CS5481: Data Engineering - Assignment 2

October 8, 2025

Instructions:

- Due on Wednesday, Nov. 12, 2025.
- You can submit your answers by **a single PDF with the code and the output files** or **a jupyter notebook with output files** containing both the answers and the code.
- For coding questions, besides the code, you are encouraged to give some descriptions of your code design and its workflow. Detailed analysis of the experimental results is also preferred.

Question 1 - LLM for Data Engineering

(20 marks) LLMs' fast and articulate answers to expert questions can help data engineers discover datasets, write and debug code, document procedures, and learn new techniques as they build data pipelines. In this question, you are required to write suitable prompts for ChatGPT to achieve the following targets.

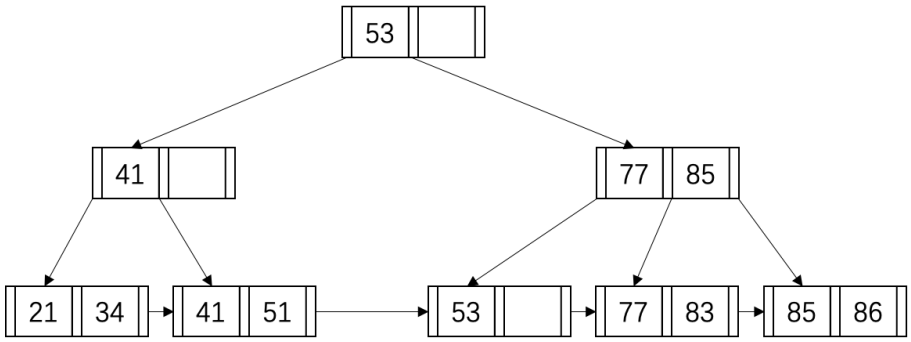
(a) (5 marks) Assume you need datasets to train a recommender system that predicts user preferences for products. Use ChatGPT to generate a synthetic dataset for training a recommender system and visualize it. List the prompts you used and the corresponding outputs from ChatGPT.

(b) (5 marks) Use ChatGPT to preprocess a sample from the **MovieLens-1M** dataset, including handling missing values and encoding categorical variables. List the prompts, inputs, and outputs from ChatGPT.

(c) (10 marks) Generate a preprocessing script using ChatGPT to format the dataset for a Collaborative Filtering model using Surprise. Correct any errors in the generated code, use the revised version to preprocess the dataset, and display the first 5 processed entries.

Question 2 - Data Indexing

(25 marks) Given the following B+ tree, please answer following questions.



(a) (5 marks) What is the value of p for this B+ tree? (Note that p is the order of a B+ tree)

(b) (6 marks) Can you re-build a taller B+ tree with the same value of p using the same set of search-key values in the leaf nodes of the given tree? If yes, show the steps by drawing a new diagram whenever the height of the tree increases.

(c) (6 marks) Insert the search-key values 32, 84, and 19 in sequence to the given B+ tree, and draw a new diagram for each insertion.

(d) (8 marks) Suggest a sequence of search-key values to be deleted from the resultant B+ tree in **(b)** to shrink the tree to 2 levels with the least number of deletions. Show the steps by drawing a new diagram whenever a node is deleted.

Question 3 - Data Querying

(25 marks) The university held a coding contest where hackers submit solutions to various tasks. Each task has a bonus for the top 3 performers. You are given the following SQL tables:

- Hackers (hacker_id: INT, name: VARCHAR, bank_account: INT)
- Tasks (task_id: INT, description: VARCHAR, bonus: INT)
- Submissions (submission_id: INT, hacker_id: INT, task_id: INT, score: INT, submission_date: DATE)

Assume that

- Each task has a bonus for the top 3 submissions with the highest scores.
- If there are multiple submissions with the same score, the earliest submission (lower submission_id) is preferred.
- Hackers can submit multiple times, but only their best submission (highest score) counts for each task.

(a) (5 marks) Write a query to print the hacker_id, name, and the number of distinct tasks each hacker participated in, but **only** for hackers who participated in more tasks than the average number of tasks per hacker. Sort the result by the number of tasks in descending order, and then by hacker_id in ascending order if there's a tie.

(b) (5 marks) Write a query to find the task_id, description, and the total bonus awarded for each task. Sort the result by task_id in ascending order.

(c) (5 marks) Write a query using nested subqueries to list the submission_id, hacker_id, name, and score of the highest-scoring submission for each task submitted on 2023-01-01. If multiple submissions have the highest score for the same task, return the submission with the smallest submission_id. Sort the result by task_id in ascending order.

(d) (5 marks) Write a query to print the hacker_id, name, the total score each hacker achieved across all tasks, the number of tasks they participated in, and their average score per task. For each task, only the hacker's best score counts. Sort the result by total_score in descending order and by hacker_id in ascending order if there's a tie.

(e) (5 marks) Write a query to find the hacker_id, name, and bank_account of hackers who did not participate in any tasks.

Question 4 - Recommender System

(30 marks)

(a) (8 marks) Please write and briefly explain two basic approaches for recommender system.

(b) (8 marks) One common challenge in recommender systems is the cold start problem. Explain what the cold start problem is and how it affects recommendation quality. Suggest at least two strategies to mitigate the cold start issue.

(c) (14 marks) Top-N recommendation is an important task for recommender systems. Try to implement a recommendation model using the Goodbooks-10k dataset to generate top-**10** book recommendations. You have two options based on your compute power:

- Option 1 (Full Dataset): For those who can access sufficient compute power, they are recommended to use the entire dataset (6 million ratings for 10,000 books by 53,000 users).
- Option 2 (Partial Dataset): For those who can only access very limited compute power, use a subset of the data, for example, the top 5,000 users and top 3,000 books based on the number of interactions (ratings). Clearly indicate that you are using the subset due to resource constraints in your submission if you choose this option.

Here are also some tips to follow:

- You can implement the model using user-based collaborative filtering, item-based collaborative filtering, or matrix factorization.
- Split your dataset into a training set and a test set.
- Display the Hit Rate and F1 score on the test set for your top-10 recommendations.