Course 6	**Big Data with Apache Spark**
Program	**1. Introduction.** 1.1. Big data and data science projects. 1.2. Big data architectures. 1.3. Data processing and analytics patterns. 1.4. Reproducible data science. • Python: Anaconda and Jupyter. • Apache Zeppelin. **2. Spark framework and APIs.** 2.1. Evolution of the Apache Hadoop ecosystem. 2.2. Spark technology stack. 2.3. Spark APIs. 2.4. Programming languages: Scala, Python and R. 2.5. Logical and physical architecture. **3. Data processing with Spark.** 3.1. The Spark programming model. 3.2. New features in Spark v2. 3.3. Spark applications. • Local mode. • Cluster: Standalone manager, YARN, Mesos. 3.4. Spark programming (I) RDDs. 3.5. Spark programming (II): DataFrames, Datasets and GraphFrames. **4. Spark Streaming.** 4.1. Overall architecture. 4.2. Structured Streaming. 4.3. Programming with Spark Streaming. • Stream data sources. • Transformations. • Output operations. • Interaction with other components.. 4.4. Case example with Spark Streaming. **5. Machine Learning with Spark MLlib.** 5.1. Spark ML and MLlib. 5.2. Pipelines. 5.3. Case example: Classification and regression. 5.4. Case example: Recommender systems. 5.5. Case example: Dynamic clustering.
Prerequisites	It is strongly recommended to use GNU/Linux (Ubuntu, Debian or any other distribution), either as the native operating system or inside a virtual machine (e.g. VirtualBox). Participants will use a self-contained environment with all software requirements already pre-installed. It is assumed some previous knowledge about programming in any language, but preferably Python 3 or Java. Basic knowledge about data mining/machine learning algorithms is also advisable.
Bibliography	- Zecevic, P., Bonaci. M. (2016). *Spark in Action*. Manning Publications. - Karau, H., Konwinski, A., Wendell, P., Zaharia, M. (2015). *Learning Spark: Lightning-Fast Big Data Analysis*. O'Reilly Media. - Guller, M. (2015). *Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis*. Apress. - Karau, H., Warren, R. (2017) *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*. O'Reilly Media. - Ryza, S., Laserson, U., Owen, S., Wills, J. (2015). *Advanced Analytics with Spark: Patterns for Learning fom Data at Scale*. O'Reilly Media.