CSCI585 Fall '17 Midterm Exam Solution & Rubrics

October 13th, 2017

CLOSED book and notes. No electronic devices. DO YOUR OWN WORK. Duration: 1 hour 50 minutes. If you are discovered to have cheated in any manner, you will get a 0 and be reported to SJACS. If you continue working on the exam after time is up you will get a 0.

Signature: _____

Problem Set	Number of Points	Your Score
Q1	5	
Q2	6	
Q3	4	
Q4	4	
Q5	6	
Q6	5	
BONUS	**1**	
Total	**30**	

Q1. (5 points total) ER Modeling

a. (1 point) Describe the difference between weak and strong relationship (according to the definition).

Answer:

A weak relationship is a relationship in which the primary keys of the related entities are independent of each other.

A strong relationship is a relationship in which a primary key of one table is part of primary key of the other table.

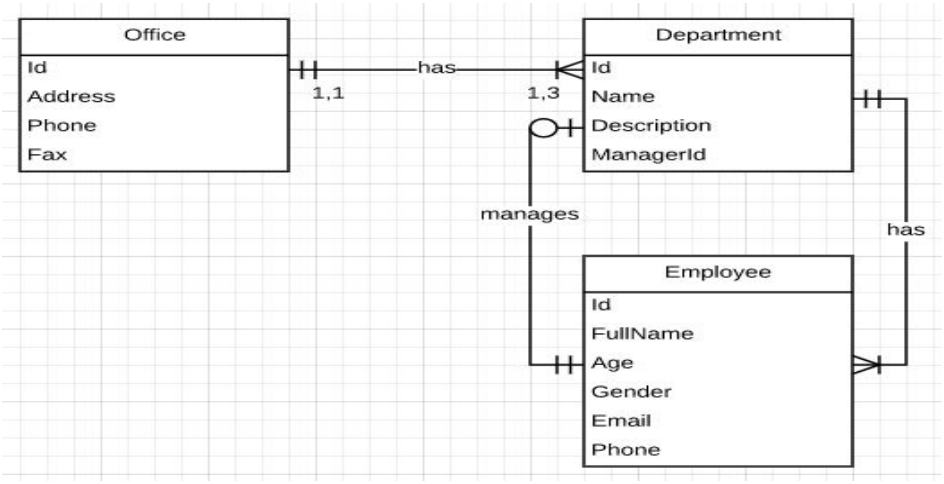
Grading Rubric : Definition must contain primary key relation mentioned in the actual answer.

b. (4 points) Draw ER Diagram based on the following description:

A company has multiple departments. Each department is located in a single office, but each office may host up to three departments.

Each department has multiple employees, but an employee can only work in one department. Each department has a manager who is also an employee.

Answer:



Grading Rubric : Each of the entities must have some attributes.

In Manager part of the problem, it's essential to have a recursive relationship between employee and department.

No specialization hierarchy required.
Any assumptions must be stated in the solution.

Q2. (6 points total) SQL

a. (1 point) Which relational operation does the following SQL query implement?

```
SELECT name
FROM driver
WHERE vehicle IN (SELECT vehicle FROM vehicles)
GROUP BY name
HAVING COUNT(*) = ( SELECT COUNT (*) FROM vehicles);
```

Answer: Division

Grading Rubric: No Partial point. 1 point for if you specifically mention 'DIVISION'

b. (2 points) Given the following enrollment table, write a SQL query to list number of students enrolled in each course (ClassID).

StudentID	ClassID	Grade
321	CSCI495	B
564	CSCI110	C
321	CSCI585	A
564	CSCI495	A
789	EE101	F
321	EE101	B

Answer:

```
SELECT ClassID, COUNT(*) AS Total
FROM enrollment
GROUP BY ClassID;
```

Grading Rubric: 2 points if your query performs the exact operation as the above solution. -1 for any trivial syntax mistakes.

c. (1 point) Explain the difference between left and right outer join.

Answer: Left outer join would include remaining (unmatched) rows from the table on the left and fill values of columns on the right with nulls for those rows. Right outer join would include the opposite: unmatched row from the table on the right and fill out values on the left with nulls.

