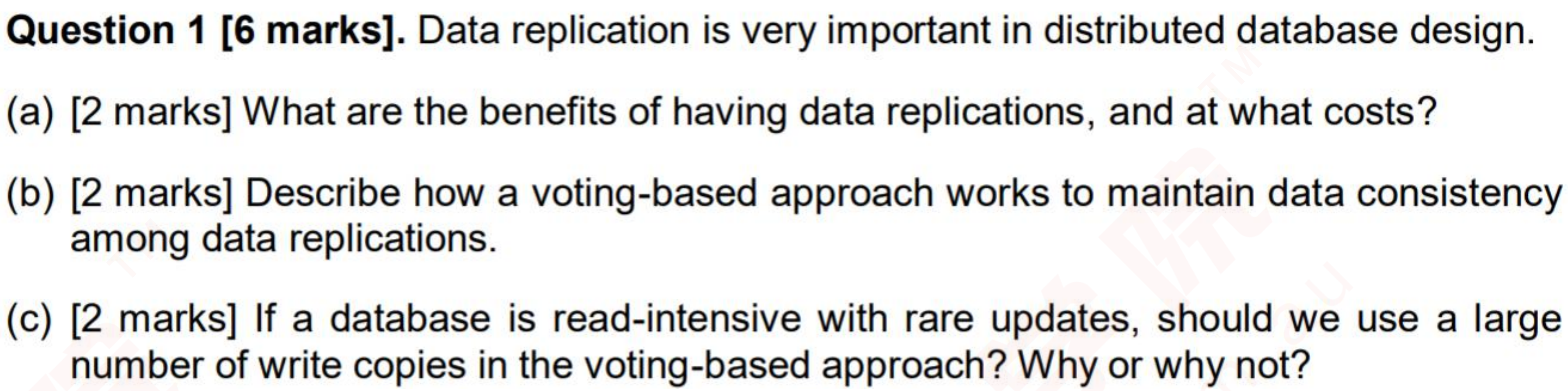
Final Exam Part 1

INFS 3200/7907

Bitwise operation

100



Question 1 [6 marks]. Data replication is very important in distributed database design.

(a) [2 marks] What are the benefits of having data replications, and at what costs?

(b) [2 marks] Describe how a voting-based approach works to maintain data consistency among data replications.

(c) [2 marks] If a database is read-intensive with rare updates, should we use a large number of write copies in the voting-based approach? Why or why not?

Pros:

Accelerated data retrieval when data is stored locally

Data backup

Cons:

