CERTIFICATION PROJECT

Country Project

About Me:

Name: RAMAN BHADAURIA
Email: 257ramanrb@gmail.com
Batch: Big Data and Hadoop Certification Training
Start Date: 26th January, 2019
Deadline: 30th April, 2019 (Extended by the support team)

---

edureka!

Table of Contents

1. Problem Statement
2. Dataset Sample
3. Dataset Description
4. Execution (Hive Queries and Results)
5. Conclusion

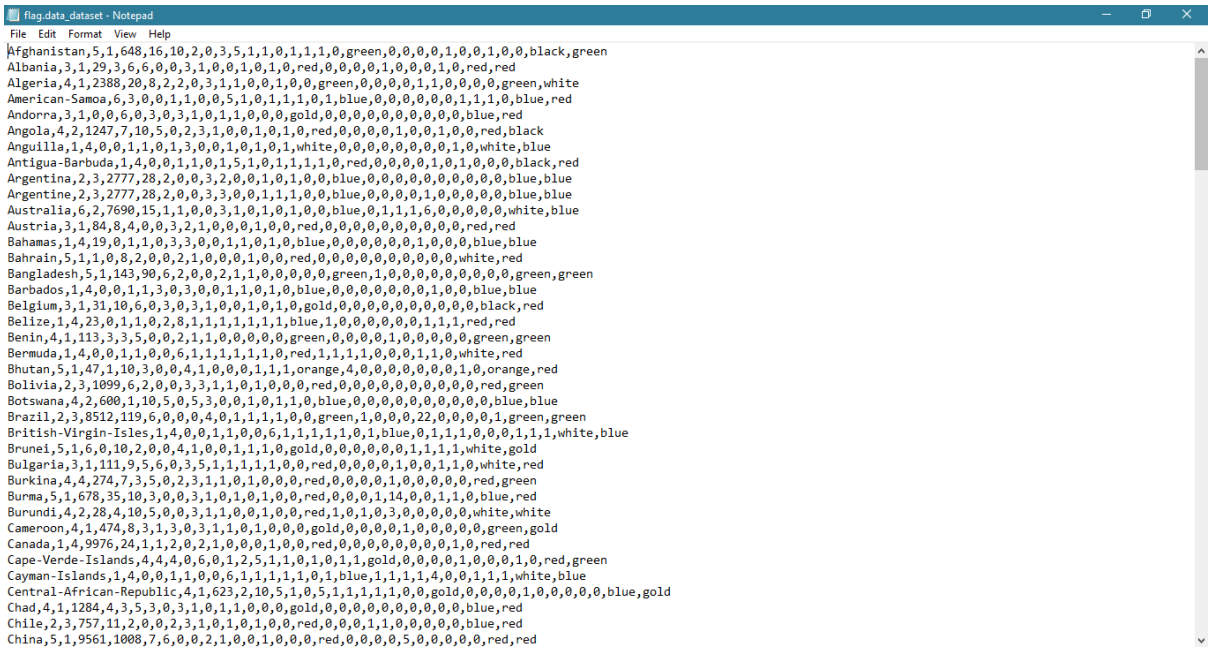
1. Problem Statement

A. Count number of countries based on landmass.
B. Find out top 5 country with Sum of bars and strips in a flag.
C. Count of countries with icon.
D. Count of countries which have same top left and top right color in flag.
E. Count number of countries based on zone.
F. Find out largest county in terms of area in NE zone.
G. Find out least populated country in S.America landmass.
H. Find out largest speaking language among all countries.
I. Find most common colour among flags from all countries.
J. Sum of all circles present in all country flags.
K. Count of countries which have both icon and text in flag.

2. Dataset

http://www.edureka.co/medias/yz5zdt174e/download?media_file_id=171471702

Sample of the dataset:



3. Dataset Description

1. Title: Flag database
2. Source Information
   -- Creators: Collected primarily from the "Collins Gem Guide to Flags":

Collins Publishers (1986).
-- Donor: Richard S. Forsyth
8 Grosvenor Avenue
Mapperley Park
Nottingham NG3 5DX
0602-621676
-- Date: 5/15/1990

