Indian Institute of Technology, Jodhpur

CSL7090: Software and Data Engineering

AY 2022-23 Semester II

Assignment No. 3

M.M.: 30

General Instructions:

1. **Plagiarized submissions will not be evaluated or awarded with an 'F' grade in the course.**
2. Create the report and upload it to google-classroom.
3. Format for the Report:
   (i) Youtube link of benchmarking: create video while performing benchmarking, upload it on youtube and
   mention the link in the report, make sure the link is accessible.
   (ii) Technical Specifications of your system/laptop, Including L1 and L2 cache size.
   (iii) Mention metadata of the benchmarking data in tabular and graphical format.

Task :

a) Create a large Dataset (>=10GB) of your choice. [Ref: 2] [5]

b) Analyze and perform queries on the selected dataset using pandas and spark library. Which library you find better? Justify your answer. [Ref 1,2] [10]

c) Perform benchmarking on the selected data (parameters should be like execution_time, Accuracy, Scalability etc.). Use the preferred library(pandas or spark) to measure at least 5 parameters while benchmarking and analyzing them with your justification. [15]

Reference:

1. https://www.databricks.com/blog/2018/05/03/benchmarking-apache-spark-on-a-single-node-machine.html
2. https://courses.cs.washington.edu/courses/csed516/20au/projects/p03.pdf