Course code: **HADOOP/ANA**

Course title: **Big Data for analysts**

Days: 2

Description:

Course intended for:

Data analysts and programmers who want to start analyzing big sets of data.

Course objective:

The training is designed to prepare participants for the role of a big data analyst. The training focuses on the smooth entry into the basics of each tool so that the participant can easily navigate the Hadoop ecosystem in the future.

Course strengths

Learning about multiple tools and programming languages; training aims to show how easy it is to analyze data without the use of the console and IDE tools

Requirements:

Basic SQL, basic programming skills, especially in: Python, R lub Java

Course parameters

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

Course curriculum:

1. Introduction to Big Data and MapReduce
2. Apache Hadoop ecosystem
3. Big Data Analyst ecosystem
4. Hadoop architecture
    I. HDFS
    II. YARN
    III. MapReduce, Tez
    IV. Basic operations in Hadoop
5. Hive
    I. Introduction

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

II. Basic commands
III. Basic SQL queries
IV. Views
V. Functions
VI. Workshop on data exploration

6. Pig
   I. Introduction
   II. Pig Latin
   III. Pig Shell
   IV. Creating ETL processes
   V. Functions
   VI. User-defined functions
   VII. Using other sources of data
   VIII. Data exploration
7. Spark
   I. Introduction
   II. RDD
   III. Transformations and actions
   IV. Spark SQL
   V. Integration with Hive
   VI. Data exploration using PySpark and Spark
8. Machine Learning
   I. Introduction
   II. Supervised and unsupervised learning
   III. Machine Learning tasks
   IV. Solving typical Machine Learning tasks
      i. Spark ML
      ii. H2O