Grading Rubric: 1 mark has been deducted if you didn't specify that there will be null or empty values in the rows which don't match in left/ right outer join. Specifying about the NULL values in unmatched rows is the requirement of the answer. Points has been awarded if 'null' values are mentioned in the d) part of the question or shown in the example.

d. (2 points) What would the query return when VENDOR and PRODUCT tables from lecture slides are joined using left outer join? When using right outer join?

Answer: Left outer join would include vendors who haven't sold any products, right outer join would include products that aren't sold by vendor (in addition to vendor list and the products each vendor is selling).

Grading Rubric: Points has been given if the explanation is correct OR the difference is shown through the means of tables or examples. -1 if you didn't specify what data from both tables will be included in the respective join table and just defined what left and right outer joins will do on both tables. The answer should give the sense that Left outer join would include vendors who haven't sold any products, right outer join would include products that aren't sold by vendor (in addition to vendor list and the products each vendor is selling) else 1 mark has been deducted.

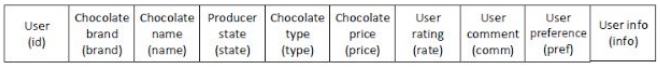
Q3. (4 points total) NORMALIZATION

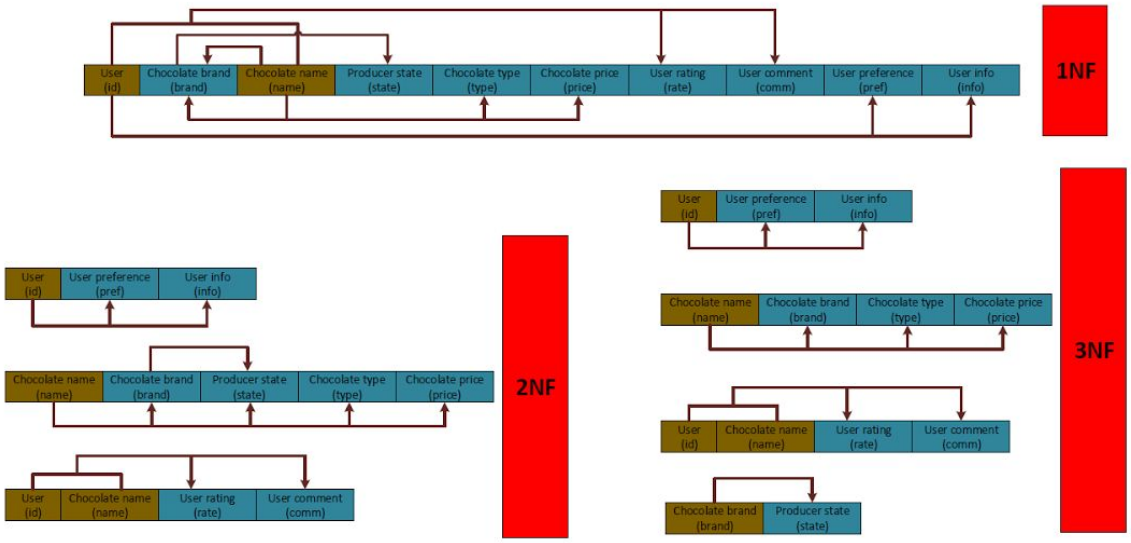
We are starting to design a new chocolate rating platform and have asked experts to taste different sweet products (sweets) and express their overall rating (0-5) and their professional comments. Each expert taster (ID) may try different products from different companies. We are only focused on US based companies and record the state of the headquarters of the company in our reports. Each product has a specific type, e.g. confection, candy, bar, etc. and also a unique price. Even though each taster can rate different types of products, we only record their top preference type for further reference. Moreover, basic personal information of each taster is available in our records. Our sample data is in the following table:

