



KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Hurtownie danych i przetwarzanie analityczne [S2Inf1-TPD>HURT]

Przedmiot

Kierunek studiów
Informatyka

Rok/Semestr
1/1

Studia w zakresie (specjalność)
Technologie przetwarzania danych

Profil studiów
ogólnoakademicki

Poziom studiów
drugiego stopnia

Język oferowanego przedmiotu
polski

Forma studiów
stacjonarne

Wymagalność
obligatoryjny

Liczba godzin

Wykład
30

Laboratorium
20

Inne
0

Ćwiczenia
0

Projekty/seminaria
45

Liczba punktów ECTS

6,00

Koordynatorzy

dr hab. inż. Robert Wrembel prof. PP
robert.wrembel@put.poznan.pl

Wykładowcy

dr inż. Paweł Boiński
pawel.boinski@put.poznan.pl

dr hab. inż. Robert Wrembel prof. PP
robert.wrembel@put.poznan.pl

Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z języków programowania, systemów operacyjnych, a w szczególności z systemów baz danych (relacyjny model danych, język SQL, indeks drzewiasty, schematy koncepcyjne i logiczne, zarządzanie transakcjami i współbieżnym dostępem do danych, podstawy optymalizacji zapytań).

Cel przedmiotu

1. Wskazanie praktycznych problemów projektowania, implementowania, wdrażania i utrzymania systemów hurtowni danych (HD), dla danych klasycznych i danych masywnych (big data). 2. Przekazanie wiedzy dotyczącej projektowania systemów HD, w zakresie: architektur technicznych, modelowania danych, projektowania warstwy integrującej i zasilającej - ETL, struktur fizycznych, optymalizacji zapytań analitycznych, technologii przetwarzania danych masywnych. 3. Przedstawienie problematyki implementowania HD i aplikacji analitycznych, w zakresie: rozszerzeń SQL do analizy danych, wykorzystania struktur fizycznych (m.in., indeksy, partycje, perspektywy zmaterializowane) w procesie optymalizacji zapytań analitycznych. 4. Rozwijanie umiejętności rozwiązywania problemów, w zakresie: projektowania i implementowania systemów HD, oceny przydatności technologii HD i analizy danych do konkretnego zastosowania, testowania zaproponowanego rozwiązania pod kątem jego efektywności i funkcjonalności. 5. Kształtowanie umiejętności pracy zespołowej w ramach projektów. Kształtowanie umiejętności realizowania projektów praktycznych z zakresu HD i analizy danych.

Przedmiotowe efekty uczenia się

Wiedza:

1. zaawansowana wiedza z zakresu: (1) architektur systemów hd (klasycznych i big data), (2) teorii modelowania danych dla zastosowań analitycznych, (3) struktur danych dla hd, (4) technik optymalizacji zapytań analitycznych, (5) narzędzi i środowisk programistycznych wykorzystywanych do budowy hd.
 2. wiedza szczegółowa z zakresu systemów hd (architektury, techniki i narzędzia integracji danych, modele logiczne i implementacyjne, struktury fizyczne, optymalizacja zapytań gwiazdowych, strojenie wydajności, serwery klasy main-memory).
 3. wiedza o trendach rozwojowych architektur i technologii hd. wiedza o istniejących problemach dotyczących projektowania i budowania systemów hd.
 4. zaawansowana wiedza o w cyklu projektowania i życia systemów hd.
 5. zaawansowana wiedza dotycząca architektur, metod, technik, struktur fizycznych w rozwiązywaniu zadań projektowania systemów hd i rozwiązywaniu nietypowych zadań analizy danych (por. zajęcia projektowe).
- zaawansowania technologii i uczenia się.
2. rozumie konieczność korzystania z najnowszej wiedzy i rozwiązań z zakresu technologii hurtowni danych (np. systemy klasy nosql, architektury przetwarzania równoległego hadoop i spark, architektury przetwarzania danych strumieniowych) w rozwiązywaniu problemów technicznych i badawczych.

