Big Data [223090-0421]

Prowadzący: **dr Mariusz Rafało** – `http://mariuszrafalo.pl`

Plan zajęć

1. Wprowadzenie do tematyki Big Data. Architektura Big data. Wybrane komponenty
2. Przetwarzanie rozproszone: koncepcja, przykłady i zastosowania
3. Wybrane komponenty ekosystemu Big Data (technologie)
4. Analizowanie danych pozbawionych struktury
5. Analityka na platformie Big Data. Integracja ekosystemu Big Data z hurtownią danych
6. Analizowanie danych w czasie rzeczywistym
7. Wybrane zagadnienia związane z etyką i prywatnością danych

Literatura

1. O'Reilly Media Inc., Big Data Now: 2012 Edition, O'Reilly Media
2. Provost, F. i Fawcett, T., Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly and Associates
3. Schutt, R. i O'Neil, C., Doing data science, O'Reilly
4. Minelli, M. Big Data, big analytics: emerging business intelligence and analytic trends for today's businesses, John Wiley and Sons, Inc.
5. Prajapati, V., Big Data Analytics with R and Hadoop, Packt Publishing Ltd.

Zasady zaliczenia przedmiotu

Zaliczenie opiera się na następujących zasadach:

- Egzamin w formie testu - należy zliczyć go na ocenę min. dostateczną
- Projekt zaliczeniowy - należy zaliczyć na ocenę min. dostateczną
- Punktacja egaminu, zgodnie z poniższa tabelą:

Od	Do	Ocena
0	59	niedostateczny
60	67	dostateczny
68	75	dostateczny plus
76	84	dobry
85	92	dobry plus
93	100	bardzo dobry

- Całkowita ocena jest średnią arytmetyczną ocen z egzminu i z projektu
- Praca podczas zajęć: każda nieobecność skutkuje obniżeniem całkowitej oceny o 0.5 (pół oceny).

Egzamin

Test wyboru (4 możliwe warianty odpowiedzi), składający się z około 25 pytań.

Zaliczenie w formie projektu

Dane

- Dane pobieramy z ogólnodostępnego repozytorium zbiorów danych, przykładowo Kaggle: `https://www.kaggle.com/datasets`
- Najlepiej gdyby dane dotyczyły działalności biznesowej, życia społecznego lub podobnych
- W oparciu o dane definiujemy problem, który chcemy zbadać, np.: analiza bankowych transakcji marketingu bezpośredniego, analiza uwarunkowań zarobków pracowników w różnych krajach, itp.

- Pracujemy na platformie *Databricks* (`https://community.cloud.databricks.com`). Należy założyć sobie konto na tej platformie.
- Pracujemy w języku *Python* i *SQL*; pracę dokumentujemy w *Markdown*
- Korzystamy z technologii poznanych na zajeciach: *Databricks*, *Hive*, *Spark*, *Kafka*

Transformacja danych

Analiza eksploracyjna powinna obejmować minimum elementy:

- Załadowanie danych do *dataframe* (format danych dowolny: CSV, JSON, strumień, itp.)
- Zapisanie danych w bazie danych Hive
- Wykonanie agregacji i transformacji danych w SQL
- Wykonanie agregacji i transformacji danych w Python
- Przedstawienie wyników na kilku wykresach

Raport

- Raport przygotowujemy w *Databricks* (eksport do HTML) i przesyłamy mailem
- Struktura raportu:
  - Źródło danych (opis, informacje o źródle)
  - Podstawowe transformacje danych (agregowanie, wyznaczenie wskaźników i kalkulacji)
  - Analiza eksploracyjna danych (wykresy i tabele)
  - Podsumowanie