Take Home Quiz

1. [6 points] Suppose we have a stream of tuples with the schema **Grades(university, courseID, studentID, grade)**

   Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., “CS101”) and likewise, studentID’s are unique only within a university (different universities may assign the same ID to different students).
   Suppose we want to answer certain queries approximately from a 1/15th sample of the original steam data, i.e., this is a sampling fixed proportion problem. The basic process to deal with sampling fixed proportion problem is: use a hash function to generate random integers in [0...14] for each incoming tuple under a certain tuple key attribute; store the tuple if the integer is 0, otherwise discard. For each of the queries below, indicate how you would construct the sample.

   That is, tell what the key attributes should be.
   (a) [2pts] For each university, estimate the average number of courses per university.
   (b) [2pts] Estimate the fraction of students who have a GPA of 3.7 or more.
   (c) [2pts] Estimate the fraction of courses where at least half the students got “A.”

   a) Key attribute - university
   b) Key attribute - (university, studentID)
   c) Key attribute - (university, courseID)
2. [2 points] Explain why the Flajolet-Martin algorithm works by showing 2^R will not be too small or too large comparing to the number of unique values in the data stream
   m - no. of unique values in the data stream
   probability that a hash func ends with r zeros = 2^-r
   prob of not seeing r zeros among m elements = (1-2^-r)^m

   prob that a hash function has atleast r trailing zeros = 1-(1-2^-r)^m = 1-(1-2^-r)^(2^r*(m*2^-r) = 1-(1-e^-m/2^r)

   If m >>> 2^r, then the probability of having atleast r trailing zeros will be approaching one

   If m <<< 2^r , then the probability of having atleast r trailing zeros will be approaching zero

   From the above 2 statements, its clear that the estimate 2^r will be neither too high nor too low.

3. [2 points] In DGIM method, explain how the buckets get updated when the value of the new bit that comes in the stream is 1?
   Ans:

1 Create a new bucket of size 1, for just this bit. End timestamp = current time

2 If there are now three buckets of size 1, combine the oldest two into a bucket of size 2

3 If there are now three buckets of size 2, combine the oldest two into a bucket of size 4

4 And so on ...