Family name: …… Given name: ……

Family name: …… Given name: ……

Family name: …… Given name: ……

Let's suppose we have a log file recording the events coming from different machines. Thus, for each event we have the following information:

(logID, traceID, eventID, duration)

The logID corresponds to the IP of the machine; the traceID identifies the transaction inside the machine (i.e., two traceIDs can coincide in different machines); the eventID identifies the kind of action performed by the machine; finally, the duration is the number of milliseconds taken to implement the action.

Assuming that we cannot keep all log entries in memory, and we decide to randomly sample them, **give the attributes** (up to three) you would use as parameters of the hash function implementing such sampling, so that each of the following queries gives a result as accurate as possible.

a) Estimate the fraction of transactions where the same kind of action appears more than once.

…….. …….. ……..

b) For each machine, estimate the average number of actions per transaction.

…….. …….. ……..

c) Estimate the number of transactions with more than two actions taking more than 100ms.

…….. …….. ……..

Family name: …… Given name: ……

Family name: …… Given name: ……

Family name: …… Given name: ……

Let's suppose we have a black-list of IP addresses (whose packages we do not want to cross our firewall), which is too long to be kept in memory (10^7 elements). Thus, we decide to implement a Bloom filter (to avoid further processing of black-listed addresses), with only 10^8 bits.

d) How many hash functions would you use?

……

e) What's the probability of a false positive in that case?

……

f) Briefly explain what is the consequence of a false positive.

……
……
……
……

Note: ln 2 ≈ 0.693

Family name: …… Given name: ……

Family name: …… Given name: ……

Family name: …… Given name: ……

Let's suppose we have a log file recording the events coming from different machines. Thus, for each event we have the following information:

(logID, traceID, eventID, duration)

The logID corresponds to the IP of the machine; the traceID identifies the transaction inside the machine; the eventID identifies the kind of action performed by the machine; finally, the duration is the number of milliseconds taken to implement the action.

Consider the following table and assume that each machine generates the same number of events and at the same pace, and use an exponentially dacaying window model with a constant 0.5.

Time	logID	traceID	eventID	duration
1	1	1	A	1
2	2	1	B	100
3	1	1	C	10
4	2	1	D	10
5	1	1	A	100
6	2	1	C	1

g) Give the milliseconds (and details of calculation) used per machine when the last event arrives.

Machine 1:

Machine 2: