

Course 8	<b>Big Data with Apache Spark</b>
Program	<ol style="list-style-type: none"> <li><b>1. Introduction.</b> <ol style="list-style-type: none"> <li>1.1. Big data and data science projects.</li> <li>1.2. Big data architectures.</li> <li>1.3. Data processing and analytics patterns.</li> <li>1.4. Reproducible data science. <ul style="list-style-type: none"> <li>• Python: Anaconda and Jupyter.</li> <li>• Apache Zeppelin.</li> </ul> </li> </ol> </li> <li><b>2. Spark framework and APIs.</b> <ol style="list-style-type: none"> <li>2.1. Evolution of the Apache Hadoop ecosystem.</li> <li>2.2. Spark technology stack.</li> <li>2.3. Spark APIs.</li> <li>2.4. Programming languages: Scala, Python and R.</li> <li>2.5. Logical and physical architecture.</li> </ol> </li> <li><b>3. Data processing with Spark.</b> <ol style="list-style-type: none"> <li>3.1. The Spark programming model.</li> <li>3.2. New features in Spark v2.</li> <li>3.3. Spark applications. <ul style="list-style-type: none"> <li>• Local mode.</li> <li>• Cluster: Standalone manager, YARN, Mesos.</li> </ul> </li> <li>3.4. Spark programming (I) RDDs.</li> <li>3.5. Spark programming (II): DataFrames, Datasets and GraphFrames.</li> </ol> </li> <li><b>4. Spark Streaming.</b> <ol style="list-style-type: none"> <li>4.1. Overall architecture.</li> <li>4.2. Structured Streaming.</li> <li>4.3. Programming with Spark Streaming. <ul style="list-style-type: none"> <li>• Stream data sources.</li> <li>• Transformations.</li> <li>• Output operations.</li> <li>• Interaction with other components..</li> </ul> </li> <li>4.4. Case example with Spark Streaming.</li> </ol> </li> <li><b>5. Machine Learning with Spark MLlib.</b> <ol style="list-style-type: none"> <li>5.1. Spark ML and MLlib.</li> <li>5.2. Pipelines.</li> <li>5.3. Case example: Classification and regression.</li> <li>5.4. Case example: Recommender systems.</li> <li>5.5. Case example: Dynamic clustering.</li> </ol> </li> </ol>
Prerequisites	<p>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.</p> <p>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.</p>
Bibliography	<ul style="list-style-type: none"> <li>- Zecevic, P., Bonaci. M. (2016). <i>Spark in Action</i>. Manning Publications.</li> <li>- Karau, H., Konwinski, A., Wendell, P., Zaharia, M. (2015). <i>Learning Spark: Lightning-Fast Big Data Analysis</i>. O'Reilly Media.</li> <li>- Guller, M. (2015). <i>Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis</i>. Apress.</li> <li>- Karau, H., Warren, R. (2017) <i>High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark</i>. O'Reilly Media.</li> <li>- Ryza, S., Laserson, U., Owen, S., Wills, J. (2015). <i>Advanced Analytics with Spark: Patterns for Learning fom Data at Scale</i>. O'Reilly Media.</li> </ul>