FINAL EXAM, 839 SPRING 2023

Write down your name here (last name first) AND student ID number:

Please wait until you are told to start working on the exam.

75 minutes, closed books and notes, no electronics, cheat sheet allowed.

If a question is ambiguous, please state your assumptions and then answer based on those assumptions. Please keep the assumptions reasonable.

You can only use the space provided in this exam. Adding more space is not necessary and may result in points being deducted.

Some questions may be harder than others. You may want to handle easier questions first, delaying the more difficult questions to later.

10 Questions, 10 Points Each, for a Total of 100 Points.

1) Consider matching two tables A and B, where each tuple describes a person. Our goal is to find tuples a in A and b in B such that a and b refer to the same real-world entity. On the surface this looks like a well-defined goal. People however often say that entity matching such as the above problem is actually subjective. What do you think they mean by this, and can you give a small example to illustrate?

2) Suppose you want to do blocking on "city", that is, if two tuples do not agree on "city", then the pair consisting of these tuples does not survive the blocking step. What happens if you have a lot of missing values in "city"? What would you do?

3) Describe two scenarios where the Sparkly blocking solution may not achieve high recall.

4) Is schema matching transitive? That is, if we have determined that Column X of Table A matches Column Y of Table B, and that Column Y of Table B matches Column Z of Table C, can we conclude that Column X of Table A matches Column Z of Table C?

5) Consider a data lake of 1000 tables. Suppose we want to match the schemas of these tables. Explain how you can reformulate this problem as an entity matching problem. Recall that entity matching is often solved with a blocking step followed by a matching step. Suggest a method to use for the blocking step. Suggest a method to use for the matching step. Do you think a solution such as Ditto would be effective in the matching step (for this problem)?

6) Explain and contrast OLTP and OLAP queries.

7) Explain and contrast data warehouse and data lake.

8) Give a small example of a star schema for a data warehouse. Name the fact table, the dimension tables, and the dimension attributes. This question only asks for the schema, you don't need to list data instances. You can describe the schema of a table like this: Emp(eid, name, salary).

9) Define data governance and give one example of a data governance problem.

10) Briefly explain the idea behind data mesh. Suppose a company wants to implement a data mesh. Discuss at least one major challenge that you think the company will face.