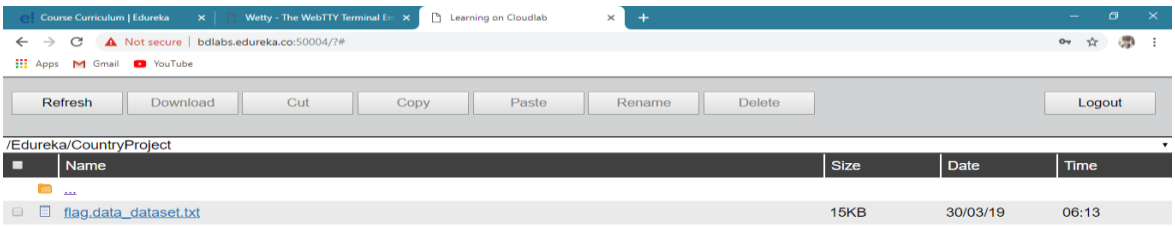
3. Past Usage:
-- None known other than what is shown in Forsyth's PC/BEAGLE User's Guide.
4. Relevant Information:
-- This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colours in its flag.
-- 10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.
5. Number of Instances: 194
6. Number of attributes: 30 (overall)
7. Attribute Information:
    1. name Name of the country concerned
    2. landmass 1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania
    3. zone Geographic quadrant, based on Greenwich and the Equator 1=NE, 2=SE, 3=SW, 4=NW
    4. area in thousands of square km
    5. population in round millions
    6. language 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
    7. religion 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu,
    5=Ethnic, 6=Marxist, 7=Others
    8. bars Number of vertical bars in the flag
    9. stripes Number of horizontal stripes in the flag
    10. colours Number of different colours in the flag
    11. red 0 if red absent, 1 if red present in the flag
    12. green same for green
    13. blue same for blue
    14. gold same for gold (also yellow)

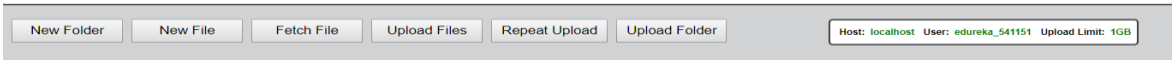
15. white same for white
16. black same for black
17. orange same for orange (also brown)
18. mainhue predominant colour in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)
19. circles Number of circles in the flag
20. crosses Number of (upright) crosses
21. saltires Number of diagonal crosses
22. quarters Number of quartered sections
23. sunstars Number of sun or star symbols
24. crescent 1 if a crescent moon symbol present, else 0
25. triangle 1 if any triangles present, 0 otherwise
26. icon 1 if an inanimate image present (e.g., a boat), otherwise 0
27. animate 1 if an animate image (e.g., an eagle, a tree, a human hand)
    present, 0 otherwise
28. text 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. topleft colour in the top-left corner (moving right to decide tie-breaks)
30. botright Colour in the bottom-left corner (moving left to decide tie-breaks)

8. Missing values: None

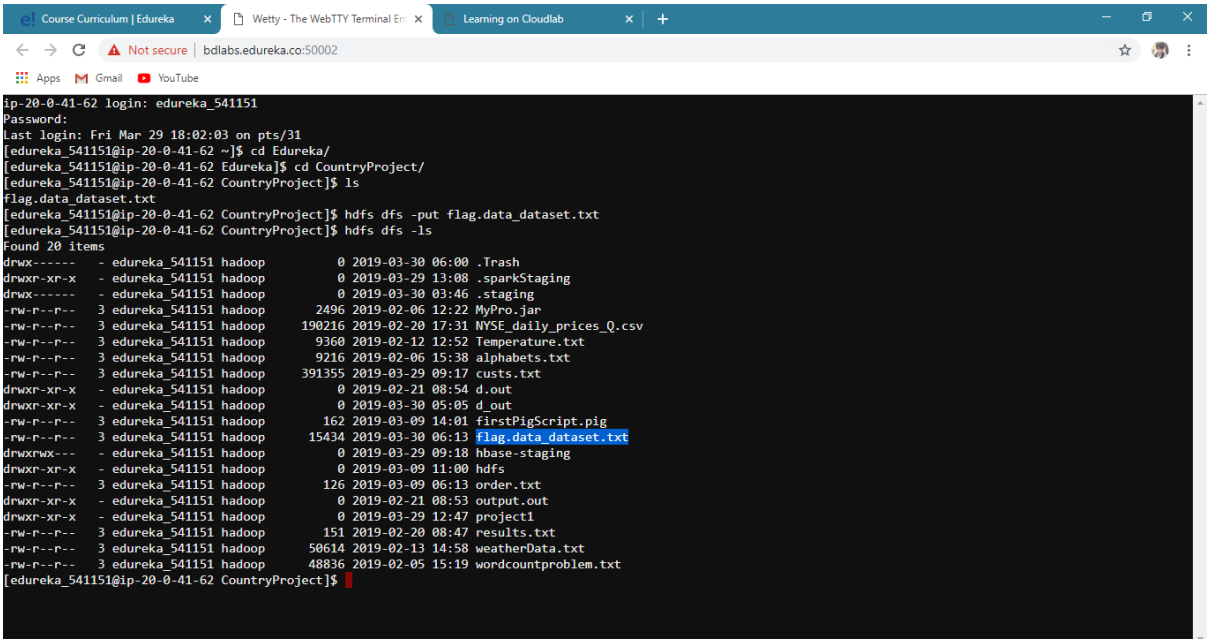
4. Execution (Hive Queries and Results)

