What is Hive?

Hive is data warehousing software, which uses similar language like SQL named HQL and it is used write queries which will be further convert to Map-Reduced code, Hive is built on the top of HDFS layer.

What is the present version of HIVE?

4.0 Alpha 2.

Does hive uses OLTP?

Hive is a data warehousing service, which uses OLAP which stands for online analytical processing.

How hive is different from RDBMS?

RDBMS stands for relational database management system, which stored data in structured manner such as tables. Data is stored in normalized form, and doesn't support partitioning.

Hive is a data warehouse system, where the scheme is not fixed it varies. Data is stored in denormalized form. We can use partitioning and bucketing on it.

Does HIVE follow ACID properties?

Hive supports ACID properties, which enables the user to use transaction, create transaction, insert, update, delete on tables.

Hive started supporting ACID properties to address few concern :

Explain HIVE architecture and its different component?

Hive is built on top of the Hdfs storage and Map-Reduce. It has two parts on top of it:

1. HIVE CLIENT: Consist of JDBC server, ODBC server, Thrift Server.

2. HIVE SERVICES: Consist of HIVE servers 2, complier, driver, optimizer, Meta store.

JDBC Server, ODBC Server, Thrift Server works as an API service.

Hive Server 2- is the successor of HIVE Server 1, where query submitted by client through JDBC, ODBC, and Thrift comes.

HIVE Driver- Gets the HQL statement submitted by the user, then it transfer it to complier.

In complier, query goes through segmantic analysis and type checking according to the metadata stored in metastore. After that, it form an plan which in in the form of DAG ,where each stage is Map Reduce, operations on HDFs, a metadata operations.

Optimizer- Optimizer performs the transformation operation on the execution plan and splits task for efficiency.

Execution- It executes the execution plan.

Meta-store- stores all the information about the data, its partition, column, column type information.

Mention what does hive Query processor do?

Hive query processor converts the hive query to map-reduce codes. It helps developers to perform analysis and semi-structured data by replacing java complex codes.

Components of Hive query processor:

1. Parse and semantic analysis
2. Metadata layer
3. Type Interface
4. Sessions
5. Map/Reduce
6. Plan Components
7. Hive Function Framework
8. Tools

What are the three modes of HIVE?

Local mode: where Hadoop is stored in pseudo mode, possessing only data node. The data size is smaller thus the process is much faster.

Map reduce Mode: Hadoop has multiple data nodes, and the data is distributed across the nodes. Used when massive set of data is used. By default, map-reduce mode is set in Hive.

Features and Limitation of Hive:

Features:

1. Hive is used for querying and managing large set of data, it is fast and scalable.
2. There isn't any fixed schema.
3. When table created schema stored in database, while the data is stored in HDFS.
4. It supports various type of file format: ORC, Parquet.
5. Hive uses similar language like SQL i.e. HQL.
6. Only difference between then is HQL is executed on the top of HDFS and SQL on traditional database.
7. Supports partitioning and bucketing.

Limitations:

1. Hive is used for OLAP, it doesn't support OLTP.
2. High latency.

How can you create database in Hive?

To create database in Hive:

Create database database\_name;

How to create a table in HIVE?

There are two types of table:

1. Internal Table:

Create table employee (

Id int,

Name varchar (60),

Salary int)

row format delimited

fields terminated by ',';

1. External Table:

Create table employee (

Id int,

Name varchar (60),

Salary int)

row format delimited

fields terminated by ','

location 'hdfs-location';

What do you mean by describe and describe extended and describe formatted with respect to database and table?

Describe: this is used for getting the details of the table

Describe formatted: this is used to even get the minute details of the table, will show table columns, data type in multiple lines.

Describe extended: this is used to even get the minute details of the table, will show table columns, data type in single lines.

How do we skip header rows from a table in Hive?

tblproperties ("skip.header.line.count"= "1");

What is hive operator? What are the different types?

Hive operators are same as the other operator which is used for data operations in between values.

Types:

* Relational Operators
* Arithmetic Operators
* Logical Operators
* Complex Operators

Explain about the Hive Built-in-functions.

Built-in-functions are those which are already created like the string function, date function etc.

Write hive DDL & DML commands.

DDL COMMANDS- CREATE, TRUNCATE, DESCRIBE, ALTER,DROP.

DML COMMANDS- SELECT, UPDATE,LOAD, INSERT, DELETE.

Explain about sort by, order by, distribute by and cluster by in HIVE.

Sort by - This is used to sort the data within the data per reducer.

Order by - This is used to sort the total order of the data.

Cluster by - repartition data and then sort data based on key in each partition.

Distributed by- repartition data based on expression.

Difference between Internal and external tables. When to use what?

