Bill Yang | Due: 2/28/19

Homework 4

CS386D Database Systems | Instructor: Daniel Miranker

1.

a)

For each c_i, there are m_i different values for the columns. Then for each column (using the result given in 14.7.3b), we need $2n\lceil \log_2(m_i - 1)\rceil$. Then the total number of bytes needed is $\frac{1}{8}\sum_{i=1}^{i=100}(2)100000000\lceil \log_2(m_i - 1)\rceil = 25000000\sum_{i=1}^{i=100}\lceil \log_2(m_i - 1)\rceil$

b)

No, it does not need one because each value of c_0 has only one row with that value which is recoverable from c_0, so it is redundant information.

c)

i.

50 columns have $n/1000$ values and 50 columns have $10,000$ values. Then for S, $n = 1000000$. For the $n/1000$ columns, we need $\frac{1}{8}50(2)(1000000)\lceil \log_2(1000000/1000 - 1)\rceil = 125000000$ bytes.
For the 10000 columns, we need $\frac{1}{8}50(2)(1000000)\lceil \log_2(10000 - 1)\rceil = 175000000$.
Total we need
300000000 bytes.

ii.

We need $\frac{1}{8}(50)(2)(100000000)(\lceil \log_2(100000000/1000 - 1)\rceil + \lceil \log_2(10000 - 1)\rceil) =$
38750000000 bytes.

d)

i.

For a single row, we need $4+50(25)+50(20)$ bytes, since we need 4 for c_0, then there are 50 columns that require an average of 25 bytes, and then other 50 columns require an average of 20 bytes.
Thus for all n rows, we need
$2254n = 2254000000$ bytes.

ii.

$2254n = 225400000000$ bytes.

iii.

4k bytes $= 2^{12} = 4096$ bytes.
$2254000000/4096 \approx 550293$ pages.

iv.

$225400000000/4096 \approx 55029297$ pages.

2.

a)

Then if there is a hit in the bloom fliter and the block is retrieved, we can verify whether the key

actually exists in storage (the block). Otherwise we wouldn't know if we actually have the file needed or if it is a false positive.

b)
i. 64Mbytes = 2^{26} bytes. $1024 = 2^{10}$. $2^{26}/2^{10} = 2^{16}$ key value pairs fit in a block. Then the number of bits required for a filter is $(10)2^{16} =$
655360 bits.

ii.
Database has 2^{43} total storage, across $128 = 2^7$ servers, so $2^{43}/2^7 = 2^{36}$ storage per server. Then there are $2^{36}/2^{26} = 2^{10}$ blocks per server. Then the number of bytes needed for all the bloom filters is $10(2^{16})(2^{10})/8 = 10(2^{23}) =$
83886080 bytes or 80Mbytes.

iii.
$\ln(2) \times m/n$, where $m/n = 10$. So the optimal number of hash functions is 7.

iv.
Probability of false positive = $(1/2)^{\ln(2)\times m/n} = .00819$.