KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Big Data i przetwarzanie rozproszone

Przedmiot

Kierunek studiów

Sztuczna inteligencja

Studia w zakresie (specjalność)

Poziom studiów

pierwszego stopnia

Forma studiów

stacjonarne

Rok/semestr

3/6

Profil studiów

ogólnoakademicki

Język oferowanego przedmiotu

angielski

Wymagalność

obieralny

Liczba godzin

Wykład

30

Laboratoria

30

Inne (np. online)

0

Ćwiczenia

0

Projekty/seminaria

0

Liczba punktów ECTS

5

Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:

dr hab. inż. Anna Kobusińska, prof. PP

e-mail: anna.kobusinska@cs.put.poznan.pl

tel:+48 61 6652964

Wydział Informatyki i Telekomunikacji

Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot/wykładowca:

mgr inż. Adam Godzinski

e-mail: adam.godzinski@cs.put.poznan.pl

Wymagania wstępne

Studenci rozpoczynający ten przedmiot powinni posiadać podstawową wiedzę z zakresu systemów

operacyjnych, sieci komputerowych, relacyjnych systemów baz danych oraz języka SQL i programowania obiektowego. Ponadto, studenci powinni 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

Celem przedmiotu jest przekazanie studentom podstawowej wiedzy z zakresu big data i przetwarzania rozproszonego. W szczególności zaprezentowane zostaną teoretyczne i praktyczne aspekty konstrukcji systemów rozproszonych oraz wyzwania związane z organizacją, zarządzaniem i przetwarzaniem Big Data w tego typu systemach

Przedmiotowe efekty uczenia się

Wiedza

1. Studenci mają ugruntowaną wiedzę teoretyczną z zakresu systemów rozproszonych, przetwarzania rozproszonego oraz klasyfikacji i zarządzania Big Data

2. Studenci znają i rozumieją podstawowe paradygmaty, techniki, metody, algorytmy i narzędzia służące do rozwiązywania problemów przetwarzania rozproszonego i przetwarzania Big Data, w tym synchronizacji czasu w przetwarzaniu rozproszonym; podejść do replikacji danych i usług, koncepcji spójności replik; implikacji wystąpienia awarii komunikacyjnych poszczególnych węzłów i sieci; wpływu dużej skali na świadczenie podstawowych usług i kompromisy wynikające ze skali i Big Data; szereg rozproszonych algorytmów, takich jak rozgłaszanie i konsensus; podejście NoSql do przetwarzania i zarządzania danymi

3. Studenci posiadają wiedzę na temat kierunków rozwoju i najważniejszych osiągnięć w dziedzinie informatyki oraz innych wybranych i pokrewnych dyscyplinach naukowych z zakresu przetwarzania rozproszonego Big Data

Umiejętności

1. Studenci rozumieją, że wiedza i umiejętności szybko się dezaktualizują w zakresie informatyki, a w szczególności przetwarzania rozproszonego Big Data i dostrzegają potrzebę ciągłego dokształcania się i podnoszenia kwalifikacji

2. Studenci potrafią analizować złożoność obliczeniową i komunikacyjną algorytmów rozproszonych

3. Studenci potrafią stosować odpowiednie metody (analityczne, symulacyjne, eksperymentalne) do rozwiązywania określonych problemów obliczeń rozproszonych

4. Studenci potrafią sprawnie planować i przeprowadzać eksperymenty, w tym pomiary i symulacje komputerowe, interpretować uzyskane wyniki i wyciągać wnioski na podstawie wyników eksperymentów w kontekście przetwarzania rozproszonego oraz problemów przetwarzania i zarządzania Big Data

5. Studenci potrafią zaprojektować i zaimplementować algorytm rozproszony, dobierając odpowiedni język do zadania oraz stosując odpowiednie techniki, metody i narzędzia

Kompetencje społeczne

1. Studenci rozumieją, że w dziedzinie informatyki wiedza i umiejętności związane z przetwarzaniem Big Data szybko się dezaktualizują

2. Studenci rozumieją znaczenie wykorzystania najnowszej wiedzy z zakresu przetwarzania rozproszonego i Big Data w rozwiązywaniu problemów badawczych i praktycznych

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w odniesieniu do wykładów - na podstawie odpowiedzi na pytania związane z omawianym materiałem podczas wykładów.

b) w odniesieniu do laboratoriów - na podstawie oceny bieżącego postępu w realizacja zadań.

Ocena podsumowująca:

a) Wykłady: weryfikacja założonych efektów uczenia się odbywa się podczas egzaminu, który ma formę testu wielokrotnego wyboru oraz zadań o zróżnicowanej charakterystyce i złożoności (proste zadania z wiedzy podstawowej, zadania trudniejsze wymagające obliczeń, zadania problemowe o dużej złożoności) dotyczących tematyki prezentowanej na wszystkich wykładach.i za ego rozwiązanie przyznawana jest określona liczba punktów. Punkty są sumowane i następujące skala jest wykorzystywana do określenia oceny: <50% - 2,0, [50%, 60%) - 3,0, [60%, 70%) - 3,5, [70%, 80%) - 4,0, [80%, 90%) - 4,5 i [90%,100%] - 5,0.

b) Zajęcia laboratoryjne: weryfikacja zakładanych efektów kształcenia odbywa się poprzez ocenę realizacji zadań związanych z danymi zajęciami laboratoryjnymi; na każdych zajęciach laboratoryjnych studenci otrzymują listę zadań do wykonania; ponadto studenci realizują trzy projekty. Studenci muszą uzyskać co najmniej 50% możliwych punktów z projektów. Za aktywność na zajęciach laboratoryjnych można uzyskać dodatkowe punkty; ocena końcowa wynika z punktów zebranych w całym semestrze.

Treści programowe

Wprowadzenie do systemów rozproszonych; zalety i wyzwania systemów rozproszonych; opóźnienia komunikacyjne i częściowe uszkodzenia; protokoły sieciowe; transparentność systemów rozproszonych; systemy klient-serwer; zdalne wywołanie procedur (RPC).

Modele i awarie systemu: synchroniczne, częściowo synchroniczne i asynchroniczne modele sieci; awarie crash-stop, crash-recovery i błędy bizantyjskie; awarie, wady i odporność na awarie; problem dwóch generałów.

Pojęcie czasu w systemach rozproszonych, zegary i kolejność zdarzeń; zegary fizyczne; UTC; synchronizacja i dryf zegara; protokół czasu sieciowego (NTP). Przyczynowość; relacja "happen-before";

czas logiczny; zegary Lamporta; zegary wektorowe. Rozgłaszanie wiadomości (FIFO, przyczynowe, globalnie uporządkowane); protokoły gossip

Wyzwania związane z przetwarzaniem Big Data: źródła Big Data, definicje i cechy Big Data, różne aspekty przetwarzania Big Data Klasyfikacje systemów przetwarzania rozproszonego Big Data, architektury systemów Big Data (Lambda, Kappa).

Wprowadzenie do baz danych NoSQL: klasyfikacja (modele wartości klucza, zorientowane na kolumny, zorientowane na dokument, zorientowane na kolumny, zorientowane na wykresy); budowa systemów NoSQL (partycjonowanie danych, równoważenie obciążenia, replikacja, wersjonowanie danych, zarządzanie członkostwem, obsługa awarii) w oparciu o rozwiązania: Google BigTable, Dynamo, Cassandra; twierdzenia CAP i PACELC

Replikacja. Kwora; idempotencja; spójność replik; replikowana maszyna stanów; linearizowalność; ostateczna spójność; konsensus; wynik FLP; wybór lidera; algorytmy konsensusu Paxos i Raft.

Przetwarzanie Big Data w oparciu o platformę Apache Spark (architektura), techniki przetwarzania z wykorzystaniem Resilient Distributed Datasets (RDD) ;relacyjne przetwarzanie danych z wykorzystaniem typów danych Spark SQL, DataFrame i Dataset, przetwarzanie danych w Spark SQL, mechanizmy optymalizacji przetwarzania

Metody dydaktyczne

1. Wykłady: prezentacja multimedialna ilustrowana przykładami podanymi na tablicy.

2. Zajęcia laboratoryjne: prezentacja multimedialna ilustrowana przykładami podanymi na tablicy oraz realicja projektów

Literatura

Podstawowa

1. Modern Operating Systems (free PDF available online) by Andrew S Tanenbaum, Herbert Bos

2. Kleppmann, M. (2017). Designing data-intensive applications. O'Reilly.

3. Tanenbaum, A.S. and van Steen, M. (2017). Distributed systems, 3rd edition. available online.

4. Cachin, C., Guerraoui, R. and Rodrigues, L. (2011) Introduction to Reliable and Secure Distributed Programming. Springer (2nd edition).

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

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

Uzupełniająca

1. Spark in Action, Bonaći M., Zečević P., Manning, 2015

2. A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2012 (online: http://infolab.stanford.edu/~ullman/mmds.html)

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

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	125	5,0
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	60	2,5
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwiów/egzaminu, wykonanie projektu)[1]	65	2,5

[1] niepotrzebne skreślić lub dopisać inne czynności