**Nazwa przedmiotu:**

Składowanie danych w systemach Big Data

**Koordynator przedmiotu:**

Dr hab. inż. Maciej Grzenda, prof. PW

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia I stopnia

**Program:**

Inżynieria i Analiza Danych

**Grupa przedmiotów:**

Wspólne

**Kod przedmiotu:**

.

**Semestr nominalny:**

7 / rok ak. 2022/2023

**Liczba punktów ECTS:**

5

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

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

1. obecność na wykładach – 30 h
2. obecność na laboratoriach – 30 h
3. obecność na projektach – 15 h
4. konsultacje – 5 h
Razem 80 h, co odpowiada 3 pkt. ECTS

**Język prowadzenia zajęć:**

polski

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

1. obecność na laboratoriach – 30 h
2. obecność na projektach – 15 h
3. przygotowanie do laboratoriów – 40 h
4. przygotowanie wyników realizacji zadań – 45 h
Razem 130 h, co odpowiada 5 pkt. ECTS

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

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

**Wymagania wstępne:**

Bazy danych, Hurtownie danych i systemy Business Intelligence, Zaawansowane programowanie obiektowe i funkcyjne, Systemy operacyjne w inżynierii danych

**Limit liczby studentów:**

.

**Cel przedmiotu:**

Celem przedmiotu jest przekazanie wiedzy na temat architektury i przykłado-wych środowisk Big Data umożliwiających pozyskiwanie i składowanie danych oraz umiejętności pozyskiwania, transformacji i składowania danych w tych środowiskach.

**Treści kształcenia:**

Wykład:
1. Przegląd zagadnień oraz oprogramowania nurtu Big Data, w tym projektów takich jak Apache Hadoop
2. Architektura systemów Big Data: pozyskiwanie danych, składowanie danych, przekazywanie danych w rozproszonej architekturze systemu, analiza danych, w tym analiza z wykorzystaniem metod uczenia maszynowego.
3. Pozyskiwanie danych z różnorodnych źródeł danych z wykorzystaniem platform Big Data.
4. Przetwarzanie wsadowe a przetwarzanie strumieni danych.
5. Programowanie rozwiązań wykorzystujących platformy Big Data z uwzględnieniem m.in. użycia równoległego i rozproszonego przetwarzania danych oraz środowisk ułatwiających tworzenie kodu przetwarzania danych w trybie wsadowym i strumieniowym.
6. Wzorce projektowe i architektoniczne np. architektura Lambda i Kappa.
7. Programowanie filtrowania i agregacji danych
8. Dobór formatów i struktur danych dla składowanych danych
Laboratorium:
1. Konfiguracja pozyskiwania danych z różnorodnych źródeł.
2. Programowanie zadań pozyskiwania danych.
3. Programowanie przetwarzania wsadowego.
4. Projektowanie architektury złożonego systemu.

**Metody oceny:**

Zaliczenie przedmiotu oparte jest o wyniki realizacji zadań punktowanych w trakcie laboratorium (60%) oraz egzamin (40%). Maksymalna liczba dostępnych punktów wynosi 100.
Wyniki zadań punktowanych są ogłaszane na stronie internetowej prowadzącego zajęcia w danej grupie laboratoryjnej lub rozsyłane do uczestników drogą mailową. W końcowej części semestru student może skorzystać z terminu poprawkowego, w trakcie którego może poprawić jedno z zadań punktowanych:
- termin poprawkowy jest ogłaszany studentom z wyprzedzeniem co najmniej jednego tygodnia,
- każdy ze studentów ma zagwarantowaną możliwość udziału w jednym terminie poprawkowym, przy założeniu, iż był obecny w oryginalnym terminie zadania lub przedstawił w terminie dokument uzasadniający nieobecność w tym terminie,
- najpóźniej tydzień przed terminem poprawkowym, studenci zainteresowani poprawianiem zadania, zobligowani są zadeklarować, które z zadań chcieliby poprawiać w terminie poprawkowym,
- ocena zadania uzyskana w terminie poprawkowym zastępuje oryginalną ocenę z tego zadania.
Ocena końcowa zależy od łącznej liczby punktów uzyskanych z zadań punktowanych oraz egzaminu 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.

**Egzamin:**

tak

**Literatura:**

1. Ellis B., Real-Time Analytics. Techniques to Analyse and Visualise Streaming Data, Wiley, 2014
2. Holmes A., Hadoop in practice, Manning Publications,2013
3. Marz N., Warren James, Big Data. Principles and best practices of scalable realtime data systems, Manning, 2015
4. 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:**

.

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka W01:**

Dysponuje wiedzą na temat danych częściowo ustrukturyzowanych i kategorii danych umieszczanych w platformach Big Data oraz relacyjnych bazach danych

Weryfikacja:

Egzamin, zadania punktowane

**Powiązane charakterystyki kierunkowe:** DS\_W12

**Powiązane charakterystyki obszarowe:** I.P6S\_WG

### Profil ogólnoakademicki - umiejętności

**Charakterystyka U01:**

Umie budować rozproszone systemy pozyskiwania i składowania danych integrujące samodzielnie stworzone komponenty oraz komponenty platform Big Data

Weryfikacja:

Zadania punktowane

**Powiązane charakterystyki kierunkowe:** DS\_U18

**Powiązane charakterystyki obszarowe:** I.P6S\_UW

**Charakterystyka U02:**

Umie pozyskiwać dane z platform Big Data i innych źródeł, jak również po opcjonalnym filtrowaniu i transformacji umieszczać je w platformach Big Data

Weryfikacja:

Zadania punktowane

**Powiązane charakterystyki kierunkowe:** DS\_U22

**Powiązane charakterystyki obszarowe:** I.P6S\_UW

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka K01:**

Rozumie potrzebę uczenia się przez całe życie i podnoszenia kompetencji zawodowych na przykładzie rozwoju platform Big Data i ewolucji systemów składowania danych.

Weryfikacja:

egzamin

**Powiązane charakterystyki kierunkowe:** DS\_K01

**Powiązane charakterystyki obszarowe:** I.P6S\_KK