Top Hive Interview Questions and Answers [Updated]

Preparing for the interview of a job that utilizes Hive software? This article covers a list of the most important and commonly asked Apache Hive interview questions and answers which will help you get an in-depth understanding of the subject to ace your next Hadoop interview or any other job interview related to Hive.

This blog will help you prepare clear and effective responses for each question that may be asked in your upcoming interview so that you can demonstrate your experience with confidence.

Apache Hive – Brief Introduction

Apache Hive is a data warehouse infrastructure tool built to process structured data in Hadoop. It accelerates data summarization, analyzing datasets, and ad-hoc queries. Apache Hive provides an easy way to structure an abundance of unstructured data and executes SQL-like queries with the given data. It can easily merge with traditional data center technologies with the help of the JDBC/ODBC interface.

In Hive, data is separated by the bucketing process. It is designed for managing and querying the structured data that is stored in the table. Hive framework has features like UDFs, and it can increase the performance of the database through effective query optimization.

The SQL-inspired queries of Hive diminish the complexity of map-reduce programming and also decrease the familiar concept of a relational database such as a row, column, schema, and table for making the learning easy.

Hive can use the directory for the partitioning of data because Hadoop’s programming works on flat flies. It improves the performance of the database queries.

Also Read>> Top SQL Interview Questions & Answers

Top Hive Interview Questions and Answers

Here are some of the most common Apache Hive interview questions that you can expect:

Q1. Explain the difference between Apache Pig and Hive.

Ans. Following are the differences:

Apache Pig	**Apache Hive**
1. It is a procedural data flow programming language 2. Apache Pig supports the Avro file format 3. This is used by researchers and programmers 4. Apache Pig works on the client-side of a cluster 5. It does not consist of a fixed metadata database	1. It is a declarative SQL Language 2. This does not support the Avro File format 3. Apache Hive is used by the data analysts 4. It is used for creating reports 5. Hive uses an exact variation of SQL DDL Language

Q2. What is the Hive variable? How do we set a variable?

Ans. Hive variables are similar to the variables in other programming languages. They are developed in the Hive environment that is assigned by the scripting language.

By using the keyword set

```
set foo=bar;
```

```
set system:foo=bar
```

Similarly, you can set the variable in command line for hiveconf namespace:

```
beeline –hiveconf foo=bar
```

Q3. What are the different modes of Hive?

Ans. There are three modes:

- Embedded Metastore
- Local Metastore
- Remote Metastore

Q4. Explain the difference between HBase and Hive.

Ans. Following are the differences between HBase and Hive:

HBase	**Hive**
1. It does not allow execution of SQL Query 2. HBase is a NoSQL database 3. This runs on top of HDFS 4. HBase is free from the schema model 5. It follows real-time processing	1. It allows execution of most SQL queries 2. Hive is a data warehouse Framework 3. It runs on top of Hadoop MapReduce 4. Hive has the schema model 5. It does not follow real-time processing

Q5. What is the use of partition in Hive?

Ans. Hive partitions are defined as the division of similar type of data on the basis of partition keys, Hive design tables into partitions.

Partition is only helpful when it has partition keys. These are the basic keys that determine the data storage in the table. It is the subdirectory in the table directory.

Q6. What are the data types supported by Hive?

Ans. Type of data types:

1. *Primitive data type*

- Numeric data type
- String data type
- Date/Time data type
- Miscellaneous data type

1. Complex data types

- Arrays
- Maps
- Structs

- Union

Q7. What is the precedence order in Hive configuration?

Ans. There is a precedence hierarchy for setting properties:

- The Hive SET command
- The command line -hiveconf option
- hive-site.xml
- hive-default.xml
- hadoop-site.xml (or, equivalently, hdfs-site.xml, core-site.xml, and mapred-site.xml)
- hadoop-default.xml (or, equivalently, hdfs-default.xml, core-default.xml, and mapred-default.xml)

Q8. Is it possible to change the default location of a managed table?

Ans. Yes, it is possible to change the default location of a managed table. We can change the location by using – LOCATION '<hdfs_path>'.

Q9. Explain the mechanism for connecting applications when we run Hive as a server.

Ans. The mechanism is done by following the below steps:

- **Thrift client:** By using thrift client, we can call Hive commands from different programming languages such as Java, Python, C++, Ruby
- **JDBC driver:** It enables accessing data and supports Type 4 JDBC driver
- **ODBC driver:** ODBC API Standards apply for the Hive DBMS. It supports ODBC protocols.

Q10. How to remove header rows from a table?

