Introduction to Apache Spark APIs for Data Processing

Luca Canali
CERN IT, Data Analytics and Spark Service

Welcome!

Instructor

- Luca Canali
  - Data Engineer with Spark and data analytics at CERN IT service since their start in 2016. Oracle DBA at CERN, since 2005.
  - Passionate about Spark: conference presentations on Spark and Big Data and minor contributions to Apache Spark upstream, see https://cern.ch/canali

Learning Objectives

- This course aims at:
  - Providing an overview of what Apache Spark can do for your data processing needs
    - Main architectural components and key abstractions in Spark
  - Present a few basic examples and code of how to use main Spark APIs:
    - DataFrame API, Spark SQL, Streaming, Machine Learning
    - Using Jupyter notebooks and PySpark
  - Demonstrate how to start using Spark at CERN

Course Outline

- Spark Architecture
- Spark Data APIs: DataFrames
  - Demo/hands on
- Spark Data APIs: Spark SQL
  - Demo/hands-on
- Spark as a data platform
  - DataFrame reader/writer and data formats
- Spark Advanced APIs
  - Intro to Streaming with demo
  - Intro to Machine Learning with demo
- Using Spark at CERN
  - Spark configuration options
  - CERN SWAN and integration with CERN Spark clusters

Disclaimer

- This is an introductory course
  - No prerequisites in data processing
  - Basic concepts will be covered
- It’s an exploration of Spark core concepts and main APIs
  - Not an extensive course, no advanced topics
  - This is intended as a launchpad to using Spark!
- The course is still evolving
  - Feedback is welcome

Website and Course Material

- Course website and repository:
  - https://sparktraining.web.cern.ch/
    - Videos and slides of the presentations
  - https://github.com/cerndb/SparkTraining
    - Notebooks with tutorials and hands-on exercises
  - Follow along and run notebooks at your pace
    - See the instructions to run on CERN SWAN
    - Or use you own Jupyter service (local, Colab, etc)

Questions?

Contact: Luca.Canali@cern.ch