Data Science for Business MId-term mini exam **February 2018**

Questions For each of the following questions indicate the letter corresponding to the right answer.

1. Consider a cluster of 4 racks (each one including 40 computers). In order to maximise robustness HDFS should
    A. store all the copies of each file chunk in the same rack
    B. store all the copies of a file chunk in the same machine of a rack
    C. never store more than one copy of a file chunk in the same rack
2. In the execution of a MapReduce job, whose input is a file partitioned and stored in 10 chunks by means of HDFS, the number of Reduce tasks
    A. must be equal to 10
    B. is 1
    C. can be either greater than 10 or less than 10 or equal to 10.
3. During the shuffle&sort process for a MapReduce job execution, couples that are returned by Map tasks
    A. are first stored in local disk files and then processed for grouping and sorting
    B. are locally grouped and sorted on the fly in RAM memory
    C. are immediately sent to machines running Reduce tasks
4. Consider an existing MapReduce program for performing the intersection between two *sets*, whose elements are respectively stored in two different files. In order to sensibly decrease the amount of information emitted by Map and sent to shuffle and sort:
    A. the only possibility is to add a Combiner
    B. neither a combiner nor a Map rewriting can be used
    C. the number of splits can be tuned

Exercise

Consider the following simple relational schema containing informations about clients and orders they made.

Customer(cid, startDate, name)

Order(#cid, total)

Note that, for the sake of simplicity, we do not have any primary key for Order. Also assume that all the fields are mandatory.

Provide the pseudo-code or Pyhton MapReduce encoding of the following SQL query.

```
SELECT  O.cid, MAX(total)
FROM Order O
WHERE year(startDate)=2017
GROUP BY O.cid
```

Is it possible to add a Combiner to improve scalability? If so, describe the combiner in Python or pseudocode.

Optional. What are the changes, if any, to your code or somewhere else in order to encode in MapReduce following variant of the previous query.

```
SELECT  O.cid, MAX(total)
FROM Order O
WHERE year(startDate)=2017
GROUP BY O.cid ORDER BY O.cid
```