Question 1.

1. Explain what re-identification means. Then, discuss a logical scenario where re-identification is necessary.

Answer:

- Explain:

Re-identification means the task that connects the public name with a data record to the de-identified record. It is necessary for re-identification to verify the contents of a record or provide relative information to the well-being of the subject of a deidentified data record. Furthermore, Re-identification needs approval and oversight during the assignment.

- Discuss:

It is sometimes necessary for scientists to discover the identify of de-identified records. For example, when the de-dentification prevents scientists from assisting people on account of the confidentiality and the privacy . More Specifically, the scientists find patients with a generic maker for a disease that is curable when the scientists conducting an analysis on a collection of deidentified data. In this situation, de-identified records can be re-identified to save lives. While the re-identification should be supervised by a third party with a confidential list maps individuals to their de-identified records.

Question 3.

3. Considering you are the data scientist hired by the national library, you have been requested to understand the behavior of the active library patrons, so that the national library can strategic their next year expansion. Discuss how relational database and/or distributed computing can assist you to achieve your goal. (5 marks)

Relational database:

In relational database, the library books records and students records could be highly structured saved. It can assist me to solve problem like data redundancy and data accessing problem. Furthermore, the data in the relational database is well organized, we can easily recognize the data types or attributes in each data tables. Every data table could contain one class of information. Consequently, if we use relational database, we could easily retrieve and update data in the database.

However, considering the massive data and some unstructured data like pictures, images, video or records should also be analyzed. We'd better use distributed computing to improve the performance of data processing. In distributed computing, the multiple computers are loosely-coupled and multiple computers could work together to solve a large. Consequently,

When we are doing OLAP task regarding processing the library and analyzing the behavior of the active library patrons, we could use distributed computing to split the task and solve each part and then combine the results to improve the processing efficiency.