Extra storage cost

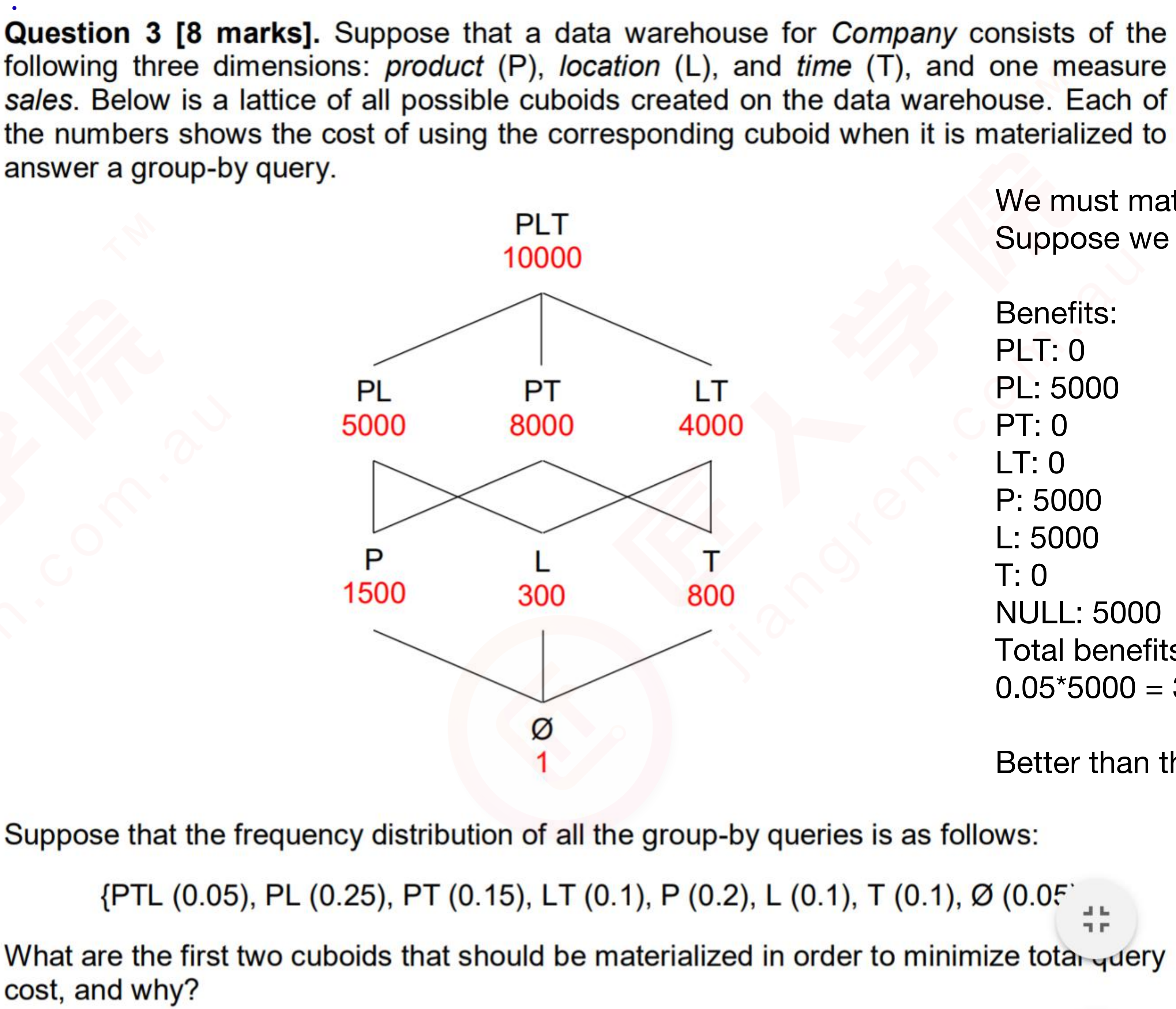
Update cost. When the original data is updated, we need to update the replicas too.

Suppose we have S replicas.

Whenever there is a need to update them, we update the majority copies, say m copies. The version number of each copy is updated at the same time too.

When we read, we need to read at least n copies such that $n + m > S$.

We should adopt the Read-Any Write-All update strategy. Since the database is read-intensive, we need to perform a lot of read operations using the voting strategy. Although we need to write many times in RAWA, the database involves few updates.



Question 3 [8 marks]. Suppose that a data warehouse for *Company* consists of the following three dimensions: *product* (P), *location* (L), and *time* (T), and one measure *sales*. Below is a lattice of all possible cuboids created on the data warehouse. Each of the numbers shows the cost of using the corresponding cuboid when it is materialized to answer a group-by query.

PLT
10000
PL
5000
PT
8000
LT
4000
P
1500
L
300
T
800
Ø
1

Suppose that the frequency distribution of all the group-by queries is as follows:

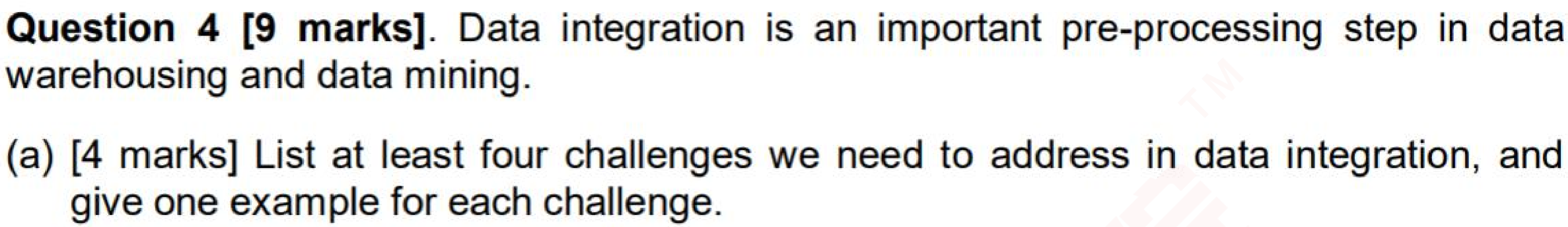
{PTL (0.05), PL (0.25), PT (0.15), LT (0.1), P (0.2), L (0.1), T (0.1), Ø (0.05

What are the first two cuboids that should be materialized in order to minimize total query cost, and why?

