



KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Big Data i przetwarzanie w chmurze [S2Inf1-TPD>BIGD]

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
0

Inne
0

Ćwiczenia
0

Projekty/seminaria
0

Liczba punktów ECTS

2,00

Koordynatorzy

dr inż. Tomasz Koszlajda
tomasz.koszlajda@put.poznan.pl

Wykładowcy

Wymagania wstępne

Efekty uczenia się ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_W1-2, K_W4, K_W6-15, weryfikowane w procesie rekrutacji na studia 2 stopnia – efekty te prezentowane są w serwisie internetowym wydziału. Efekty uczenia się ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_U1-2, K_U4, K_U7-8, K_U14-20, K_U22-23, K_U26, weryfikowane w procesie rekrutacji na studia 2 stopnia – efekty te prezentowane są w serwisie internetowym wydziału. Efekty uczenia się ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_K1-9, weryfikowane w procesie rekrutacji na studia 2 stopnia – efekty te prezentowane są w serwisie internetowym wydziału. Ponadto w zakresie kompetencji społecznych student musi prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

Cel przedmiotu

1.Przekazanie studentom podstawowej wiedzy z nowych dziedzin zastosowań systemów baz danych i nowych modeli systemów baz danych, w zakresie przetwarzania danych w chmurach obliczeniowych, a w szczególności przetwarzania ogromnych zbiorów danych - Big Data. 2.Rozwijanie u studentów umiejętności rozwiązywania problemów analizy, projektowania i implementacji aplikacji nowych generacji baz danych.

Przedmiotowe efekty uczenia się

Wiedza:

ma zaawansowaną i pogłębioną wiedzę z zakresu szeroko rozumianych systemów informatycznych, podstaw teoretycznych ich budowania oraz metod, narzędzi i środowisk programistycznych wykorzystywanych do ich implementacji (k2st_w1)

ma wiedzę dotyczącą problematyki wydajności, odporności na awarie oraz spójności przetwarzania w rozproszonych, replikowanych i równoległych platformach obliczeniowych, bazującą na podstawach teoretycznych: teorii kolejek, modeli spójności przetwarzania replikowanych danych (k2st_w2)

ma zaawansowaną wiedzę szczegółową dotyczącą wybranych zagadnień z zakresu informatyki (k2st_w3)

ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach informatyki i innych, wybranych, pokrewnych dyscyplin naukowych (k2st_w4)

zna zaawansowane metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich i prowadzeniu prac badawczych w wybranym obszarze informatyki (k2st_w6)

Umiejętności:

potrafi pozyskiwać informacje z literatury, baz danych oraz innych źródeł (w języku ojczystym i angielskim), w zakresie dotyczącym alternatywnych rozwiązań przedstawianych na zajęciach problemów; (k2st_u1)

potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne, symulacyjne oraz eksperymentalne (k2st_u4)

potrafi integrować wiedzę z różnych obszarów informatyki, np. systemów baz danych lub systemów operacyjnych (k2st_u5)

potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć oraz nowych produktów informatycznych, np. w doborze odpowiedniego systemu klasy nosql; (k2st_u6)

potrafi dokonać krytycznej analizy istniejących rozwiązań technicznych oraz zaproponować ich ulepszenia, np. w kwestii równoważenia obciążenia; (k2st_u8)

potrafi ocenić przydatność metod i narzędzi służących do rozwiązania zadania inżynierskiego, polegającego np. na właściwych rozwiązaniach do analizy danych big data; (k2st_u9)

potrafi rozwiązywać złożone zadania informatyczne, np. wymagające wielokrotnych interakcji analizy danych; (k2st_u10)

Kompetencje społeczne:

rozumie, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe, (k2st_k1)

rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu informatyki w rozwiązywaniu problemów badawczych i praktycznych (k2st_k2)

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

- uczestnictwo w wykładach, aktywność w trakcie wykładów: szukanie odpowiedzi na pytania zadawane przez wykładowcę, krytyczne podejście do tłumaczenia wykładowców, zainteresowanie rozszerzeniem zakresu wykładów, znajdowanie błędów w materiałach wykładowych.

