

Course code: **PENTAHO/DW**

Course title: **DWH solutions design using Pentaho Business Intelligence**

Days: 5

Description:

Course intended for:

The training is aimed at developers, architects and application administrators who want to create or maintain data warehouses (DWH) using the Pentaho Business Intelligence Suite.

Course objective:

Course participants will learn how to design, implement, monitor, start-up, tune data warehouse components, know the general assumptions of the data warehouse such as the most popular relational/ROLAP data warehouse schemas. 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.

Course strengths:

The program includes both a general introduction to the subject of ETL, DWH, OLAP, as well as an overall presentation of the Pentaho Business Intelligence product stack. The training is unique because its subject is not fully recognized in the literature and knowledge of the DWH 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:

5 * 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

Course curriculum:



1. Program of training

2. Introduction

I. Data warehouses basic concepts:

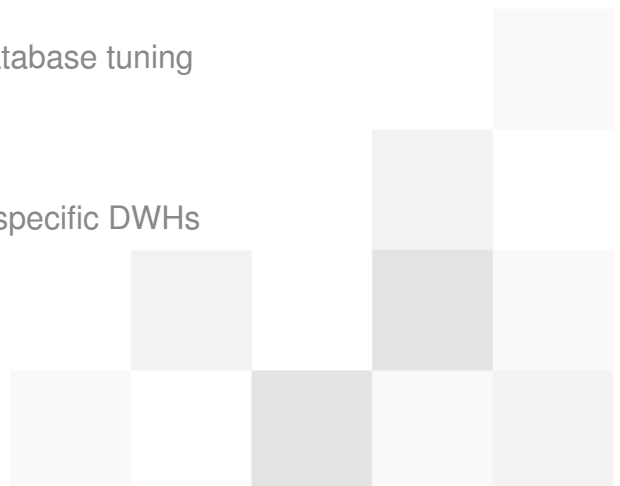
- i. OLTP, OLAP, database, data mart, data warehouse
- ii. ROLAP, MOLAP, HOLAP
- iii. Normalization, aggregation, facts, dimensions
- iv. SQL, MDX, XML/A
- v. ETL
- vi. BigData, BigTable, NoSQL, non-relational databases and data warehouses
- vii. Others

II. Pentaho BI Suite

3. DWH physical data structures

- I. Fact and dimension database tables
- II. Indices, views and materialized views
- III. ROLAP: star, snowflake, constellation database schema
- IV. The TPC Benchmark H
- V. ROLAP vs MOLAP
- VI. Database performance tuning in DWH
- VII. Pentaho Aggregate Designer and database tuning
- VIII. Time dim in DWH
- IX. GeoSpatial DWH and other domain specific DWHs

4. ETL and Pentaho Data Integration (PDI)



I. ETL

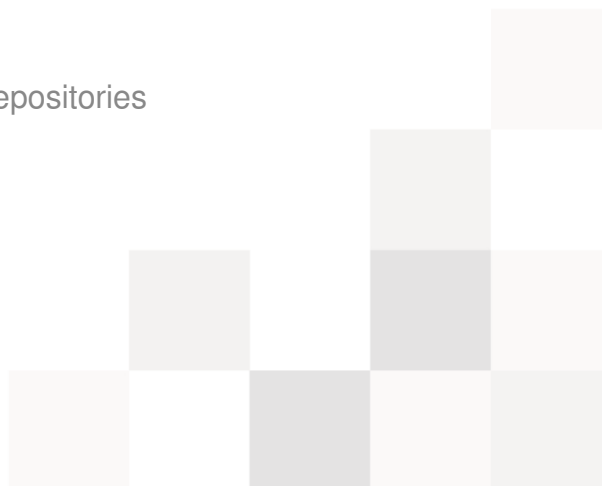
- i. Extraction of data
- ii. Transformation, cleaning, replenishment od data
- iii. Loading
- iv. Data quality
- v. Staging
- vi. Real-time DWH
- vii. ETL performance problems
- viii. ETL tools

