## Hadoop i Spark

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

## Plan zajęć

1. Wprowadzenie do ekosystemu Apache Hadoop Techniki i technologie przetwarzania danych
2. Wprowadzenie do platformy Databricks
3. Wybrane technologie ekosystemu Big Data
4. Formaty plików w ekosystemie Apache Hadoop
5. Przetwarzanie danych w czasie rzeczywistym
6. Wprowadzenie do technologii Apache Kafka
7. Aplikacja: integracja Kafka i Spark Streaming

## Literatura


1. Spark. Zaawansowana analiza danych, S. Ryza, U. Laserson, S. Owen, J. Wills, Helion, 2015
2. Big Data: A Revolution That Will Transform How We Live, Work, and Think, V. Mayer-Schönberger, K. Cukier, Eamon Dolan/Mariner Books, 2014
3. Schutt, R. i O'Neil, C., Doing data science, O'Reilly
4. Mastering Apache Spark, M. Frampton, Packt Publishing Ltd., 2017
5. Prajapati, V., Big Data Analytics with R and Hadoop, Packt Publishing Ltd.


## Projekt zaliczeniowy

Podstawą zaliczenia jest projekt wykonany w technologiach Big Data.

**Dane**

- Dane pobieramy z ogólnodostępnego repozytorium zbiorów danych, przykładowo UCI: `http://mlr.cs.umass.edu/ml/datasets.html`
- 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: *HDFS*, *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
- Pobieranie danych w czasie rzeczywistym
- 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