Project 1: Analysis on Olympic dataset using HDFS and Hive

In this project, I'm going to perform an analysis on the Olympic dataset took from Kaggle. This dataset contains a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. Dataset contains 2,71,116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events).

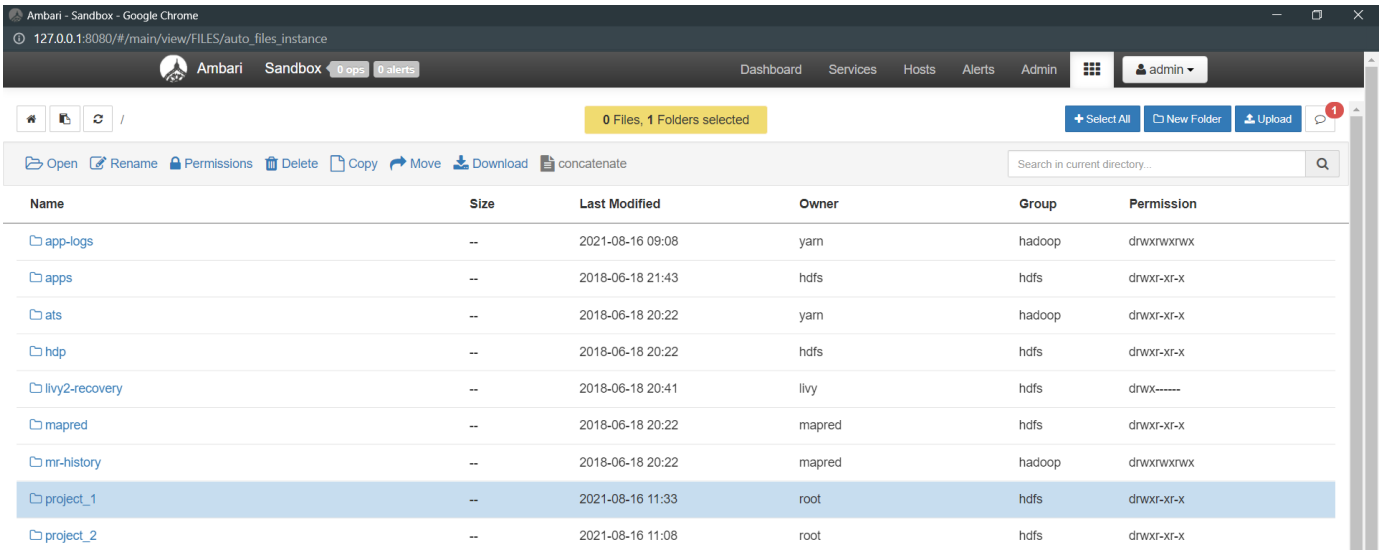
The columns are:

1) ID - Unique number for each athlete
2) Name - Athlete's name
3) Sex - M or F
4) Age - Integer
5) Height - In centimeters
6) Weight - In kilograms
7) Team - Team name
8) NOC - National Olympic Committee 3-letter code
9) Games - Year and season
10) Year - Integer
11) Season - Summer or Winter
12) City - Host city
13) Sport - Sport
14) Event - Event
15) Medal - Gold, Silver, Bronze, or NA

For performing analysis I'm going to use Hadoop and Apache Hive as data warehousing software. Hive gives an SQL-like interface to query data stored in various databases and file systems in this project I'm going to use the HDFS file system for data storage.

- ## Create Directory in HDFS :

