CSCI585 Summer ’18 Final Exam

June 26th, 2018

CLOSED book and notes. No electronic devices. DO YOUR OWN WORK. Duration: 2 hours. 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.

Solutions are marked in red font.

Signature: ______________________________

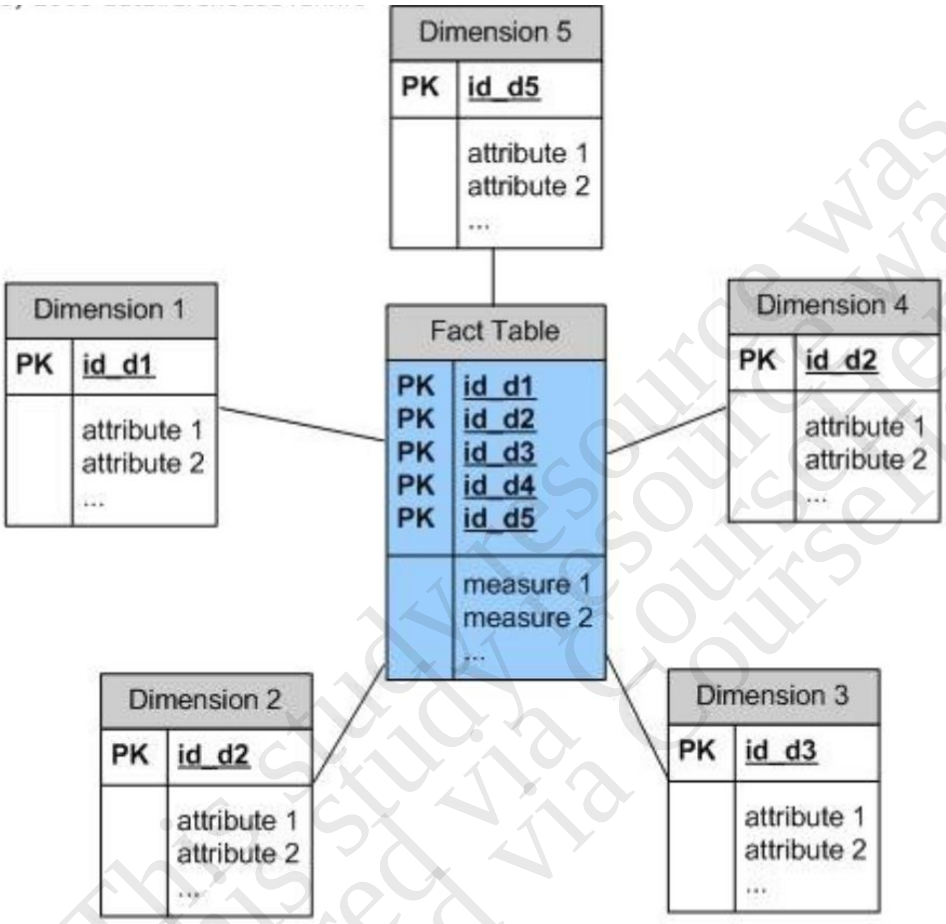
Problem Set	Number of Points
Q1	5
Q2	5
Q3	5
Q4	5
Q5	5
Q6	5
Q7	5
Total	**35**

Q1. (5 points total) Business Intelligence

A (2 points) What aspects does decision support data differ from operational data in?

Time span, granularity, and dimensionality. (2 points if two out of three are listed).

B (1 point) What kind of schema does the ER diagram demonstrate?



Star schema.

C (2 points) In OLAP, what are two kinds of SQL extensions for generating the aggregated dimensional data?

ROLLUP and CUBE extensions.

