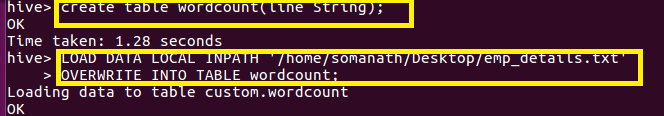
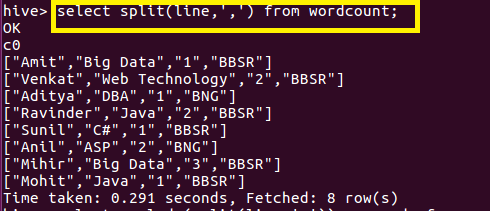
In order to perform wordcount using hive

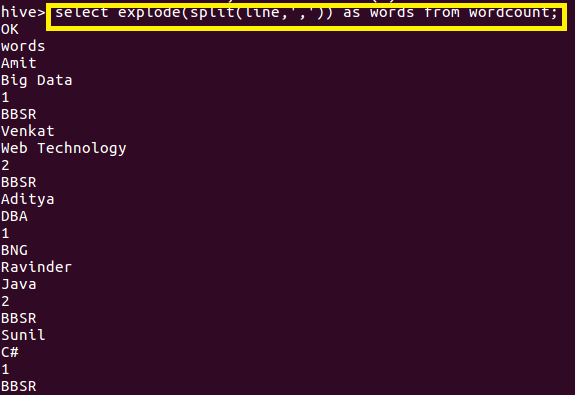
**Step1: load the entire row data as a string in a table**



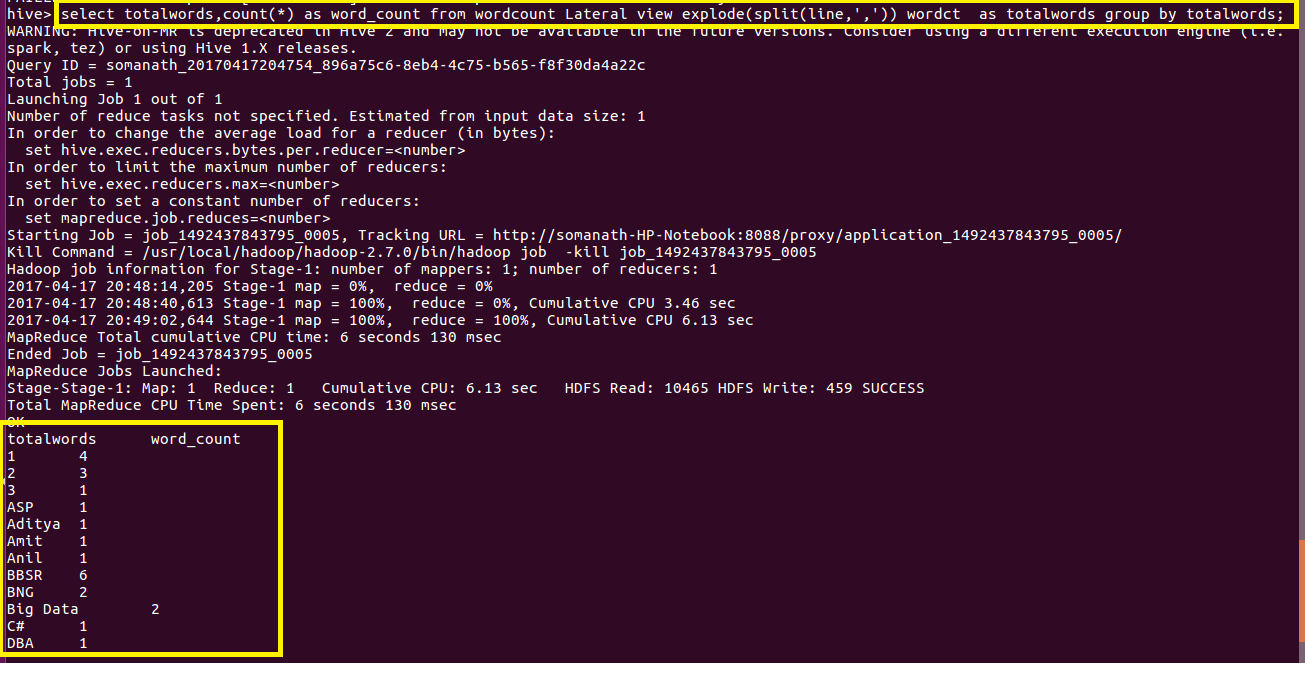
**Step2:used Split function to Separate the line by comma to get as an array of words as shown**



**Step 3 :Used explode command which is used to separate the array element in rows**



**Step4:Since we want word count I used lateral view which will create a temporary table “ wordct” in which this explode and give a alias name as Totalwords and group by totalwords which will group by each words and used select totalwords,count(\*) function from this word\_ct view table**



