**Project**

**Name:**

**Student Number:**

**Note: Read the following carefully**

* Any late submission will be considered as a grade of 0%

### I do not accept your report through email.

### Upload your video in google drive and make it public and give me access

### Provide all details (screen shots and explanation) for each question.

**Note**

1. You can use any platform Linux (Ubuntu, RedHat, CentOS, EC2), EMR, Hortonworks, Cloudera
2. This is a group project.
3. Troubleshooting is part of the evaluation. Please research and troubleshoot your issue.

**Task 1: Pig**

1. Use the dataset in the following URL: [UCI Machine Learning Repository: Flags Data Set](http://archive.ics.uci.edu/ml/datasets/Flags)
2. Review the Attribute information
3. Use Pig and answer the following questions:
4. Count number of countries based on landmass.

**flag\_details = LOAD '/project/flag.data' USING PigStorage(',')**

**as (name:chararray, landmass:int, zone:int, area:float, population:int, language:int,religion:int, bars:int, stripes:int, colours:int,**

**red:int, green:int, blue:int, gold:int, white:int, black:int, orange:int, mainhue:chararray, circles:int, crosses:int, saltires:int, quarters:int, sunstars:int, crescent:int, triangle:int ,icon:int, animate:int, text:int, topleft:chararray,botright:chararray);**

**group\_landmass = GROUP flag\_details by landmass;**

**group\_landmass\_count = FOREACH group\_landmass GENERATE COUNT($1),(flag\_details.landmass);**

**Dump group\_landmass\_count;**

1. Find out top 5 country with Sum of bars and strips in a flag.

**bars\_plus\_stripe = FOREACH flag\_details GENERATE (bars+stripes) as bars\_stripe, name as name;**

**bars\_plus\_stripe\_ordered = ORDER bars\_plus\_stripe by bars\_stripe DESC;**

**top\_bars\_plus\_stripe\_ordered = Limit bars\_plus\_stripe\_ordered 5;**

**dump top\_bars\_plus\_stripe\_ordered**

1. Count of countries with icon.

**filter\_with\_icon = FILTER flag\_details BY icon == 1;**

**filter\_with\_icon\_grp= Group filter\_with\_icon ALL;**

**filter\_with\_icon\_count = FOREACH filter\_with\_icon\_grp GENERATE COUNT(filter\_with\_icon.name);**

1. Count of countries which have same top left and top right color in flag

**filter\_with\_top = FILTER flag\_details BY topleft == botright;**

**filter\_with\_top\_grp = Group filter\_with\_top All;**

**filter\_with\_top\_count = FOREACH filter\_with\_top\_grp GENERATE COUNT(filter\_with\_top.name);**

1. Count number of countries based on zone.

**group\_zone = GROUP flag\_details by zone;**

**group\_zone\_count = FOREACH group\_zone GENERATE COUNT($1),(flag\_details.zone);**

**Dump group\_zone\_count;**

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

**filter\_with\_area = FILTER flag\_details BY zone == 1;**

**filter\_with\_area\_col = FOREACH filter\_with\_area GENERATE (name), (area);**

**dump filter\_with\_area\_col;**

**filter\_with\_area\_ordered = ORDER filter\_with\_area\_col by area DESC;**

**filter\_with\_area\_ordered\_top= limit filter\_with\_area\_ordered 1;**

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

2=S.America,

**filter\_with\_landmass = FILTER flag\_details BY landmass == 2;**

**filter\_with\_landmass\_col = FOREACH filter\_with\_landmass GENERATE (name), (population);**

**filter\_with\_landmass\_ordered = ORDER filter\_with\_landmass\_col by population;**

**filter\_with\_landmass\_ordered\_top= limit filter\_with\_landmass\_ordered 20;**

1. Find out largest speaking language among all countries.

**group\_language = GROUP flag\_details by language;**

**group\_language\_col= FOREACH group\_language GENERATE SUM(flag\_details.** population**) as population, (flag\_details.language) as language;**

**group\_language\_col\_ordered= ORDER group\_language\_col by population DESC;**

**largest\_group\_language\_col\_ordered= limit group\_language\_col\_ordered 1**

1. Find most common color among flags from all countries.

**grp\_all=Group flag\_details ALL;**

**most\_common\_color = FOREACH grp\_all GENERATE SUM(flag\_details.red) as sum\_red, SUM(flag\_details.blue) as sum\_blue, SUM(flag\_details.gold) as gold, SUM(flag\_details.black) as sum\_black, SUM(flag\_details.white), SUM(flag\_details.orange) as sum\_orange ;**

**dump most\_common\_color;**

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

**grp\_all = Group flag\_details ALL;**

**sum\_circles= FOREACH grp\_all GENERATE SUM(flag\_details.circles);**

**dump sum\_circles;**

What do you need to upload?

1. Softcopy of the codes for Part A to J (Notepad)
2. Record 2-4 minutes video and demonstrate your work for Part A to J (The video should have your voice and show your customized background and timestamp. Upload your video to BB. No YouTube or any other link. If you cannot upload the upload it in google drive and provide me a link to download it

**Task 2: Hive**

1. Upload Book.xml in HDFS then create a table in Hive and parse it.
2. RUN Three queries (on your choice) i.e run a query to display all authors.

What do you need to upload?

1. Softcopy of the codes for Part B (Notepad)
2. Record 2-4 minutes video and demonstrate your work for Part A and B (The video should have your voice and show your customized background and timestamp ,….Upload your video to BB. No YouTube or any other link. If you cannot upload the upload it in google drive and provide me a link to download it

**Note:** Processing XML data in Hive using XML SerDe XML has been one of the most important data structures and has been used for quite a long time for data transfers and storage. Parsing XML data and then processing it is always a tricky task as parsing XML is one of the costliest operations. Hive does not have any built-in support for XML data processing, but many organizations and individuals have made open-source contributions to XML SerDe .

**Task 3: Sqoop**

1. Export your table from Task 2 (Hive) by using Sqoop to any DB (either build in or new installation)
2. RUN Three queries (on your choice) i.e run a query to display all authors.

What do you need to upload?

1. Softcopy of the codes for Part A and B (Notepad)
2. Record 2-4 minutes video and demonstrate your work for Part A and B (The video should have your voice and show your customized background and timestamp. Upload your video to BB. No YouTube or any other link. If you cannot upload the upload it in google drive and provide me a link to download it

**Submission Requirement**

1. **Add all the three videos in Task 1,2 and 3 then upload only one video file.**
2. **Use one note pad file add accommodate all the three tasks soft copy.**

**So, you just need to upload one video and one notepad.**