Apache Hadoop Interview Questions & Answers

Beginner-Level Questions

1. What is Apache Hadoop?

Answer: Apache Hadoop is an open-source framework that allows distributed processing of large datasets across clusters of computers.

2. What are the main components of Hadoop?

Answer: HDFS, YARN, MapReduce, and Common Utilities.

3. What is HDFS?

Answer: Hadoop Distributed File System stores large files across multiple nodes with replication.

4. What is the default block size?

Answer: 128 MB in Hadoop 2.x/3.x.

5. What is the role of NameNode and DataNode?

Answer: NameNode manages metadata; DataNode stores actual data blocks.

6. What happens if a DataNode fails?

Answer: Data is retrieved from other replicas.

7. What is MapReduce?

Answer: A programming model with Map and Reduce phases for distributed data processing.

8. What is YARN?

Answer: Resource manager for Hadoop. Components: ResourceManager, NodeManager, ApplicationMaster.

9. Default replication factor?

Answer: 3.

10. Difference between RDBMS and Hadoop?

Answer: Hadoop handles unstructured data and scales horizontally.

Intermediate-Level Questions

11. How does Hadoop achieve fault tolerance?

Answer: Through data replication in HDFS.

12. Difference between Mapper and Reducer?

Answer: Mapper outputs key-value pairs; Reducer aggregates results.

13. Role of Secondary NameNode?

Answer: Merges edit logs and fsimage for checkpointing.

14. How does Hadoop handle data locality?

Answer: Moves computation close to data.

15. What is a Combiner?

Answer: Local reducer that minimizes data transfer.

16. Hadoop daemons?

Answer: NameNode, DataNode, ResourceManager, NodeManager, Secondary NameNode.

17. Data flow in MapReduce?

Answer: HDFS -> Mapper -> Shuffle/Sort -> Reducer -> HDFS.

18. What are input splits?

Answer: Logical chunks of data for each Mapper.

19. Common file formats?

Answer: Text, Avro, Parquet, ORC, SequenceFile.

20. What is speculative execution?

Answer: Runs duplicate tasks to mitigate slowness.

Advanced-Level Questions

21. How does YARN manage resources?

Answer: Through ResourceManager and ApplicationMaster scheduling.

22. How does Hadoop ensure data consistency?

Answer: Single-writer model; append-only files.

23. Integrating Hadoop with Spark or Hive?

Answer: Hive for SQL queries; Spark for fast in-memory processing.

24. Hadoop tuning?

Answer: Optimize block size, enable compression, balance cluster load.

25. Common bottlenecks?

Answer: Disk/Network I/O, skewed data. Solutions: compression, data locality.

26. MapReduce limitations vs Spark?

Answer: Spark is in-memory and faster; MapReduce is disk-based.

27. What is rack awareness?

Answer: Data replicas placed on different racks for reliability.

28. Handling small files?

Answer: Combine into SequenceFiles or use HAR files.

29. Checkpointing and recovery?

Answer: Secondary NameNode merges logs to create checkpoints.

30. Real-world use case?

Answer: Log analysis with HDFS storage, Hive queries, Pig transformations.

Scenario-Based Questions

31. Frequent DataNode failures?

Answer: Check logs, replication factor, disk health.

32. Slow Map task?

Answer: Data skew or hardware issue; enable speculative execution.

33. Efficient data ingestion?

Answer: Use Flume/Sqoop with compression and partitioning.

34. Monitoring cluster health?

Answer: Use Ambari, Ganglia, or Cloudera Manager.

35. Securing Hadoop?

Answer: Use Kerberos, HDFS encryption, and ACLs.