

**Bucketting**

In order to Increase the performance of queries Partitions are introduced in hive.So if there is a huge dataset regarding “world population” and suppose if we want to filter data by each country using where hive has to scan the entire dataset.**To ensure faster querying Partioning is made on country name** and

**Now for each country a directory will be created on hive/warehouse and the querying can be faster**

**Limitation with hive partitions:**

**No1:**

**If the dataset is so large and now for each country a directory will be created which will cause a increased overload on namenode**

**No2:**

**Now think of the above example where partitions are nade on country.Since population varies the Data say 100 GB for say 100 countries will not be equal.So again the processing on these partition will increase time if we use a group by like operation So to encounter these issues hive provides BUCKETTING**

**First problem is encountered as bucketing creates this much number of buckets so whatever may be the size the entire data will be divided among these buckets**

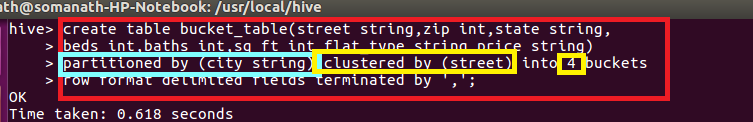
**2 Problem is encountered by since hashcode is code in the range of 1 to 10 say 10000 records all these data will be divided within these equally as 1000**

**Similarly If we want to further classify partitioned data,bucketing can be made over partitioned data and the bucketed record will be stored as files within the directory**

**So from above scenario we can define bucketing as**

**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 this example I am creating a realestable table for easy querying I am partitioning by City and bucketing by street as shown

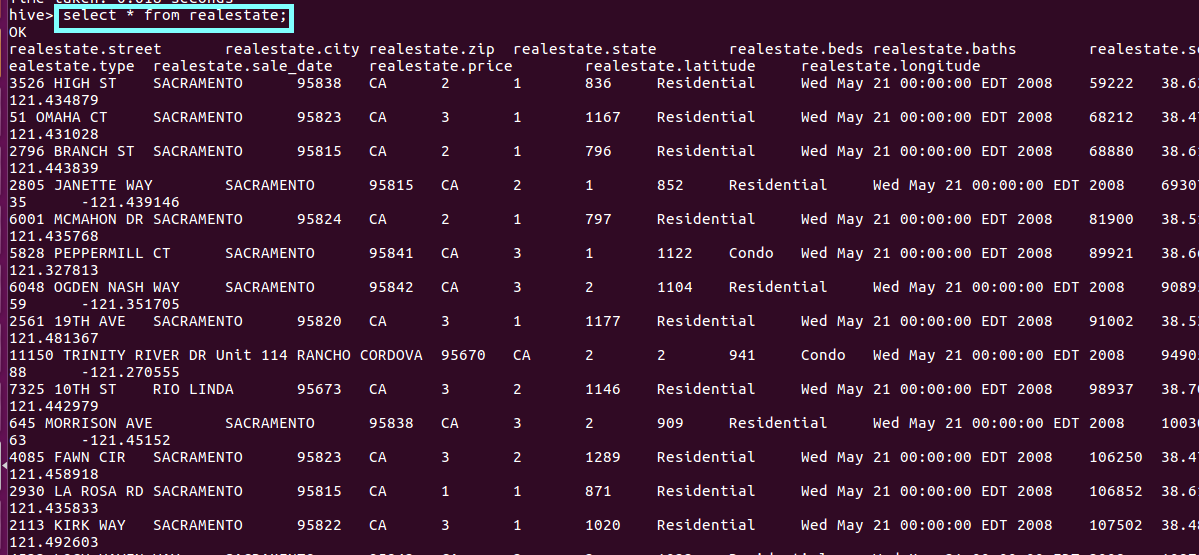


The point to note is that a directory will be created for each city and within each directory Four files will be created as 4 buckets is given based on street

