LIVE ONLINE TRAINING WITH ANDY PETRELLA AND XAVIER TORDOIR
MARCH 1–3, 2016 | 9:00–11:00AM PST

Building Distributed Pipelines for Data Science using Kafka, Spark, and Cassandra

Learn how to introduce a distributed data science pipeline in your organization

WHAT YOU'LL LEARN | INSTRUCTORS | SCHEDULE **REGISTER**

Building a distributed pipeline is a huge—and complex—undertaking. If you want to ensure yours is scalable, has fast in-memory processing, can handle real-time or

streaming data feeds with high throughput and low-latency, is well suited for ad-hoc queries, can be spread across multiple data centers, is built to allocate resources efficiently, *and* is designed to allow for future changes, join Andy Petrella and Xavier Tordoir for this immensely practical hands-on course.

Andy Petrella

Xavier Tordoir

What you'll learn—and how you can apply it

By the end of this course, you'll have a solid understanding of:

- The most important technologies for a distributed pipeline, when they should be used—and how
- How to integrate scalable technologies into your company's existing data architecture
- How to build a successful, scalable, elastic, distributed pipeline using a lean approach

This course is for you if...

- You're a data scientist with experience with data modeling, business intelligence, or a traditional data pipeline and need to deal with bigger or faster data
- You're a software or data engineer with experience in architecting solutions in Scala, Java, or Python and you need to integrate scalable technologies in your company's architecture

Prerequisites:

- Intermediate knowledge of an object-oriented language and basic knowledge of a functional programming language, as well as basic experience with a JVM
- Understanding of classic web architecture and service-oriented architecture
- Basic understanding of ETL, streaming data, and distributed data architectures
- Intermediate understanding of Docker and UNIX, as well as some basic knowledge about networks (IP, DNS, SSH, etc.)

About the instructors

Andy Petrella (@noootsab) is a mathematician

turned into a distributed computing entrepreneur, in addition to being a Scala and Spark trainer. Andy participated in many projects built using Spark, Cassandra, and other distributed technologies, in various fields including geospatial, IoT, automotive, and smart cities projects.

Andy is the creator of the Spark Notebook, the only reactive and fully Scala notebook for Apache Spark.

In 2015, Andy founded Data Fellas, working on an integrated and reactive distributed data science toolkit orchestrated from within the Spark Notebook.

Xavier Tordoir (@xtordoir) started his career as a researcher in experimental physics, focused on data processing. He took part in projects in

finance, genomics, and software development for academic research, working on time series, prediction of biological molecular structures and interactions, and applied machine learning methodologies. He developed solutions to manage and process data distributed across data centers.

Xavier founded and works at Data Fellas, a company dedicated to distributed computing and advanced analytics, leveraging Scala, Spark, and other distributed technologies.

Back to top

Schedule

Participants will be provided with Spark Notebooks on the first day, and will follow along using the notebooks throughout the course.

Day 1 (9:00–11:00AM PST)

- Introduction, Spark, Spark Notebook, and Kafka
- Assignment #1

Day 2 (9:00–11:00AM PST)

- Streaming: Spark, Kafka, and Cassandra
- Data analysis and external libraries
- Assignment #2

Day 3 (9:00–10:00AM PST)

- Microservices, cluster management, job orchestration, and live demo of end-to-end distributed pipeline
- Final discussion & wrap up

Back to top

Register now; March 1 is just around the corner.

Individual ticket: $299

Participate in this workshop from the convenience of your home, your office… whatever environment you find most comfortable and conducive to an intensive educational experience.

With additional post-course support: $799

Individual ticket plus the ability to correspond with the instructors about the content of the course for 2 weeks after the course ends. (Consulting for specific use cases is not included.)

Group ticket: $799

Project the workshop on a screen in a meeting room and invite your professional colleagues to participate. Learning alongside each other is a great team-building experience.

Group size is limited to 8 attendees. For larger groups, please contact onlinetraining@oreilly.com.

With additional post-course support: $1299

Group ticket plus the ability to correspond with the instructors about the content of the course for 2 weeks after the course ends. (Consulting for specific use cases is not included.)

Once you have registered, further details about joining the workshop will be available in your members.oreilly.com account. After the event concludes, access to the recording of the event will be added to your account.

Back to top