**Assignment 26.6**

* **Explain the differences between static and dynamic partitioning in hive and their working procedures?**

**Partitioning:**

* Working of Partitioning

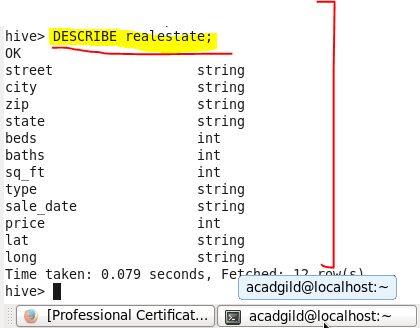
While writing query in hive for a large dataset, for a simple query say we want HOUSES with beds=2 then it will search the entire database and find the result which will take a large time .So in order to optimize querying time , Partition is introduced.

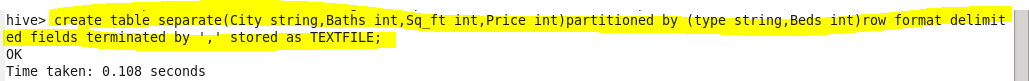
In Partitioning, the data is divided into directories based on column specified under Partition so that while writing query the hive searches from that directory for exp. if we use people from beds=2, It will go directly to that directory and the operation will be performed

* 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 type

****

****

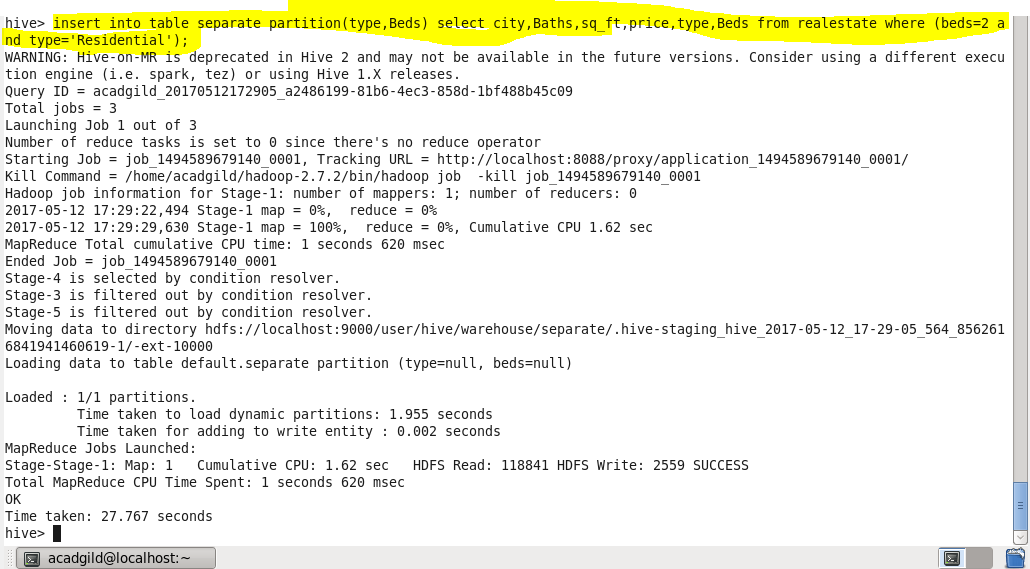
* **Static Partitioning**

**Scenario:**

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

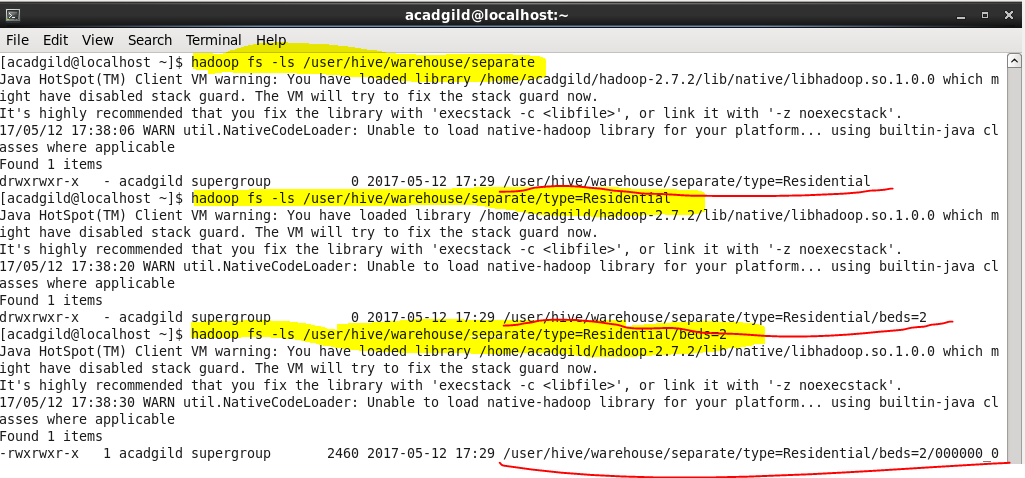
So we will create a static partitioning of bed=2

