CS6847 - Cloud Computing

Assignment 3

Note - Submission by Google Classroom.

Problem Description

You need to setup a spark cluster and run a few programs on the given dataset. The objective is to understand the integration of HDFS with Spark programming framework and how to use it for Big Data processing and Data Engineering.

- Setup a spark 1 node cluster on your machines along with HDFS.
- Put your datasets into the HDFS.
- Run the Spark ALS algorithm provided by Spark MLlib on "ALS.txt" dataset. The data set contains three columns (user_id, movie_id, movie_rating).
- Run the Frequent Pair Matching *FPgrowth* algorithm provided by Spark MLlib on "FP_Part-1.csv".
- Run *FPgrowth* on "FP_Part-2.csv". The data of "FP_Part-2.csv" is not the default format. You must write a spark program to extract the data in required format for *FPgrowth* and save it in a different file (**formatted.csv**) in HDFS. Then, "formatted.csv" will be your input for *FPgrowth*.

Evaluation

- For *FPgrowth* give first five frequently occurring pairs for both the datasets as "FP_out1.txt", "FP_out2.txt", with relevant support and confidence parameters in the report.
- Give the ALS output in "ALS_out.txt", fine tune the number of iterations and regularization parameter to minimize the RMS error. The default program divides the data set into 80:20 for training and testing, Experiment by changing this ratio also (min test data size 10%) to minimize error. Provide the error rate for all the experiments in report.
- Screenshots of spark setup need to be included in the report.
- Three different programs files for ALS, FPgrowth, Data Cleaning are required.
- Efficiency of your Data Cleaning program will be considered while evaluation.

Submission guidelines

- Submit the source code of Spark program for the assignment. All other supporting files logs, etc. should also be placed in the zip file (Roll_number.zip).
- Submit a README file containing the necessary details for running your program.
- Create a report explaining your setup and results.

Academic Honesty

WARNING ABOUT ACADEMIC DISHONESTY: Do not share your work with anyone else. The work YOU submit SHOULD be the result of YOUR efforts.