**ASSIGNMENT 26.6**

**2. Problem Statement**

1. Explain the differences between static and dynamic partitioning in hive and their working procedures. Olympic data analysis

2. Use static partitioning in hive and evaluate the below problem statements - Find the number of medals india won year wise - Find the number of medals india won in swimming year wise - Find the number of gold and silver medals india won year wise

3. Use dynamic partitioning in hive and evaluate the below problem statements - Find the total number of medals won by each country. - Find the number ot medals each country won in Athletics year wise - Find the average age of atheltes participated from each country in olympics year wise You can downlaod the olympics data set from the below link https://drive.google.com/drive/folders/0ByJLBTmJojjzVGNsWmpUUUxTZDA

**Solution-**

1. **Partitioning-**

• Hive organizes tables horizontally into partitions.

• It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, department etc.

• Using partition, it is easy to query a portion of the data.

• Partitioning can be done based on more than column which will impose multi-dimensional structure on directory storage.

• In Hive, partitioning is supported for both managed and external tables.

The partition statement lets Hive alter the way it manages the underlying structures of the table’s data directory.

• In case of partitioned tables, subdirectories are created under the table’s data directory for each unique value of a partition column.

• When a partitioned table is queried with one or both partition columns in criteria or in the WHERE clause, what Hive effectively does is partition elimination by scanning only those data directories that are needed.

• If no partitioned columns are used, then all the directories are scanned (full table scan) and partitioning will not have any effect.

There are two types of partitioning-

1. Static Partitioning- Static partitioning needs to be applied when we know data (supposed to be inserted) belongs to which partition. Static Partition saves your time in loading data compared to dynamic partition You “statically” add a partition in table and move the file into the partition of the table.

We can alter the partition in static partition

You can get the partition column value form the filename, day of date etc without reading the whole big file. If you want to use Static partition in hive you should set property

set hive.mapred.mode = strict  
This property set by default in hive-site.xml Static partition is in Strict Mode You should use where clause to use limit in static partition You can perform Static partition on Hive Manage table or external table.

1. Dynamic Partitioning- In static partitioning, every partitioning needs to be backed with individual hive statement which is not feasible for large number of partitions as it will require writing of lot of hive statements. In that scenario dynamic partitioning is suggested as we can create as many number of partitions with single hive statement.

Usually dynamic partition load the data from non partitioned table

Dynamic Partition takes more time in loading data compared to static partition

When you have large data stored in a table then Dynamic partition is suitable.

If you want to partition number of column but you don’t know how many columns then also dynamic partition is suitable

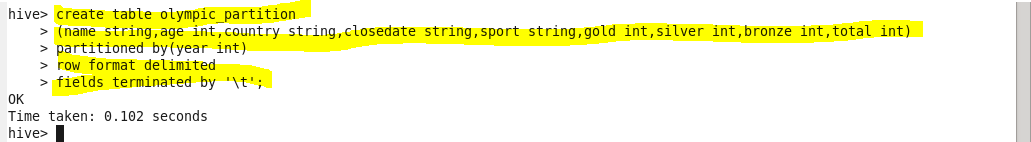
Dynamic partition there is no required where clause to use limit. we can’t perform alter on Dynamic partition

You can perform dynamic partition on hive external table and managed table If you want to use Dynamic partition in hive then mode is in nonstrict mode Here is hive dynamic partition properties you should allow

