Politechnika Warszawska

Wydział Matematyki i Nauk
Informacyjnych

# Metody Data Science

## Moduł pozyskiwania danych

*Autorzy:*

Piotr Izert
Przemysław Rząd
Anna Zawadzka

9 kwietnia 2016

# 1 Wstęp

Niniejszy projekt ma na celu wykonanie uczącego się systemu, który na podstawie pobieranych danych po odpowiednim ich przetworzeniu przygotowuje analizę w postaci prognozowania oraz klasyfikacji. Pierwszym etapem projektu jest stworzenie modułu pozyskiwania danych.

# 2 Dane

Źródłem danych są zbiory udostepnione przez ministerstwo transportu rządu Wielkiej Brytanii na stronach internetowych:

- `https://data.gov.uk/dataset/road-accidents-safety-data`
- `https://data.gov.uk/dataset/dft-eng-srn-routes-journey-times`

Dane obejmują informacje na temat wypadków drogowych na terenie Wielkiej Brytanii oraz średnich prędkościach przejazdu samochodów na danych odcinkach dróg w latach 2009-2014. Przechowywane są w plikach o formacie CSV. Dane pobierane są jako zamknięty zbiór rekordów, natomiast wykorzystane będą jako dane napływające w czasie rzeczywistym.

# 3 Moduł pozyskiwania danych

Użyte narzędzia:

- maszyna wirtualna Hortonworks Sandbox (HDP 2.4)
- Flume
- Spark
- HDFS

W katalogu *rawData* na maszynie wirtualnej umieszczone zostaną pliki w formacie CSV pięciu kategorii: wypadki, samochody, marki i modele samochodów, ofiary oraz średnie prędkości na poszczególnych odcinkach dróg. Dane o wypadkach, ofiarach, samochodach oraz ich markach i modelach pogrupowane są w pliki ze względu na rok, natomiast dla danych o prędkościach istnieją osobne pliki dla każdego miesiąca w danym roku.

Następnie za pomocą skryptu napisanego w języku Python zostaną połączone pliki z danymi o wypadkach, ofiarach, samochodach oraz ich markach i modelach dla każdego roku (za pomocą instrukcji SQL). Pliki wynikowe (osobne dla każdego roku) będą zawierać dane o wypadkach, samochodach w nich uczestniczących wraz z informacją o markach i modelach oraz ofiarach tychże wypadków.

Do każdego wypadku przyporządkowanych jest zawsze kilka samochodów, natomiast nie dla wszystkich pojazdów istnieje informacja o ofiarach (poszkodowani przyporządkowani są do konkretnego pojazdu) oraz o marce i modelu.

Nowo utworzone pliki zostaną automatycznie skopiowane do katalogu *spoolDir1* (*spooling directory*), który jest specjalnym katalogiem monitorowanym przez Flume'a. Dane o prędkościach przejazdów będą automatycznie umieszczane w katalogu *spoolDir2* bez żadnych zmian. Jeżeli w katalogach *spooling* pojawiają się w nim nowe pliki, Flume rozpoczyna ich przetwarzanie.

Konfiguracja Flume'a zakłada istnienie dwóch źródeł (*source*), którymi są *spoolDir1* i *spoolDir2*, oraz czterech ujść (*sink*), po dwa dla HDFS'a i dla Spark'a, odpowiednio dla skonsolidowanych danych o wypadkach oraz dla danych o prędkościach przejazdów. Ujścia skierowane do HDFS'a odwoływać się będą do katalogów *Accidents* oraz *Speeds*, natomiast ujścia do Spark'a odwołują się do odpowiedniego Job'a, który przetworzy dane w sposób strumieniowy.

W systemie HDFS istnieje dodatkowo katalog *Dictionaries*, w którym umieszczone są pliki słownikowe, wykorzystywane w dalszych etapach projektu. Pliki w tym katalogu umieszczone będą ręcznie i jednorazowo.

Dane ze wszystkich katalogów w HDFS (*Accidents*, *Speeds*, *Dictionaries*) pobierane będą przez Job'a (Spark) w sposób wsadowy i używane do generowania prognoz.

Przetwarzanie strumieniowe także korzystać będzie z danych znajdujących się w HDFS w katalogu Dictionaries i będzie realizować zadanie klasyfikacji.