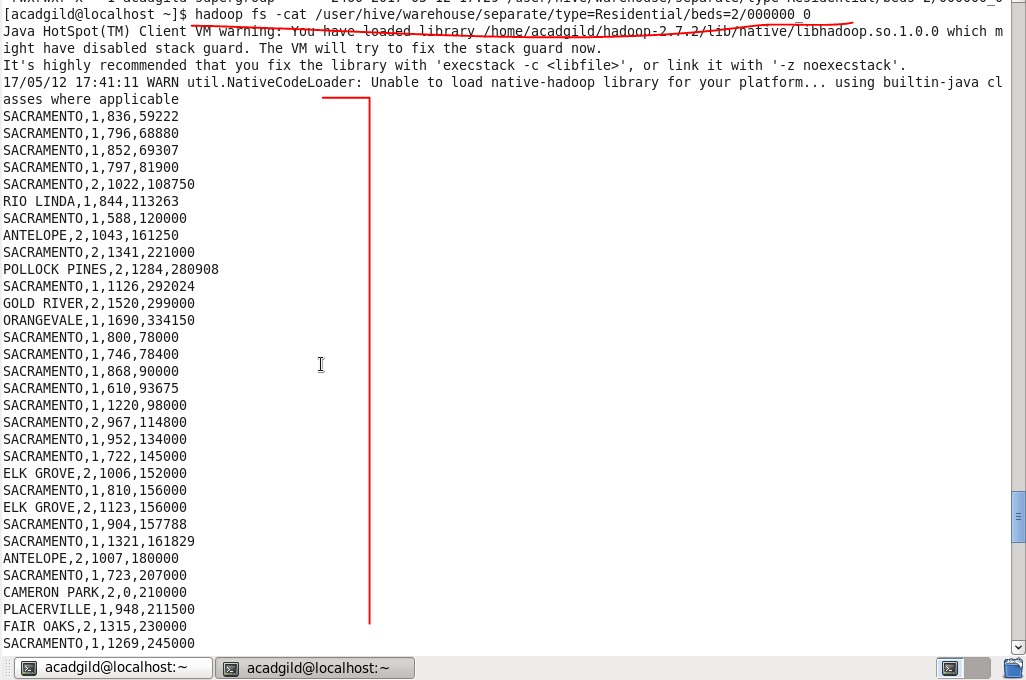
**Thus in static partitioning, we know the type of data for example if we know the data is of bedroom with 2 we will create a partitioning into the partitioned table by giving the following command by giving type as residential and beds as 2**



**Output:-**

We can see that a separate directory for as type = ‘Residential’ and beds=2 is created which contains all the data of beds=2

****



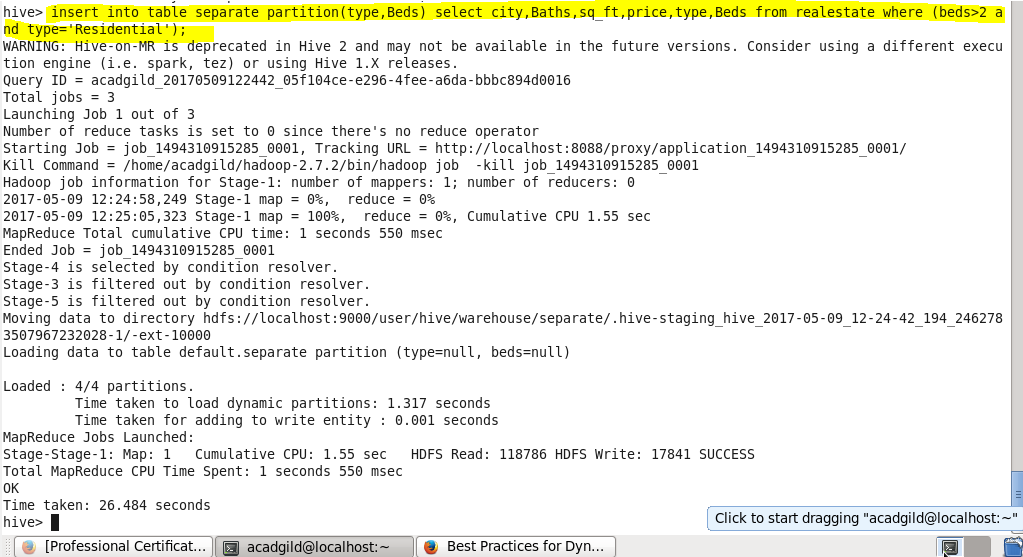
**Drawback:** the major problem with static partitioning is that suppose the dataset is large and contains house with bedroom from 2 to 6 .

