VIVA Question and Answers

Week-I

Hadoop Basics & Ecosystem

1. What is Big Data?   
   Big Data refers to large volumes of structured, semi-structured, or unstructured data that are difficult to process using traditional tools.
2. What is Hadoop?   
   Hadoop is an open-source framework that enables distributed storage and processing of large data sets across clusters of computers.
3. What are the core components of Hadoop?
   * HDFS (Hadoop Distributed File System)
   * MapReduce (Computational model)
   * YARN (Resource management)
4. What does HDFS do?   
   HDFS stores vast amounts of data across multiple machines while ensuring fault tolerance and high throughput.
5. What are the two main components of HDFS?
   * NameNode: Manages metadata and the file system
   * DataNode: Physically stores data blocks

Week-2

MapReduce & YARN

1. What is MapReduce?   
   MapReduce is a programming paradigm that splits processing into two phases—Map and Reduce—to handle large-scale data.
2. What are Mapper and Reducer roles?
   * Mapper: Takes input and generates key-value pairs
   * Reducer: Aggregates those pairs to produce the final output
3. What is YARN?   
   YARN stands for Yet Another Resource Negotiator; it manages and schedules application resources across the Hadoop cluster.
4. How does Hadoop ensure fault tolerance?   
   HDFS replicates data across multiple nodes (default is 3), so data remains available even if nodes fail.

Week 3: Viva Questions – MapReduce

10. Q: What is MapReduce?

A: MapReduce is a programming model used for processing and generating large datasets with a distributed algorithm on a cluster.

11. Q: What are the main components of MapReduce?

A: Mapper and Reducer.

12. Q: What is the role of a Mapper?

A: It processes input data and produces a set of intermediate key-value pairs.

13. Q: What is the role of a Reducer?

A: It aggregates the output of the Mapper and provides the final output.

14. Q: What is the input format in MapReduce?

A: Typically, it is TextInputFormat which treats each line of input as a record.

15. Q: What is the output format in MapReduce?

A: TextOutputFormat which writes records in plain text.

16. Q: How does MapReduce handle failure?

A: By rerunning the failed task on another node.

17. Q: What is the purpose of the combiner?

A: It performs local aggregation before data is shuffled to the reducer.

18. Q: What is the shuffle and sort phase?

A: It is the process between Map and Reduce that transfers and sorts intermediate data.

19. Q: What is a key-value pair in MapReduce?

A: It is the data format used to pass data between Mapper and Reducer.

20. Q: How do you run a MapReduce job?

A: Using the command `hadoop jar <jarfile> <mainclass> <input> <output>`.

21. Q: What are Partitioner and their use?

A: They decide which reducer will receive a particular key.

22. Q: What is the default number of reducers?

A: 1 unless specified otherwise.

23. Q: What happens if the output directory already exists?

A: The job fails unless the directory is deleted beforehand.

24. Q: How do you view the job status?

A: Using the ResourceManager UI or CLI via `mapred job -status <jobid>`.

25. Q: What is a counter in MapReduce?

A: A mechanism to measure progress or track custom metrics.

26. Q: What is speculative execution?

A: It runs duplicate instances of slow tasks to finish the job faster.

27. Q: What is the role of InputSplit?

A: It defines the chunks of input for each Mapper.

28. Q: Can we have multiple mappers or reducers?

A: Yes, depending on the input size and configuration.

29. Q: What are common errors in MapReduce jobs?

A: File already exists, memory errors, or permission denied.

Week 4: Viva Questions – Hive

30. Q: What is Hive?

A: Hive is a data warehouse tool used for querying and analyzing large datasets using a SQL-like language called HiveQL.

31. Q: What is HiveQL?

A: It is the query language used in Hive, similar to SQL.

32. Q: What is the default database in Hive?

A: The default database is 'default'.

33. Q: How do you create a table in Hive?

A: Using `CREATE TABLE table\_name (col1 type, col2 type);`

34. Q: What is the difference between internal and external tables?

A: Internal tables manage data, external tables just store metadata.

35. Q: How do you load data into a Hive table?

A: Using `LOAD DATA LOCAL INPATH 'file' INTO TABLE table\_name;`

36. Q: What is the default file format in Hive?

A: TextFile format.

37. Q: What is a partition in Hive?

A: It divides a table into related parts based on column values.

38. Q: What is bucketing in Hive?

A: It divides data into equal parts (buckets) based on a hash function.

39. Q: How do you view tables in Hive?

A: Using `SHOW TABLES;`

40. Q: How do you describe a table in Hive?

A: Using `DESCRIBE table\_name;`

41. Q: How can you run a query in Hive?

A: Using the Hive shell or Beeline CLI.

