CLOUD COMPUTING

Practical 6: Pig Latin Script

School of Computer Science
University College Dublin

TO BE GRADED: YES

In the world of Big data, it is crucial to be able to process the data quickly and efficiently. Processing data quickly means one should use massive-parallelism, and processing data efficiently means the execution of processes should be fault-tolerant and linearly scalable. The Hadoop ecosystem is an Open Source set of frameworks designed around this concept. Through its components, the Hadoop ecosystem enables developers to focus on solving their Big Data problems rather than developing ad hoc solutions to managing massive data, as in Big Data applications. Pig is an analysis platform, which provides a dataflow language called Pig Latin. In this practical, we will model and write some scripts in Pig Latin to perform some data analysis.

Exercise 1

Suppose that we have a large data log about user's movie preferences. The file meta-data contains the following information (*login_time*, *logout_time*, *user_name*, *rated_movie*, *rated_point*, *completed_movie*, *incompleted_movie*, *searched_movie*, *purchased_movie, …*). The log file sample is called "Movie_Log".

For example, two rows in the log file are:

login_time	**logout_time**	**user_name**	**rated_ movie**	**rated_ point**	**completed _movie**	**incompleted _movie**	**searched_ movie**	**purchased _movie**	**…**
09:15:20 6 Nov 2019	16:27:53 6 Nov 2019	John_Smith	Ground hog Day	4	Doctor Sleep	Inception	Blade Runner	The Avengers	…
08:26:11 7 Nov 2019	11:22:42 7 Nov 2019	Peter_Kety	The Godfat her	5	Raging Bull	Null	Citizen Kane	The Wizard of Oz	…

Use Pig Latin script to answer the following questions:

1. How to load the file and print results in Pig Latin script?
2. How to get the list of movies grouped by ratings.
3. What is the primitive "DESCRIBE" in Pig Latin?
4. Extend the previous script to process the clickstream data into user sessions.
5. How can I use FOREACH statement in Pig Script?

6. How to get the top-rated movies in each group.
7. Select only the clicks, which correspond to starting, browsing, completing, or purchasing movies.

Exercise 2

Let "students.csv" is a file that contains students' data. We assume that the data values are separated by "comma".

1. Describe the structure of students file in Pig Latin script.
2. Write a pig script to assign names to the data fields of the students.csv data. The output file should be called "students_details".

Assume that we have another file recording the students' attendance: "students_attendance.csv".

3. Perform any 3 operations on the file "students_attendance.csv". The output of the 3rd operation in this case should be called "SA_details".
4. Write a script to find the sum of hours attended by each student.
5. Write a script to join StudentID, Name, hours attended.
6. Print the results on the screen.

Submission Instructions

All submissions must be done via **Brightspace** with deadline: **8 November 2021, 23:55**. Your submission should consist of one file (PDF), which contains the answers, to the above questions. The submitted file should be named following the format:

COMP*xxxxx_Surname_FirstName_StudentNo_* ***Pracitcal****06*. pdf Example: *(**COMP**47780_Smith_John_12345 _**Practical**06.pdf)*