

Tutorial 4: System Integration

INFS3200/7907
Advanced Database Systems

1

Tutorial 4 - Overview

- Semantic Heterogeneity
 - Semantic heterogeneity occurs when there are differences in the meaning, interpretation, and intended use of the same or related data.
 - (UQ) courses vs. (QUT) subjects (synonyms)
 - (Bed) room vs. (Teaching) room (homonyms)

2

Background

- We will consider a portion of the Olympics system involving five bodies
 - Beijing Organizing Committee (BOC)
 - International Olympic Committee (IOC)
 - FINA, the international sporting body governing swimming
 - The Olympic Village catering contractor (OVC)
 - A supplier to OVC called CSP

3

Question 1

- Schemas
 - BOC (2008 Olympic Games)
 - Results (EventID, CompID, Position, Time)
 - IOC (All Games except the 2008)
 - Competitors (CompID, Country, Name)
 - OlympicRecords (EventID, CompID, Olympiad, Time)
 - Time – the **best records** ever in Olympic Games.
 - FINA
 - Athletes (AthID, Country, Name)
 - WorldRecords (EventID, AthID, Year, Time)
 - Time – the **best records** ever in World Championships

4

Q1a

There is insufficient information to construct the global schema from the local schema. What is missing the what can be done about it?

- No problem in **semantic heterogeneity** in terms of:
 - Athletes
 - Events
 - Records
- However, there could be different representations of athletes (IOC's CompID vs. FINA's AthID) among different organizations.
- Solution: a modification of table Competitors in IOC is needed
 - Competitors (CompID, Country, Name, **SportingFederationID**)

5

Q1b

Adding the solution to Q1a to the above schema, construct a view GoldMedallist (at Beijing Olympics) as specified

- The Schemas
 - BOC: Results (EventID, CompID, Position, Time)
 - IOC: OlympicRecords (EventID, CompID, Olympiad, Time)
 - FINA: WorldRecords (EventID, AthID, Year, Time)
- Solutions
 - GoldMedallists (CompID, EventID, Time, **OlympicRecord, WorldRecord**)
 - Workings:
 - Create VIEW GoldMedallist
(CompID, EventID, Time, OlympicRecord, WorldRecord) AS
SELECT B.CompID, B.EventID, B.Time, IR.Time, WR.Time
FROM Results B, **OlympicRecords** IR, **WorldRecords** WR
WHERE IR.EventID = B.EventID AND
WR.EventID = B.EventID AND
B.Position = 1
- The Olympics and World Record holders are not necessary to be the same athletes

6

Q1c

Assume we want GoldMedallist to accurately represent the Olympic and World record times. Assume further that GoldMedallist is maintained by the BOC, with any new records updated to the IOC Olympic. What guarantees do the IOC and FINA need to make so that GoldMedallist will be accurate as specified?

- View GoldMedallist is composed of geographically distributed tables Results in BOC, OlympicRecords in IOC and WorldRecords in FINA respectively.
- The GoldMedallist accuracy primarily relies on the tables' currency.

7

Question 2

- Assumption:
 - We have a view computed at the ABC Television site
 - NewRecord (EventID, CompID, Record, Time) where Record is either "World" or "Olympic"
 - In the context of swimming,

8

Q2a

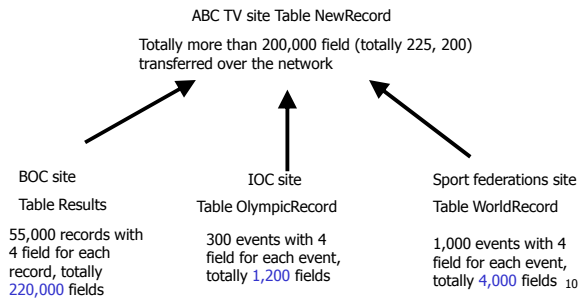
Show an SQL query computing NewRecord. Hint – First compute a view NewWorldRecord, then a view NewOlympicRecord excluding those results in NewWorldRecord. NewRecord is a union of the two with the appropriate constants added.

