Two-Marks Questions

1. **Give two limitations of traditional data processing systems that led to Big Data technologies.**
2. **Who developed Hadoop, and what problem was it originally designed to solve?**
3. **What are the main components of Apache Hadoop?**
4. **Explain Hadoop Streaming in brief.**
5. **List any two components of the Hadoop Ecosystem and their purpose.**
6. **What is IBM's Big Data Strategy? Mention any two focus areas.**
7. **What is BigInsights? Mention any two features.**
8. **What are BigSheets? Mention any two uses.**
9. **State any two design goals of HDFS.**
10. **Mention two Hadoop command-line interface (CLI) commands and their usage.**
11. **What is Hadoop Data Flow? Explain in two points.**
12. **What is data ingestion? Name any two tools used in Hadoop for ingestion.**
13. **What are Hadoop Archives (HAR)? Why are they used?**
14. **What is the purpose of compression in Hadoop I/O? Mention any one codec.**
15. **What is Avro? State any two advantages of using Avro.**
16. **Describe the key steps in a MapReduce job execution (any two).**
17. **What is the role of the JobTracker (or ResourceManager in YARN)?**
18. **What is shuffle and sort in MapReduce?**
19. **Differentiate between Text Input Format and KeyValue Input Format.**
20. **What is Pig Latin? Mention any two characteristics.**
21. **List the execution modes of Apache Pig.**
22. **What are User Defined Functions (UDFs) in Pig? Give one use.**
23. **What is Hive Metastore? Mention its role.**
24. **Compare HBase with RDBMS in two points.**

Four-Mark Questions

1. **Explain any four characteristics of Big Data.**
2. **Write any four challenges associated with Big Data processing.**

4. **What are the main advantages of using Hadoop for Big Data? (Any four)**

7. **Describe how the awk command helps in data analysis with two examples.**
8. **List four steps involved in analysing data using the Hadoop framework.**
9. **Explain any four benefits of using Hadoop Streaming.**

10. **List any four components of the Hadoop ecosystem with their functions.**
11. **Write four focus areas of IBM's Big Data Strategy.**
12. **Explain four features of IBM BigInsights.**

13. **Explain the importance of block size in HDFS and list three advantages of large block sizes.**
14. **List and briefly explain any four Hadoop shell commands.**

16. **Explain HDFS data flow during file read or write operations (any four steps).**
17. **Give four differences between Flume and Sqoop.**
18. **Write any four uses of Hadoop Archives (HAR).**

19. **List four advantages of using compression in Hadoop.**
20. **Explain any four features of Avro serialization.**
21. **Describe four features of SequenceFiles.**

22. **List and explain any four features of MapReduce.**
23. **Write four causes of task failure in MapReduce jobs.**
24. **Explain the importance of shuffle and sort phase (any four points).**

25. **Explain the two execution modes of Pig with two advantages of each.**

    .

26. **Explain any four data processing operators in Pig (e.g., FILTER, FOREACH, GROUP, JOIN).**

28. **Briefly explain the role of Hive Metastore in four points.**
29. **Write any four differences between Hive and traditional RDBMS.**
30. **Explain four features of HiveQL.**

Five-Mark Questions

1. **Explain structured, semi-structured, and unstructured data with examples. How does each type impact Big Data processing?**
2. **Discuss the 5 V's of Big Data with suitable examples.**
3. **Explain the importance of Big Data Analytics in modern enterprises with at least three real-world applications.**

4. **Describe the evolution of Hadoop from Google's architecture, highlighting GFS and MapReduce influences.**
5. **Explain the architecture of Hadoop 1.x. How does it differ from Hadoop 2.x (YARN)?**
6. **Discuss the major components of Apache Hadoop and their roles.**

7. **Explain how Unix tools like grep, awk, and sort are used to analyse large datasets. Provide examples.**
8. **Discuss the steps involved in analysing data with Hadoop.**
9. **Explain Hadoop Streaming and illustrate how it enables the use of non-Java programming languages.**

