# MapReduce, Hadoop Streaming, Hive
# Zestaw 2 – nyc-taxi

Pochodzenie danych to `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`
Uwaga! Dane pobieramy z miejsca wskazanego w opisie projektu

## Dwa zbiory danych

1. `datasource1` – informacje na przejazdów taksówek (1)

Dane mają format CSV. Pliki nie posiadają nagłówka.

Pola w pliku:

- Kod wskazujący dostawcę TPEP, który dostarczył rekord. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
- Data i godzina włączenia taksometru.
- Data i godzina wyłączenia taksometru.
- Liczba pasażerów w pojeździe. Jest to wartość wprowadzona przez kierowcę.
- Odległość, w milach, podana przez taksometr.
- Ostateczny kod stawki obowiązujący na koniec podróży. 1= stawka standardowa 2=JFK 3=Newark 4=Nassau lub Westchester 5=Taryfa do negocjacji 6=Jazda grupowa
- Wskazane czy rekord podróży był przechowywany w pamięci pojazdu przed wysłaniem do dostawcy, czyli "store and forward", ponieważ pojazd nie miał połączenia z serwerem. Y = tak N = nie
- TLC Taxi Zone, w której taksometr był włączony – `PULocationID`
- Strefa taksówek TLC, w której taksometr został wyłączony – `DOLocationID`
- Kod numeryczny oznaczający sposób zapłaty za podróż przez pasażera. 1=Karta kredytowa, 2=Gotówka, 3=Bez opłat, 4=Spór, 5=Nieznany, 6=Przejazd anulowany
- Opłata za przejazd czasowo-dystansowy obliczona przez licznik. Dodatkowe dodatki i dopłaty. Obecnie obejmuje to tylko opłaty w godzinach szczytu w wysokości 0,50 USD i 1 USD oraz opłaty za nocleg.
- Podatek MTA w wysokości 0,50 USD, który jest uruchamiany automatycznie na podstawie stawki licznika w użyciu.
- Kwota napiwku — to pole jest automatycznie wypełniane w przypadku napiwków zapłaconych za pomocą kart kredytowych. Napiwki gotówkowe nie są uwzględniane.
- Łączna kwota wszystkich opłat drogowych uiszczanych podczas podróży.
- $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
- Całkowita kwota pobierana od pasażerów. Nie obejmuje napiwków gotówkowych.

2. `datasource4` – informacje na temat stref taksówek (4)

Dane mają format CSV, każdy z plików ma wiersz nagłówka.

Pola w pliku

- `LocationID` – identyfikator lokalizacji TLC Taxi Zone
- `Borough` – dzielnica
- `Zone` – nazwa strefy taksówek
- `service_zone` – nazwa strefy usługowej

Krzysztof Jankiewicz

## Program MapReduce (2)

Działając na zbiorze `datasource1` należy dla każdego miesiąca (w określonym roku) wyznaczyć liczbę pasażerów, którzy wsiedli do taksówek w poszczególnych strefach (strefa włączenia taksometru) i opłacili swój przejazd gotówką.

W wynikowym zbiorze (3) powinny znaleźć się atrybuty:

- miesiąc,
- strefa rozpoczęcia podróży oraz
- liczba pasażerów, jaka rozpoczęła swoją podróż w podanej strefie w podanym miesiącu

## Program Hive (5)

Działając na wyniku zadania MapReduce oraz zbiorze danych `datasource4` należy dla każdego miesiąca wyznaczyć trzy dzielnice (`Borough`), w których wsiadło najwięcej pasażerów.

Wynik (6) powinien zawierać następujące atrybuty:

- `month` – miesiąc
- `borough` – dzielnica
- `passengers` – liczba pasażerów jaka rozpoczęła swoją podróż w podanej dzielnicy w podanym miesiącu