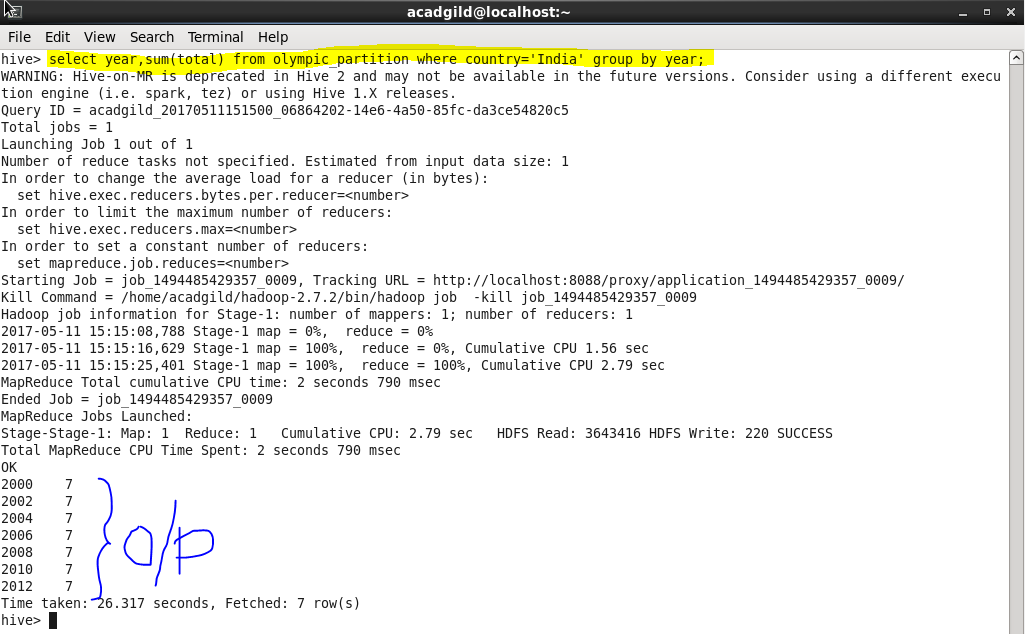
SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

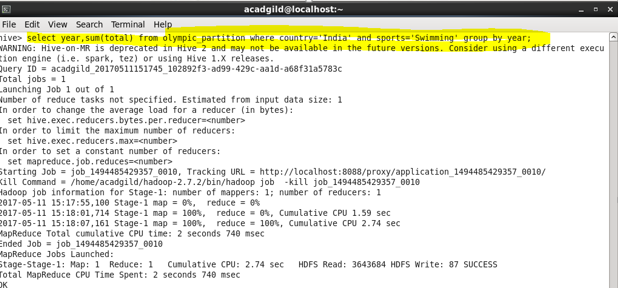
1. **Static Partitioning-**
2. **Find the number of medals India won year wise.**
3. Table created for static partitioning.
4. Table is partitioned by year.
5. It is a static partitioning, so data is loaded manually according to each year.
6. Select query to count total number of medals won by India.

