Course code: **PENTAHO/ETL**

Course title: **ETL solutions design using the Pentaho Data Integration (PDI)**

Days: 3

Description:

Course intended for:

The training is aimed at developers, architects and application administrators who want to create or maintain ETL (extract, transform, and load) processes using the Pentaho Data Integration (PDI). The training is also directed to people who want to supplement their knowledge of concepts related to data warehouses (DWH) and their implementation using the Pentaho Business Intelligence Suite.

Course objective:

Course participants will learn how to design, implement, monitor, start-up, tune ETL processes. After the training, participants will be able to choose the right set of tools and techniques for their projects. In addition to a general introduction to DWH concepts, training is focusing on the Pentaho Business Intelligence Suite and the Pentaho Data Integration (PDI).

Course strengths:

The training program includes both general introduction to the subject of ETL and DWH, as well as the overall presentation of the Pentaho Data Integration product stack. The training is unique because its subject is not fully recognized in the literature and knowledge of ETL and PDI is highly fragmented. Training program is constantly updated due to a rapid development of ETL solutions.

Requirements:

The participants are required basic knowledge of databases, basic programming skills in Java.

Course parameters:

3*7 hours of lectures and workshops at a ratio of 1:3. During workshops, in addition to simple exercises, participants will solve problems by implementing its ETL processes, model DWH data structures, perform basic administrative tasks. Group size: max. 8-10 people.

Program szkolenia

sages

Course curriculum:

1. Introduction
    I. Data warehouses basic concepts:
        i. OLTP, OLAP, database, data mart, data warehouse
        ii. ROLAP, MOLAP, HOLAP
        iii. Normalization, aggregation, facts, dimensions
        iv. SQL, MDX, XML/A
        v. ETL
        vi. BigData, BigTable, NoSQL, non-relational databases and data warehouses
        vii. Others
    II. Pentaho BI Suite
2. ETL
    I. Extraction of data
    II. Transformation, cleaning, replenishment od data
    III. Loading
    IV. Data quality
    V. Staging
    VI. Real-time DWH
    VII. ETL performance problems
    VIII. ETL tools
3. Pentaho Data Integration
    I. Architecture

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

i. Kettle
ii. Spoon
iii. Pan
iv. Kitchen
v. Carte

4. Working witch Spoon
   I. Installation, starting up, look & feel
   II. Variables
   III. Hops
   IV. Working with XML files and repositories
   V. Data sharing
5. Transformations
   I. Working with data sources
      i. Inputs and Outputs
      ii. Table input/output
      iii. Text file input/output
      iv. XML file input/output
      v. Deserialize from/Serialize to
      vi. Others
   II. Validation
      i. Data Validator
      ii. XSD Validator
      iii. Others

III. Replenishment
    i. Database/Web service/Stream lookup
    ii. HTTP/REST client
    iii. Combination lookup/update
    iv. Dimension lookup/update
    v. Others
IV. Transformation
    i. Transform
    ii. Joins
    iii. Mapping
    iv. Flow
    v. Filter
V. Optimization
    i. Bulk loading
    ii. Statistics
    iii. Parallel processing
    iv. Partitioning
    v. Clustering
VI. Custom code
    i. Java Expression, Java Class
    ii. Java Script
    iii. SQL Script
    iv. Regex

VII. Utilities
    i. Syslog
    ii. Mail
    iii. SSH
    iv. Others
VIII. Monitoring
IX. Versioning

6. Jobs
I. Jobs (kjb) and transformations (ktr)
II. Complex jobs
III. Custom code
    i. Java Script
    ii. SQL Script
    iii. Shell
IV. Workflows
    i. Conditions
V. Files
    i. XML
    ii. File transfer
    iii. File encryption
    iv. File management
VI. Monitoring
VII. Versioning

7. Kitchen and Pan
    I. Running jobs and transformations
    II. Scheduling
    III. Error handling
    IV. IO redirection
8. Cartle
    I. Running jobs and transformations remotely
9. Pentaho Data Integration Marketplace and Pentaho Data Integration Plug-Ins