KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Przetwarzanie masywnych danych	**1010514381010519249**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**4 / 8**
Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
-	**polski**	**obieralny**

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

Godziny	Liczba punktów
Wykłady: **16** Ćwiczenia: **-** Laboratoria: **16** Projekty/seminaria: **-**	**3**

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	**3 100%**

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Krzysztof Dembczyński
email: krzysztof.dembczynski@put.poznan.pl
tel. 61 6652936
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Student rozpoczynający ten przedmiot powinien posiadać wiedzę z zakresu podstaw systemów zarządzania bazami danych, algorytmów i struktur danych, rachunku prawdopodobieństwa oraz statystyki i analizy danych.
2	**Umiejętności:**	Powinien również rozumieć konieczność poszerzania swoich kompetencji i wykazywać 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:

Przekazanie studentom podstawowej wiedzy w zakresie przetwarzania masywnych danych (bardzo dużych zbiorów danych): podstawowych metod organizacji, zarządzania, przechowywania, dostępu do danych oraz efektywnych algorytmów przetwarzania masywnych danych.

Rozwijanie u studentów umiejętności rozwiązywania problemów dotyczących przetwarzania masywnych danych.

Przedmiot przedstawia zagadnienia związane z obecnie popularnymi hasłami Big Data oraz Data Science.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. Ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie podstawowych metod przetwarzania masywnych danych. - [K_W4]
2. Ma szczegółową wiedzę związaną z zagadnieniami takimi jak: efektywna organizacja, przechowywanie i dostęp do masywnych danych w modelu relacyjnym, wielowymiarowym oraz nierelacyjny (NoSQL), podstawowe struktury i algorytmy przetwarzania masywnych danych (funkcje i tabele mieszające, indeksy, filtry Blooma) - [K_W5]
3. Ma szczegółową wiedzę związaną z zagadnieniami takimi jak: przybliżone przetwarzanie zapytań, dokładne i przybliżone wyszukiwanie najbliższych sąsiadów, przetwarzanie strumieni danych, paradygmat MapReduce. - [K_W5]
4. Zna podstawowe metody, techniki i narzędzia stosowane przy rozwiązywaniu prostych zadań informatycznych z zakresu przetwarzania masywnych danych. - [K_W8]

Umiejętności:

1. Zaprojektować ? zgodnie z zadaną specyfikacją ? oraz zrealizować prosty system informatyczny, używając właściwych metod, technik i narzędzi. - [K_U21]
2. Sformułować i zaimplementować algorytm z użyciem przynajmniej jednego z popularnych narzędzi. - [K_U22]

Kompetencje społeczne:

1. Rozumie, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe. - [K_K1]
2. Zna przykłady i rozumie przyczyny wadliwie działających systemów informatycznych, które doprowadziły do poważnych strat finansowych, społecznych lub też do poważnej utraty zdrowia, a nawet życia. - [K_K4]

Sposoby sprawdzenia efektów kształcenia

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:

- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na poprzednich wykładach;

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 różnej charakterystyce i złożoności problemów do rozwiązania (proste zadania dotyczące wiedzy podstawowej, zadania trudniejsze wymagające obliczeń lub symulacji algorytmów, zadania problemowe o dużej złożoności); łączna liczba pytań na egzaminie to ok. 10; wszystkie pytania są podobnie punktowane, łącznie można otrzymać 100 punktów; zaliczenie egzaminu jest od 50 punktów; ostateczna ocena jest średnią ważoną z egzaminu pisemnego i laboratorium.
- omówienie wyników egzaminu,

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

- ocenę realizacji zadań związanych z danymi zajęciami laboratoryjnymi; podczas każdych zajęć laboratoryjnych student otrzymuję listę zadań do wykonania; zadania dzielą się na niepunktowane i punktowane do realizacji na zajęciach laboratoryjnych oraz punktowane zadania domowe; możliwe jest uzyskanie dodatkowych punktów za aktywność podczas zajęć.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