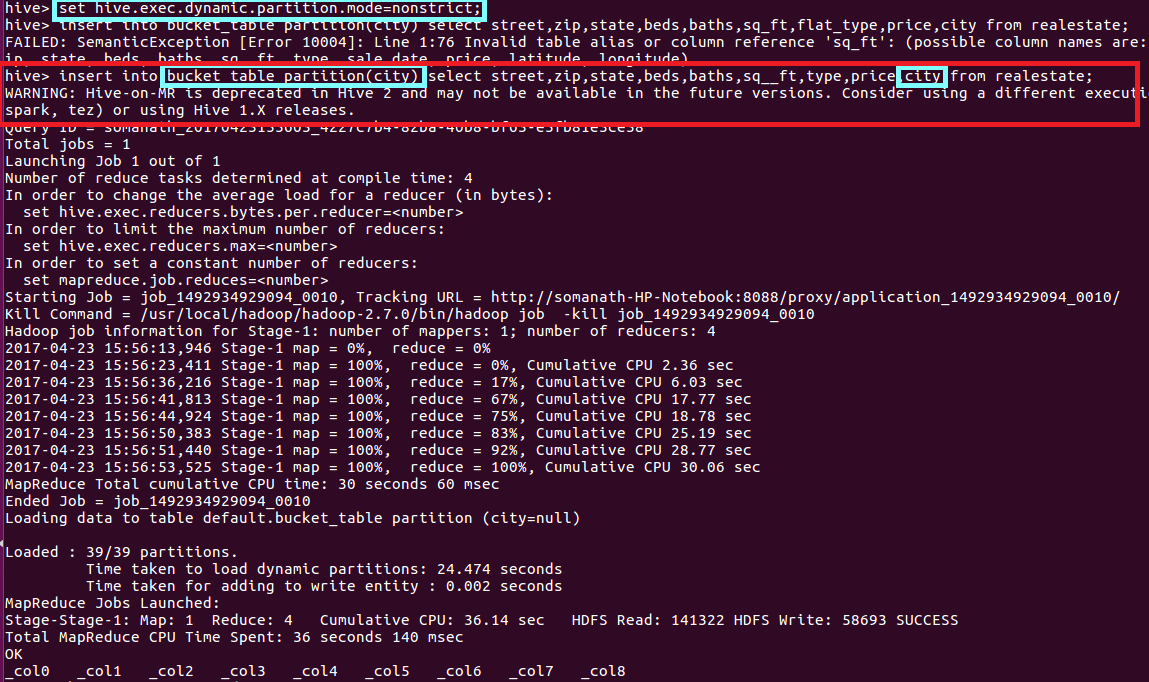
In order to allow bucketing **set.hive.exec.dynamic.partitions=nonstrict**

**In order to insert data into bucketed table we must have a table and so we are inserting data from unpartititioned realestate table into this bucket table**

**First I am checking realestate table**

****

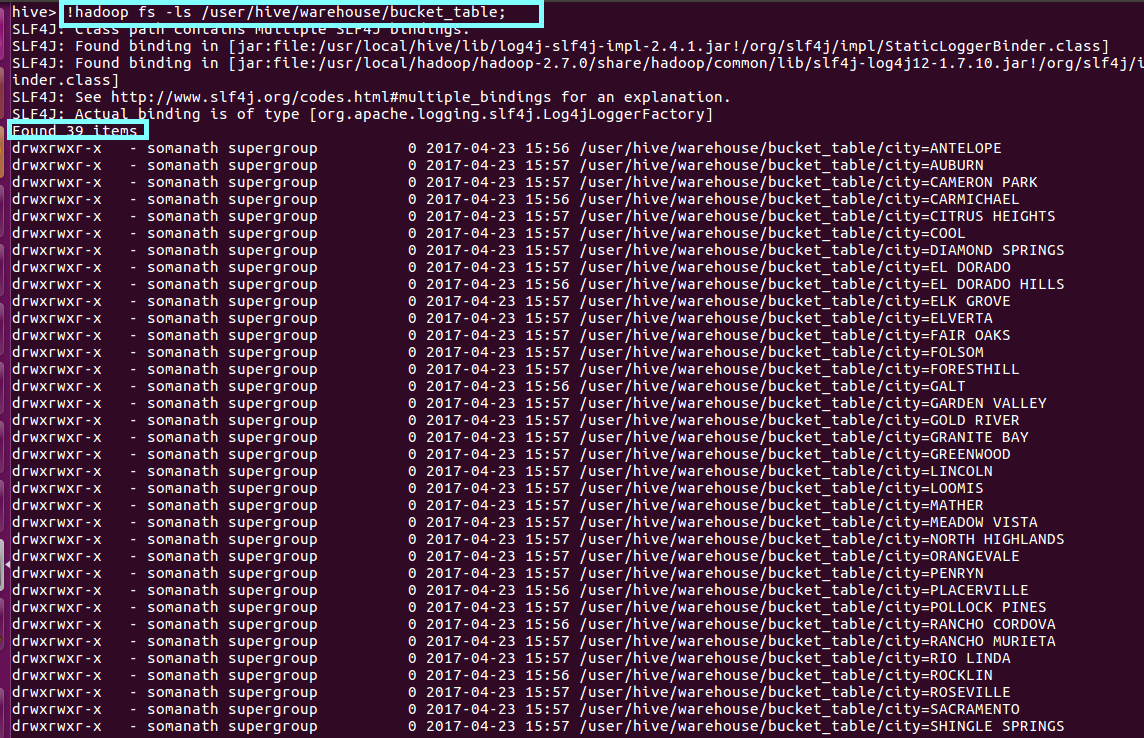
**Since we are using dynamic partition the last column should be city column(based on which partition is made)**

