collections examined in Course 1 (Introduction to Scala and functional programming), and it will be shown how it can be used not only for the processing of very large data sets, but also for the parallel and distributed analysis of large or otherwise computationally-intensive models. We will see how Spark can be used both interactively and as a Scala library, producing compiled Spark applications for submission to a Spark cluster. We will also cover the use of Spark's DataFrame for more convenient processing of tabular data.

This suite of 4 half-day courses is aimed at statisticians and data scientists already familiar with a dynamic programming language (such as R, Python or Octave). Scala is a free modern, powerful, strongly-typed, functional programming language. It is fast and efficient, runs on the Java virtual machine (JVM), and is designed to easily exploit modern multi-core and distributed computing architectures. Scala is a favourite language for data engineering teams and others wanting to work with data at scale in an efficient, safe and timely fashion. For similar reasons, it is also very well suited to the development of robust data science, machine learning and statistical applications.

The courses can be taken independently, but do have pre-requisites which are detailed within the Prior Knowledge summaries. They will be delivered through a combination of lectures, live demos and hands-on practical sessions. The courses will be delivered by Prof Darren Wilkinson, a well-known expert in computational Bayesian statistics and a leading proponent of the use of strongly-typed FP languages (such as Scala) for scalable statistical computing. Participants will be expected to use their own laptops and to have a recent version of Java pre-installed. Other set-up instructions will be provided in advance to registered participants.

Prior Knowledge

Course 3 will assume a basic familiarity with Scala programming at least equivalent to that provided by Course 1 (Introduction to Scala and functional programming). It is not suitable for people completely new to Scala. Some familiarity with the use of Scala for data science will also be useful, such as provided by Course 2 (Scala for data science and machine learning). Note that we will not cover pySpark in this course, so no familiarity with Python is required.

Contact

hello@jumpingrivers.com