- Problem eksplozji danych we współczesnym świecie; rozróżnienie systemów informatycznych pod względem wykorzystywania danych na systemy operacyjne oraz na systemy analityczne; zastosowania metod eksploracji danych oraz pułapki związane z przetwarzaniem masywnych danych.
- Historia i ewolucja systemów baz danych; modele danych w rozróżnieniu na rodzaje systemów przetwarzania danych: model relacyjny, wielowymiarowy i nierelacyjny (NoSQL).
- Struktury i algorytmy przetwarzania masywnych danych: przypomnienie wiedzy teoretycznej z zakresu funkcji i tabel mieszających, indeksy stosowane w przetwarzaniu masywnych danych, filtry Blooma, podstawowe zagadnienia dotyczące partycjonowania danych, przetwarzanie zapytań.
- Przetwarzanie przybliżone zapytań: próbkowanie danych, sygnatury i szkice (ang. sketches), algorytmy szybkiego zliczania, znajdowania najczęstszej wartości, przybliżenie wartości funkcji agregujących.
- Poszukiwanie najbliższych sąsiadów: struktury danych do dokładnego wyszukiwania najbliższych sąsiadów, przybliżone algorytmy bazująca na teorii lokalnie wrażliwych funkcji mieszających (ang. locality-sensitive hashing).
- Przetwarzanie strumieni danych: próbkowanie strumieni danych, filtrowanie strumieni danych, zliczenia unikatowych elementów w strumieniu, estymacja momentów.
- Wprowadzenie do MapReduce: paradygmat MapReduce, implementacja na przykładzie oprogramowania Hadoop lub Spark, podstawowe algorytmy takie jak zliczanie, operacje algebry relacji (projekcja, selekcja, grupowanie, łączenie), oraz mnożenie macierzy.

Zajęcia laboratoryjne prowadzone są w formie piętnastu dwugodzinnych ćwiczeń, odbywających się w laboratorium. Ćwiczenia realizowane są indywidualnie, z wyjątkiem niektórych zadań, które mogą być realizowane w zespołach dwuosobowych. Program laboratorium obejmuje następujące zagadnienia:

- Proste zadania z rachunku prawdopodobieństwa, które mają na celu pokazanie pułapek dotyczących analizy dużych zbiorów danych.
- Organizacja danych w systemie informatycznym dla przykładowego dużego zbioru danych, np. z dziedziny systemów rekomendacyjnych.
- Implementacja wybranych algorytmów i struktur danych związanych z przetwarzaniem masywnych danych, np. filtrów Blooma, szybkiego zliczania, wyszukiwania najczęstszego elementu w zbiorze, obliczania przybliżonych wartości funkcji agregujących; zastosowanie tych algorytmów do analizy przykładowego dużego zbioru danych.
- Implementacja algorytmu minhash oraz innych zagadnień związanych z lokalnie wrażliwymi funkcjami mieszającymi; zastosowanie tych algorytmów do analizy przykładowego dużego zbioru danych.
- Wprowadzenie do MapReduce: przedstawienie podstawowych zagadnień technicznych oraz implementacja prostych algorytmów w tym paradygmacie programowania, takich jak zliczanie, operacje algebry relacji, mnożenie macierzy; zastosowanie technologii MapReduce do analizy przykładowego dużego zbioru danych.
- Przykłady innych systemów przetwarzania masywnych danych.

Metody dydaktyczne:

1. Wykład: prezentacja multimedialna, prezentacja ilustrowana przykładami podawanymi na tablicy, dyskusja i analiza problemów.
2. Ćwiczenia laboratoryjne: rozwiązywanie zadań, dyskusja, praca w zespole.

Literatura podstawowa:

1. Mining of Massive Datasets, A. Rajaraman, J. D. Ullman, Cambridge University Press, 2012 (podręcznik jest legalnie dostępny w wersji elektronicznej: http://www.mmds.org/)

2. Systemy baz danych. Kompletny podręcznik. Wydanie II, Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom

Literatura uzupełniająca:

1. Data-Intensive Text Processing with MapReduce, J.Lin, Ch. Dyer, Morgan and Claypool Publishers, 2010 (podręcznik jest legalnie dostępny w wersji elektronicznej: http://lintool.github.com/MapReduceAlgorithms/)

2. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, R. Kimball, M. Ross, John Wiley & Sons 2002

3. Introduction to Information Retrieval, Ch. D. Manning, P. Raghavan, H. Schütze, Cambridge University Press 2008, (podręcznik jest legalnie dostępny w wersji elektronicznej: http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html)

4. Hurtownie danych: logiczne i fizyczne struktury danych, Z. Królikowski, Wydawnictwo Politechniki Poznańskiej 2007

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)

1. Udział w zajęciach laboratoryjnych:	16
2. Przygotowanie do ćwiczeń laboratoryjnych:	8
3. Zadanie domowe:	8
4. Udział w konsultacjach związanych z realizacją procesu kształcenia.	2
5. Udział w wykładach	16
6. Zapoznanie się ze wskazaną literaturą/materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 100 stron	10
7. Przygotowanie do zaliczenia	16

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
Łączny nakład pracy	76	3
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	34	1
Zajęcia o charakterze praktycznym	32	1