**ASSIGNMENT - 26.6**

**Problem Statement:**

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

**Olympic data analysis:**

1. 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.

1. Use dynamic partitioning in hive and evaluate the below problem statements:

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

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

- Find the average age of athletes participated from each country in Olympics year wise.

**Input:**

You can download the Olympics data set from the below link:

<https://drive.google.com/drive/folders/0ByJLBTmJojjzVGNsWmpUUUxTZDA>

**Data Set Description:**

The data set consists of the following fields:

Athlete: Name of the athlete

Age: Age of the athlete

Country: The name of the country participating in Olympics

Year: The year in which Olympics is conducted

Closing Date: Closing date of Olympics

Sport: Sports name

Gold Medals: No. of gold medals

Silver Medals: No. of silver medals

Bronze Medals: No. of bronze medals

Total Medals: Total no. of medals

**Solution:**

**Problem Statement 1:**

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

**Answer:**

**Difference between Static and Dynamic Partitioning:**

In Hive There is two types of Partitions:

1. Static Partition in Hive
2. Dynamic Partition in Hive

* **Static Partition in Hive:**

Insert input data files individually into a partition table is Static Partition

Usually when loading files (big files) into Hive tables static partitions are preferred

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.

Usually when loading files (big files) into Hive tables static partitions are preferred. That 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. Since the files are big they are usually generated in HDFS.

You can get the partition column value form the filename, day of date etc. without reading the whole big file.

**For example,**

CREATE TABLE cityreport(cityid string,creport string, ctover string)

partitioned BY (city string)

row format delimited

FIELDS terminated BY ‘|’

stored AS textfile;

**Load Data in table by using Static Partition:**

hive> LOAD DATA LOCAL inpath ‘/home/mahesh/hive-related/hyderabad.log’ INTO TABLE cityreport partition (city = ‘hyderabad’);

* **Dynamic Partition in Hive:**

Single insert to partition table is known as dynamic partition

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 no strict mode

Here is hive dynamic partition properties you should allow.

In case of dynamic partition whole big file i.e. every row of the data is read and data is partitioned through a MR Job into the destination tables depending on certain field in file. So usually dynamic partition are useful when you are doing sort of an ETL flow in your data pipeline.

E.g. you load a huge file through a move command into a Table X. then you run an inert query into a Table Y and partition data based on field in table X say day, country. You may want to further run an ETL step to partition the data in country partition in Table Y into a Table Z where data is partitioned based on cities for a particular country only.

**For example,**

SET hive.exec.dynamic.partition = true;

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

**Load data into Table by using Dynamic Partition:**

hive> INSERT INTO TABLE t2 PARTITION(country) SELECT \* FROM T1;

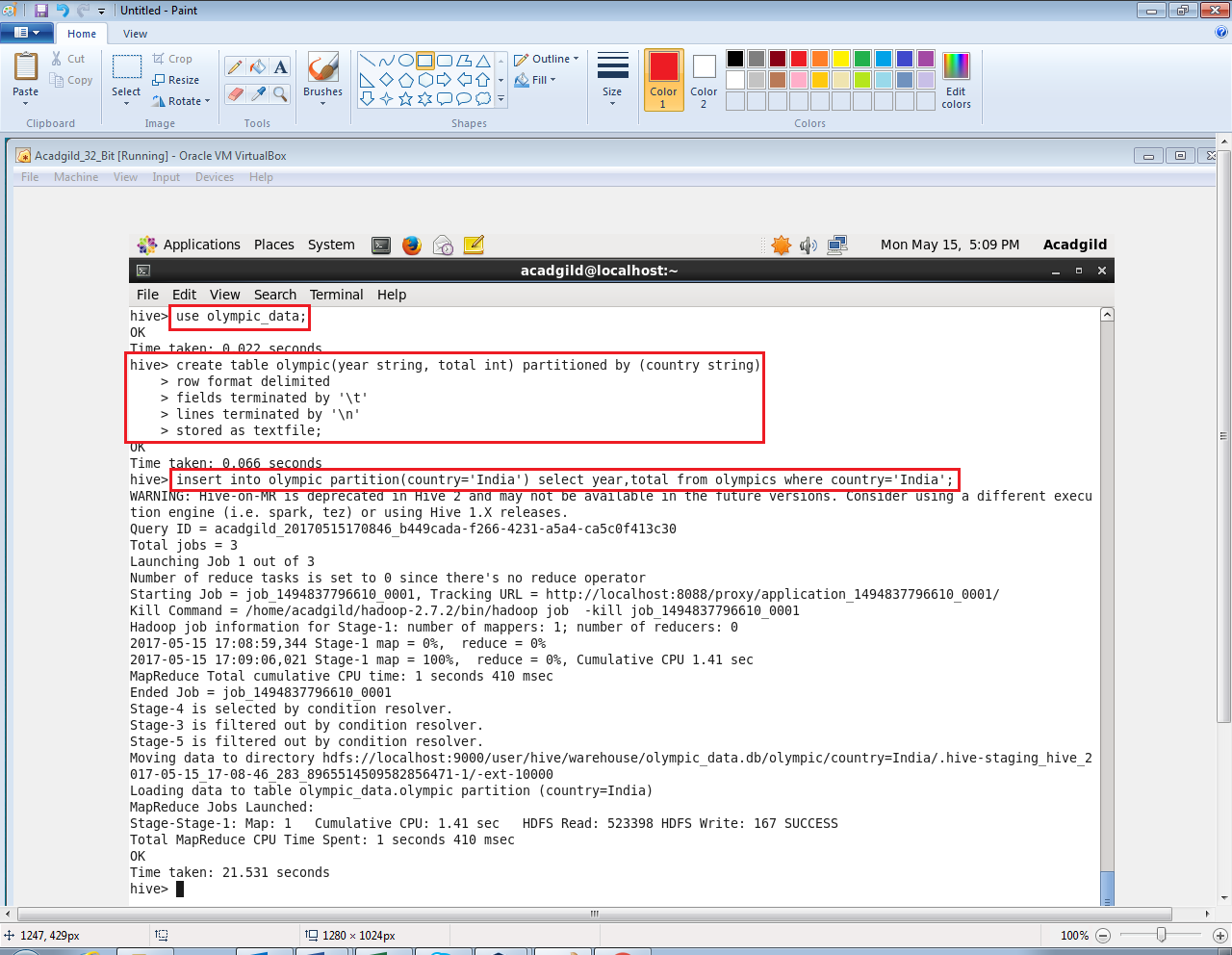
**Problem Statement 2:**

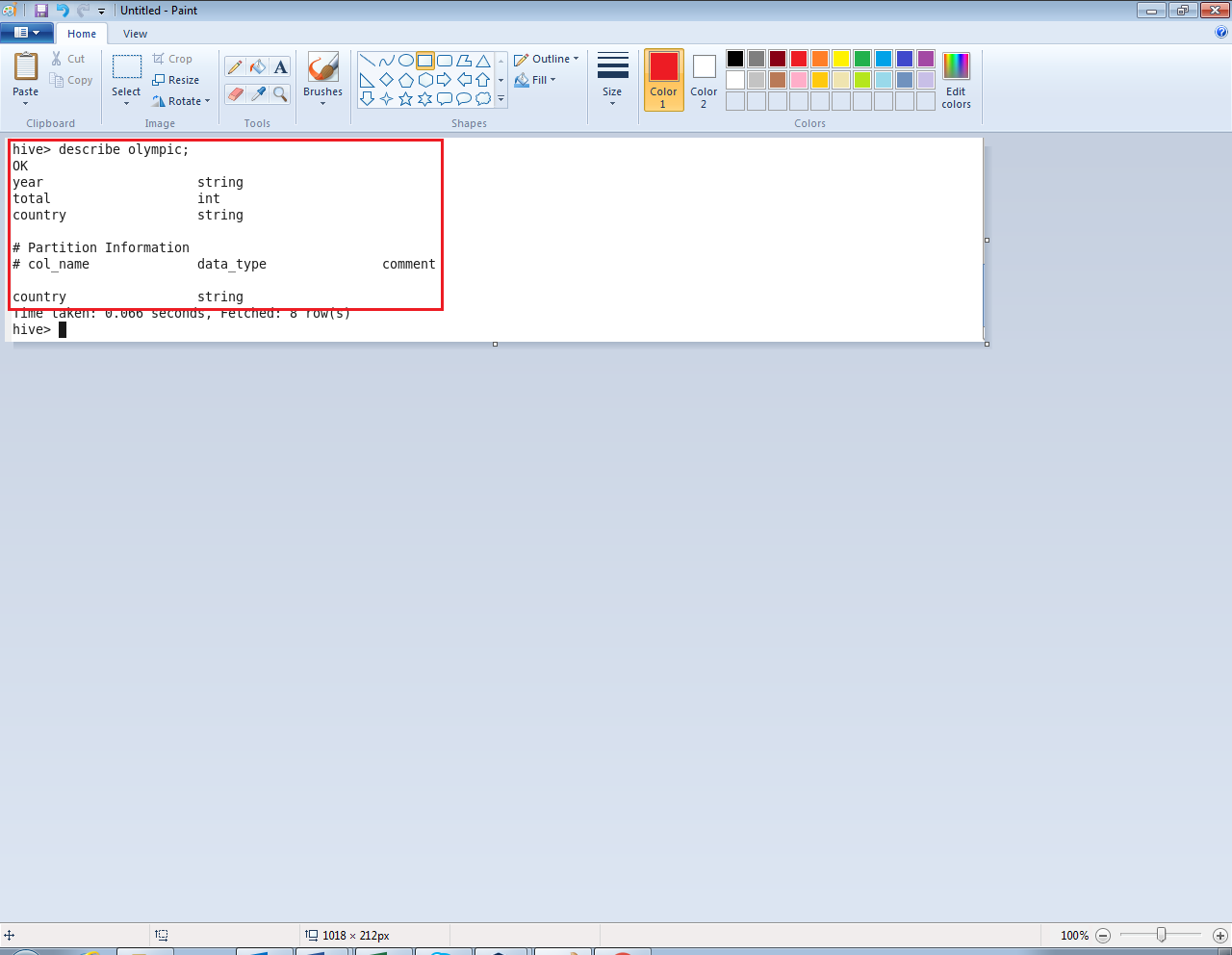
Use static partitioning in hive and evaluate the below problem statements:

