**Problem Statement**

Explain in breif with an example

● Bucketing

● Bucketing V/S Partitioning

● Sampling

# Bucketing

1. Hive bucketing is responsible for dividing the data into number of equal parts.

2. If we want to use bucketing in hive then you should use CLUSTERED BY (Col)

command while creating a table in Hive.

3. We can perform Hive bucketing concept on Hive Managed tables or External tables.

4. We can perform Hive bucketing optimization only on one column only not more than one.

5. The value of this column will be hashed by a user-defined number into buckets.

6. Bucketing works well when the field has high cardinality and data is evenly distributed among buckets.

7. If we want to perform queries on Date or Timestamp or other columns which are having high records fields at that time Hive bucketing concept is perfectible.

8. We can assign number of number buckets while creating the table.

9. Bucketing also very useful in doing efficient map-side joins etc.

10. Bucketing on the other hand, will result with a fixed number of files, since we specify the number of buckets.

Bucketing example:

CREATE TABLE mytable (

name string,

city string,

employee\_id int )

CLUSTERED BY (employee\_id) INTO 256 BUCKETS

Here we are creating buckets based on the employee.

# Bucketing V/S Partitioning

a) Bucketing:

-Fixed number of buckets

-Almost of the same size

-Stored in files

-Based on hash value of the column

-Used for optimizing the lookup, joins, sampling

b) Partitioning

-Unknown number of partitions

-Sizes are unknown

-Stored in directories.

-Based on actual value of the column

-Optimizes retrieve or scan

Partitioning example:

CREATE TABLE example (

name string,

city string,

employee\_id int )

PARTITIONED BY (year STRING, month STRING, day STRING)

CLUSTERED BY (employee\_id) INTO 256 BUCKETS

/user/hive/warehouse/example/y=2015/m=12/d=02

Here we can see that directories are created for every combination of the day, moth and year.

# Sampling

-Sampling is concerned with the selection of a subset of data from a large dataset to run queries and verify results. The dataset may be too large to run queries on the whole data.

-There are various ways of sampling the data

1. Bucket Sampling

We can use TABLESAMPLE clause to bucket the table on the given column and get data from only some of the buckets.

TABLESAMPLE (BUCKET x OUT OF y [ON colname])

colname indicates the column to be used to bucket the data into y buckets[1-y]. All the rows which are in the bucket x are returned.

If the table is not bucketed on the column(s) used in sampling, TABLESAMPLE will scan the entire table and fetch the sample.

If the hive table is bucketed on some column(s), then we can directly use that column(s) to get a sample. In this case Hive need not read all the data to generate sample as the data is already organized into different buckets using the column(s) used in the sampling query. Hive will read data only from some buckets as per the size specified in the sampling query.

b)Block Sampling

Block sampling allows Hive to select at least n% data from the whole dataset. Sampling granularity is at the HDFS block size level. If HDFS block size is 64MB and n% of input size is only 10MB, then 64MB of data is fetched.

1. Limit

Limit will limit the records to top n records. This traverses the whole data.

Hence is one of the insufficient way of sampling.

1. Row Sampling

Row sampling samples entire dataset and similar to that the limit but the rows are chosen from entire dataset and not the top samples. It is also insufficient as whole of the dataset is scanned.