DSC 102: Systems for Scalable Analytics

Programming Assignment 2: Grading Scheme

Programming Correctness (100)

For each task, we will run several tests on it with our hidden datasets. Your code must pass all the tests to be counted pass for the task.

Task No.	Task Description	Score (pass/fail)
1	Combine tables and group-by aggregations	15/0
2	Flatten schema and handle array and map type	10/0
3	Flatten schema and conduct self-joins	20/0
4	Typecasting and data imputation	10/0
5	Apply word2vec on string data	10/0
6	One-hot encoding and PCA on categorical data	15/0
7	Train a decision tree regression model	5/0
8	Hyperparameter tuning for the decision tree regression model	15/0

Timeout: each submission (all eight tasks put together, running sequentially) will be given at most 60 minutes to finish on a four-worker cluster. Your score will be the total scores of tasks passed before timeout. We will also deduct points if it takes longer as per the following.

- Between 60 min and 90 min: -10 points
- Over 90 min: -30 points

For instance, if your code took 75 min to run and during the first 60 min it passed Task 1-4, your final score will be 45 = 15 (Task 1) + 10 (Task 2) + 20 (Task 3) +10 (Task 4) - 10 (Overtime penalty).

If any task fails you might still get partial credits for the task.

Extra Credit (10)

Your code will be timed with all tasks together. If it manages to pass all the tests for Programming Correctness, you may receive extra credits as showed in the following table according to the runtime.

Runtime t	Credits
$t < 15$ min	10
15 min $\leq t <$ 30 min	6
30 min $\leq t <$ 45 min	3
45 min $\leq t$	0