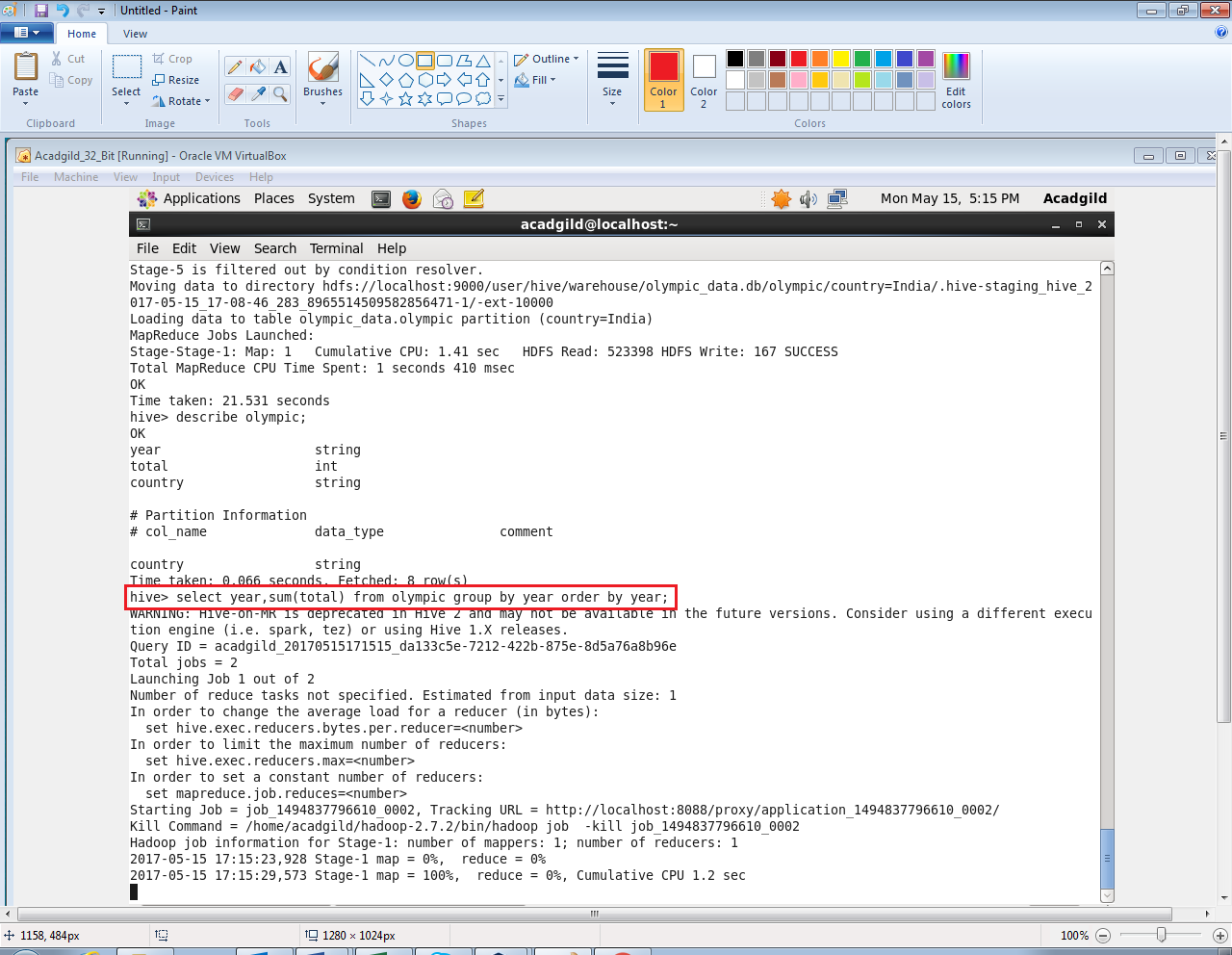
* + 1. Find the number of medals India won year wise.
    2. Find the number of medals India won in swimming year wise.
    3. Find the number of gold and silver medals India won year wise.

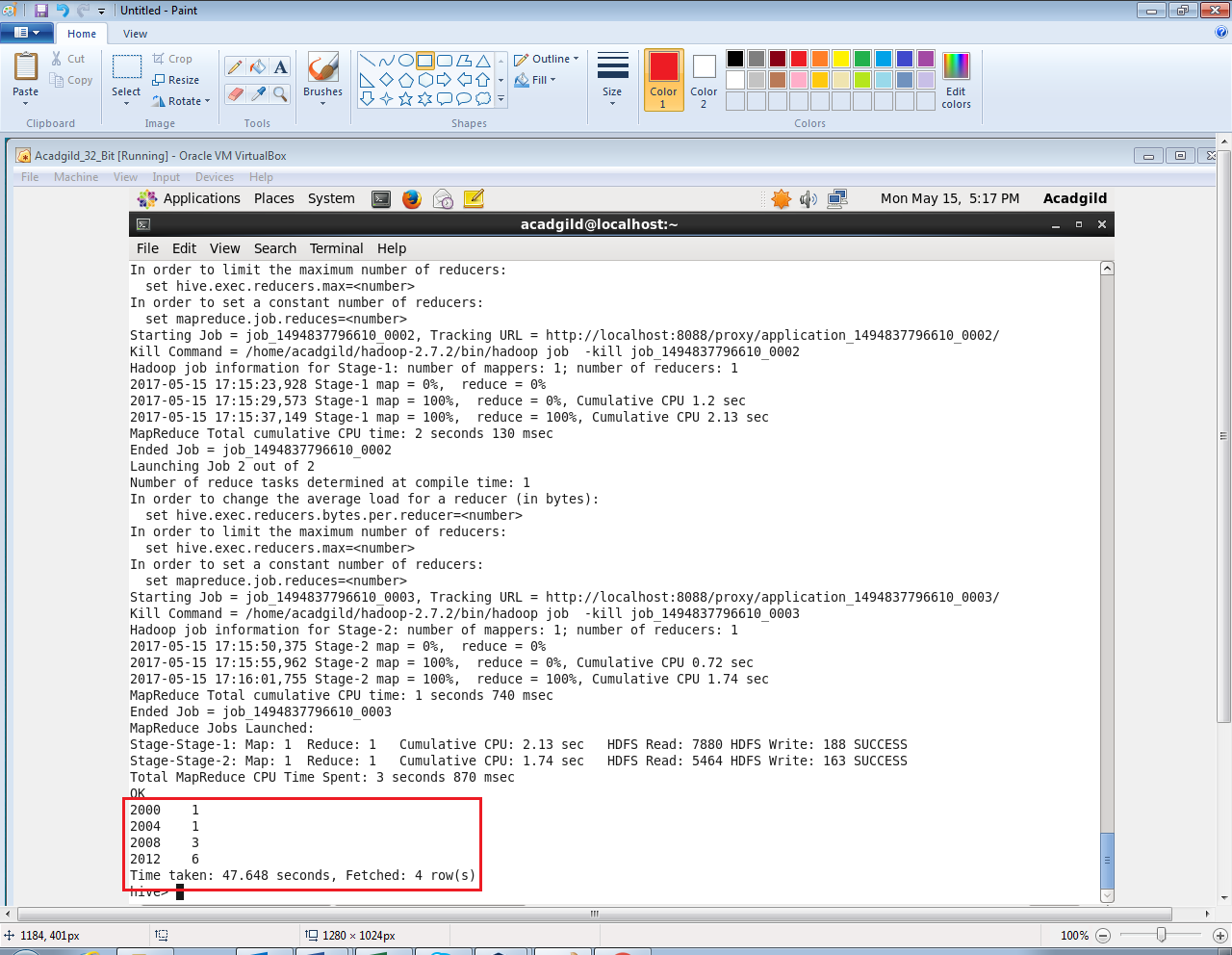
1. **Find the number of medals India won year wise:**

