**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.
5. Find out top 5 country with Sum of bars and strips in a flag.
6. Count of countries with icon.
7. Count of countries which have same top left and top right color in flag.
8. Count number of countries based on zone.
9. Find out largest county in terms of area in NE zone.
10. Find out least populated country in S.America landmass.
11. Find out largest speaking language among all countries.
12. Find most common color among flags from all countries.
13. Sum of all circles present in all country flags.

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.**