hdfs dfs -mkdir /project_1

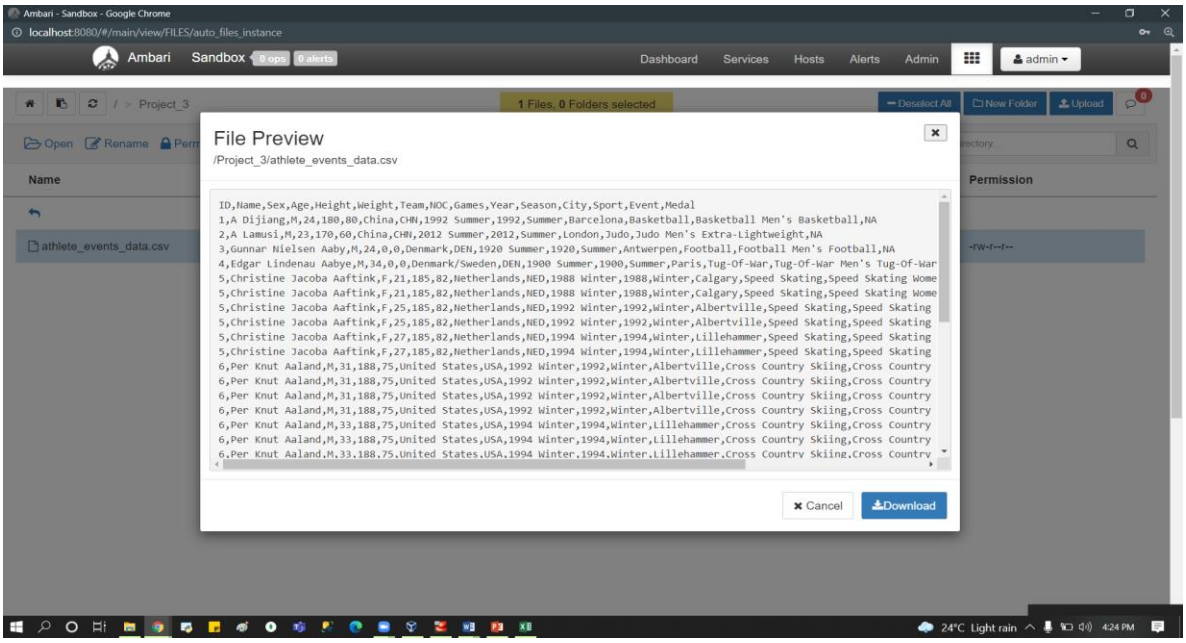


- ## Stored data inside project_1 directory

HDFS DFS –copyFromLocal ‘source file path’ ‘dest file path’

NOTE: In this case I’m using sandbox therefore I’m not able to do this so I direct uploaded data through ambari.

- ## Look Dataset which is present in HDFS



- Start with hive:

Hive

```
[root@sandbox-hdp ~]# hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
hive>
```

- Create a database:

create database project_1

```
hive> create database project_1;
OK
Time taken: 6.73 seconds
hive> show databases;
OK
bigdata
default
foodmart
project_1
Time taken: 0.546 seconds, Fetched: 4 row(s)
hive>
```

- Create a table inside database and storing data which is present in HDFS in the form of CSV

create table game_data(ID int,Name string,Sex string,Age int,Height int,Weight int,Team string,NOC string,Games string,Year int,Season string,City string,Sport string,Event string,Medal string) row format delimited fields terminated by ',' stored as textfile location '/project_3/' TBLPROPERTIES ("skip.header.line.count"="1");

- ## Describe the table

```
hive> desc game_data;
OK
col_name        data_type       comment
id                      int
name                    string
sex                     string
age                     int
height                  int
weight                  int
team                    string
noc                     string
games                   string
year                    int
season                  string
city                    string
sport                   string
event                   string
medal                   string
Time taken: 0.54 seconds, Fetched: 15 row(s)
hive>
```

- ## Glance to the top 10 records

Select ID,name,sex,age,team,cityMedal from game_data limit 10;

```
hive> Select ID,name,sex,age,team,city,Medal from game_data limit 10;
OK
id      name    sex     age     team    city    medal
1       A Dijiang       M       24      China   Barcelona       NA
2       A Lamusi        M       23      China   London  NA
3       Gunnar Nielsen Aaby     M       24      Denmark Antwerpen       NA
4       Edgar Lindenau Aabye    M       34      Denmark/Sweden  Paris   Gold
5       Christine Jacoba Aaftink        F       21      Netherlands     Calgary NA
5       Christine Jacoba Aaftink        F       21      Netherlands     Calgary NA
5       Christine Jacoba Aaftink        F       25      Netherlands     Albertville     NA
5       Christine Jacoba Aaftink        F       25      Netherlands     Albertville     NA
5       Christine Jacoba Aaftink        F       27      Netherlands     Lillehammer     NA
5       Christine Jacoba Aaftink        F       27      Netherlands     Lillehammer     NA
Time taken: 0.176 seconds, Fetched: 10 row(s)
hive>
```

1. Find how many medals are wined by player

select medal AS Medal_type,count(medal) AS Count from game_data group by medal;

```
OK
medal_type      count
Bronze  13295
Gold    13372
NA      231333
Silver  13116
Time taken: 5.078 seconds, Fetched: 4 row(s)
hive>
```

