KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu
Przetwarzanie masywnych danych

Przedmiot

Kierunek studiów
Informatyka

Rok/semestr
1/1

Studia w zakresie (specjalność)
Sztuczna inteligencja

Profil studiów
ogólnoakademicki

Poziom studiów
drugiego stopnia

Język oferowanego przedmiotu
polski

Forma studiów
stacjonarne

Wymagalność
obligatoryjny

Liczba godzin

Wykład	Laboratoria	Inne (np. online)
30	30	
Ćwiczenia	Projekty/seminaria	

Liczba punktów ECTS

4

Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:
dr hab. inż. Anna Kobusińska
email: Anna.Kobusinska@cs.put.poznan.pl
tel. 61 665-2964
Instytut Informatyki, Wydział Informatyki i Telekomunikacji
ul. Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot/wykładowca:
dr inż. Krzysztof Jankiewicz
email: Krzysztof.Jankiewicz@cs.put.poznan.pl
tel: 61 6652960
Instytut Informatyki, Wydział Informatyki i Telekomunikacji
ul. Piotrowo 2, 60-965 Poznań

Wymagania wstępne

Student rozpoczynający przedmiot Przetwarzanie masywnych danych powinien posiadać podstawową wiedzę z zakresu kształcenia ze studiów I stopnia zdefiniowanych w Uchwale Senatu PP weryfikowane w procesie rekrutacji na studia 2 stopnia, efekty te prezentowane są w serwisie internetowym wydziału www.cat.put.poznan.pl. W szczególności student rozpoczynający przedmiot powinien posiadać podstawową wiedzę z zakresu systemów operacyjnych, przetwarzania rozproszonego, sieci komputerowych, relacyjnych systemów baz danych oraz języka SQL i obiektowych języków programowania .

Ponadto, student powinien posiadać także umiejętność pozyskiwania informacji ze wskazanych źródeł, jak również rozumieć konieczność poszerzania swoich kompetencji i mieć gotowość do podjęcia współpracy w ramach zespołu.

Cel przedmiotu

Przekazanie studentom podstawowej wiedzy związanej z wyzwaniami przetwarzania masywnych danych w zakresie prezentacji teoretycznych i praktycznych aspektów konstrukcji systemów przetwarzania masywnych danych oraz wyzwań związanych z organizacją, zarządzaniem i przetwarzaniem masywnych danych. Rozwijanie u studentów umiejętności rozwiązywania problemów przetwarzania masywnych danych w systemach rozproszonych dużej skali.

Przedmiotowe efekty uczenia się

Wiedza

1. ma zaawansowaną wiedzę szczegółową dotyczącą wybranych zagadnień z zakresu informatyki takich jak: architektura i klasyfikacja systemów przetwarzania masywnych danych, narzędzia programowania w środowiskach przetwarzania masywnych danych [K2st_W3]

2. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach informatyki i innych, wybranych, pokrewnych dyscyplin naukowych w zakresie przetwarzania masywnych danych [K2st_W4]

3 ma zaawansowaną i szczegółową wiedzę o procesach zachodzących w cyklu życia systemów informatycznych sprzętowych lub programowych [K2st_W5]

4. zna zaawansowane metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich i prowadzeniu prac badawczych w zakresie przetwarzania masywnych danych [K2st_W6)

Umiejętności

1. potrafi pozyskiwać informacje z literatury, baz danych oraz innych źródeł (w języku polskim i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie [K2st_U1]

2. potrafi planować i przeprowadzać eksperymenty, w tym pomiary i symulacje komputerowe, interpretować uzyskane wyniki i wyciągać wnioski oraz formułować i weryfikować hipotezy związane ze złożonymi problemami inżynierskimi i prostymi problemami badawczymi przetwarzania masywnych danych [K2st_U3]

3. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych związanych z przetwarzaniem masywnych danych, metody analityczne, symulacyjne oraz eksperymentalne [K2st_U4]

4. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć (metod i narzędzi) oraz nowych produktów informatycznych w kontekście przetwarzania masywnych danych [K2st_U6]

5. potrafi - stosując m.in. koncepcyjnie nowe metody - rozwiązywać złożone zadania przetwarzania masywnych danych, w tym zadania nietypowe oraz zadania zawierające komponent badawczy [K2st_U10]

Kompetencje społeczne

1. rozumie, że w informatyce wiedza i umiejętności związane z przetwarzaniem masywnych danych bardzo szybko stają się przestarzałe [K2st_K1]

2.rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu przetwarzania masywnych danych 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:

a) w zakresie wykładów - na podstawie odpowiedzi na pytania dotyczące materiału omówionego na 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 w odpowiedziach na pytania o różnej charakterystyce i złożoności problemów do rozwiązania (proste zadania dotyczące wiedzy podstawowej, zadania trudniejsze wymagające obliczeń, zadania problemowe o dużej złożoności), które pojawią się w ramach dwóch kolokwiów zaliczeniowych; każde kolokwium musi być zaliczone na powyżej 50% możliwych do zdobycia punktów; ocena końcowa wynika z wyników uzyskanych z obu kolokwiów.

