Datenbanksysteme WS 2016/17
Prof. Dr.-Ing. Sebastian Michel
BSc. Peter Brucker
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
http://dbis.informatik.uni-kl.de
Übungsblatt 3: Ausgabe 15.11.2016, Abgabe 21.11.2016

Aufgabe 1: V-Optimale Histogramme, Wavelets, Sketches (1 P.)

(a) Gegeben folgende Tabelle mit Zahlen und ihren Häufigkeiten. Berechnen Sie das V-Optimale Histogramm für $B = 2$ Zellen unter Anwendung des in der Vorlesung vorgestellten Algorithmus. Geben Sie dabei verwendete Hilfsstrukturen wie P und PP an, und die DP-Tabelle $SSE^*(i, k)$, insbesondere für Schritte, in denen eine günstigere Möglichkeit gefunden wurde. Geben Sie am Ende die optimalen Histogrammgrenzen an.

Wert	**Häufigkeit**
1	5
2	9
3	1
4	3
5	2

Zeigen Sie ferner

$$SSE([i,j]) = \sum_{i \le k \le j} (F[k]^2) - (j - i + 1) * AVG([i,j])^2$$

(b) Berechnen Sie für folgende kumulative Verteilung über den Zahlen 0 bis 7 die Wavelet-Transformierte und berechnen Sie den Mean-Squared-Error (MSE) in der Approximation durch die Wavelet-Transformierte, die durch Weglassen von Koeffizienten ≤ 1 entsteht.

$$[4, 5, 7, 10, 11, 15, 19, 20]$$

(c) Gegeben folgende Tabelle mit Elementen $x \in X$ und den Hashwerten für die Hashfunktion $h : X \to [0, 1]$.

x	$h(x)$
Apfel	0,34
Birne	0,85
Kirsche	0,04
Pflaume	0,85
Pfirsisch	0,31

x	$h(x)$
Brombeere	0,49
Erdbeere	0,75
Stachelbeere	0,27
Banane	0,63
Ananas	0,83

Berechnen Sie mittels KMV-Sketch für $k \in 1, 2, 3, 4, 5$ den Schätzwert, den jeder einzelne Sketch liefert sowie jeweils den absoluten Fehler. Wird der Fehler mit größerem k kleiner? Wie verhält sich dies generell für den KMV-Sketch?

Aufgabe 2: Statistiken in PostgreSQL (1 P.)

Das Datenbankmanagementsystem PostgreSQL verwaltet für jede Spalte in jeder Tabelle statistische Daten, die zur Anfrageoptimierung verwendet werden können. Betrachten Sie folgenden Auszug der Statistiken zur Tabelle *lineitem* aus dem TPC-H-Datensatz:

```
SELECT * FROM pg_stats WHERE tablename = 'lineitem'
```

(a) Was bedeuten die einzelnen Spalten von *pg_stats*?

attname	null_frac	n_distinct	most_common_vals	most_common_freqs
l_linenumber	0	7	{1,2,3,4,5,6,7}	{0.25,0.21,0.17,0.14,0.10,0.07,0.03}
l_orderkey	0	419664	{1703939,3644579,507137, 703172,712326,770882, 971014}	{0.000133333,0.000133333,0.0001, 0.0001,0.0001,0.0001, 0.0001}
l_extendedprice	0	-0.11964	{13747.3,25165.2, 26677.8,32274.0, 33265.3,35938.8,50370.3 }	{0.0001,0.0001,0.0001,0.0001, 0.0001,0.0001,0.0001 }

(b) Für welche der folgenden Anfragen können diese Statistiken nützlich sein?

- ☐ `SELECT * FROM lineitem`
- ☐ `SELECT * FROM lineitem WHERE l_orderkey IS NOT NULL`
- ☐ `SELECT * FROM lineitem WHERE l_orderkey = 406631`
- ☐ `SELECT * FROM lineitem WHERE l_extendedprice = 32274.0`
- ☐ `SELECT * FROM lineitem WHERE l_extendedprice <= 9500.0`

(c) Schätzen Sie so gut es geht die Anzahl der Ergebnistupel für die Anfragen aus (b)

Im Folgenden sehen Sie eine weitere Spalte von *pg_stats*, in der ein Histogramm über die Spaltenwerte definiert wird. Achtung: NULL-Werte sowie die Werte aus *most_common_vals* fließen nicht in die Histogrammberechnung ein![1]

attname	histogram_bounds
l_extendedprice	{906.00,1528.60,2062.06,3035.16,3639.44,4444.04,5265.52,5920.60, 6751.85,7518.63,8252.80,9032.32,9777.68,10540.68,11267.46, 11921.76,12652.00,13413.40,14078.28 (...) }

d) Bei welchen Anfragen aus b) sind die Informationen aus der Spalte *histogram_bounds* hilfreich? Schätzen Sie erneut die Anzahl der Ergebnistupel.

Aufgabe 3: Joinordering (1 P.)

Gegeben folgende Query über der TPC-H Datenbank:

```
SELECT *
FROM part p, partsupp ps, lineitem l
WHERE p.p_partkey = ps.ps_partkey
  AND ps.ps_partkey = l.l_partkey AND ps.ps_suppkey = l.l_suppkey
```

(a) Zeichnen Sie den Anfragegraphen.

(b) Finden Sie heraus, wie groß die Join-Selektivitäten in dem in der Vorlesung vorgestellten TPC-H Datensatz sind.

(c) Berechnen Sie für alle möglichen Join-Bäume ohne Kreuzprodukte die Kosten C_{out}.

[1] histogram_bounds: "A list of values that divide the column's values into groups of approximately equal population. The values in most_common_vals, if present, are omitted from this histogram calculation." – `http://www.postgresql.org/docs/9.6/static/view-pg-stats.html`