Przetwarzanie masywnych danych

Ćwiczenia przed kolokwium

Imię i nazwisko: _____

Numer indeksu: _____

Data: **28 maja 2019**

W odpowiedziach na pytania należy być **bardzo zwięzłym** oraz **pisać starannie**. Za każde pytanie można dostać do 10 punktów.

1. Sieć hoteli postanowiła założyć hurtownię danych. Na początek postanowiono stworzyć hurtownię danych dla procesu analizy zajętości pokoi hotelowych przy zabranych następujących informacjach. W ramach sieci rozróżnia się kategorię hotelu (standardowy, motel, zajazd, itp.). O każdym hotelu przechowuje się takie informacje jak jego kategorię, adres, kraj, region, położenie (centrum miasta, poza miastem, itp.). Pokoje w każdym hotelu opisane są rozmiarem, typem (apartament, pojedynczy, podwójny, itp.), liczbą łóżek, maksymalną liczbą osób, wyposażeniem dodatkowym (lodówka, kuchenka, itp.). Każdy hotel ma określoną liczbę pokoi danego typu.

   Hurtownia danych ma umożliwiać podanie sumarycznej liczby zajętych pokoi w różnych okresach czasu, z podziałem na kategorię hotelu, typ pokoju, dodatkowe wyposażenie, itp..

   Zaproponuj schemat gwiazdy, wskaż tablicę faktów, tablice wymiarów, zaznacz podkreśleniem klucze główne oraz linią falowaną klucze obce, zaznacz miary.
   Odp.: (10 pkt.)

2. Operatory agregacji można poklasyfikować ze względu na sposób ich obliczania. Rozróżnia się operatory rozproszone, algebraiczne oraz holistyczne. Przyporządkuj następujące operatory do odpowiednich grup: count(), max(), mediana(), min(), rank(), sum(), ave(), std(): (10 pkt.)

   - operatory rozproszone: _____
   - operatory algebraiczne: _____
   - operatory holistyczne: _____

3. W systemie składowana jest następująca perspektywa zmaterializowana (ang. *materialized view*) oznaczona jako V:

   ```
   SELECT marka, model, rok, sum(cena) as cena
   FROM Sprzedaz, Samochod
   WHERE Sprzedaz.id_samochod = Samochod.id_samochod
   GROUP BY marka, model, rok;
   ```

   Do systemu zostało wysłane poniższe zapytanie:

   ```
   SELECT model, sum(cena)
   FROM Sprzedaz, Samochod
   WHERE Sprzedaz.id_samochod = Samochod.id_samochod
   AND Sprzedaz.marka = 'Ford'
   GROUP BY model;
   ```

   Pokaż jak system może wykonać przepisanie zapytania (ang. *query re-write*)) do równoważnego zapytania, które wykorzystuje powyższą perspektywę V (10 pkt.):

4. Dwuwymiarowe dane przedstawione są w postaci poniższej macierzy 4×4:

8			
	8	32	
13	52		1
			15

   Przedstaw sposób bezstratnej kompresji, który pozwoli przechować powyższą macierz na 9 bajtach przy założeniu, że do przechowania każdej liczby potrzebny jest 1 bajt. Metadane takie jak rozmiar macierzy, nazwy wymiarów, atrybutów i ich wartości zapamiętywane są osobno.
   Odp.: (10 pkt.)

5. Zaprojektuj w technologii MapReduce algorytm projekcji. Dla dużego zbioru krotek przechowywanego w rozproszonym systemie plików należy wybrać atrybuty ze zbioru A. Dla pojedynczej krotki t, będącej wejściem procedury mapowania, oznaczmy jej projekcję jako $t(A)$. Przedstaw wyjście procedury mapowania oraz wejście i wyjście procedury redukcji, zakładając że wynikiem operacji ma być relacja (czyli zbiór krotek).

Procedura mapowania: (5 pkt.)

Procedura redukcji: (5 pkt.)

6. Krótko wyjaśnij co robi poniższy kod w jęzuku Scala używający techonologii Spark, jeśli obiekt `data` (typu RDD) jest zbiorem krotek w formacie (grupa, liczba). Funkcja `pow(base, exponent)` podniosi `base` do potęgi `exponent`.

```
data.map(a => (a._1, (a._2, 1)))
.reduceByKey((a,b) => (a._1 * b._1, a._2 + b._2))
.map(a => (a._1, pow(a._2._1, 1/a._2._2))
```

Odp.: (10 pkt.)