2. Count of medal distribution based on year

select year, count(medal) from game_data group by year;

```
year    medals
1896    380
1900    1936
1904    1301
1906    1733
1908    3101
1912    4040
1920    4292
1924    5693
1928    5574
1932    3321
1936    7401
1948    7480
1952    9358
1956    6434
1960    9235
1964    9480
1968    10479
1972    11959
1976    10502
1980    8937
1984    11588
1988    14676
1992    16413
1994    3160
1996    13780
1998    3605
2000    13821
2002    4109
2004    13443
2006    4382
2008    13602
2010    4402
2012    12920
2014    4891
2016    13688
Time taken: 11.745 seconds, Fetched: 35 row(s)
hive>
```

3. Find count of sport year wise

SELECT year, COUNT (DISTINCT sport) FROM game_data GROUP BY year;

```
year    sports
1896    9
1900    20
1904    18
1906    13
1908    24
1912    17
1920    25
1924    30
1928    25
1932    25
1936    32
1948    29
1952    27
1956    27
1960    27
1964    31
1968    30
1972    33
1976    33
1980    33
1984    35
1988    37
1992    41
1994    12
1996    31
1998    14
2000    34
2002    15
2004    34
2006    15
2008    34
2010    15
2012    32
2014    15
2016    34
Time taken: 14.59 seconds, Fetched: 35 row(s)
hive>
```

4. Year wise count of Events and count of Teams participated in first 10 Olympic games.

select year AS year,count(DISTINCT event) AS events,count(DISTINCT team) AS teams from game_data group by year limit 10;

```
----------------------------------------------
OK
year    events  teams
1896    43      18
1904    95      790

1906    74      52
1912    107     102

1920    158     72
1924    148     93
1928    136     85
1932    145     72
Time taken: 10.043 seconds, Fetched: 10 row(s)
hive>
```

5. Count the participation of players after olympic 2000 based on gender

select year,sex, count(sex) as count from game_data where year >= 2000 group by year,sex;

```
----------------------------------------------
OK
year    sex     count
2000    F       5431
2000    M       8390
2002    F       1582
2002    M       2527
2004    F       5546
2004    M       7897
2006    M       2625

2008    F       5816
2010    F       1847

2010    M       2555
2012    F       5815
2012    M       7105
2014    F       2023
2014    M       2868
2016    F       6223
2016    M       7465
Time taken: 5.651 seconds, Fetched: 18 row(s)
hive>
```

6. Find the count of medal between year 1980 to 2000 based on season and gender

```
select year AS year,season as season,sex ,count(medal) AS count
        from game_data
        where year between 1980 and 2000
group by year,season,sex;
```

```
OK-----------------------------------------

year    season  sex     count
1980    Summer  F       1756
1980    Summer  M       5435
1980    Winter  F       430
1980    Winter  M       1316
1984    Summer  F       2447
1984    Summer  M       7007
1984    Winter  F       536
1984    Winter  M       1598
1988    Summer  F       3543
1988    Summer  M       8494
1988    Winter  F       680
1988    Winter  M       1959
1992    Summer  F       4124
1992    Summer  M       8853
1992    Winter  F       1054
1992    Winter  M       2382
1994    Winter  F       1105
1994    Winter  M       2055
1996    Summer  F       5008
1996    Summer  M       8772
1998    Winter  F       1384
1998    Winter  M       2221
2000    Summer  F       5431
2000    Summer  M       8390
Time taken: 7.45 seconds, Fetched: 24 row(s)
hive>
```

7. Top 3 countries who win more medals in Olympic 2016

select year AS Year,team,count(team) AS country from game_data where year=2016 group by Year,team sort by country DESC limit 3;

```
-----------------------------
OK
year    team    country
2016    United States   699
2016    Brazil  571
2016    Germany 528
hive>
```

8. Find top 20 countries who win more medals in olympic games

select team as Country, count(medal) AS Medals from game_data group by team sort by Medals DESC limit 20;

```
-----------------------------------------------
OK
country medals
"United States" 4111
"Soviet Union"  1926
"Germany"       1367
"Great Britain" 1216
"Australia"     1127
"Canada"        1040
"Italy" 937
"Russia"        908
"Sweden"        893
"France"        855
"Japan" 792
"East Germany"  771
"China" 762
"Netherlands"   730
"Hungary"       619
"Norway"        605
"Finland"       581
"Romania"       473
"South Korea"   450
"Spain" 432
Time taken: 10.792 seconds, Fetched: 20 row(s)
```

9. Find the name of top 3 players based on Gold medals they achieved.

select name, count(medal) as Total from game_data where medal = 'Gold' group by name sort by Total DESC limit 3;