**Partioning:**

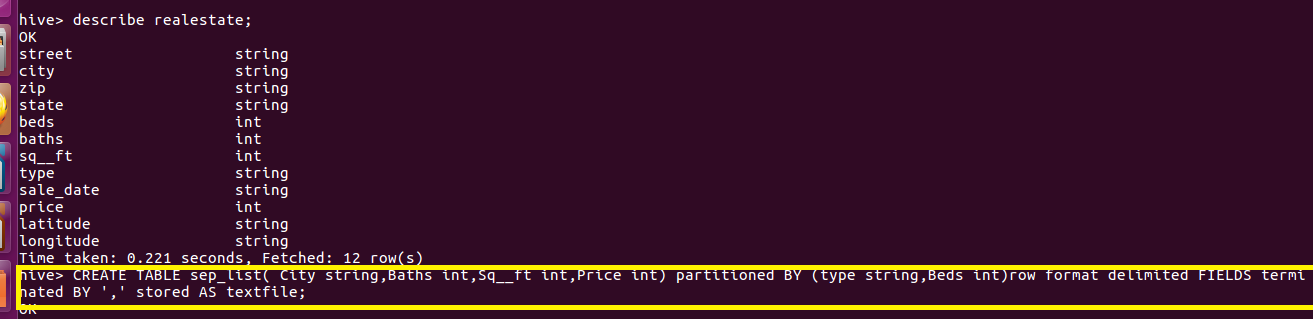
Why do we need partitioning?

Since during querrying in hive for a large dataset,for a simplequerry say we want HOUSES with beds=2 it want to search the entire database and find the result which will take a large time .So in order to optimize querrying time Partition is introduced

In Partitioning,the data is divided into directories based on column specified under Partition so that while querring the hive searches form that directory (eg ) if we use people from beds=2, It will go directly to that directory and the operation will be performed

**Creating Partitioned Table**:It is similar to ordinary table but just adding Partitioned By() where column based on which partitioning needs to be done

Here we will create table with partition for which we will add data from a large realestate data based on bedrooms and flat\_type