7. Udowodnij poprzez indukcję, że w k-tej iteracji próbkowania *resorvoir* prawdopodobieństwo wyboru bieżącego oraz każdego wcześniejszego elementu do próbki jest równe $\min(1, s/k)$, gdzie s jest rozmiarem losowanej próbki. (10 pkt.)

8. Baza danych o adresach URL zawiera n elementów. Sformułuj problem minimalizacji prawdopodobieństwa p fałszywych pozytywnych odpowiedzi (ang. *false positives*) dla filtra Blooma o wielkości $m = 10n$. Co w tym problemie jest zmienną decyzyjną? Jak można przeformułować ten problem, aby był łatwiejszy w optymalizacji? (10 pkt.)

9. Miara odległości oparta na podobieństwie Jaccarda spełnia wszystkie 4 warunki metryki. Podobną miarą podobieństwa jest współczynnik nałożenia (ang. *overlap coefficient*) zdefiniowany następująco: $f(x, y) = |x \cap y| / \min(|x|, |y|)$, gdzie x i y są zbiorami. Wykaż, że funkcja $1 - f(x, y)$ nie jest metryką. Wskazówka: wykaż, że jeden z warunków nie jest spełniony. (10 pkt.)

10. Dla poniższych danych przedstawionych w postaci macierzy charakterystycznej,

Element	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

sprawdź jak dobrze technika funkcji minhaszowych przybliża podobieństwo Jaccarda.

Wykorzystaj poniższe funkcje mieszające do obliczenia signatury i przybliżenia wartości podobieństwa Jaccarda:

- $h_1(r) = r + 1 \mod 5$, $h_2(r) = 3r + 1 \mod 5$,
- $h_3(r) = 2r + 4 \mod 5$, $h_4(r) = 2r + 1 \mod 5$.

Macierz sygnatur: (6 pkt.)

h	S_1	S_2	S_3	S_4
$h_1(r)$				
$h_2(r)$				
$h_3(r)$				
$h_4(r)$				

Oszacowane podobieństwa: (2 pkt.)

$\hat{J}(S_i, S_j)$	S_1	S_2	S_3	S_4
S_1				
S_2				
S_3				
S_4				

Prawdziwe podobieństwa: (2 pkt.)

$J(S_i, S_j)$	S_1	S_2	S_3	S_4
S_1				
S_2				
S_3				
S_4				

11. W problemie szukania najbliższych sąsiadów względem odległości Hamminga w binarnej przestrzeni 100-wymiarowej wykorzystano funkcje mieszające, które losowo wybierają element wielowymiarowego wektora opisującego dany obiekt. Jaką charakterystykę (d_1, d_2, p_1, p_2) ma taka funkcja mieszająca dla odległości $d_1 = 5$ i $d_2 = 95$? Co się zmieni jeżeli zastosujemy 50 takich funkcji w następującej konfiguracji: funkcje zostały podzielone na 5 grup, każda zawiera 10 elementów. Wewnątrz grup zastosowano operację OR, natomiast pomiędzy grupami operację AND.

    Podaj charakterystykę (d_1, d_2, p_1, p_2) pojedynczej funkcji mieszającej dla $d_1 = 5$ i $d_2 = 95$: (5 pkt.)

    Podaj charakterystykę (d_1, d_2, p_1, p_2) dla 50 funkcji mieszających w konfiguracji podanej powyżej (również dla $d_1 = 5$ i $d_2 = 95$): (5 pkt.)

12. W problemie szukania najbliższych sąsiadów względem odległości kosinusowej wykorzystano następujące 4 losowe wektory:

- $v_1 = (1, 1, 1, -1)$, $v_2 = (1, 1, -1, 1)$
- $v_3 = (1, -1, 1, 1)$, $v_4 = (-1, 1, 1, 1)$

Oblicz sygnatury dla poniższych dwóch punktów (6 pkt.):

- $x = (2, 3, 4, 5)$, $y = (1, 2, 3, 4)$

    Sygnatura dla punktu x: _____

    Sygnatura dla punktu y: _____

Jeżeli metoda LSH używa otrzymane sygnatury łącząc poszczególne ich elementy operacją AND, czy powyższe punkty staną się parą kandydacką? Odpowiedź uzasadnij. (4 pkt.)