INFS3200 Advanced Database Systems

Assignment (25%)

Semester 1, 2023

Deadline: 4pm Friday, 26 May 2023
Submit: Online Submission on the Blackboard INFS3200 Course Website

Introduction

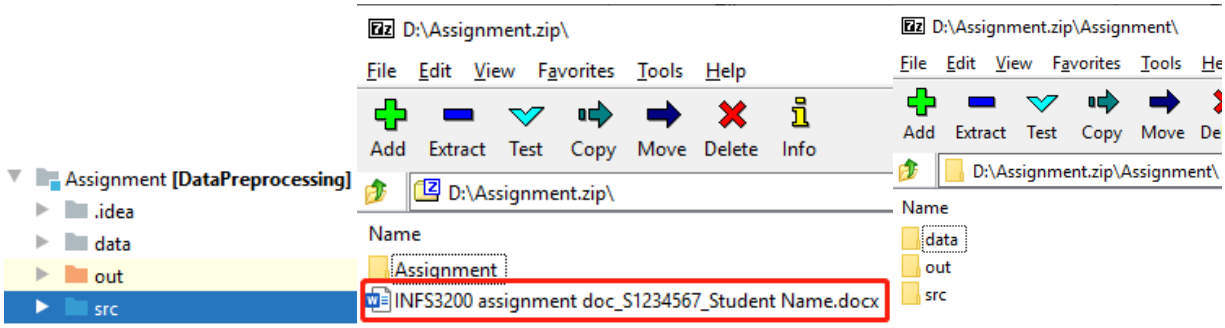
The assignment contains four parts with seven questions (total marks 25 for 25% of the course) to demonstrate your understanding of multiple topics, including distributed database, data warehousing, data integration and data quality management. Meanwhile, coding is required for some questions to show your problem-solving ability. This assignment must be performed individually.

Important Notes:

1. As UQ has provided the Lab environment for this assignment, you don't need to install the required software systems on your own machine. The software environment problems on your own computer machine cannot be used to ask for extension of submission.
2. Each dataset used in this assignment contains thousands of records, which is hard to be checked record-by-record manually. Therefore, it is recommended to have a handy text editor tool (e.g. Microsoft Excel, Notepad++ or Sublime Text on Windows) to view and search the contents in CSV files. Please use search function (i.e., *Ctrl+F* keys) in text editor to look through values. Also, please don't change the data unintentionally while viewing or searching, as it may affect your assignment results.
3. You should complete *Prac 3* before working on the coding part of this assignment (i.e., *Part 4* of this assignment). Although the assignment is independent to the three practicals, the code introduced in *Prac 3* can be a starting point of this assignment as the tasks are similar.
4. You implement your code in SQL, Java or Python, you may choose the ones that you feel comfortable. The code must be companied by minimum comments so that tutors can understand the structure of your coding and the objective of each snippet. If you performed this assignment on your own Laptop machine instead of the UQ provided software environment, you must ensure that the codes submitted by you are all able to execute correctly on UQ provided Lab environment, either remotely via Internet connection, or locally in UQ GPS Building Lab 78-116.

Submission Requirements:

Please include all your answers in a word/pdf document. Pack the documents with your code folder (which contains at least "src" and "data" folders, shown as below) into a **.zip/.rar** file and submit it to the Blackboard INFS3200 course Website. The name of both the zip file and the document should contain your student ID, your name and "Assignment", shown as follows:



Please format your document nicely, in terms of consistent font, font size and spacing. The answers are suggested to follow the below structure (No need to repeat questions if not necessary, fonts and spacing are not limited):

…

Part 1.

Question 1: Your answers…

Question 2: Your answers…

Part 2.

…

WARNING: *This assignment must be completed individually,* ***Artificial Intelligence tools cannot be used to generate any part of solutions for this assignment****. Any form of answer-sharing with other people is not acceptable and, once identified, will be penalized. Contract cheating will be investigated and it will result in heavy penalty.*

Preliminary: Dataset Description

In this assignment, we have four datasets about book information from four different sources. The data schemas are listed 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: [6 marks] **Database Schema and Fragmentation**

Read the above schemas carefully and understand the meaning of the attributes. If you don't know the meaning of a certain attribute, check the data under it or Google its meaning (especially for some abbreviations, like ISBN). 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 "*Find the top 15 books that have the highest ratings and 10 books that have the lowest ratings, 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 two queries that are most frequently asked on ***A*** are:

- *Find all books whose publisher name is "XXX" (or among multiple publishers), return their book titles and author info.*
- *Find all books that are published in a given year, return their book IDs, ~~languages~~, number of pages, HardcoverPrice and EbookPrice.*

Answer the following questions: Verticle

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

(2) [2 marks] Assuming that we horizontally fragment the table into three fragments based on the following predicate:

Fragment 1: pages ≤ 200

Fragment 2: 200 < pages ≤ 600

Fragment 3: pages > 800

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). Also, explain the insertion process for a new record after the valid predicate set is made.