We must materialize PLT
Suppose we pick PL

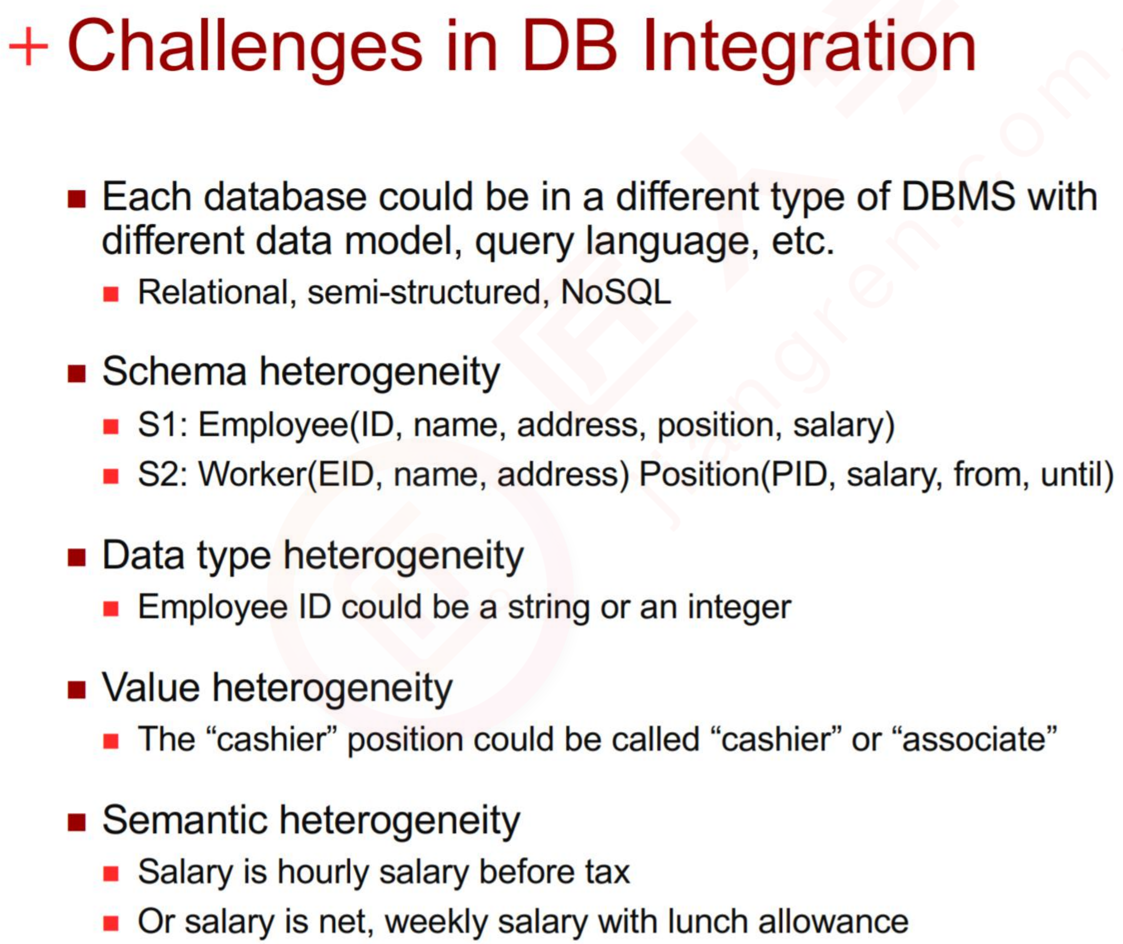
Benefits:
PLT: 0
PL: 5000
PT: 0
LT: 0
P: 5000
L: 5000
T: 0
NULL: 5000
Total benefits: 0.25*5000 + 0.2*5000 + 0.1*5000 + 0.05*5000 = 3000