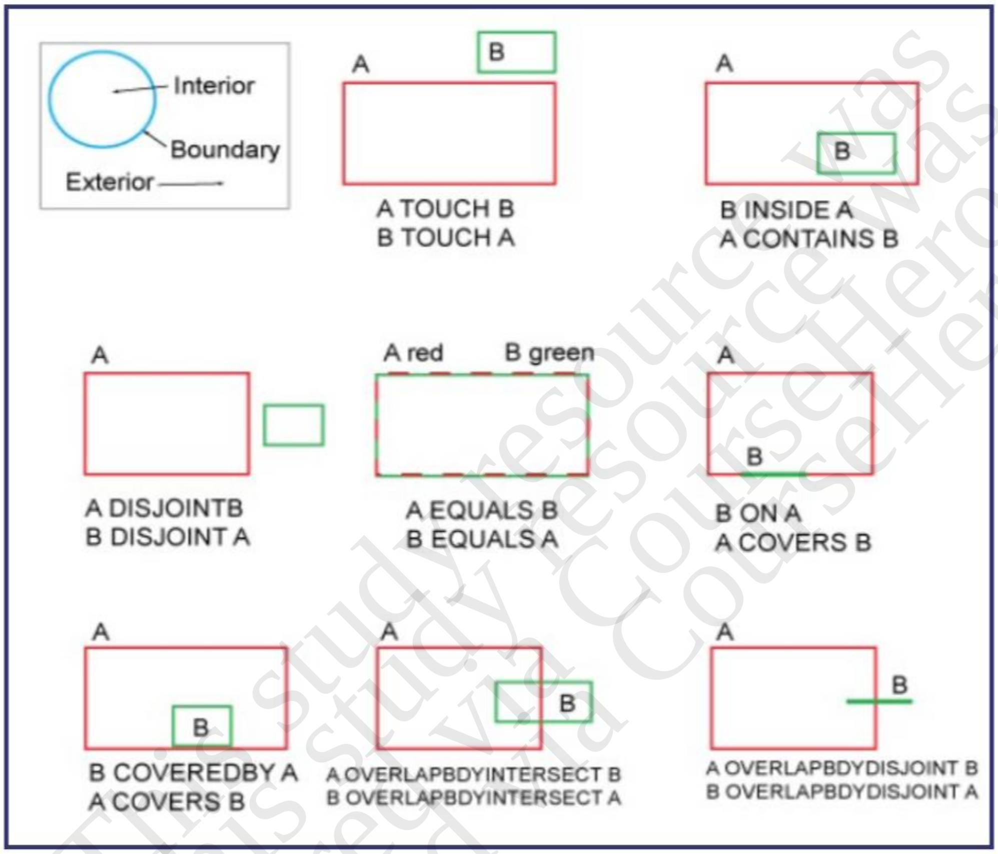
Q2. (5 points total) Spatial Databases

A. (3 points) Spatial databases store the data about the objects in spaces. What are those objects modeled in the spatial databases?

Points, lines, and polygons.

B. (2 points) Draw and explain two kinds of spatial relationships in 2D environment

Any two from the figure below.



Q3. (5 points total) NoSQL

A. (4 points) What are the differences between key-value and document databases? Give an example database of each type.

Key-value database:

- Store data as key-value pair
- Only support index on key
- memcached, Redis

Document databases:

- Store data as document. Each document has a multiple fields.
- Support secondary indexes.
- MongoDB.

B. (1 point) What is NewSQL database? How is it different than NoSQL?

NewSQL databases are databases that provide ACID compliant transactions across a highly distributed infrastructure. It is different than NoSQL which provides BASE properties.

This study resource was shared via CourseHero.com

Q4. (5 points) MapReduce

A. (3 points) Briefly explain how MapReduce works with the problem “Count the number of occurrences of each word in a large collection of documents.”

Answer:

“This is a sample. This sample is a sample.”

Map phase: Documents are partitioned into Map tasks.

Given a set of words, a Map task outputs the number of occurrences of each word.

(“this”, 1), (“is”, 1), (“a”, 1), (“sample”, 1)

(“this”, 1), (“sample”, 1), (“is”, 1), (“a”, 1), (“sample”, 1)

Reduce phase: Reduce tasks pull the Map tasks outputs to process.

Given a set of words and their occurrences from Map task, aggregate to produce the final result.

(“this”, 2), (“is”, 2”), (“a”, 1), (“sample”, 3)

B (2 points) Assume there are M Map tasks and R Reduce tasks. How many outputs Map phase produces? How many outputs Reduce phase produces?

Answer:

MxR

R

Q5. (5 points) Big Data / Data Science Intro

A. (3 points) What factors or situations make the Big data be so "big" now?

1) So many data sources: social network, web browsing history

2) The ability to store and compute this data in the "cloud" is virtually unlimited

3) Based on Hadoop/Mapreduce, we can do process it efficiently.

The answers will vary and are expected to correspond to three V's: volume, velocity, variety of data. (1 point per example).

B. (2 points) Usually, data mining algorithms fit into 4 categories: classification, clustering, regression, and rule extraction. List which category or categories are supervised and unsupervised algorithms.

Supervised: classification, regression

Unsupervised: clustering, rule extraction

(1 point for supervised and 1 point for unsupervised.)
If logical explanation as to why rule extraction could be considered supervised and everything else is correct, accept answer.

Q6. (5 points) Machine Learning

A (3 points) List three examples of applications of neural networks (problems which can be solved using NN).

Answers will vary. (1 point per correct example)

Here are some examples:

Speech Recognition

Detect anomalies

Find objects from images.

B (2 points) Given the descriptions of items for sale, a company wants to use machine learning to automatically extract properties of items from the descriptions.

For example, given the description:
“Iphone 7 Silver 32 GB for sale! Never used. Price is $649. Shipping is free”.

One desired result could be:
(“Category”, “Smartphone”), (“Model, “Iphone 7”), (“Brand”, “Apple”), (“Color”: “Silver”), (“Storage”, “32 GB”), (“Condition”, “New”), (“Price”: “$649”), (“Shipping”, “Free”)

If the company hires you, describe how would you apply machine learning to solve this problem?

Answers will vary. (1 point for acceptable technique + 1 point for detailed explanation)

Q7. (5 points) Data Visualization

Explain how you'd visualize the following data for the US. You can use the given map of the US for some examples if you so choose.

A (1 point) Number of Starbucks shops in each state

Answers will vary. (1 point for acceptable technique with reasonable explanation)

B (1 point) Number of earthquakes last year (including small ones)

Answers will vary. (1 point for acceptable technique with reasonable explanation)

C (1 point) Locations of particle accelerators (proportional to their sizes)

Answers will vary. (1 point for acceptable technique with reasonable explanation)

D (1 point) annual potato production (in tonnes) over the last decade

Answers will vary. (1 point for acceptable technique with reasonable explanation)

E (1 point) Major hurricane and flood zones

Answers will vary. (1 point for acceptable technique with reasonable explanation)

BONUS (1 point) According to “Creating Data Driven Enterprise with DataOps” book shared in class, what is the streaming analytical stack powered by at Uber in 2017 (which technologies are used)?
Powered by Kafka and Samza. (1 point if either Kafka or Samza were mentioned)