

SZKOLENIE ŚREDNIO ZAAWANSOWANE

Techniczne aspekty budowania wydajnych procesów ładowania hurtowni danych z wykorzystaniem Pentaho Data Integration

PENTAHO/ETL

Czas trwania: 3 dni

Cele szkolenia

- Zdobycie przekrojowej wiedzy dotyczącej projektowania, implementowania, monitorowania, uruchamiania, strojenia procesów ETL
- Poznanie założeń hurtowni danych tj najpopularniejsze schematy relacyjnych hurtowni danych
- Zdobycie umiejętności wybierania właściwego zestawu narzędzi i technik dla swoich projektów
- Skupienie na stosie produktowym wybudowanym wokół Pentaho Business Intelligence a w szczególności na Pentaho Data Integration (PDI), poza ogólnym wprowadzeniem do pojęć teoretycznych

Zalety

- Program obejmuje zarówno ogólne wprowadzenie w tematykę ETL i DWH, jak i całościowe przedstawienie stosu produktowego Pentaho Data Integration
- Szkolenie jest unikalne, gdyż tematyka poruszana w jego trakcie nie jest wyczerpująco ujęta w dostępnej literaturze, a wiedza na ten temat jest mocno rozproszona
- Kameralne grupy - szkolenia technologiczne prowadzimy w grupach liczących do 8 osób. Pozwala to na indywidualne podejście oraz aktywizację każdego uczestnika
- Praktyka przed teorią - wszystkie szkolenia technologiczne prowadzone są w formie warsztatowej. Konieczna teoria jest wyjaśniana na przykładzie praktycznych zadań
- Konkretnie umiejętności - w ramach każdego szkolenia rozwijamy praktyczne umiejętności związane z daną technologią i tematyką
- Nauka z praktykami - wszyscy trenerzy na co dzień pracują w projektach, gwarantuje to dostęp do eksperckiej wiedzy i praktycznego know-how

Dla kogo?

- Szkolenie jest adresowane do programistów, architektów oraz administratorów aplikacji, którzy chcą tworzyć lub utrzymywać procesy ekstrakcji, transformacji i ładowania danych (ETL) z wykorzystaniem Pentaho Data Integration (PDI)



- Szkolenie jest także kierowane do osób, które chcą uzupełnić swoją wiedzę o pojęcia związane z hurtowaniami danych (DWH) oraz ich realizacją z wykorzystaniem oprogramowania Pentaho Business Intelligence Suite
- Program jest ciągle uaktualniany ze względu na szybki rozwój rozwiązań, których dotyczy szkolenie

Wymagania

- Podstawowa znajomość baz danych
- Podstawowa umiejętność programowania w języku Java



Program

1. Wstęp

- a. Wprowadzenie do hurtowni danych
 - OLTP, OLAP, bazy danych, hurtownie danych, data marty
 - ROLAP, MOLAP, HOLAP
 - Normalizacja, agregacja, fakty, wymiary
 - SQL, MDX, XML/A
 - ETL
 - BigData, BigTable, NoSQL, nierelacyjne hurtownie danych
 - Pozostałe
- b. Platforma Pentaho BI Suite

2. ETL

- a. Ekstrakcja danych
- b. Transformacja, oczyszczenie i uzupełnianie danych
- c. Ładowanie
- d. Wyznaczniki i ocena jakości danych
- e. Staging
- f. Hurtownie czasu rzeczywistego
- g. Kwestie wydajnościowe w kontekście procesów ETL
- h. Przegląd narzędzi ETL dostępnych na rynku

3. Pentaho Data Integration

- a. Architektura
 - Kettle
 - Spoon
 - Pan
 - Kitchen
 - Carte

4. Praca ze Spoon

- a. Instalacja, uruchomienie, Look & Feel (perspektywy, zakładki etc.)
- b. Zmienne
- c. Hopy
- d. Praca z repozytorium i plikami XML
- e. Współdzielenie obiektów tj. połączenia z bazami danych, partycje, klastry

5. Praca ze Spoon i z transformacjami

- a. Czytanie i zapisywanie danych z różnych źródeł danych
 - Przegląd Inputs and Outputs
 - Table input/output w szczegółach
 - Text file input/output w szczegółach
 - XML file input/output w szczegółach
 - Deserialize from/Serialize to file w szczegółach
 - Pozostałe
- b. Walidacja danych
 - Data Validator



- XSD Validator
 - Pozostałe
 - c. Uzupełnianie danych
 - Database/Web service/Stream lookup
 - HTTP/REST client
 - Combinaton lookup/update
 - Dimension lookup/update
 - Pozostałe
 - d. Transformacja danych
 - Transform
 - Joins
 - Mapping
 - Flow
 - Filter
 - e. Optymalizacja
 - Bulk loading
 - Statistics
 - Przetwarzanie równoległe
 - Partycjonowanie
 - Klastrowanie
 - f. Wykorzystanie własnego kodu w transformatach
 - Java Expression, Java Class
 - Java Script
 - SQL Script
 - Regex
 - g. Utilities
 - Syslog
 - Mail
 - SSH
 - Pozostałe
 - h. Monitorowanie wydajności i analiza logów
 - i. Wersjonowanie i praca grupowa
6. Praca ze Spoon i z zadaniami
- a. Zadania (kjb) a transformaty (ktr)
 - b. Tworzenie złożonych zadań
 - c. Wykorzystanie własnego kodu w zadaniach
 - Java Script
 - SQL Script
 - Shell
 - d. Przepływ pracy
 - Conditions
 - e. Praca z plikami
 - XML
 - File transfer



- File encryption
- File management

- f. Monitorowanie wydajności i analiza logów
- g. Wersjonowanie i praca grupowa

7. Praca z Kitchen i Pan

- a. Uruchamianie zadań i transformat z plików i z repozytorium
- b. Scheduling
- c. Obsługa błędów
- d. Przekierowanie wejścia/wyjścia

8. Praca z Cartle

- a. Zdalne uruchomienie transformat i zadań

9. Pentaho Data Integration Marketplace i Pentaho Data Integration Plug-Ins