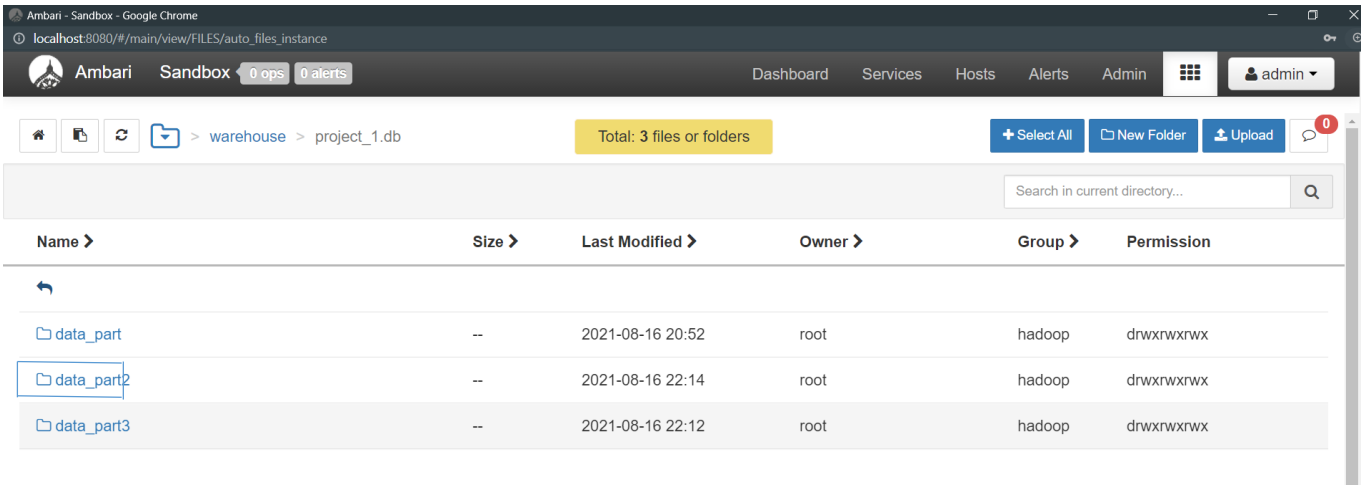
```
name    total
Michael Fred Phelps  II 23
"Raymond Clarence ""Ray"" Ewry" 10
Larysa Semenivna Latynina (Diriy-)      9
Time taken: 13.006 seconds, Fetched: 3 row(s)
hive>
```

10. Find the total metals achieved by each player and store name and count of medal into file in HDFS.

INSERT OVERWRITE DIRECTORY "/Query_result" ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT name,count(medal) as medals FROM game_data group by name order by medals DESC;

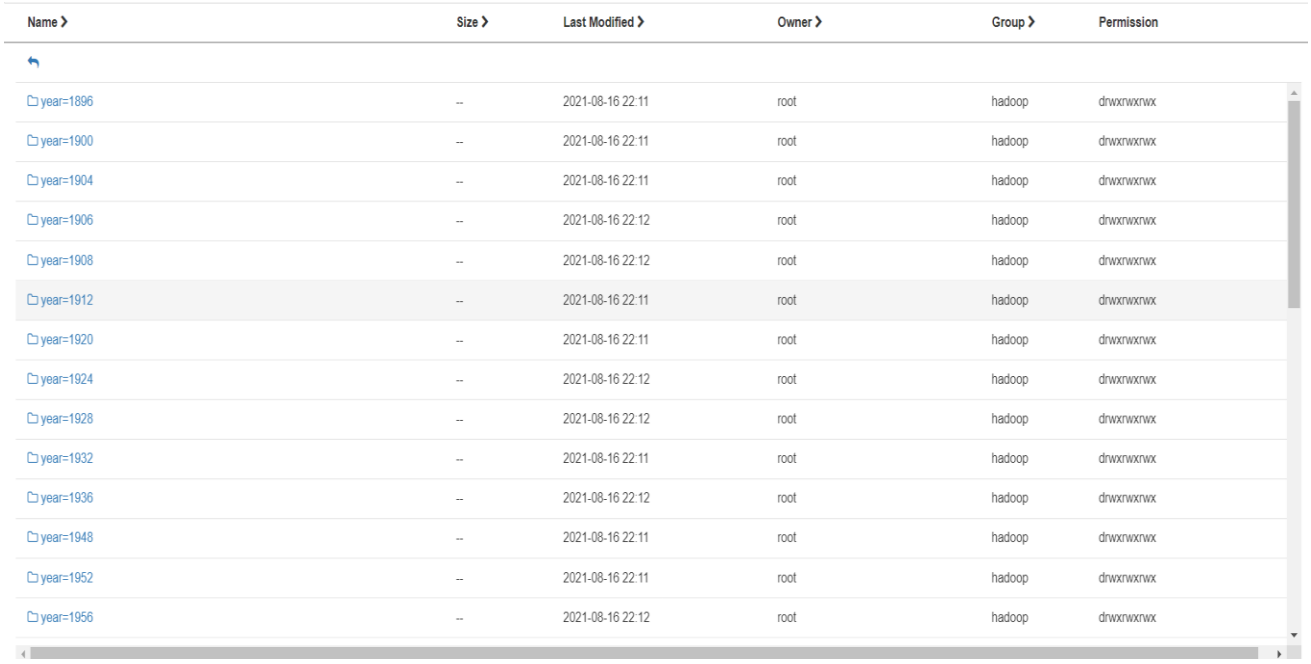
11. Create table for partition for performing partition on dataset based on year and city