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

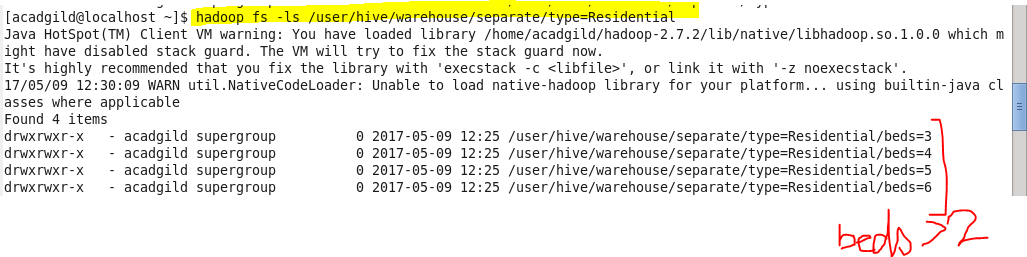
**So for this we use dynamic partitioning in which we will not specify the beds as 2 but we will just mention the columns on which partition to be done and the columns will be added as last two column in select statement so that hive will automatically do partitioning**

* **Dynamic Partitioning**

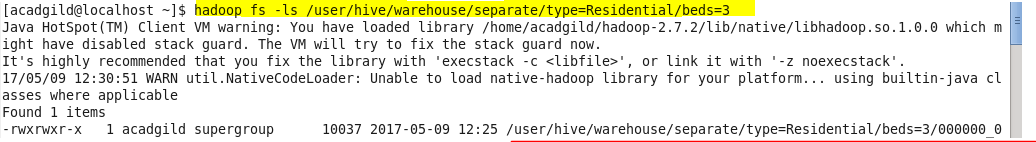
**So if I want the partitioning to be done on houses on bedrooms greater than 2 I will just specify the columns based on which partitioning need to be done and the hive will automatically do partitioning with beds=3, beds=4, beds=5, beds=6 as shown**

