**Nazwa przedmiotu:**

Metody i systemy analizy Big Data

**Koordynator przedmiotu:**

Dr hab. inż. Maciej Grzenda, prof. PW

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Inżynieria i Analiza Danych

**Grupa przedmiotów:**

Wspólne

**Kod przedmiotu:**

1120-DS000-MSP-0112

**Semestr nominalny:**

2 / rok ak. 2019/2020

**Liczba punktów ECTS:**

4

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. godziny kontaktowe – 50 h; w tym
a) obecność na wykładach – 15 h
b) obecność na projekcie – 30 h
c) konsultacje – 5 h
2. praca własna studenta – 45 h; w tym
a) przygotowanie projektu – 25 h
b) zapoznanie się z literaturą – 10 h
c) przygotowanie do kolokwium końcowego – 10 h
Razem 95 h, co odpowiada 4 pkt. ECTS

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

a) obecność na wykładach – 15 h
b) obecność na projekcie – 30 h
c) konsultacje – 5 h
Razem 50 h, co odpowiada 2 pkt. ECTS

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

a) przygotowanie projektu – 25 h
b) obecność na projekcie – 30 h
Razem 55 h, co odpowiada 2 pkt. ECTS

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład:  | 15h |
| Ćwiczenia:  | 0h |
| Laboratorium:  | 0h |
| Projekt:  | 30h |
| Lekcje komputerowe:  | 0h |

**Wymagania wstępne:**

Znajomość relacyjnych baz danych i systemów składowania danych, umiejętność programowania w językach Java lub Python, podstawowa znajomość metod uczenia maszynowego.

**Limit liczby studentów:**

.

**Cel przedmiotu:**

Celem przedmiotu jest przekazanie wiedzy i pozyskanie przez studentów umiejętności związanych z przetwarzaniem i analizą danych wielkoskalowych, w tym wykorzystaniem metod uczenia maszynowego do analizy strumieni danych.

**Treści kształcenia:**

Wykład:
1. Porównanie systemów składowania danych Big Data, w tym takich jak Apache Hadoop z systemami relacyjnymi
2. Architektura systemów Big Data i rola analizy danych w tym analizy strumieniowej w tych systemach.
3. Programowanie filtrowania i agregacji danych w trybie wsadowym i strumieniowym
4. Przetwarzanie strumieni danych i kluczowe aspekty tego przetwarzania, w tym przetwarzanie w układzie okien czasowych oraz identyfikacja sesji
5. Uczenie maszynowe a przetwarzanie wsadowe i przetwarzanie strumieni danych
6. Uczenie maszynowe: metody i środowiska dedykowane dla przetwarzania danych wielkoskalowych
7. Zaawansowane zagadnienia uczenia maszynowego i analizy danych w trybie strumieniowym (ang. stream mining), w tym m.in. zmienność modelowanych procesów i problem opóźnionych etykiet
8. Wzorce projektowe i architektoniczne.
Projekt:
Celem zajęć projektowych będzie samodzielna analiza złożonego zagadnienia i przygotowanie systemu bazującego na gotowych systemach Big Data i dedykowanych komponentach. Realizacja projektów będzie uwzględniać następujące aspekty:
● Samodzielna analiza literatury przedmiotu związanej z tematem projektu
● Projektowanie architektury systemu.
● Programowanie zadań strumieniowego przetwarzania i analizy danych
● Uczenie maszynowe z wykorzystaniem danych wielkoskalowych
● Okresowa prezentacja wyników prac nad projektem

