1a) $4,000,000 \times 40\,MB = 160,000,000\,MB$

1b) $B_{k \times m} \Rightarrow k \times m = 27,000,000$ bits

$= 3.375\,MB$

1c) $FPR = \prod_{i=1}^{k} P[B_i(h_i(x)) = 1]$

$= \left[1 - \left(1 - \frac{1}{m}\right)^n\right]^k$

$k = 30 \quad n = 4,000,000$

$.03 = \left[1 - \left(1 - \frac{1}{m}\right)^{4,000,000}\right]^{30}$

$m = 1,814,524$

1d) $= 30 \times 1,814,524 = 54,435,720$ bits

$6.8\,MB$

about double the space but still very short of if it was deterministic

2a) $.5 \times 102{,}000 = 61{,}000$ seconds

2b)
- look for URL in Bloom Filter
- if it returns for malicious
  - check in database to know for sure (this is deterministic)
- if its not there, do nothing else

2c) $102{,}000 \times .03 = 3060$ false positives

$(2000 + 3060)(.5 \text{ seconds}) + (102000 - 3060)(1 \text{ ms})$

2530 seconds + 98,94 seconds

2628.94 seconds

2d) Bloom filter + database deterministic approach is 4% of the run time of the purely database approach

Bloom filter is much better approach

Question 2)

Because there is only one row the

$$FPR = \left[1 - \left(1 - 1/m\right)^n\right]$$

instead of

$$FPR = \left[1 - \left(1 - 1/m\right)^n\right]^k$$ because there

are no longer k rows in the matrix

b)

if FPR = .05

$k = 30$ $n = 4{,}000{,}000$

$.05 = [1 - (1 - 1/m)^n]$

$.95 = (1 - 1/m)^{4{,}000{,}000}$

$\log(.95) = 4{,}000{,}000 \log(1 - 1/m)$

$-1.28 \times 10^{-8} = \log(1 - 1/m)$

$-1.28 \times 10^{-8} = \log(e^{-1/m})$

$-1.28 \times 10^{-8} = -1/m$

$m = 77{,}982{,}909$ bits

space = 9.747 MB

modified version of bloom filter is worse

$.05 = [1 - (1 - 1/m)^n]^k$

$.05 = [1 - (1 - 1/m)^{4{,}000{,}000}]^{30}$

$\frac{\log(.05)}{30} = \log(1 - (1 - 1/m)^{4\ldots})$

$-.099 = \log(1 - (1 - 1/m)^{4{,}000{,}000})$

$e^{-.099} = (1 - (1 - 1/m)^{4000000})$

$.09 = (1 - 1/m)^{4{,}000{,}000}$

$\frac{\log(.09)}{4{,}000{,}000} = \log(e^{-1/m})$

$-6.0198 \times 10^{-7} = -1/m$

$m = 1{,}661{,}184$

$k = 30$

total space = $m * k$

= 6.22 MB

Coding Challenge

original FPR = .0359

FPR with $k=10$ $m=8000$ and $t=(1, m-1) = 1.0$

FPR with $k=10$ $m=79,000$ and $t(1, m-1) = .0339$

ratio from 2b = $\frac{77,982,909}{1661,184} = 46.944$

ratio from coding challenge = $\frac{79,000}{8000} = 9.875$

I think the ratio for the homework is smaller because $n = 10,000$ but for the derivation $n = 4,000,000$, k was also larger for the derivation