**Output:**



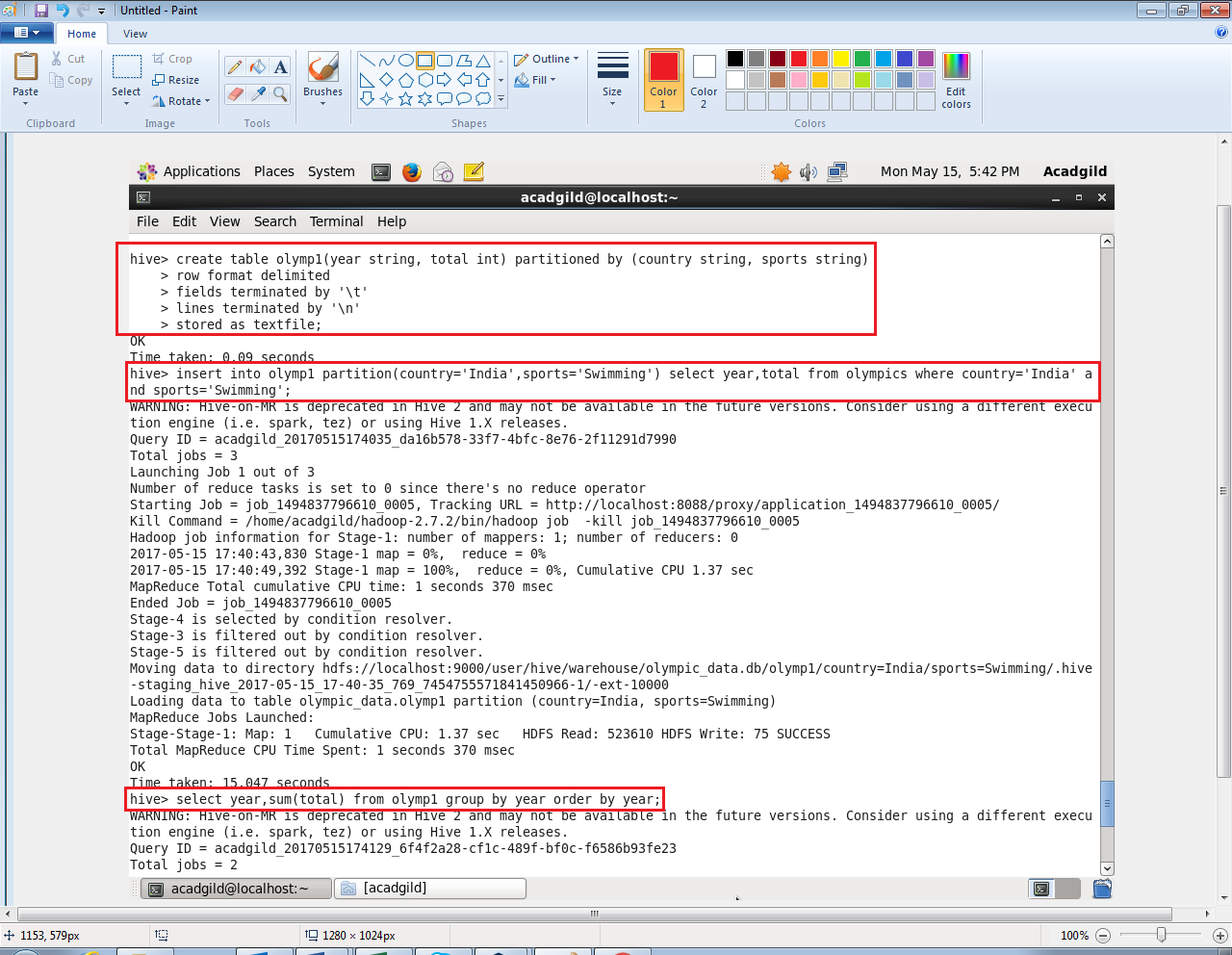


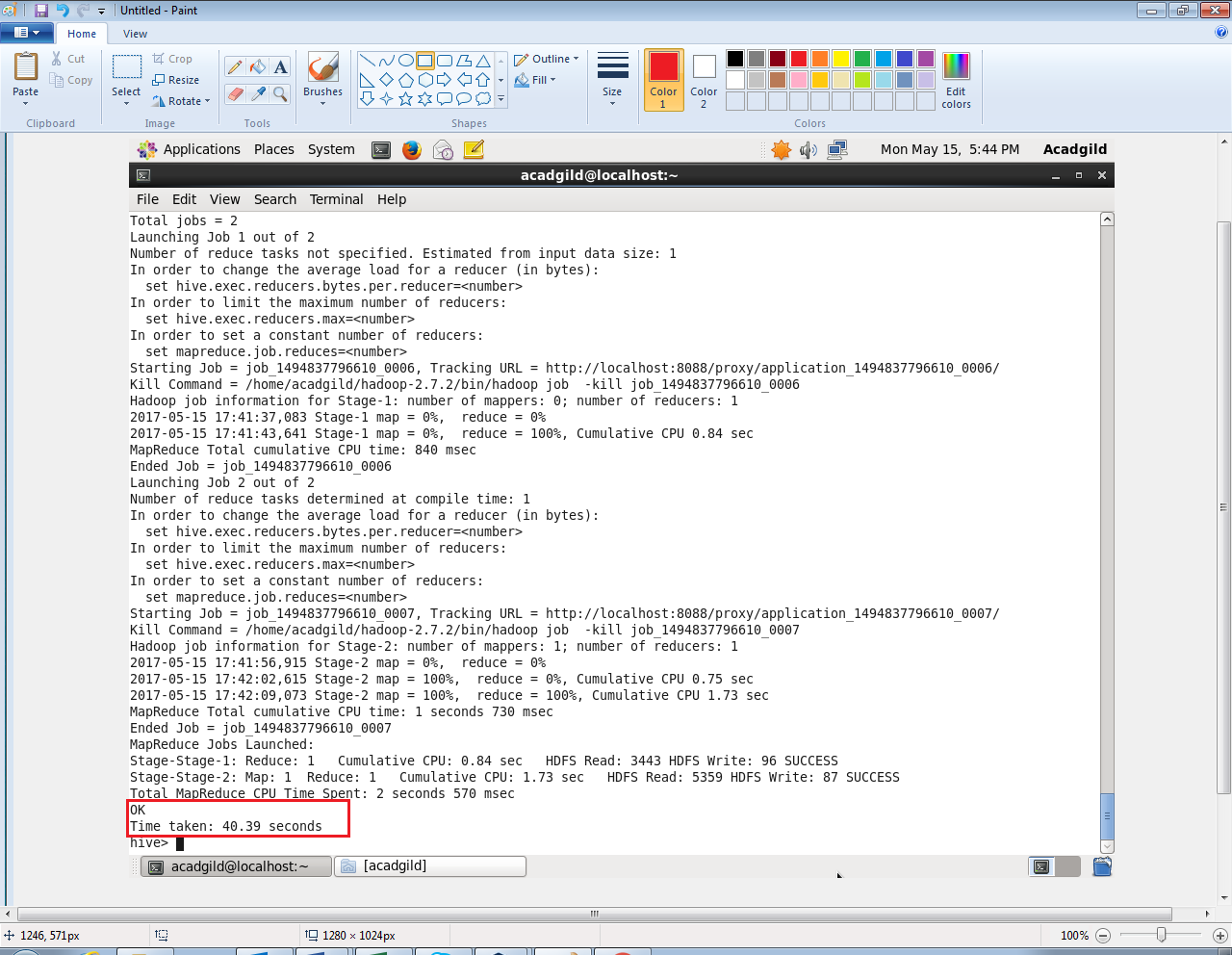




1. **Find the number of medals India won in swimming year wise:**

