POLITECHNIKA GDAŃSKA

Karta przedmiotu

Nazwa i kod przedmiotu	Big data, PG_00068194		
Kierunek studiów	Automatyka, cybernetyka i robotyka		
Data rozpoczęcia studiów	październik 2026 r.	Rok akademicki realizacji przedmiotu	2029/2030
Poziom kształcenia	I stopnia - inżynierskie	Grupa zajęć	Grupa zajęć fakultatywnych Grupa zajęć powiązanych z prowadzonymi badaniami naukowymi w dziedzinie nauki związanej z kierunkiem - profil ogólnoakademicki
Forma studiów	stacjonarne	Sposób realizacji	na uczelni
Rok studiów	4	Język wykładowy	polski
Semestr studiów	7	Liczba punktów ECTS	3.0
Profil kształcenia	ogólnoakademicki	Forma zaliczenia	zaliczenie
Jednostka prowadząca	Wydziały Politechniki Gdańskiej -> Wydział Elektroniki, Telekomunikacji i Informatyki -> Katedra Systemów Decyzyjnych i Robotyki		
Imię i nazwisko wykładowcy (wykładowców)	Odpowiedzialny za przedmiot	dr inż. Jakub Wszołek	
	Prowadzący zajęcia z przedmiotu	dr inż. Jakub Wszołek	

Formy zajęć							
	Forma zajęć	Wykład	Ćwiczenia	Laboratorium	Projekt	Seminarium	RAZEM
	Liczba godzin zajęć	15.0	0.0	15.0	15.0	0.0	45
	W tym liczba godzin zajęć na odległość: 0.0						

Aktywność studenta i liczba godzin pracy					
	Aktywność studenta	Udział w zajęciach dydaktycznych, objętych planem studiów	Udział w konsultacjach	Praca własna studenta	RAZEM
	Liczba godzin pracy studenta	45	3.0	27.0	75

Cel przedmiotu

Przedmiot wprowadza studentów w zagadnienia związane z przetwarzaniem, składowaniem i analizą dużych, rozproszonych zbiorów danych. Obejmuje podstawy architektury systemów Big Data, paradygmaty programowania równoległego, modelowanie danych oraz praktyczne wykorzystanie narzędzi wykorzystywanych we współczesnych systemach analitycznych i przetwarzania strumieniowego.

Studenci poznają zarówno aspekty teoretyczne, jak i praktyczne, przygotowując się do pracy w środowiskach enterprise i cloud, takich jak Hadoop, Spark, systemy Lakehouse oraz nowoczesne podejścia do budowy platform danych.

Efekty uczenia się przedmiotu

Efekt kierunkowy	Efekt z przedmiotu	Sposób weryfikacji i oceny efektu
[K6_W21] zna i rozumie podstawowe metody podejmowania decyzji oraz metody i techniki projektowania i eksploatacji systemów regulacji automatycznej i sterowania, zastosowania komputerów do sterowania i monitorowania systemów dynamicznych.	Student zna i rozumie podstawowe metody projektowania oraz realizacji systemów przetwarzania dużych zbiorów danych, w tym architekturę systemów rozproszonych, modele obliczeń równoległych oraz techniki realizacji procesów ETL/ELT.	[SW3] Ocena wiedzy zawartej w opracowaniu tekstowym i projektowym
[K6_U07] potrafi wykorzystać metody wspomagania procesów i funkcji, specyficzne dla kierunków studiów	Student potrafi wykorzystać narzędzia oraz platformy Big Data do realizacji procesów analitycznych, przetwarzania strumieniowego oraz budowy pipeline'ów danych, zgodnie ze specyfiką kierunku studiów.	[SU3] Ocena umiejętności wykorzystania wiedzy uzyskanej w ramach przedmiotu
[K6_U08] potrafi przy identyfikacji i formułowaniu specyfikacji zadań inżynierskich związanych z kierunkiem studiów oraz ich rozwiązywaniu: − wykorzystać metody analityczne, symulacyjne i eksperymentalne, − dostrzegać ich aspekty systemowe i pozatechniczne, − dokonać wstępnej oceny ekonomicznej proponowanych rozwiązań i podejmowanych działań inżynierskich	Student potrafi zastosować metody analityczne oraz narzędzia Big Data do formułowania i rozwiązywania problemów inżynierskich, a także oceniać efektywność i ograniczenia proponowanych rozwiązań.	[SU4] Ocena umiejętności korzystania z metod i narzędzi

Treści przedmiotu

Treści przedmiotu - wykład

1. Wprowadzenie do Big Data definicje, charakterystyka danych, 5V.
2. Architektura systemów rozproszonych modele obliczeń, skalowalność, odporność na awarie.
3. Ekosystem Hadoop HDFS, MapReduce, YARN.
4. Apache Spark architektura, model wykonania, RDD vs DataFrame.
5. Spark SQL optymalizacja zapytań, Catalyst, Tungsten.
6. Format danych w systemach Big Data Parquet, ORC, Avro; kompresja i kolumnowość.
7. Delta Lake transakcyjność, ACID, time travel, zarządzanie metadanymi.
8. Przetwarzanie strumieniowe Spark Structured Streaming, źródła, zlewy, okna czasowe.
9. Apache Kafka architektura, partycjonowanie, kolejki komunikatów, integracja ze Spark.
10. Projektowanie i budowa pipelineów ETL/ELT.
11. Architektury danych: Data Lake, Data Warehouse, Lakehouse, podstawy Data Mesh.
12. Bezpieczeństwo danych kontrola dostępu, audit, jakość danych, linie rodowodowe.
13. Skalowanie i optymalizacja systemów Big Data partycjonowanie, cache, shuffle.
14. Współczesne platformy danych (cloud + on-prem) przegląd rozwiązań.

