Explain in brief with an example

● Bucketing

Bucketing is a concept similar to partitioning here the only difference is that the number of partitions to be divide is decided by the user only. Hence the problem of high cardinality is looked after automatically. This is also one reason why bucketing is used for high cardinality datasets.

The keyword used for bucketing is CLUSTER BY. Bucketing uses the hash code for deciding the which entry goes to which bucket.

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

|  |  |
| --- | --- |
| Bucketing | Partitioning |
| Fixed number of buckets | Unknown number of partitions |
| Based on hash value of the column | Based on actual value of the column. |
| Stored in files | Stored in directories. |
| Almost of the same size | Sizes are unknown |
| Used for optimizing the lookup, joins, sampling | Optimizes retrieve or scan. |

Partitioning example.

CREATE TABLE mytable (

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/mytable/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 in hive results in displaying the small part of the data. There are various ways of sampling the data

1. Limit
2. Block Sampling
3. Row sampling
4. Bucket sampling

Limit will limit the records to top n records. This traverses the whole data hence is one of the insufficient way of sampling

Block sampling is done using the tablesample but it gives part of the rows which is mentioned in the percentage. That is only mentioned percentage that is in table.

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.

Bucket sampling- here the total dataset is not scanned but only the selected bucket is scanned and the records from that bucket is given hence this sampling is the efficient way of sampling of the data.