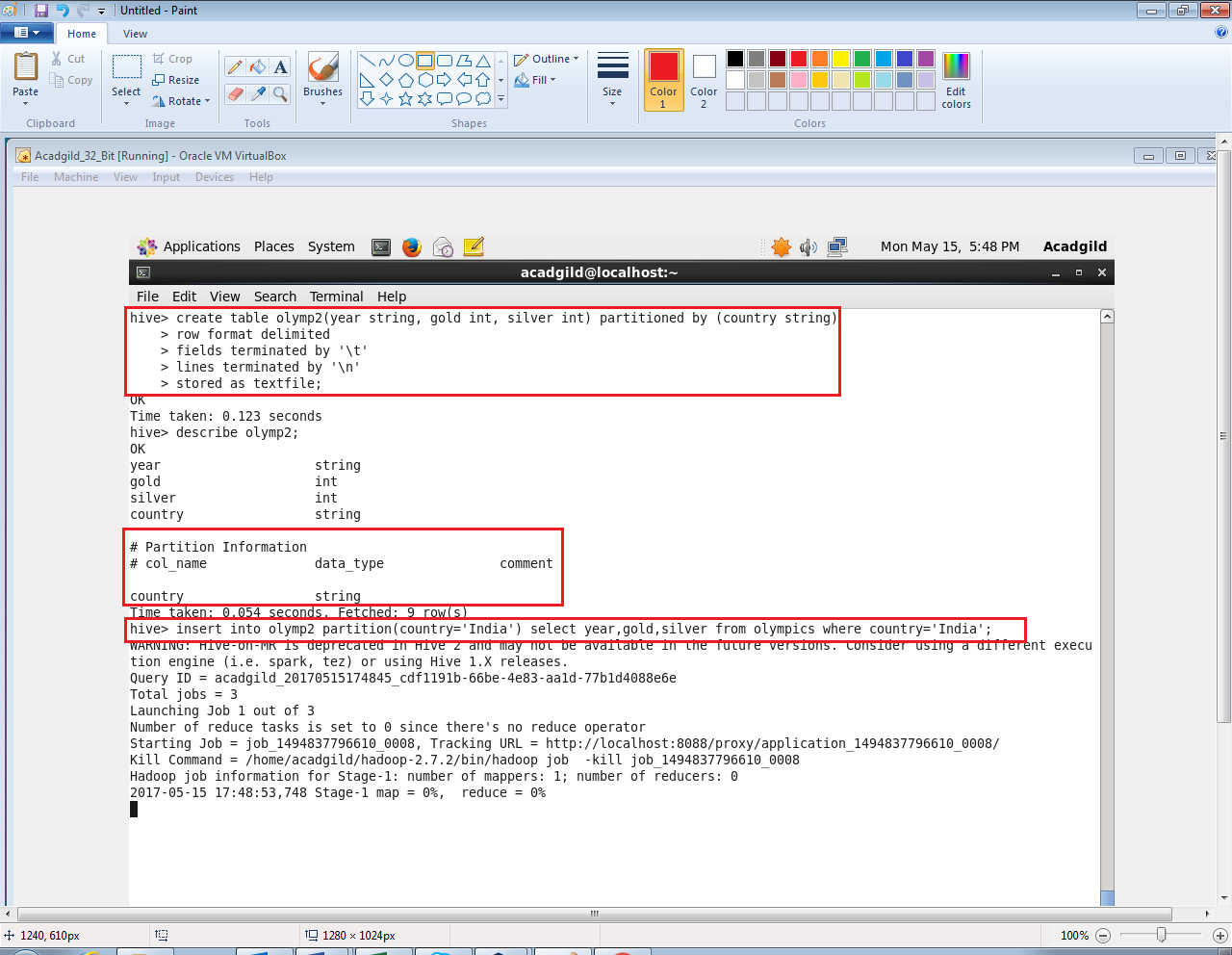
**Output:**

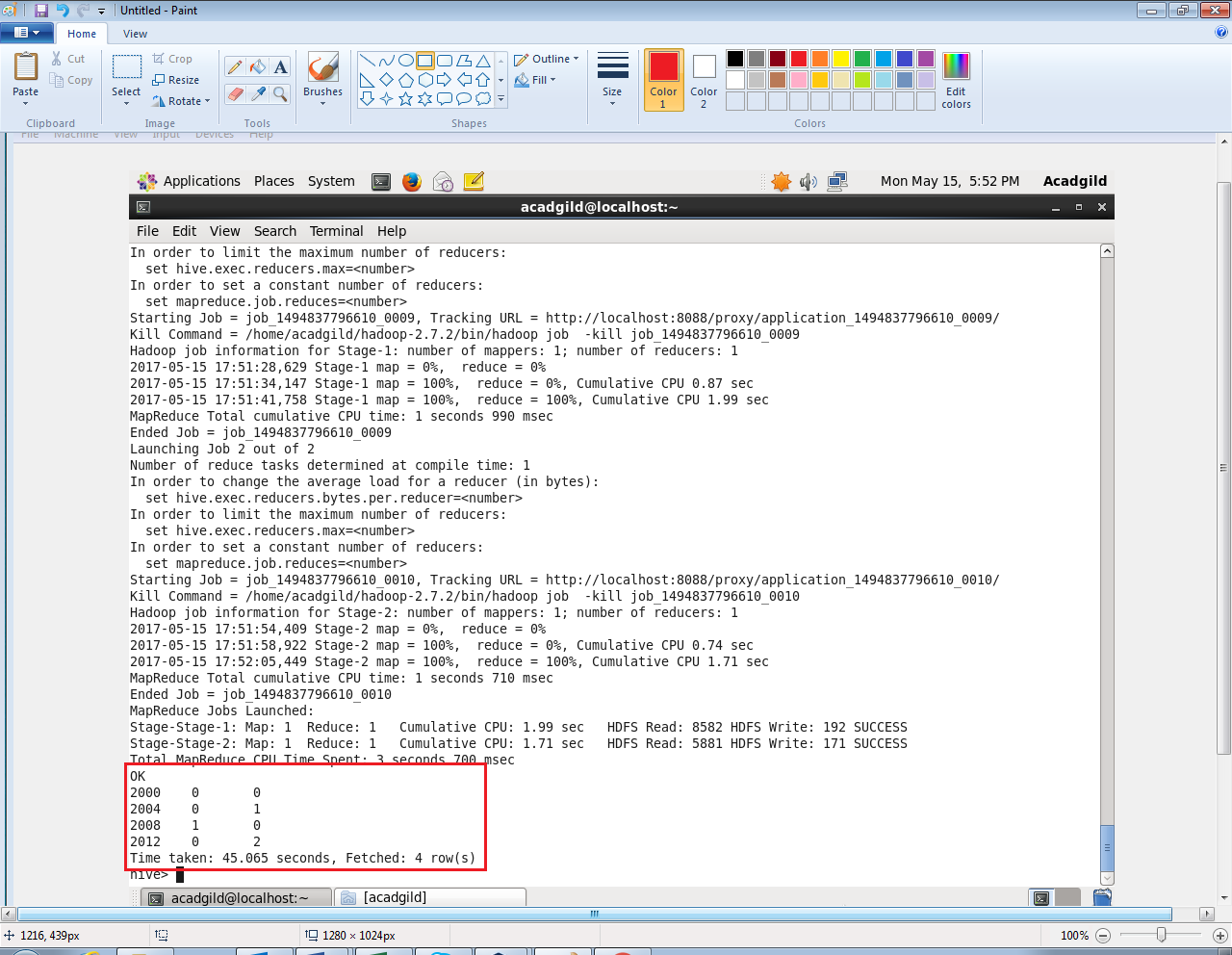
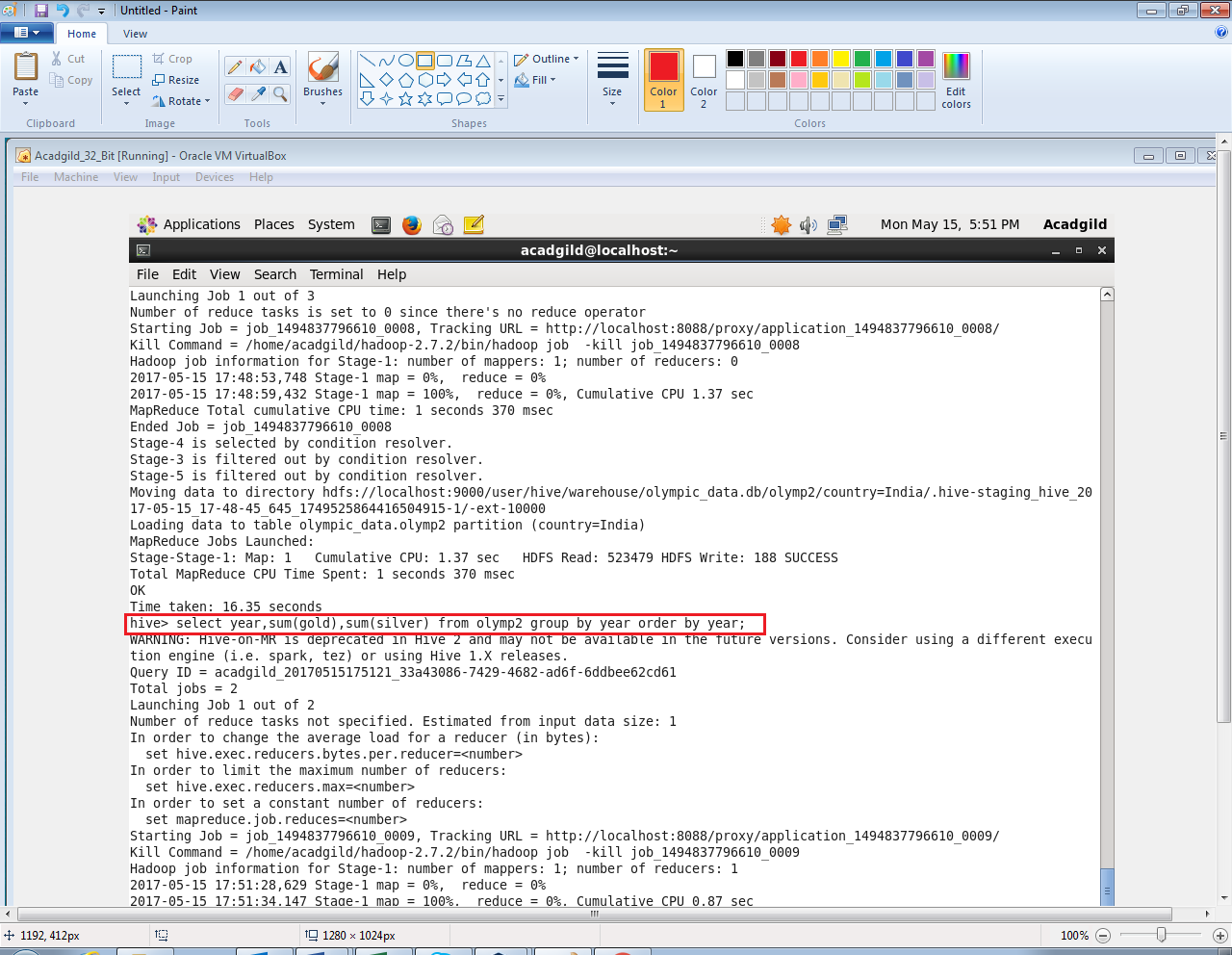