Ans. By using the TBLPROPERTIES clause, we can remove N number of rows from the top or bottom from a text file without using the Hive file. TBLPROPERTIES clause can provide multiple features that we can set as per our needs. It can be used when files are generated with additional header or footer records.

Following are the header records in a table:

System=…

```
Version=…

Sub-version=…
```

To skip the header lines in a Hive file, we can use the following table property:

```
CREATE EXTERNAL TABLE employee (

name STRING,

job STRING,

dob STRING,
```

Also Read>> Top Hadoop Interview Questions and Answers

Q11. Explain the need for buckets in Apache Hive.

Ans. The concept of bucketing provides a way of differentiating Hive table data into various files or directories. It provides effective results only if –

- There are a limited number of partitions
- Partitions are of almost similar sizes

To solve the problem of partitioning, Hive provides a bucketing concept. It is an effective way to decompose tables into manageable parts.

Q12. How can you recursively access subdirectories?

Ans. We can access subdirectories recursively by using the following command:

```
hive> Set mapred.input.dir.recursive=true;

hive> Set hive.mapred.supports.subdirectories=true;
```

The Hive tables can be directed to the higher level directory and it suitable for the directory structure:

/data/country/state/city/

Q13. How Hive distributes rows into buckets?

Ans. Rows can be divided into buckets by using:

hash_function (bucketing_column) modulo (num_of_buckets)

Here, Hive lead the bucket number in the table

Function used for column data type:

hash_function

Function used for integer data type:

hash_function (int_type_column)= value of int_type_column

Q14. What are the commonly used Hive services?

Ans. Following are the commonly used Hive services:

- Command Line Interface (cli)
- Printing the contents of an RC file with the use of rcfilecat tool
- HiveServer (hiveserver)
- Hive Web Interface (hwi)
- Metastore
- Jar

Q15. Is it possible to change the settings within the Hive session?

Ans. Yes, it is possible to change the settings in the Hive session by using the SET command. It helps with the change in Hive job settings for an exact query.

Following command shows the occupied buckets in the table:

```
hive> SET hive.enforce.bucketing=true;
```

By using SET command, we can see the present value of any property

```
hive> SET hive.enforce.bucketing;
```

```
hive.enforce.bucketing=true
```

We cannot target the defaults of Hadoop with the above command, so we can use –

```
 SET -v
```

Q16. Mention the components used in the Hive query processor.

Ans. Following are the components:

- Parse and Semantic Analysis (ql/parse)
- Map/Reduce Execution Engine (ql/exec)
- Optimizer (ql/optimizer)
- Sessions (ql/session)
- Hive Function Framework (ql/udf)
- Plan Components (ql/plan)
- Type Interfaces (ql/typeinfo)
- Metadata Layer (ql/metadata)
- Tools (ql/tools)

Q17. What are the Trim and reverse functions?

Ans. The trim function removes the spaces related to the strings.

Example:

```
TRIM(‘ NAUKRI ‘);
```

Output:

```
NAUKRI
```

To remove the leading space:

LTRIM(‘NAUKRI’);

To remove the trailing space:

RTRIM(‘NAUKRI ‘);

The reverse function will reverse the characters into strings.

Example:

REVERSE(‘NAUKRI’);

Output:

IRKUAN

Q18. Explain the default metastore?

Ans. It provides an embedded Derby database instance that can only be supported by one user where it can store metadata. If you run your Hive query by using the default derby database. Then, you will get a default subdirectory in your current directory with the name metastore_db . It will also create the metastore if it does not already exist. Here, the property of interest is javax.jdo.option.ConnectionURL.

And the default value is jdbc:derby:;databaseName=metastore_db;create=true. This value identifies that you are using embedded derby as your Hive metastore, and its location is metastore_db.

Also Read>> 7 Trending Tech Skills to Master in 2020

Q19. Is multiline comment supported?

Ans. No, Hive can only support single-line comments.

Q20. What is the possible way to improve the performance with ORC format tables?

Ans. We can improve the performance by using the ORC file format by storing data in a highly efficient manner. The performance can also be improved by using ORC files while writing, reading, and processing data.

```
Set hive.compute.query.using.stats-true;

Set hive.stats.dbclass-fs;

CREATE TABLE orc_table (

idint,

name string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\:’

LINES TERMINATED BY ‘\n’

STORES AS ORC;
```

Q21. Why the Hive does not store metadata information in HDFS?

Ans. The Hive does not store the metadata information in HDFS and stores it in RDBMS instead. This is done to attain low latency because HDFS read/write operations are time-consuming. HDFS is not meant for regular updates, thus, RDBMS is used as it provides low latency and random updates.

Q22. What is the difference between local and remote metastore?