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez ocenę realizacji zadań związanych z danymi zajęciami laboratoryjnymi; podczas każdych zajęć laboratoryjnych student otrzymuje listę zadań do wykonania, ponadto student realizuje dwa projekty w połowie i pod koniec semestru; zaliczenie laboratorium wymaga uzyskania 50% możliwych do zdobycia punktów w pierwszej oraz w drugiej połowie semestru; możliwe jest uzyskanie dodatkowych punktów za aktywność podczas zajęć; ocena końcowa wynika z punktów zebranych w ramach całego semestru.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

1. Przedstawienie wyzwań związanych z przetwarzaniem masywnych danych: źródła masywnych danych, definicje masywnych danych, aspekty przetwarzania masywnych danych

2. Wprowadzenie do systemów przetwarzania masywnych danych: klasyfikacje systemów przetwarzania masywnych danych, architektury systemów Big Data (Lambda, Kappa).

3. Wprowadzenie do tematyki baz danych NoSQL: klasyfikacja ze względu na modele danych (key value, column-oriented, document-oriented, column-oriented, graph-oriented); budowa i działanie systemów NoSQL (data partitioning, load balancing, replication, data verioning, membership management, failure handling) na przykładzie Google BigTable, Cassandra, Neo4j; twierdzenie CAP, PACELC

4. Zarządzanie zasobami w systemach masywnych danych - na przykładzie systemy zarządzania zasobami Mesos i YARN (architektura, algorytmy lokowania zasobów)

5.Przechowywanie masywnych danych - rozproszone systemy plików na przykładzie Google File System (architektura, stosowane algorytmy)

6. Techniki przetwarzanie masywnych danych grafowych - na przykładzie systemu Pregel

7. Współbieżne przetwarzanie masywnych danych - na przykładzie platformy Apache Spark (architektura), techniki przetwarzania z wykorzystaniem Resilient Distributed Datasets (RDD)

8. Relacyjne przetwarzanie danych z wykorzystaniem Spark SQL, typy danych DataFrame i Dataset, przetwarzanie danych w Spark SQL, mechanizmy optymalizacji przetwarzania

9. Nowoczesne techniki przetwarzanie danych strumieniowych - na przykładzie platform przetwarzania strumieniowego : Apache Flink, Apache Kafka, Apache Spark Streaming

Program laboratorium obejmuje następujące zagadnienia:

1. Zapoznanie się ze środowiskami wykorzystywanymi na laboratoriach - instalacja, konfiguracja, interfejs programistyczny, typy danych, podstawowe operacje dostępne w danym systemie.

2. Praktyczne wykorzystanie systemów masywnych danych:

- realizacja zadań w środowisku Cassandra (zapoznanie się z bazą danych Cassandra, CQL, Java Api)

- realizacja zadań w środowisku Apache Flink

- realizacja zadań w środowisku Apache Kafka

- realizacja zadań w środowisku Apache Spark (zapoznanie się z platformą Apache Spark, Spark SQL, Spark Structured Streaming)

Metody dydaktyczne

1. Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy.

2. Zajęcia laboratoryjne: prezentacja multimedialna ilustrowana przykładami podawanymi na tablicy oraz demonstracja, dyskusja, warsztaty, ćwiczenia praktyczne, praca w zespole.

Literatura

Podstawowa

1. J.Berman, Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information, Morgan-Kaufman, 2013

2. N. Marz, J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Pubications Co., 2015

3. M. Zaharia, B. Chambers, Spark: The Definitive Guide, O'Reilly Media, 2018

4. A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2012 (podręcznik dostępny w wersji elektronicznej: http://infolab.stanford.edu/~ullman/mmds.html)

5. P. Sadalage, M. Flower, NoSQL distilled, Addison-Wesley, 2013

Uzupełniająca

1. S. Ryza, U. Lasersson, S. Owen, J. Wills, Spark. Zaawansowana analiza danych, Helion, 2015

2. J. S. Damji et al., Learning Spark - Lightning-Fast Data Analytics, O'Relly Media, 2020

3. I. Robinson, J. Webber, E. Eifrem, Graph Databases: New Opportunities for Connected Data, O'Reilly Media, Inc., 2015

4. A. Kobusińska, C. Leung, C.-H. Hsu, S. Raghavendra , V. Chang, Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing, Future Generation Computer Systems, 87, 2018

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	100	4,0
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	60	2,0
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych, przygotowanie do kolokwiów, wykonanie projektów)[1]	40	2,0

[1] niepotrzebne skreślić lub dopisać inne czynności