1. **Find the number of gold and silver medals India won year wise.**

**Output:**



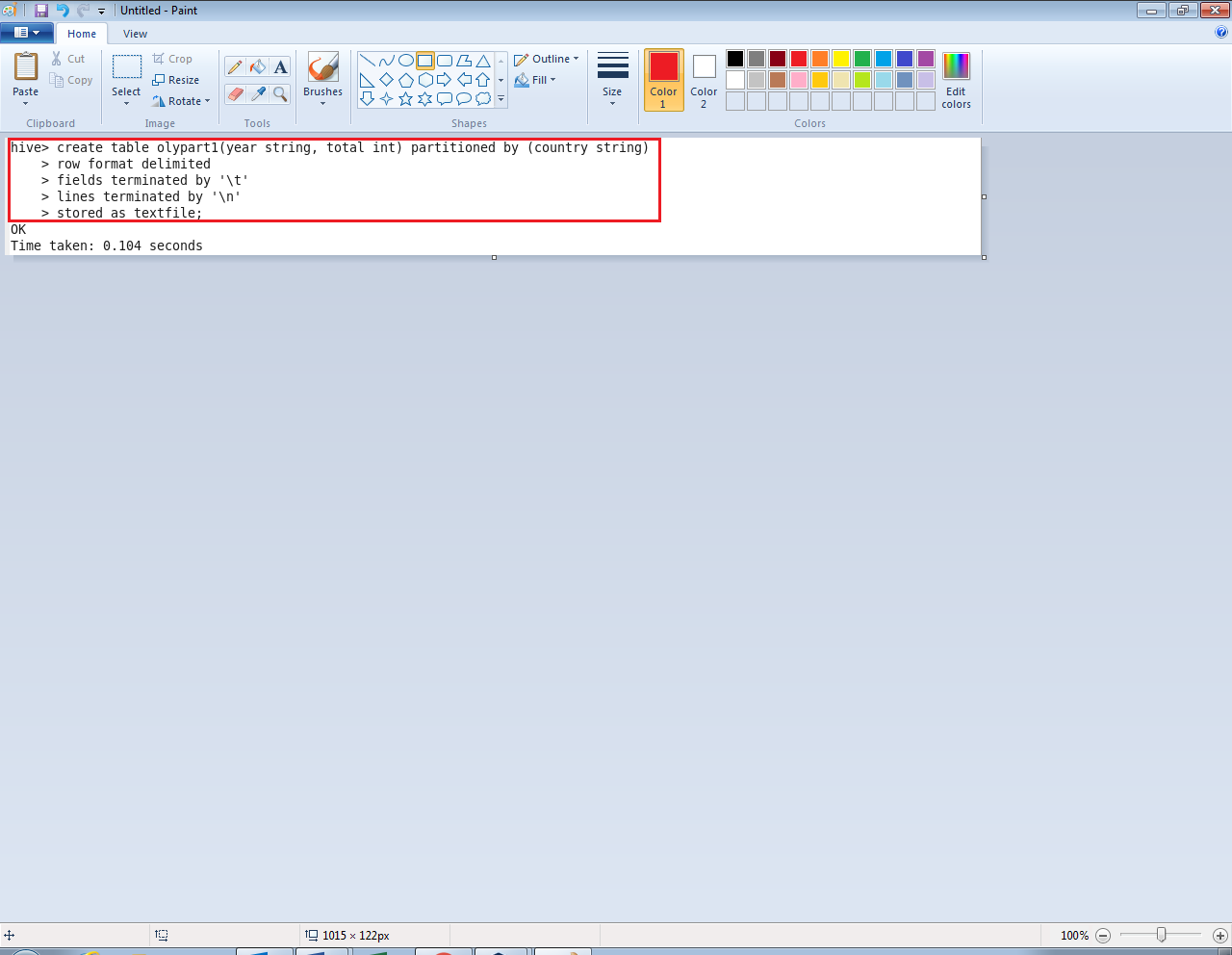


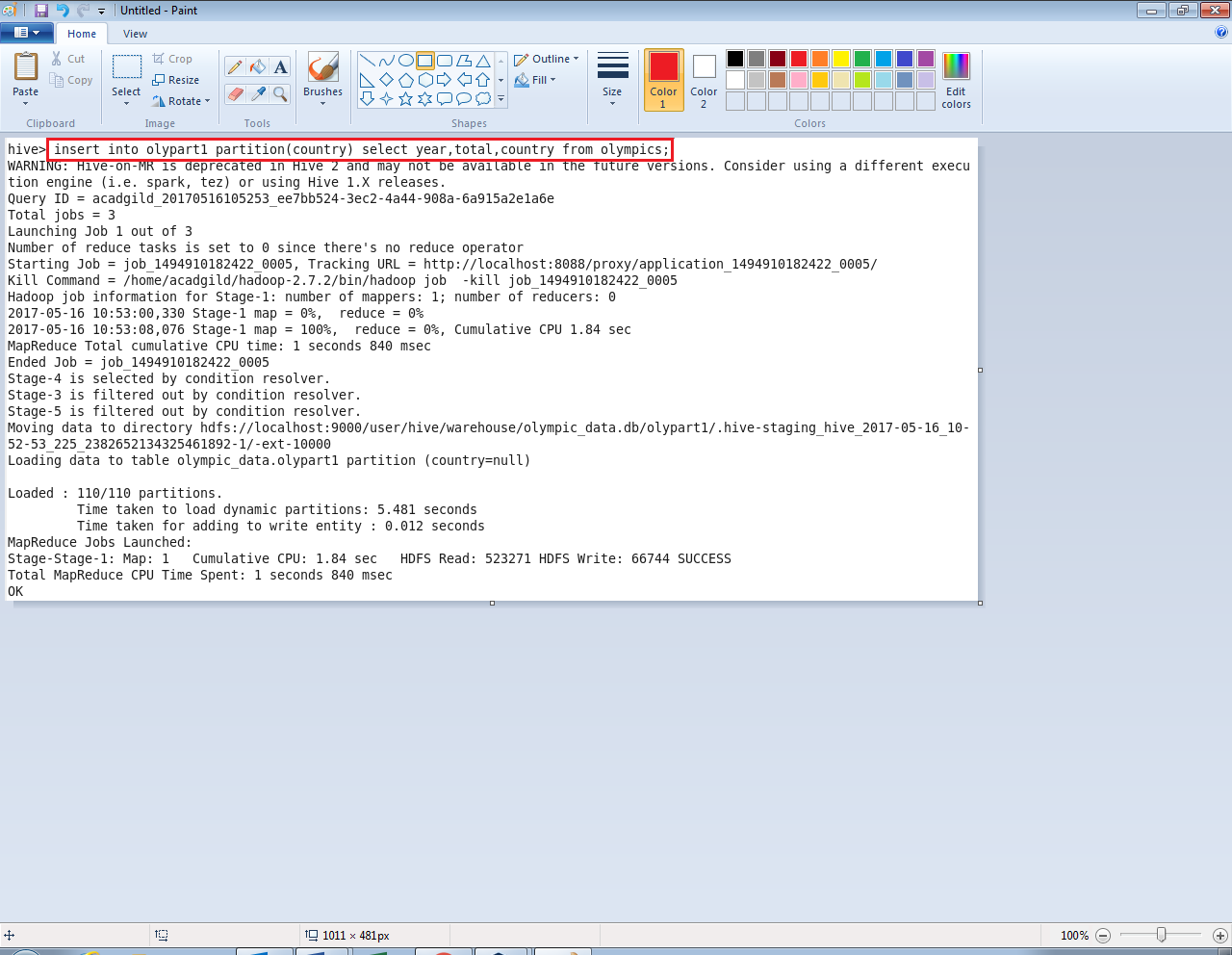
**Problem Statement 3:**

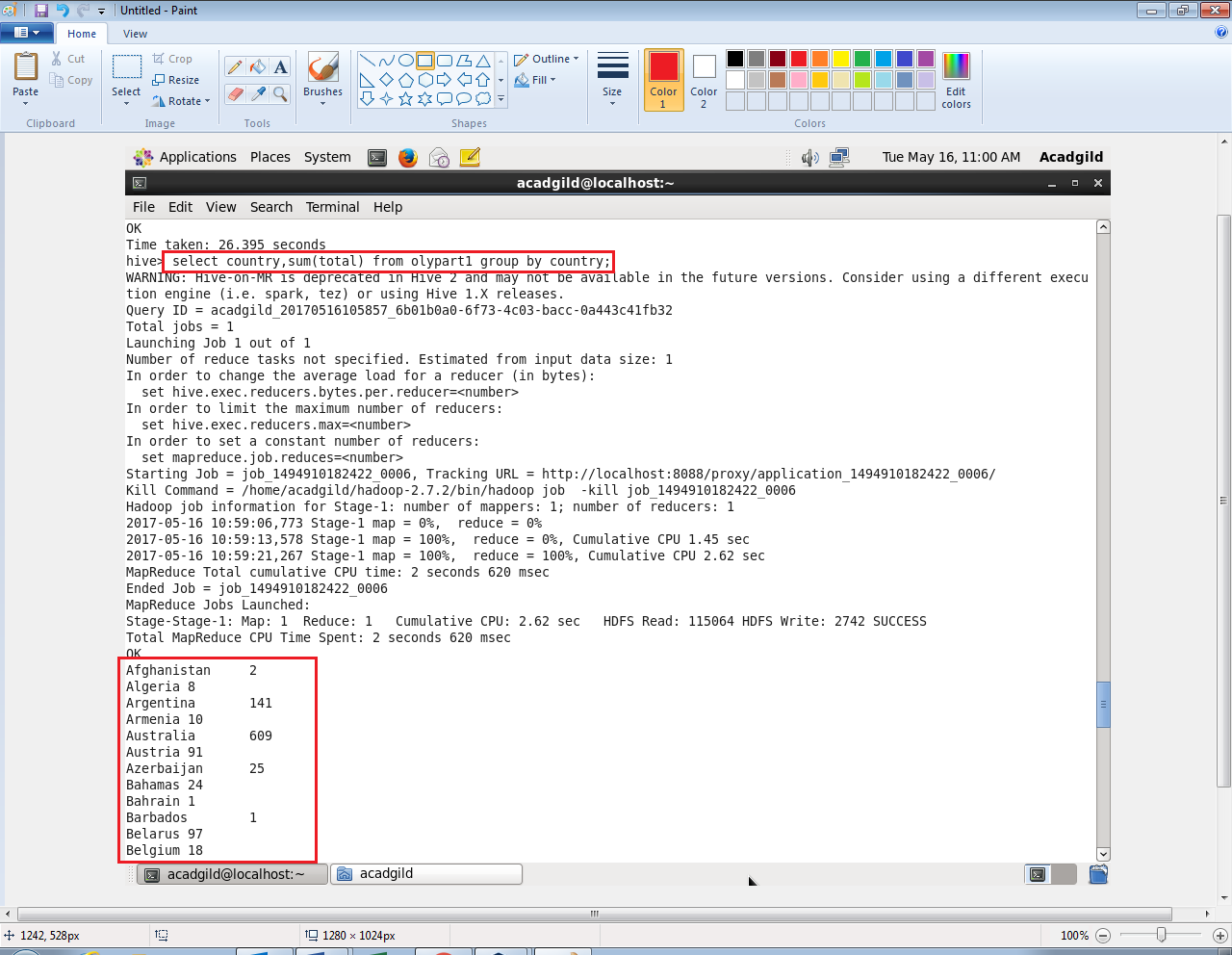
Use dynamic partitioning in hive and evaluate the below problem statements:

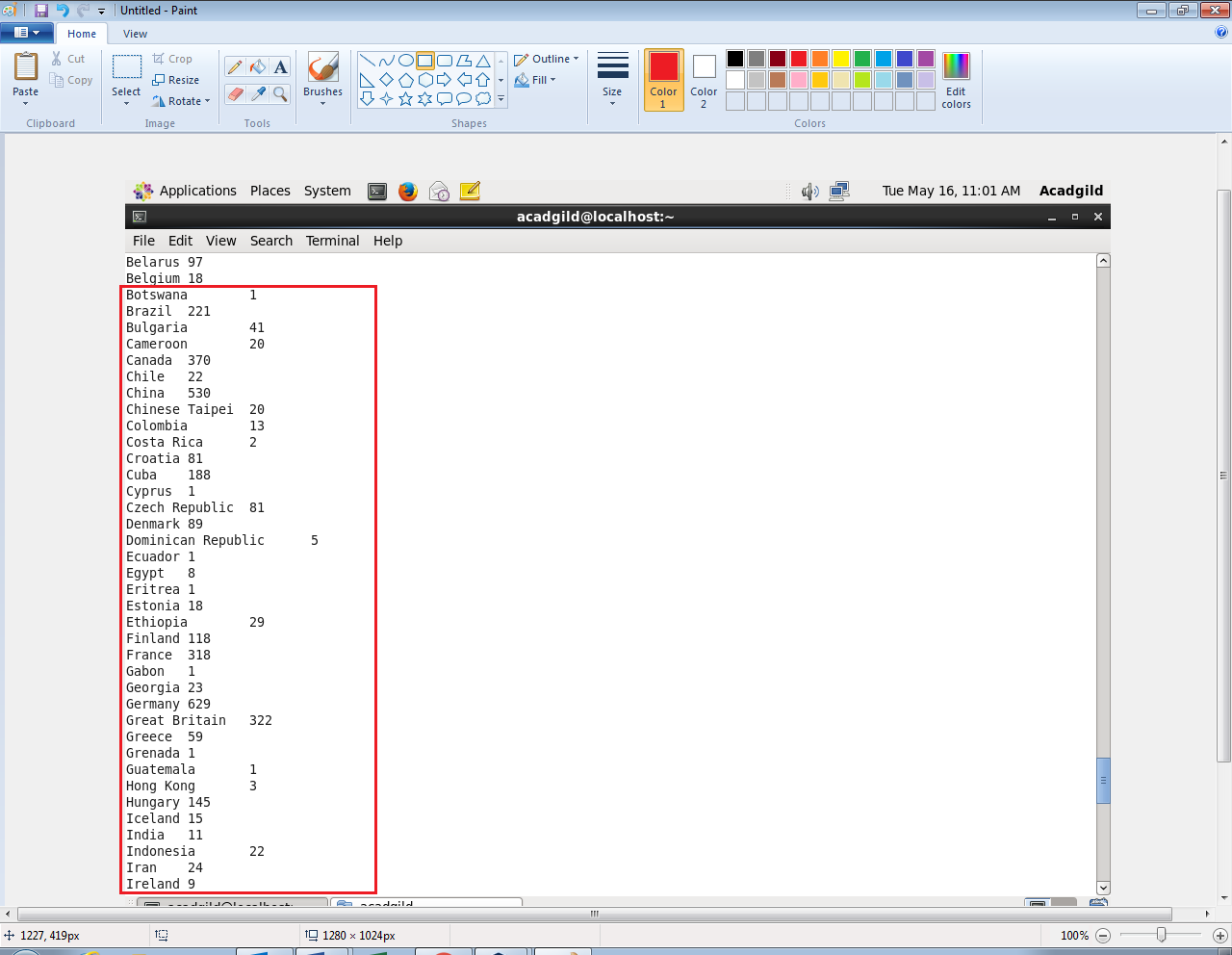
* 1. Find the total number of medals won by each country.
  2. Find the number of medals each country won in Athletics year wise.
  3. Find the average age of athletes participated from each country in Olympics year wise.
* **Find the total number of medals won by each country:**