****

**Output**

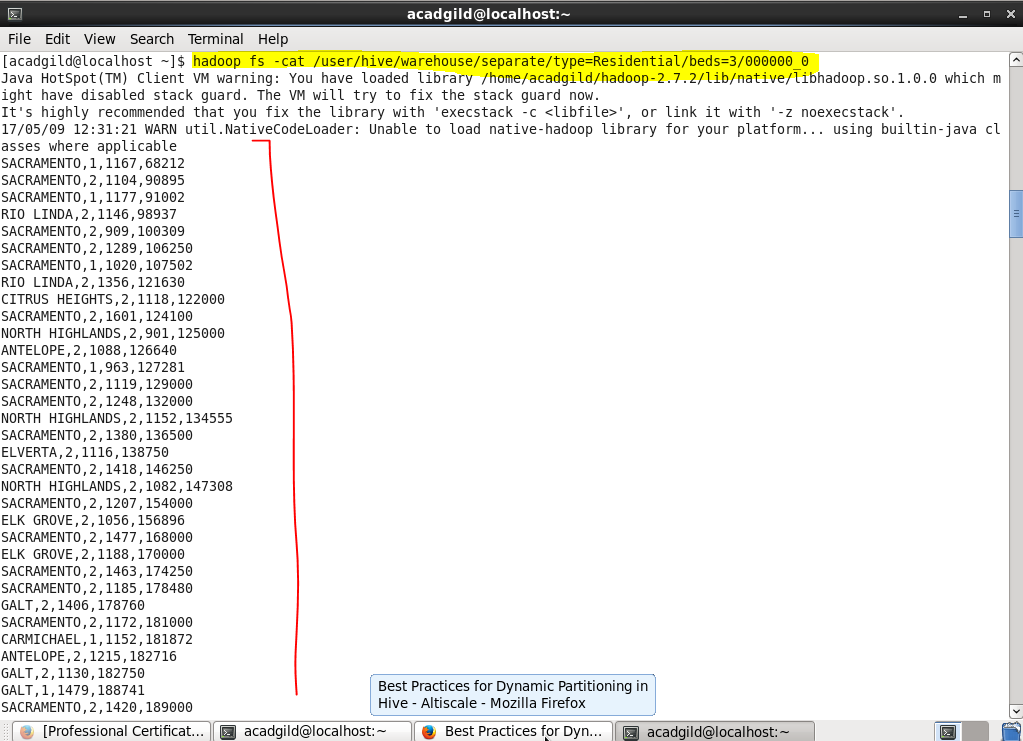


**So our separate list for bed =3 will be under this directory as shown**

****

**Now we can check for the result using cat command in HDFS.**

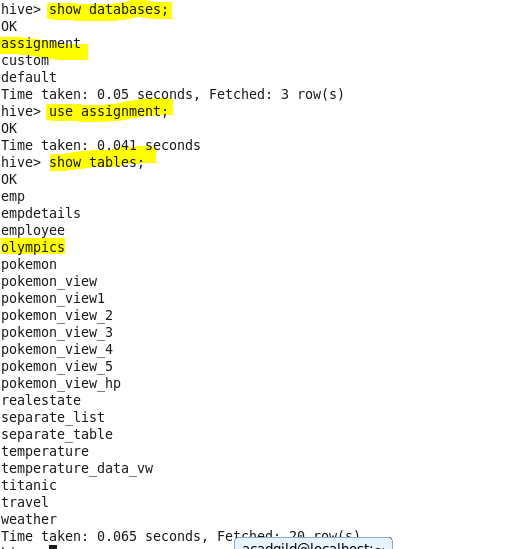
**Sample Output for bed =3**

****

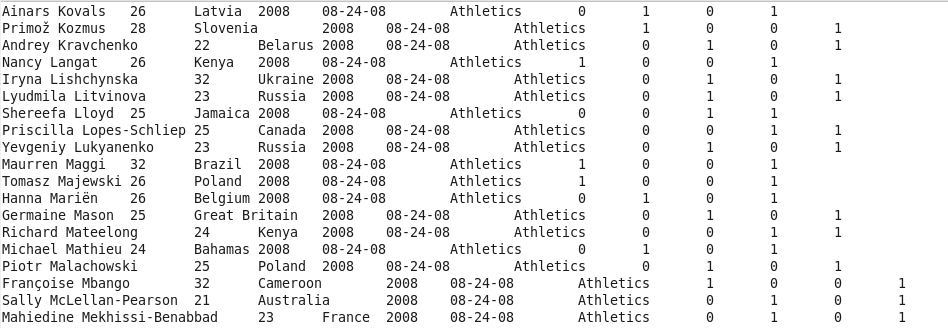
**So on..**

* **Use static partitioning in hive and evaluate the below problem statements**

Using The Database And The Already Created Table Inside It

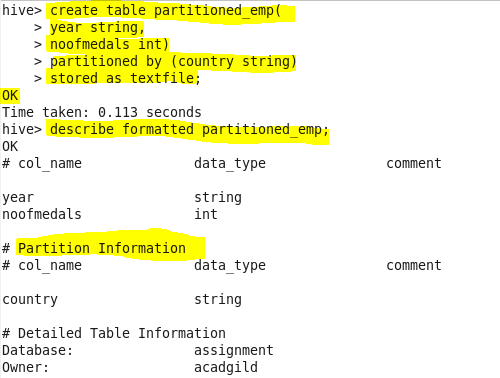


**DATASET:**

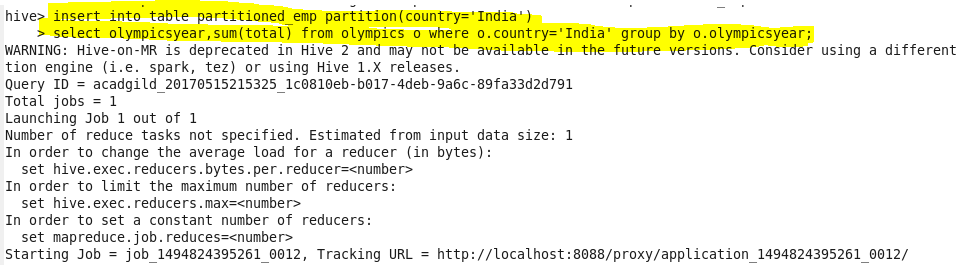


* **Find the number of medals india won year wise**

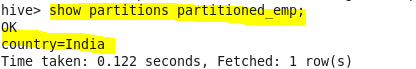
**First we create a partitioned table to generate the data based on the country**



