CS585 Final exam

2016-06-23

Duration: 1 hour.

Last Name: ___________________________________

First Name: ___________________________________

Student ID: ___________________________________

Email: ___________________________________

Hi.. there are 11 questions (10 plus a bonus), one per page. Please read each question carefully before answering. You don't to elaborate on anything, so you won't need additional sheets.

The exam is CLOSED book/notes/devices/neighbors(!) but 'open mind' :)

If you are observed cheating, or later discovered to have cheated in any manner, you will get a 0 on the test and also be reported to SJACS.

When we announce that the time is up, you NEED to stop writing immediately, and turn in what you have; if you continue working on the exam, we will not grade it (ie. you will get a 0).

Have fun, and good luck! Hope you do well.

Saty

Q1 (2+1+1=4 points).

a. **In the context of database performance tuning, what is an 'access plan'?**

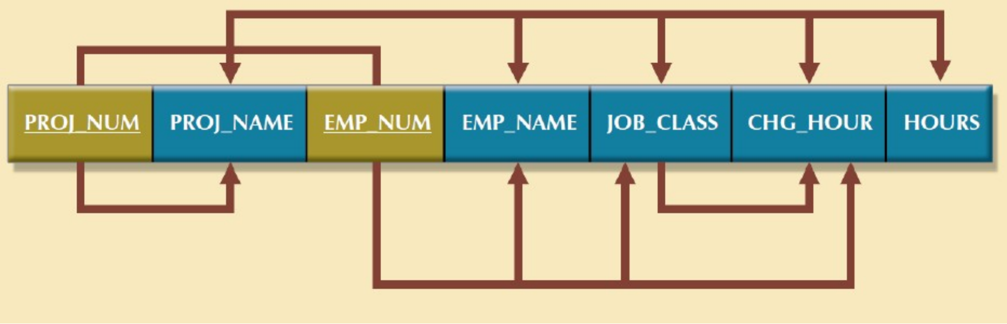
A series of complex, low-level, I/O operations for reading and writing data from/to database tables.

□

b. What are a couple of ways using which a SQL programmer can enhance her queries (make them be executed efficiently)?

Use simple columns or literals as operands; prefer numeric comparisons to char/string ones (and, transform conditionals to use expressions, do equality comparisons instead of inequality....).

Q2 (1+3=4 points). A 1NF table, such as the one shown below (we covered this in class on great detail), is analyzed to detect problems (related to unwanted dependencies), which are then systematically eliminated (the table is converted to 2NF, then 3NF).



a. What is the diagram (shown above) called?

Dependency diagram.

b. How does the diagram aid in normalization? Explain briefly, using the above diagram (you can mark it up if you want).

We use the diagram to first eliminate **partial dependencies** (1NF → 2NF), then we eliminate **transitive dependencies** (2NF → 3NF).

Q3 (4 points). What is the acronym for the creation part of a data warehouse? Briefly describe the steps involved in this.

ETL.

E: Extract – identifying data from multiple sources, gathering (eg using SQL queries), consolidating.

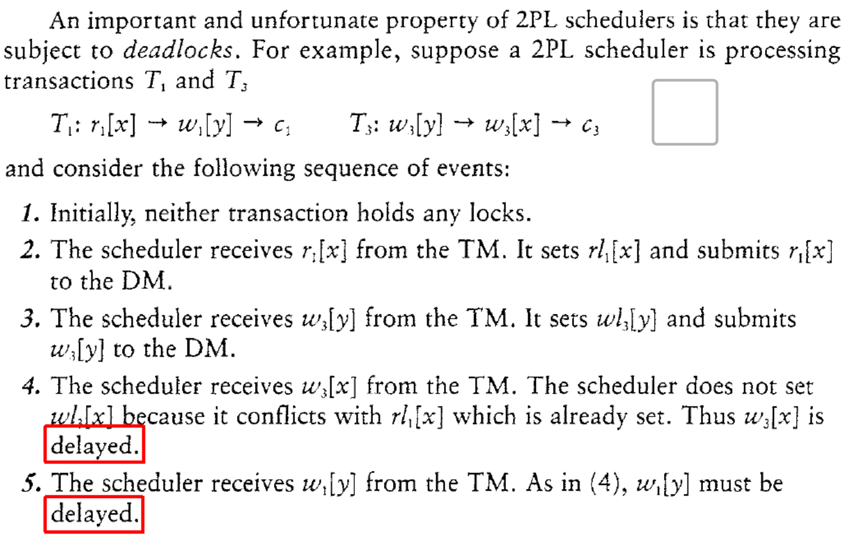
T: Transform – cleaning, filtering, validating, etc. of data.

