39 Hadoop interview questions to ask your next candidate

Questions

1. Can you explain what Hadoop is and its primary components?
2. How does Hadoop handle fault tolerance?
3. What is the role of NameNode and DataNode in HDFS?
4. What are the main advantages of using Hadoop for big data processing?
5. How does Hadoop ensure data security?
6. Can you explain what YARN is and its function within the Hadoop ecosystem?
7. What is the difference between HDFS and a traditional file system?
8. How would you optimize a MapReduce job for better performance?
9. What are the different types of data formats that Hadoop can process, and how do they differ?
10. Can you describe the concept of 'data locality' in Hadoop and its importance?
11. How do you manage and monitor the performance of a Hadoop cluster?
12. What are the key differences between a MapReduce job and a Spark job?
13. Can you explain the function and significance of the Secondary NameNode?
14. What are some common challenges you might face while working with Hadoop, and how would you address them?
15. How do you implement data ingestion in Hadoop, and what tools would you use?
16. What are the differences between batch processing and stream processing in Hadoop?
17. Can you explain the difference between a job, task, and a map/reduce in the context of Hadoop?
18. How would you handle data versioning when working with Hadoop?
19. Can you explain the differences between MapReduce and Apache Spark in terms of processing speed and ease of use?
20. What are the key benefits of using Apache Hive for data processing?
21. How does Apache Pig complement Hadoop's processing capabilities?
22. What is Apache Flink, and how does it differ from other processing frameworks?
23. Describe the role of Apache HBase in the Hadoop ecosystem.
24. How does Apache Storm facilitate real-time computation, and what are its main features?
25. What are the advantages of using Apache Tez over traditional MapReduce?
26. How does Apache Kafka integrate with Hadoop, and what are its use cases?
27. What is Apache NiFi, and how does it streamline data flow management in Hadoop?
28. Can you explain the block storage concept in HDFS and its advantages?
29. How does HDFS handle small files, and what are the potential issues?
30. What is the purpose of the fsck command in HDFS?
31. Explain the concept of rack awareness in Hadoop. How does it affect data storage?
32. What is the difference between the 'put' and 'copyFromLocal' commands in HDFS?
33. How does Hadoop ensure data integrity in HDFS?
34. What is the significance of the replication factor in HDFS?
35. Can you describe the process of data balancing in HDFS?
36. What are storage policies in HDFS, and how are they useful?
37. How does HDFS handle append operations?
38. What is the role of the edit log and fsimage in HDFS?