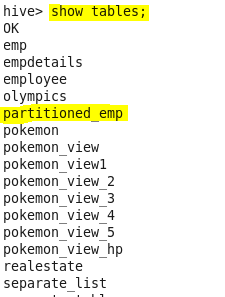
Now we insert the data according to the query



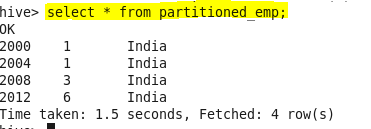
We can check our partition by the following command



**The table has been created as follows:**

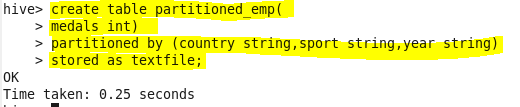


**Output:**

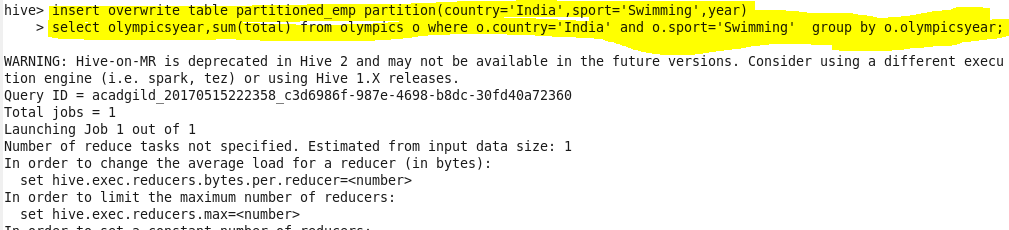


* **Find the number of medals india won in swimming year wise**

**Creating table:**



**Inserting data:**



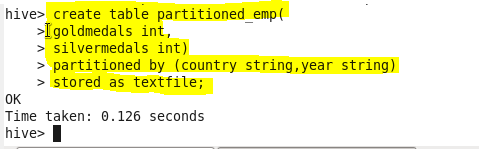
**Output:**



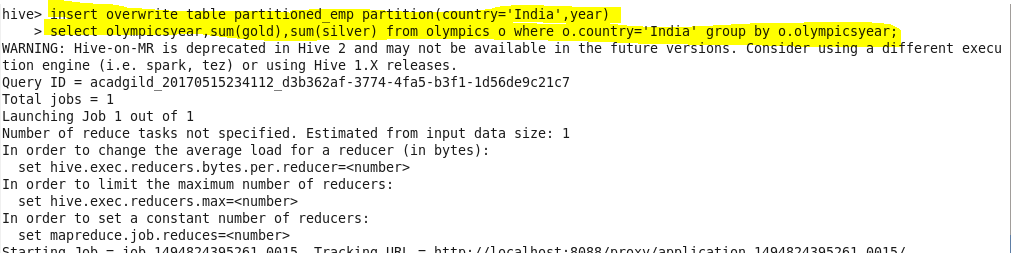
**This means there was no year when india won a medal in swimming**

