Course code: **SPARK**

Course title: **Big Data processing with Apache Spark**

Days: 2

Description:

Course intended for

Course main audience are software engineers and data analysts, who want to learn the basis of Big Data processing, which surpasses the capabilities of traditional processing, using tools of Apache Spark family. The course is both for people interested in starting to work with Big Data, as well as, people with previous experience in other Big Data systems, such as Apache Hadoop, who want to learn new technology.

Course objective

The attendees will learn about new problems that arise during Big Data analysis from various sources using Apache Spark family tools. During the course a general set of typical Big Data problems and their solutions using Apache Spark will be presented. Moreover, the attendees will have a general overview of the pros and cons of using Apache Spark for their business problem solving. In addition, the course allows the attendees to familiarize themselves with fast-moving Big Data processing field and the novel approach to problem solving that Apache Spark presents.

Course strengths

The course is conducted by people that have practical work experience with Big Data problems in their everyday practice. Hence, the material often goes beyond the common textbook information, that are often fragmented. Moreover, the content of the training is continuously updated following the modern advancements in the field. After the course the graduate will have a broad view of Big Data problem solving using Apache Spark damily tools for their specific business cases.

Requirements

The course requires experience in programming in Java, Scala or Python; preferred training language is Scala. Useful skills are: experience in data processing, functional programming, distributed processing, *nix systems.

Course parameters

2 working days, 2*7 working hours, group 8-10 people. The course contains presentations and coding workshops.

Course curriculum:

1. Introduction to Big Data
    - I. Definition
    - II. What is Big Data?
    - III. History of Big Data
    - IV. Stakeholders in Big Data project
    - V. Big Data problems
    - VI. Big Data processing types
        - Batch
        - Stream
2. Apache Spark
    - I. Introduction
    - II. History
    - III. Spark vs Hadoop
    - IV. MapReduce paradyme
    - V. Resilient Distributed Datasets (RDDs)
    - VI. Processing in memory vs from disk
    - VII. Architecture
    - VIII. Operation variants
        - Spark build-in cluster
        - Apache Mesos
        - Apache YARN
    - IX. Administration
3. Spark Core
    - I. Introduction
    - II. Java vs Spark vs Python
    - III. Connecting to cluster
    - IV. Dataset distribution
    - V. RDD operations
        - Transformations
        - Actions
    - VI. Shared variables
    - VII. Execution and testing
    - VIII. Job tuning
        - Serialization
        - Memory
4. Spark SQL
    - I. Introduction
    - II. Spark SQL vs Hive
    - III. Basic operation
    - IV. Data and schema
    - V. Queries

    VI. Hive integration
    VII. Execution and testing

5. Spark Streaming
    I. Introduction
    II. Basic operation
    III. Streams
        - Input
        - Transformation
        - Output
    IV. Execution and testing
6. Other Apache Spark family tools
    I. MLlib
    II. GraphX