

KARTA OPISU MODUŁU KSZTAŁCENIA		
Nazwa modułu/przedmiotu Przetwarzanie masywnych danych		Kod 1010514381010509249
Kierunek studiów Informatyka	Profil kształcenia (ogólnoakademicki, praktyczny) ogólnoakademicki	Rok / Semestr 4 / 8
Ścieżka obieralności/specjalność -	Przedmiot oferowany w języku: polski	Kurs (obligatoryjny/obieralny) obieralny
Stopień studiów: I stopień	Forma studiów (stacjonarna/niestacjonarna) niestacjonarna	
Godziny Wykłady: 16 Ćwiczenia: - Laboratoria: 16 Projekty/seminaria: -		Liczba punktów 4
Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny) kierunkowy		(ogólnouczelniany, z innego kierunku) z danego kierunku
Obszar(y) kształcenia i dziedzina(y) nauki i sztuki nauki techniczne nauki techniczne		Podział ECTS (liczba i %) 4 100% 4 100%
Odpowiedzialny za przedmiot / wykładowca:		
dr hab. 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 podstawową wiedzę z zakresu systemów baz danych, algorytmiki, metod probabilistycznych oraz statystycznej analizy danych.
2	Umiejętności:	Powinien posiadać umiejętności programistyczne (zwłaszcza w zakresie systemów baz danych), rozwiązywania zadań z algorytmiki, metod probabilistycznych i statystycznej analizy danych.
3	Kompetencje społeczne	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 w zakresie organizacji, zarządzania i przetwarzania masywnych danych (bardzo dużych zbiorów danych).		
2. Rozwijanie u studentów umiejętności rozwiązywania problemów dotyczących organizacji, zarządzania i przetwarzania masywnych danych.		
Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia		
Wiedza:		
1. Ma uporządkowaną i podbudowaną teoretycznie wiedzę ogólną w zakresie przetwarzania dużych wolumenów danych oraz wiedzę szczegółową w zakresie wybranych zagadnień dotyczących tego obszaru informatyki. - [K1st_W4]		
2. Ma wiedzę o istotnych kierunkach rozwoju i najważniejszych osiągnięciach dokonanych w przetwarzaniu masywnych danych. - [K1st_W5]		
3. Zna podstawowe techniki, metody oraz narzędzia wykorzystywane w przetwarzaniu masywnych danych, głównie o charakterze inżynierskim. - [K1st_W7]		
Umiejętności:		

1. Potrafi pozyskiwać informacje z różnych źródeł, w tym z literatury oraz baz danych, zarówno w języku polskim jak i w języku angielskim, właściwie je integrować, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski, oraz wyczerpująco uzasadniać sformułowane przez siebie opinie - [K1st_U1]
2. Potrafi odpowiednio posługiwać się technikami przetwarzania masywnych danych, znajdującymi zastosowanie na różnych etapach realizacji przedsięwzięć informatycznych. - [K1st_U2]
3. Potrafi, formułując i rozwiązując zadania przetwarzania masywnych danych, zastosować odpowiednio dobrane metody, w tym metody analityczne, symulacyjne lub eksperymentalne. - [K1st_U4]
4. Potrafi - zgodnie z zadaną specyfikacją - zaprojektować oraz zrealizować projekt dotyczący przetwarzania masywnych danych, dobierając odpowiednie metody, techniki i narzędzia programistyczne. - [K1st_U10]
5. Ma umiejętność formułowania algorytmów przetwarzania danych masywnych i ich implementacji z użyciem przynajmniej jednego z popularnych narzędzi programistycznych. - [K1st_U11]
6. Potrafi planować i realizować proces własnego permanentnego uczenia się oraz zna możliwości dalszego dokształcania się (studia II i III stopnia, kursy i wykłady dostępne w Internecie). - [K1st_U19]

Kompetencje społeczne:

1. Rozumie, że wiedza i umiejętności dotyczące przetwarzania masywnych danych bardzo szybko stają się przestarzałe - [K1st_K1]
2. Ma świadomość znaczenia wiedzy w rozwiązywaniu problemów inżynierskich z zakresu przetwarzania danych masywnych oraz 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. - [K1st_K2]

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 wykładach.
- b) w zakresie laboratoriów / ćwiczeń:
 - 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 każdego zajęcia laboratoryjnego student otrzymuje listę zadań do wykonania (składającą się z zadań niepunktowanych, zadań punktowanych oraz zadań domowych); zaliczenie laboratorium jest od 50% zdobytych punktów podczas całego semestru; możliwe jest uzyskanie dodatkowych punktów za aktywność podczas zajęć.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

- Problem eksplozji danych w 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, przybliżanie wartości funkcji agregujących.
- Poszukiwanie najbliższych sąsiadów: struktury danych do dokładnego wyszukiwania najbliższych sąsiadów, przybliżone algorytmy bazujące 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 paradygmatu MapReduce na przykładzie oprogramowania Hadoop i 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:

<ul style="list-style-type: none"> - 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 na przykładzie oprogramowania Hadoop i Spark: przedstawienie podstawowych zagadnień technicznych oraz implementacja prostych algorytmów, takich jak zliczanie, operacje algebry relacji, mnożenie macierzy <p>Część wymienionych wyżej treści programowych jest realizowana w ramach pracy własnej studenta.</p> <p>Metody dydaktyczne:</p> <ol style="list-style-type: none"> 1. wykład: prezentacja multimedialna ilustrowana przykładami podawanymi na tablicy, dyskusja i analiza problemów. 2. ćwiczenia laboratoryjne: rozwiązywanie zadań, dyskusja, praca w zespole. 		
<p>Literatura podstawowa:</p> <ol style="list-style-type: none"> 1. Mining of Massive Datasets, A. Rajaraman, J. D. Ullman, Cambridge University Press, 2012 (podręcznik jest legalnie dostępny w wersji elektronicznej: http://infolab.stanford.edu/~ullman/mmds.html) 2. Systemy baz danych. Kompletny podręcznik. Wydanie II, Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom 		
<p>Literatura uzupełniająca:</p> <ol style="list-style-type: none"> 1. Hurtownie danych: logiczne i fizyczne struktury danych, Z. Królikowski, Wydawnictwo Politechniki Poznańskiej 2007 2. Hadoop in Action, Ch. Lam, , Manning Publications Co., 2011. 3. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, R. Kimball, M. Ross, John Wiley & Sons 2002 4. 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) 5. Projektowanie hurtowni danych, Zarządzanie kontaktami z klientami (CRM), Ch. Todman, Wydawnictwa Naukowo-Techniczne 2003 		
Bilans nakładu pracy przeciętnego studenta		
Czynność		Czas (godz.)
1. Udział w zajęciach laboratoryjnych/ćwiczeniach		16
2. Dokończenie (w ramach pracy własnej) zadań z ćwiczeń laboratoryjnych		16
3. Zadanie domowe		10
4. Udział w konsultacjach związanych z realizacją procesu kształcenia (częściowo mogą być realizowane drogą elektroniczną)		2 10
5. Przygotowanie do zajęć z obowiązkowymi zadaniami punktowanymi		16
6. Udział w wykładach		20
7. Zapoznanie się ze wskazaną literaturą i materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 200 stron		10
8. Przygotowanie do egzaminu		
Obciążenie pracą studenta		
forma aktywności	godzin	ECTS
Łączny nakład pracy	100	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	34	1
Zajęcia o charakterze praktycznym	42	2