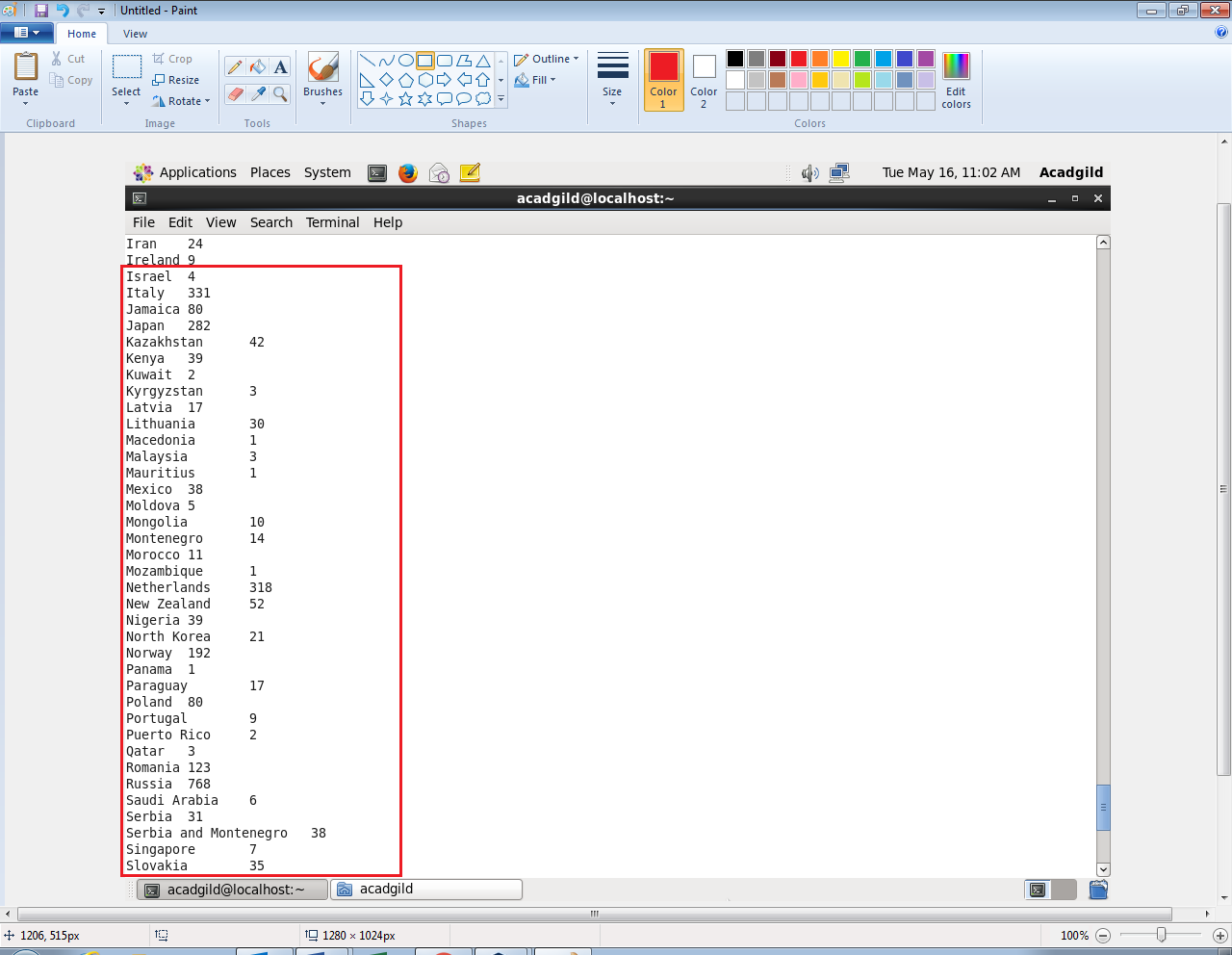
**Output:**

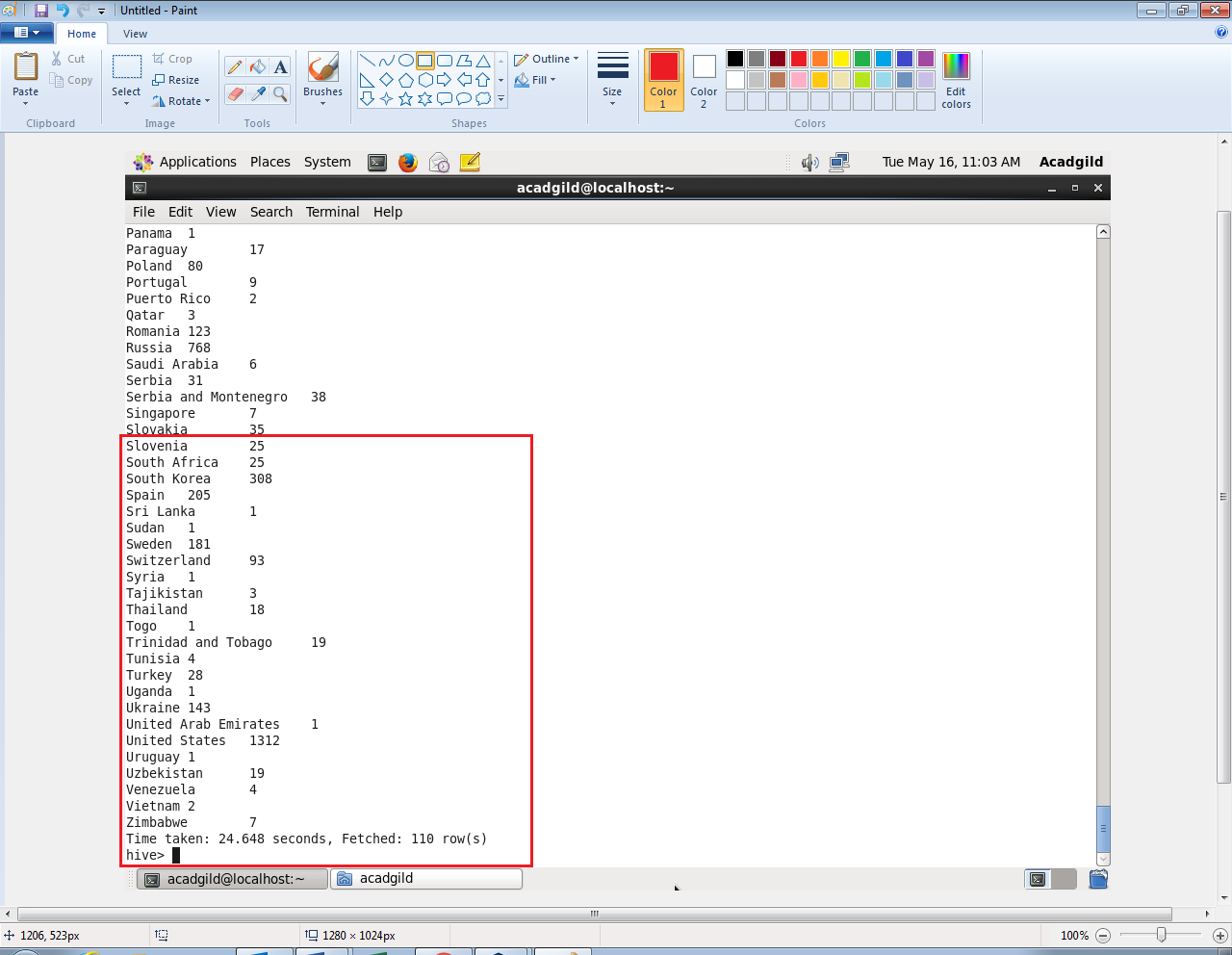






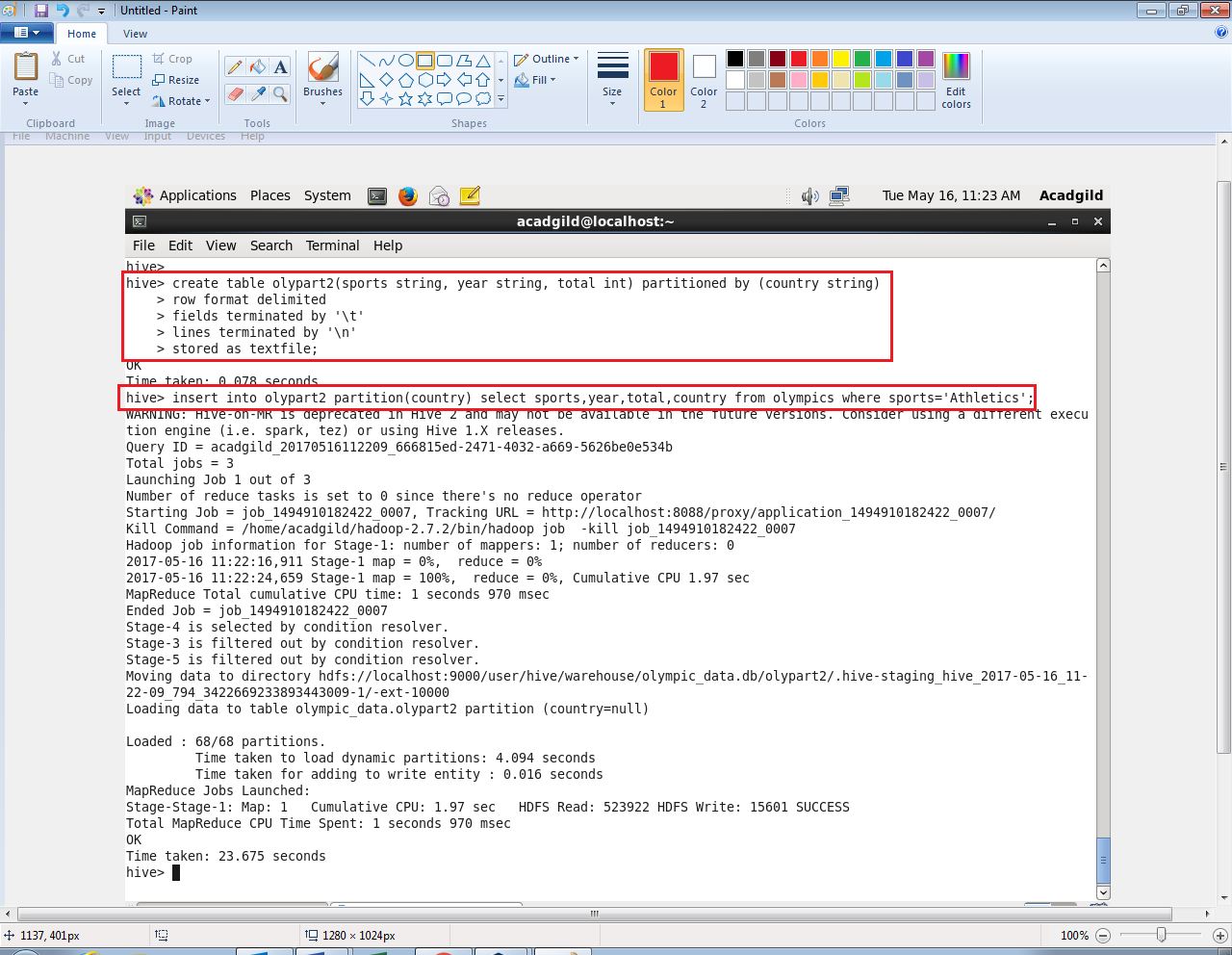


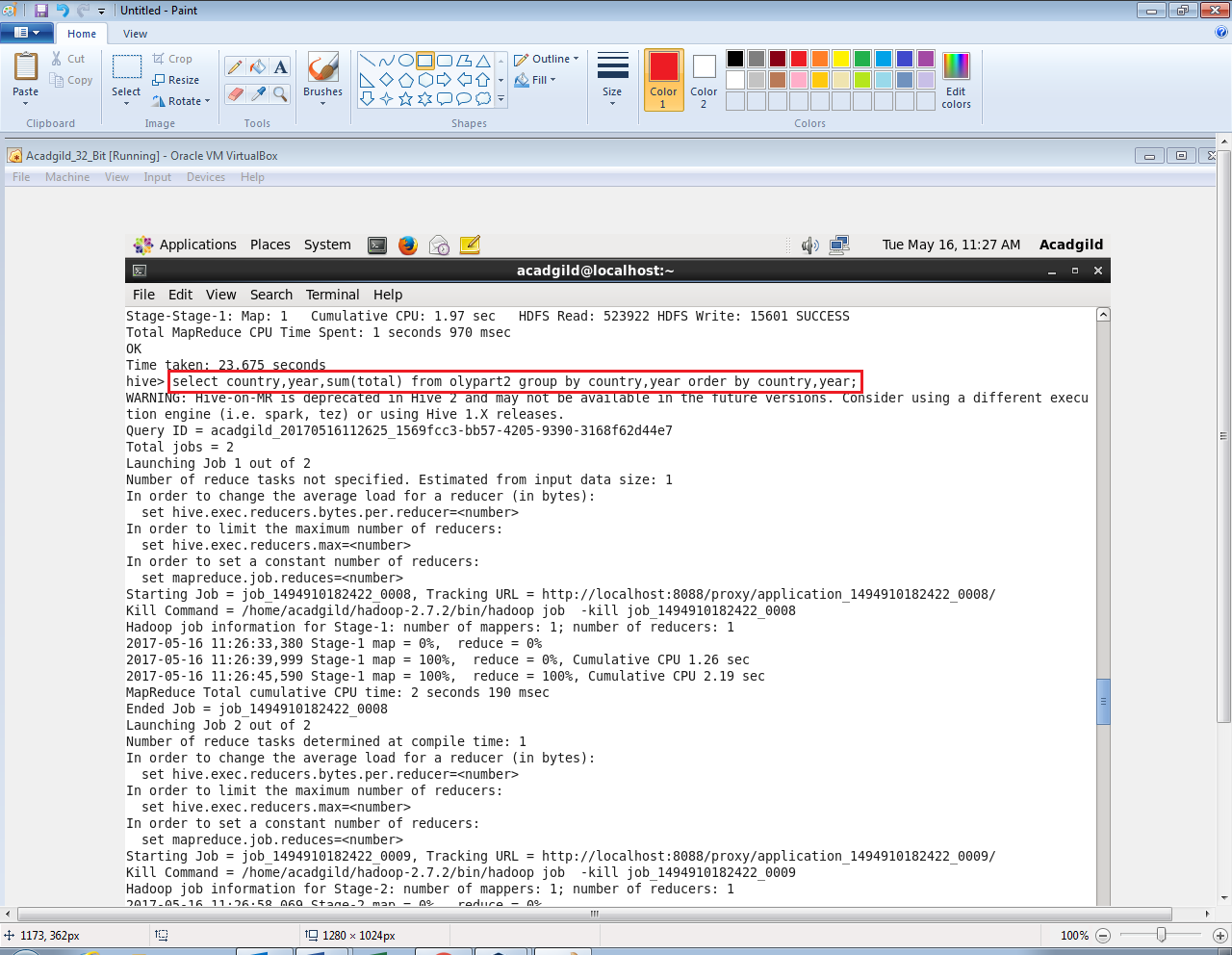