Internal table are those where the is directly loaded in the table, schema is created and data is loaded into it from local or hdfs. The table's metadata is stored in metastore and the data in hdfs If the table is deleted, then the data is also lost.

External Table are those where only the schema is created and the data will not be stored in Hive, there will a external reference to the raw data. If deleted, data will remain intact.

External table is used where the data is permanent and is used outside of the hive as well.

Internal table is used for temporary data files.

Where does the HIVE data gets stored?

Hive data gets stored in HDFS.

Is it possible to change the location of managed table?

Default location is ---- hive/users/warehouse

Yes, we can change it by using the location clause.

What is Metastore? What is the default database which Apache Hive provides?

Meta store is the place where all the information about the data, table and the column is stored.

Apache Hive provides a default database by the name of 'default

Why Hive stores data in metastore instead of storing in HDFS metastore?

Hive stores all the insight of the data in its metastore using RDBMS instead of HDFS, because of its low latency as the read and write operation in HDFS is a time consuming process.

What is a partition in Hive? And why do we perform partitioning in hive?

Partitioning in hive is a process of dividing the data on the basis of certain values of a column.

It is done so that the search process gets speeded up.

22. What is the difference between dynamic partitioning and static partitioning?

Static partitioning - where you externally mention the value of the column on which partition will be created.

Dynamic Partitioning - where Hive will fetch the unique values of the column and will create separate tables on those values.

.*How do you check if a particular partition exists?*

SHOW PARTITIONS table\_name

PARTITION (partitioned\_column=’partition\_value’)

. How can you stop a partition form being queried?

By using the enable offline clause after alter statement.

Why do we need buckets? How Hive distributes the rows into buckets?

The bucketing in Hive is a data organizing technique. It is similar to partitioning in Hive with an added functionality that it divides large datasets into more manageable parts known as buckets.

* The concept of bucketing is based on the hashing technique.
* Here, modules of current column value and the number of required buckets is calculated (let say, F(x) % 3).
* Now, based on the resulted value, the data is stored into the corresponding bucket.

In Hive, how can you enable buckets?

set hive.enforce.bucketing = true;

How does bucketing help in the faster execution of queries?

Bucketing in hive is the concept of breaking data down into ranges, which are known as buckets, to give extra structure to the data so it may be used for more efficient queries. The range for a bucket is determined by the hash value of one or more columns in the dataset (or Hive metastore table)

How to optimise Hive Performance? Explain in very detail.

By using partitioning and bucketing.

What is the use of Hcatalog?

HCatalog is a tool that allows you to access Hive metastore tables within Pig, Spark SQL, and/or custom MapReduce applications.

Explain about the different types of join in Hive.

There are different types of joins given as follows:

* JOIN
* LEFT OUTER JOIN
* RIGHT OUTER JOIN
* FULL OUTER JOIN

Is it possible to create a Cartesian join between 2 tables, using Hive?

Yes, Cartesian join is possible.

The CARTESIAN JOIN is also known as CROSS JOIN. In a CARTESIAN JOIN there is a join for each row of one table to every row of another table. This usually happens when the matching column or WHERE condition is not specified.

What is the difference between order by and sort by which one we should use?

In hive, order by clause is same as in SQL.

* There are some limitations in the "order by" clause. In the strict mode (i.e., [hive.mapred.mode](https://cwiki.apache.org/confluence/display/Hive/Configuration+Properties#ConfigurationProperties-hive.mapred.mode)=strict), the order by clause has to be followed by a "limit" clause.
* The limit clause is not necessary if you set hive.mapred.mode to nonstrict. The reason is that in order to impose total order of all results, there has to be one reducer to sort the final output.
* If the number of rows in the output is too large, the single reducer could take a very long time to finish.

In hive, sort by is similar to SQl-

* Hive uses the columns in *SORT BY* to sort the rows before feeding the rows to a reducer. The sort order will be dependent on the column types.
* If the column is of numeric type, then the sort order is also in numeric order. If the column is of string type, then the sort order will be lexicographical order.

What is the usefulness of the DISTRIBUTED BY clause in Hive?

DISTRIBUTE BY clause is used to **distribute the input rows among reducers**. It ensures that all rows for the same key columns are going to the same reducer. So, if we need to partition the data on some key column, we can use the DISTRIBUTE BY clause in the hive queries.

How does data transfer happen from HDFS to Hive?

For hive, there isn't any transfer of data you can create a schema in hive and load the data in the schema from hive and can even create a schema and point it to the data stored in hdfs.

Wherever (Different Directory) I run the hive query, it creates a new metastore\_db, please explain the reason for it?