ID	Brand	Name	State	Type	Price	Rate	Comm	Pref	Info
29001	Hershey	Reese's	PA	Confection	1.29	3	Good peanut butter!	Bar	Michael Mast
29001	Hershey	NutRageous	PA	Bar	1.39	4	Packaging isn't great	Bar	Michael Mast
29001	Hershey	Pieces	PA	Candy	0.99	2	Worst candy ever	Bar	Michael Mast
29001	Hershey	York	PA	Confection	0.99	3	Too sweet	Bar	Michael Mast
29001	Mars	Snickers	WA	Bar	1.45	3	Not a fan of almonds	Bar	Michael Mast
29001	Mars	Twix	WA	Bar	1.45	5	Best product ever	Bar	Michael Mast
29001	Mars	M&M	WA	Candy	1.79	3	The coating is too hard	Bar	Michael Mast
1202	Hershey	York	PA	Confection	0.99	3	Just a failed cop of PBCs	Candy	Paul Bulcke
1202	Hershey	Snack Barz	PA	Bar	0.89	3	Too hard	Candy	Paul Bulcke
1202	Mars	3Musketeers	WA	Bar	1.69	2	Best quality	Candy	Paul Bulcke
1202	Mars	Bounty	WA	Bar	1.55	3	Too much coconut	Candy	Paul Bulcke
1202	Mars	Twix	WA	Bar	1.45	3	Too much nuts	Candy	Paul Bulcke
1202	Mars	M&M	WA	Candy	1.79	5	The best taste ever	Candy	Paul Bulcke

1202	Hershey	Kisses	PA	Chips	0.23	2	Poor quality chocolate	Candy	Paul Bulcke

a. (2 points) Identify dependencies and draw dependency diagram.
b. (2 points) Normalize tables. Show all steps and resulting tables in 3NF.
Answer:





Grading Rubric: Note: Taking (ID, Brand, Name) as the primary key is also an acceptable solution (As long as the normalization process and final normal forms correlate with this assumption).

Part a)

2 points: All (and only) correct dependencies included in the dependency diagram.

1 point: 4 or more correct dependencies in the dependency diagram.

0 points: Less than 4 correct dependencies in the dependency diagram.

Part b)

2 points: Both 3NF and 2NF are correct.

1 point: Either 3NF or 2NF not shown / Either 2NF or 3NF is incorrect / Process of normalization is correct, but the final tables were incorrect due to incorrect dependencies in part a / Process of normalization did not match the dependency diagram but the final tables match the correct solution / Some part of normalization process does not match an assumption made in the dependency diagram.

0 points: Neither 3NF nor 2NF are correct.

Q4. (4 points) TRANSACTION MANAGEMENT

a. (2 points) What does 'A' stand for in 'ACID'? Describe this property using its definition or example.

Answer: Atomicity, all parts of each transaction must complete or the transaction needs to be rolled back.

Transfer from checking to savings contains 4 parts: start, update to balance on checking, update to balance on savings, and commit.

Grading Rubric: 1 mark for "Atomicity" 0 for "Availability"/other/no answer. 1 mark for " All or none", "rollback if interrupted" 0 for No answer or explained deduction in comment.

b. (2 points) Use transaction log below to recover database assuming the system crashed right before TR_ID 365 Commit executed. Show recovery steps in order.

TRL_ID	TRX_NUM	Table	ROW ID	Attribute	BEFORE VALUE	AFTER VALUE
341	101	****Start Transaction				
352	101	PRODUCT	1558-QW1	PROD_QOH	25	23
363	101	CUSTOMER	10001	CUST_BALANCE	525.75	615.73
365	101	****Commit				

Answer: Logs are read from bottom up.

1. Update CUST_BALANCE value for customer row ID 10001 back to original balance $525.75.
2. Update PRODUCT value for product row ID 1558-QW1 back to original quantity 25.

Grading Rubric: 1 for correct balance =525.75 AND correct quantity=25, 1 for correct order of recovery i.e. first balance is recovered then quantity. 0 otherwise. No Partial Marking.