****

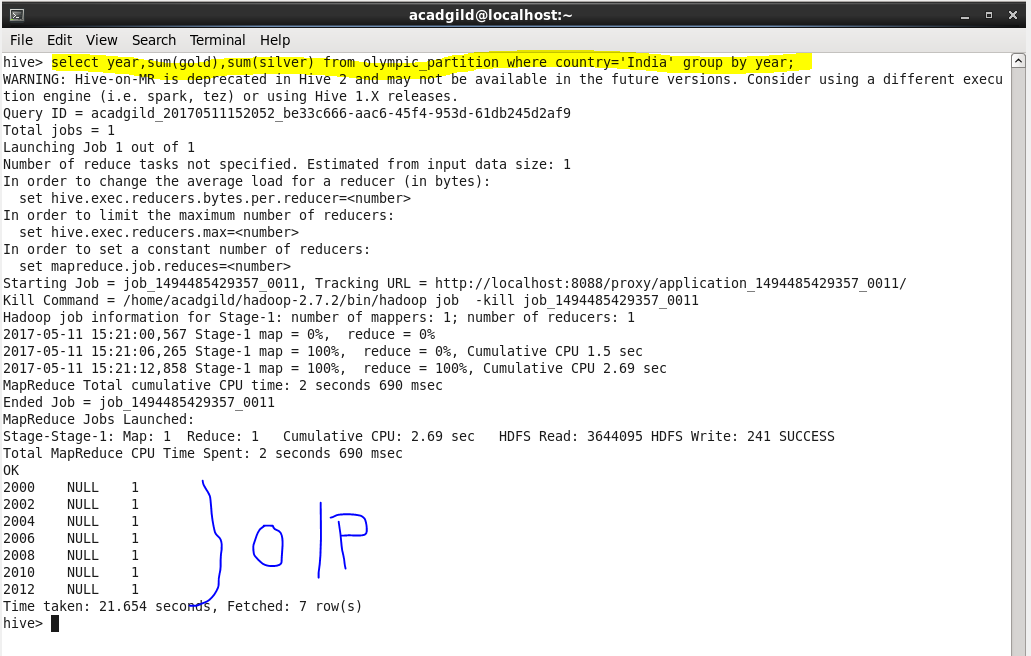
****

****

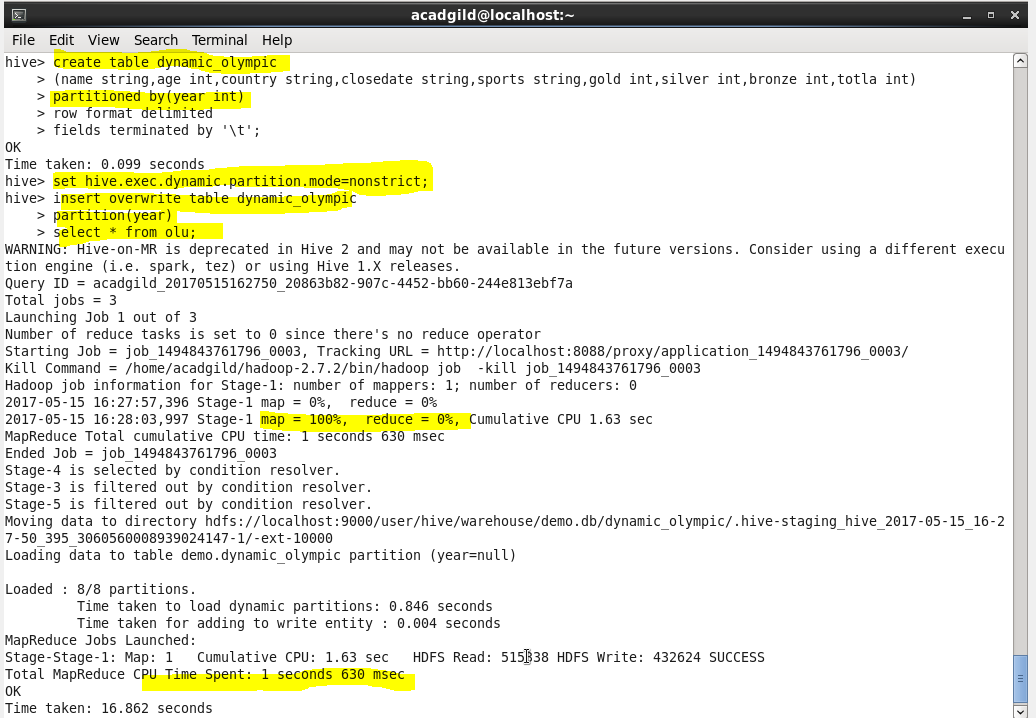
1. **Number of medals India won in Swimming year wise.**
2. Select query to find total number of medals India won in Swimming year wise.
3. It is a static extraction, so the query is fired on the previous created table.

