

Tutorial 7: Data Cleansing

INFS3200/7907
Advanced Database Systems

Data Cleansing

- Data Cleansing is the act of **detecting** and **correcting** (or removing) corrupt or inaccurate records from a record set.
- Dirty data
 - Missing data
 - Inconsistent data
 - Semantic heterogeneity
 - Obsolete data
 - Schema drift

Question 1

- Consider the following schema
 - Results (EventID, CompID, Pos, Time)
 - Competitors (CompID, Country, CompName, NMedals, NEventsEntered)
 - Records (EventID, Country, CompName)
- Are any of these likely to suffer from data quality? if so, which and why? if not, why not?

Solution

- All the attributes in all three tables are essential to the operation of the Olympics.
- Any errors in any of them would be detected and corrected almost certainly.
- So they are not likely to suffer from **data quality** problems.

Question 2 –

Tickets to Olympic events are allocated to the member countries for sale to their residents. Consider a breakdown of attendance at an event by country derived from the country to which the tickets were allocated. Is this data likely to be clean? Why or why not?

- Can an American buy a ticket in Australia for 2008 Olympics Game?
- Yes
- Therefore, the country to which the tickets were issued is **not a clean measure** of the country of **residence** of the audience.

Question 3

Consider question 3b from the Data Warehousing tutorial. Using **dirty data concepts**, explain why it does not make sense to pivot on Olympiad and Venue in the Olympics data warehouse

- The **venues** are entirely different from Olympiad to Olympiad.
- For example, Beijing National Stadium (91,000 seats) vs. Athens Olympic Stadium (71,030 seats)
- So **attendance by venue** is not comparable from Olympiad to Olympiad.
- An example of *Schema Drift*
 - When the organisation changes and the data schemas change with it, some old data is no longer comparable with current data

Question 4

Consider a tabulation of hits on the Olympics website by the IP address of the ISP of the person making the hit. There is an easy mapping of IP address to country. Suppose you wanted a tabulation of number of hits by country of residence.
a. Is the country determined by the **IP address** a clean way to get country of residence?

- An unreliable way of identifying country of residence.
 - it is possible to disguise one's IP address, e.g. proxy server.

Question 4

Consider a tabulation of hits on the Olympics website by the IP address of the ISP of the person making the hit. There is an easy mapping of IP address to country. Suppose you wanted a tabulation of number of hits by country of residence.

b. Can you think of a way to get cleaner country of residence data? What is its cost?

- Simply asking the user.
 - but he/she can cheat.
- A rewarding mechanism such as lottery
 - higher accuracy
 - significant costs
 - Sufficiently attractive

Question 4

Consider a tabulation of hits on the Olympics website by the IP address of the ISP of the person making the hit. There is an easy mapping of IP address to country. Suppose you wanted a tabulation of number of hits by country of residence.

c. Does the additional value of cleaner data justify the increased cost of obtaining it?

- Hard to see how the value of accurate country of residence data could justify the very significant costs of obtaining it by the methods of b.

Question 5

- Consider the following schema.
 - Result (EventID, OCompID, Pos, Time)
 - WorldRecord (EventID, SFCompID, Time)
- Where Result is managed by the IOC and WorldRecord is managed by a sporting federation. The two organization have different identifier for competitors.

Q5a

Show the schema for a synonym table grouping the different identifiers for athletes.

- Syn (Group, Identifier, SourceOfIdentity)
 - Where SourceOfIdentity is the organization providing the identifier (IOC, FINA, etc.)

Q5b

How would you use the **synonym table** to determine the Olympic competitors who are current world record holders in their Olympic events? Show an SQL query.

```
SELECT R.EventID, R.OCompID
FROM Result R, WorldRecord W, Syn S1 S2
WHERE R.EventID = W.EventID AND
      W.SFCompID = S1.Identifier AND
      S1.Group = S2.Group AND
      S2.Identifier = R.OCompID
```

Q5c

How would you use the synonym table to update the WorldRecord table for events held in the Olympics? Show an SQL query

```
UPDATE WorldRecord W
SET W.Time =
  SELECT R.TIME
  FROM Result R, Syn S1 S2
  WHERE R.EventID=W.EventID AND
        W.SFCompID=S1.Identifier AND
        S1.Group = S2.Group AND
        S2.Identifier = R.OCompID
WHERE (W.EventID, W.SFCompID) IN
  (SELECT R.EventID, S1.Identifier
   FROM Result R, Syn S1 S2
   WHERE R.EventID = W.EventID AND
         W.SFCompID = S1.Identifier AND
         S1.Group = S2.Group AND
         S2.Identifier = R.OCompID AND
         R.Time < W.Time)
```