* **Find the number of gold and silver medals india won year wise**

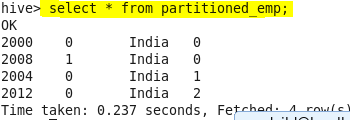
**Creating table:**



**Inserting data**



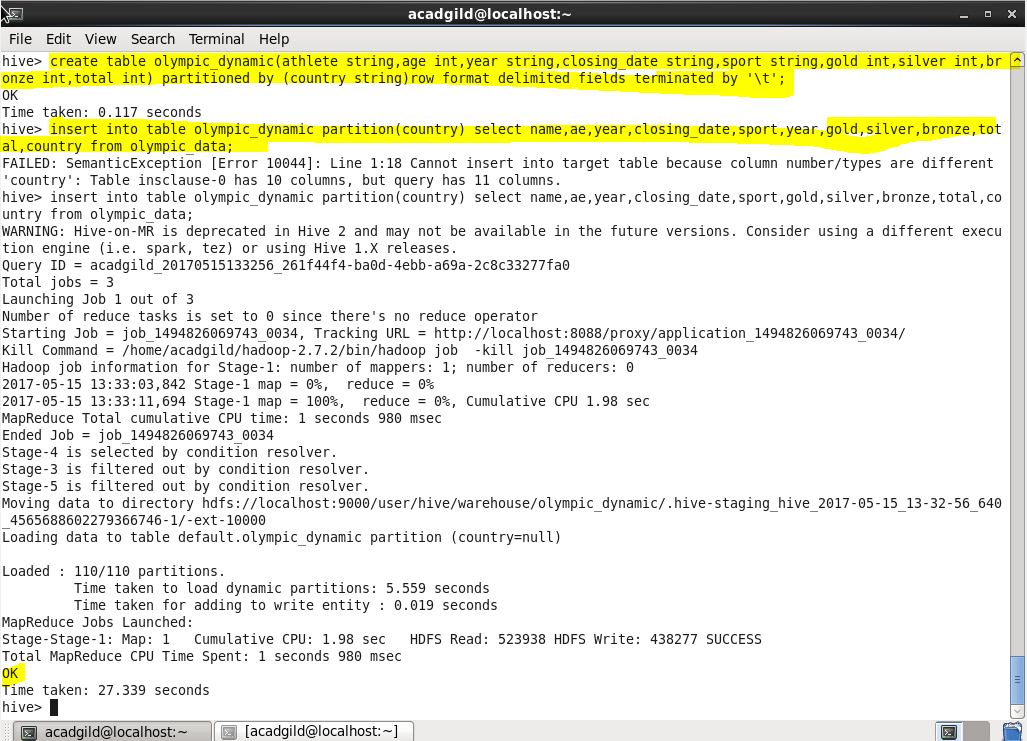
**Output:**



* **Use dynamic partitioning in hive and evaluate the below problem statements**

**Creating the dynamic partitioned table**

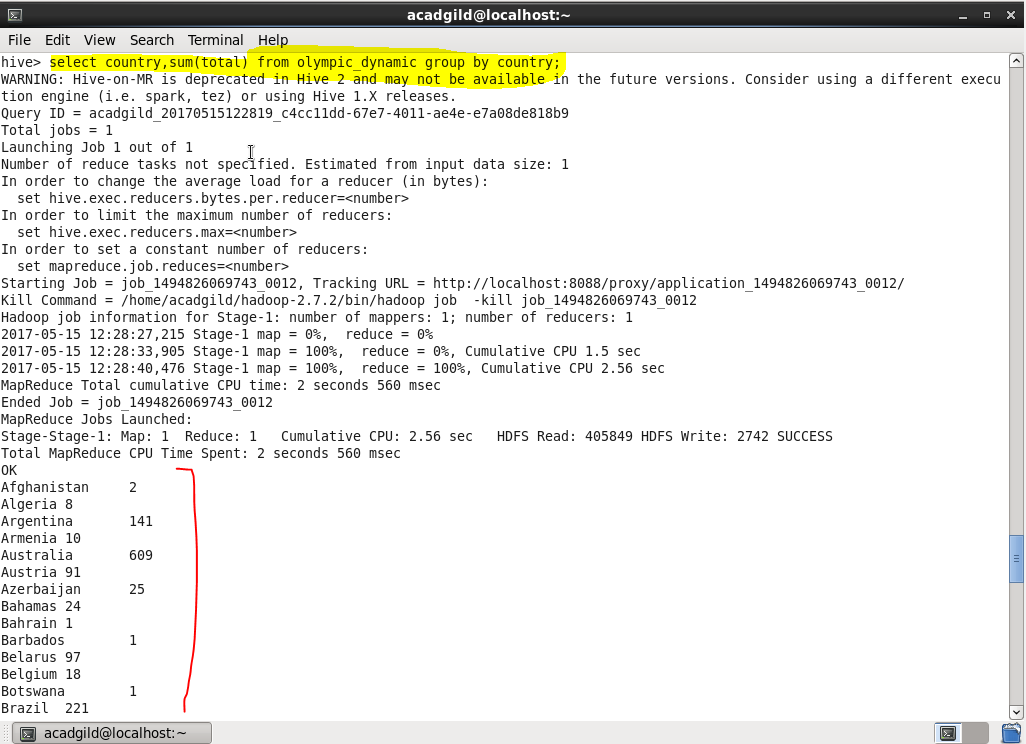
Creating and Loading data dynamically based on country:-

