Przetwarzanie masywnych danych

Ćwiczenia przed kolokwium

Imię i nazwisko: _____

Numer indeksu: _____

Data: **15 stycznia 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. Zaproponuj algorytm wykorzystujący indeks segmentowy (ang. *bit-sliced index*) dla warunków zakresowych postaci $l < A < r$, gdzie l i r definiują lewy i prawy koniec zakresu dla wartości atrybutu A. Niech wszystkie liczby będą zapisane na n bitach (najbardziej znaczący bit to $n-1$, najmniej znaczący bit to 0). Do i-tego bitu liczb l i r odwołuj się jak do tablic, tzn. poprzez $l[i]$ i $r[i]$. Indeks segmentowy składa się z n bitmap $B[i]$, gdzie $i = 0, \ldots, n-1$. Wynikiem jest bitmapa R. Pomocnicze bitmapy oznacz dużymi literami. **Odp**.: (10 pkt.)

4. 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
   WHERE 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.):

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. 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.)

7. 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.)

8. Wykorzystaj algorytm Flajoleta-Martina do oszacowania unikatowej liczby elementów w zbiorze. Dla każdego elementu wykorzystywana jest 8-bitowa funkcja mieszająca o postaci:

$$h(x) = 3x + 1 \mod 2^8,$$

   a następnie brana jest reprezentacja binarna wartości tej funkcji. Do określenia *niezwykłości* danego elementu brana jest liczba ostatnich zer (najmniej znaczących) reprezentacji binarnej. Dla ciągu liczb 2, 3, 2, 9, 11, 24, 15, 3 jaką otrzymamy odpowiedź? (10 pkt.)

9. 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	___	___	___	___

10. 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.)