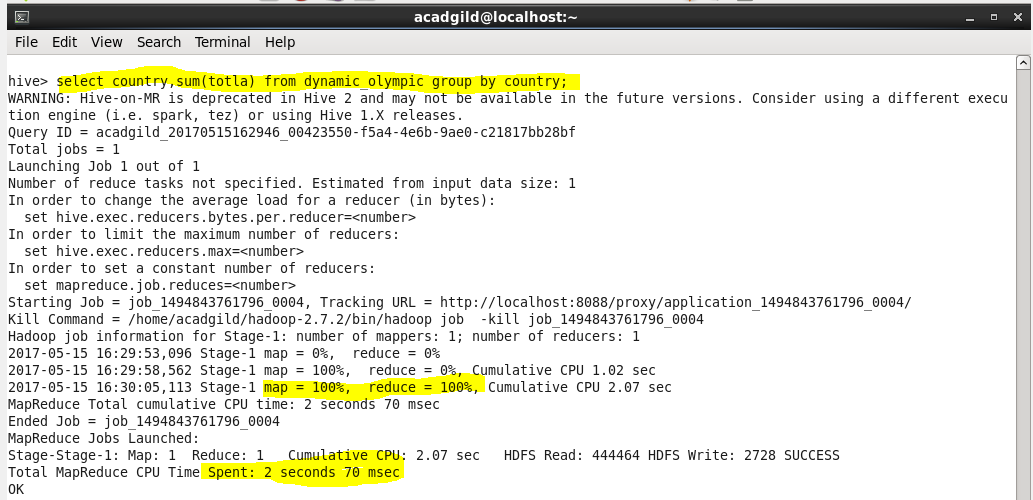
****

1. **Find the number of gold and silver medals India won year wise.**
2. Select query to find total number of gold and silver medals by India year wise.

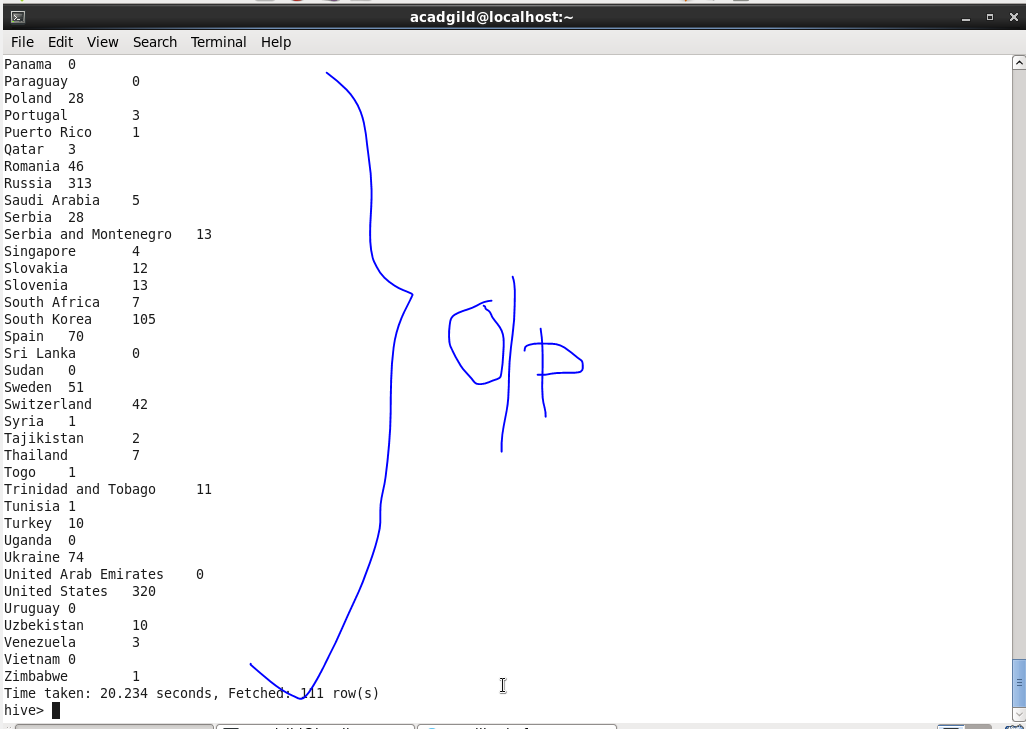
****

1. **Dynamic Partitioning-**
2. **Find the total number of medals won by each country.**
3. A table is created named dynamic\_olympic, which is partitioned by year.
4. It is a dynamic partition on the field year.
5. Hive property is set for dynamic partitioning.
6. Loaded data in the table.
7. Select query to find the total number of medals won by each country.