****

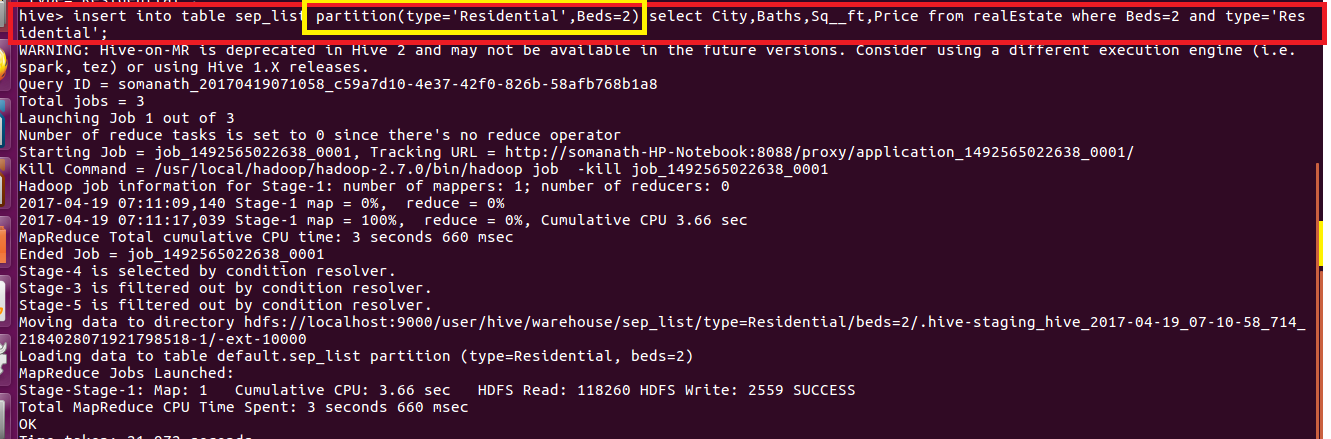
**Static Partitioning**

**Scenario:**

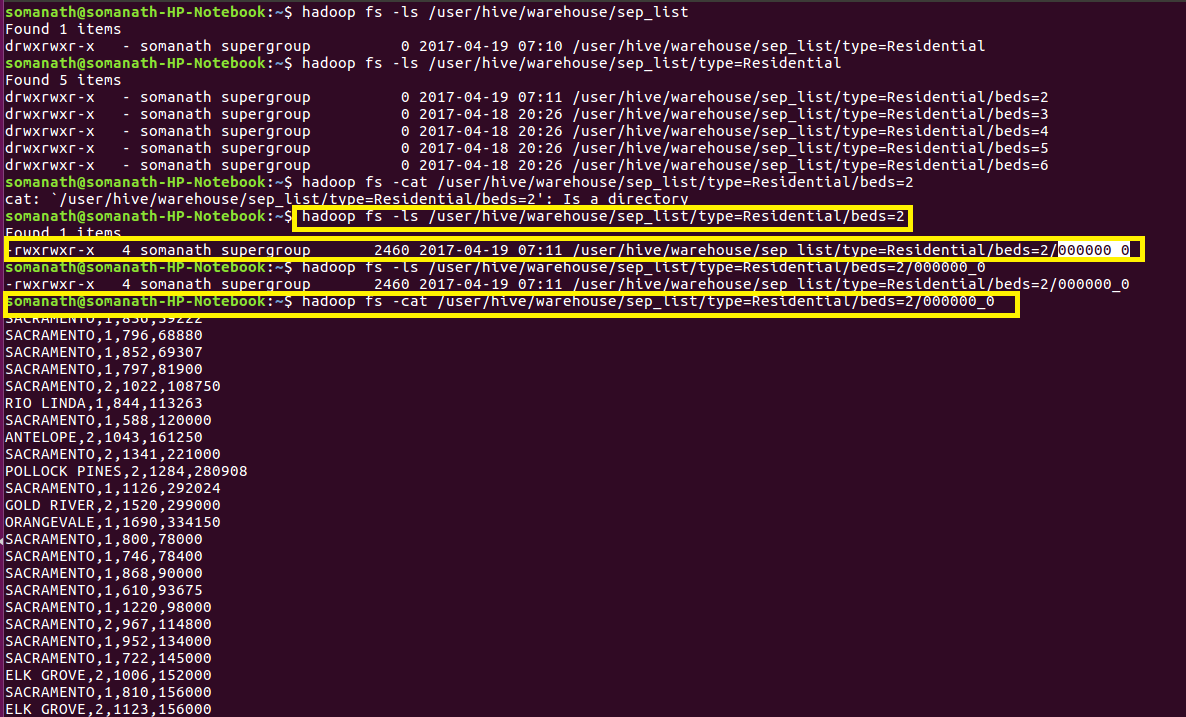
**WE have a real estate database for which we want a separate list of only data with bedroom=2**

**So we will create a static partitioning of bed=2**

**Thus in static partitioning,we know the type of data for example if we know the data is of bedroom with 2 we will create a partitioning into the partitioned table by giving the following command by giving type as residential and beds as2**

****

**Output:we can see that a separate directory for residential as type and bed=2 is created which contains all the data of beds=2**

****

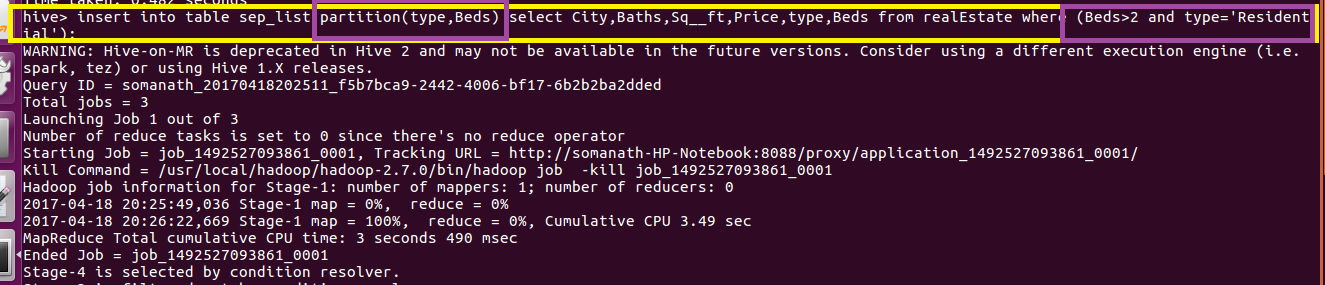
**Drawback: the major problem with static partitioning is that suppose the dataset is large and contains house with bedroom from 2 to 6 .**

**So we want to type the same SQL querry 5 times with beds=3, beds=4, beds=5, beds=6 if we use static partitioning**

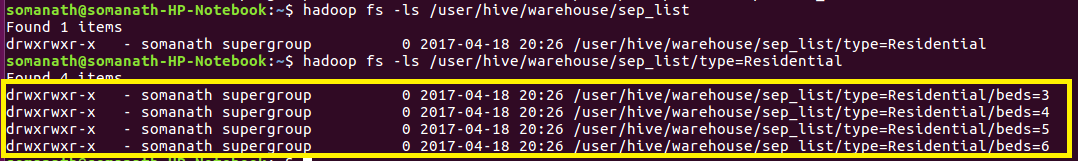
**So for this we use dynamic partitioning in which we will not specify the beds as 2 but we will just mention the columns on which partition to be done and the columns will be added as last two column in select statement so that hive will automatically do partitioning**

**Dynamic Partitioning**

**So if I want the partitioning to be done on houses on bedrooms greater than 2 I will just specify the columns based on which partitioning need to be done and the hive will automatically do partitioning with beds=3, beds=4, beds=5, beds=6 as shown**

****

**Output**

****