INFS3200 Advanced Database Systems

Individual Assignment (25%)

Semester 2, 2022

Due: Week 13 – Friday October 28th at 4pm
Submission: via Blackboard → `Assessment > Assignment > Submission`

Introduction

In this assignment, you will complete four distinct tasks with a total of seven questions (totaling 25 marks) to demonstrate your understanding of multiple topics including distibuted databases, data warehousing, data integration, and data quality management. You will also be required to implement programs to demonstrate your problem solving ability.

Warning: This is an **individual assignment.** You must not share your answers with anyone else. Sharing answers or plagiarism is unacceptable and will be penalized.

Marking Scheme

- The marks associated with each component are listed in square brackets.
- Marks may be reduced for poorly formatted submissions, or submissions which do not run on the lab environment. Please double check your submission.

Submission Format

Compile your results into a document, with a subsection dedicated to each component of the assignment. Keep your explanations terse but make sure you contain all necessary information. Export your report into a **PDF document.** Copy your code into a subdirectory named `src` and copy your data into a subdirectory named `data`. Package your code, data, and report into a single, appropriately named zip file (for example, `JohnnyExample_s1234567890_assigment.zip`**).** Please format your document and code nicely to assist the tutor's marking process. A poorly formatted document may receive a reduced mark. **Due 4pm on Friday the 28th of October, 2022.**

Late Penalties (from the ECP)

"*Where an assessment item is submitted after the deadline, without an approved extension, a late penalty will apply. The late penalty shall be 10% of the maximum possible mark for the assessment item will be deducted per calendar day (or part thereof), up to a maximum of seven (7) days. After seven days, no marks will be awarded for the item. A day is considered to be a 24 hour block from the assessment item due time. Negative marks will not be awarded.*"

Preliminaries

Tips

1. It is recommended that you complete *practical 3* before working on the coding part (that is, part 4) of this assignment. Although the assignment is independent to the practicals, the code introduced in *practical 3* can be used as a starting framework for the tasks required in this assignment.

2. The datasets used in this assignment contain thousands of records. As such, it can be difficult to manually check your answers record-by-record. As such, we recommend using tools such as a fast text editor, a spreadsheet program, or command line tools like `grep` and `awk` to view, sort, and search the raw data as appropriate. Please be careful not to change any of the original data, as it may affect your results.

3. You are free to implement your code in Java, Python, or directly as SQL scripts. The code must be readible, and must be accompanied with comments, to facilitate easy marking. It is also good practice to write clear code with appropriate comments.

Dataset Description

In this assignment, there are four datasets containing information pertaining to books. The data can be found on the LMS (`Assessment > Individual Assignment > Assignment Specification and Data`). These datasets are derived from different sources. Each dataset schema is outlined below:

Book1 (id, title, authors, pubyear, pubmonth, pubday, edition, publisher, isbn13, language, series, pages)

Book2 (id, book_title, authors, publication_year, publication_month, publication_day, edition, publisher_name, isbn13, language, series, pages)

Book3 (ID, Title, Author1, Author2, Author3, Publisher, ISBN13, Date, Pages, ProductDimensions, SalesRank, RatingsCount, RatingValue, PaperbackPrice, HardcoverPrice, EbookPrice, AudiobookPrice)

Book4 (ID, Title, UsedPrice, NewPrice, Author, ISBN10, ISBN13, Publisher, Publication_Date, Pages, Dimensions)

Part 1: Database Schema and Fragmentation [6 marks total]

Read the above schemas carefully and understand the meaning of the attributes. If you don't know the meaning of a certain attribute, inspect the data under it (within the data files) or try to find the meaning via a web search engine (especially for some abbreviations, like ISBN, for example).

Answer the following questions based on your understanding.

Question 1: [2 marks]
Given four datasets that are stored in one relational database as separate relations.

1. Write an SQL query to solve the following problem: *Find the top 5 books with the highest ratings, and 5 books that have the lowest rating, return their ranks (sorted in descending order), titles, publishers and number of pages.*

2. Which table schema(s) is/are used to answer the above query?

Question 2: [4 marks]
Given that **Book3** is stored in a distributed database **A**, and the two most frequent queries are:

- *Find all books whose publisher name is [Some Name] (or, a list of names), return their book titles and author info.*

- *Find all books that are published in a given year, and return their book IDs, ~~languages~~, number of pages, HardcoverPrice, and EbookPrice.*

Answer the following questions:

1. [2 marks] If the goal of database **A** is to handle each query via a dedicated local site (with no information needed from the other site(s)), which fragmentation strategy should be used to fragment the **Book3** table? If only two fragments are generated, write their schemas (if vertically fragmented) or predicates (if horizontally fragmented), respectively. (*Note: there are many valid fragmentation solutions; just provide one of them.*)

2. [2 marks] Assume that we *horizontally* fragment the table into three fragments based on the following predicates:

   - Fragment 1: `RatingsCount ≤ 25`
   - Fragment 2: `25 < RatingsCount ≤ 120`
   - Fragment 3: `RatingsCount > 125`

   Is this set of predicates valid?

   - **If so**, please explain (using plain English) the insertion process if we want to insert a new record into **Book3**.
   - **If not**, please generate a valid predicate set using minterm predicates (show the calculation process). Then, explain the insertion process for a new record after the valid predicate set is made.