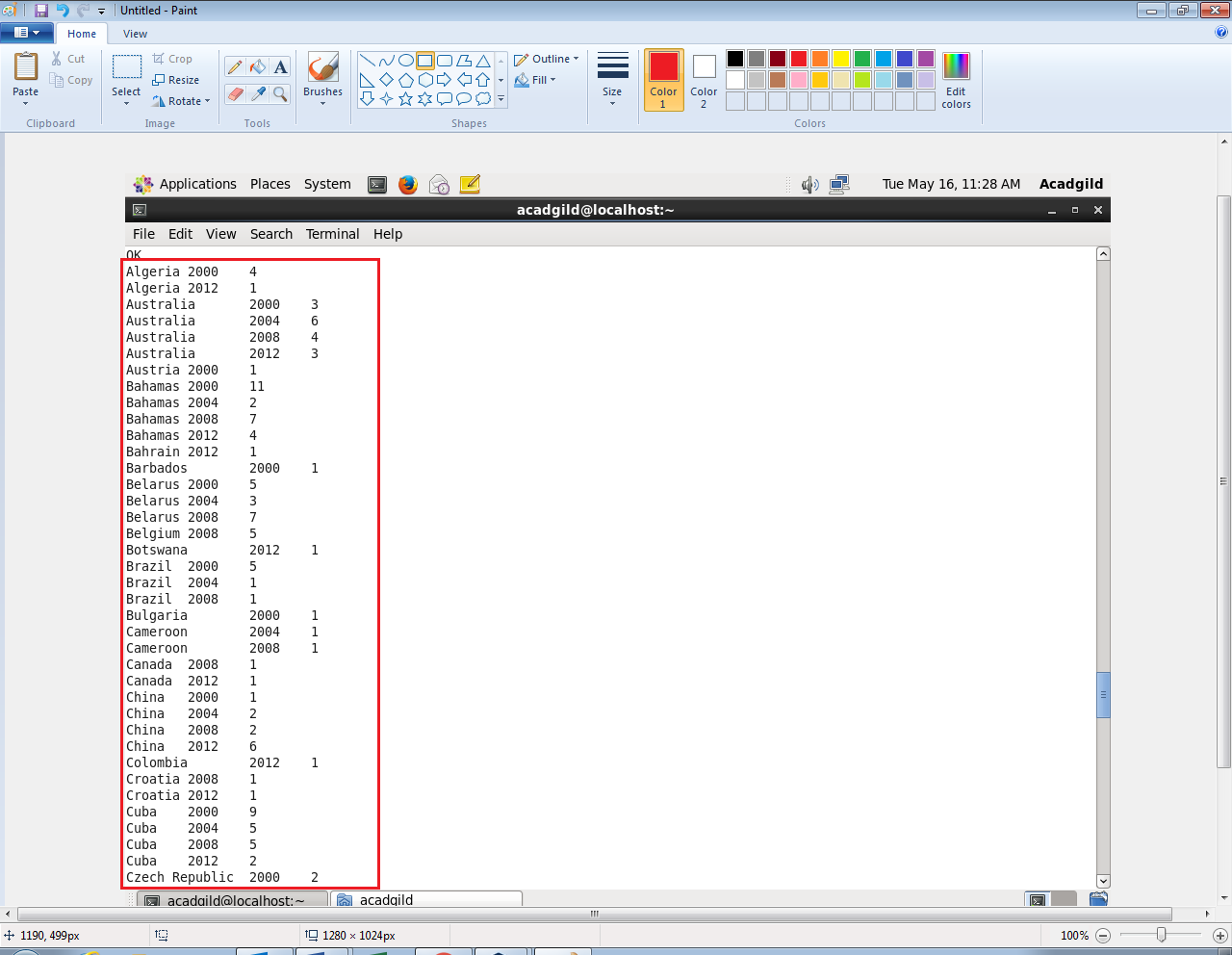


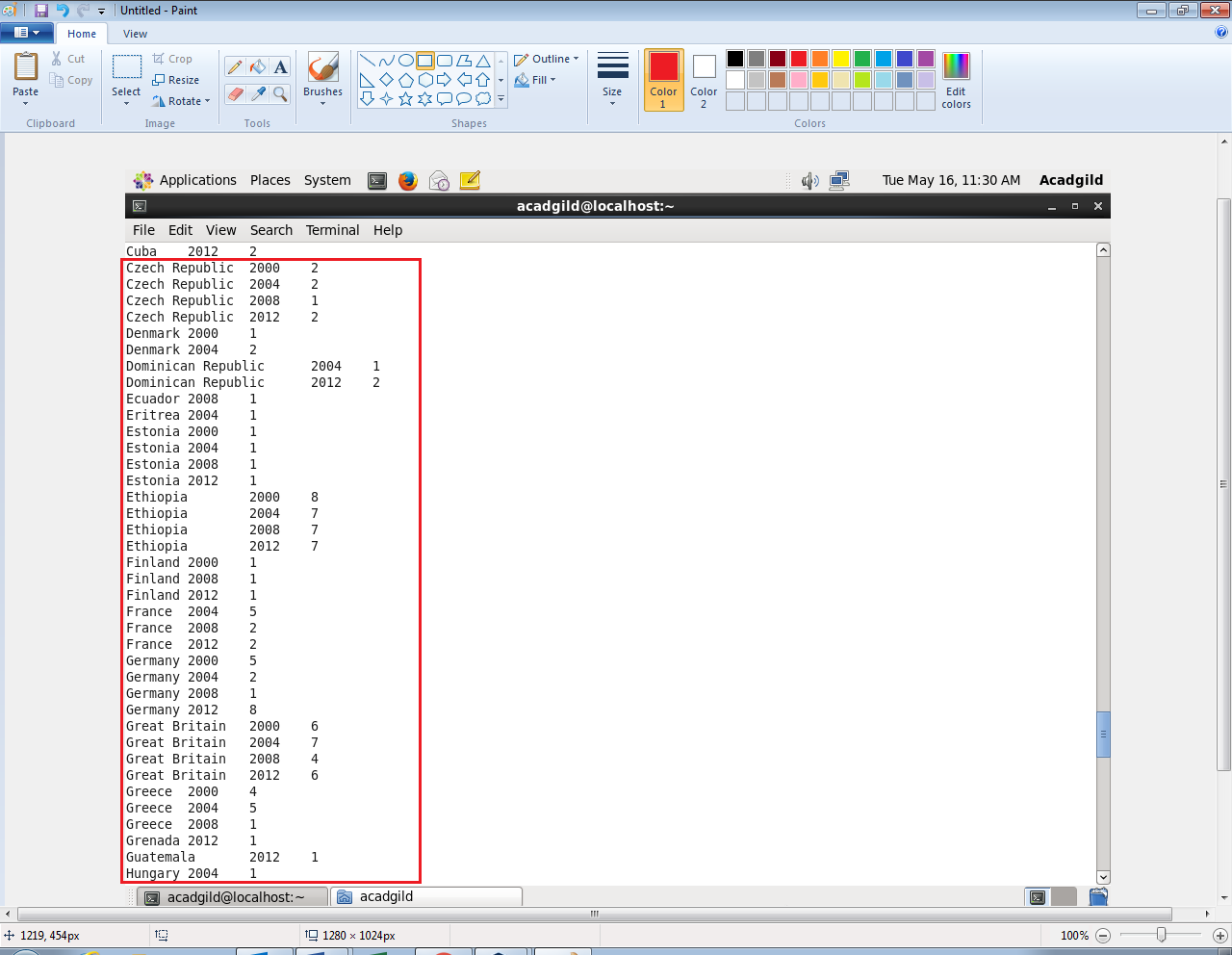
* **Find the number of medals each country won in Athletics year wise.**

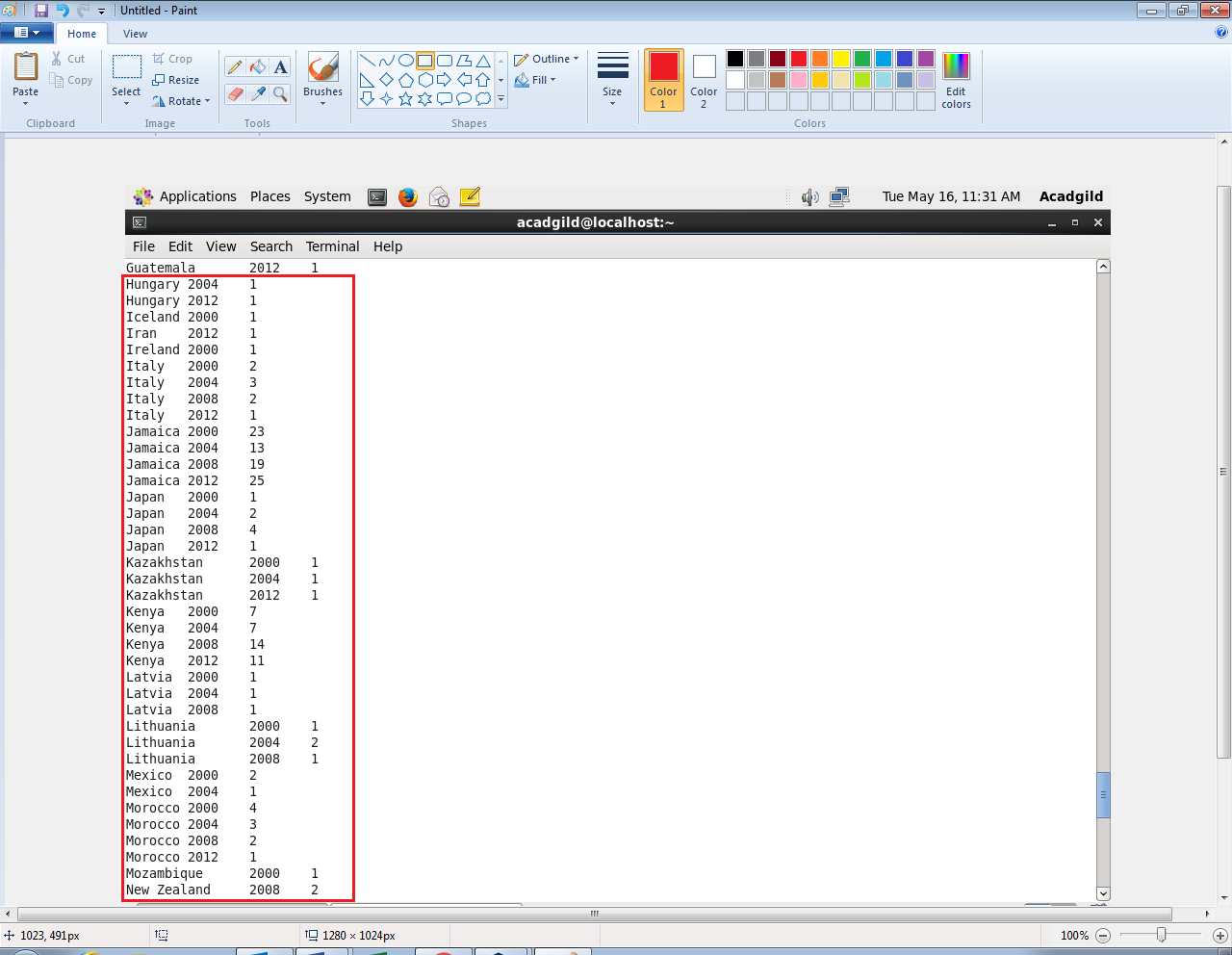