42. Q: What is the command to list databases?

A: SHOW DATABASES;

43. Q: What is Metastore?

A: It is the central repository to store metadata information in Hive.

44. Q: What are managed tables?

A: Tables whose data is managed by Hive.

45. Q: How do you drop a table in Hive?

A: Using `DROP TABLE table\_name;`

46. Q: What is the difference between Hive and RDBMS?

A: Hive is for batch processing, while RDBMS is for transaction-based operations.

47. Q: What are SerDe in Hive?

A: Serializer and Deserializer used to read/write data in Hive.

48. Q: What is the use of LIMIT in Hive?

A: To limit the number of output rows.

49. Q: What is dynamic partitioning in Hive?

A: Partitions are created automatically while loading data.

Week 5: Viva Questions – Pig

50. Q: What is Apache Pig?

A: Pig is a high-level platform for creating MapReduce programs used with Hadoop.

51. Q: What is Pig Latin?

A: It is the language used to write Pig scripts.

52. Q: What is a relation in Pig?

A: A relation is a bag of tuples.

53. Q: What is the difference between Pig and Hive?

A: Pig is procedural and used for ETL, Hive is declarative and used for querying.

54. Q: What are the modes of Pig execution?

A: Local mode and MapReduce mode.

55. Q: What is a bag in Pig?

A: An unordered collection of tuples.

56. Q: What is a tuple in Pig?

A: An ordered set of fields.

57. Q: How do you load data in Pig?

A: Using `LOAD 'file' USING PigStorage(',');`

58. Q: What is the purpose of DUMP in Pig?

A: It displays the content of a relation on the screen.

59. Q: What is the purpose of STORE in Pig?

A: It saves the output to HDFS.

60. Q: What is the purpose of FILTER in Pig?

A: To filter data based on a condition.

61. Q: How do you group data in Pig?

A: Using the GROUP operator.

62. Q: How do you join datasets in Pig?

A: Using the JOIN operator.

63. Q: What is a UDF in Pig?

A: User Defined Function written in Java or Python.

64. Q: What is the SPLIT operator?

A: Used to split a relation into two or more relations.

65. Q: What is the difference between FOREACH and GENERATE?

A: FOREACH is used to iterate over tuples; GENERATE selects and transforms data.

66. Q: Can Pig handle unstructured data?

A: Yes, Pig can handle both structured and semi-structured data.

67. Q: What is the use of DESCRIBE in Pig?

A: To describe the schema of a relation.

68. Q: How do you remove duplicates in Pig?

A: Using DISTINCT operator.

69. Q: What is the purpose of ORDER BY?

A: To sort the data in ascending or descending order.

Week 6: Viva Questions – Sqoop, Flume, and Workflows

70. Q: What is Apache Sqoop?

A: Sqoop is a tool designed for transferring data between Hadoop and relational databases.

71. Q: What is the purpose of Flume?

A: Flume is used to efficiently collect, aggregate, and move large amounts of log data.

72. Q: How do you import data from MySQL using Sqoop?

A: Using `sqoop import --connect <JDBC\_URL> --table <table\_name> --username <user> --password <pass>`

73. Q: How do you export data from Hadoop to RDBMS using Sqoop?

A: Using `sqoop export` command.

74. Q: What are Sqoop connectors?

A: They allow Sqoop to connect to external databases.

75. Q: What is the default file format for Sqoop imports?

A: Text files (CSV format).

76. Q: What is an agent in Flume?

A: A process that hosts the components: source, channel, and sink.

77. Q: What is a source in Flume?

A: It receives data into the Flume agent.

78. Q: What is a sink in Flume?

A: It outputs the data to the desired destination.

79. Q: What is a channel in Flume?

A: It acts as a bridge between the source and sink.

80. Q: What is a Flume configuration file?

A: A file that defines the source, sink, and channel of an agent.

81. Q: What is a workflow in Oozie?

A: A sequence of actions arranged in a control dependency DAG.

82. Q: What is the use of Oozie in Big Data?

A: To schedule and manage Hadoop jobs.

83. Q: What languages are used in Oozie workflows?

A: XML for workflow definitions.

84. Q: What is the command to run a workflow in Oozie?

A: `oozie job -oozie <URL> -config <properties\_file> -run`

85. Q: What is a coordinator job in Oozie?

A: It triggers workflows based on time or data availability.

86. Q: Can Sqoop import specific columns?

A: Yes, using the `--columns` parameter.

87. Q: Can Flume handle streaming data?

A: Yes, it is designed for continuous data ingestion.

88. Q: Can Oozie trigger multiple workflows?

A: Yes, using coordinator jobs.

89. Q: What is the importance of scheduling in Big Data workflows?

A: It automates job execution, improving efficiency and repeatability.