Treści przedmiotu - laboratoria

1. Środowisko pracy konfiguracja klastra Spark lokalnie lub na serwerze.
2. Wczytywanie i eksploracja dużych zbiorów danych (CSV, JSON, Parquet).
3. Praca z RDD, DataFrame oraz Spark SQL.
4. Transformacje danych, agregacje, funkcje okienkowe.
5. Optymalizacja zapytań i analiza planów wykonania (explain).
6. Implementacja procesów ETL/ELT w Spark.
7. Praca z formatami kolumnowymi i Delta Lake tworzenie tabel, operacje ACID.
8. Przetwarzanie strumieniowe budowa prostego pipelineu streamingowego.
9. Integracja Spark Structured Streaming z Apache Kafka.
10. Walidacja jakości danych i monitoring pipelineów.

Treści przedmiotu - projekt

1. Opracowanie koncepcji systemu przetwarzania dużych zbiorów danych (problem + architektura).
2. Zaprojektowanie i implementacja pipelineu ETL/ELT na wybranej platformie Big Data (np. Spark + Delta Lake).
3. Budowa i konfiguracja środowiska projektowego (lokalnie lub w chmurze).
4. Integracja danych z wielu źródeł (batch i/lub streaming).
5. Implementacja logiki analitycznej agregacje, raporty, modele statystyczne lub ML (opcjonalnie).
6. Dokumentacja architektury, metadanych i wyników.
7. Prezentacja i obrona projektu.

Wymagania wstępne i dodatkowe

Student powinien posiadać podstawową wiedzę z zakresu programowania (preferowane Python lub Java), podstaw algorytmiki oraz znajomość zasad działania systemów operacyjnych i sieci komputerowych. Wskazana jest również podstawowa znajomość baz danych (SQL) oraz umiejętność pracy w środowisku Linux. Mile widziane doświadczenie w analizie danych lub pracy z większymi zbiorami danych.

<table>
<tr><td>Sposoby i kryteria oceniania osiąganych efektów uczenia się</td><td colspan="2"><table>
<tr><th>Sposób oceniania (składowe)</th><th>Próg zaliczeniowy</th><th>Składowa oceny końcowej</th></tr>
<tr><td>Kolokwium sprawdzające wiedzę teoretyczną z zakresu architektury systemów Big Data, przetwarzania rozproszonego i narzędzi (Hadoop, Spark, Delta Lake).</td><td>50.0%</td><td>40.0%</td></tr>
<tr><td>Ocena wykonania ćwiczeń laboratoryjnych obejmujących implementację zadań w Apache Spark, analizę danych, pracę z formatami kolumnowymi oraz przetwarzanie strumieniowe.</td><td>50.0%</td><td>30.0%</td></tr>
<tr><td>Ocena projektu zespołowego obejmującego zaprojektowanie i implementację pipeline’u Big Data (ETL/ELT), integrację danych oraz prezentację wyników.</td><td>50.0%</td><td>30.0%</td></tr>
</table></td></tr>
<tr><td>Zalecana lista lektur</td><td>Podstawowa lista lektur</td><td>1. Tom White, Hadoop: Definitywne wprowadzenie (Hadoop: The Definitive Guide), OReilly.
2. Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee, Spark: Definitywne wprowadzenie (Spark: The Definitive Guide), OReilly.
3. Bill Chambers, Matei Zaharia, Learning Spark: Lightning-Fast Big Data Analysis, OReilly.
4. Dokumentacja projektów Apache Spark, Hadoop, Delta Lake (materiały online).</td></tr>
<tr><td></td><td>Uzupełniająca lista lektur</td><td>1. Martin Kleppmann, Projektowanie systemów odpornych na błędy. Przetwarzanie danych na dużą skalę
2. (Designing Data-Intensive Applications), Helion.
3. Tyler Akidau, Slava Chernyak, Przetwarzanie strumieniowe. Podejście praktyczne
4. (Streaming Systems), OReilly.
5. Ben Stopford, Kafka: wzorce architektoniczne
6. (Designing Event-Driven Systems), OReilly.
7. Andrew Psaltis, Streaming Data: Understanding the Real-Time Pipeline, Manning.
8. Zbiory danych publicznych (np. Kaggle, UCI ML Repository) materiały do projektów.
9. Dokumentacje online: Apache Kafka, Apache Airflow, Delta Lake, Spark Streaming.</td></tr>
<tr><td></td><td>Adresy eZasobów</td><td></td></tr>
<tr><td>Przykładowe zagadnienia/ przykładowe pytania/ realizowane zadania</td><td colspan="2">• Omów architekturę Apache Spark i wyjaśnij rolę drivera, executorów oraz klastra.
• Wyjaśnij różnice między RDD, DataFrame i Dataset w Spark.
• Porównaj formaty danych Parquet, ORC i Avro. W jakich przypadkach są najbardziej efektywne?
• Omów sposób działania Apache Kafka, w tym rolę partycji i offsetów.
• Opisz proces optymalizacji zapytań w Spark SQL (Catalyst, Tungsten).
• Zaimplementuj pipeline ETL z wykorzystaniem Spark DataFrame API.
• Przeprowadź analizę dużego zbioru danych z użyciem funkcji okienkowych.
• Zbuduj prostą aplikację streamingową (Spark Structured Streaming + Kafka).
• Przeprowadź operacje ACID w Delta Lake i pokaż zastosowanie time travel.
• Zaprojektuj architekturę systemu Big Data dla podanego przypadku biznesowego.</td></tr>
<tr><td>Zajęcia praktyczne w ramach przedmiotu</td><td colspan="2">Nie dotyczy</td></tr>
</table>

Dokument wygenerowany elektronicznie. Nie wymaga pieczęci ani podpisu.