**HIVE INTERVIEW QUESTIONs – SOLUTIONS**

1. Hive is a data warehousing system for processing large datasets. The current version is Apache Hive 3.1.2.

2. Hive is not suitable for OLTP systems as it is designed for batch processing of large datasets.

3. Hive is different from RDBMS as it uses SQL-like syntax for querying distributed data. It does not support full ACID transactions.

4. Hive architecture consists of a driver, a query compiler, a metastore, and Hadoop Distributed File System (HDFS).

5. The Hive query processor executes queries and consists of a parser, an optimizer, and an executor.

6. Hive can operate in local mode, MapReduce mode, or Spark mode.

7. Features of Hive include SQL-like querying, support for large datasets, and integration with Hadoop. Limitations include slower query performance and lack of ACID transactions.

8. To create a database in Hive, use the CREATE DATABASE command.

9. To create a table in Hive, use the CREATE TABLE command.

10. DESCRIBE displays metadata about a database or table. DESCRIBE FORMATTED displays metadata in a more structured format.

11. Use the TBLPROPERTIES clause with the 'skip.header.line.count' property to skip header rows from a table.

12. Hive operators are used in queries to perform operations on data. Types of operators include comparison, arithmetic, and logical.

13. Hive built-in functions perform operations on data in queries. Types of functions include mathematical, string, and date/time.

14. DDL commands create and modify database and table structures. DML commands manipulate data within tables.

15. SORT BY orders data within a reducer. ORDER BY orders data globally. DISTRIBUTE BY controls how data is partitioned. CLUSTER BY is a shorthand for DISTRIBUTE BY and SORT BY.

16. Internal tables are managed by Hive and stored in HDFS. External tables are registered with Hive but data is stored outside of HDFS.

17. Data of a Hive table is stored in HDFS.

18. Yes, the default location of a managed table can be changed using the LOCATION clause in the CREATE TABLE statement.

19. Metastore stores metadata about databases, tables, and columns in Hive. The default database provided by Hive for metastore is Derby.

20. Hive does not store metadata information in HDFS because it is not designed for transactional updates.

21. A partition in Hive is a way to organize data within a table based on a specific column or columns. Partitioning allows for faster querying of specific subsets of data.

22. Dynamic partitioning creates partitions as data is inserted into a table. Static partitioning creates predefined partitions before data is inserted into a table.

23. Use the SHOW PARTITIONS command to check if a particular partition exists.

24. Use the ALTER TABLE command with the EXCLUDE clause to prevent a partition from being queried.

25. Buckets are used to improve performance by reducing data size. Hive distributes rows into buckets based on a hash function applied to a specific column.

26. To enable buckets in Hive, use the CLUSTERED BY clause in the CREATE TABLE statement.

27. Bucketing helps in faster query execution by reducing the amount of data that needs to be scanned.

28. To optimize Hive performance, use partitioning, bucketing, and indexing. Additionally, use efficient query design and consider using Tez or LLAP as execution engine.

29. HCatalog is a metadata and table management system for Hadoop that provides a unified view of data stored in Hive, HBase, and Pig.

30. Joining tables in Hive can be done using different types of joins, including inner join, full outer join, and left outer join.

31. Yes, it is possible to create a Cartesian join between 2 tables, using Hive.

32. Sort-Merge-Bucket join (SMB join) is a type of join in Hive that uses bucketing to improve performance.

33. SORT BY orders data within reducers while retaining grouping. ORDER BY sorts data globally. Use ORDER BY for single reducers and SORT BY for scaling.

34. The DISTRIBUTED BY clause in Hive controls how data is partitioned within a reducer.

35. Data is transferred from HDFS to Hive through a mapping layer called Hive Input Format.

36. Hive creates a new metastore\_db wherever the query is run because it uses a local instance of an embedded Derby database to store the metastore.

37. If 'SET hive.enforce.bucketing=true;' is not issued before bucketing a table in Hive, the table will not be properly bucketed.

38. Yes, a table can be renamed in Hive using the ALTER TABLE command.

39. Use the ALTER TABLE command with the ADD COLUMNS and BEFORE clauses to insert a new column in a table at a specific position.

