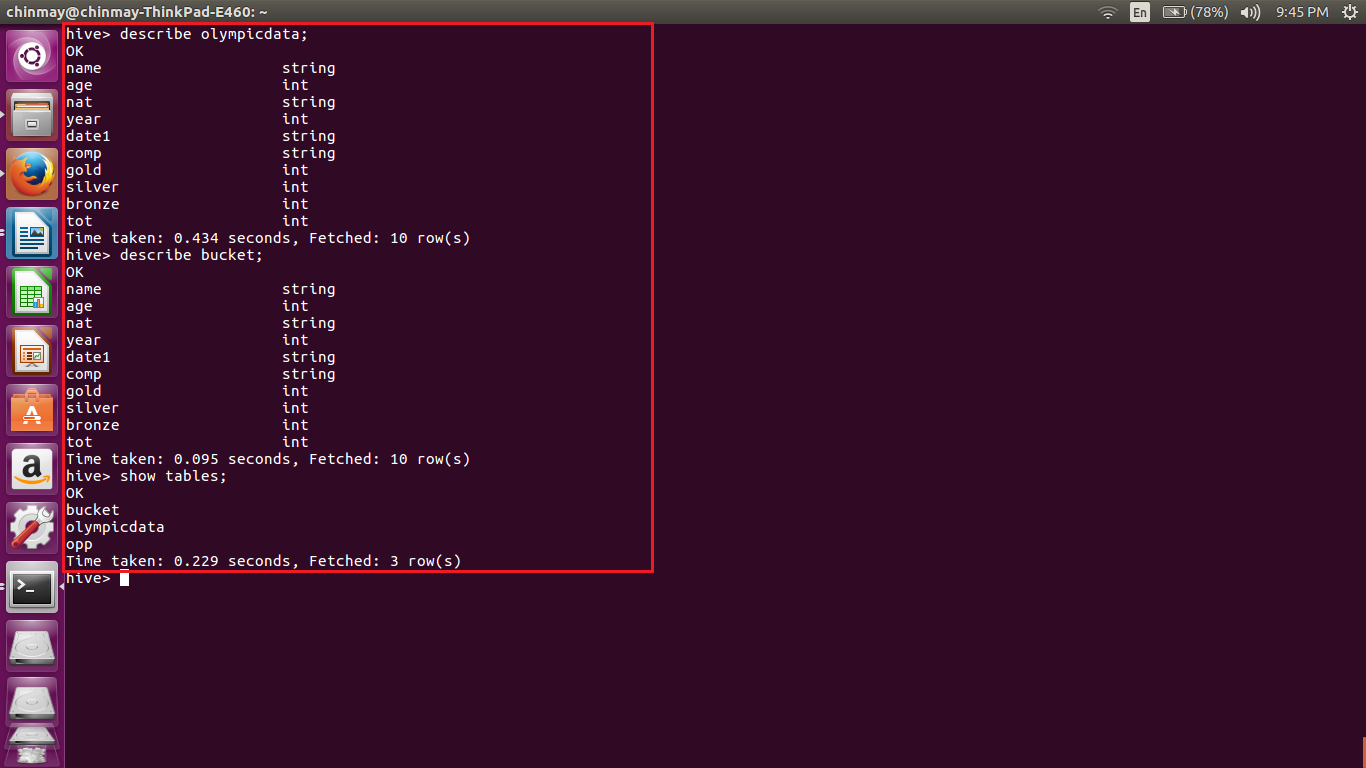
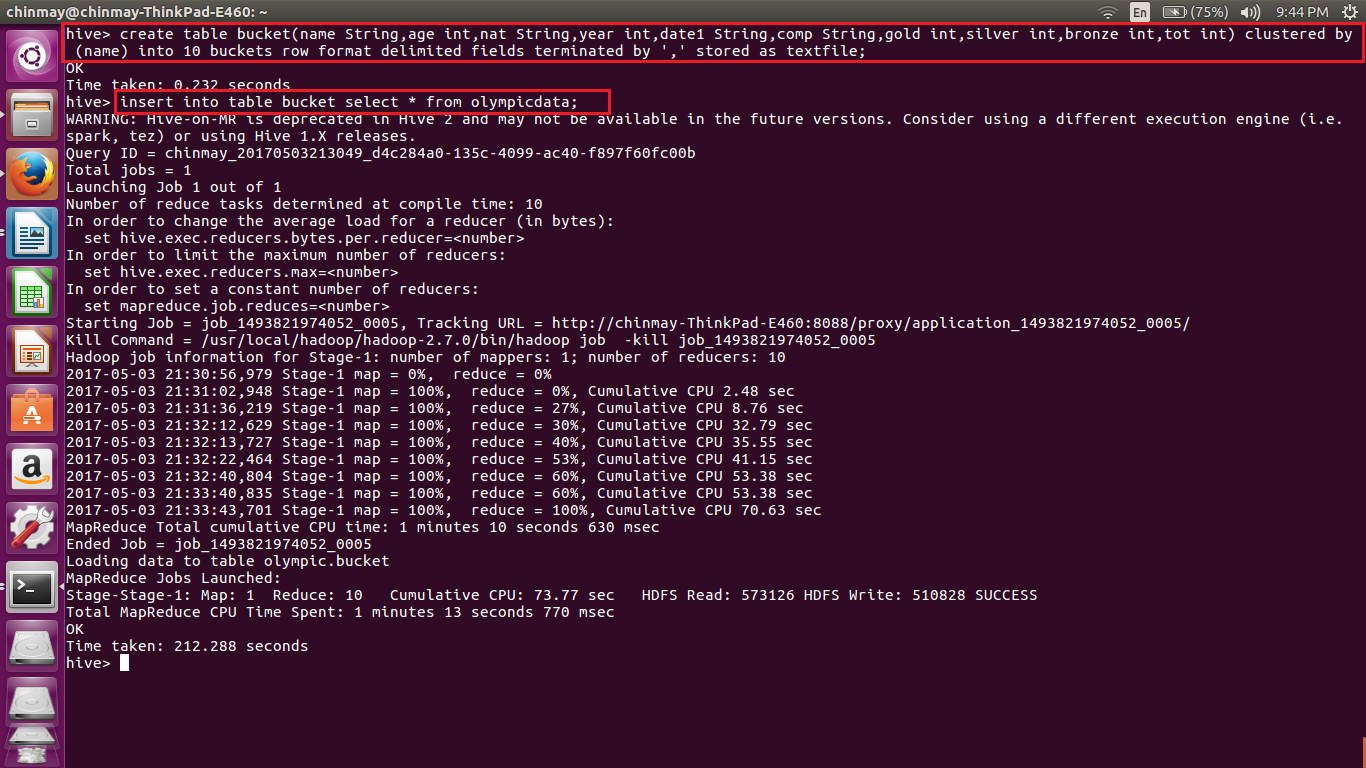
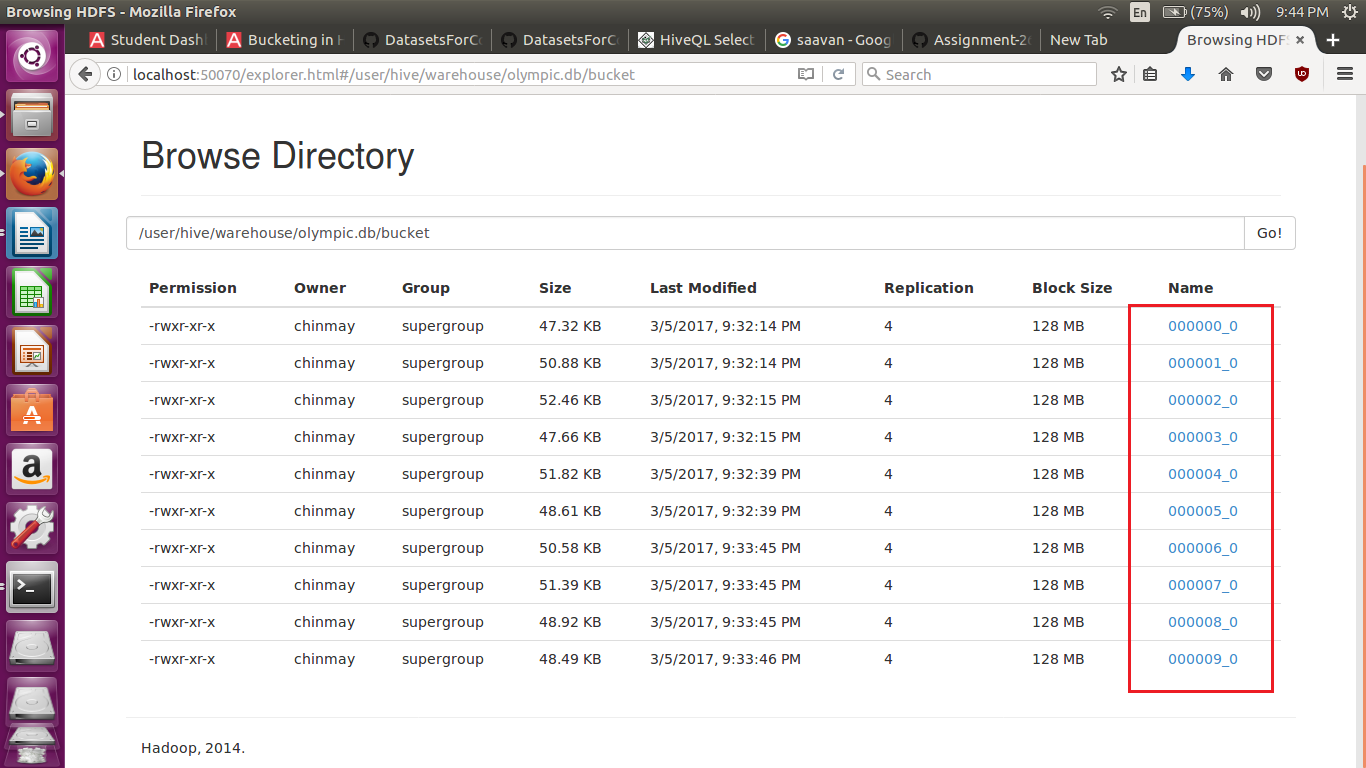
Explain the key concepts of Bucketing and perform bucketing operations using our attached Blog. Share and explain the commands used with the final result.





Here we are creating bucket using the field and distributed into 10 buckets.



In the hdfs we can see that 10 buckets have been created

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. |