**SECTION 2 : BUCKETING**

**set hive.enforce.bucketing = true;**

-----------------------------------------------------------------

**CREATE TABLE Migrants\_Buckets (**

**Web\_ID INT,**

**Incident\_Region STRING,**

**Reported\_Date STRING,**

**Reported\_Year INT,**

**Reported\_Month VARCHAR(5),**

**No\_Dead INT,**

**Estd\_No\_Missing INT,**

**Tot\_Dead\_Missing INT,**

**No\_Survivors INT,**

**No\_Females INT,**

**No\_Males INT,**

**No\_Children INT,**

**Death\_Cause String,**

**Loc\_Desc String,**

**Info\_Src String**

**)clustered by (Incident\_Region) into 7 buckets;**

-----------------------------------------------------------------

**from migrants**

**INSERT INTO migrants\_buckets**

**SELECT \*** -----------------------------------------------------------------

**select \* from migrants\_buckets tablesample(bucket 4 out of 7 on incident\_region);**

**select \* from migrants\_buckets tablesample(bucket 2 out of 5 on incident\_region);** #tablesample dynamically collapses the data into 5 buckets

**select \* from migrants\_buckets tablesample(bucket 7 out of 12 on incident\_region);** #tablesample dynamically expands the data into 12 buckets

------------------------------------------------------------------

NOTE : No change in original bucketed data...7 buckets created at time of creation stays as it is.

BUCKETED TABLE

**hive> SELect \* FROM migrants\_buckets where incident\_region = "US-Mexico Border" or incident\_region = "Central America";**

*Time taken: 0.084 seconds, Fetched: 440 row(s)*

**hive> ANALYZE TABLE default.migrants\_buckets compute statistics;**

*Table default.migrants\_buckets stats: [numFiles=7, numRows=1509, totalSize=262546, rawDataSize=261037]*

*MapReduce Jobs Launched:*

*Stage-Stage-0: Map: 1 Cumulative CPU: 3.0 sec HDFS Read: 266998 HDFS Write: 86 SUCCESS*

*Total MapReduce CPU Time Spent: 3 seconds 0 msec*

*OK*

*Time taken: 56.037 seconds*-----------------------------------------------------------------

NON - BUCKETED TABLE

**hive> SELect \* FROM migrants where incident\_region = "US-Mexico Border" or incident\_region = "Central America";**

*Time taken: 0.136 seconds, Fetched: 440 row(s)*

**hive> ANALYZE TABLE default.migrants compute statistics;**

*Table default.migrants stats: [numFiles=1, numRows=1509, totalSize=252934, rawDataSize=249913]*

*MapReduce Jobs Launched:*

*Stage-Stage-0: Map: 1 Cumulative CPU: 3.0 sec HDFS Read: 256824 HDFS Write: 78 SUCCESS*

*Total MapReduce CPU Time Spent: 3 seconds 0 msec*

*OK*

*Time taken: 52.066 seconds*

------------------------------------------------------**THE END** --------------------------------------------