40. SerDe (Serializer/Deserializer) is a mechanism in Hive for transforming data to and from external storage formats.

41. Hive deserializes data into an internal format when reading from external storage. It serializes data into external formats when writing data to external storage.

42. The default built-in SerDe in Hive is LazySimpleSerDe.

43. Custom SerDe is used in Hive when the built-in SerDe does not support a specific input or output format.

44. Complex data types in Hive include arrays, maps, and structs.

45. Hive queries can be executed from script files using the command: 'hive -f file\_name.hql'

46. The default record delimiter is '\n' and the default field delimiter is '\t' for Hive text files.

47. Use the SHOW DATABASES command with a LIKE clause to list all databases in Hive whose name starts with s.

48. LIKE is used to match patterns while RLIKE uses regular expressions for matching in Hive.

49. Use the ALTER TABLE command with the CHANGE COLUMN clause to change column data types in Hive.

50. Use the CAST function to convert the string '51.2' to a float value in the particular column.

51. When 'abc' (string) is cast as INT, the result will be NULL.

52. The query inserts data from the staged\_employees table into the employees table partitioned by country and state.

53. Use the INSERT OVERWRITE TABLE command followed by a SELECT statement to overwrite data in a new table from an existing table.

54. The maximum size of a string data type in Hive is 2GB. Hive supports binary formats through the SerDe mechanism.

55. Hive supports file formats like ORC, Avro, Parquet, and applications like JDBC, ODBC.

56. ORC format tables have column-level compression and predicate pushdown, leading to better performance.

57. Hive can avoid mapreduce by using Tez, a DAG-based execution engine, and vectorization.

58. A view is a virtual table that acts as a filter, and indexing improves query speed by creating a smaller index table.

59. Yes, a view can have the same name as a hive table, but the schema should differ.

60. The costs associated with creating indexes are increased storage requirements and overhead for index maintenance during updates.

61. The command SHOW INDEXES ON table\_name; can be used to see the indexes on a table.

62. Subdirectories can be accessed recursively in Hive queries by setting the property mapred.input.dir.recursive to true.

63. A select \* query in Hive doesn't run MapReduce if the table metadata is cached in memory.

64. Hive Explode is used to transform arrays and maps into separate rows to make them queryable.

65. JDBC and ODBC drivers can be used to connect applications to Hive when run as a server.

66. Yes, the default location of a managed table can be changed in Hive by setting the table's location property.

67. Hive ObjectInspector is a class that inspects and returns the metadata of Hive table columns.

68. A UDF in Hive, or User-Defined Function, is a custom function that can be used in Hive queries.

69. The LOAD DATA INPATH query can be used to extract data from HDFS to Hive.

70. TextInputFormat reads data as plain-text, while SequenceFileInputFormat reads data stored in Sequence Files.

71. A large job can be prevented from running for a long time by optimizing the query, partitioning the data, and setting resource quotas.

72. Explode is used to unnest and transform complex data types in Hive queries.

73. Hive can process different data formats, but it performs best with columnar formats like ORC, Parquet, and Avro.

74. Hive creates a new metastore\_db due to its architecture that stores the metadata of the tables.

75. Yes, the ALTER TABLE query can be used to change the data type of a column in a Hive table.

76. By specifying the hdfs:// prefix before the HDFS file path in Hive's LOAD DATA clause.

77. Hive configuration has a precedence order: Session > HiveServer2 Configuration > hive-site.xml file > System Defaults.

78. The Hive metastore is accessed using the thrift interface, which communicates with the metastore service.

79. Yes, JSON can be compressed in the Hive external table using compression codecs like Gzip, Bzip2, etc.

80. Local metastores run on the same machine as Hive, while remote metastores run on a separate machine.

81. Archiving tables in Hive can help in the management of historical data.

82. DBPROPERTY function is used to retrieve metadata about a Hive database.

83. Local mode runs Hive queries on a single machine, while MapReduce mode uses Hadoop clusters to process the queries.