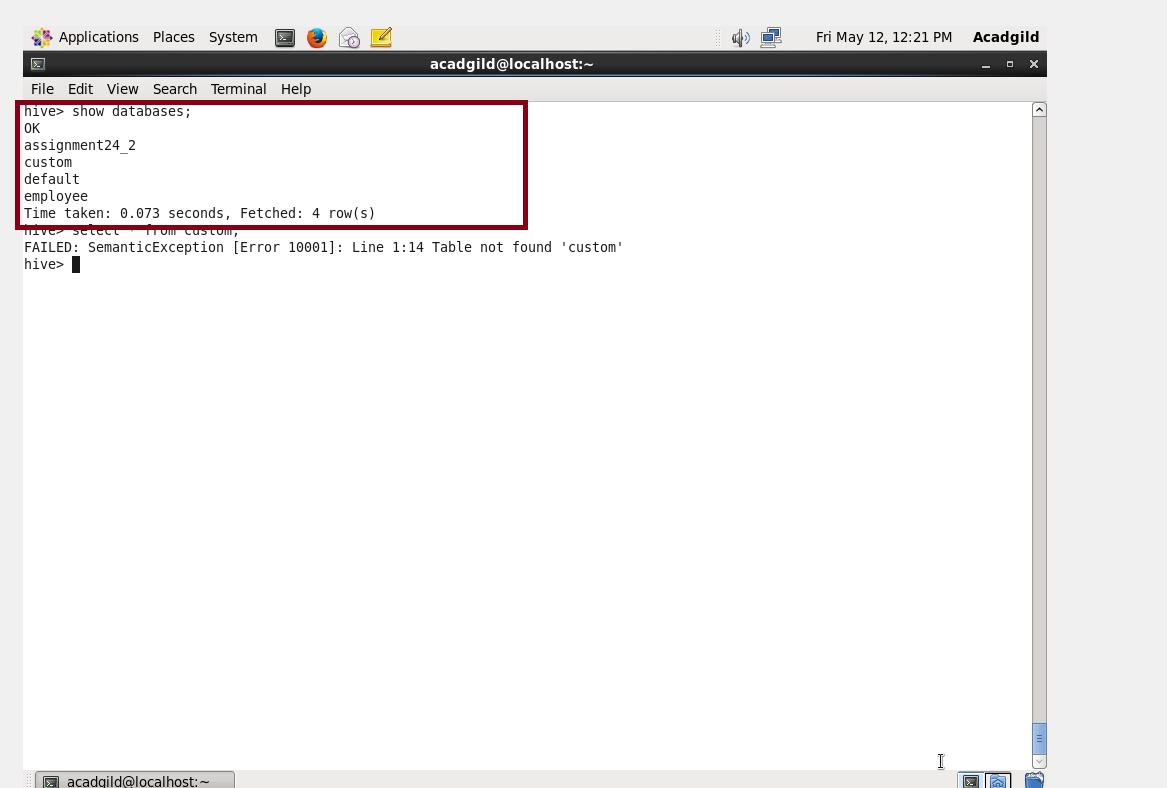
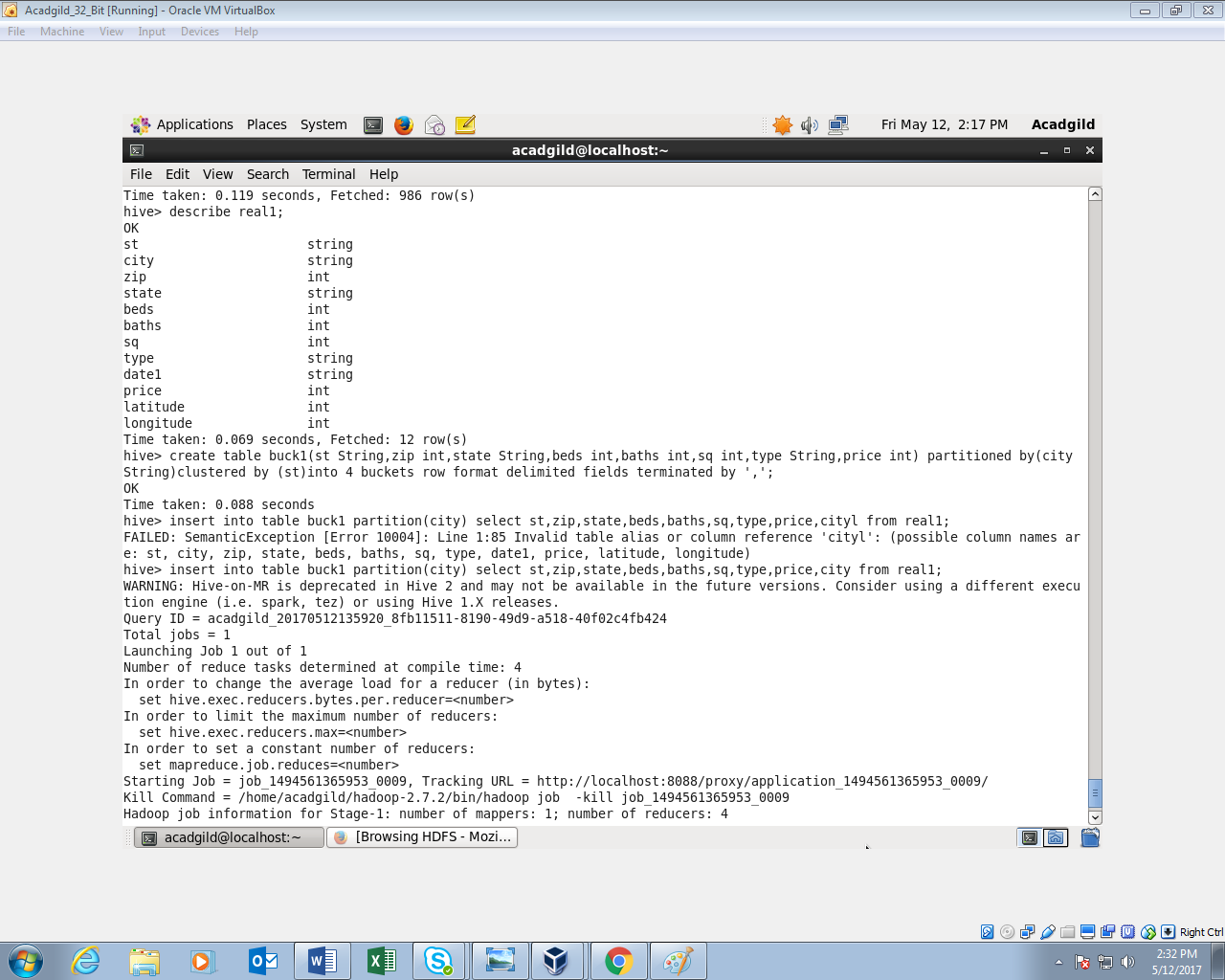
***Problem Statement:***

