Family name:…… Given name:……

1) (35%) Consider a left-deep process tree corresponding to a query, where each internal node is a join, and every leaf a data source (e.g., relational table). Knowing that the tree contains 9 nodes (including leaves), the system has infinite parallelism capacity in pipelining mode (no other kind of parallelism is available), which is the occupancy if the overall cost of the query is 4 seconds? Explicit any assumption you need to make.

   Assumptions:
   a) ……
   b) ……

   Answer:

   ……
   ……
   ……

2) (20%) Consider an HDFS cluster with 100 data nodes, without replication. If I upload a file with 10 chunks and 10 blocks each, answer the following questions and briefly justify your answer:
   a. Which is the maximum number of machines containing data?
      ……
      ……
   b. Which is the probability of the maximum number of machines actually contain data?
      ……
      ……
      ……
      ……

3) (20%) Briefly explain how you would implement an intersection (i.e., T∩S) with MapReduce. Clearly explicit which will be the key and which the value. Since it is a binary operation, assume the existence of a function "input: key -> [T|S]", as well as an operation "⊕" that concatenates attributes as needed.

   ……
   ……
   ……
   ……

4) (25%) 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 the pace of processing all log entries, and we decide to randomly sample them, briefly explain how would you implement the sample to bound the error of the following query: "Return the average sum of the duration per transaction".

   ……
   ……
   ……
   ……