Course code: **BIGDATA/F**

Course title: **Introduction to Big Data technology**

Days: 1

Description:

Course intended for:

The training is intended for analysts and programmers, who would like to make their first step towards getting familiar with the Big Data technology, where the processed data volume is of the highest priority and exceeds the capabilities of traditional architecture and systems such as relational databases or even data warehouses.

Course objective:

The training participants will acquire basic knowledge of the Big Data scale problems, understand the MapReduce algorithm, get to know the BigTable, the NoSQL databases using the example of HBase and HDFS distributed file systems, they will get familiar with the Pig and Hive analytical tools. The users will be able to identify strengths and weaknesses of specific technologies, they will know when to use a given technology.

Course strengths:

The program offers a quick review of the basic technologies of the Apache Hadoop ecosystem. Apart from presentations, the participants will be able to attend a workshop and explore data sets on their own.

Requirements:

The training participants are required to have basic knowledge of SQL, bash, Python (or a different script language), Java.

Course parameters:

8 hours (7 net hours) of lectures and workshops (with a visible emphasis on workshops).

Group size: no more than 8-10 persons

Course curriculum:

1. Introduction to Big Data
    I. Definition
    II. BI, Big Data and data warehouses
    III. Genesis and history, BigTable, MapReduce, GFS
    IV. Problem classification
    V. The concepts of real time, batch in the context of data processing
    VI. Data storage – files, databases of NoSQL
    VII. A review of Big Data systems and platforms
    VIII. A review of the Hadoop ecosystem
    IX. Big Data distributions
2. Introduction to MapReduce – the example of Hadoop platform
    I. Architecture
    II. HDFS and YARN
    III. Map Reduce Framework
    IV. Map Reduce Streaming
    V. Workshop
        HDFS
        Map Reduce
3. Introduction to data processing – the example of Pig
    I. Architecture
    II. Work modes
    III. Data types, keywords
    IV. Syntax

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

V. The Pig workshop

4. Introduction to data analysis – the example of Hive
    I. Architecture
    II. Work modes
    III. Data types
    IV. Syntax
    V. Data formats
    VI. Comparison with Pig
    VII. The Hive workshop
5. Introduction to NoSQL on the basis of HBase
    I. What is NoSQL, NoSQL vs. relational databases
    II. A review of non-relational databases, CAP theorem
    III. Designing of non-relational databases
    IV. HBase Architecture
    V. Data model
    VI. Use
    VII. CLI
    VIII. Data saving, reading
    IX. HBase workshop
6. Cluster monitoring and management – the example of Ambari
    I. CLI
    II. A review of Apache Ambari