



KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Przetwarzanie strumieni danych w systemach Big Data [S2Inf1>PSD]

Przedmiot

Kierunek studiów
Informatyka

Rok/Semestr
1/1

Studia w zakresie (specjalność)
Technologie przetwarzania danych

Profil studiów
ogólnoakademicki

Poziom studiów
drugiego stopnia

Język oferowanego przedmiotu
polski

Forma studiów
stacjonarne

Wymagalność
obligatoryjny

Liczba godzin

Wykład	Laboratorium	Inne
30	30	0
Ćwiczenia	Projekty/seminaria	
0	0	

Liczba punktów ECTS

5,00

Koordynatorzy

dr inż. Krzysztof Jankiewicz
krzysztof.jankiewicz@put.poznan.pl

Wykładowcy

dr inż. Krzysztof Jankiewicz
krzysztof.jankiewicz@put.poznan.pl

Wymagania wstępne

Student rozpoczynający przedmiot Przetwarzanie strumieni danych w systemach Big Data powinien posiadać podstawową wiedzę z zakresu kształcenia ze studiów I stopnia zdefiniowanych w Uchwale Senatu PP weryfikowane w procesie rekrutacji na studia 2 stopnia, efekty te prezentowane są w serwisie internetowym wydziału www.cat.put.poznan.pl. W szczególności student rozpoczynający przedmiot powinien posiadać podstawową wiedzę z zakresu systemów operacyjnych, przetwarzania rozproszonego, sieci komputerowych, relacyjnych systemów baz danych oraz języka SQL i obiektowych języków programowania, a także systemów przetwarzania masywnych danych (Big Data) w zakresie wsadowego przetwarzania danych. Ponadto, student powinien posiadać także umiejętność pozyskiwania informacji ze wskazanych źródeł, jak również rozumieć konieczność poszerzania swoich kompetencji i mieć gotowość do podjęcia współpracy w ramach zespołu.

Cel przedmiotu

Przekazanie studentom podstawowej wiedzy związanej z wyzwaniem przetwarzania strumieni danych, w tym przetwarzania strumieni masywnych danych w zakresie prezentacji teoretycznych i praktycznych aspektów konstrukcji systemów przetwarzania strumieni masywnych danych oraz wyzwań związanych z organizacją, zarządzaniem i przetwarzaniem strumieni masywnych danych. Rozwijanie u studentów umiejętności rozwiązywania problemów przetwarzania strumieni masywnych danych w systemach rozproszonych dużej skali.

Przedmiotowe efekty uczenia się

Wiedza:

1. ma zaawansowaną wiedzę szczegółową dotyczącą wybranych zagadnień z zakresu informatyki takich jak: architektura i klasyfikacja systemów przetwarzania strumieni masywnych danych, narzędzia programowania w środowiskach przetwarzania strumieni masywnych danych [k2st_w3]
2. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach informatyki i innych, wybranych, pokrewnych dyscyplin naukowych w zakresie przetwarzania strumieni masywnych danych [k2st_w4]
3. ma zaawansowaną i szczegółową wiedzę o procesach zachodzących w cyklu życia systemów informatycznych sprzętowych lub programowych [k2st_w5]
4. zna zaawansowane metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich i prowadzeniu prac badawczych w zakresie przetwarzania strumieni masywnych danych [k2st_w6]

Umiejętności:

1. potrafi pozyskiwać informacje z literatury, baz danych oraz innych źródeł (w języku polskim i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie [k2st_u1]
2. potrafi planować i przeprowadzać eksperymenty, w tym pomiary i symulacje komputerowe, interpretować uzyskane wyniki i wyciągać wnioski oraz formułować i weryfikować hipotezy związane ze złożonymi problemami inżynierskimi i prostymi problemami badawczymi przetwarzania strumieni masywnych danych [k2st_u3]
3. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych związanych z przetwarzaniem strumieni masywnych danych, metody analityczne, symulacyjne oraz eksperymentalne [k2st_u4]
4. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć (metod i narzędzi) oraz nowych produktów informatycznych w kontekście przetwarzania strumieni masywnych danych [k2st_u6]
5. potrafi - stosując m.in. koncepcyjnie nowe metody - rozwiązywać złożone zadania przetwarzania strumieni masywnych danych, w tym zadania nietypowe oraz zadania zawierające komponent badawczy [k2st_u10]

Kompetencje społeczne:

1. rozumie, że w informatyce wiedza i umiejętności związane z przetwarzaniem strumieni masywnych danych bardzo szybko stają się przestarzałe [k2st_k1]
2. rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu przetwarzania strumieni masywnych danych w rozwiązywaniu problemów badawczych i praktycznych [k2st_k2]

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się 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 - 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 w odpowiedziach na pytania o różnej charakterystyce i złożoności problemów do rozwiązania (proste zadania dotyczące wiedzy podstawowej, zadania trudniejsze wymagające obliczeń, zadania problemowe o dużej złożoności), które pojawią się w ramach wykładowego

sprawdzianu zaliczeniowego; sprawdzian musi być zaliczony na co najmniej 50% możliwych do zdobycia punktów; ocena końcowa wynika ze średniej oceny wykładowego sprawdzianu zaliczeniowego oraz oceny z laboratorium.

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez ocenę realizacji zadań związanych z bieżącymi zajęciami laboratoryjnymi oraz znajomości poruszanych w nich zagadnień; podczas każdego zajęcia laboratoryjnego 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

Program obejmuje: wprowadzenie do przetwarzania strumieni danych, podstawy scentralizowanych systemów przetwarzania strumieni danych, omówienie systemów i zagadnień związanych z przetwarzaniem strumieni masywnych danych, systemy wymiany wiadomości, systemy zarządzania przetwarzaniem strumieni danych.

Tematyka zajęć

Program wykładu obejmuje następujące zagadnienia:

1. Przedstawienie wyzwań związanych z przetwarzaniem strumieni danych oraz strumieni masywnych danych: źródła strumieni danych, definicje strumieni danych, aspekty przetwarzania strumieni danych.
2. Wprowadzenie do systemów przetwarzania strumieni danych, podstawowe pojęcia, architektura systemów, poziomy interfejsów programistycznych na przykładzie scentralizowanych platform Esper i Oracle.
3. Wprowadzenie do systemów przetwarzania strumieni masywnych danych, pozyskiwanie strumieni danych, kolejkowe systemy wymiany wiadomości, rozwiązania publisher/subscriber na przykładzie platformy Apache Kafka.
4. Generacje systemów przetwarzania strumieni masywnych danych, obsługa znaczników czasowych zdarzeń, danych nieuporządkowanych, danych spóźnionych, przetwarzanie stanowe danych, wyzwalacze, na przykładzie bibliotek Kafka Streams oraz Spark Structured Streaming.
5. Przetwarzanie strumieni masywnych danych za pomocą interfejsów programistycznych wysokiego poziomu, Table API, Complex Event Processing na przykładzie platformy Apache Flink.
6. Platformy do budowy deklaracyjnych systemów przepływów danych z obsługą strumieni masywnych danych, architektury, pojęcia, koncepcje, na przykładzie Apache NiFi

Program laboratorium obejmuje następujące zagadnienia:

1. Zapoznanie się ze środowiskami wykorzystywanymi na laboratoriach - instalacja, konfiguracja, interfejs programistyczny, typy danych, podstawowe operacje dostępne w danym systemie.
2. Praktyczne wykorzystanie wybranych platform przetwarzania strumieni danych oraz przetwarzania strumieni masywnych danych:
 - realizacja zadań w narzędziach Esper oraz Oracle
 - realizacja zadań w środowisku Apache Kafka
 - realizacja zadań w środowisku Apache Spark
 - realizacja zadań w środowisku Apache Flink
 - realizacja zadań w środowisku Apache NiFi

Metody dydaktyczne

1. Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy a także z wykorzystaniem notatników interaktywnych i programowania "na żywo".
2. Zajęcia laboratoryjne: dyskusja, warsztaty, ćwiczenia praktyczne, praca w zespole.

Literatura

Podstawowa

1. M. Zaharia, B. Chambers, Spark: The Definitive Guide, O'Reilly Media, 2018
2. Tyler Akidau, Slava Chernyak, Reuven Lax, Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing, O'Reilly, 2018
3. Fabian Hueske, Vasiliki Kalavri, Stream Processing with Apache Flink. Fundamentals, Implementation,

and Operation of Streaming Applications, O'Reilly Media, 2019

4. Gwen Shapira, Todd Palino, Rajini Sivaram, Krit Petty, Kafka - The Definitive Guide: Real-time data and stream processing at scale, O'Reilly Media; Wydanie II, 2022

5. A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2012 (podręcznik dostępny w wersji elektronicznej: <http://infolab.stanford.edu/~ullman/mmds.html>)

6. P. Sadalage, M. Flower, NoSQL distilled, Addison-Wesley, 2013

Uzupełniająca

1. S. Ryza, U. Lasersson, S. Owen, J. Wills, Spark. Zaawansowana analiza danych, Helion, 2015

2. J. S. Damji et al., Learning Spark - Lightning-Fast Data Analytics, O'Reilly Media, 2020

3. A. Kobusińska, C. Leung, C.-H. Hsu, S. Raghavendra, V. Chang, Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing, Future Generation Computer Systems, 87, 2018

4. Dokumentacja systemów/platform wykorzystywanych w ramach kursu dostępna on-line

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	125	5,00
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	60	2,50
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu)	65	2,50