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

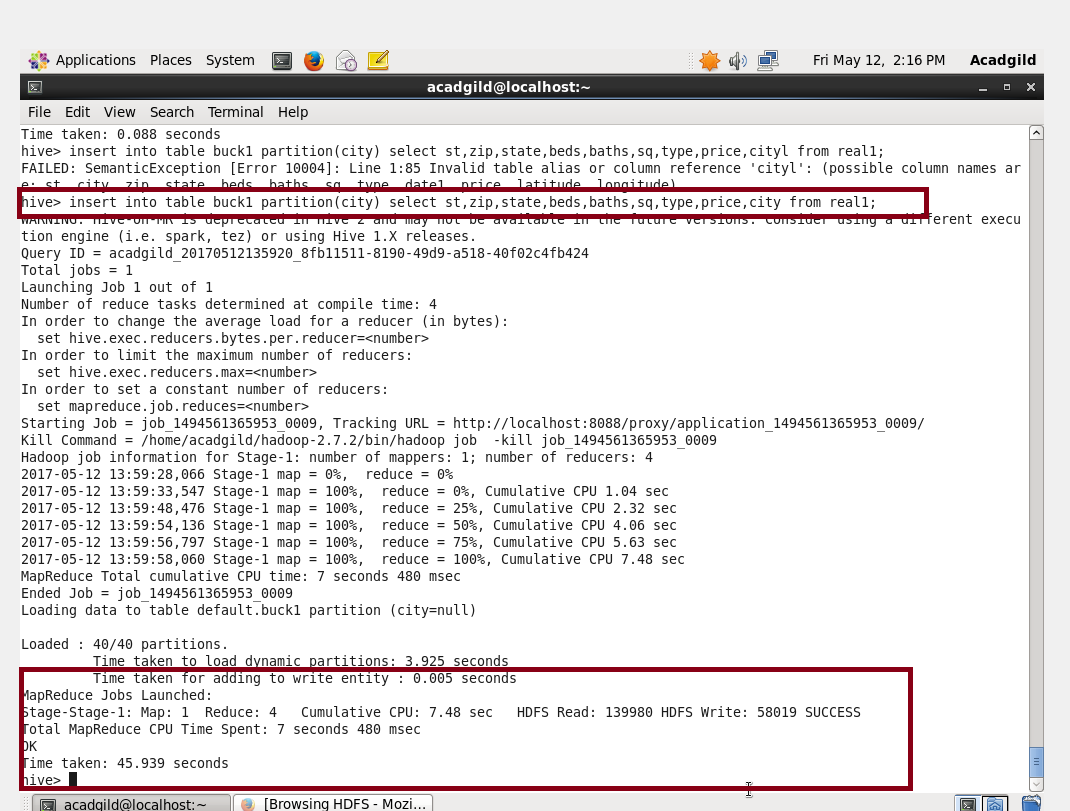
All the data base were shown:

