Info 290T: Human-in-the-loop Data Management **Spring 2020**

Assignment 2

Due Monday, May 4 at 11:59pm

General Instructions

- Feel free to talk to other members of the class in doing the homework. You should, however, write down your solutions yourself. *List the names of everyone you worked with at the top of your submission.*
- Keep your solutions brief and clear.
- Please use Piazza if you have questions about the homework but do not post answers. Feel free to use private posts or come to the office hours.
- Here is the UC Berkeley student code of conduct: `sa.berkeley.edu/code-of-conduct`. We will treat any violations of the code of conduct with the seriousness it deserves.

Homework Submission

- Late Policy: You are allowed up to four late days, to be used in any way, without any explanation necessary. Any submission after the deadline will be rounded up to the closest day (i.e., 24 hour window)—so even one hour late will count as the use of a full day. After four late days, students will lose 1% of the grade per day late.
- We will be using BCourses for collecting the homework assignmens. Please submit your answers via BCourses. Hard copies are not accepted.
- Contact Doris Lee if you are having technical difficulties in submitting the assignment; attempt to submit well in advance of the due date/time.
- The homework must be submitted in **pdf** format. Scanned handwritten and/or hand-drawn pictures in your documents won't be accepted.
- Please do not zip the answer document (PDF) so that the instructors can read it directly on BCourses. You need to submit one answer document, named as **hw2_CalID.pdf**.

Transactional Processing (20 pts)

Consider a simple relation Student(s integer) initially containing three tuples [1], [2], and [3] with values of s as 1, 2, and 3.
Consider the following two concurrent transactions.

```
Transaction T1: (a) update Student set s = s + 2;
                (b) insert into Student (select * from Student where s > 3);
                 commit;
Transaction T2: (a) select sum(s) from Student;
                (b) select sum(s) from Student;
                commit;
```

Assume that any interleaving between concurrent transactions is at the statement level. In other words, transaction T2 cannot perform any read operations within the execution of transaction T1's individual modification statements. However, in some cases transaction T2 may perform its read operations between the execution of transaction T1's modification statements. Similarly, transaction T1 cannot perform any modification operations within the execution of transaction T2's individual queries, although in some cases transaction T1 may perform its modifications between the execution of transaction T2's queries.
Recall that in class that you can set a custom ISOLATION LEVEL for each transaction to relax the default serializable requirement. In this question, you will consider various scenarios of isolation levels for T2 and explain how they would affect the result of T2. You can additionally assume for all parts of this problem that transaction T1 executes with the default isolation level serializable.

1. [**5pts**] If transaction T2 executes with isolation level "serializable", what are all the possible sets of values selected by T2's two queries? Clearly list each set separately.
ANSWER:

2. [**5pts**] If transaction T2 executes with isolation level "read committed", what are all the possible sets of values selected by T2's two queries? Clearly list each set separately.
ANSWER:

3. [**5pts**] If transaction T2 executes with isolation level "repeatable read", what are all the possible sets of values selected by T2's two queries? Clearly list each set separately.
ANSWER:

4. [**5pts**] If transaction T2 executes with isolation level "read uncommitted", what are all the possible sets of values selected by T2's two queries? Clearly list each set separately.
ANSWER:

Online Analytical Processing – OLAP (25 pts)

Consider a fact table that records student exam scores over time:
`Student(studentID, semester, courseID, examID, score)`
Suppose the following four materialized views have been created on this table:

```
CREATE MATERIALIZED VIEW V1 As
  SELECT studentID, semester, Sum(score)
  FROM Student
  GROUP BY studentID, semester

CREATE MATERIALIZED VIEW V2 As
  SELECT studentID, semester, courseID, Sum(score)
  FROM Student
  GROUP BY studentID, semester, courseID

CREATE MATERIALIZED VIEW V3 As
  SELECT studentID, semester, courseID, Sum(score)
  FROM Student
  GROUP BY CUBE (studentID, semester, courseID)

CREATE MATERIALIZED VIEW V4 As
  SELECT studentID, semester, courseID, Sum(score)
  FROM Student
  GROUP BY ROLLUP (studentID, semester, courseID)
```

For each of the following 5 queries, state which of the aforementioned views is most efficient to use instead of accessing the relation Student to answer the following query. If none of the views can be used, write NONE.
When comparing efficiency, consider primarily how many tuples are required to compute the query answer: the fewer, the better. In case of ties, consider the size of the accessed view secondarily: the smaller, the better.

5. [**5pts**] Which of the aforementioned views are most efficient to compute the following query? Provide a brief explanation to justify your choice.

```
SELECT studentID, courseID, Sum(score)
FROM Student
GROUP BY studentID, courseID
```

Answer:

6. [**5ts**] Which of the aforementioned views are most efficient to compute the following query? Provide a brief explanation to justify your choice.

```
SELECT studentID, Sum(score)
FROM Student
GROUP BY studentID
```

Answer:

7. [**5pts**] Which of the aforementioned views are most efficient to compute the following query? Provide a brief explanation to justify your choice.

```
SELECT courseID, examID, Sum(score)
FROM Student
GROUP BY courseID, examID
```

Answer:

8. [**5pts**] Which of the aforementioned views are most efficient to compute the following query? Provide a brief explanation to justify your choice.

```
SELECT Sum(score)
FROM Student
```

