Course code:	**HADOOP/ADM**
Course title:	**Hadoop Administration**
Days:	3

Description:

Course intended for:

Course is intended for architects and administrators of big data clusters based on Apache Hadoop, who want to create and manage systems for which the volume of processed data has the highest priority and exceeds the capabilities of traditional architectures and systems, such as relative data bases or even data warehouses.

Course objective

Participants will gain necessary knowledge to build and manage big calculation clusters, based on Apache Hadoop. Discussed topics will include the preparation of the machines, cluster structure, installation of operating systems, installation of each environment component in Apache Hadoop, their basic usage and management in everyday work.

Course strengths

Course curriculum covers both an introduction to the subject and a comprehensive presentation of production stack around Apache Hadoop. The training is unique since the issues presented during it are not sufficiently covered in the available literature and the information on this subject is dispersed. The curriculum is constantly updated due to the rapid development of these solutions. Presented knowledge is the result of several years of practice of trainers in building systems based on Apache Hadoop platform.

Requirements

Participants are expected to have basic working knowledge of the Linux operating system.

Course parameters

3*8 hours (3*7 net hours) of lectures and workshops. During workshops participants will build their own calculation cluster and will learn how to use and manage it. Group size: max. 8-10 people

Course curriculum:

1. Introduction to Big Data and Apache Hadoop ecosystem
2. Infrastructure organization and calculating cluster topography
3. Cluster structure
    - I. Automated installation of a system by network (PXE)
    - II. Configuration of the operating system
4. Administration and management in Hadoop
    - I. Installation and configuration of HDFS
    - II. Use of HDFS from command line
    - III. Distributed copying of data with DistCP
    - IV. HDFS monitoring in operating system
    - V. Creating Snapshots and backup copies for HDFS
    - VI. YARN and MapReduce configuration and management
    - VII. Differences between basic types of files
    - VIII. Compression of data
5. Administration and management of Apache HBase
    - I. Installation on a cluster
    - II. Use of HBase Shell and administration tools
    - III. Migration of data from other sources
    - IV. Creating and recovery of backup copy
    - V. Monitoring and diagnostics
6. Hive and Pig installation for fast MapReduce tasks creation and optimization with Apache Tez
7. Aggregation of logs with Apache Flume
8. Management of workflow with Apache Oozie
9. Adding of graphic user's interface with the use of Hue
10. Apache Spark Installation and integration with YARN and HDFS
11. Automated installation of clusters
    - I. Ambari
    - II. deb rpm packets
    - III. Ready-made distributions (Hortonworks, Cloudera, MapR, etc)
    - IV. Puppet and Chef
12. Cluster performance tuning
13. System of high reliability / High availability (HA) in Hadoop ecosystem
14. Solving of common problems
15. Adding new calculation knots
16. Data safety and certifications
17. Administration and management
    - I. Ambari
    - II. Ganglia
    - III. Nagios