10. **Describe any five components of the Hadoop Ecosystem and their significance.**
11. **Explain IBM's Big Data Strategy. How does it address volume, variety, and velocity?**
12. **Explain the key features of IBM BigInsights and how BigSheets supports business analytics.**

13. **Explain the design principles behind HDFS. Why is it suitable for Big Data workloads?**
14. **c**
15. **Explain the Hadoop command-line interface (CLI) with examples of at least four important commands.**

16. **Describe the data flow in Hadoop from input to output in MapReduce.**
17. **Explain data ingestion in Hadoop using Flume and Sqoop with architecture diagrams.**
18. **What are Hadoop Archives (HAR)? Describe their structure, purpose, and benefits.**

19. **Explain the importance of compression in Hadoop. Compare at least three compression codecs.**
20. **Describe Avro's data serialization system, its schema evolution, and benefits.**
21. **Write short notes on: SequenceFiles, MapFiles, and Parquet.**

22. **Explain the complete anatomy of a MapReduce job execution from job submission to completion.**
23. **Discuss different types of failures in MapReduce and how Hadoop handles them.**
24. **Explain job scheduling in Hadoop and compare FIFO, Fair, and Capacity schedulers.**
25. **Describe the shuffle and sort mechanism in MapReduce and its importance.**
26. **Discuss various MapReduce input and output formats with examples.**

27. **Explain the architecture of Apache Pig and its execution modes.**
28. **Compare Pig Latin with SQL, highlighting similarities and differences.**
29. **Explain the role of User Defined Functions (UDFs) in Pig with examples.**

30. **Explain the Hive architecture including Hive Shell, Metastore, Drivers, and Execution Engine.**

10 Ten-Mark Questions

1. **Discuss in detail the architecture of Apache Hadoop, explaining the roles and interactions of HDFS, MapReduce, and YARN. Include a neat labelled diagram.**

2. **Explain the design of HDFS with emphasis on block replication, fault tolerance, heartbeats, rack awareness, and data locality. How do these features make HDFS suitable for large-scale data storage?**

3. **Describe the complete lifecycle of a MapReduce job, from job submission to completion. Explain job initialization, task assignment, shuffle and sort, task execution, and job finalization in detail.**

4. **Write a detailed note on Hadoop I/O. Discuss compression techniques, serialization (Writable types), Avro architecture, and common Hadoop file-based data structures such as SequenceFile, MapFile, and Parquet.**

5. **Explain the architecture and functioning of Apache Pig. Discuss execution modes, features of Pig Latin, data types, major operators, and the role of User Defined Functions (UDFs). Provide suitable examples.**

6. **Describe the architecture of Apache Hive in detail, including Hive Shell, Driver, Compiler, Optimizer, Execution Engine, Metastore, and the role of HiveQL. Compare Hive with traditional RDBMS on at least five parameters.**

7. **Explain the architecture of HBase. Discuss HBase data model, storage mechanism (HFiles, MemStore, WAL), RegionServers, Master, ZooKeeper coordination, and how HBase differs from RDBMS. Include a diagram.**

8. **Compare and contrast data ingestion using Flume and Sqoop. Discuss their architectures, components, workflows, use cases, advantages, and limitations. Provide example command flows for both tools.**

9. **Explain IBM BigInsights platform in detail. Discuss its architecture, components (BigSheets, Big SQL, Analytics Accelerator), integration with Hadoop ecosystem, and use cases of IBM's Big Data Strategy.**

10. **Discuss job scheduling in Hadoop. Explain FIFO, Fair Scheduler, Capacity Scheduler, their architectures, scheduling policies, advantages, limitations, and suitable use cases.**

