# 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: We will run your code (all eight tasks put together running one after another) on a three-worker cluster and impose a timeout of 2 hours and will kill your process at that point. Only tasks that finish within the timeout will be considered for the correctness portion. If any of the tasks fails, you will still get partial credits for the other tasks. 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 it only passed Tasks 1-4, your final score will be 45 = 15 (Task 1) + 10 (Task 2) + 20 (Task 3) +10 (Task 4) - 10 (Overtime penalty).

## 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 |
|-------------|---------|
| $t \leq 10$ min | 10 |
| $10$ min $< t \leq 15$ min | 5 |