**Output:**

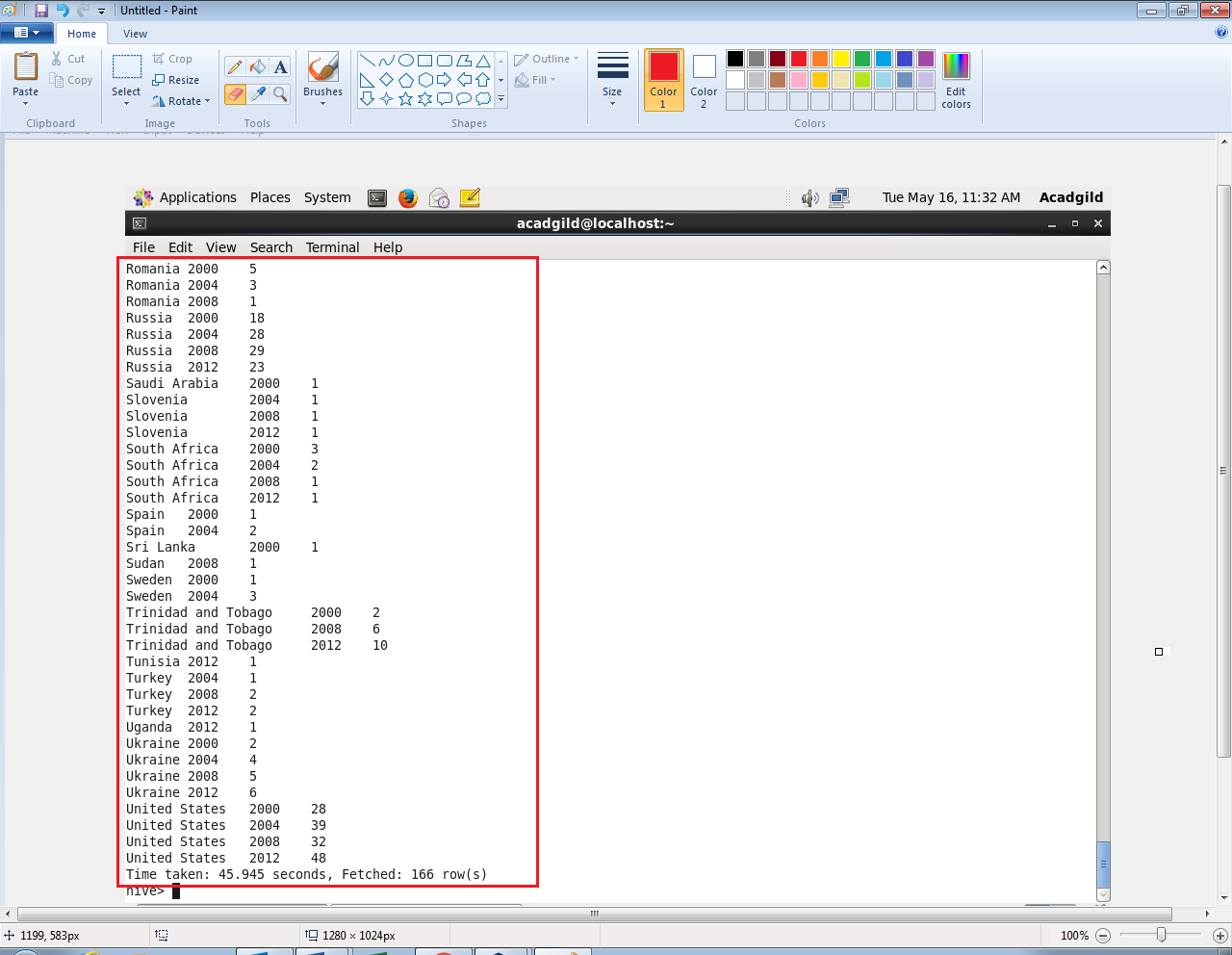












* **Find the average age of athletes participated from each country in Olympics year wise.**

**Output:**