Umiejętności:

1. pozyskuje informacje z różnych źródeł wiedzy technicznej i naukowej (w j. pol. i ang.) na temat zagadnień objętych programem przedmiotu. potrafi integrować i konfrontować te informacje, dokonywać ich interpretacji i krytycznej oceny. potrafi uzasadnić wybór rozwiązania dla zadanego problemu (por. zajęcia projektowe).
2. posługuje się technikami informacyjno-komunikacyjnymi do realizowania projektów.
3. potrafi: (1) projektować i przeprowadzać eksperymenty, (2) interpretować uzyskane wyniki i wyciągać wnioski, (3) formułować i weryfikować hipotezy, w zakresie zadań deweloperskich systemów hd. potrafi zrealizować proste projekty badawcze w oparciu o technologie hd (por. zajęcia projektowe).
4. potrafi wykorzystać metody analityczne, symulacyjne i eksperymentalne do formułowania i rozwiązywania zadań technicznych i prostych zadań badawczych w zakresie systemów hd.
5. przy rozwiązywaniu zadań technicznych i badawczych integruje wiedzę z różnych obszarów informatyki (bazy danych, hurtownie danych, systemy operacyjne, systemy rozproszone, języki programowania, teoria złożoności obliczeniowej, technologie internetowe).
6. ocenia przydatność i możliwość wykorzystania nowych koncepcji, technologii i oprogramowania dla systemów hd (np. systemy klasy nosql, technologie strumieniowe, hadoop, spark).
7. potrafi ocenić przydatność i możliwość wykorzystania metod i narzędzi służących do rozwiązania zadania technicznego/inżynierskiego, polegającego na: zaprojektowaniu, zaimplementowaniu, lub ocenie wybranych komponentów systemu hd.
8. stosując nowe techniki, technologie, oprogramowanie, rozwiązuje złożone zadania projektowania, implementowania, wdrożenia wybranych komponentów systemu hd, często o charakterze badawczym (por. zajęcia projektowe).
9. na podstawie otrzymanej lub opracowanej samodzielnie analizy wymagań, potrafi zaprojektować system hd lub jego fragment. w tym celu wykorzystuje/przystosowuje właściwe dla problemu metody,

techniki, oprogramowanie lub opracowuje własne rozwiązanie (por. zajęcia projektowe).
10. potrafi pracować w zespole, przyjmując w nim różne role (por. zajęcia projektowe). potrafi dokonać analizy wymagań odnośnie do projektowanego systemu hd w interakcji z klientem.

Kompetencje społeczne:

1. rozumie, że obszarze technologii hurtowni danych (podobnie jak w innych obszarach informatyki)

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Wykłady: weryfikowanie założonych efektów kształcenia jest realizowane przez ocenę wiedzy i umiejętności wykazanych na kolokwium pisemnym o charakterze problemowym i testowym (student może korzystać z dowolnych materiałów dydaktycznych). Kolokwium składa się z 5-6 zadań problemowych i 6-8 pytań testowych jedno- lub wielokrotnego wyboru. Maksymalnie można uzyskać 40 punktów, z czego 6-8 za pytania testowe. Nie przyznaje się punktów ułamkowych. Wykłady uznaje się jako zaliczone od 21 punktów. Przyjmuje się następującą skalę ocen i punktów:
0-20: ndst, 21-24: dst, 25-28: dst+, 29-32: db, 33-36: db+, 37-40: bdb

Zajęcia projektowe: weryfikowanie założonych efektów kształcenia jest realizowane przez: (1) ocenę wiedzy i umiejętności związanych z realizacją zadań projektowych poprzez co-tygodniowe spotkania ze studentami, także z wykorzystaniem narzędzi videokonferencji, (2) okresowe prezentacje studentów z postępów w realizacji projektów, , także z wykorzystaniem narzędzi videokonferencji, (3) okresowe weryfikowanie jakości i zawartości dokumentacji technicznej, (4) obronę projektu przez studentów, (5) ocenę wyniku projektu, (6) ocenę dokumentacji projektowej.

Uwaga: projekty realizowane dla firm są oceniane także przez opiekunów projektów ze strony firm; w obronie projektu biorą udział przedstawiciele firmy.

Za projekt maksymalnie można uzyskać 100 punktów, z czego 50 za wynik projektu, 40 - za dokumentację techniczną, 10 - za prezentację końcową. Nie przyznaje się punktów ułamkowych. Projekt uznaje się z zaliczony od 51 punktów. Przyjmuje się następującą skalę ocen i punktów:
0-50: ndst, 51-60: dst, 61-70: dst+, 71-80: db, 81-90: db+, 91-100: bdb

W zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę realizacji zadań zleczanych na każdym zajęciach,
- ocenę wiedzy i umiejętności związanych z realizacją zadań laboratoryjnych poprzez rozwiązanie sprawdzianu (w formie testu i pytań otwartych) na koniec semestru.
- uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:
 - omówienia dodatkowych aspektów zagadnienia,
 - uwagi związane z udoskonaleniem materiałów dydaktycznych.

