**Quiz**

**Question 1: What does relational data model consist of?**

Tables, rows, columns and values

**Question 2: Imagine that in your application you need to associate various data with users, i.e. their preferences, behavioural information and so on. You are willing to persist user profiles on the disk. Given the choice between CSV and JSON, which format would you choose?**

JSON, obviously

**Question 3: Why are columnar file formats used in data warehousing? Mark all the correct answers.**

Columnar stores allow more efficient slicing of data (both horizontal and vertical).

Columnar stores occupy less disk space due to compression.

**Question 4: Compared to text formats, why is the SequenceFile format faster? Mark all the correct statements.**

Simplified grammar (not tracking paired quotes, brackets, and so on) leads to a streamlined, unconditional code.

Scalar values are serialized/deserialized with a simple copy.

Serialized data occupies less disk space thus saving I/O time.

**Question 5: True or False? Optimizing the computation itself (optimizing the computation time) will not help to reduce the completion time of an I/O-bound process.**

False

**Question 6: True or False? Switching from compressed to uncompressed data for a CPU-bound process may increase the completion time despite the saved CPU time.**

True

Correct: It is always about the exact numbers. Removing the compression may drastically increase I/O times so that the process will become I/O bound with a worse completion time. Yet adding the compression makes the process CPU-bound -- but with a better completion time. It is always about the numbers.

**Question 7: What fact is more relevant to the horizontal scaling of the filesystems than to the vertical scaling?**

Usage of commodity hardware

Correct: Yes, it's possible to use simple and commodity hardware

**Question 8: The operation 'modify' files is not allowed in distributed FS (GFS, HDFS). What was NOT a reason to do it?**

Increasing reliability and accessibility

Correct: Yes, it was not a reason, because reliability and accessibility are mostly achieved by the replication.

**Question 9: How to achieve uniform data distribution across the servers in DFS?**

By splitting files into blocks

Correct: Yes, splitting large files into blocks increases data items granularity and allows to fill the servers more evenly

**Question 10: What does a metadata DB contain?**

File permissions

Correct: Yes, permissions are administrative information about a file, they are stored in the metadata DB.

File creation time

Correct: Yes, file creation time is administrative information about a file, it is stored in the metadata DB.

Location on the file blocks

Correct: Yes, locations of the blocks are administrative information about a file, they are stored in the metadata DB.

**Question 11: Select the correct statement about HDFS:**

Namenode is a master server, it stores files metadata

Correct: That's right, metadata is stored in the database on the Namenode.

**Question 12: If you have a very important file, what is the best way to protect it in HDFS?**

Both ways are allowed and implemented in HDFS

Correct: Yes, a replication is used for better reliability and you should also remember about the appropriate access to your data.

**Question 13: You were told that two servers in HDFS were down: Datanode and Namenode, your reaction:**

Restore Namenode first

Correct: Namenode is a master server, so it should be restored first.

**Question 14: What the block size in HDFS does NOT depend on?**

The block size on the local Datanodes filesystem

Correct: Yes, HDFS block size mostly depends on Namenode RAM amount and Datanode disks performance