CS257 Final Fall 2020

Name:
StudID:

Instructions:

1. This final is due 11:59pm PST, Dec 16, 2020.
2. To complete the final, print it out, fill in your answers on the final, and scan it back (or take pictures of the pages and make a pdf) into a file Final.pdf where the total size is less than 10MB.
3. If you don't have a printer, copy and paste each problem into a word processor document. Then after each problem write your solution. Make a less than 10MB Final.pdf file of the result and submit that.
4. Use the same submit mechanism as for the homeworks to submit your completed final.
5. Each problem on this final is worth the same amount (3pts).
6. If you have a question on the interpretation of a problem on the final, you can email me at chris@pollett.org.
7. Due to the coronavirus this is an open book, open internet final.
    a. What that means is that you can consult any static (on the order of static for weeks) source of information related to the final material.
    b. You cannot directly or indirectly ask another person how to do any problem off the final.
    c. To receive credit on problems that make use of your personal information, you need to have correctly filled in that personal information.
    d. When you submit your completed final, you are asserting all of the work in the final is your own.

1. Give an example situation framed in terms of a databases where: (a) you would use a bitmap index over a Bloom filter (0.5 example, 1pt why example works), (b) you would use a Bloom Filter over a Bitmap index (0.5 example, 1pt why example works).

2. Explain and give a concrete example (involving a database of your childhood toys) of how to create a database and a collection in MongoDB (1pt). Give commands to insert several items into this collection, and give an example of query these items and returning the result (1pt). Explain how map reduce aggregation can be done in MongoDB (1pt).

3. Suppose R had 2,000,000 tuples and 8 fit into a block of (rightmost digit of your id +1)*1024 bytes. (a) How many blocks and bytes does R take to store? (b) If the key is 16 bytes long and the record pointer 8 bytes long, approximately how many index records can fit in a block? (c) If we have a sparse index on R, how many blocks and bytes would the index file take?

4. Consider 10 people: Person 0, Person 1, ..., Person 9. For i=0,..., 8, Person {i} Knows Person {i+1} and Person {i} Knows Person {i} (they know themselves). Let j be the next-to-rightmost digit of your student id. We also have Person 9 Knows Person {j}. Show the CYPHER commands needed to create this graph in Neo4j (1pt). Express as a CYPHER query, everyone known by at least two people (1pt). Express as a CYPHER query everyone who knows someone who knows Person {k} where k is the last digit of your student id (1pt).

5. Explain and give an example using your name of how to use XQUERY FLWOR expressions to (a) return the results of computing an XPath expression where a salary attribute of an employee tag is greater than the year you were born in as specified in a LET clause (1pt), (b) format the results of computing a query in <answer> tags (1pt), (c) compute the join on some attributes of two XML documents (1pt).

6. Explain and give an example of the following concepts: (a) consistent hashing (1pt), (b) stabilization (as related to key value stores) (1pt), (c) gossiping (as related to key value stores) (1pt).

7. For the Inmon 1996 definition of data warehouse given in class, for each of the parts of the definition, give an explanation and concrete example (1pt). Give a SQL query on a data warehouse involving the cube operator that could be used by management in a decision-making process (say why) (1pt). Briefly explain what a star schema is with regard to data warehouses (1pt).

8. Modify the Hadoop Map Reduce job from class (you can cut and paste that code as your starting point) so that it computes for each term the number of the documents that have more than the first non zero digit of your student id occurrences of that term (1pt map, 1pt reduce). Explain how to compile and run your program (1pt).

9. What are the four steps in the Total Data Quality Management cycle? (1pt) Make up an example scenario involving airline data. Walk through the four steps of TDQM in terms of your airline scenario (1pt). Briefly explain the orchestration pattern used in process management (1pt).

10. Briefly explain the difference between predictive and descriptive analytics (1pt). Give an example of a technique connected with each (1pt) and explain how that technique works (1pt).