Answer:

9. [**5pts**] Which of the aforementioned views are most efficient to compute the following query? Provide a brief explanation to justify your choice.

```
SELECT studentID, semester, Sum(score)
FROM Student
GROUP BY studentID, semester
```

Answer:

Map Reduce (15 pts)

Consider a file of temperature sensor readings: `S(time,temp)`. Assume times are integers, temperatures are reals, and there is exactly one tuple for timesteps $0, 1, 2, ..., n$. You want to produce an output file: `Smoothed(time,temp)`
`Smoothed` contains one record for each timestep t in `S` starting at 1 and ending at n-1. The `temp` value in `Smoothed` is the average of all `temp` values in `S` whose time is within ±1 of t.
10. [**20pts**] Express this query in the MapReduce framework. Specifically, fill in the Map and Reduce functions, so that the output of a MapReduce job using your functions produces the desired smoothed result.

- Remember that a Map function takes an item as input, and returns zero or more ⟨key, value⟩ pairs as output. For this problem, each input item is a tuple from table `S`.
- Remember that a Reduce function takes as input a ⟨key, set-of-values⟩ pair, and produces one or more items as output. For this problem, each output item should be a tuple in the query result.

```
Map(<time, temp>):
Reduce(<k,{v_1, v_2,..., v_n}>):
```

We are not concerned about syntax—any understandable pseudo-code will do, just make sure to be very clear about what is returned by each function.
ANSWER:

Inverted Index (15 pts)

An inverted index consists of a list of all the unique words that appear in any document, and for each word, a sorted list of the documents in which it appears.

11. [**10pts**] Imagine that you are building a text search engine for all the articles on Wikipedia, where we only intend to use queries, such as w_1 AND/OR w_2 AND/OR ... AND/OR w_n. Which of the following strategies is more effective for improving the search engine efficiency with minimal impact to search effectiveness? Consider factors such as the size of the inverted index and the amount of time it would take to identify relevant documents given a search query.

1. remove k common words
2. remove k rare words
3. build the index based on bigrams (e.g., two word combinations) instead of simply a single word
4. remove k documents that are very long and contain many words

Briefly explain why each strategy would or would not be effective.

Answer:

12. [**5pts**] Imagine that you have a large collection of images and you want to be able to perform image retrieval on the collection of images (e.g., given the query 'cat', return images containing cats). Given what you know about inverted indexes, explain how you might use an inverted index to facilitate image retrieval. List one advantage and one disadvantage for using an inverted index for image search.

Answer:

Different types of Data Systems (25 pts)

Consider the following data systems and technology that we covered in class this semester.

- RDBMS row-store (e.g., PostgreSQL)
- RDBMS column-store (e.g., Vertica)
- Dataframes (e.g., Pandas)
- MapReduce (e.g., Hadoop)
- Document Store (e.g., Mongo)

For each of the hypothetical scenarios below, pick the most suitable data system that would support the use case. Please list **at least two** reasons why the system is suitable for the application and **one** reason why you might want to consider something else. The reason should be largely be based on the description in the scenario or realistic extensions to this use case based on your imagination.

13. [**5pts**] You are building a data system to support large social media applications like Reddit with forums and commenting capabilities. There are a lot of free-form text in the website to be recorded. There is also a long tail of users who have an account but have very little activity over time (post, upvotes). Due to Reddit's constant UI change, the kind of information to be recorded changes constantly. There is minimal need for joins across different bits of information. The system needs to support large amounts of concurrent transactions (such as upvotes on a post) when there is a post that is trending. However, it is okay if the database is inconsistent at times, i.e., we lose a few individual upvotes or the upvotes does not show up to all users immediately.

Answer:

14. [**5pts**] You are analyzing a dataset consisting of 100 million records of time series information about traffic flow in various cities in the US. This dataset is historical and collected from 2013-2018. There are no plans for adding new information to this dataset. For your analysis, you might have to join the dataset with other information, such as weather.

Answer:

15. [**5pts**] You are working with land satellite imagery data stored across many servers. You can think of each of these images as a large matrix with storing numerical values of temperature intensity for each pixel. You want to process this data once to build an machine learning model for finding images that correspond to a particular land type (i.e., images with high average intensity values).
Answer:

16. [**5pts**] You are building a data system to support an emergency room at a hospital. You need to record the intake of each patient and updates to their medication information. The system does not aim to store the free-form clinician's note, but each patient record contains a link to an external software that allows clinicians to share, annotate, and edit their notes. There may be multiple types of users, such as front-desk staff, nurses, doctors, interacting with the database at the same time. The database needs to be reliable and consistent at all times to support real-time medical decision making.
Answer:

17. [**5pts**] You have a 30-megabyte CSV file containing information about student information and grades for past semesters in an undergraduate database course. Your goal is to build a machine learning model to predict whether a student might struggle with a course based on early indicators. The dataset has a mix of string values, such as the student's major, class standing (freshmen, sophomore, etc.), other courses that they have also taken, as well as numeric values, such as the student's grades in past courses and their GPA. You want to clean up the dataset, such as performing data imputation and wrangling the data into a form that can be used as input to a machine learning model.
Answer: