



KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Przetwarzanie masywnych danych - BigData

Przedmiot

Kierunek studiów

Informatyka

Studia w zakresie (specjalność)

Poziom studiów

pierwszego stopnia

Forma studiów

niestacjonarne

Rok/semestr

4/8

Profil studiów

ogólnoakademicki

Język oferowanego przedmiotu

Polski

Wymagalność

obieralny

Liczba godzin

Laboratoria

16

Wykład

16

Liczba punktów ECTS

4

Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:

dr inż. Krzysztof Jankiewicz

email: Krzysztof.Jankiewicz@cs.put.poznan.pl

tel: 61 6652960

wydział: Wydział Informatyki i Telekomunikacji

adres: ul. Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot/wykładowca:

Wymagania wstępne

Znajomość relacyjnych systemów baz danych. Znajomość języka SQL. Podstawowa znajomość obiektowych języków programowania np. Java.

Cel przedmiotu

1. Przekazanie studentom podstawowej wiedzy w zakresie organizacji, zarządzania i przetwarzania Big Data (bardzo dużych zbiorów danych).
2. Rozwijanie u studentów umiejętności rozwiązywania problemów dotyczących organizacji, zarządzania i przetwarzania Big Data.

Przedmiotowe efekty uczenia się

Wiedza

Ma wiedzę o istotnych kierunkach rozwoju i najważniejszych osiągnięciach dokonanych w przetwarzaniu Big Data. (K1st_W5)



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)

Zna podstawowe techniki, metody oraz narzędzia wykorzystywane w przetwarzaniu Big Data, głównie o charakterze inżynierskim. (K1st_W7)

Umiejętności

Potrafi, formułując i rozwiązując zadania przetwarzania Big Data, zastosować odpowiednio dobrane metody, w tym metody analityczne, symulacyjne lub eksperymentalne. (K1st_U4)

Potrafi odpowiednio posługiwać się technikami przetwarzania Big Data, znajdującymi zastosowanie na różnych etapach realizacji przedsięwzięć informatycznych. (K1st_U2)

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 formułowane przez siebie opinie (K1st_U1)

Potrafi - zgodnie z zadaną specyfikacją - zaprojektować oraz zrealizować projekt dotyczący przetwarzania Big Data, dobierając odpowiednie metody, techniki i narzędzia programistyczne. (K1st_U10)

Potrafi planować i realizować proces własnego permanentnego uczenia się oraz zna możliwości dalszego kształcenia się (studia II i III stopnia, kursy i wykłady dostępne w Internecie). (K1st_U19)

Ma umiejętność formułowania algorytmów przetwarzania Big Data i ich implementacji z użyciem przynajmniej jednego z popularnych narzędzi programistycznych. (K1st_U11)

Kompetencje społeczne

Rozumie, że wiedza i umiejętności dotyczące przetwarzania Big Data bardzo szybko stają się przestarzałe (K1st_K1)

Ma świadomość znaczenia wiedzy w rozwiązywaniu problemów inżynierskich z zakresu przetwarzania Big Data 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)

Metody weryfikacji efektów uczenia się i kryteria oceny

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 sprawdzianie zaliczeniowym wykładowym 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 to ok. 10; wszystkie pytania są podobnie punktowane, łącznie można



otrzymać 100 punktów; zaliczenie sprawdzianu wykładowego jest od 50 punktów; ostateczna ocena jest średnią ważoną ze sprawdzianu wykładowego pisemnego i laboratorium.

- omówienie wyników sprawdzianu wykładowego,

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez ocenę realizacji zadań związanych z bieżącymi zajęciami laboratoryjnymi; podczas każdego zajęcia 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; możliwe jest uzyskanie dodatkowych punktów za aktywność podczas zajęć; ocena końcowa wynika z punktów zebranych w ramach całego semestru.

Treści programowe

Wykład:

Program wykładu obejmuje następujące zagadnienia:

- Wprowadzenie do systemów Big Data, motywacje, definicje, problemy świata Big Data, typy przetwarzania narzędzia. Architektury systemów Big Data (Lambda, Kappa). Modele baz danych noSQL, BASE, twierdzenie CAP.
- Platforma Hadoop, rozproszone systemy plików na przykładzie HDFS, systemy szeregowania zadań w systemach Big Data na przykładzie YARN, silniki przetwarzania wsadowego danych na przykładzie MapReduce, techniki optymalizacji przetwarzania MapReduce, dekomponowanie złożonych problemów na sekwencje działań MapReduce, Hadoop Streaming
- Narzędzia programistyczne wyższego poziomu na przykładzie systemów Pig i Hive, architektura, techniki optymalizacji przetwarzania, Pig Latin, Hive SQL. Fizyczne organizacje danych, format pliku ORC, filtr Blooma.
- Wprowadzenie do programowania funkcyjnego Scala
- Nowoczesne silniki przetwarzania Big Data na przykładzie platformy Spark, architektura, techniki przetwarzania danych niestrukturalnych z wykorzystaniem RDD, obsługa RDD par klucz-wartość, optymalizacja przetwarzania RDD.
- Relacyjne przetwarzanie danych z wykorzystaniem Spark SQL, typy danych DataFrame i Dataset, przetwarzanie danych w Spark SQL, mechanizmy optymalizacji przetwarzania.

Laboratoria:
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:

- Zapoznanie się ze środowiskami wykorzystywanymi na laboratoriach
- Hadoop - wprowadzenie, MapReduce
- HDFS, YARN
- Wysokopoziomowe wsadowe przetwarzanie danych - Pig
- Wysokopoziomowe wsadowe przetwarzanie danych - Hive
- Wprowadzenie do języka Scala
- Platforma Spark - wprowadzenie
- Spark - RDD - podstawy
- Spark - RDD - klucz-wartość
- Spark - RDD - wydajność



- Spark - DataFrames
- Spark - Datasets

Metody dydaktyczne

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

Literatura

Podstawowa

1. N. Marz, J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Publications Co., 2015. (lub tłumaczenie)
2. T. White, Hadoop. Kompletny przewodnik. Analiza i przechowywanie danych, Helion, 2015. (lub oryginał)
3. Matei Zaharia, Bill Chambers, Spark: The Definitive Guide, O'Reilly Media, 2018
4. M. Odersky, L. Spoon, B. Venners, Programming in Scala, 3rd edition, Artima Inc, 2016. (są dostępne legalne wersje online)
5. 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>)
6. Systemy baz danych. Kompletny podręcznik. Wydanie II, Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom

Uzupełniająca

1. S. Ryza, U. Lasersson, S. Owen, J. Wills, Spark. Zaawansowana analiza danych, Helion, 2015. (lub oryginał)
2. C. Horstmann, Scala for the Impatient, Addison-Wesley, 2016.
3. Hurtownie danych: logiczne i fizyczne struktury danych, Z. Królikowski, Wydawnictwo Politechniki Poznańskiej 2007
4. Hadoop in Action, Ch. Lam, , Manning Publications Co., 2011.
5. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, R. Kimball, M. Ross, John Wiley & Sons 2002
6. 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>)
7. Projektowanie hurtowni danych, Zarządzanie kontaktami z klientami (CRM), Ch. Todman, Wydawnictwa Naukowo-Techniczne 2003



Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	101	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	33	1.3
Udział w zajęciach laboratoryjnych/ćwiczeniach ¹	16	0.6
Dokończenie (w ramach pracy własnej) zadań z ćwiczeń laboratoryjnych	19	0.8
Zadanie domowe: 5 x 2 godz.	10	0.4
Udział w konsultacjach związanych z realizacją procesu kształcenia (częściowo mogą być realizowane drogą elektroniczną)	1	0.0
Przygotowanie do zajęć z obowiązkowymi zadaniami punktowanymi	19	0.8
Udział w wykładach	16	0.6
Zapoznanie się ze wskazaną literaturą i materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 100 stron	10	0.4
Przygotowanie do sprawdzianu zaliczeniowego wykład	10	0.4

¹niepotrzebne skreślić lub dopisać inne czynności