Create table data_part3(ID int,Name string,Sex string,Age int,Height int,Weight int,Team string,NOC string,Games string, Season string,Sport string,Event string,Medal string) partitioned by (Year int,city string);

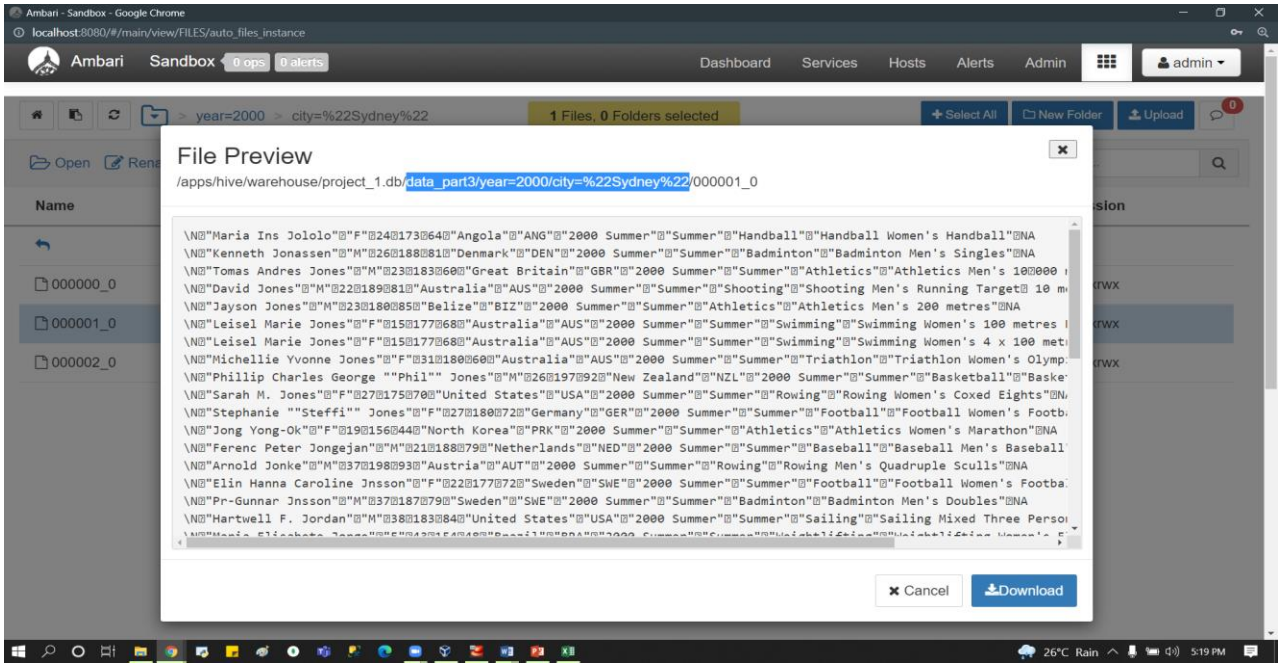


12. Storing the data

insert overwrite table data_part3 partition(year,city) select ID,Name,Sex, Age,Height,Weight,Team,NOC,Games, Season,Sport,Event,Medal,year,City from game_data;



13. Our data is partitioned by year and city, here take a look at data which is partitioned by year 2000 and city ‘Sydney’



ANY QUESTION ?

Thank you!