**ASSIGNMENT-26.2:**

**QUESTION** : Explain in brief the following :

1. **BUCKETING**:

Bucketing is a method to evenly distributed the data across many files. Create multiple buckets and then place each record into one of the buckets based on some logic mostly some hashing algorithm.

Bucketing feature of Hive can be used to distribute/organize the table/partition data into multiple files such that similar records are present in the same file. While creating a Hive table, a user needs to give the columns to be used for bucketing and the number of buckets to store the data into. Which records go to which bucket are decided by the Hash value of columns used for bucketing.

[Hash(column(s))] MOD [Number of buckets]

Hash value for different columns types is calculated differently. For int columns, the hash value is equal to the value of int. For String columns, the hash value is calculated using some computation on each character present in the String.

Data for each bucket is stored in a separate HDFS file under the table directory on HDFS. Inside each bucket, we can define the arrangement of data by providing the SORT BY column while creating the table.For partitioned table, we will have n buckets created per partition.

Bucketing concept also provides the flexibility to keep the records in each bucket to be sorted by one or more columns. This makes map-side joins even more efficient, since the join of each bucket becomes an efficient merge-sort.

1. **BUCKETING V/S PARTITIONING**

**PARTITIONING :-**

1. Partitioning is used to divide the table into different partitions. Each partition is stored as a different directory.
2. A partition is created for each unique value of the partition column.
3. Hierarchical partitioning can be done by specifying the partitioning columns in a sequence as per the hierarchy like Country, State, City.
4. We cannot control the number of partitions if the value of partitioning columns have a very high cardinality.
5. Partitioning allows hive to avoid full table scan if partition columns are used in the where clause of hive query. A query containing partition columns in the where clause will scan directories for specific partition only.

**BUCKETING :-**

1. Bucketing is used to distribute/organize the data into fixed number of buckets.
2. Each bucket is stored as a file under the Table/Partition directory.
3. The number of buckets are fixed at the table creation time. All the data will be distributed into these buckets based on the hash value of the bucketing columns.
4. Which records go to which bucket are decided by the Hash value of columns used for bucketing.
5. A Bucket will have all the records for same value of bucketing columns.
6. A Bucket will have all the records for same Hash value of bucketing columns. So records having different value of bucketing columns but having same hash value will go into the same bucket.
7. Bucketing is used for efficient map-side joins between bucketed tables and for effectively executing sampling queries.
8. **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. Therefore in development and testing phases it is a good idea to run queries on a sample of dataset.

## TABLESAMPLE Clause

We can run Hive queries on a sample of data using the TABLESAMPLE clause. Any column can be used for sampling the data. We need to provide the required sample size in the queries.

**Sampling by Bucketing**

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.

**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.