**ASSIGNMENT27.3**

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

**Bucketing in Hive**

Hive partition divides table into number of partitions and these partitions can be further subdivided into more manageable parts known as Buckets or Clusters. The Bucketing concept is based on Hash function, which depends on the type of the bucketing column. Records which are bucketed by the same column will always be saved in the same bucket. In Hive Buckets, each bucket will be created as file. Bucketing can also be done even without partitioning on Hive tables.

**Advantages:**

* Bucketed tables allows much more efficient [sampling](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+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.

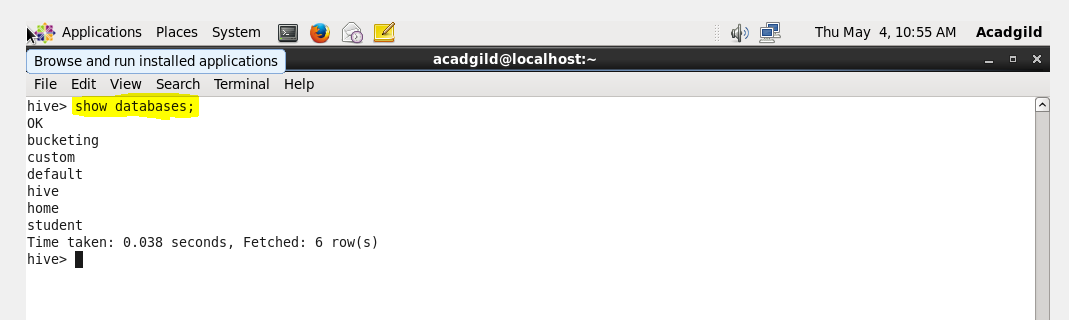
***Example:***

***Input Dataset to Perform Bucketing Operation:***



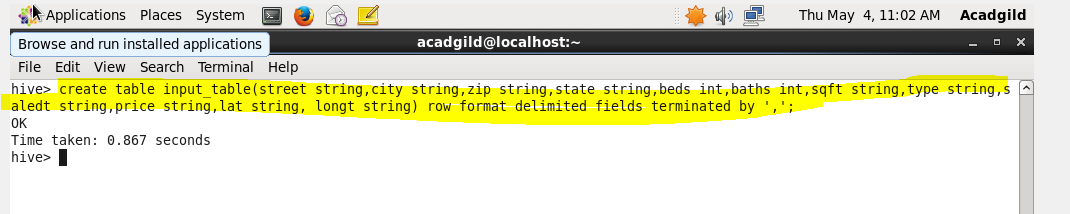
***Selecting the Database:***

we have used a database named ***‘****bucketing****’***which already existsin the Hive metastore database and we will create new tables in the ‘*bucketing*’ database, to perform Bucketing operation on the newly created tables.



***Creating a New Input Table:***

Next, we have created a new input table with name ***‘input\_table’,***where we will save the contents of the input dataset.



***Load the Input Dataset:***

Now, in the below image, we can see that we have loaded the contents of ***real\_state*** into input***\_table*** by using the load command.

