Course code: **BIGDATA/STR**

Course title: **Stream Data Processing**

Days: 2

Description:

Course intended for:

Training course is intended mainly for programmers/developers and data analysts who want to learn the basics of Big Data processing. Tools from Apache Storm and Spark family can be used to process very big sets of data in (almost) real time. The training offers a good background knowledge both for those who want to start working with Big Data stream processing and for those who are already experienced in such systems, e.g. Apache Hadoop, and want to learn this new technology.

Course objective:

Participants will learn about the problem of analysis of very big sets of data (Big Data) in (almost) real time from various sources. During the course a basic set of problems with stream processing of Big Data will be presented along with the solutions using the tools from Apache Storm and Spark family. What is more, participants will be made aware of the advantages and disadvantages of these technologies when approaching real-world, business-related problems.

Course strengths

Training is conducted by experts who solve Big Data problems on a daily basis and thus have practical experience in this field.

Requirements:

Participants are expected to have at least basic Java programming experience, Scala or Python; Java is the preferred programming language of the training. It would be helpful if the participants have the following additional skills: basic data processing concepts, functional programming, distributed processing, *nix systems.

Course parameters

2*8 hours (2*7h net hours) of lectures and workshops. Group size: max 8-10 participants.

Course curriculum:

sages

1. Introduction to Big Data
    - Definition
    - What is Big Data?
    - Origin and history of Big Data
    - Webpages in Big Data projects
    - Big Data problems
    - Types of Big Data processing
        - Batch
        - Stream
    - When Hadoop is not enough
    - Data processing in (almost) real time
        - Definition
        - Advantages and disadvantages
        - Examples
    - Types of message delivery guarentee
        - at-most-once
        - at-least-once
        - exactly-once
2. Apache Storm
    - Introduction
    - History
    - Architecture
    - Run variants
        - Own claster
        - Apache Mesos
        - Apache YARN
    - Administraction
    - How it works
        - Topologies
        - Streams
        - Spouts
        - Bolts
        - Data models
        - Stream grouping
        - Mixing of programming languages
        - Guaranted message processing
    - Running and testing
    - RPC server
    - Entry queues
        - Kestrel
        - Apache Kafka
    - Trident
        - How it works
        - Data model
        - State
        - Running and testing

Sages sp. z o.o., ul. Nowogrodzka 62c, 02-002 Warszawa
tel. +48 22 203 56 00 fax: +48 22 203 56 01
e-mail: office@sages.io web: www.sages.io

        - Apache Spark
    - Introduction
    - History
    - Resilient Distributed Datasets (RDDs)
    - Processing from memory and from disc
    - Architecture
    - Variations of cluster running
        - Own Spark cluster
        - Apache Mesos
        - Apache YARN
    - Administration
    - Spark Core
        - Introduction
        - Java vs Scala vs Python
        - Cluster connection
        - Distributed data
        - RDD operations
        - Transformation
        - Actions
        - Shared variables
        - Running and testing
    - Spark Streaming
        - Introduction
        - How it works
        - Streams
        - Enter
        - Transformation
        - Exit
        - Running and testing
3. Other related stream technologies
    - Apache Flume
    - Amazon Kinesis
    - Akka
    - Apache Samza