### Course 8: Big Data with Apache Spark

#### Program

1. **Introduction.**
   1.1. Big data and data science projects.
   1.2. Big data architectures: batch, streaming and interactive.
   1.3. From MapReduce to in-memory processing.
   1.4. Reproducible data science.
      - Anaconda (Python distribution).
      - Jupyter.
      - Zeppelin.

2. **Spark framework and APIs.**
   2.1. Evolution of the Apache Hadoop ecosystem.
   2.2. Spark technology stack.
   2.3. Spark APIs.
      - Scala, Java, Python, R.
      - MLlib.
      - GraphX.
   2.4. Logical architecture.
   2.5. Physical architecture.

3. **Data processing with Spark.**
   3.1. The Spark programming model.
   3.2. Spark applications.
      - Local mode.
      - Cluster deployment.
      - Resource managers: standalone, YARN, Mesos.
   3.3. Spark programming: RDDs.
   3.4. Spark programming: Spark SQL, DataFrames and Datasets.

4. **Spark Streaming.**
   4.1. Overall architecture.
   4.2. Micro-batch execution model.
   4.3. Programming with Spark Streaming.
      - Stream data sources.
      - Transformations.
      - Output operations.
      - Interaction with other Spark 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. Classification and regression.
   5.4. Recommender systems.
   5.5. Clustering.

#### Prerequisites

It is advisable to use GNU/Linux (Ubuntu, Debian or any other distribution), either as the native operating system or inside a virtual machine (e.g. VirtualBox). Installation instructions for Apache Spark and related components will be provided. It is assumed some previous knowledge about programming (in any language, but preferably Python or Java). Basic knowledge about data mining/machine learning algorithms is also recommendable.

#### Bibliography


---

**Note:** The content above is a natural representation of the document, accurately transcribed. It is formatted to emphasize the structure and key points of the course outline and prerequisites. The bibliography is provided as a list of references that would typically be included in a full course material.