Hive Partitioning

• Hive organizes tables horizontally into partitions.

• It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, department etc.

• Using partition, it is easy to query a portion of the data.

• Partitioning can be done based on more than column which will impose multi-dimensional structure on directory storage.

• In Hive, partitioning is supported for both managed and external tables.

• The partition statement lets Hive alter the way it manages the underlying structures of the table’s data directory.

• In case of partitioned tables, subdirectories are created under the table’s data directory for each unique value of a partition column.

• When a partitioned table is queried with one or both partition columns in criteria or in the WHERE clause, what Hive effectively does is partition elimination by scanning only those data directories that are needed.

• If no partitioned columns are used, then all the directories are scanned (full table scan) and partitioning will not have any effect.

Static vs dynamic partitioning

When to use static partitioning

• Static partitioning needs to be applied when we know data (supposed to be inserted) belongs to which partition.

When to use dynamic partitioning

• In static partitioning, every partitioning needs to be backed with individual hive statement which is not feasible for large number of partitions as it will require writing of lot of hive statements.

• In that scenario dynamic partitioning is suggested as we can create as many number of partitions with single hive statement.