Better than the benefits of materializing LT or PT



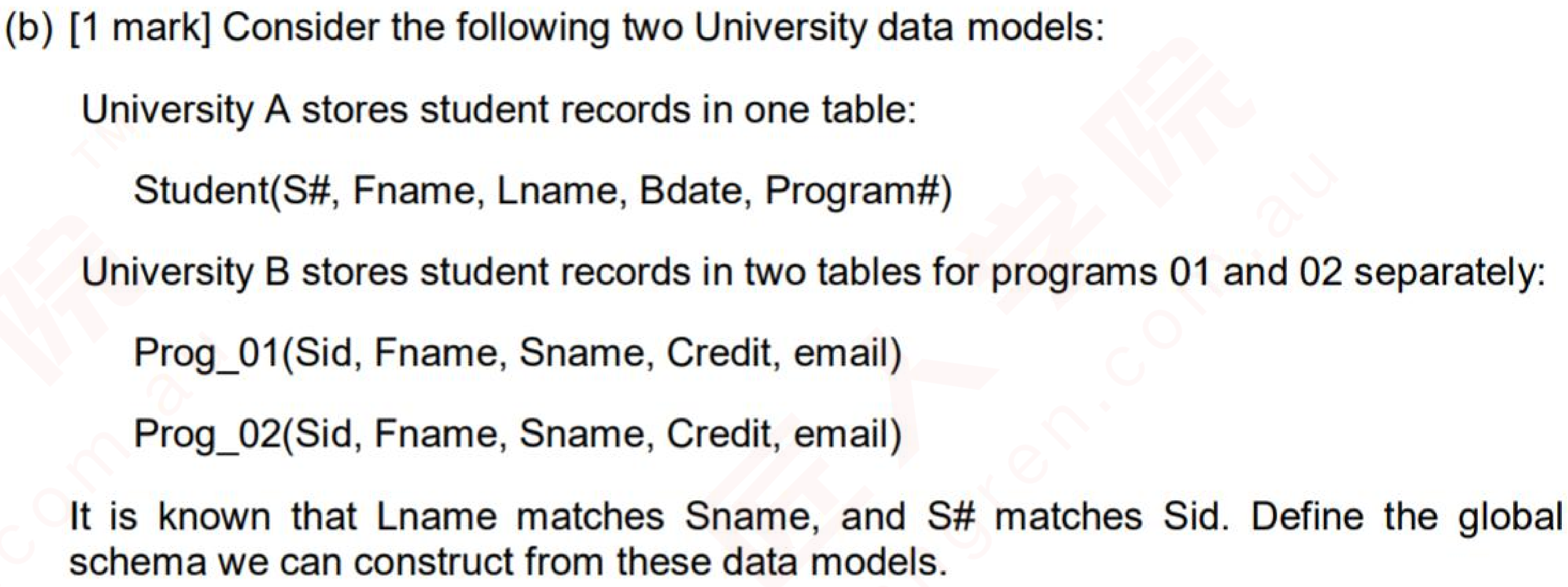
Question 4 [9 marks]. Data integration is an important pre-processing step in data warehousing and data mining.

(a) [4 marks] List at least four challenges we need to address in data integration, and give one example for each challenge.



+ Challenges in DB Integration

- Each database could be in a different type of DBMS with different data model, query language, etc.
  - Relational, semi-structured, NoSQL
- Schema heterogeneity
  - S1: Employee(ID, name, address, position, salary)
  - S2: Worker(EID, name, address) Position(PID, salary, from, until)
- Data type heterogeneity
  - Employee ID could be a string or an integer
- Value heterogeneity
  - The “cashier” position could be called “cashier” or “associate”
- Semantic heterogeneity
  - Salary is hourly salary before tax
  - Or salary is net, weekly salary with lunch allowance



(b) [1 mark] Consider the following two University data models:

University A stores student records in one table:

Student(S#, Fname, Lname, Bdate, Program#)

University B stores student records in two tables for programs 01 and 02 separately:

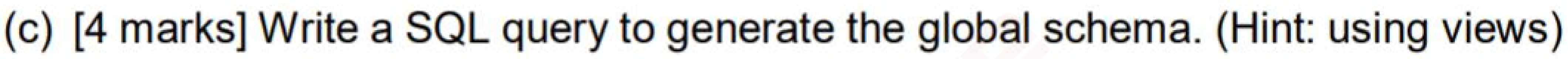
Prog_01(Sid, Fname, Sname, Credit, email)

Prog_02(Sid, Fname, Sname, Credit, email)

It is known that Lname matches Sname, and S# matches Sid. Define the global schema we can construct from these data models.

