**5 Days FDP on Big Data Analytics**

**Day 3-Lecture 11 Demo**

**“Apache Hive Query Language (HQL)- Sport Data Analysis”**

**Problem Statement for *‘Sports.csv’* dataset**

Sports data analytics are used not only in cricket but many [other sports](https://adtmag.com/blogs/dev-watch/2017/07/sports-analytics.aspx) such as ‘*Swimming,*’ ‘*Tennis,’ ‘Shooting’, ‘Rowing’* etc. This can be used for improving the overall team performance and maximizing winning chances. Consider a ***‘Sports.cs*v’** dataset of International Games which is provided as an input to Apache Hive, analyzed and partitioned based on different given categories. Write ***Apache Hive Query Language (HQL) Script*** to answer the following queries at ***Hadoop CLI shell mode:***

**Data set: ‘Sports.csv’**

**Task 1:** Write HQL Script to ***design ‘schema’*** and create the required ***‘table’*** for ***‘sports.csv’***.

**Answer:**

……………………………………………………………………………………………….

**Create Table Syntax:**

1. **Create Table Syntax at the hive CLI.**

|  |
| --- |
| hive> create table if not exists table3(col1 int,col2 array<string>,col3 string,col4 int)row format delimited fields terminated by',' collection items terminated by':' lines terminated by'\n' stored as textfile; |

**The above command creates a table name ‘table3’ with dataset schema.**

* **row format delimited:** Means Input file to be loaded to be treated as delimiter type with each field separated by comma (,)
* **collection items terminated by':'-** it means inside array the elements are separated by double column
* **lines terminated by'\n' stored as text file:** Each line is start with new row and this data to be store in the form of text file. By default is also text file format.

……………………………………………………………………………………………….

**Query 1 Solution:**

**hive>** create table if not exists FDPSportTable(AthleteName string, Age int, Country string, Year int, ClosingDate string, Sport string, Gold\_Medals int, Silver\_Medals int, Bronze\_Medals int, Total\_Medals int) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;

**Check table is created or not using following command**

**hive> show tables;**

**To check the schema of table**

**hive> describe FDPSportTable;**

**Load Data into Table**

**hive>** load data local inpath '/home/cloudera/Downloads/Sports.csv'into table FDPSportTable;

**Crosscheck data is loaded successfully or not**

**hive>** select \* from FDPSportTable;

Graphical user interface, application

Description automatically generated with medium confidence

**Task 2**: Write HQL Script to analyze and list the total number of medals won by each country for ‘***Swimming’*** sport category***.***

**hive>** select sum(total\_medals),Country from FDPSportTable where Sport=="Swimming" group by Country;

Graphical user interface, text, application, email

Description automatically generated

**Task 3:** Write HQL Script to calculate and display the total number of ***‘Gold Medals’*** won by India. (Assume all sports category).

**hive>** SELECT SUM(Gold\_Medals) FROM FDPSportTable WHERE Country =="India";

**Task 4:** Write HQL Script to list the number of medals won by ***‘China’*** in ***‘Shooting’***.

**hive>** select count(total\_medals) from FDPSportTable where Country=='China' and Sport='Shooting'

**Task 5:** Write HQL Script to calculate and count of the total number of medals each country won.

**hive>** select count(Total\_Medals) from FDPSportTable group by Country;

Graphical user interface, text, application

Description automatically generated

**Task 6:** Write HQL Script to display year and countries name that won medals in ‘***Shooting’***.

**hive>** select Year, Country from FDPSportTable where Sport=='Shooting';

Graphical user interface, text, application

Description automatically generated

**Task 7:** **Write HQL Script to group and display the medals won from 2000 to 2008.**

**hive>**select \* from FDPSportTable where Year>2000 and Year<2008;