AG Heterogene Informationssysteme
Prof. Dr.-Ing. Stefan Deßloch
Fachbereich Informatik
Technische Universität Kaiserslautern

Middleware for Heterogeneous and Distributed Information Systems – Exercise Sheet 8

Wednesday, December 17, 2008 – 10:00 to 11:30 – Room 48-379

Forms of Heterogeneity

Information systems are called heterogeneous if they do not provide the exact same methods, models, and structures for data access. Heterogeneity may occur on both, the data level and metadata level. A classification of heterogeneity has been presented in class. The classification distinguishes between technical heterogeneity, data model heterogeneity, syntactic heterogeneity, structural heterogeneity, schematic heterogeneity, and semantic heterogeneity.

US Weather Data Source

Date	City	State	Description	Temperature	Air Humidity
12/11/2008	Los Angeles	California	Fair	59℉	22%
12/11/2008	Paris	Tennessee	Partly Cloudy	31℉	74%
12/11/2008	New York	New York	Light Rain	38℉	89%

European Weather Data Source

Date	CityId	Description	Temperature	Humidity
11.12.2008	1001	Light Snowfall	0℃	90%
11.12.2008	1002	Cloudy	2℃	93%
11.12.2008	1003	Cloudy	0℃	97%

CityId	Name	Country	Population
1001	Berlin	Germany	3.416.300
1002	London	United Kingdom	7.355.400
1003	Paris	France	12.067.000

1. Consider the two (relational) data sources shown above. What forms of heterogeneity do you see? What heterogeneity occurs on the level of data; what heterogeneity occurs on the level of metadata?
2. Recall the middleware infrastructure and the related query languages that have been discussed in previous lectures such as database gateways, federated database systems, SQL, SQL/XML, XQuery, object persistence infrastructure, CORBA, and web services. What forms of heterogeneity can be resolved by these types of middleware?

Mediator-based Information Systems: Garlic

Garlic is a middleware system that provides an integrated view of a variety of legacy data sources, without changing how or where data is stored[1]. Key components of the Garlic architecture are wrappers that encapsulate foreign data sources. Wrappers mediate between the data source and the Garlic middleware and participate in query planning and execution.

Schema of the geographic data source

```
interface Country {
       attribute string name;
       attribute string airlines_served;
       attribute boolean visa_required }
```

```
interface City {
       attribute string name;
       attribute long population;
       attribute boolean airport;
       attribute Country country }
```

Schema of the hotel data source

```
interface Hotel {
       attribute readonly string name;
       attribute readonly short category;
       attribute readonly double daily_rate;
       attribute readonly string location;
       attribute readonly string city }
```

Table 1: Sample Garlic Schema Definitions

Table 1 shows sample Garlic schema definitions for two data sources. Assume that each of these sources is encapsulated by a Garlic wrapper. Say, Garlic is asked to retrieve five star hotels close to the beach in Greek towns with less then 500 inhabitants. Describe the query planning process, the resulting query plan, and the query execution process for the following wrapper capabilities!

1. The geographic wrapper and the hotel wrapper are unable to evaluate predicates. The geographic wrapper does not support joins, i.e. neither the `plan_join()` method nor the `plan_bind()` method are implemented.
2. The hotel wrapper cannot handle equality predicates on strings because it does not adhere to SQL semantics for string equality. However, it treats the predicate `location = 'beach'` as `location LIKE '%beach%'`, which provides a superset of the results of the equality predicate. The geographical wrapper supports local joins, i.e. the `plan_join()` method is implemented while the `plan_bind()` method is not.
3. The hotel wrapper is able to evaluate equality predicates and the geographic wrapper supports both, joins and bind joins.

[1] M. T. Roth, P. Schwarz, Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources, VLDB 97