1. Uploading dataset under user/edureka_541151/Edureka/CountryProject





2. Put dataset on HDFS



3. Create table raman_country

```
hive> create table raman_country(name varchar(100), landmass int, zone int, area int, pop int, lang int, relig int, ver_bars int, horz_stripes int, no_of_colors in
t, red int, green int, blue int, gold_yellow int, white int, black int, orange_brown int,mainhue varchar(20), circles int, up_crosses int, saltires_diag_cross int,
 quart int, sunstars int, crescent_moon int, triangle int, icon int, animate int, text int, topleft varchar(20), botright varchar(20))
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 2.194 seconds
```

4. Load data into table

```
hive> load data inpath 'flag.data_dataset.txt' into table raman_country;
Loading data to table default.raman_country
Table default.raman_country stats: [numFiles=1, totalSize=15434]
OK
Time taken: 0.527 seconds
```

5. Describe table

```
hive> describe raman_country;
OK
name                    varchar(100)
landmass                int
zone                    int
area                    int
pop                     int
lang                    int
relig                   int
ver_bars                int
horz_stripes            int
no_of_colors            int
red                     int
green                   int
blue                    int
gold_yellow             int
white                   int
black                   int
orange_brown            int
mainhue                 varchar(20)
circles                 int
up_crosses              int
saltires_diag_cross     int
quart                   int
sunstars                int
crescent_moon           int
triangle                int
icon                    int
animate                 int
text                    int
topleft                 varchar(20)
botright                varchar(20)
Time taken: 0.097 seconds, Fetched: 30 row(s)
hive>
```

6. Problem Statements:

A. Count number of countries based on landmass.

Hive Query:

```
hive> SELECT landmass, case landmass
    > when 1 then 'N.America'
    > when 2 then 'S.AMerica'
    > when 3 then 'Europe'
    > when 4 then 'Africa'
    > when 5 then 'Asia'
    > when 6 then 'Oceania'
    > END AS Name, '\t', COUNT(*)
    > from raman_country
    > group by landmass;
Query ID = edureka_541151_20190330071111_89af35d6-0a00-479c-b811-cbaae458637a
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 910 msec
OK
1       N.America                       31
2       S.AMerica                       17
3       Europe                  35
4       Africa                  52
5       Asia                    39
6       Oceania                 20
Time taken: 16.012 seconds, Fetched: 6 row(s)
hive>
```

B. Find out top 5 countries with Sum of bars and strips in a flag.

Hive Query:

```
hive> SELECT name, '-', ver_bars + horz_stripes AS sum from ramu_country ORDER BY sum DESC LIMIT 5;
Query ID = edureka_541151_20190330071313_6d84ad49-cf4b-428d-a0c8-9ea4d0786133
```

Result:

```
Total MapReduce CPU Time Spent: 3 seconds 860 msec
OK
Malaysia        -       14
USA     -       13
Liberia -       11
Uruguay -       9
Greece  -       9
Time taken: 16.023 seconds, Fetched: 5 row(s)
hive>
```

C. Count of countries with icon.

Hive Query:

```
hive> SELECT 'NUMBER OF COUNTRIES WITH ICON', '-', COUNT(*) as Count from ramu_country WHERE icon=1;
Query ID = edureka_541151_20190330070606_6ff9e031-cdf8-406e-a0d3-62303c448eb2
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 200 msec
OK
NUMBER OF COUNTRIES WITH ICON    -       49
Time taken: 24.019 seconds, Fetched: 1 row(s)
hive>
```

D. Count of countries which have same top left and top right colour in flag.

Hive Query:

```
hive> SELECT 'Number Of Reqd. Countries are', COUNT(*) from raman_country WHERE topleft=botright;
Query ID = edureka_541151_20190330072222_57ac389d-e137-4a90-a8e9-c3e9a3eab341
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 260 msec
OK
Number Of Reqd. Countries are   76
Time taken: 15.063 seconds, Fetched: 1 row(s)
hive>
```