- In general, the world records are always faster than the Olympic records. We assume that if Olympic records were faster than world record, the world record would be updated instantly.
- Solution
 - CREATE VIEW NewRecord (EventID, CompID, Record, Time) AS
 - SELECT EventID, CompID, "World", Time
 - FROM GoldMedallist G
 - WHERE Time < G.WorldRecord
 - UNION
 - SELECT EventID, CompID, "Olympic", Time
 - FROM GoldMedallist G
 - WHERE Time > G.WorldRecord AND Time < G.OlympicRecord

9

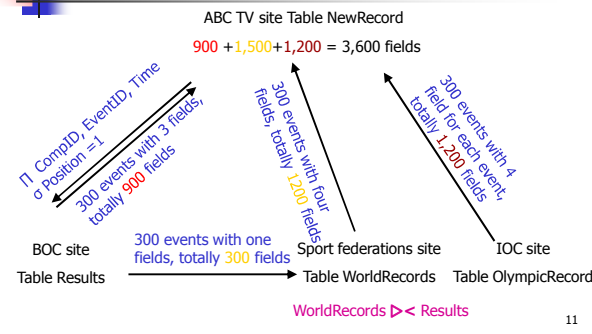
Q2b Continues

Show two different query plans for computing the query Q2a taking into account the sites at which the data resides. Assume the query originates at the ABC TV site, so the result must end up there. One of the plans would move all data to the ABC TV site and perform the query operations there, while the other would attempt a minimum movement of data, making maximum use of semijoins.



Q2b

Show two different query plans for computing the query Q2a taking into account the sites at which the data resides. Assume the query originates at the ABC TV site, so the result must end up there. One of the plans would move all data to the ABC TV site and perform the query operations there, while the other would attempt a minimum movement of data, making maximum use of semijoins.



11

Q2c

Which of the two query plans of Q2b is more economical to compute? How do you know this? In particular, what information is held by which bodies to enable you to make the evaluation? Which of the sites must support semijoin queries? Is it plausible that they would reveal the necessary information in this application? Is it plausible that they would support semijoin queries in this application?

- The second plan is far more efficient than the first. How did we know it in advance?
 - We need to know the table size (Results, WorldRecords and OlympicRecords) in terms of the attributes and rows of records from system catalogs.
 - We also need to know the selectivity of join attributes from global schemas since the join attributes is a key of all three tables–
 - the higher selective, the less likely the tables are used for semijoin because it results in more values transferred over the network.
 - On the other hand, the lower selective, the higher likely the tables are used for semijoin because it results in less values transferred over the network.

12

Selectivity of Join Attributes

Site 1: **Staff**

Staff ID	Name	Dept ID
S123	John	D-11
S234	Ann	D-11
S345	James	D-12
S456	David	D-12

Low Selectivity: more duplicate values under a field

Site 2: **Dept**

Dept ID	Name	Location
D-11	ITEE	Bldg78
D-12	AIBN	Bldg80
D-13	IMB	Bldg85
D-14	ACMC	Bldg69

High Selectivity: less/no identical values under a field

- There are two alternatives for semi-join: Staff ▷< Dept OR Dept ▷< Staff
- Which one is more economical ?
- Yes, Staff ▷< Dept. Why?
- For Staff ▷< Dept, the number of the values under Dept ID across from site 1 to site 2 is two, and the number of the values returned across from site 2 to site 1 is also two, totally 4 fields.
- On the other hand, for Dept ▷< Staff the number of the values under Dept ID across from site 2 to site 1 is four, the returned number is two, so totally 6 fields.

Q2C Continues

Which of the two query plans of Q2b is more economical to compute? How do you know this? In particular, what information is held by which bodies to enable you to make the evaluation? Which of the sites must support semijoin queries? Is it plausible that they would reveal the necessary information in this application? Is it plausible that they would support semijoin queries in this application?

- To construct an optimal query plan, all three sites must expose at least table sizes from their system catalogs.
- If the join attributes on both tables are **primary keys** that uniquely identify each row. Under such circumstance, we will semi-join from the smaller table to the larger table, rather than the other way round, because it results in a smaller outcome table from the semijoin.

14

Question 3

Is it **always** possible to integrate different systems (construct a global schema and have a global query processing system)? Why ? Why not?

- No.
 - Semantic heterogeneity** because systems have their own autonomies such as independent database designs and terminologies.
 - Performance suffering because the system does not know how big the schemas and facilities from other systems are exposed.

15