Big Data Analytics with PySpark

Level: Intermediate
Duration: 6 hours

Tools such as pandas offer a powerful way to manipulate and analyse data in Python. However, if you need to process a large dataset, a single machine might not cut it. Apache Spark is an analytics engine for processing large volumes of data on a computer cluster. It comes with a Python interface, PySpark, enabling those familiar with Python to easily get started with Spark for big data. This course will introduce data science at scale with the PySpark DataFrames API and Spark MLlib.

Course Outline

- **Introduction and motivation**: A brief overview of popular data analysis tools in Python, and when to consider using PySpark.
- **Distributed computing with Spark**: An introduction to the concepts of distributed computing, the Spark architecture, and Spark data structures.
- **DataFrames API**: Data analysis and manipulation with Spark DataFrames.
- **Spark SQL**: Querying DataFrames with SQL statements
- **Pandas and Spark**: Leveraging the familiar pandas API on distributed data with pandas-on-spark
- **Machine learning with MLlib**: Model fitting, data pipelines, hyperparameter searches and cross validation on distributed data with Spark's machine learning library.

Learning Outcomes

Session 1

By the end of session 1, participants will...

- have an understanding of the advantages of distributed computing
- be familiar with the Spark cluster architecture and distributed data structures
- be able to analyse and manipulate data with the DataFrames API
- be able to use Spark SQL together with DataFrames to query data

Session 2

By the end of session 2, participants will...

- understand how to effectively use the pandas API for Spark
- be able to train machine learning models with Spark MLlib
- know how to speed up hyperparameter searches and cross validation by taking advantage of data and model distributed machine learning.
- understand some of the caveats of Spark and distributed data

This course does not include:

- advanced usage of Spark SQL
- how to use Spark's low level RDD API
- how to use Spark Structure Streaming for computations on streaming data

Prior Knowledge

This course assumes familiarity with the Python programming language and common data structures. Some exposure to machine learning concepts would be an advantage but not essential. Attendance of the Introduction to Python and Programming with Python courses or equivalent experience should be sufficient.

Attendee Feedback

- "Very interesting course and plenty of resources to look into"
- "Highly knowledgeable trainers"
- "Flowed well, easy to follow, and clear instructions"

Contact

hello@jumpingrivers.com