**Metody oceny:**

Zaliczenie przedmiotu oparte jest o wyniki realizacji projektu (60%) oraz kolokwium końcowe (40%). Maksymalna liczba dostępnych punktów wynosi 100. Ocena projektu jest sumą punktów wynikających z oceny realizacji poszczególnych etapów projektu. Wyniki oceny kolejnych etapów projektu oraz kolokwium są ogłaszane na stronie internetowej prowadzącego zajęcia w danej grupie projektowej lub rozsyłane do uczestników drogą mailową. Ocena końcowa zależy od łącznej liczby punktów uzyskanych z projektu oraz kolokwium końcowego i jest wyznaczana zgodnie z poniższymi regułami: 0-50 pkt – 2.0, 51-60 pkt – 3.0, 61-70 pkt – 3.5, 71-80 pkt – 4.0, 81-90 pkt – 4.5, 91-100 pkt – 5.0.
Do uzyskania pozytywnej oceny końcowej konieczne jest uzyskanie co najmniej 50% punktów z projektu i co najmniej 50% punktów z kolokwium końcowego.

**Egzamin:**

nie

**Literatura:**

1. Ankam, V., Big Data Analytics, Packt, 2016,
2. Ellis B., Real-Time Analytics. Techniques to Analyse and Visualise Streaming Data, Wiley, 2014
3. Kleppmann, M., Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O’Reilly, 2017
4. Marz N., Warren James, Big Data. Principles and best practices of scalable realtime data systems, Manning, 2015
5. Provost F., Facett T., Data Science for Business. What you need to know about data mining and data-analytic thinking, O’Reilly, 2013

**Witryna www przedmiotu:**

.

**Uwagi:**

program 4 semestralny - 2 semestr
program 3 semestralny - 1 semestr

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka W01:**

Dysponuje wiedzą na temat metod i algorytmów przetwarzania danych Big Data stosowanych w trybie wsadowym i strumieniowym

Weryfikacja:

kolokwium końcowe

**Powiązane charakterystyki kierunkowe:** DS2\_W01

**Powiązane charakterystyki obszarowe:** I.P7S\_WG

**Charakterystyka W02:**

Zna podstawowe środowiska przetwarzania danych Big Data w trybie wsadowym i strumieniowym

Weryfikacja:

kolokwium końcowe

**Powiązane charakterystyki kierunkowe:** DS2\_W02, DS2\_W05

**Powiązane charakterystyki obszarowe:** I.P7S\_WG

### Profil ogólnoakademicki - umiejętności

**Charakterystyka U01:**

Umie zaprojektować i wykonać komponenty stosowane do analizy danych w trybie wsadowym i strumieniowym

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_U01, DS2\_U14

**Powiązane charakterystyki obszarowe:** I.P7S\_UW

**Charakterystyka U02:**

Umie stworzyć zintegrowany system pozyskiwania i analizy danych, wykorzystujący zarówno uniwersalne, jak i dedykowane podsystemy i komponenty

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_U10, DS2\_U02, DS2\_U09

**Powiązane charakterystyki obszarowe:** I.P7S\_UW

**Charakterystyka U03:**

Umie zaprojektować i zrealizować podsystem analizy danych wykorzystujący metody uczenia maszynowego, z uwzględnieniem użycia zewnętrznych środowisk analizy danych

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_U01, DS2\_U04

**Powiązane charakterystyki obszarowe:** I.P7S\_UW

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka K01:**

Posiada świadomość potrzeby ciągłego kształcenia w dziedzinie systemów składowania i analizy danych

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_K01, DS2\_K07

**Powiązane charakterystyki obszarowe:** I.P7S\_KK

**Charakterystyka K02:**

Ma świadomość odpowiedzialności za wspólnie realizowane zadania na przykładzie zadania budowy złożonego systemu pozyskiwania i analizy danych przez zespół programistów

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_K04

**Powiązane charakterystyki obszarowe:** I.P7S\_KR

**Charakterystyka K03:**

Potrafi przekazać w sposób zrozumiały założenia i rozwiązania złożonego systemu przetwarzania i analizy danych

Weryfikacja:

Ocena punktowa etapów projektu

**Powiązane charakterystyki kierunkowe:** DS2\_K06

**Powiązane charakterystyki obszarowe:** I.P7S\_KO