1. What is the definition of Hive? What is the present version of Hive?

Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. Current version of Hive is 3.1.2.

1. Is Hive suitable to be used for OLTP systems? Why?

No Hive does not support insert and update at row level. So it is not suitable for Online Transaction Processing (OLTP) system.

1. How is HIVE different from RDBMS? Does hive support ACID transactions. If not then give the proper reason.

RDBMS: (Relational Database Management system) It is used to maintain database. It uses SQL (Structured Query Language). Tables in RDBMS are sparse. Schema is fixed in RDBMS. Normalized data is stored. It doesn’t support partitioning. No partition method is used.

Hive: It is used to maintain data warehouse. It uses Hive Query Language (HQL). Tables in Hive are dense. Schema varies in it. Normalized and de-normalized both type of data are stored. It supports automation partition. Sharding method is used for partition.

Currently, Hive supports ACID transactions on tables that store ORC file format.

ACID – Atomicity, Consistency, Isolation, Durability

ORC – Optimized Row Columnar

4. Explain the hive architecture and the different components of a Hive

architecture?

<https://www.interviewbit.com/blog/hive-architecture/#:~:text=The%20most%20important%20part%20of,compiler%20generates%20the%20Execution%20Plan>.

5. Mention what Hive query processor does? And Mention what are the

components of a Hive query processor?

<https://cwiki.apache.org/confluence/display/Hive/DeveloperGuide#DeveloperGuide-QueryProcessor>

6. What are the three different modes in which we can operate Hive?

2 modes: Local and map reduce mode

**When to use Local mode:**

* If the Hadoop installed under pseudo mode with having one data node we use Hive in this mode
* If the data size is smaller in term of limited to single local machine, we can use this mode
* Processing will be very fast on smaller data sets present in the local machine

**When to use Map reduce mode:**

* If Hadoop is having multiple data nodes and data is distributed across different node we use Hive in this mode
* It will perform on large amount of data sets and query going to execute in parallel way
* Processing of large data sets with better performance can be achieved through this mode

In Hive, we can set this property to mention which mode Hive can work? By default, it works on Map Reduce mode and for local mode you can have the following setting.

Hive to work in local mode set

SET mapred.job.tracker=local;

7. Features and Limitations of Hive.Features:

Hive supports MapReduce, Tez, and Spark computing engine.

Hive is a stable batch-processing framework built on top of the Hadoop Distributed File system and can work as a data warehouse.

Hive uses HIVE query language to query structure data which is easy to code. The 100 lines of java code we use to query a structure data can be minimized to 4 lines with HQL.

HQL is a declarative language like SQL means it is non-procedural.

The table, the structure is similar to the RDBMS. It also supports partitioning and bucketing.

Partition, Bucket, and tables are the 3 data structures that hive supports.

Apache hive supports ETL i.e. Extract Transform and Load. Before Hive python is used for ETL.

Hive supports users to access files from HDFS, Apache HBase, Amazon S3, etc.

Hive is capable to process very large datasets of Petabytes in size.

We can easily embed custom MapReduce code with Hive to process unstructured data.

JDBC/ODBC drivers are also available in Hive

Since we store Hive data on HDFS so fault tolerance is provided by Hadoop.

We can use a hive for data mining, predictive modeling, and document indexing.

Limitations:

Hive is not designed for the OLTP (Online transaction processing). We can use it for OLAP.

Hive does not support update and delete operation on tables.

Subqueries are not supported.

The latency in the apache hive query is very high.

Hive is not used for real-time data querying since it takes a while to produce a result.

8. How to create a Database in HIVE?

Hive> create database db-name; (or) create schema db-name;

To check if it has got created: show databases;

9. How to create a table in HIVE?

Create [external] table table\_name(column\_name data\_type)

Row format delimited

Fields terminated by ‘,’

Lines terminated by ‘\n’

Stored as texfile;

10.What do you mean by describe and describe extended and describe

formatted with respect to database and table?

Describe shows the list of columns (including partition columns) for the given table. Extended, formatted keyword is optional. Extended shows all the metadata for the specified table. Formatted shows the metadata in tabular format.

11.How to skip header rows from a table in Hive?By adding below command at the end of table creation,

tblproperties(“skip.header.line.count”=”1”);

12.What is a hive operator? What are the different types of hive operators?

* Hive provides Built-in operators for Data operations to be implemented on the tables present inside Hive warehouse.
* Types of Built-in Operators in HiveQL are:
* Relational Operators
* Arithmetic Operators
* Logical Operators
* Operators on Complex types
* Complex type Constructors

<https://www.guru99.com/hive-query-language-built-operators-functions.html#:~:text=Hive%20provides%20Built%2Din%20operators,Relational%20Operators>