Warunkiem zaliczenia laboratorium jest przesłanie do systemu ekursy projektu zrealizowanego w ramach laboratorium.

W zakresie laboratorium przyjmuje się następującą skalę ocen w zależności od liczby uzyskanych punktów: <0;50%>: ndst., (50%;60%>: dst, (60%;70%>: dst+, (70%;80%>: db, (80%;90%>: db+, (90%;100%>: bdb.

Treści programowe

Program wykładów:

Problematyka integracji danych

Architektury integracji danych

Architektury systemów hurtowni danych dla zastosowań klasycznych

Zasilanie hurtowni danych - ETL/ELT

Modelowanie hurtowni danych

Struktury fizyczne dla hurtowni danych

Optymalizacja zapytań gwiazdzystych

Systemy klasy main-memory

Architektury przetwarzania danych masywnych

Program zajęć laboratoryjnych podzielono na następujące części:

1. Wprowadzenie do środowiska ćwiczeniowego
 - studium przypadku,

- źródła danych, schemat hurtowni danych,
 - podstawy metodyki Agile BI.
 - 2. Wprowadzenie do obsługi narzędzia Pentaho Data Integration
 - podstawowe pojęcia,
 - repozytorium,
 - transformacja oparta na jednym źródle danych,
 - transformacja podrzędna.
 - 3. Obsługa wielu źródeł danych
 - rozbudowa istniejących transformacji i transformacji podrzędnych o dodatkowe źródło danych,
 - sterowanie ścieżką przepływu danych,
 - metody łączenia danych.
 - 4. Dodatkowe transformacje
 - metody eliminowania duplikatów,
 - automatyczne generowanie danych dla wymiarów,
 - zasilanie tabeli faktów.
 - podstawy metodyki Agile BI.
 - 5. Zaawansowane transformacje
 - źródła danych oparte na plikach CSV, wykrywanie zmian w źródłach danych,
 - operacyjna składnica danych, odświeżanie hurtowni danych.
 - 6. Nowoczesne źródła danych
 - dokumenty XML, usługi sieciowe.
 - 7. Profilowanie i czyszczenia danych, dane historyczne
 - wykrywanie błędów w danych (dane referencyjne, wzorce danych),
 - automatyczne poprawianie błędów, naprawianie błędów w źródłach danych,
 - modyfikacja transformacji w celu przechowywania danych historycznych dla zmieniających się wymiarów.
 - 8. Poprawa wydajności procesu ETL, tematyczne hurtownie danych
 - masowe ładowanie danych (Oracle, PostgreSQL, MySQL)
 - wyliczanie agregatów z danych, przykład tematycznej hurtowni danych.
 - 9. Przetwarzanie danych w hurtowniach danych za pomocą języka SQL i jego rozszerzeń.
- Zajęcia są prowadzone w formie zajęć ćwiczeniowych przy komputerach, przy czym każdy student pracuje samodzielnie.

Każde zadanie jest poprzedzone krótką prezentacją a następnie omówione zagadnienia są ćwiczone w praktyce.

Tematyka zajęć

Program wykładów:

Problematyka integracji danych

Architektury integracji danych

Architektury systemów hurtowni danych dla zastosowań klasycznych

Zasilanie hurtowni danych - ETL/ELT

Modelowanie hurtowni danych

Struktury fizyczne dla hurtowni danych

Optymalizacja zapytań gwiazdzystych

Systemy klasy main-memory

Architektury przetwarzania danych masywnych

Program zajęć laboratoryjnych podzielono na następujące części:

1. Wprowadzenie do środowiska ćwiczeniowego

- studium przypadku,

- źródła danych, schemat hurtowni danych,

- podstawy metodyki Agile BI.

2. Wprowadzenie do obsługi narzędzia Pentaho Data Integration

- podstawowe pojęcia,

- repozytorium,

- transformacja oparta na jednym źródle danych,

- transformacja podrzędna.