Q5. (6 points) OPTIMIZATION

a. (1 point) What is the result of the following query (what does it do)?

```
SELECT id, name
FROM viterbi_Students
WHERE branch = 'Computer Science' AND courseTaken = 'CS 585';
```

Answer: Lists all Viterbi Computer Science students who have taken CS 585.

Grading Rubric: 1 mark for correct explanation. No partial points.

a. (2 points) Optimize the query above and justify your answer.

Answer:

```
SELECT id, name
FROM viterbi_Students
WHERE courseTaken = 'CS 585' AND branch = 'Computer Science';
```

Assuming the number of students taking CS 585 would be much lesser than the number of students who are enrolled in Computer Science. For AND we should write the condition that is more likely to be false first.

Grading Rubric: 1 mark for correct query. 1 mark for correct explanation. No points for explanation if query is wrong. No points for solutions related to indexing.

c. (1 point) What is the likely result of the following query (what does it do)? Model refers to a model of a cell phone.

```
SELECT model.modelCompany, AVG(model.modelPrice *
company.ratingCompany)
FROM model INNER JOIN company ON model.modelCompany = company.name
WHERE model.modelPrice > 400
GROUP BY model.modelCompany;
```

Answer: Lists cell phone model and avg model price value weighted by company rating for models that are more expensive than $400.

Grading Rubric: 1 point for correct explanation. No partial points.

d. (2 point) Approximately 20000 cell phones are added to the inventory weekly. How would you optimize this query? Explain your answer.

Answer:

Consider having an attribute in model which stores model.modelPrice * company.ratingCompany.

```
SELECT modelCompany, AVG(priceBase)
FROM model
WHERE modelPrice > 400
GROUP BY modelCompany;
```

Grading Rubric: 2 points for correct explanation of how we can optimize it. -1 for solutions that pre-calculate average completely as it requires update for each company and models in that company > 400. No points for solutions related to indexing, creating trigger, optimizing join, updating the entire query to something new or "suggesting we don't calculate modelPrice * ratingCompany" as its expensive. This has been covered with lots of examples in class and hence no partial grading for the cases mentioned above.

Q6. (5 points) DISTRIBUTED DATABASES

A large bank makes many investments on annual basis. They use PROJECT table below to keep track of the investments. The accountants check how much money was invested every year in all projects. They only care about annual spending (dollar signs $$$ per year) and are not interested in any project related data. Recently, the bank is considering changing to a distributed database to store their investment data and has hired you to help them.

PROJECT

Project_id	PName	Budget	Location	Manager

a. (2 points) Which data fragmentation technique would you recommend they use to meet the requirement of the financial department? Show new design.

Answer: Vertical fragmentation.

Project_id	Budget

Project_id	PName	Location	Manager

Grading Rubric:

1 point for vertical fragmentation (awarded point for explaining the definition exactly if the word is missing). 1 point for correct design as mentioned above.

b. (1 point) If there was an account at each site to manage money for the local projects only, would your recommendation change? Which data fragmentation technique would you recommend in this case?

Answer: Horizontal fragmentation. Each sub-table stores its state's project records (rows).

Grading Rubric: 1 point for horizontal fragmentation (awarded point for explaining the definition exactly if word is missing)

c. (2 points) Describe the steps of the Two Phase Commit protocol.

Answer:

Phase 1: Coordinator asks subordinates if ready to write to log (write ahead protocol), if yes: write, if no or fails: abort.
Phase 2: Coordinator asks subordinates if ready to commit, if yes: commit, if no: abort, if fails: undo.
DO-UNDO-REDO protocol rolls transactions back and forward with the help of the system's transaction log entries.

Grading Rubric:

2 points: Correct explanation

0 points : For students who have written 2 phase locking protocol.

1 mark : If only DO-UNDO-REDO is mentioned with relevant explanation.

BONUS Q. (1 point total) BUSINESS INTELLIGENCE (BI)

Which data schema is widely used in data warehouses for BI? What are the key characteristics of this schema?

Answer:

Star schema. Snowflake schema is also star schema.

It maps multidimensional decision support data (facts) into a relational database. Many-to-one relationship between fact table and each dimension table.