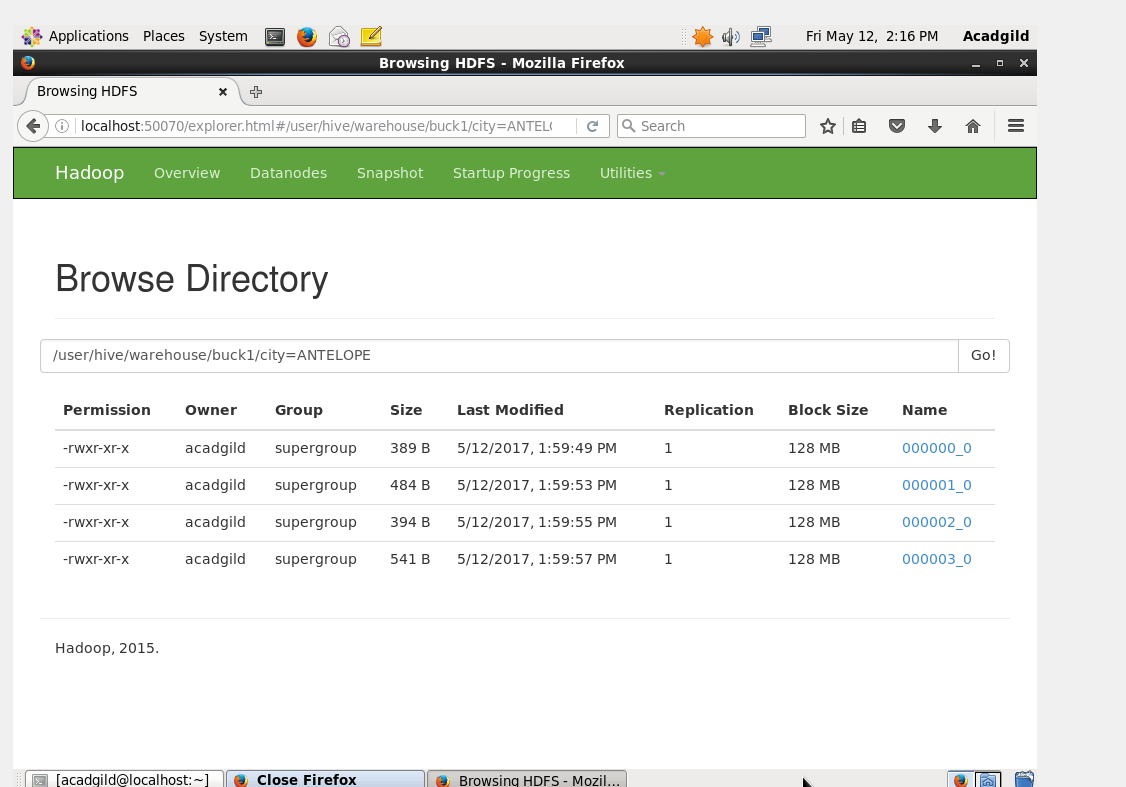


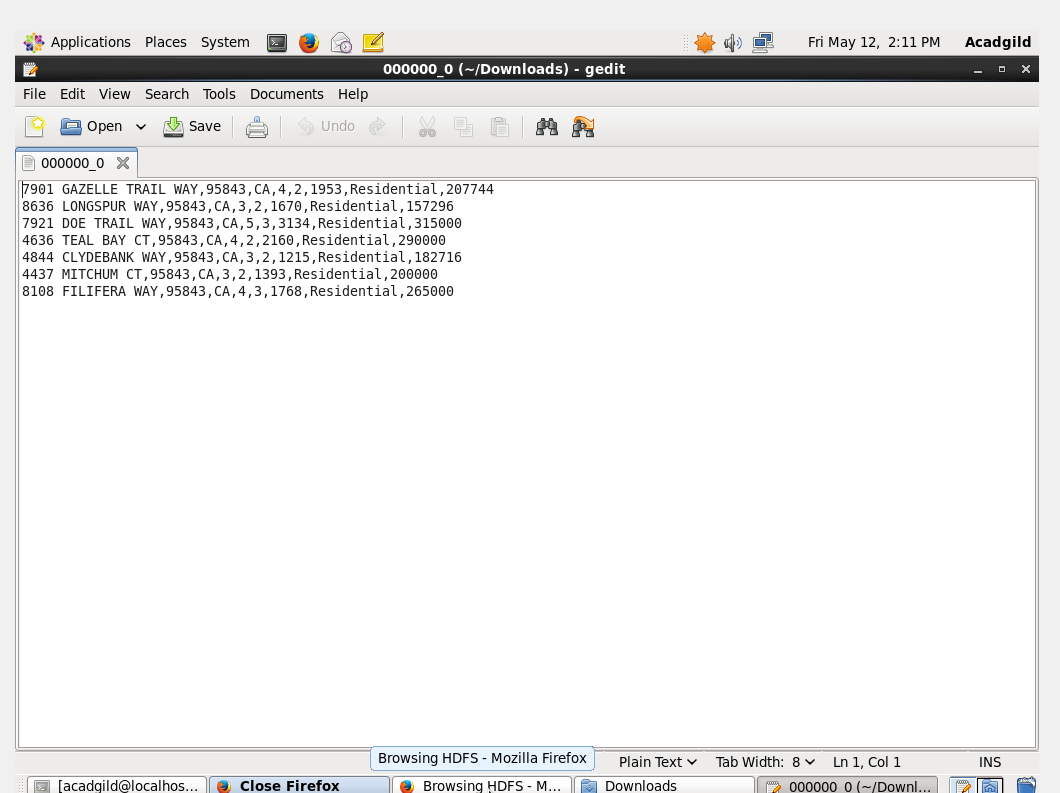
* Real\_estate data was already load for previous problem.
* And so described to know the content

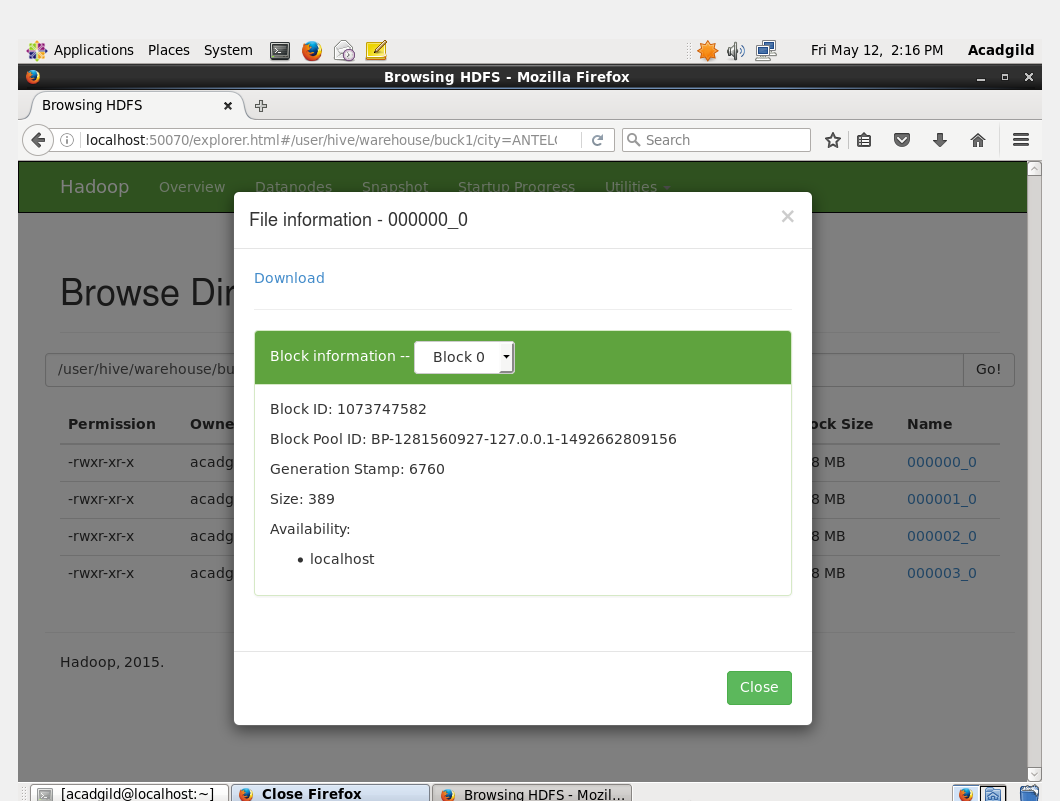


* To the existing table the data was partitioned using city.
* And other necessary details were selected.









In the hdfs we can see that 8 buckets have been created

**Bucketing**

* Bucketing is similar to partitioning concept.
* the main difference is the number of partitions that need to be divide will be decided only by the user. Hence the problem of high cardinality is looked after automatically.
* This is also one reason why bucketing is used for high cardinality datasets.
* CLUSTER BY is the keyword used for bucketing is.
* 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. |