THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

INFS3200 Advanced Database Systems

Data Pre-processing Assignment (25%)

Semester 1, 2021

Introduction

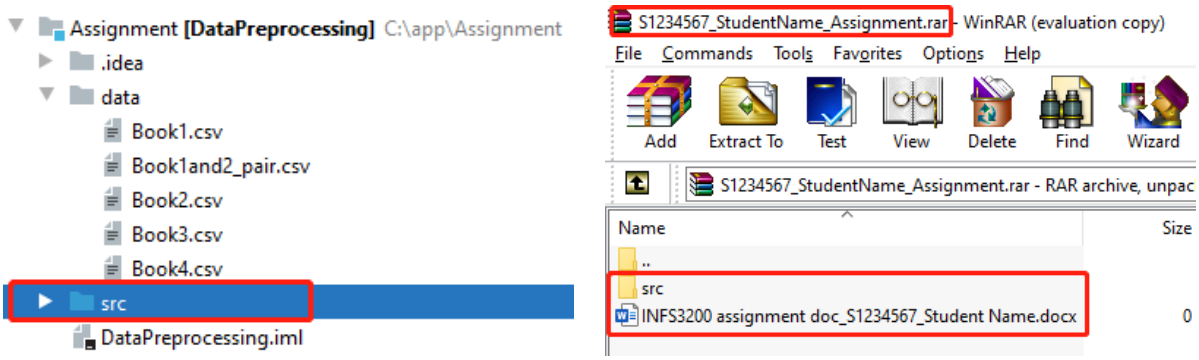
During this assignment, you are going to deal with a real-world data integration and data quality challenge, including answering a series of questions to demonstrate your level of understanding on these topics as well as your abilities of problem-solving and implementation.

Tips & Suggestions:

1. It is **highly recommended** to complete Prac 3 before working on the implementation section of this assignment (Part 4). The assignment is independent to Prac 3, but you will benefit a lot from the Prac 3 solutions in terms of the coding part.
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 fully utilize the search functionality (usually is CTRL+F) in text editor to look for certain values, tuples or characters. Also, please avoid changing the data unintentionally while viewing or searching as it may affect your assignment results.
3. The programming language is not limited to Java or Python, choose the one you feel comfortable with and stick to it until you finish the assignment. The code **must contain basic comments** so that tutors are able to understand the structure of your code and the objective of each snippet.

Assessment:

The assignment will due **at 16:00 pm, May 28th**, please include all your answers in a word/pdf document. Pack the document with your code folder ("src" folder as shown below) into a .zip/.rar file and submit it to the Blackboard. 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: Please complete this assignment **individually**. Any type of answer-sharing among classmates is not acceptable and, once identified, will be penalized.

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: Data Schema Questions [6 marks]

Read the above schemas carefully and understand the meaning of each attribute. 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: Given that all four book datasets were stored in different relational databases, respectively:

(1) **[1 marks]** Can attribute authors be the primary key for Book1? Why?
(2) **[1 marks]** Given a query "Find the top 99 books which are sold the most, return their ranks (sorted in ascending order), titles, PaperbackPrice, HardcoverPrice, EbookPrice, AudiobookPrice of those books.", which schema is capable of answering such query? Write down the corresponding SQL query on that schema.

Question 2: Given that Book2 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 and number of pages.

Answer the following questions:

(1) **[2 marks]** If the goal of A is to handle each query by a dedicated local site, which fragmentation strategy should be used to fragment Book2 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 fragmented the table into three fragments based on the following predicate:

Fragment 1: $1 \leq$ publication_day ≤ 10
Fragment 2: $11 \leq$ publication_day ≤ 20
Fragment 3: $21 \leq$ publication_day ≤ 31

If we want to insert a new record into Book2, please explain the insert process in plain English (you can use an example to demonstrate the process).

Part 2: Data Warehouse Design [7 marks]

In this part, we aim to design a data warehouse on the book sales system. Specifically, we obtained the data from the given datasets and create a table which contains the total sales on each publisher, each day and each language. An example table is shown as follows:

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: Given the above example, answer the following questions:

(1) **[1 marks]** For each of the above four columns, identify if that column is a dimension column or a fact column.

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

(1) **[2 marks]** What are the advantages of building a bitmap index? Which type of column is 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 "Spanish" books published by "AAAI Press".

Part 3: Schema-Based Data Integration [4 marks]

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, the data sources only send you their schemas (shown in Preliminary), and you are asked to design your integration plan based on that (records in Book 1,2,3,4 are not available at this stages).

Question 5: Now we plan to define a global conceptual schema (Global as a View) which can integrate data from all four sources.

(1) **[2 marks]** Design a global conceptual schema which combining the common attributes from each schema together. Your design should include any information that is represented in all four schemas. If a column can not be found or derived in each of the schemas, then it should be left out of your global conceptual schema.

(2) **[2 marks]** Identify two possible heterogeneity issues that may occur during your integration, each heterogeneity should be illustrated by an example in our schemas, together with the possible solution.

Part 4: Data Quality Management [8 marks]

Now you are provided with the actual data from each source, namely “Book1.csv”, “Book2.csv”, “Book3.csv” and “Book4.csv”. As it is very common that the same book is recorded by multiple sources, it is crucial to identify and merge duplicated records during the data integration process, which relies on the data linkage technique 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)

Note that in a CSV file, the attributes are separated by comma (,). If two commas appear consecutively, it means the value in the corresponding field between two commas is NULL. Furthermore, if an attribute field contains comma naturally, the field will be enclosed by a double quote ("") to distinguish the commas inside the attribute with 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 double quote. 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 writing your answers based on the code results. Please store all the code you wrote during this part and submit them to Blackboard.

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 in the sample set (code required)?

(2) **[1 mark]** Among the samples found in 6.1, how many fields containing NULL are present (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 code 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 edit distance and Jaccard distance)? 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. 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).