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

Bucketed tables allows much more efficient sampling than the non-bucketed tables. With sampling, we can try out queries on a section of data for testing and debugging purpose when the original data sets are very huge. Here, the user can fix the size of buckets according to the need.

Bucketing concept also provides the flexibility to keep the records in each bucket to be sorted by one or more columns. Since the data files are equal sized parts, map-side joins will be faster on the bucketed tables.

To perform the bucketing operation on a dataset, we require an input dataset of real\_state.

Dataset Description:

Column 1 : Street

Column 2 : City

Column 3 : Zip

Column 4 : State

Column 5 : Beds

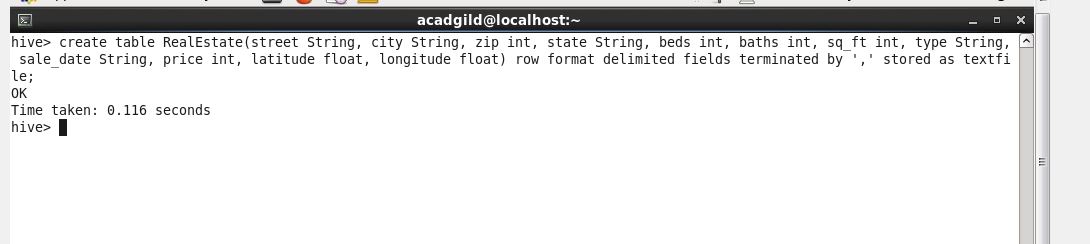
Column 6 : Baths

Column 7 : Sq\_feet

Column 8 : flat\_type

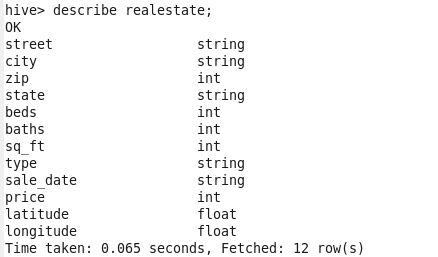
Column 9 : Price

● Create table RealEstate in Hive database, with all the columns present in dataset.

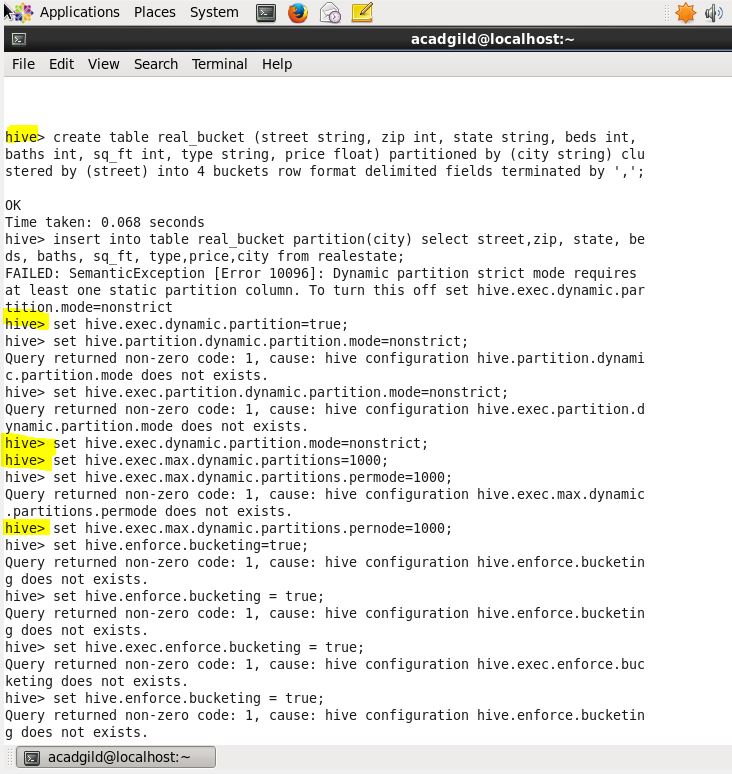


● Load dataset into table realEstate.





Creating Bucket Table:



From the above image, we can see that we have created a new bucket table with name ‘bucket\_table’, which is partitioned by ‘city’ and clustered by ‘street’ field with the bucket size of ‘4’

Here, we have decomposed Hive Buckets into ‘4’ parts.

