1. Explain the role of the NameNode in a Hadoop cluster and how it manages the storage and replication of data blocks in HDFS.
2. What is the significance of Hadoop in the big data ecosystem?
3. What are the key components of Hadoop?
4. Explain the Hadoop Distributed File System (HDFS) architecture.
5. What is the purpose of NameNode in HDFS?
6. What is the role of DataNode in HDFS?
7. How does data replication work in HDFS?
8. What is the default block size in HDFS?
9. How does Hadoop ensure fault tolerance in HDFS?
10. What is the purpose of the MapReduce framework in Hadoop?
11. Describe the basic flow of a MapReduce job.
12. What is the difference between the Map and Reduce functions in MapReduce?
13. What are the input and output formats in MapReduce?
14. Explain the concept of combiners in MapReduce.
15. What is a partitioner in MapReduce?
16. How does data shuffling occur in MapReduce?
17. What is the purpose of the JobTracker in Hadoop?
18. What is the role of TaskTracker in Hadoop?
19. How does Hadoop handle task scheduling and monitoring?
20. Explain speculative execution in Hadoop.
21. What is the purpose of the YARN framework in Hadoop 2.x?
22. What are the advantages of using YARN over the classic MapReduce framework?
23. What is the role of ResourceManager in YARN?
24. What is a NodeManager in YARN?
25. How does YARN handle resource allocation and job scheduling?
26. What are the different types of schedulers available in YARN?
27. Explain the concept of speculative execution in YARN.
28. What is the purpose of Hadoop Streaming?
29. How can you optimize Hadoop jobs for better performance?
30. What is the role of the Hadoop MapReduce Shuffle and Sort phase?
31. What is the purpose of the Hadoop Oozie workflow scheduler?
32. How can you secure a Hadoop cluster?
33. Explain the role of the secondary NameNode in HDFS.
34. What is the purpose of HBase in the Hadoop ecosystem?
35. How does HBase differ from traditional relational databases?
36. What is the purpose of Apache Hive in Hadoop?
37. What are the key features of Apache Pig in Hadoop?
38. Explain the concept of data locality in Hadoop.

39. How does Hadoop handle data processing failures?
40. What is the purpose of the Hadoop Distributed Cache?
41. How does Hadoop handle data compression?
42. What is the purpose of the Hadoop Command-Line Interface (CLI)?
43. Explain the concept of speculative execution in Hadoop.
44. How can you monitor the performance of a Hadoop cluster?
45. What is the role of ZooKeeper in Hadoop?
46. How does Hadoop handle input/output (I/O) operations?
47. What is the purpose of the Hadoop High Availability (HA) feature?
48. Explain the difference between block-level and file-level storage in HDFS.
49. How does Hadoop handle data replication across different racks?
50. What are the different file permissions available in Hadoop?
51. How can you integrate Hadoop with other tools and technologies in the big data ecosystem?
52. Describe how data is split and processed in a Hadoop MapReduce job, including the role of the JobTracker and TaskTracker.
53. Explain how the Hadoop YARN framework improves the scalability and flexibility of Hadoop by allowing different processing frameworks and workloads to run on the same cluster.
54. Describe the different types of data storage formats supported in Hadoop and the use cases for each one, such as Avro, Parquet, and ORC.
55. Discuss the different methods of data ingestion and transportation in Hadoop and the trade-offs between them, such as using Flume, Kafka, and Sqoop.
56. Explain how Hadoop security can be implemented to protect data at rest and in transit and the role of HDFS encryption zones, Hadoop KMS and Sentry.
57. Analyze the performance and scalability of a Hadoop cluster and suggest ways to improve it, such as increasing the number of nodes, using data compression, or changing the block size.
58. Explain how Hadoop Distributed File System (HDFS) works and the design principles behind it, including data replication, block size, and fault tolerance.
59. Describe how Hadoop MapReduce works, including the role of the Mapper and Reducer, how data is partitioned and sorted, and how intermediate data is handled.
60. Explain how the Hadoop ecosystem has evolved over time, including the introduction of YARN, the emergence of Spark and other big data processing frameworks, and the role of Hadoop in the broader data ecosystem.
61. Analyze the different approaches to data governance and management in Hadoop, such as using Apache Atlas or Apache Ranger, and explain the trade-offs between them.

62. Discuss the different use cases for Hadoop and the Hadoop ecosystem, including big data analytics, data warehousing, real-time data processing, and machine learning.
63. Explain the different Hadoop deployment models and the trade-offs between them, such as on-premises, cloud-based, and hybrid deployments.
64. Describe the different types of data processing available in the Hadoop ecosystem, such as batch processing, real-time processing, and interactive querying, and explain the use cases for each one
65. Scenario: A company is using Hadoop for batch processing of large amounts of sensor data. They are running into performance issues and need to optimize their cluster.
    a. What steps can the company take to improve the performance of their Hadoop MapReduce jobs?
    b. How can the company increase the scalability of their Hadoop cluster?
    c. What configuration changes can the company make to their HDFS to improve performance?
    d. How can the company monitor and troubleshoot their Hadoop cluster?

66. Scenario: A company is storing sensitive data in a Hadoop cluster and needs to secure it.
    a. What steps can the company take to encrypt data at rest in HDFS?
    b. How can the company implement fine-grained access control to data stored in HDFS?
    c. What measures can the company take to protect data in transit within the Hadoop cluster?
    d. How can the company audit access to data stored in HDFS?

67. Scenario: A company is using Hadoop for data warehousing and business intelligence. They have a large amount of structured and unstructured data and need to decide which Hadoop ecosystem component to use for their analysis.
    a. What Hadoop ecosystem component is best suited for processing structured data?
    b. What Hadoop ecosystem component can be used for real-time data processing?
    c. What Hadoop ecosystem component can be used for data warehousing and SQL-like querying?
    d. What Hadoop ecosystem component can be used for machine learning?

68. Scenario: A company is using Hadoop for data warehousing and business intelligence. They are storing large amounts of data in HDFS, but they are running out of storage space.
    a. How can the company scale their HDFS cluster to add more storage capacity?

b. What options are available for archiving or purging old data to free up space in HDFS?
c. How can the company optimize the storage usage in HDFS, such as by using data compression or changing the block size?

69. Scenario: A company is using Hadoop for big data analytics and they are running into performance issues when querying large datasets.
a. What steps can the company take to improve the performance of their Hive or Pig queries?
b. How can the company use indexes or partitioning to improve query performance?
c. What are the trade-offs between using Hive, Pig, and other SQL-on-Hadoop frameworks for querying large datasets?

70. Scenario: A company is using Hadoop for real-time data processing and they need to implement a data pipeline to ingest and process data from multiple sources.
a. How can the company use tools like Flume, Kafka, and Sqoop to ingest data into HDFS?
b. What are the trade-offs between different data ingestion and transportation methods in Hadoop?
c. How can the company use Spark Streaming or Storm to process real-time data in Hadoop