L: Load – loading (populating) the data into the warehouse tables.

Other ways to explain are OK, as long as they align with the steps above.

Q4 (4 points). Two-phase locking (2PL) is a popular concurrency control scheme for managing transactions. Unfortunately, however, it cannot prevent deadlocks from occurring. Using two transactions T1 and T2, show (using a sequence of events you come up with) how a deadlock can occur between them.

From the notes (simpler cases are OK too):



An important and unfortunate property of 2PL schedulers is that they are subject to *deadlocks*. For example, suppose a 2PL scheduler is processing transactions T_1 and T_3

T_1: $r_1[x] \to w_1[y] \to c_1$ T_3: $w_3[y] \to w_3[x] \to c_3$

and consider the following sequence of events:

1. Initially, neither transaction holds any locks.

2. The scheduler receives $r_1[x]$ from the TM. It sets $rl_1[x]$ and submits $r_1[x]$ to the DM.

3. The scheduler receives $w_3[y]$ from the TM. It sets $wl_3[y]$ and submits $w_3[y]$ to the DM.

4. The scheduler receives $w_3[x]$ from the TM. The scheduler does not set $wl_3[x]$ because it conflicts with $rl_1[x]$ which is already set. Thus $w_3[x]$ is delayed.

5. The scheduler receives $w_1[y]$ from the TM. As in (4), $w_1[y]$ must be delayed.

Q5 (4 points). What role do minimum bounding rectangles (MBRs) play, in spatial query processing (how are they used/helpful)?

Spatial queries are performed using a two step filter+refine process, and MBRs help in the filtering step (which acts as a preprocessor for refining).

Q6 (2 points). 'Ensemble methods' are often used in machine learning - what is the single biggest benefit of using this technique?

To minimize or eliminate any variances or biases between the individual learners in the ensemble.

Q7 (3+1=4 points). Weka and R are both highly capable of helping analyze tabular data.

a. In Weka, what is the file format that is used to read a table? Name the format, and provide a very small example.

ARFF: .arff.

Example:

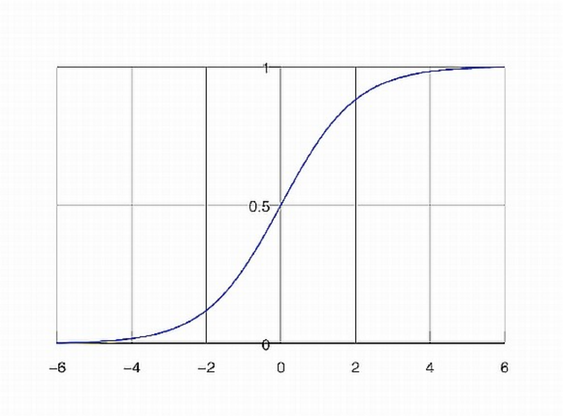
```
@RELATION iris

@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
….
….
```

Alternate examples (eg. with a class attribute) are OK, too. But, there does need to be at least one @ATTRIBUTE specified, and there needs to be a @DATA section that follows it.

b. In R, what is the name of the datatype that can hold a table of data?

Data frame.

Q8. (3+3=6 points). The 'sigmoid' function/curve shown below, is useful in at least two techniques of Machine Learning. What are the two techniques, and briefly, how is the curve used in each?