E. Count number of countries based on zone.

Hive Query:

```
hive> SELECT zone, case zone
    > when 1 then 'NE'
    > when 2 then 'SE'
    > when 3 then 'SW'
    > when 4 then 'NW'
    > END AS zone,'-', COUNT(*) from raman_country GROUP BY zone;
Query ID = edureka_541151_20190330072929_96596795-7311-4669-9729-b96e7d100c07
```

Result:

```
Total MapReduce CPU Time Spent: 5 seconds 200 msec
OK
1       NE      -       91
2       SE      -       29
3       SW      -       16
4       NW      -       58
Time taken: 15.97 seconds, Fetched: 4 row(s)
```

F. Find out largest county in terms of area in NE zone.

Hive Query:

```
hive> SELECT 'Country with Max area in NE:' , name, 'Area:', area from raman_country where zone=1 order by area DESC LIMIT 1;
Query ID = edureka_541151_20190330073333_20b72d9c-4fd4-4b3b-b78b-6b951562fc91
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 400 msec
OK
Country with Max area in NE:    USSR    Area:    22402
Time taken: 14.918 seconds, Fetched: 1 row(s)
hive>
```

G. Find out least populated country in S.America landmass.

Hive Query:

```
hive> SELECT 'Least populated country in S.America:', name, 'Population:', pop from raman_country where landmass=2 order by pop LIMIT 1;
Query ID = edureka_541151_20190330074343_9258dcaf-4fe0-46b5-9e99-0afc2c11c274
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 860 msec
OK
Least populated country in S.America:    Falklands-Malvinas      Population:     0
Time taken: 15.118 seconds, Fetched: 1 row(s)
```

H. Find out largest speaking language among all countries.

Hive Query:

```
hive> SELECT 'Largest speaking language:', case lang
    > when 1 then 'English'
    > when 2 then 'Spanish'
    > when 3 then 'French'
    > when 4 then 'German'
    > when 5 then 'Slavic'
    > when 6 then 'Other-Indo-European'
    > when 7 then 'Chinese'
    > when 8 then 'Arabic'
    > when 9 then 'Japanese/Turkish/Finnish/Magyar'
    > when 10 then 'Others'
    > END AS language, 'Lang Id:', lang, 'No. of countries:', COUNT(*) as count
    > from raman_country
    > GROUP By lang
    > ORDER By count DESC
    > LIMIT 1;
Query ID = edureka_541151_20190330080303_3c782b8c-4082-4d5f-ad7a-66aa83de7db0
```

Result:

```
Total MapReduce CPU Time Spent: 7 seconds 970 msec
OK
Largest speaking language:      Others  Lang Id:        10      No. of countries:       46
Time taken: 53.375 seconds, Fetched: 1 row(s)
```

I. Find most common colour among flags from all countries.

Hive Query:

```
hive> SELECT 'Most common colour:', mainhue, '-', COUNT(*) as count from raman_country GROUP BY mainhue ORDER BY count DESC LIMIT 1;
Query ID = edureka_541151_20190330080707_02a65b12-222f-4389-9e5c-16f9e6600d99
```

Result:

```
Total MapReduce CPU Time Spent: 6 seconds 870 msec
OK
Most common colour:     red     -       71
Time taken: 35.011 seconds, Fetched: 1 row(s)
hive>
```

J. Sum of all circles present in all country flags.

Hive Query:

```
hive> SELECT 'SUM OF CIRCLES: ', SUM(circles) from raman_country;
Query ID = edureka_541151_20190330081616_62682e8e-c39d-482a-ad96-9795c88ec522
```

Result:

```
Total MapReduce CPU Time Spent: 3 seconds 600 msec
OK
SUM OF CIRCLES:         33
Time taken: 21.688 seconds, Fetched: 1 row(s)
hive>
```

K. Count of countries which have both icon and text in flag.

Hive Query:

```
hive> SELECT 'COUNTRIES HAVING BOTH ICON AND TEXT - ', COUNT(*) from raman_country where icon=1 and text=1;
Query ID = edureka_541151_20190330081919_cb5cc3de-1f1a-4d39-b1d5-83a292d0d866
```

Result:

```
Total MapReduce CPU Time Spent: 4 seconds 510 msec
OK
COUNTRIES HAVING BOTH ICON AND TEXT -   13
Time taken: 14.94 seconds, Fetched: 1 row(s)
hive>
```

5. Conclusion

The “Country Project” has been completed successfully and all the problem statements are solved successfully by obtaining the required results.