INFS3200 Advanced Database Systems

Tutorial 6: Data Integration & Data Linkage

Semester 1, 2021

Question 1: Discuss different roles database views play in the following systems:

- Relational database
- Distributed database
- Data warehouse
- Data integration

You should give at least one type of use for each type of systems. Construct simple examples to illustrate your discussions.

Question 2: Consider the following three parties related to the *Olympics* information system involving swimming events:

- Local Organisation Committee (LOC) for an Olympiad
- International Olympic Committee (IOC)
- FINA, the international sporting body governing swimming

Assume the following tables are owned and managed by the LOC, IOC, and FINA:

- LOC: *Result*(EventID, CompID, Position, Time)
- IOC: *Competitor*(ID, Country, Name)
  OlympicRecord(EventID, CompID, Olympiad, Time)
- FINA: *Athlete*(ID, Country, Name)
  WorldRecord(EventID, AthID, Year, Time)

Suppose that all competitors participating the Game organised by the LOC are registered with the IOC already (i.e., LOC.CompID is a foreign key to IOC.ID), and that each competitor is also already registered by the relevant international sporting body, in this case FINA.

(a) Identify possible semantic heterogeneity when integrating these three independently developed databases into *GoldMedallist*, and discuss possible solutions.
(b) Now the LOC wants to integrate these three databases into the following table that shows all swimming event Gold medallists in the Game:

GoldMedallist(CompID,EventID,Time,OlympicRecord,WorldRecord)

Use SQL to construct *GoldMedallist*. (Hint: using views)

(c) Assume the ABC Television wants to create the following table to show all the swimming records set at the Game organised by the LOC:

NewRecord(EventID, CompID, Record, Time)

where *Record* is either "World" or "Olympic". Show an SQL query computing the table *NewRecord*.

(d) Assume that *GoldMedallist* is maintained by the LOC, with any new records updated to the *OlympicRecord* and *WorldRecord* tables being done by the IOC and FINA respectively. It is a requirement that the Olympic records and World records in *GoldMedallist* must be accurate all the time. What "**quality of service**" guarantees do the IOC and FINA need to make to ensure such accuracy in *GoldMedallist*?

Question 3: Answer the following questions about string similarity measures.

(a) What is Edit distance/Levenshtein Metric?

(b) What is Jaccard coefficient?

(c) Calculate the *similarity* between the following strings using Edit distance and Jaccard coefficient (3-gram) respectively.

- "University Queensland"
- "Queensland University"

(d) Which one is better for measuring the similarity between the above two strings, Edit distance or Jaccard coefficient, and why?

---ooo000O000ooo---