KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Big Data i przetwarzanie w chmurze	**1010512311010510161**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**1 / 1**

Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
Technologie przetwarzania danych	**polski**	**obligatoryjny**

Stopień studiów:	**Forma studiów** (stacjonarna/niestacjonarna)
II stopień	**stacjonarna**

Godziny	Liczba punktów
Wykłady: **30** Ćwiczenia: **-** Laboratoria: **30** Projekty/seminaria: **-**	**5**

Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny)	(ogólnouczelniany, z innego kierunku)
kierunkowy	**z danego kierunku**

Obszar(y) kształcenia i dziedzina(y) nauki i sztuki	**Podział ECTS (liczba i %)**
nauki techniczne	**5 100%**
nauki techniczne	**5 100%**

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Tomasz Koszlajda
email: Tomasz.Koszlajda@cs.put.poznan.pl
tel. 61 6652960
Informatyki
ul. Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Krzysztof Jankiewicz
email: Krzysztof.Jankiewicz@cs.put.poznan.pl
tel. 61 6652960
Informatyki
ul. Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z systemów baz danych, systemów operacyjnych, architektury systemów komputerowych oraz matematyki w zakresie rozkładów zmiennych losowych.
2	**Umiejętności:**	Powinien posiadać umiejętność rozwiązywania podstawowych problemów występujących w dziedzinie architektury systemów komputerowych, systemów baz danych i systemów operacyjnych oraz posiadać umiejętność pozyskiwania informacji ze wskazanych źródeł. Powinien również rozumieć konieczność poszerzania swoich kompetencji i mieć gotowość do podjęcia współpracy w ramach zespołu.
3	**Kompetencje społeczne**	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.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. ma zaawansowaną i pogłębioną wiedzę dotyczącą platform programistycznych do analizy danych Big Data; - [K2st_W1]

2. ma wiedzę dotyczącą fundamentalnych problemów informatyki: wydajności, odporności na awarie i poprawności przetwarzania w kontekście rozproszonych i replikowanych danych, bazującą na podstawach teoretycznych: teorii kolejek i modeli spójności przetwarzania replikowanych danych; - [K2st_W2]

3. ma wiedzę dotyczącą algorytmów fragmentacji, partycjonowania i replikacji danych; algorytmów równoważenia obciążenia w chmurach obliczeniowych oraz algorytmów zrównoleglania przetwarzania na platformach Big Data; - [K2st_W3]

4. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach informatyki w dziedzinach analiz danych Big Data oraz skalowalnego i rozproszonego przetwarzania danych; - [K2st_W4]

5. zna metody przetwarzania danych Big Data na platformie Spark Hadoop. - [K2st_W6]

Umiejętności:

1. 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]

2. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne, symulacyjne oraz eksperymentalne - [K2st_U4]

3. potrafi integrować wiedzę z różnych obszarów informatyki, np. systemów baz danych lub systemów operacyjnych - [K2st_U5]

4. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć oraz nowych produktów informatycznych, np. w doborze odpowiedniego systemu klasy NoSQL; - [K2st_U6]

5. potrafi dokonać krytycznej analizy istniejących rozwiązań technicznych oraz zaproponować ich ulepszenia, np. w kwestii równoważenia obciążenia; - [K2st_U8]

6. potrafi ocenić przydatność metod i narzędzi służących do rozwiązania zadania inżynierskiego, polegającego np. na właściwych rozwiązań do analizy danych Big Data; - [K2st_U9]

7. potrafi rozwiązywać złożone zadania informatyczne, np. wymagające wielokrotnych interakcji analizy danych; - [K2st_U10]

8. potrafi zaprojektować apliikacje do złożonej analizy danych Big Data, odpowiednio do specyfiki tych danych; - [K2st_U11]

Kompetencje społeczne:

1. rozumie, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe, - [K2st_K1]

2. rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu informatyki w rozwiązywaniu problemów badawczych i praktycznych - [K2st_K2]

Sposoby sprawdzenia efektów kształcenia

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:

- 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,

b) w zakresie laboratoriów:

- na podstawie oceny bieżącego postępu realizacji zadań,

Ocena podsumowująca:

a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:

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

- omówienie wyników egzaminu,

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę opanowanej wiedzy i umiejętności studenta w realizacji zajęć laboratoryjnych za pomocą testowych sprawdzianów,

- ocenę wiedzy i umiejętności związanych z realizacją zadań projektowych, uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za: aktywne uczestnictwo w zajęciach polegające na rozwiązywaniu zaproponowanych zadań, efektywność zastosowania zdobytej wiedzy podczas rozwiązywania zadanego problemu, uwagi związane z udoskonaleniem materiałów dydaktycznych, wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenie procesu dydaktycznego,

- rozkład punktów zdobywanych w ramach testowych sprawdzianów oraz innych form weryfikacji założonych efektów kształcenia to 50/50; dla uzyskania oceny dostatecznej należy uzyskać ponad 50% możliwych do zdobycia punktów; każde kolejne 10% możliwych do zdobycia punktów podnosi ocenę o pół stopnia.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

1. Przesłanki dla technologii chmur baz danych. Usługa przetwarzania danych ? DaaS. Przetwarzanie danych Big Data. Technologia rozproszonych baz danych: fragmentacja, partycjonowani 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.

Zajęcia laboratoryjne prowadzone są w formie piętnastu 2-godzinnych ćwiczeń, odbywających się w laboratorium. Program laboratorium obejmuje następujące zagadnienia:

1. Wprowadzenie do platformy Hadoop
2. Wprowadzenie do modelu przetwarzania MapReduce
3. Wprowadzenie do HDFS
4. Wprowadzenie do YARN
5. Pig
6. Hive
7. Wprowadzenie do Sparka
8. Wprowadzenie do programowania funkcyjnego (Scala)
9. Przetwarzanie danych w formacie RDD
10. Rozszerzenia funkcjonalności Sparka: Spark SQL ? przetwarzanie danych w formatach DataSets i DataFrames
11. Rozszerzenia funkcjonalności Sparka: GraphX, MLib, Spark Streaming

Metody dydaktyczne:

1.wykład: prezentacja multimedialna, prezentacja ilustrowana przykładami podawanymi na tablicy, rozwiązywanie zadań,

2.ćwiczenia laboratoryjne: rozwiązywanie zadań, ćwiczenia praktyczne, wykonywanie eksperymentów, dyskusja, praca w zespole, studium przypadków.

Literatura podstawowa:

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

Literatura uzupełniająca:

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)
1. udział w zajęciach laboratoryjnych	30
2. przygotowanie do ćwiczeń laboratoryjnych	7
3. dokończenie (w ramach pracy własnej) sprawozdań z ćwiczeń laboratoryjnych	7
4. udział w konsultacjach związanych z realizacją procesu kształcenia, w szczególności ćwiczeń laboratoryjnych / projektu	5
5. napisanie programów, uruchomienie I weryfikacja (czas poza zajęciami laboratoryjnymi)	10
6. przygotowanie do sprawdzianów testowych	10
7. udział w wykładach	30
8. zapoznanie się ze wskazaną literaturą / materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 100 stron	10
9. omówienie wyników egzaminu	2
10. przygotowanie do egzaminu i obecność na egzaminie: 18 godz. + 2 godz.	20

Obciążenie pracą studenta		
forma aktywności	**godzin**	**ECTS**
Łączny nakład pracy	121	5
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	69	3
Zajęcia o charakterze praktycznym	54	2