

ASSIGNMENT 9.1 Advance Hive

Problem Statement:

Given Olympic data set, write a hive program for the below tasks.

Created table and loaded the data using below commands:

1.create table Olympic_data(name String, age Int, country String, year String, Cdate String, sport String, GMedal int, SMedal int, BMedal int, Totmedals int) row format delimited fields terminated by '\t';

2. LOAD DATA LOCAL INPATH '/home/acadgild/olympix_data.csv' into table Olympic_data;

Task 1:

1. Write a Hive program to find the number of medals won by each country in swimming.

Query for the above task:

select country , sum(totmedals) from olympic_data where sport = "Swimming" group by country;

output:

```
-----  
Argentina      1  
Australia      163  
Austria 3  
Belarus 2  
Brazil 8  
Canada 5  
China 35  
Costa Rica 2  
Croatia 1  
Denmark 1  
France 39  
Germany 32  
Great Britain 11  
Hungary 9  
Italy 16  
Japan 43  
Lithuania 1  
Netherlands 46  
Norway 2  
Poland 3  
Romania 6  
Russia 20  
Serbia 1  
Slovakia 2  
Slovenia 1  
South Africa 11  
South Korea 4  
Spain 3  
Sweden 9  
Trinidad and Tobago 1  
Tunisia 3  
Ukraine 7  
United States 267  
Zimbabwe 7  
Time taken: 38.62 seconds, Fetched: 34 row(s)
```

2. Write a Hive program to find the number of medals that India won year wise.

Query for the above task:

```
select year , sum(totmedals) from olympic_data where country = "India" group by year;
```

output:

```
-----  
2000      1  
2004      1  
2008      3  
2012      6
```

3. Write a Hive Program to find the total number of medals each country won.

```
select country , sum(totmedals) from olympic_data group by country;
```

Output:

4. Write a Hive program to find the number of gold medals each country won.

```
select country , sum(GMedals) from olympic_data group by country;
```

Output:

Task 2:

Write a hive UDF that implements functionality of string `concat_ws(string SEP, array<string>)`.

This UDF will accept two arguments, one string and one array of string.

It will return a single string where all the elements of the array are separated by the SEP.

Steps to create a udf in Hive:

2. For Adding the import `org.apache.hadoop.hive.ql.exec.UDF` without error add the below external jar files as below:

`install/hive/bin/hive-exec-2.3.2`

3. Once after fixing all the code issues create a jar file(`concat_ws1.jar`) using eclipse IDE.

Source code is attached as separate file explaining the code.

4. Move the `concat_ws1.jar` file to the VM local file system.

5. Add the jar file and create temporary function as below:

ADD JAR /home/acadgild/concat_ws1.jar;

CREATE TEMPORARY FUNCTION concat1 AS 'concat_ws.concat_ws';

7. Now use the created temporary file to concat1 to create the below output.

select concat1('-', zip,temp) from temperature_data;

Output:
