Question Bank (BCADS)

Section A: Introduction to Big Data (1–15)

1. Define Big Data and explain its key characteristics (Volume, Velocity, Variety, Veracity, Value).
2. Describe real-life examples where Big Data is used and how it impacts decision-making.
3. Explain the different types of Big Data with suitable examples.
4. Describe how IoT contributes to the generation and growth of Big Data.
5. Explain the evolution from traditional data processing to Big Data processing.
6. Discuss the challenges faced when working with Big Data.
7. List and explain Big Data use cases in various domains.
8. Explain the need for a Big Data strategy in modern organizations.
9. Describe the role of interconnected devices in Big Data growth.
10. List and describe the components of the open-source Hadoop ecosystem.
11. Explain the future directions and advancements expected in Hadoop and Big Data.
12. Compare structured, semi-structured, and unstructured data with examples.
13. Discuss the benefits of using Big Data over traditional data warehousing systems.
14. Illustrate the shift in data analytics techniques due to Big Data tools.
15. Describe the importance of parallel processing in Big Data.

Section B: RDBMS and HDFS (16–30)

16. Explain the difference between DDL, DML, and DCL commands with examples.
17. Write SQL queries to demonstrate the use of SELECT, INSERT, UPDATE, and DELETE commands.
18. Describe the architecture of HDFS and its core components.
19. Explain the function and roles of NameNode and DataNode in HDFS.
20. How are files split and stored in HDFS? Explain with a diagram.
21. Describe HDFS replication and its importance in fault tolerance.
22. Write and explain basic HDFS commands to create, read, move, and delete files.
23. Compare HDFS with traditional file systems.
24. Explain the benefits of HDFS in a distributed computing environment.
25. Discuss the need for high internode network speed in Hadoop clusters.
26. Demonstrate block-level file storage in HDFS with a suitable example.
27. Describe the write and read path in HDFS.
28. Explain the HDFS command-line interface with syntax and examples.
29. Explain the purpose of metadata in HDFS.
30. Compare local file system and HDFS in terms of scalability and reliability.

Section C: Hortonworks and Ambari (31–45)

31. Describe the components and architecture of Hortonworks Data Platform (HDP).
32. What are IBM-added value components in HDP and their purposes?
33. Explain the role and benefits of Apache Ambari in a Hadoop environment.
34. Describe the overall architecture of Ambari.

35. Explain how Ambari integrates with other services in a Hadoop cluster.

36. List and explain Ambari's main features for system monitoring.

37. Discuss how to install and configure Apache Ambari.

38. Describe how to start and stop Hadoop services using Ambari Web UI.

39. Compare Ambari with other cluster management tools.

40. Explain Ambari alerts and their role in cluster management.

41. Describe user and role management in Ambari.

42. Demonstrate the service monitoring dashboard in Ambari.

43. Explain the architecture and functionality of Ambari Metrics System.

44. Describe how to use Ambari to manage configurations.

45. What is Ambari Views and how can it be used to manage Big Data?

Section D: Hive and Pig (46–60)

46. Explain the architecture of Hive and its integration with Hadoop.

47. What is bucketing in Hive? Explain with example.

48. What is partitioning in Hive? Explain static vs dynamic partitioning.

49. Write HiveQL commands to create a table with bucketing and partitioning.

50. Explain how Hive executes queries on HDFS data.

51. Discuss data types supported by Hive and their usage.

52. Compare Hive and RDBMS.

53. Describe how to load data into a Hive table.

54. Explain Hive query optimization techniques.

55. What is Pig? Describe its role in Big Data analytics.

56. Compare Pig and Hive.

57. Explain Pig's data model with examples.

58. Write a Pig script to read a file and perform filtering and aggregation.

59. Describe the architecture of Apache Pig.

60. Explain the flow of execution in a Pig script.

Section E: MapReduce and Spark (61–80)

61. Describe the MapReduce programming model with an example.

62. Write a MapReduce job in Python to count words in a file.

63. Explain the role of Mapper and Reducer classes in Hadoop.

64. Describe the Hadoop v1 architecture and its limitations.

65. Explain the improvements introduced in Hadoop v2 (YARN).

66. Compare Hadoop v1 with Hadoop v2.

67. Describe the architecture of YARN.

68. What is Apache Spark? Explain its role in Big Data processing.

69. Describe the Apache Spark unified stack and its components.

70. Explain the concept of RDD in Spark.

71. Compare RDD, DataFrame, and Dataset in Spark.

72. Describe how Spark achieves fault tolerance.

73. List and describe Apache Spark libraries (Spark SQL, MLlib, GraphX, Streaming).
74. Demonstrate the use of Spark Shell (Scala) to filter data from a CSV.
75. Describe Spark transformations and actions with examples.
76. Write a Spark script in Python to find the average of numbers in a file.
77. Explain the difference between batch and stream processing in Spark.
78. Compare Spark with MapReduce.
79. Describe how Spark handles lazy evaluation.
80. Explain SparkContext and SparkSession.

Section F: Practical & Case-Based Questions (81–100)

81. Write and explain a Java MapReduce program to find max temperature in weather data.
82. Install Hadoop in pseudo-distributed mode and demonstrate basic HDFS operations.
83. Create a Hive table, load data, and run sample SELECT queries.
84. Use Spark to count word frequency from a text file in PySpark.
85. Demonstrate Ambari's process to start and stop Hadoop services.
86. Show how to monitor cluster health using Ambari metrics.
87. Write and explain a Pig script for student marks analysis.
88. Design a Big Data pipeline using Hive and Pig.
89. Demonstrate Spark SQL by querying structured data in a DataFrame.
90. Show an example of bucketing and partitioning combined in Hive.
91. Create and run a job in Spark that filters sales data by date range.

92. Describe a real-world use case of IoT contributing to Big Data.

93. Show a use case for real-time processing using Spark Streaming.

94. Perform a file operation in HDFS using command line and interpret logs.

95. Design a cluster layout and explain network communication strategy.

96. Analyze a Big Data case study from healthcare or finance.

97. Write a comparative report on Hadoop distributions: Cloudera vs Hortonworks.

98. Show how to troubleshoot a failed Hadoop service using Ambari logs.

99. Simulate data ingestion in Hive using HDFS and validate data integrity.

100. Create a step-by-step guide to installing and configuring Apache Spark.