GlobalUni(Sid, Fname, Lname)

Student U Prog_01 U Prog_02

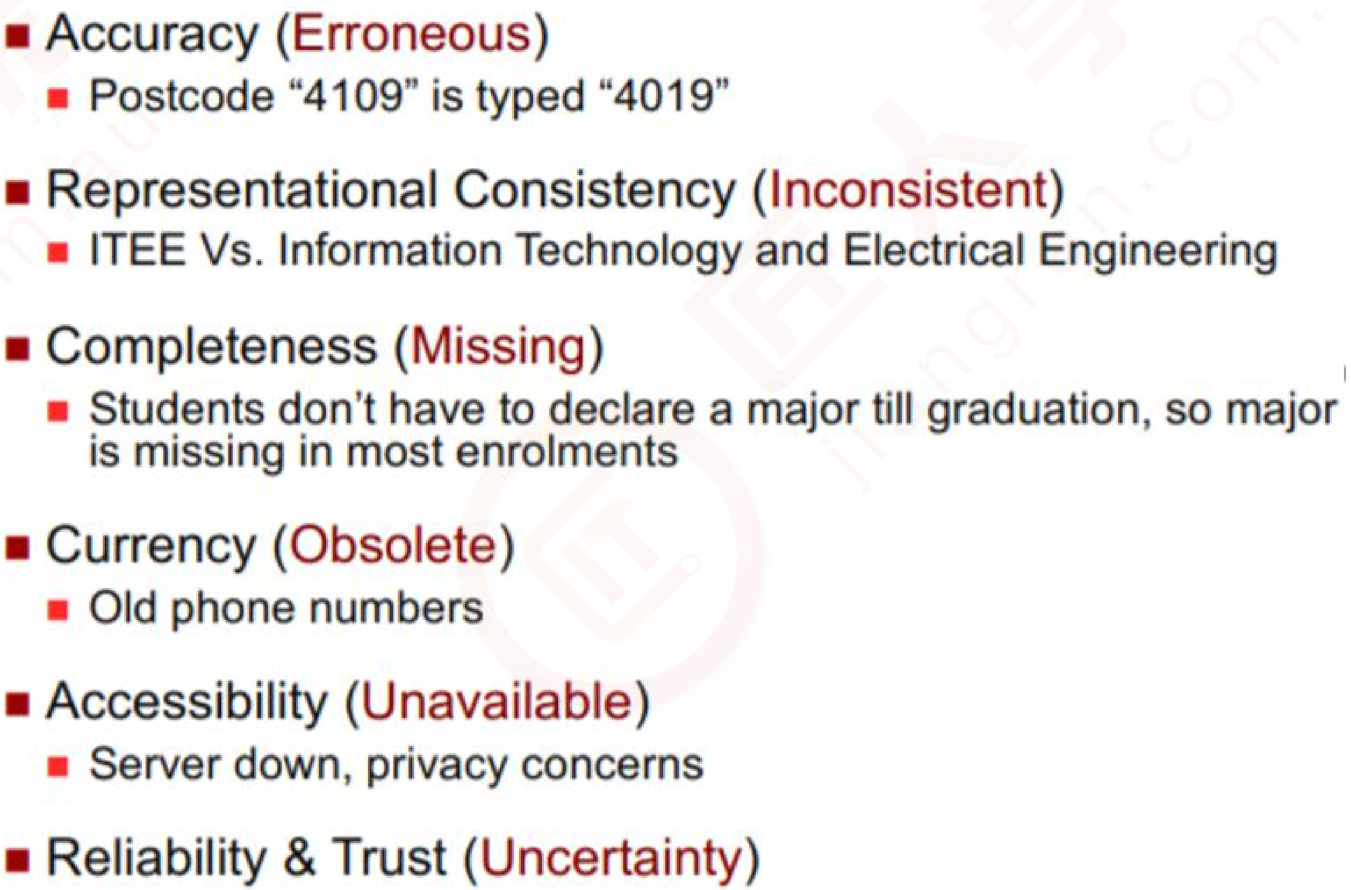


(c) [4 marks] Write a SQL query to generate the global schema. (Hint: using views)