Part 2: [7 marks] **Data Warehouse Design**

In this part, we design a Data Warehouse on book sales w.r.t. the Book1, Book2, Book3, and Book4 datasets. Particularly, we need to use data from the given assignment 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 shows just an example:

Day	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	Spanish	2
05/01/2008	Springer International Publishing	Spanish	13
11/19/2012	Springer London	Spanish	5
08/06/2014	IEEE Computer Society Press	Spanish	22

Question 3: Design a Data Warehouse Schema that can accommodate the above example, 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?

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 bitmap index?

(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 "English" books published by "AAAI Press".

Part 3: [4 marks] **Data Integration**

Given 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, those database sources (i.e., owners) only give you their schemas (shown in Preliminary part), and you are asked to design an integrated schema based on the given schemas (i.e., the data records within tables Book 1,2,3,4 are supposedly not available for you at this stages).

Question 5: Now you 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) **[1 marks]** Identify structural heterogeneity issues that may occur during your integration by an example in the schemas together with the possible resolution.

(3) **[1 marks]** Identify semantic heterogeneity issues that may occur during your integration by an example in the schemas together with the possible resolution.

Part 4: [8 marks] **Data Quality Issues**

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 provided datasets). 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 eliminate the duplicated records during the data integration process, which relies on the data linkage techniques to be used. In this regard, we provide a human-labelled gold-standard dataset (refer to **Prac 3 Part 2.2** for more information about gold-standard), 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)

In a CSV file, you need to note that the attributes are separated by comma (,). If two commas appear consecutively, it means the value in the corresponding field between two commas is NULL (i.e., absent). Furthermore, if an attribute field contains comma naturally, the field will be enclosed by a double quote ("") to differentiate the actual comma notation inside attribute from the outside comma separator. For example, a record in 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 Book 2 schema, we can infer the following fields:

id=1725,

book_title=Informix Unleashed,

authors= John McNally, Jose Fortuny, Jim Prajesh, Glenn Miller,

…

isbn13=9.78E+12

language=NULL,

series=Unleashed Series,

pages=1195.

Here, since there are commas in "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 your answers based on the code results. Please store all the code you wrote during this part and submit them to Blackboard Course Website as a part of your assignment submission.

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

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

(2) **[1 mark]** Among the samples found in Question 6-(1), how many fields containing NULL values are presented (code required)?

(3) **[2 marks]** Calculate the Empo (error per million opportunities) according to your samples (only NULL value is considered). (**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 mentioned in ***Prac 3***:

(1) **[2 marks]** Given two author strings from Book1 and Book2 that refer to the same author list:

a. "Richmond Shee, Kirtikumar Deshpande and K. Gopalakrishnan;"

b. "K. Gopalakrishnan, Kirtikumar Deshpande, and Richmond Shee"

Which distance function is more likely to regard them as similar (between two approaches of *edit distance* and *Jaccard distance*)? And Why?

(2) **[2 marks]** Perform the data linkage between Book1 and Book2 data. When linking their results, use *Jaccard coefficient* with 3-gram tokenization as the similarity measure and perform the comparison only on the "book title" field. The book pairs whose similarity is higher than 0.75 are regarded as matched pairs. Compare your output with the gold-standard dataset and write down the **precision**, **recall** and **F-measure** (code required).

--- The End of Assignment ---

The Jaccard distance function is more likely to regard the two author strings as similar.

This is because the Jaccard distance function measures similarity based on the intersection over the union of sets. In this case, the two author strings contain the same three author names, just in a different order. Since the Jaccard distance function does not take into account the order of the elements in the sets, it would consider the two author strings to have the same set of authors and therefore a high degree of similarity.

On the other hand, the edit distance function measures similarity based on the number of edits required to transform one string into the other. In this case, the two author strings are not identical and require multiple edits to transform one into the other. Therefore, the edit distance function is less likely to regard them as similar.