Hive Interview Q & A

# Generic Hive questions

1. Where do you store the changes to reflect permanently while loading hive?

In $HOME/.hiverc file

1. Which is the property to be set for JDBC connetion?

javax.jdo.option.ConnectionURL

1. Where to place the JDBC driver for Hive?

$HIVE\_HOME/lib or hadoop lib directory

1. Where does hive store the commands history?

In $HOME/.hivehistory

1. How to run the shell cmds from hive cli?

Use !

Shell pipes and globs and masking (\* for file name matching) do not work here

1. How is dfs in hive faster than hadoop dfs from shell?

Dfs from hive runs on the same JVM whereas hadoop dfs creates a new JVM on each load.

1. Which hive property will display the headers in select query results in cli?

hive.cli.print.header

1. Which property stores information about the database directory?

hive.metastore.warehouse.dir

1. What are the URI scheme for hadoop, MapR and amazon S3?

Hdfs, maprfs and hdfs/s3n(preferable)/s3(older)

1. Which command will let you check the current database?

There is no such command, you can use USE command for safer side.

But you can set the property hive.cli.print.current.db to true to always show current database. Similarly, there is no way to unset or delete a DB property.

1. Difference between RESTRICT and CASCADE in drop database command?

CASCADE will drop the underlying tables first and then drop the database, but RESTRICT does not, it fails and forces you to drop the tables before dropping the database.

1. What will happen when the underlying schema of the table does not match in hive while you try to create a table with IF NOT EXISTS clause?

Hive will ignore the discrepancy.

1. Which are the 2 table properties Hive automatically adds?

last\_modified\_by holds the username of the last user to modify the table, and last\_modified\_time holds the epoch time in seconds of that modification

1. How can you search for a table in a database without changing to that database?

By using : SHOW TABLES IN mydb;

However, the regexp will not work here

1. Difference between extended and formatted

Using formatted shows detailed formatted output and a more verbose one than extended.

1. Difference between Hive managed and external tables.

Managed tables are more or less handled by Hive, i.e. dropping a table deletes all its data, but dropping an external table only deletes the metadata not the data in the table as it is stored externally outside the control of Hive.

1. How can you copy a schema of an existing table in Hive?

By using CREATE TABLE <table\_name> like <another\_table\_whose\_schema\_you\_need>

1. How can you list tables in another database?

SHOW tables in <some\_db>

1. What is meant by partition in Hive, and why do we do it?

Partition is a sub-directory in the table directory in Hive. Hive organizes tables into partitions for similar data types based on a column or partition key. Partitioning helps reduce query latency, as it scans only the relevant partitioned data.

1. Default location of hive tables : /user/hive/warehouse
2. How to force hive to use the predicate clause on partitioned tables?

By setting hive.mapred.mode property to true

1. List partitions on a table;

SHOW PARTITIONS <table\_name>

For specific partitions,

SHOW PARITIONS <table\_name> PARTITION(<partition\_column>=<partition\_value>)

1. For external paritions, how do you define the location of each partition?

Partitions on external tables are created separately and location clause is given with the alter table clause to add partition at a specific location

1. Does archiving help in space reduction?

No, by archiving tables in hive, we reduce the number of files in the file system and reduce the load on name node but no space savings eg as through compression.

1. Hive architecture components:

UI, Driver, compiler, metastore, execution engine.

1. Properites to be set for ACID transactions in HIVE.

Hive.txn.manager = org.apache.hadoop.ql.lockmgr.DbTxnManager

Hive.support.concurrency = true

Set table’s TBLPROPERTIES(‘transactional’=’true’)

1. To have bucketing implemented dynamically which property needs to be updated

Hive.enforce.bucketing=true;

1. How to let hive avoid using map reduce?

You can set the property : hive.exec.mode.local.auto = true.

1. Which are the dynamic partition parameters?

hive.exec.dynamic.partition

hive.exec.dynamic.partition.mode

hive.exec.max.dynamic.partitions.pernode

hive.exec.max.dynamic.partitions

hive.exec.max.created.files

1. Limitation of CTAS format for populating a table

It cannot be used with external tables, because for external tables we need to use alter table to point the partition to correct location.

1. How do you export data from table to local/any directory location?

Use INSERT OVERWRITE LOCAL DIRECTORY '/tmp/ca\_employees'

SELECT name, salary, address

FROM employees

WHERE se.state = 'CA';

The number of files created will be determined by the number of reducers configured.

For populating multiple directories:

FROM staged\_employees se

INSERT OVERWRITE DIRECTORY '/tmp/or\_employees'

SELECT \* WHERE se.cty = 'US' and se.st = 'OR'

INSERT OVERWRITE DIRECTORY '/tmp/ca\_employees'

SELECT \* WHERE se.cty = 'US' and se.st = 'CA'

INSERT OVERWRITE DIRECTORY '/tmp/il\_employees'

SELECT \* WHERE se.cty = 'US' and se.st = 'IL';

1. What is table sampling in hive?

We sample the data based on buckets or percent to retrieve only a small portion or full data.