Ans. The differences between local and remote metastore are:

Local Meta-store: In this, the meta-store service runs in the same JVM in which the Hive service runs. It associates with a database running in a different JVM, either on a similar machine or a remote machine.

Remote Meta-store: In this, the meta-store service runs alone separating JVM and not in the Hive benefit JVM. Thus, different procedures communicate with the metastore server utilizing Thrift Network APIs. Having at least one meta-store server for this situation will provide greater accessibility.

Q23. When would you use SORT BY instead of ORDER BY?

Ans. SORT BY clause should be used instead of ORDER BY when one has to sort huge datasets. SORT BY sorts the data using multiple reducers while ORDER BY sorts all data together using a single reducer. Thus, if ORDER BY is used against a large number of inputs, then the execution will be time-consuming.

Q24. Explain dynamic partitioning?

Ans. In Hive Dynamic Partitioning, the data is inserted into the respective partition dynamically without you having explicitly creating the partitions in the table. The values for partition columns are known in the runtime. Typically, the data is loaded from the non-partitioned table in the dynamic partition loads and takes more time in loading data compared to static partition.

Q25. What is indexing? Explain its use.

Ans. Hive index is a Hive query optimization technique to access a column or set of columns in a Hive database. We use Hive indexing as it improves the speed of query lookup on certain columns of a table as the database system does not need to read all rows in the table.

Also Read>> Trending Tech Skills in 2020: Cloud, Game development and DevOps

Q26. Explain the use of Hcatalog?

Ans. HCatalog is a table storage management layer for Hadoop that is used to share data structures with external systems. It provides access to the Hive metastore to users of other Hadoop or data processing tools, such as Pig and MapReduce so that they can easily read and write data on Hive's data warehouse.

Q27. Explain the components of a Hive architecture.

Ans. The different components of Hive architecture are:

- User Interface: It provides an interface between user and hive. User Interface allows users to submit queries to the system. It creates a session handle to the query and sends it to the compiler to generate an execution plan for it. It supports Hive web UI and Hive command line.
- Compiler: It generates the execution plan.
- Execute Engine: It manages the dependencies for submitting each of these stages to the relevant component.
- Metastore: It sends the metadata to the compiler for the execution of the query on receiving the send metadata request.

Q28. What is ObjectInspector functionality?

Ans. The ObjectInspector helps in analyzing the structure of individual columns and the internal structure of the row objects in the Hive. It offers access to complex objects that can be stored in multiple formats in the memory. The ObjectInspector describes the structure of the object as well as the ways to access the internal fields inside the object.

Q29. What are the different types of join in Hive?

Ans. The different types of join in Hive are:

- Join: It gives the cross product of both the table's data as output. It is similar to the Outer Join in SQL.
- Full Outer Join: It fulfills the join condition and gives the cross product of both the left and right outer tables without match condition.
- Left Outer Join: It returns all the rows from the left table even if there are no matches in the right table.
- Right Outer Join: It returns all the rows from the right table even if there are no matches in the left table.

Q30. Which classes are used to read and write HDFS files in Hive?

Ans. The following classes are used:

- TextInputFormat: Reads data in plain text file format.
- HiveIgnoreKeyTextOutputFormat: Writes data in plain text file format.
- SequenceFileInputFormat: Reads data in hadoop SequenceFile format.
- SequenceFileOutputFormat: Writes data in Hadoop SequenceFile format.

Q31. What is HiveServer2 (HS2)?

Ans. The HiveServer2 (HS2) is a server interface that allows remote clients to execute queries against the Hive. HS2 retrieves the results of the mentioned queries. It supports multi-client concurrency and authentication and aims to provide better support for open API clients like JDBC and ODBC.

Q32. What Hive is composed of?

Ans. The Hive consists of 3 components:

- Clients
- Services
- Storage and Computing

Q33. What are the different types of tables in Hive?

Ans. There are two types of tables available in Hive:

- Managed Table: Both the data and schema are under the control of the Hive.
- External Table: Only the schema is under the control of the Hive.

Q34. What do you mean by describe and describe extended.

Ans. The Describe command/query displays the name of the database, the root location on the file system and comments.

The Describe extended command/query provides the details of the database or schema in a detailed manner. It displays the detailed information of the table such as the list of columns, the data type of the columns, table type, table size, and more.

Q35. How can you optimize Hive performance?

Ans. Hive performance can be optimized to run queries faster in the following ways:

- Using vectorization
- Using ORCFILE
- Cost-based query optimization.
- Enable Compression
- Enable Tez Execution Engine

- Optimize LIMIT operator
- Using Parallel Execution
- Enable Mapreduce Strict Mode

Visit Naukri Learning website for more