Ocena podsumowująca:

- ocena wiedzy i umiejętności wykazanych na egzaminie pisemnym o charakterze problemowym (student może korzystać z ograniczonego zbioru materiałów dydaktycznych); do uzyskania oceny 3.0 wymagane jest uzyskanie co najmniej 50% punktów. W ocenie finalnej uwzględniana jest również ocena z aktywność w trakcie wykładów.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

1. Technologia rozproszonych baz danych: fragmentacja, partycjonowanie i sharding danych.
2. Wydajność działania chmur – równoważenie obciążenia w chmurach obliczeniowych; Zarządzanie współbieżną realizacją dużych zadań obliczeniowych.
3. Poprawność działania baz danych z replikacją danych. .
4. Równoległe bazy danych. Architektury równoległych baz danych.
5. Technologia BigData. Model i architektura przetwarzania Map-Reduce. Technologia baz danych w pamięci operacyjnej.
6. Nowa generacja baz danych klasy NoSQL.

Tematyka zajęć

Tematyka wykładu obejmuje następujące zagadnienia:

1. Przesłanki dla technologii chmur baz danych. Usługa przetwarzania danych – DaaS. Przetwarzanie danych Big Data. Technologia rozproszonych baz danych: fragmentacja, partycjonowanie i sharding danych, podstawy fragmentacji danych - Consistent Hashing.
2. Wydajność działania chmur – równoważenie obciążenia w chmurach obliczeniowych; podstawowe pojęcia z teorii kolejek, notacja Kendalla; prawo Little'a; formuła Kingmana; protokoły równoważenia obciążenia w chmurze; protokoły szeregowania zadań. Wpływ zmienności wielkości zadań i częstości przedkładania zadań na jakość równoważenia obciążenia i szeregowania zadań; systemy kolejkowe typu G/G/N. Zarządzanie maszynami wirtualnymi – algorytm Distributed Resource Scheduler. Zarządzanie współbieżną realizacją dużych zadań obliczeniowych. Algorytmy sprawiedliwego przydziału zasobów: Max-min fairness i Dominant Resource Fairness.
3. Poprawność działania baz danych z replikacją danych. Spójność replikowanych baz danych: twierdzenie Brewera, klasyfikacja iPACeLC; modele spójności replikowanych baz danych; metody utrzymania replik Primary Copy, MultiMaster Copies i Kworum. Algorytmy utrzymywania replik; zegary logiczne, wektory wersji, protokół Paxos i algorytm RAFT.
4. Równoległe bazy danych. Architektury równoległych baz danych. Metody partycjonowania danych. Algorytmy równoległego przetwarzania baz danych.
5. Technologia BigData. Model i architektura przetwarzania Map-Reduce: HDFS , YARN i ZooKeeper. Platforma Spark: struktury danych i funkcjonalność. Technologia baz danych w pamięci operacyjnej; algorytmy i struktury danych: red-black tree, AVL-tree, T-tree, haszowanie liniowe.
6. Nowa generacja baz danych klasy NoSQL. Nowe modele logiczne: klucz-wartość, rodziny kolumn, dokumentowy i grafowy model danych. Paradygmat przetwarzania CRUD. Wydajność systemów baz danych z rodziny NoSQL. Sharding i replikacja w systemach NoSQL.
7. NewSQL połączenie relacyjnego modelu danych z technologiami shardingu i replikacji danych.

Metody dydaktyczne

Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy.

Literatura

Podstawowa

1. Big data: efektywna analiza danych, Mayer-Schonberger, MT Biznes 2017
2. Big data: najlepsze praktyki budowy skalowalnych systemów obsługi danych w czasie rzeczywistym, N. Marz, J. Warren, Helion 2016
3. Cloud Computing: Theory and Practice, D. Marinescu, Morgan Kaufmann 2013
4. Principles of Distributed Database Systems, M. Özsu, P. Valduriez, Springer 2011
5. Spark. Zaawansowana analiza danych, S.Ryza, U.Laserson, S.Owen, J.Wills, Helion 2016
6. Hadoop. Kompletny przewodnik. Analiza i przechowywanie danych, T. White, Hekion 2016
7. Performance Modeling and Design of Computer Systems, M. Harchol-Balter, Cambridge University 2013

Uzupełniająca

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	50	2,00
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	30	1,50
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu)	20	0,50