3. Obsługa wielu źródeł danych

- rozbudowa istniejących transformacji i transformacji podrzędnych o dodatkowe źródło danych,

- sterowanie ścieżką przepływu danych,
 - metody łączenia danych.
 - 4. Dodatkowe transformacje
 - metody eliminowania duplikatów,
 - automatyczne generowanie danych dla wymiarów,
 - zasilanie tabeli faktów.
 - podstawy metodyki Agile BI.
 - 5. Zaawansowane transformacje
 - źródła danych oparte na plikach CSV, wykrywanie zmian w źródłach danych,
 - operacyjna składnica danych, odświeżanie hurtowni danych.
 - 6. Nowoczesne źródła danych
 - dokumenty XML, usługi sieciowe.
 - 7. Profilowanie i czyszczenia danych, dane historyczne
 - wykrywanie błędów w danych (dane referencyjne, wzorce danych),
 - automatyczne poprawianie błędów, naprawianie błędów w źródłach danych,
 - modyfikacja transformacji w celu przechowywania danych historycznych dla zmieniających się wymiarów.
 - 8. Poprawa wydajności procesu ETL, tematyczne hurtownie danych
 - masowe ładowanie danych (Oracle, PostgreSQL, MySQL)
 - wyliczanie agregatów z danych, przykład tematycznej hurtowni danych.
 - 9. Przetwarzanie danych w hurtowniach danych za pomocą języka SQL i jego rozszerzeń.
- Zajęcia są prowadzone w formie zajęć ćwiczeniowych przy komputerach, przy czym każdy student pracuje samodzielnie.

Każde zadanie jest poprzedzone krótką prezentacją a następnie omówione zagadnienia są ćwiczone w praktyce.

Metody dydaktyczne

Wykłady: prowadzone z wykorzystaniem slajdów ppt. Dla wykładów z fizyczną obecnością studentów, część problemów jest dodatkowo omawiana z wykorzystaniem ilustracji na tablicy; inicjowane są dyskusje nad rozwiązaniami omawianych problemów.

Zajęcia projektowe: prowadzone albo w laboratorium na terenie uczelni albo w infrastrukturze dostarczonej przez firmę, dla której jest realizowany projekt. Studenci rozwiązują praktyczne zadania projektowe, głównie zlecane przez firmy zewnętrzne. Zakres tematów jest każdego roku inny, zależnie od zainteresowań firm. Tematyka projektów obejmuje m.in. projektowanie warstwy ETL, ocenę efektywności struktur danych w różnych systemach zarządzania hurtowniami danych (m.in. Oracle, DB2, SQL Server), zaawansowaną analizę danych, projektowanie i implementowanie aplikacji analitycznych, składowanie i analizę danych NoSQL, techniki optymalizacji wykonania procesów ETL, architektury data lake. Zadania projektowe są realizowane w grupach 2-4 osobowych, w zależności od złożoności zadania. Dotychczasowe projekty realizowano na zlecenie następujących firm: Roche, IBM, Pearson/IOKI, Capgemini, Kogeneracja Zachód, PKO BP, Santander.

Laboratoria: prezentacja multimedialna, prezentacja jest uzupełniana krótkimi przykładami prezentowanymi w sposób tradycyjny z wykorzystaniem tablicy, wykonywanie ćwiczeń w hurtowni danych, omawianie trudniejszych ćwiczeń przy tablicy, odpowiedzi na pytania na bieżąco, rozwiązywanie problemów na bieżąco.

Literatura

Podstawowa

Vaisman A., Zimanyi E.: Data Warehouse Systems - Design and Implementation. Springer Verlag, 2014
 Jarke M., Lenzerini M., Vassiliou Y., Vassiliadis P.: Fundamentals of Data Warehouses. Springer, 2010, ISBN-13: 978-3642075643

Golfarelli M., Rizzi S.: Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill Osborne, 2009, ISBN-13: 978-0071610391

Uzupełniająca

Jiang B.: Constructing Data Warehouses with Metadata-driven Generic Operators, and more: Architecture, Methodology, and Paradigm; Concepts, Algorithms, and Operators; Principles, Recommendations, and Exercises. DBJ Publishing, 2011, ISBN-13: 978-3033029200

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	150	6,00
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	95	4,00
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu)	55	2,00