Apache Spark – Zestaw 2 – nyc-taxi

Misja główna

Cel przetwarzania

Dla każdego miesiąca, rozumianego jako okres czasu czyli miesiąc w roku (`YYYY-MM`), należy wyznaczyć trzy dzielnice (`Borough`), w których wsiadło najwięcej pasażerów. Oprócz wyliczenia liczby pasażerów należy wyznaczyć także sumaryczną kwotę jaką ci pasażerowie zapłacili za te przejazdy, sumaryczną odległość tych przejazdów, a także trzy dni (w tym miesiącu i dzielnicy), w których liczba odjazdów była największa. Ograniczamy naszą analizę tylko do tych przejazdów, które zostały opłacone gotówką. Nie należy uwzględniać takich miesięcy, dla których liczba wszystkich przejazdów (niezależnie od metody zapłaty) nie przekroczyła 1000.

Wynik przetwarzania powinien zawierać następujące atrybuty:

- `month` – miesiąc w formacie `YYYY-MM`
- `borough` – dzielnica
- `passengers` – liczba pasażerów jaka rozpoczęła swoją podróż w podanej dzielnicy w podanym miesiącu
- `total_amount` – sumaryczna kwota zapłacona za przejazdy
- `trip_distance` – sumaryczna odległość przejazdów
- `top_days` – trzy dni z największą liczbą pasażerów (3 elementowa tablica rekordów o dwóch polach `day` i `passByDay`)

Sugerowany schemat wyniku

```
root
 |-- month: string (nullable = true)
 |-- borough: string (nullable = true)
 |-- passengers: long (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- top_days: array (nullable = false)
 |    |-- element: struct (containsNull = false)
 |    |    |-- day: string (nullable = true)
 |    |    |-- passByDay: long (nullable = true)
```

Uwagi

- Do obliczania sumarycznej kwoty uwzględniamy całkowitą kwotę pobieraną od pasażerów.

Misje poboczne

Misja 1

Wyznacz trzy dni, w których liczba przejazdów była największa. Dla każdego z dni wyznacz: dzień w formacie YYYY-MM-DD, liczbę przejazdów, liczbę pasażerów, sumaryczną długość przejazdów. Pomiń w obliczeniach te przejazdy, które odbyły się bez opłat (`Payment_type=3`) oraz te, które zostały anulowane (`Payment_type=6`)

Wynik ma zawierać kolumny:

- `day` – dzień w formacie YYYY-MM-DD
- `trips` – liczba przejazdów
- `passengers` - liczba pasażerów
- `distance` – sumaryczny przejechany dystans

Misja 2

Dla każdej pary dzielnic wyznacz ile przejazdów zostało wykonanych pomiędzy nimi, jaka liczba pasażerów została w wyniku tych przejazdów obsłużona, jaka sumaryczna całkowita kwota (Total_amount) została zapłacona za te przejazdy.

Uwzględnij tylko te przejazdy, które został opłacone gotówką lub kartą kredytową (`Payment_type in (1,2)`).

Pomiń te pary dzielnic, pomiędzy którymi liczba przejazdów była mniejsza niż 10 000.

Wynik ma zawierać kolumny:

- PUBorough – dzielnica wyjazdu
- DOBorough – dzielnica docelowa
- `trips` – liczba przejazdów
- `passengers` - liczba pasażerów
- `totalAmount` – sumaryczna całkowita kwota