II. Pentaho Data Integration

- i. Architecture
 - A. Kettle
 - B. Spoon
 - C. Pan
 - D. Kitchen
 - E. Carte

III. Working with Spoon

- i. Installation, starting up, look & feel
- ii. Variables
- iii. Hops
- iv. Working with XML files and repositories
- v. Data sharing
- vi. Transformations



- A. Working with data sources
 - a. Inputs and Outputs
 - b. Table input/output
 - c. Text file input/output
 - d. XML file input/output
 - e. Deserialize from/Serialize to
 - f. Others

vii. Jobs

- A. Jobs (kjb) and transformations (ktr)
- B. Complex jobs
- C. Custom code jobs
- D. Workflows
- E. Working with files
- F. Monitoring
- G. Versioning

IV. Kitchen and Pan

- i. Running jobs and transformations
- ii. Scheduling
- iii. Error handling
- iv. IO redirection

V. Cartle

5. OLAP and Pentaho Analysis Services (Mondrian)

I. OLAP cubes



II. Pentaho Schema Workbench (PSW)

III. Logical and physical structures

- i. cubes
- ii. metrics
- iii. dimensions, hierarchies, levels
- iv. tables
- v. relations
- vi. aggregations
- vii. expressions

IV. MDX

V. Slice, Dice, Drill

VI. OLAP tuning

6. Pentaho Report Designer (PRD)

I. Defining reports with PRD

- i. Data sources
- ii. Queries
- iii. Data extraction
- iv. Filtering and narrowing
- v. Visualization

II. Embedding reports

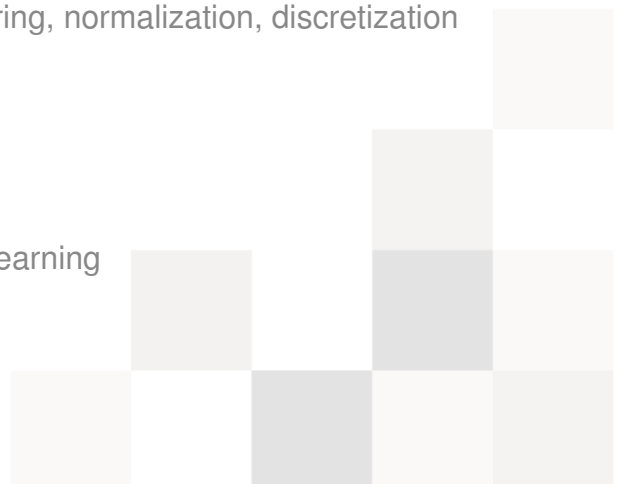
III. PDF, HTML, RTF export

IV. Report Wizard

7. Ad-hoc reporting and analytics



- I. Pentaho BI Platform/Portal (BA Platform)
 - II. Pentaho Interactive Reporting (PIR)
 - III. Pentaho Metadata Editor (PME)
 - IV. Pentaho Analyzer (PAZ)
8. Pentaho Dashboard Designer (PDD)
- I. Dashboards PDD
 - i. simple bar, line, area, pie, dial charts
 - ii. Tables
 - iii. Reports
 - iv. Parameters
 - v. Templates
 - II. Embedding dashboards
9. Pentaho Data Mining (WEKA)
- I. Architecture
 - II. WEKA Explorer
 - III. Preprocessing
 - i. ARFF data format
 - ii. Preparing data do mining
 - iii. Choosing proper data attributes (data correlation)
 - iv. Filtering with WEKA np.: filtering, normalization, discretization
 - v. Visualization
 - vi. Large dataset on 32bit JVM
 - vii. Stream mining, incremental learning



IV. Data mining

- i. Classification
- ii. Grouping
- iii. Association rules
- iv. Reducing dimensions of data
- v. Other techniques

V. Extending WEKA

VI. Using WEKA with Pentaho BI Suite

10. Pentaho Mobile BI