11. Apply MapReduce Model for Matrix- Vector Multiplication.

3	4	1	*	1
5	6	2		2
				3

12. Solve using MapReduce Model for Matrix- Matrix Multiplication by showing all the steps

3	4	*	1	2	7
5	6		8	9	0

Application Based Questions

1. A healthcare company wants to analyse patient records (images, text, lab reports). Explain which types of digital data are involved and how Big Data techniques can help.
2. A social media platform receives millions of posts daily. Explain why Hadoop is a suitable framework for processing this data.
3. A startup wants to store 500 TB of video data. Explain how HDFS replication and block size decisions would impact performance and reliability.
4. A retail company wants to collect live clickstream data from its e-commerce site. Explain how Flume can be configured for this use case.
5. A bank wants to migrate historical customer data from MySQL to Hadoop for fraud analytics. Explain how Sqoop can be used.
6. A data engineer using Unix tools wants to quickly extract error logs from large server files. Explain how `grep`, `awk`, and `sort` help in this process.
7. A telecom company wants to calculate call drop frequency from large log files. Design an application scenario using MapReduce.
8. A data lake stored in Hadoop has large JSON files. Suggest suitable compression codecs and justify your selection.
9. A media company wants efficient columnar queries for user metrics. Recommend a file format (e.g., Parquet) and justify why it fits the requirement.
10. A university wants to process student activity logs (login time, attendance, submissions). Explain how Pig Latin scripts can be used.
11. A business intelligence team wants to run analytical queries on structured warehouse data. Explain how Hive helps them.
12. A mobile company needs fast random reads/writes for user profiles. Explain why HBase is suitable.
13. A smart city project generates sensor data on traffic, pollution, and waste management. Describe how Big Data Analytics can help decision-making.
14. A logistics company wants to track vehicle movement in real time. Explain how Hadoop can handle large streaming data.
15. A video platform wants to store high-resolution media files. Explain how HDFS fault tolerance helps in uninterrupted user access.
16. A company uploads large CSV datasets into HDFS daily. Describe the data flow of a file write operation from the client to DataNodes.
17. A data engineer needs to manage huge directories of log files. Explain the use of HDFS CLI commands (`ls`, `du`, `put`, `rm`) in this context.
18. An AI team wants to preprocess and clean data before model building. Recommend four Hadoop ecosystem tools and justify their usage.
19. A newspaper company wants to find the most commonly used words in online articles. Explain how a MapReduce job can be designed.
20. A cloud company has multiple users submitting Hadoop jobs. Explain how Fair Scheduler ensures resource sharing.
21. A fraud detection system needs high-speed read/write. Recommend efficient file formats (e.g., SequenceFile) and justify.

22. A retail chain needs to summarize daily sales records. Explain how Pig's GROUP and FOREACH operators can help.
23. A business analyst wants to run SQL-like queries on massive structured data. Explain how HiveQL simplifies the task.
24. A messaging app stores chat history of millions of users. Explain how HBase enables real-time querying.
25. A financial organization handles PDFs, transaction logs, and emails. Categorize each as structured, semi-structured, or unstructured, explaining why.
26. A weather forecasting system collects terabytes of satellite data daily. Explain how Big Data technologies help improve predictions.
27. An insurance company needs scalable storage and parallel processing. Explain how Hadoop 2.x (YARN) architecture supports this need.
28. A ride-sharing company wants to ingest real-time trip data. Explain a suitable ingestion pipeline using Flume.
29. An enterprise migrates its existing Oracle DB to Hadoop. Explain the Sqoop workflow for bulk import and incremental load.
30. A robotics lab wants tools for scripting, search, and data modification. Explain how Unix tools assist in data preparation.
31. A stock exchange collects tick-by-tick data. Recommend a serialization framework (e.g., Avro) and justify your choice.
32. An online exam platform wants to calculate average student performance. Explain the MapReduce stages required.
33. A data warehouse wants fast analytical queries on big tables. Explain why Parquet columnar format is a suitable solution.
34. A transport company wants to analyze GPS logs for route optimization. Explain how Pig operators can be applied.
35. A marketing team wants to run complex analytical queries like joins and aggregations. Explain how Hive supports this requirement.
36. A global e-commerce platform needs fast lookup for product IDs. Explain how HBase handles this with its architecture.