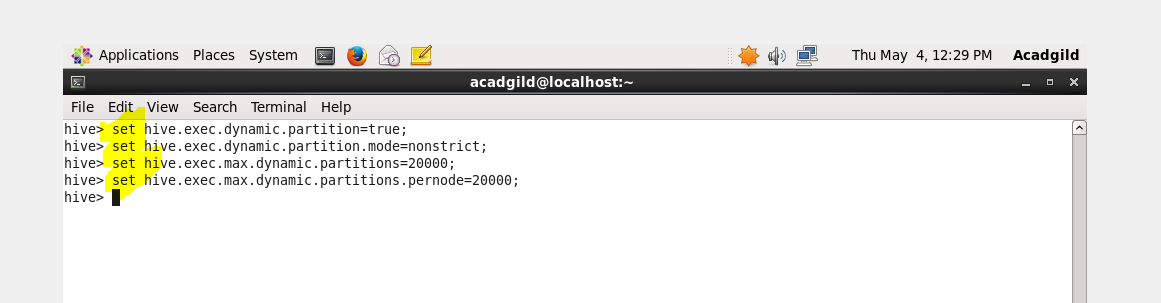


***Display the Contents of Table input\_table to Ensure Whether the Input File has been Loaded Successfully or Not:***

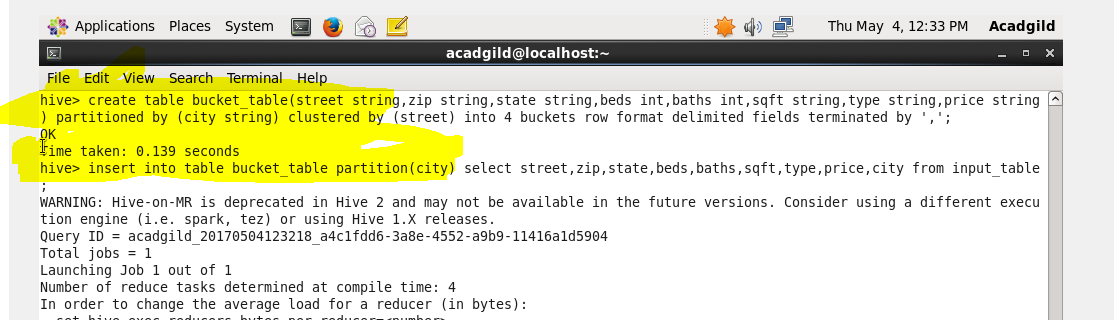
By using ‘*select’* statement, we can check the contents of the ***input\_table,***whether we have successfully loaded the contents or not.