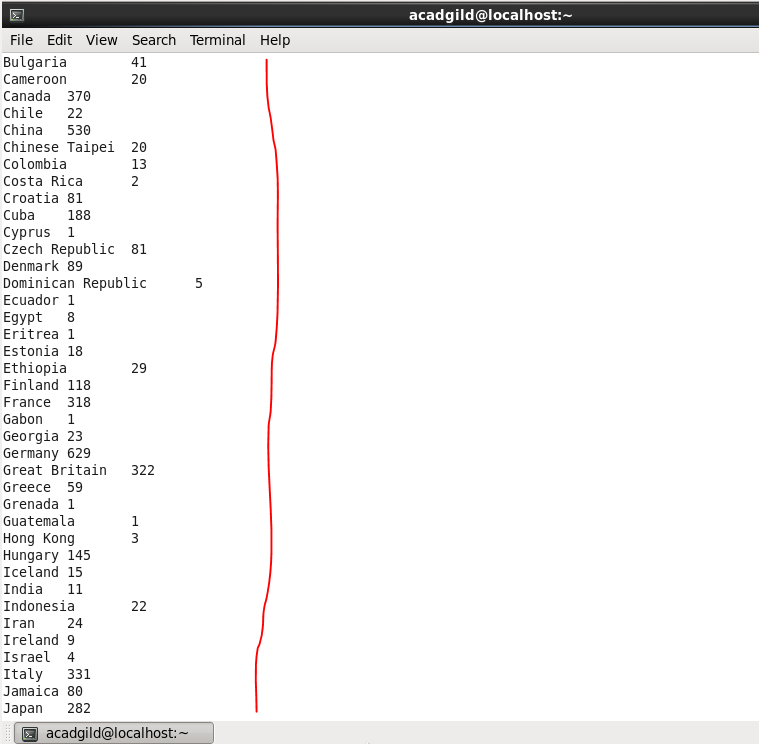


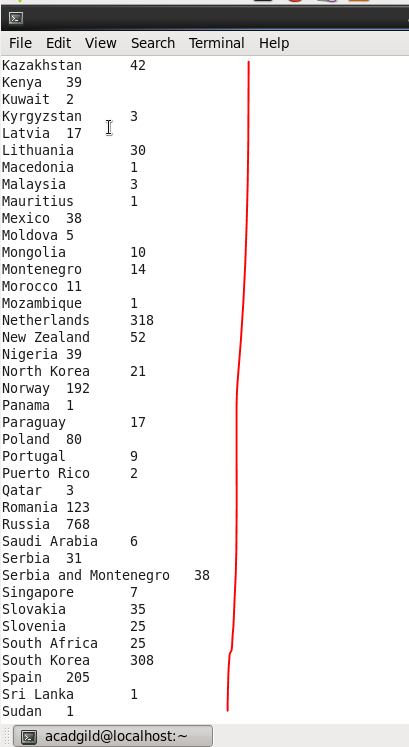
Now we have loaded the data into our dynamic partitioned table

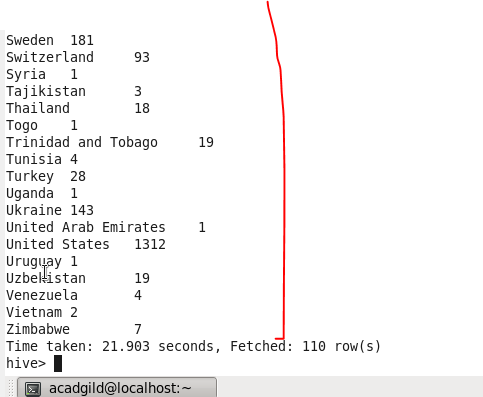
* **Find the total number of medals won by each country.**

Since we want total medals won by each country, group by country and the sum of total medals using sum() function.