Like:

1M: SELECT \* FROM source TABLESAMPLE(100M) s;

20 rows: SELECT \* FROM source TABLESAMPLE(20 ROWS) s;

Bucket 1 out of 3 on location: SELECT \* FROM source TABLESAMPLE(BUCKET 1 out of 3 on location) s;

2 percent: SELECT \* FROM source TABLESAMPLE(2 percent) s;

1. What is the use of nodrop?

Enable this facility on a table or partition to prevent accidental dropping.

Alter table emp\_table enable nodrop;

1. How can you prevent a table/partition from being queried?

By using offline property

ALTER TABLE emp\_table (PARTITION partition\_spec) ENABLE OFFLINE

1. Types of partitioning in hive

Static and dynamic

In static partitioning, the partitioned column cannot be a part of table columns

To enable dynamic partition:

Set hive.exec.dynamic.partition : true

Set hive.exec.dynamic.partition.mode : nonstrict

This will create partitions automatically from the table based on the column given. The column on which partition is to be done should be mentioned as the last column in the select within insert statement.

1. Perils of over partitioning

HDFS was designed for many millions of large files not billions of small files. Each partition is stored as a separate directory and the directory metadata is stored in the name node which is ultimately stored in the memory. Similarly for mapreduce each task creates a new JVM, so for small files there will be an overhead of JVM start up and tear down.

1. How do you use variables in hive query?

Set variable with hive :

hive -hiveconf df=2011-01-01

Now dt variable can be used in hive query as below:

SELECT distinct(ip) from weblogs where hitdate=’$(hiveconf:dt)’;

# Hive QL

1. How to refer to array, map and struct elements in hive.

For array: use 0-based indexing, like array\_name[0]

For map: use name based indexing, like map\_name[“map-key-name”]

For struct: use dot notation with key, like struct.sruct\_key

1. How do you use regular expressions in column names display:

SELECT symbol, `price.\*` FROM stocks;

Displays symbol column and all columns starting with price.

1. How do you improve performance on aggregations in hive?

By setting below property to true:

hive.map.aggr

This does top level aggregation in the map phase, which would be aggregation performed after group by if this value is set to false. This setting requires more memory though.

1. Difference between parse\_url and parse\_url\_tuple

Parse\_url returns only the specified part whereas tuple returns tuple of parameters passed as arguments to the command

1. What is the function str\_to\_map used for?

Str\_to\_map(s,delim1,delim2) is used to convert string s into a map where delim1 is the delimiter between key-value pairs and delim2 is the key value separator.

1. What is a map side join?

If all tables are small but 1 is large, the largest table can be streamed through the mappers while the small tables are cached in memory. Hive can do all the joining map-side, since it

can look up every possible match against the small tables in memory, thereby eliminating the reduce step required in the more common join scenarios.

1. How does bucketed table optimization work?

The data must be bucketed on the keys used in the ON clause and the number of buckets for one table must be a multiple of the number of buckets for the other table. When these conditions are met, Hive can join individual buckets between tables in the map phase, because it does not need to fetch the entire contents of one table to match against each bucket in the other table.

1. Which properties are to be set for turning on the bucket table optimization?

set hive.optimize.bucketmapjoin=true;

set hive.input.format=org.apache.hadoop.hive.ql.io.BucketizedHiveInputFormat;

set hive.optimize.bucketmapjoin.sortedmerge=true;

1. Difference between order by and sort by

Order by does total ordering while sort by does local reducing in each reducer. Both have their limitations, order by takes time to sort and sort by does not give completely sorted results if there are more than 1 reducers.

1. What is the significance of distribute by clause?

When you want to use multiple reducers and the overall sorted order does not matter, you can pass the columns on which you want grouping to be done so that the reducers receive rows grouped by a certain specific column.

1. What is another way of sorting with distribute clause?

Using cluster by clause, this will help achieve overall ordering of the file.

# Hive Views

# Hive Indexes

# Schema design gotchas

1. Table-by-Day anti pattern of database world can be replaced with Hive’s partitions concept
2. Over partitioning should be avoided because it creates unnecessary number of files and folders. If there are multiple small files, when map reduce creates tasks that require a JVM start-up and tear down each time, which would be more time consuming then original processing time
3. Making multiple passes over same data can be an overhead, hive can achieve it in one steo as below;

FROM history

INSERT OVERWRITE sales SELECT \* WHERE action='purchased'

INSERT OVERWRITE credits SELECT \* WHERE action='returned';

Instead of:

hive> INSERT OVERWRITE TABLE sales

SELECT \* FROM history WHERE action='purchased';

hive> INSERT OVERWRITE TABLE credits

SELECT \* FROM history WHERE action='returned';

1. When you create folders for partitions in hdfs, you need to repair your table in hive as below:

Msck repair table partitioned\_table

1. What is bucketing?

Data organizing technique in hive.

All the same column values of a bucketed column go into same bucket

Can be used alone or alongwith partitioning

Hashing is used for bucketing, so it is good to use an integer column for this. All records with column satisfying the hash values fall in same bucket.

Number of buckets are given by clustered by statement and all buckets are folders.

For bucketed map joins, both joining tables must be bucketed on same column(joining column) and have equal number of buckets.

1. The Case for Partitioning Every Table ------------

# Tuning

# File formats and compressions

# Developing

# Functions

# Streaming

# Customizing hive file formats and record formats

# Hive thrift server

# Storage handlers and NoSql

# Security

# Locking

# Integration with Oozie

# Hive and AWS