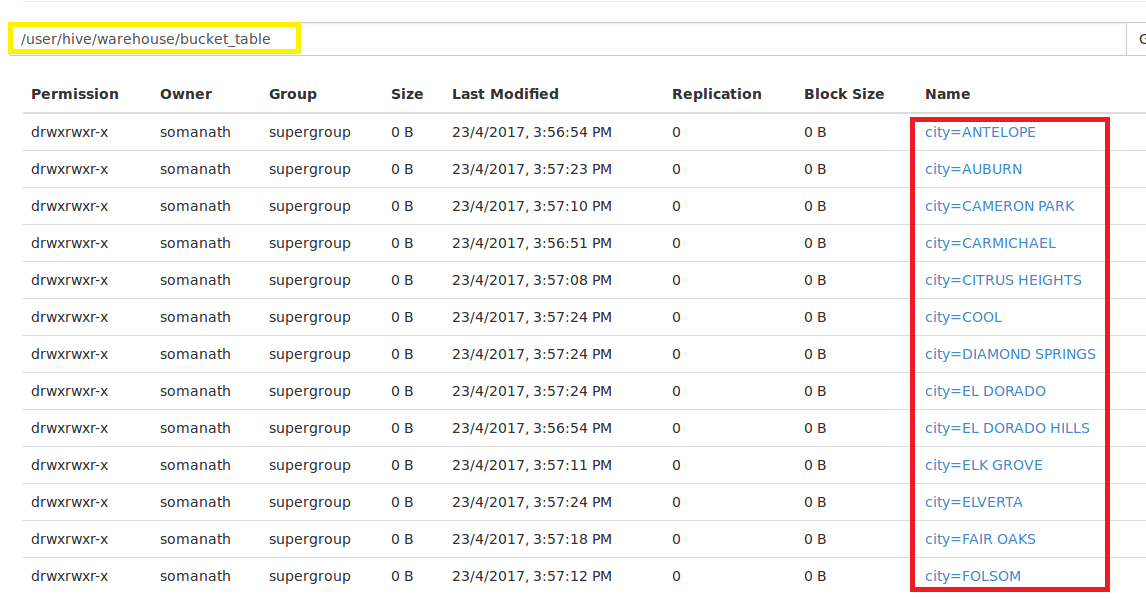
****

**Output**

**We can see that as expected a directory is created for each city as shown below**

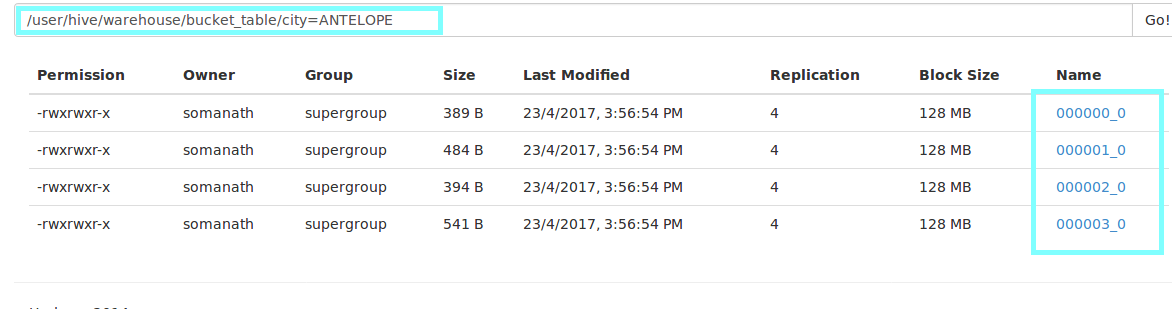
**Since there are 39 distinct cities 39 directories have been created**

