Thursday Quiz

1. [1+1 points] Explain why we want to compute moments for data streams. Also give the formula for calculating the second moment.

Answer-
0th moment =number of distinct elements
1st moment =count of the numbers of elements = length of the stream, Easy to compute
2nd moment = surprise number S=a measure of how uneven the distribution is

Formula-

- 2^{nd} moment is $S = \sum_i m_i^2$

2. [2 points] Explain the three properties of buckets that need to be maintained while bucketizing data streams in DGIM.
   Either one or two buckets with the same power-of-2 number of 1s
   Buckets do not overlap in timestamps
   Buckets are sorted by size.
3. [4 points] In the Flajolet-Martin Algorithm, Suppose our stream consists of the integers 3, 1, 4, 5, 9, 2. Our hash functions will all be of the form h(x) = ax+b mod 32 for some a and b. You should treat the result as a 5-bit binary integer. Determine the tail length (i.e. length of longest trailing zeros) among all stream elements and the resulting estimate of the number of distinct elements if the hash function is:
   (a) h(x) = 2x + 3 mod 32.
   (b) h(x) = 5x mod 32
   Ans:
   Length of Longest Zeros for (a) = 0 [1 point]
   Estimated number of distinct elements for (a)= 1 [1 point]
   Length of Longest Zeros for (b) = 2 [1 point]
   Estimated number of distinct elements for (b) = 4 [1 point]
   a)
   h(3) = 2*3 + 3 mod 32 = 9 = 01001
   h(1) = 2*1 + 3 mod 32 = 5 = 00101
   h(4) = 2*4 + 3 mod 32 = 11 = 01011
   h(5) = 2*5 + 3 mod 32 = 13 = 01101
   h(9) = 2*9 + 3 mod 32 = 21 = 10101
   h(2) = 2*2 + 3 mod 32 = 7 = 00111
   tail length = 0 and hence number of distinct elements = 2^0 = 1

   b)
   h(3) = 5*3 mod 32 = 15 = 01111
   h(1) = 5*1 mod 32 = 5 = 00101
   h(4) = 5*4 mod 32 = 20 = 10100 ←

h(5) = 5*5 mod 32 = 25 = 11001
h(9) = 5*9 mod 32 = 13 = 01101
h(2) = 5*2 mod 32 = 10 = 01010
tail length = 2 and hence number of distinct elements = 2^2 = 4

4. [2 points] What are the two different strategies when it comes to sampling data from a data stream. What advantage does one have over the other?.
   Sample a fixed proportion of elements in the stream (say 1 in 10)
   Maintain a random sample of fixed size over a potentially infinite stream.

   Sampling fixed proportion becomes too large when we have too big data
   In the second method, we hash consistently, that means remove all or none of the occurrences of a query