Neural nets – to determine the output value of a neuron; logistic regression – to classify the incoming (unknown) data point into one of two classes, depending on the sigmoid function's value being <0.5 or >0.5.

Q9 (2+2 = 4 points). Fraudulent credit card purchase detection relies on using a binary (yes/no) classifier to analyze card transactions. Name two algorithms (we covered four) that could be used for this purpose, explain very briefly how each works.

SVM, Neural nets, decision tree, k-means clustering.. Refer to the notes for brief explanations.

Q10 (1*4=4 points). A very straightforward question - name (the) 4 types of NoSQL DBs, and provide an example (an open source or commercial implementation) of each.

Key-value. Dynamo. □

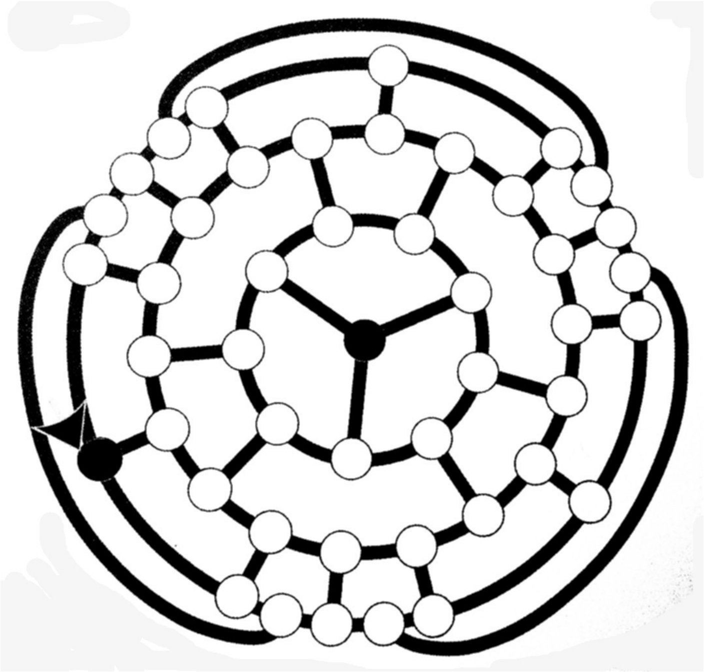
Column Family. Cassandra.

Document. MongoDB.

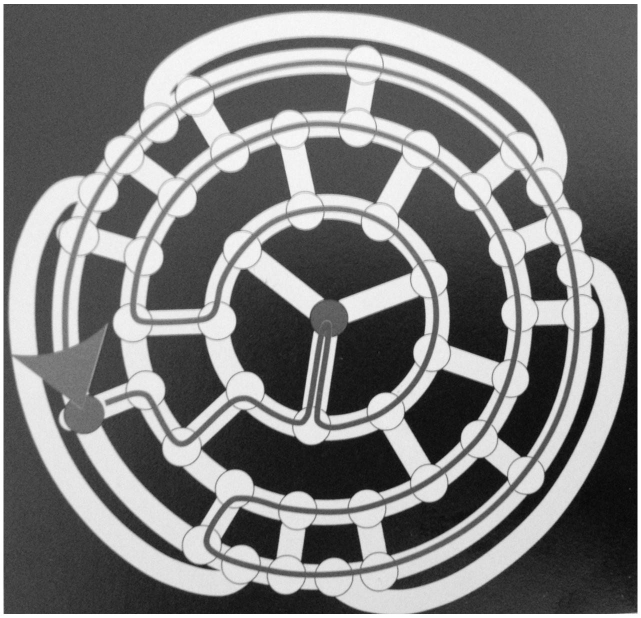
Graph. Giraph.

Several alternates are acceptable, for the example DBs.

Bonus question (1 point). In the diagram shown below, trace a path that starts and ends at the arrowed circle. The path needs to cover (pass through) every circle, which means it needs to cover every link (links are the lines that connect adjacent circles) – you are allowed to backtrack (redraw over) just ONE link.



No points for this, but another question – in graph theory, what is such a path called?



Hamiltonian circuit/cycle.