****

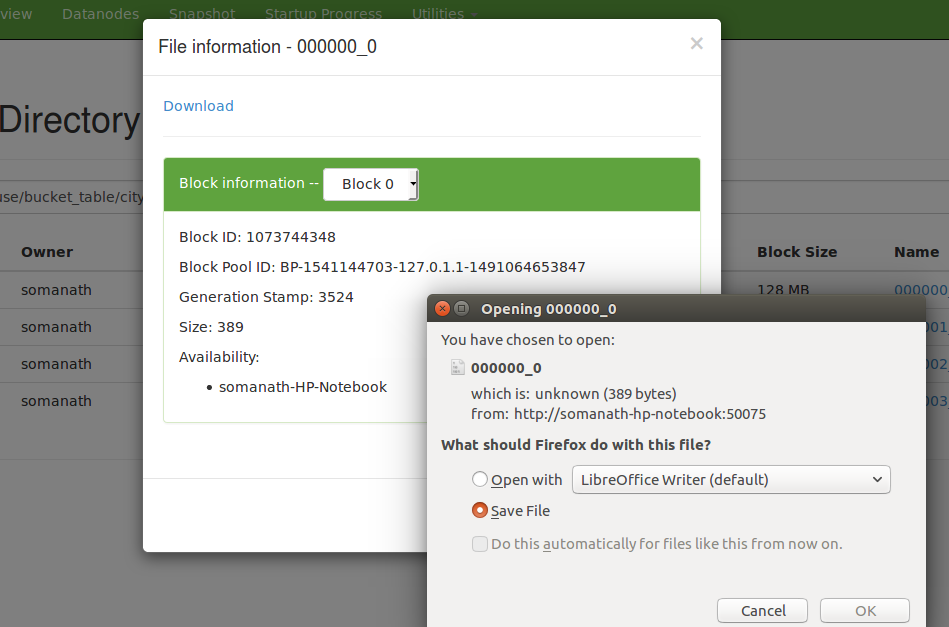
****

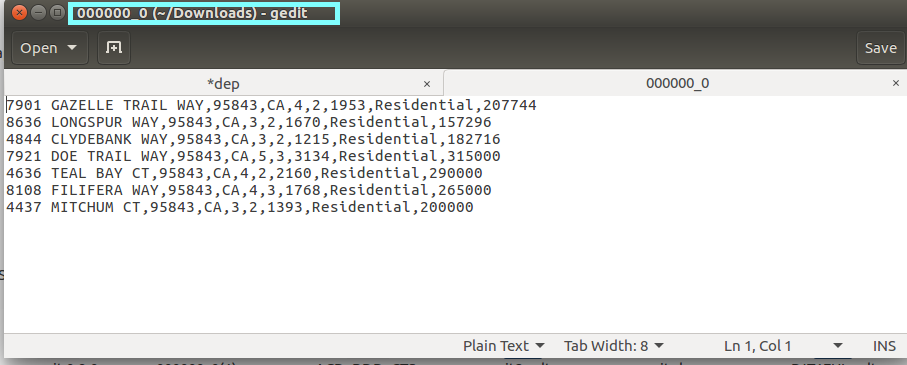
Now when we go inside directory we will see 4 file will be created as 4 buckets are mentioned to which data will be sent based hashmodulo of the street

For example if we see Inside antelope directory

****

Now we are accessing the file by downloading the file as shown

****

****

**Conclusion:**

Creation of partitioning results in creation of directory which result in memory overhead if file is large .So we are creating partitioning based on city (major)and then further managing partitioning by creating buckets based on streets.