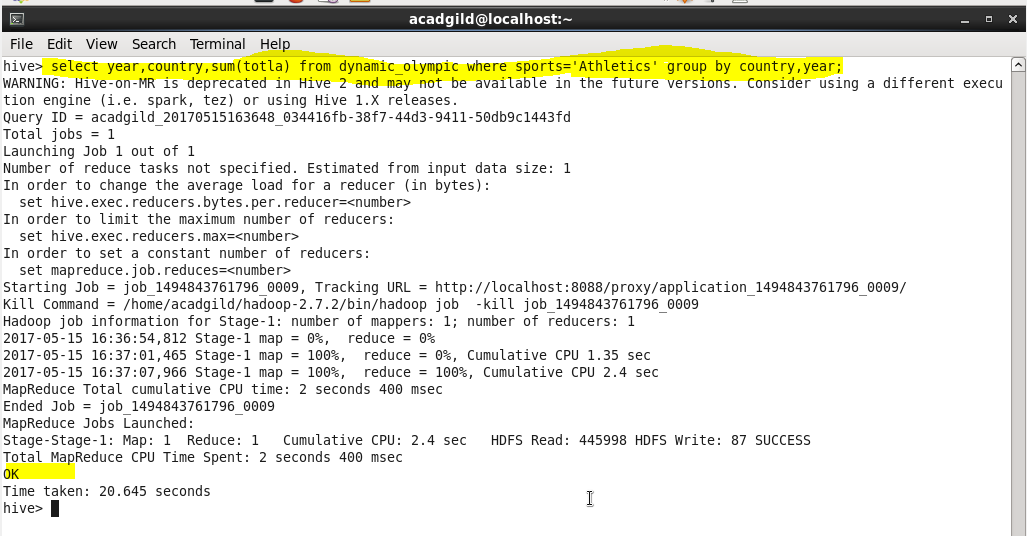
****

****

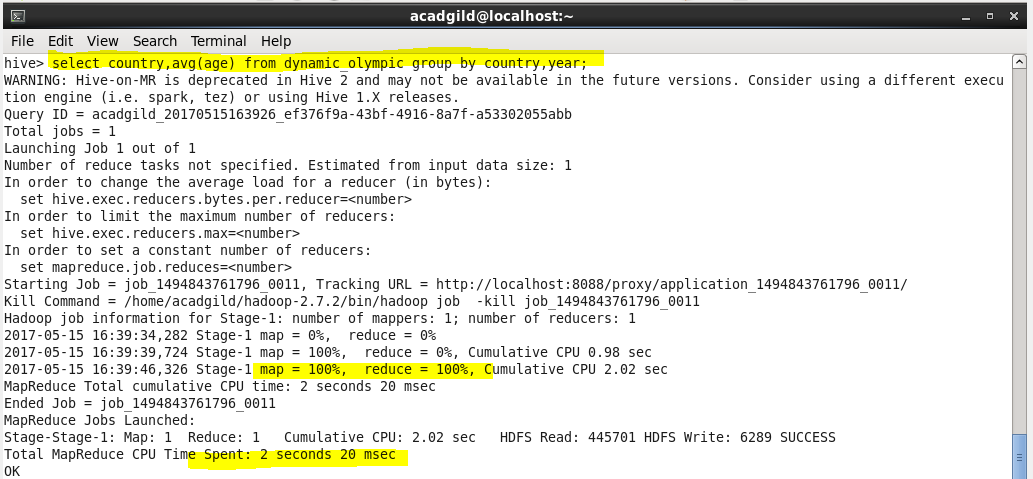
**Output-**

****

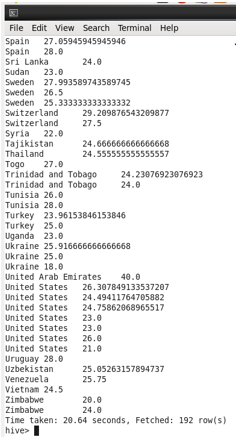
1. **Find the number of medals each country won in athletics year wise.**
2. Select query on the created table to count total number of medals where sports is athletics for each country.

****

1. **Find the average age of all the athletes participated country wise.**
2. Select query to find the average age of all the participants of each country year wise.

****

**Output-**

****