Big Data Technologies BE0M33BDT: What you need to know

1. What is typical for big data (three V-features)?
2. What is Hadoop?
3. What important components of Hadoop do you know? What are they for?
4. Which distributions of Hadoop do you know? Why is it good to use a distribution?
5. What is HDFS and how does it differ from common filesystems?
6. What is HDFS block and what size does it have usually?
7. How does a replication in HDFS work? What is the replication factor?
8. Do we prefer many small files, or few big files on HDFS? Why?
9. Is Hadoop suitable for transactional data (often updating and deleting)? Why?
10. What are main types of nodes in Hadoop cluster?
11. What is the Secondary NameNode for? How does it co-operate with the NameNode?
12. What is the name of the native resource manager in Hadoop?
13. What file formats are typically used for storing in Hadoop?
14. What are main advantages of column-oriented storage format?
15. What is a SequenceFile and what is its purpose?
16. What compression algorithms are typically used in Hadoop?
17. What SQL operations are not supported in Hive? Why? How do we solve this?
18. What is a difference between external and internal (managed) tables in Hive?
19. What is a purpose of partitioning in Hive?
20. What is the Hive metastore?
21. What are the main steps of MapReduce algorithm? Which of them makes the biggest load on cluster resources?
22. What is Apache Spark and what is its main advantage over MapReduce processing?
23. What is the difference between transformation and action in Spark?
24. How can we make data stay in memory in Spark?
25. Explain the terms job, stage, driver, executor in Spark.
26. What modes can be used for launching jobs in Spark?
27. What is Spark RDD and Spark DataFrame? How do they differ?
28. What programming languages can be used in Spark?