1. What is Impala and how is it different from hive?

Impala is an sql query engine which hive is a data warehousing software built on top of hadoop.

Impala does not use map reduce which makes it faster than hive.

Hive generates query expressions at compile time whereas Impala does runtime code generation for “big loops”

Hive supports complex types but Impala does not.

1. How can you list all tables in a database without switching to it?

Show tables in db\_name

1. How can you list all tables in a database that start with customer?

Show tables in db like ‘customer\*’

1. How can you move a table from one database to another?

Alter table table1 rename to exp.table1;

1. How can you run an sql file containing sql statements from impala shell?

impala-shell -i localhost -f setup.sql

1. How can you directly execute a query in impala shell?

impala-shell -i impala-host -q 'select count(\*) from customer\_address'

1. Insert query in impala

INSERT OVERWRITE TABLE tab3 SELECT id, col\_1, col\_2, MONTH(col\_3), DAYOFMONTH(col\_3) FROM tab1 WHERE YEAR(col\_3) = 2012;

1. Join in impala

SELECT tab1.col\_1, MAX(tab2.col\_2), MIN(tab2.col\_2) FROM tab2 JOIN tab1 USING (id) GROUP BY col\_1 ORDER BY 1 LIMIT 5;

1. Create a partitioned table in impala

use external\_partitions;

alter table logs rename to logs\_original;

create external table logs (field1 string, field2 string, field3 string) partitioned by (year string, month string, day string, host string) row format delimited fields terminated by ',' location '/user/impala/data/logs';

1. How can you fetch data from hdfs in an impala table directly if there are any modifications in the files?

refresh tablename;

1. What needs to be done in impala shell if there is any update/change in hive tables or directories?

Refresh or INVALIDATE\_METADATA

1. While copying files to HDFS, what option can you use to make sure that each file resides in single HDFS block? Why is this important?

sudo -u hdfs hdfs dfs -Ddfs.block.size=253m -put \*.parq /user/impala/ staging/airlines

For optimal query performance, it is better if the entire file is in single HDFS block

1. How can you tell impala to identify columns and metadata from files?

Using LIKE PARQUET 'path\_to\_any\_parquet\_file' clause means we skip the list of column names and types; Impala automatically gets the column names and data types straight from the data files. (Currently, this technique only works for Parquet files.)

$ impala-shell

> CREATE DATABASE airlines\_data;

> USE airlines\_data;

> CREATE EXTERNAL TABLE airlines\_external LIKE PARQUET 'hdfs:staging/airlines/4345e5eef217aa1bc8f16177f35fd983\_1150363067\_data.0.parq' STORED AS PARQUET

LOCATION 'hdfs:staging/airlines';

WARNINGS: Impala does not have READ\_WRITE access to path 'hdfs:// myhost.com:8020/user/impala/staging'

1. What is the significance of external table in impala?

The clause external table in impala means that dropping the table in impala will not delete the HDFS files.

1. Difference between NDV and Count distinct?

NDV is the function that returns number of distinct values in a column, it is an estimate when there are lots of different values in the column, but is precise when the cardinality is less than 16 K, so you can use multiple NDVs in a statement but only 1 count distinct.

SElECT NDV(carrier), NDV(flight\_num), NDV(tail\_num), NDV(origin), NDV(dest) FROM airlines\_external;

1. How to collect statistics in impala(tablename=airlines)?

COMPUTE INCREMENTAL STATS airlines;

1. Which are the startup options for impalad daemons to specify timeout?

--idle\_query\_timeout: Specifies the time in seconds after which an idle query is cancelled. A zero value for --idle\_query\_timeout disables query timeouts.

--idle\_session\_timeout Specifies the time in seconds after which an idle session expires. The default value of 0 specifies sessions never expire.

1. Which option can you use to configure the number of Thrift connection retries?

Impala connections to the backend client are subject to failure in cases when the network is momentarily overloaded. To avoid failed queries due to transient network problems, you can configure the number of Thrift connection retries using the following option:

The --backend\_client\_connection\_num\_retries option specifies the number of times Impala will try connecting to the backend client after the first connection attempt fails. By default, impalad will attempt three re-connections before it returns a failure.

The --backend\_client\_rpc\_timeout\_ms option can be used to specify the number of milliseconds Impala should wait for a response from the backend client before it terminates the connection and signals a failure. The default value for this property is 300000 milliseconds, or 5 minutes.

1. What are scratch directories in impala?

When there are large joins impala stores the intermediate files and removes them after operation finishes.

You can specify locations of the intermediate files by starting the impalad daemon with the ##scratch\_dirs="path\_to\_directory" configuration option. By default, intermediate files are stored in the directory /tmp/impala-scratch.

