**Nazwa przedmiotu:**

Techniki i technologie Big Data

**Koordynator przedmiotu:**

Rajmund Kożuszek

**Status przedmiotu:**

Fakultatywny dowolnego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

TBD

**Semestr nominalny:**

3 / rok ak. 2021/2022

**Liczba punktów ECTS:**

4

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

1. liczba godzin kontaktowych – 47 godz.,w tym:
a. obecność na wykładach: 30 godz.,
b. udział w konsultacjach związanych z treścią wykładu: 2 godz.,
c. udział w spotkaniach projektowych: 15 godz.,
2. praca własna studenta – 59 godz., w tym:
a. przygotowanie do wykładów (przejrzenie materiałów z wykładu i dodatkowej literatury, próba rozwiązania ćwiczeń domowych sformułowanych na wykładzie): 7 godz.,
b. realizacja projektu: 8 godz. (zapoznanie się z literaturą i oprogramowaniem) + 32 godz. (wykonanie zadań projektowych) + 4 godz. (sporządzenie dokumentacji) = 44 godz.,
c. przygotowanie do kolokwiów: 8 godz.
Łączny nakład pracy studenta wynosi: 106 godz., co odpowiada 4 pkt. ECTS.

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

1,75 pkt. ECTS, co odpowiada 47 godz. kontaktowym

**Język prowadzenia zajęć:**

polski

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

2,25 pkt. ECTS, co odpowiada 60 godz. realizacji projektu

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

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

**Wymagania wstępne:**

 Umiejętność programowania w przynajmniej jednym z języków: Scala, Python, R, Java.
 Znajomość relacyjnych baz danych i język SQL.
 Podstawowa wiedza z zakresu uczenia maszynowego.

**Limit liczby studentów:**

45

**Cel przedmiotu:**

Przekazanie wiedzy na temat zagadnień związanych z przetwarzaniem danych w środowisku rozproszonym i wykorzystaniem narzędzi z obszaru Big Data.

**Treści kształcenia:**

Wykład:
1. Wprowadzenie do zagadnień przetwarzania Big Data. (2 godz). Wprowadzenie do przedmiotu, omówienie spraw organizacyjnych, cele zajęć oraz ich program. Omówienie przyczyn zmian w podejściu do retencji danych. Przedstawienie rysu historycznego metod składowania oraz analizy dużych zbiorów danych. Omówienie najważniejszych zagadnień związanych z Big Data: architektura lambda, przetwarzanie strumieniowe vs przetwarzanie batchowe, orkiestracja, serializacja i deserializacja danych, bazy NoSQL. Przedstawienie czołowych projektów z obszaru Big Data: Hadoop, Spark, Cassandra.
2. Podstawowe komponenty ekosystemu Hadoop: wprowadzenie do YARN i HDFS. (2 godziny)
3. Sposób organizacji danych: pojęcie jeziora danych (data lake), analityka oraz bazy klucz-wartość, serializacja i deserializacja danych, formaty (ORC, Parquet), Cassandra, HBase, pojęcie schematu danych i ewolucji na przykładzie Avro, wsparcie dla ACID na przykładzie DeltaLake czy Iceberg. Porównanie wydajności różnych konfiguracji dla rzeczywistych przypadków użycia, pochodzących z projektów badawczych (4 godziny)
4. Apache Spark. Omówienie koncepcji i zastosowań RDD (historycznie) i DataFrame. Architektura Spark (cluster manager, executor’y). Porównanie przetwarzania z Hadoop MapReduce oraz dyskusja dotycząca optymalizacji. Powiązanie z platformą Hadoop poprzez Resource Manger’a oraz wersja standalone (local). Omówienie API na podstawie przykładowego job’a. (4 godziny)
5. Analityka Big Data - SQL w środowisku Big Data (na przykładzie SparkSQL, Hive). Analiza danych z Hadoop za pomocą R i innych środowisk analitycznych (pyspark + jupyter). Przykłady potoków przetwarzania wykorzystywanych w projektach badawczych prowadzonych przez prowadzących (m.in. zastosowania w analizie danych z sekwencjonowania DNA). (4 godziny)
6. Wizualizacja danych - środowisko R/Python + narzędzia D3, Leaflet, Vega, deck.gl. Omówienie podstawowych pojęć i strategii wizualizacji danych z uwzględnieniem przede wszystkim danych ilościowych i danych geograficznych. (2 godziny)
7. Przetwarzanie strumieniowe. Wprowadzenie do narzędzi służących przetwarzaniu strumieni danych: Kafka, Spark Streaming, Apache Flink, Apache Beam. Przykład algorytmu strumieniowego z wykorzystaniem struktur danych takich jak: count min sketch, bloom filter. (4 godziny)
8. Big Data w chmurze. Przedstawienie architektur chmurowych, wirtualizacji i kontenerów, systemów zarządzania chmurą (OpenShift, K8S), Przykładowe osadzenie projektu big data w chmurze. Przedstawienie heterogenicznych środowisk obliczeniowych i zarządzania zasobami oraz ich izolacji za pomocą konteneryzacji. (2 godziny)
9. Zapewnienie bezpieczeństwa w środowisku rozproszonym - autentykacja(Kerberos) oraz scentralizowana autoryzacja dostępu do zasobów (Apache Ranger. Integracja z istniejącymi systemami bezpieczeństa, wykorzystanie impersonacji uzytkowników, bezpieczeństwo w środowisku kontenerowym. (2 godziny)
10. Uczenie maszynowe w środowiskach rozproszonych z wykorzystaniem bilbiotek TensorFlow. Wprowadzenie do uczenia maszynowego na przykładzie sieci neuronowych oraz głębokich sieci neuronowych. Wprowadzenie do TensorFlow oraz przykłady implementacji rozproszonej m.in. w oparciu o rozwiązania chmurowe. (4 godziny)

**Metody oceny:**

Realizacja przedmiotu obejmuje następujące formy zajęć:
 wykład prowadzony w wymiarze 2 godz. tygodniowo,
 interaktywna formuła wykładu,
 projekt realizowany w zespołach,
 wymóg konsultacji interpretacji tematu i zakresu projektu
 wymóg przedstawienia do oceny wstępnej dokumentacji projektu
 w ramach projektu studenci będą mieli możliwość rozwiązywać, rzeczywiste problemy badawcze, występujące m. in. w badaniach biomedycznych.
 autorzy najbardziej wartościowych projektów, będą mieli możliwość kontynuacji pracy z prowadzącymi i wspólnego opracowania wyników w formie artykułu. Publikacja będzie mogła być zaliczona do Priorytetowego Obszaru Badawczego: Cyberbezpieczeństwo i Analiza Danych.
 konsultacje z prowadzącymi.
Sprawdzanie założonych efektów kształcenia realizowane jest przez:
 ocenę wiedzy i umiejętności wykazanych na kolokwium,
 ocenę wiedzy i umiejętności związanych z realizacją zadań projektowych – ocena wykonanych prac implementacyjnych i jakości dokumentacji.

**Egzamin:**

nie

**Literatura:**

1. Hadoop. Komplety przewodnik. Analiza i przechowywanie danych, Tom White, O’Reilly/Helion, 2015
2. Designing Data Intensive Applications, Martin Kleppman, O’Reilly, 2014
3. High Performance Spark, Holden Karau, O’Reilly, 2016
4. Edward Tufte, The Visual Display of Quantitative Information, Graphics Press, 2001
5. David McCandless, Knowledge is beautiful, Harper Collins, 2014

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103A-INSZI-MSP-TBD

**Uwagi:**

(-)

## Charakterystyki przedmiotowe