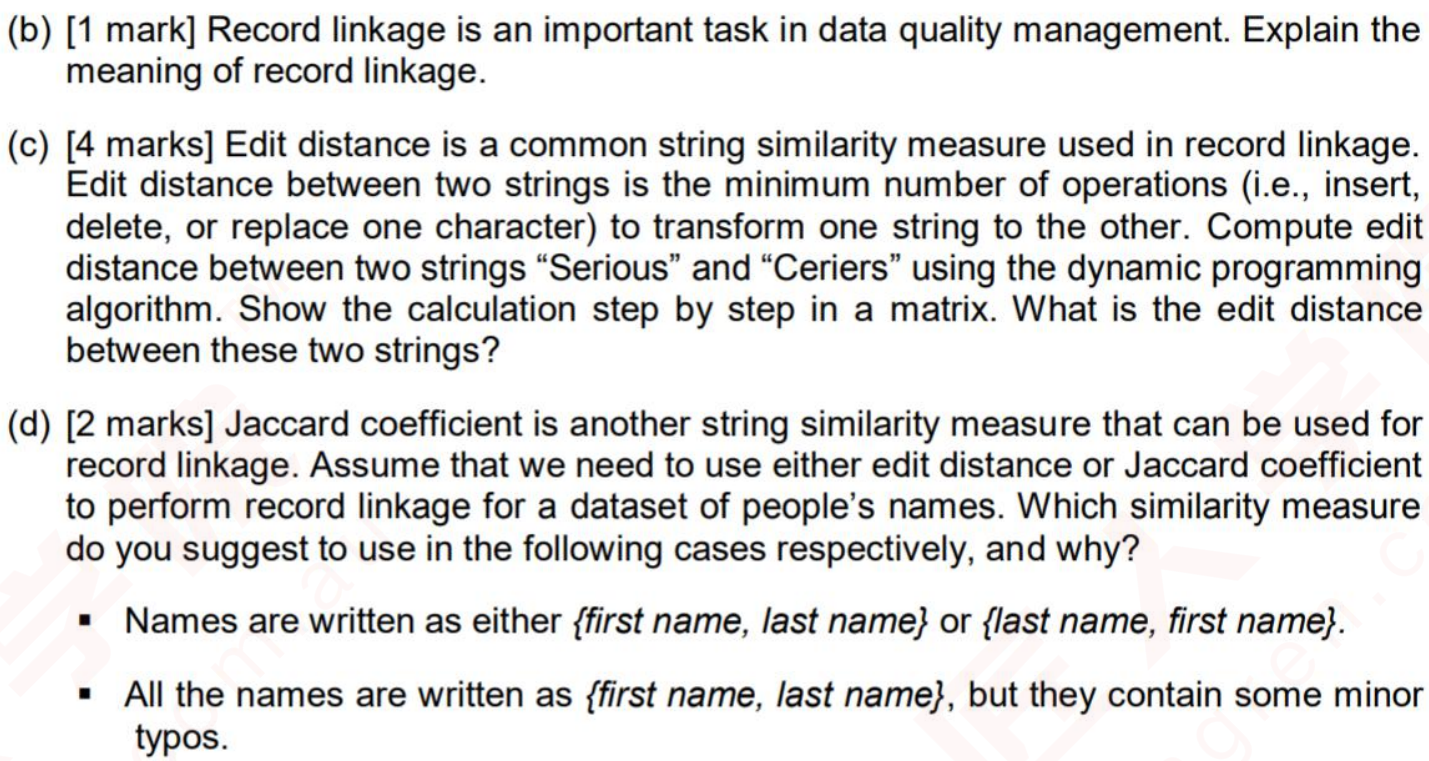
```
CREATE VIEW GLOBALUNI AS
SELECT A.S#, A.Fname, A.Lname
FROM STUDENT AS A
UNION
SELECT B.Sid, B.Sname, B.Sname
FROM PROG_01 AS B
UNION
SELECT C.Sid, C.Sname, C.Sname
FROM PROG_02 AS C
```

Question 5 [15 marks]. Data quality issues need to be addressed before the data can be released for use by other data analysis applications.

(a) [4 marks] Data quality can be measured from various dimensions. Please list at least four data quality dimensions and give one example of data quality problem for each of these dimensions.



