McGill | School of Continuing Studies | École d'éducation permanente

Assessment
Data at Scale
YCBS-257

Assignment 1 – Part 1

15% - Part 1 of 2 due on 23:59 Sunday Feb 03 2019

Hadoop Data Analysis using Pig

Provided files:

9 Bixi Data files (one file for Stations, 8 files for rides)
1 protocols file
Download these files to your machine in order to perform the assignment

1. Analyzing structured data using Pig 50%

Apache Pig makes it possible to write complex queries over big datasets using a language named Pig Latin. The queries are transparently compiled into MapReduce jobs and run with Hadoop.

While doing this assignment, you are invited to check the official documentation http://pig.apache.org/docs/r0.12.0/).

Connect to your Virtual Machine and copy the files to a local directory on your machine.

Note: Depending on your machine resources, you may choose to run pig in **local** mode or in **mapreduce** mode.

Question 1:

Write a Pig script to load the files into two relations named **data** (all rides) and **stations** (stations), and remove the header lines of the csv files from the Pig relations.

Question 2:

Write a Pig script to compute and print on the screen the number of rows for each relations (data, stations).

Question 3:

Write a Pig script to split the data relation into two relations **members** and **notmembers**. The members relation should contain only peoples that are members and the notmembers relation should contains all non-member people.

Question 4:

Write a Pig script to compute and print on the screen the number of rows for each relation (members, notmembers).

Question 5:

Write a Pig script that given some station's code (hard-coded constant) will return the stations record if found.

Question 6:

Write a Pig script that will calculate the number of rides departures per station. The output does not have to be sorted.

Question 7:

Write a Pig script that will calculate the count of rides per station (start station) and the MIN, MAX, AVG of the ride's duration for members and non-members.

Question 8:

Write a Pig script that will list the Top 5 names of the start station for members and for non-member

2. Analyzing unstructured data using Pig 50%

You are given an unstructured text file to import it into a structured data analyzing tool such as Hive or Impala.

Your task is:

- To remove all the comments lines (lines starting with #)
- To remove the header line
- To output all the columns of the file to disk.

What to submit:

One Word file

○ <your name> - W2019A1P1.docx

✓ Sections inside should be separated and named

✓ Comments can be submitted in French or in English in the text file