1. Design

City(City, County)
County(County, Country)
Country(Country, Continent)

and

Location(City, County, Country, Continent)

a. Describe the pros and cons of these two approaches.

b. What is a materialized view? Compare/contrast with ROLAP.

c. Consider that our customers routinely compute a complex tax calculation over the tuples in a database. Compare/contrast the use of stored procedures versus client side functions.
2. Temporal databases

Consider a booking system for three sports fields. Members are allowed to book fields for one hour blocks. Members may cancel only if they do so 5 hours before the start time of their booking.

a. Design a schema that can handle that can model this problem.

b. For your schema, in SQL or Datalog, write the following queries:
   * “Give the current booking of field 3 from 3-4 pm on June 18th.”
   * “Give the people who have canceled a field.”
   * “Give the members who have booked two fields at the same time and then canceled one of them less that 7 hours before their booking started.”
3. Multi-dimensional data models

On the temporal database design in problem 2, design a ROLAP system that allows for trend analysis on bookings. (e.g. "show the number of bookings per week throughout the 3 months of the Fall"). Assume that we have gender and age data on members.

Discuss how you would populate this system from your schema/database of problem 2.
4. Datalog queries

Here we model water supplies and their contamination. Bodies of water have a given volume and an amount that rains or springs into them from the ground each year. Bodies of water flow into each other as well, placing a fraction of their total volume ‘down stream’ every year. Unfortunately, there is an amount of contamination (fully dissolved) that was initially present in bodies of waters at time zero.

Given this, we may have the following situation:

BodyOfWater(1,’Pond1’,10, 10).
BodyOfWater(2,’Lake2’,100, 5).
BodyOfWater(3,’Lake3’,200, 10).
BodyOfWater(4,’River4’,50, 0).
BodyOfWater(5,’Lake5’,500, -25).

FlowsInto(1, 2, 1.0).
FlowsInto(2, 3, 0.5).
FlowsInto(3, 4, 0.5).
FlowsInto(4, 5, 1.0).

Contaminated(1, 1000).
Contaminated(2, 100).

So here for example, we have ‘Lake2’ being a volume of 100, being fed by an underground spring 5 units per year. 50% of the water in ‘Lake2’ flows into ‘Lake3’ each year. ‘Lake 1’ has 1000 units of contamination at time zero.

a. Define the predicate WaterReaches(id1,id2) that expresses when water from one body (ultimately) flows into the other.

b. Define the predicate TotalContamination that gives the total amount of contaminant in the environment (1100 in the example above)
– Write an LDL program that fully models the diffusion process. The predicate \textit{AmountOfContamination(id, amount, $year)} is what you need to answer queries over. (This should require at least 7-8 rules).
5. DTD/XML

Give a DTD and a corresponding XML document for the schema/data of problem 4.
6. Association rule mining

You are given the following transactions

1. \{a, d, e\}
2. \{a, d\}
3. \{b, e\}
4. \{b, c, e\}
5. \{a, d, e\}
6. \{a, d, e\}
7. \{a, e\}
8. \{b, c\}

Using the A priori algorithm find the associations rules with support above \(1/4\) (that is \(\beta > 1/4\)) and confidence over \(2/3\) (that is \(\alpha > 2/3\)). Start with calculating large item sets. Indicate which item sets have counts calculated over the database.

a. Calculate large item sets:

Pass 1

Pass 2

Pass 3

Pass 4

Pass 5

b. Gives rules with sufficient confidence and support.