- Accuracy (Erroneous)
  - Postcode “4109” is typed “4019”
- Representational Consistency (Inconsistent)
  - ITEE Vs. Information Technology and Electrical Engineering
- Completeness (Missing)
  - Students don’t have to declare a major till graduation, so major is missing in most enrolments
- Currency (Obsolete)
  - Old phone numbers
- Accessibility (Unavailable)
  - Server down, privacy concerns
- Reliability & Trust (Uncertainty)

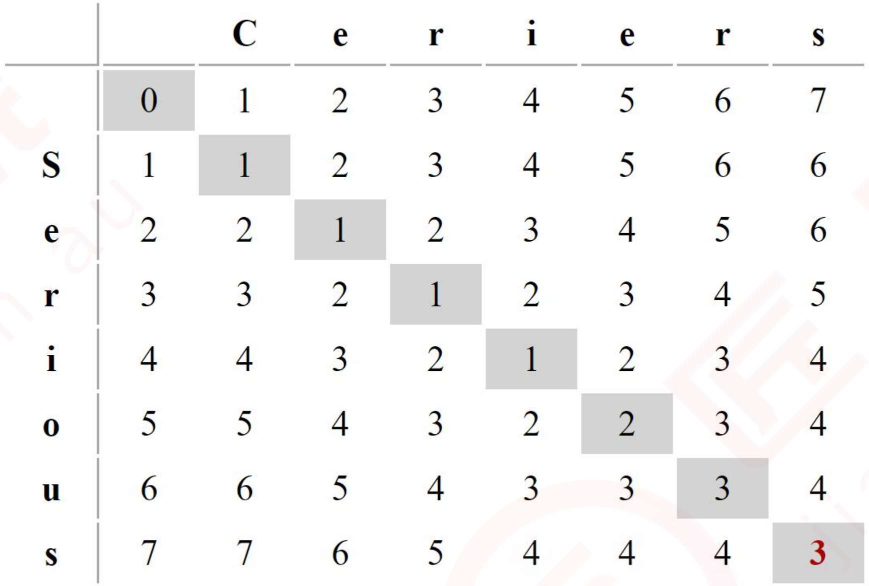


(b) [1 mark] Record linkage is an important task in data quality management. Explain the meaning of record linkage.

(c) [4 marks] Edit distance is a common string similarity measure used in record linkage. Edit distance between two strings is the minimum number of operations (i.e., insert, delete, or replace one character) to transform one string to the other. Compute edit distance between two strings “Serious” and “Ceriers” using the dynamic programming algorithm. Show the calculation step by step in a matrix. What is the edit distance between these two strings?

(d) [2 marks] Jaccard coefficient is another string similarity measure that can be used for record linkage. Assume that we need to use either edit distance or Jaccard coefficient to perform record linkage for a dataset of people’s names. Which similarity measure do you suggest to use in the following cases respectively, and why?

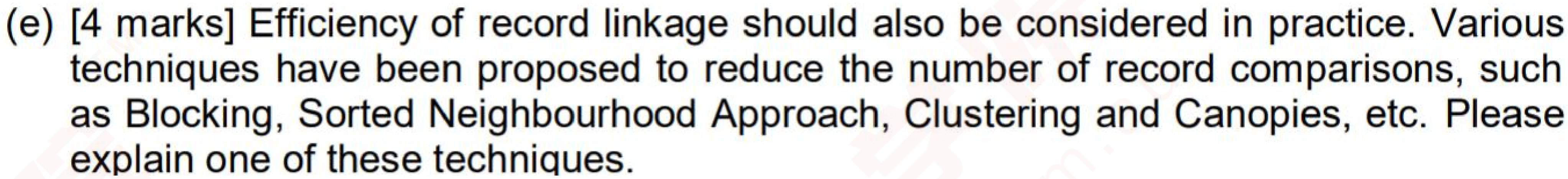
- Names are written as either *{first name, last name}* or *{last name, first name}*.
- All the names are written as *{first name, last name}*, but they contain some minor typos.



