THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

INFS3200/7907 Advanced Database Systems

Data Pre-processing Assignment (25%)

Semester 2, 2020

Introduction

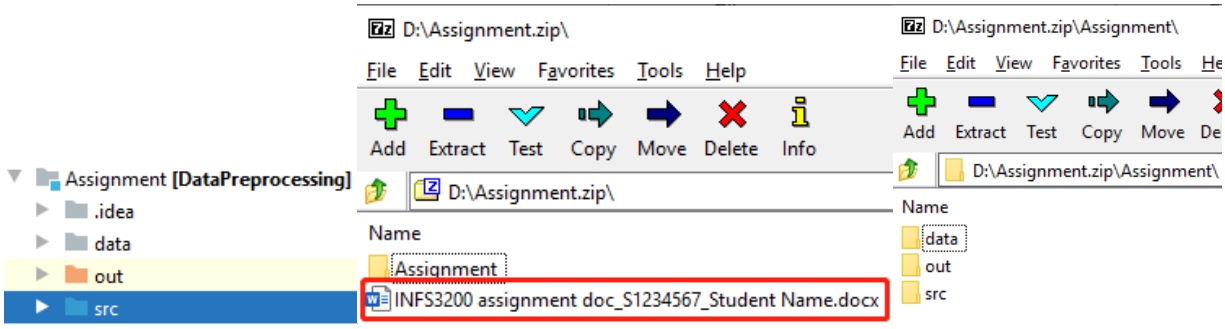
In this assignment, you are facing a real-world data preprocessing task. During the assignment, you are asked to answer several questions to demonstrate your level of understanding on multiple topics, including distributed database, data warehousing, data integration and data quality management. Meanwhile, coding is required for some questions to show your problem-solving ability.

Tips & Suggestions:

1. It is **highly suggested** to complete Prac 3 before working on the coding part of this assignment (Part 4). Although the assignment is independent to pracs, the code introduced in Prac 3 can be the starting point of this assignment as they work on similar tasks.
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. Implement your code in SQL, Java or Python, choose the one you feel comfortable with and stick to it till the end of 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 11:59 pm, November 1st**, please include all your answers in a word/pdf document. Pack the document 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. 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**. The reuse of code from practicals are allowed, but any form 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 Management and Query [6 marks]

Read the above schemas carefully and understand the meaning of each attribute. If you fail to understand some of them, 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 that four datasets are stored in one relational database as separate relations. For a query "Find top 100 books that have the best sales, return their ranks (sorted in ascending order), titles, publishers and number of pages.", which schema(s) can answer such query? Write down the corresponding SQL query.

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 (no information needed from the other 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.

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

Fragment 1: pages ≤ 100
Fragment 2: $100 \leq$ pages ≤ 800
Fragment 3: pages ≥ 600

Is this predicate set valid? If so, please explain the insert process if we want to insert a new record into Book2 (using plain English). If not, please generate a valid predicate set using min-term predicate (show the calculation process). Also, explain the insert process for a new record after the valid predicate set is made.

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 mark]** 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 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: Data Integration and Quality Management [12 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. The actual data of book lists are given as CSV files, namely “Book1.csv”, “Book2.csv”, “Book3.csv” and “Book4.csv”. 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 the “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 through programming (if “code required” is specified). Your answers to the questions must be based on code results. Please save all the code you wrote, and submit them to Blackboard. Do not paste your code to the answer sheet, instead, when answering a question, please specify the location of the corresponding code for that question or name your file as “Question5”, “Question6”, to direct tutor to the correct file.

Question 5: As the book list schemas provided in Preliminary, design a global conceptual schema which combines the common attributes among all four schemas. Your design should include every piece of information that four schemas share in common. In other words, if a column can be **found** or **derived** from every schema, it must be included in your global conceptual schema.

(1) **[2 marks]** Write down the global conceptual schema. The format should be similar to the schemas in Preliminary.

(2) **[3 marks]** Integrate "Book3.csv" and "Book4.csv" data according to the global schema you defined (**code required**). The data should be sorted by ISBN13 in ascending order. You should use either of the two approaches mentioned below.
   a. If you perform the integration on Oracle database, please create a table named "FullBookList" using your schema and insert data into it. Take a screenshot of your table schema from SQL Developer, and another screenshot of **first 20 records in the table**. Both screenshots should include your student ID as database username. Add the screenshots to your solution document, and include your SQL scripts as a text file in your final submission.
   b. If you perform the integration using Java/Python, write the integrated dataset to a CSV file named as "FullBookList.csv". Include this file in your final submission. Take a screenshot of your CSV file that shows first 20 records and add the screenshot to your solution document.

Normally, we would expect to have various data quality issues in an integrated dataset. For example, by checking ISBN13 code in "FullBookList", we can find multiple pairs of books with the same ISBN13 code, like "9781296126568", "9780679887911", "9781298248848", etc. 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 record linkage technique.

In this regard, question 6 asks you to perform a record linkage task on "Book1.csv" and "Book2.csv". 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)

Question 6: **[4 marks]** Perform data linkage on Book1 and Book2 using the methods mentioned in Prac 3. 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 (double quotes that are used to enclose book titles should be removed before the linkage). 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).

Question 7: [3 marks] In addition to the duplication issue, we want to explore other data quality issues remained in datasets. Create a sample dataset from "Book3.csv" containing all records whose id is the multiple of 100 (i.e. 100, 200, 300, …). Among all samples, how many fields (a field is a cell in the table) containing NULL are present (here, NULL is recorded as an empty value in this field)? Calculate the Empo (error per million opportunities) according to your samples (Empo= number of NULLs / number of fields) (code required). (**Hint:** you can sample the records manually to validate the correctness of your code results)