WQD7007 Big Data Management

Lab Test (Pig)

Instructions: Answer the following question by using command line interface or Ambari interface. Explain adequately how you get the answers that can include codes used in command line interfaces, codes used in Ambari, a number of print screens and related files. You can create a readme file (either in .txt or .docx) to explain your answer. Zip all the files and submit to the spectrum "Lab test" submission page.

Part 1: (5 marks)

1. Create a file named: "lab_(YourInitial).txt". E.g. lab_hwl.csv
2. Insert the following content into the file created.

> This is a dummy text file for WQD7007.
> This file contains 4 lines.
> Please remember to save this file.

3. Upload the file to HDFS in directory "/user/hdfs/lab_test/".

Part 2: (7 marks)

1. Create a "cars_small.csv" file using the following data. Load the data and display all the record in Pig script.

```
Chevrolet Chevelle Malibu;18.0;8;307.0;130.0;3504.;12.0;70;US
Buick Skylark 320;15.0;8;350.0;165.0;3693.;11.5;70;US
Plymouth Satellite;18.0;8;318.0;150.0;3436.;11.0;70;US
AMC Rebel SST;16.0;8;304.0;150.0;3433.;12.0;70;US
Ford Torino;17.0;8;302.0;140.0;3449.;10.5;70;UK
```

Attributes used:

Car (String); Miles Per Gallon, MPG (double), Cylinders (int), Displacement (double); Horsepower (double); Weight (double); Acceleration (double) ;Model (int); Origin (String).

2. Update the record that have horsepower = 150 to 120 and delete the record that have weight = 3433. Then, display the record that displacement smaller than 320 and origin from US.

Part 3: (3 marks)

1. Download "cars.csv" from spectrum. Then, load the file in Pig script and display first 10 records. Hint: you will need to change some settings and/or edit the csv file manually in order for you can upload to upload the content successfully.
2. Display maximum number of cylinders for each car model using Pig script.