Database Management System
Quiz

Question # 1

Assume an input file named as `players.csv` containing the following fields
`surname,team,position,minutes,shots,passes,tackles,saves`
An example tuple could be `{Abdoun,Algeria,midfielder,16,0,6,0,0}`
The following program (lines are numbered) is based on this input file. Explain in English what each of the following line does, if there are commented lines, write the code for that line.

```
1. from pyspark.sql import SparkSession
2. df = spark.read.load("players.csv", format="csv", sep=":",
   inferSchema="true", header="true")
3. df.createOrReplaceTempView("players")
4. results = spark.sql("SELECT surname FROM players")
5. results.show()
6. // write code to answer the question: which players(surname only)
   played less than 200 minutes and made more than 100 passes?
7. // show the result of the query in line # 6
8. // Write code to answer the query: What is the average number of
   passes made by forwards? By midfielders? Write one query that
   gives both values.
9. // show the result of the query in line #
```

Question # 2

i. Spark is 10-100 times faster than Hadoop because of
   a. Disk-based data processing.
   b. Memory-based data processing.
   c. Both a and b
   d. None of the above.

ii. Which of the following are NoSQL databases?
   a. Key-value
   b. Column
   c. Document
   d. All of the above

iii. NoSQL databases is used mainly for handling large volumes of _____ data.
   a. Unstructured
   b. Structured
   c. None of the above
   d. all of the above

iv. Which one of the following is not a valid MapReduce step?
   a. Map
   b. Reduce
   c. Shuffle
   d. Save