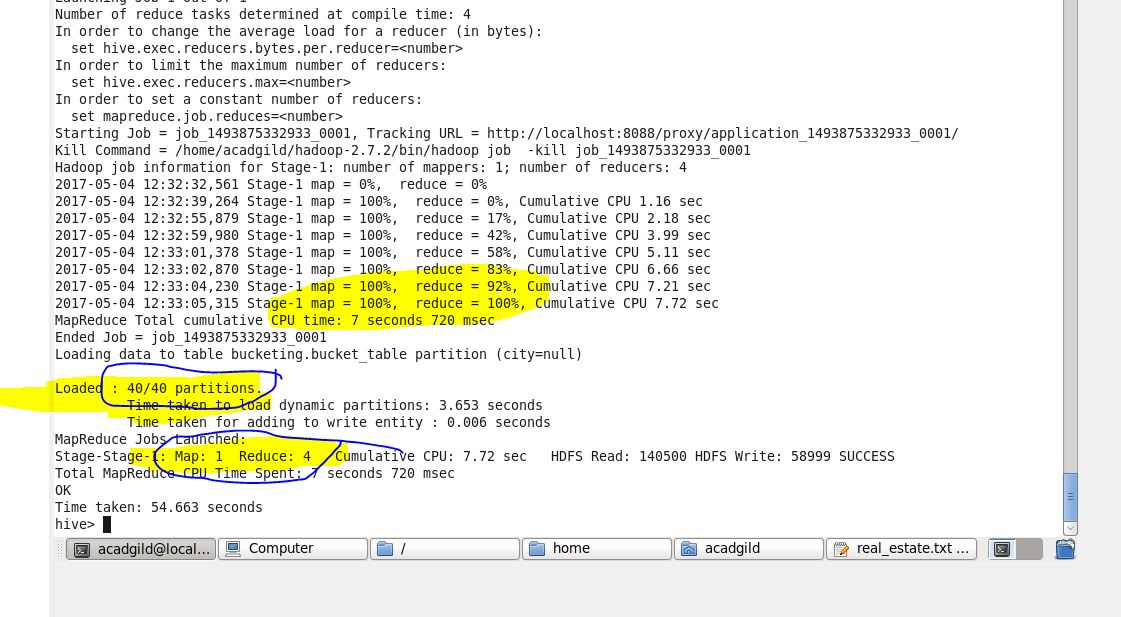


***Setting the Below Properties in Hive Command Line***

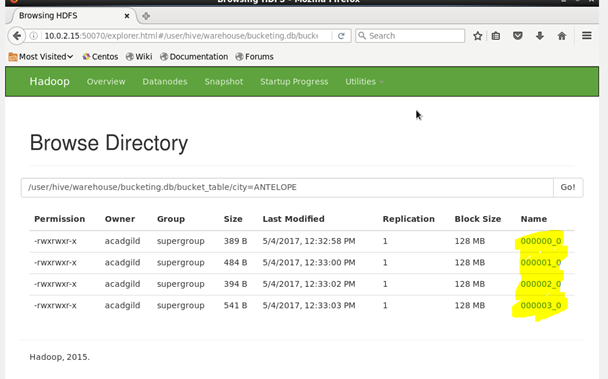


***Creating Bucket Table &Query to Retrieve Data from Bucketed Table:***

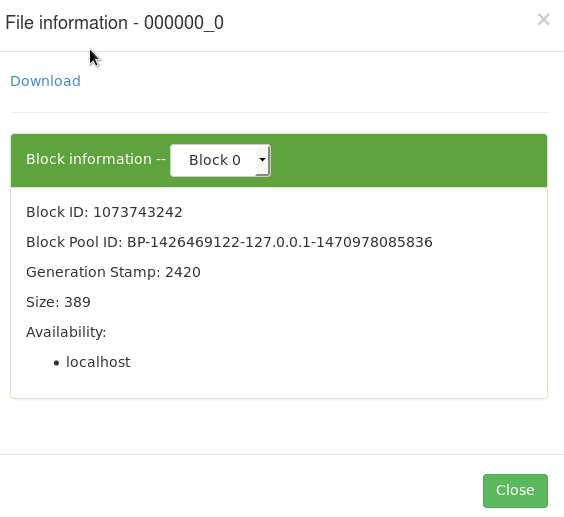


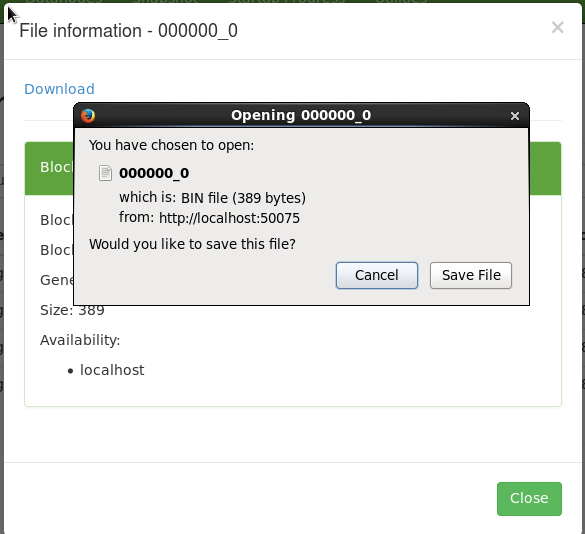


**Browse Directory:**



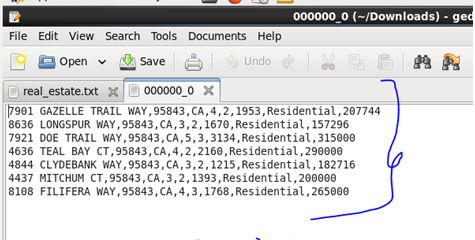
***Browsing the file:***



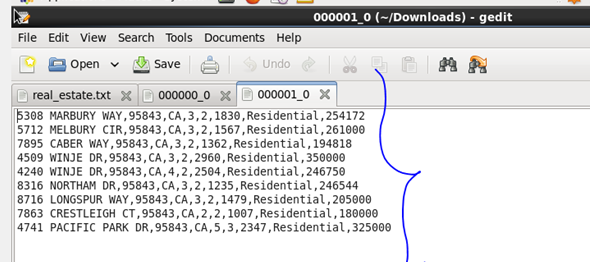


**Output:**

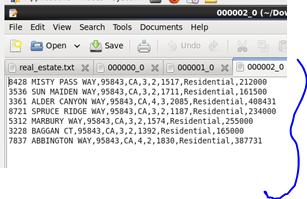
**Bucket-1**

****

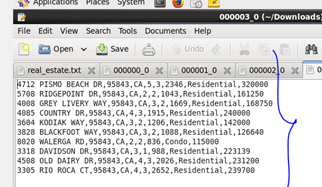
**Bucket-2**

****

**Bucket-3**

****

**Bucket-4**

****