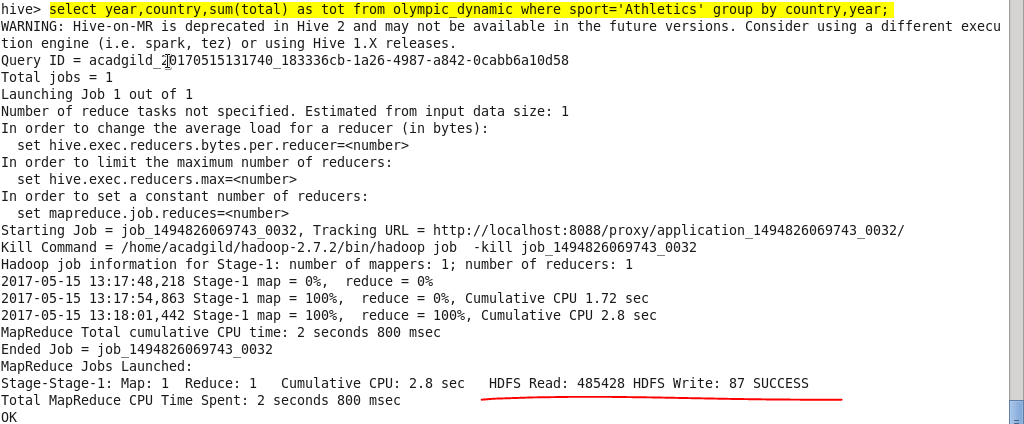




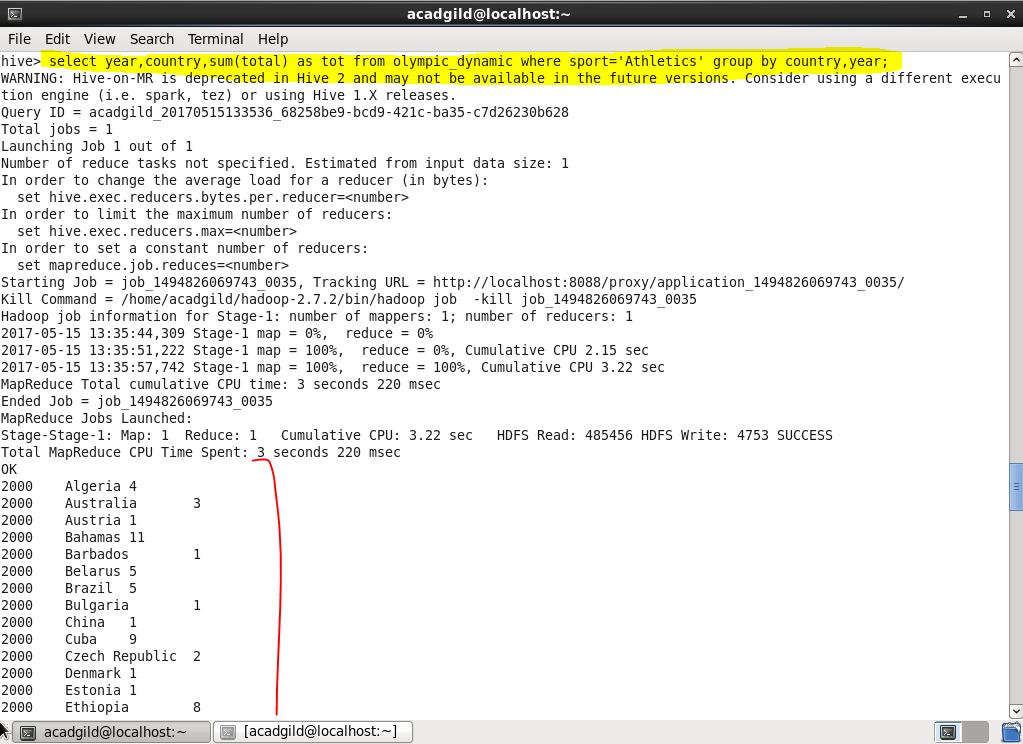


* **Find the number of medals each country won in Athletics year wise**

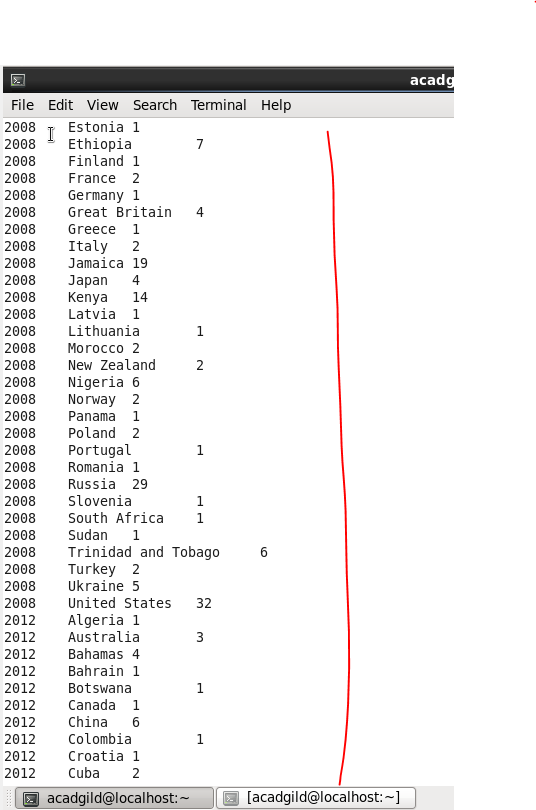
**Since we want medals won by athletics group by country and year and filter by sport=athletics**



**Output :**









* **Find the average age of athletes participated from each country in olympics year wise**

Since we want average age of athletes in Olympics for each country yearwise,

I grouped by country and year and used avg() on age for finding the average age of athletes