Part 2: Data Warehousing [7 marks total]

In this section, we design a Data Warehouse on book sales with respect to the aforementioned **Book1, Book2, Book3**, and **Book4** datasets. In particular, we need to use the data from the given datasets and create a Data Warehouse Schema. The designed Data Warehouse will contain summary data, such as the total sales of each publisher, for each day and each language. The following table shows an example:

Date	Publisher	Language	Sales
07/15/1984	AAAI Press	English	11
05/05/1990	Springer International Publishing	English	23
06/04/1995	Springer London	English	15
12/11/2000	IEEE Computer Society Press	English	30
04/03/2004	AAAI Press	Japanese	2
05/01/2008	Springer International Publishing	Japenese	13
11/19/2012	Springer London	Japanese	5
08/06/2014	IEEE Computer Society Press	Japanese	22

Question 3: Design a Data Warehouse Schema that can accommodate the above example, and then answer the following questions:

1. [1 mark] Show the schema and point out the dimensions and fact table. Given that we have a dimension table for each dimension and there are 4000 records in the fact table; Among all dimension tables and the fact table, which table has the most records? Why? Explain your answer.

Question 4: Now we want to the create bitmap indices for the given model:

1. [2 marks] What are the advantages of building a bitmap index? Which type of column is not suitable for a bitmap index? Why?

2. [2 marks] Suppose the `Publisher` column only contains four distinct values and `Language` only contains two, which are all shown in the above example. Please create bitmap indices for both `Publisher` and `Language`.

3. [2 marks] Explain how to use the bitmap indices to find the total sales of `Japanese` books published by `AAAI Press`.

Part 3: Data Integration [4 marks total]

Assuming that the data warehouse loads data from the above four sources (**Book 1,2,3,4**), you are asked to integrate their data and address various data quality issues. In this part, the database sources (i.e., owners) only give you their schemas (shown in the *Dataset Description* section of the *Preliminaries* on page 2), and you are asked to design an integrated schema based on the given schemas. You must assume that the data records within tables **Book 1,2,3,4** are not available for you at this stage – you only have access to the schemas

Question 5: Define a global schema (using the approach namely, Global as a View) which can integrate data from all four sources.

1. [2 marks] Design a global schema which will combine the common attributes from each schema together. Your design should include any information that is represented in all four schemas. If an attribute cannot be found or derived in the given schemas, then it should be left out of your global schema.

2. [2 marks] Identify any structural or semantic heterogeneity issues that may occur during your integration. Using the data, provide a concrete example of each, if applicable, and outline a possible resolution.

Part 4: Data Quality [8 marks total]

Now assume you are provided with the actual data from each source, namely `Book1.csv`, `Book2.csv`, `Book3.csv`, and `Book4.csv` (see the `Assignment-Data.zip` file). As it is very common that the same book is recorded by different sources, it is crucial to identify the redundant information by merging and eliminating the duplicated records during the data integration process, which relies on data linkage techniques.

To assist you with your analysis, we provide a human-labelled gold-standard dataset (refer to Prac 3 Part 2.2 for more information about gold-standard data), named as `Book1and2_pair.csv`, which lists all correct matchings between **Book1** and **Book2**. It will be used in the following tasks. Its schema is as follows:

`Book1and2_pair``(Book1_ID, Book2_ID)`

All of the datasets are provided as CSV files, and hence have the fields separated by the comma character. Thus, if two commas appear consecutively, it means the value in the corresponding field between two commas is `NULL` (absent). Furthermore, if an attribute field contains a comma naturally, the field will be enclosed by a double quote ("") to differentiate the actual comma notation inside attribute from the outside comma separator. An example record from **Book2** is as follows:

```
1725,Informix Unleashed,"John McNally, Jose Fortuny, Jim Prajesh,
Glenn Miller",97,6,28,1,Sams,9.78E+12,,Unleashed Series,1195
```

According to the **Book2** schema, we can infer the following fields:

id	1725
book_title	Informix Unleashed
authors	John McNally, Jose Fortuny, Jim Prajesh, Glenn Miller
publication_year	97
publication_month	6
publication_day	28
edition	1
publisher_name	Sams
isbn13	9.78E+12
language	NULL
series	Unleashed Series
pages	1195

Here, since there are commas within the authors field, the whole field is enclosed by a notation of double quotes. Also, since there are two consecutive commas before `Unleashed Series`, it means that the language is `NULL`.

In this part, you are asked to answer the following questions by writing code to complete the tasks (**if "code required" is specified**) and provide answers based on the results of your programs. Make sure you maintain a copy of your code as you will be required to submit it along with the report; refer to the *Submission Format* section on page one for specific instructions on packaging your code for submission.

Question 6: Sample records from `Book3.csv` to measure its data quality:

1. [1 mark, code required] By sampling records whose `id` is the multiple of 100 (that is: 0, 100, 200, 300, ...), how many records are there in the sample set?

2. [1 mark, code required] Among the samples found in the previous question (question 6.1), how many fields containing `NULL` values are present?

3. [2 marks] Calculate the *defects per million opportunities* (DPMO) according to your samples. You can assume that any `NULL` value is an error, and that the remaining values are valid. *Hint: you can sample the records manually to validate the correctness of your program results.*

Question 7: Perform data linkage on **Book1** and **Book2** using the methods shown in *Practical 3*:

Given two author strings from **Book1** and **Book2** that refer to the same author list:

```
s1 = "Peter Rob, Carlos Coronel"
s2 = "Carlos Coronel;Steven Morris;Peter Rob;"
```

1. [1 mark] Compute the edit distance between **`s1`** and **`s2`** – What is the value?

2. [1 mark] Compute the Jaccard distance between **`s1`** and **`s2`** – What is the value?

[2 marks, code required] Write a program to link the data between **Book1** and **Book2** using the *Jaccard coefficient* with *3-gram tokenization* as the similarity measure, performing the comparison only on the `book title` field, and using a matching threshold of 0.75 (that is, only return matches with a similarity of 0.75 or more). Compare your output with the gold-standard dataset and compute the precision, recall and F-measure.

Changelog

v1: Original document
v2: Added specific details on where the data can be accessed
v3: Removed languages from Q1.2