* When running [Hive](https://www.projectpro.io/top%20hive%20commands%20with%20examples-tutorial/hive-commands) in embedded mode, it creates a local metastore.
* When you run the query, it first checks whether a metastore already exists or not.
* The property javax.jdo.option.ConnectionURL defined in the hive-site.xml has a default value jdbc: derby: database Name=metastore\_db; create=true.
* The value implies that embedded derby will be used as the Hive metastore and the location of the metastore is metastore\_db which will be created only if it does not exist already.
* The location metastore\_db is a relative location so when you run queries from different directories it gets created at all places from wherever you launch hive.

What will happen in case you have not issued the command: ‘SEThive.enforce.bucketing=true;’ before bucketing a table in Hive?

* The command: 'SET hive. enforce. bucketing=true;' allows one to have the correct number of reducer while using 'CLUSTER BY' clause for bucketing a column.
* In case it's not done, one may find the number of files that will be generated in the table directory to be not equal to the number of buckets.

Can a table be renamed in Hive?

It's same as in SQL, we can use the alter command.

Write a query to insert a new column (new\_col INT) into a hive table a position before an existing column (x\_col) .

ALTER TABLE h\_table  
  
CHANGE COLUMN new\_col INT  
  
BEFORE x\_col;

What is serde operation in HIVE?

SerDe is short for Serializer/Deserializer. Hive uses the SerDe interface for IO. The interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing.

Serialization is a process of converting an object into bytes that can be stored in a file or transmitted over the network.

Deserialization is a process converting bytes to an objects.

Explain how Hive Deserializes and serialises the data?

We know what serialization and deserialization means.

Java understands objects and hence object is a deserialized state of data. When you use the same concept, Hive understands “columns” and hence if given a “row” of data, the task of converting that data into columns is the Deserialization part of Hive SerDe. In short

A select statement creates deserialized data (columns) that is understood by Hive. An insert statement creates serialized data (files) that can be stored into an external storage like HDFS.

Write the name of the built-in serde in hive.

* **TextInputFormat/HiveIgnoreKeyTextOutputFormat**
* **SequenceFileInputFormat/SequenceFileOutputFormat**
* **MetadataTypedColumnsetSerDe**
* **LazySimpleSerDe**
* **Thrift SerDe in Hive**
* **Dynamic SerDe in Hive**

Can you write the name of a complex data type (collection data types) in Hive?

Hive also support complex data types which includes Array, Map, Struct and union. Array is used to store the list of elements. Map is used to store key/value pair. Struct is for parent and child assosiations.

Can hive queries be executed from script files? How?

It is possible by using the source command.

For example −

Hive> source /path/to/file/file\_with\_query.hql

What are the default record and field delimiter used for hive text files?

The default record delimiter is − \n and the filed delimiters are − \001,\002,\003

. How do you list all databases in Hive whose name starts with s?

**Show databases** or **Show schemas** statement is lists all the database names in Hive metastore. In that statement, the optional LIKE clause allows the list of databases to be filtered using a regular expression.

show databases like 'product\*'; will return database whose name stats with product.

show databases like '\*payment'; will return databases name ends with payment.

show databases like 'product\*|\*payment'; will return databases either with a mentioned name.

show databases like '\*sellers\*'; --- will display databases with the mentioned pattern.

What is the difference between LIKE and RLIKE operators in Hive?

**LIKE** is an operator similar to LIKE in SQL. We use LIKE to search for string with similar text.

**RLIKE** (Right-Like) is a special function in Hive where if any substring of A matches with B then it evaluates to true. It also obeys Java regular expression pattern. Users don't need to put % symbol for a simple match in RLIKE.

How will you convert the string ’51.2’ to a float value in the particular column in hive?

Hive CAST function converts from data type to another type, below is the syntax.

CAST(from\_datatype AS to\_datatype);

What will be the result when you cast ‘abc’ (string) as INT?

If you have a non-numerical value (or a non-integer) value in a column and you attempted to convert that to an Integer. Hive will be forgiving in that regard. It will simply substitute null and continue processing.

52. What does the following query do?

INSERT OVERWRITE TABLE employees

PARTITION (country, state)

SELECT ..., se.cnty, se.st

FROM staged\_employees se;

It will overwrite data in employees by selecting the data from from staged\_employees and by creating partition.

53. Write a query where you can overwrite data in a new table from the existing table.

from sales\_data\_v2 insert overwrite table sales\_data\_pq select;

54. What is the maximum size of a string data type supported by Hive? Explain how Hive supports binary formats.

* Hive supports two more primitive data types, BOOLEAN and BINARY.
* Similar to Java's Boolean, BOOLEAN in hive stores true or false values only. BINARY is an array of Bytes and similar to VARBINARY in many RDBMS.

55. What File Formats and Applications Does Hive Support?

* File Format ---
* Text files
* Sequence Files
* Avro Files
* Parquet files
* ORC files(optimized row column)
* Hive support all application written in these languages:
* Java
* Python
* Ruby
* C++
* PHP

56. How do ORC format tables help Hive to enhance its performance?

ORC is columnar file format, its seek time is less and the compression is more, thus it enhances its performance.

57. How can Hive avoid map reduce while processing the query?

You can make Hive avoid Map Reduce to return query results by setting the hive.exec.mode.local.auto property to ‘true’.

To enable Mapreduce we keep it false.

58. What is view and indexing in hive?

View is a virtual table and Indexes are a pointer or reference to a record in a table as in relational databases.

59. Can the name of a view be the same as the name of a hive table?

No, it can't be same.

60. What types of costs are associated in creating indexes on hive tables?

Basically, there is a processing cost in arranging the values of the column on which index is created since Indexes occupies.

61. Give the command to see the indexes on a table.

show indexes from yourtable;

63.If you run a select \* query in Hive, why doesn't it run MapReduce?

It’s an optimization technique. hive.fetch.task.conversion property can (FETCH task) minimize latency of mapreduce overhead. When queried SELECT, FILTER, LIMIT queries, this property skip mapreduce and using FETCH task. As a result Hive can execute query without run mapreduce task.

64. What are the uses of Hive Explode?

The explode function **explodes an array to multiple rows**. Returns a row-set with a single column (col), one row for each element from the array.19-Jan-2023

66. Can the default location of a managed table be changed in Hive?

/user/hive/warehouse directory is the default location.

Yes, you can do it by using the clause – LOCATION '<hdfs\_path>' we can change the default location of a managed table.

68. What is UDF in Hive?

**User Defined Functions and User Defined Aggregate Functions** allow you to completely customize how Hive evaluates data and manipulate queries.

69. Write a query to extract data from hdfs to hive.

70. What is TextInputFormat and SequenceFileInputFormat in hive?

TextInputFormat -This is very familiar input format in the Hadoop. The input will be given as key and value to Mapper, where key and value are generated in record reader. The record reader is just like a multiplexing.

SequenceFileInputFormat -The sequence file is the file has lot of importnace in hadoop.It has equal importance as "AVRO file" .Hadoop is the technology best suited for file with huge size,than many number of files with small size.

71. How can you prevent a large job from running for a long time in a hive?

This can be achieved by setting the MapReduce jobs to execute in strict mode set hive.

73. Can Hive process any type of data formats? Why? Explain in very detail.

No.

The data types it can process:

1. Integer Types: Tinyint, smallint, int, bigint.

2. String: Varchar and Char

3. Timestamp: "YYYY-MM-DD HH:MM:SS"

4. DATE: YYYY-MM-DD"

5. LITERALS:

* Floating Point Types
* Decimal Type
* Null value
* Complex Type
* Arrays
* Maps
* Struct

74. Whenever we run a Hive query, a new metastore\_db is created. Why?

Because it acts as central schema repository which can be also access tools like spark and pig as well.

75. Can we change the data type of a column in a hive table? Write an entire query.

Yes, we can change the data type of the column by using alter.

76. While loading data into a hive table using the LOAD DATA clause, how do you specify it is a hdfs file and not a local file?

load data local inpath 'file: ///config/workspace/department\_details.csv' into deparment\_data;

load data inpath '/tmp/depart\_data.csv' into department\_data;

77. What is the precedence order in Hive configuration?

In Hive we can use following precedence order to set the configurable properties.

* Hive SET command has the highest priority
* -hiveconf option from Hive Command Line
* hive-site.xml file
* hive-default.xml file
* hadoop-site.xml file
* hadoop-default.xml file

78. Which interface is used for accessing the Hive metastore?

WebHCat API web interface.

80. What is the difference between local and remote metastore?

Local Metastore: - Here metastore service still runs in the same JVM as Hive but it connects to a database running in a separate process either on same machine or on a remote machine.

Remote Metastore:- Metastore runs in its own separate JVM not on hive service JVM.

81. What is the purpose of archiving tables in Hive?

You can use Hadoop archiving to reduce the number of hdfs files in the Hive table partition. Hive has built in functions to convert Hive table partition into Hadoop Archive (HAR).

82. What is DBPROPERTY in Hive?

The DB properties are nothing but mentioning the details about the database created by the user. Suppose the name of the user, the type of the database and the tables it has, the date on which the database is created etc.

83. Differentiate between local mode and Map Reduce mode in Hive.

Local mode is actually a local simulation of MapReduce in Hadoop's LocalJobRunner class. MapReduce mode (also known as Hadoop mode)