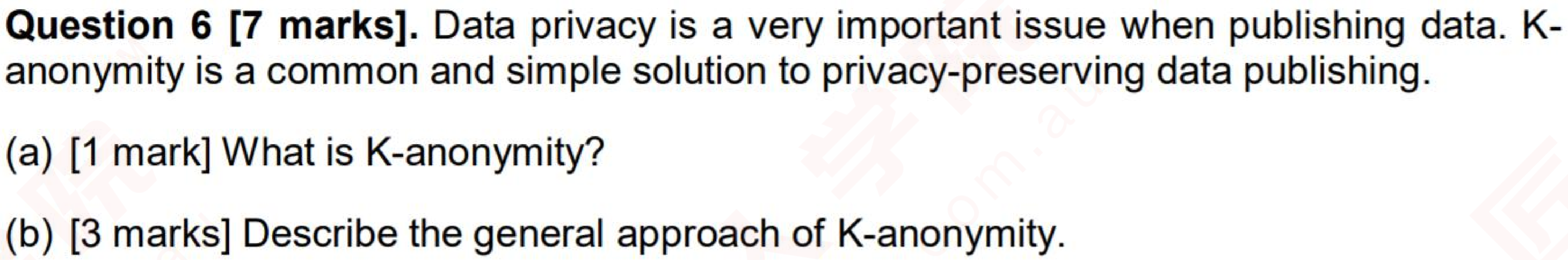
		C	e	r	i	e	r	s
	0	1	2	3	4	5	6	7
S	1	1	2	3	4	5	6	6
e	2	2	1	2	3	4	5	6
r	3	3	2	1	2	3	4	5
i	4	4	3	2	1	2	3	4
o	5	5	4	3	2	2	3	4
u	6	6	5	4	3	3	3	4
s	7	7	6	5	4	4	4	3

Record linkage (RL) is the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, and databases).

Jaccard coefficient. When measuring string similarity using JD, strings are broken into a set of Q-grams. The order of elements in a set does not matter. The edit distance between the same names are large if the first and last names are written in different order.
Edit Distance. Last and first names are written in the same order. ED is defined to be the minimum number of operations to transform one string to another. However, a minor type can cause huge change in jaccard distance since their Q-grams are be completely different. (e.g. “Emily Smith” VS “Emmily Smith”)



(e) [4 marks] Efficiency of record linkage should also be considered in practice. Various techniques have been proposed to reduce the number of record comparisons, such as Blocking, Sorted Neighbourhood Approach, Clustering and Canopies, etc. Please explain one of these techniques.



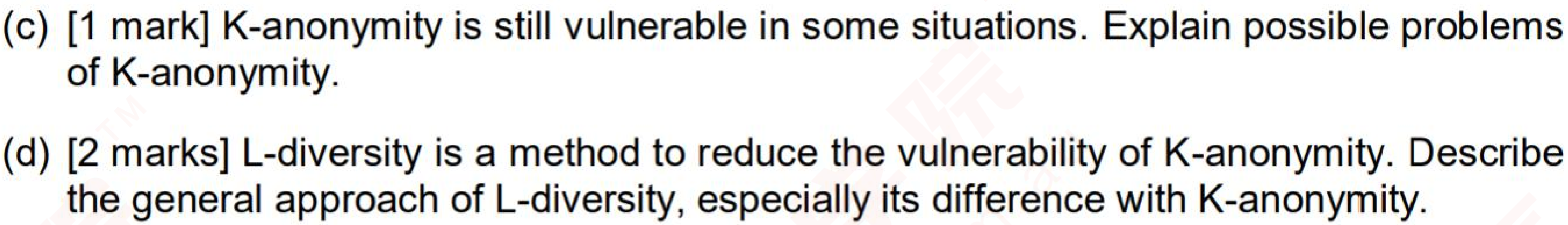
Question 6 [7 marks]. Data privacy is a very important issue when publishing data. K-anonymity is a common and simple solution to privacy-preserving data publishing.

(a) [1 mark] What is K-anonymity?

(b) [3 marks] Describe the general approach of K-anonymity.

K-anonymity is a key concept that was introduced to address the risk of re-identification of anonymised data through linkage to other datasets.

Generalization: individual values of attributes are replaced with a broader category. For example, the value “19” of the attribute “age” can be replaced by “under 20”.



(c) [1 mark] K-anonymity is still vulnerable in some situations. Explain possible problems of K-anonymity.

(d) [2 marks] L-diversity is a method to reduce the vulnerability of K-anonymity. Describe the general approach of L-diversity, especially its difference with K-anonymity.

(c)After processing the data with the k-anonymity technique, each record from the adversary’s knowledge corresponds to at least k records of the table containing sensitive information. However, the sensitive values of the k records can be identical. In such case, the sensitive value can still be predicted.

(d) l-diversity introduces some intra-group diversity to the k records containing sensitive information. L-diversity specifies that the sensitive column of the k records must have at least l different values, whereas k-anonymity specifies that each record known by the adversary can linked to at k records of the table containing sensitive information

1 2 一.

3 4 二.

一 1 2

$2^2=4$ 二 3 4

一 二 12 14 23 24

∅ ∅ ∅