PEER DATA MANAGEMENT SYSTEM

Philippe Cudré-Mauroux
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA – USA
pcm@csail.mit.edu

SYNONYMS

PDMS; Decentralized Data Integration System

DEFINITION

A Peer Data Management System (PDMS) is a triple $\mathcal{S} = \langle \mathcal{P}, \mathcal{S}, \mathcal{M} \rangle$ where $\mathcal{P}$ is a set of autonomous peers, $\mathcal{S}$ a set of heterogeneous schemas used by the peers to represent their data, and $\mathcal{M}$ a set of schema mappings, each enabling the reformulation of queries between a given pair of schemas.

MAIN TEXT

A Peer Data Management System (PDMS) is a distributed data integration system providing transparent access to heterogeneous databases without resorting to a centralized logical schema. Instead of imposing a uniform query interface over a mediated schema, PDMSs let the peers define their own schemas and allow for the reformulation of queries through mappings relating pairs of schemas (see Figure 1). PDMSs typically exploit the schema mappings transitively in order to retrieve results from the entire network.

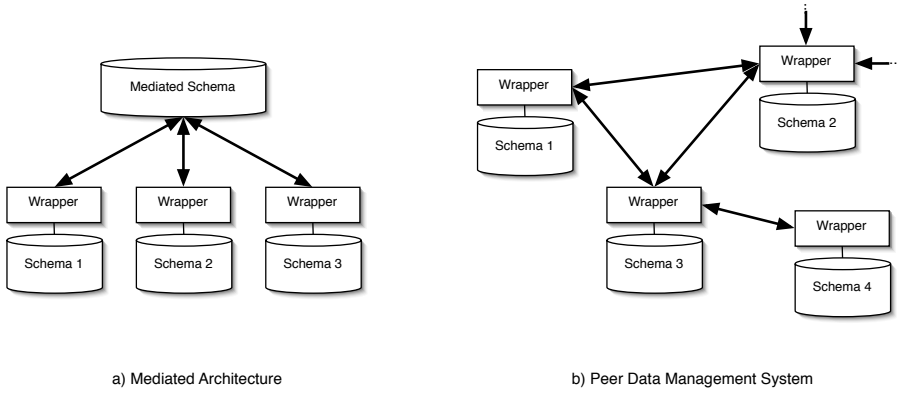

Figure 1: Contrary to the mediated architecture (a), Peer Data Management Systems (b) do not impose any form of centralization but consider instead networks of heterogeneous data sources related to each other through pairwise schema mappings.

Compared to centralized data integration systems, PDMSs suggest a scalable, decentralized and easily extensible integration architecture where any peer can contribute data, schemas, and mappings. Peers with new schemas simply need to provide a mapping between their schema and any other schema already used in the system to be part of the network.
The languages used to define the mappings in PDMSs may vary, but are typically derived from GLAV formulae with extensions to support both inclusion and equality mappings. The Piazza system [1] proposes new algorithms to retrieve certain answers in this context. Hyperion [2] is a PDMS system focusing on relating data not only at the schema level, but also at the instance level through mapping tables. GridVine [3] provides distributed probabilistic analyses in order to automatically detect mapping inconsistencies in PDMS settings.

PDMSs generally use Peer-to-Peer overlay networks to support their distributed operations. Some use unstructured overlay networks [1, 2] to organize the peers into a random graph and use flooding or random walks to contact distant peers. Other PDMSs maintain a decentralized yet structured Peer-to-Peer network [3] to allow any peer to contact any other peer by taking advantage of a distributed index.
The lack of global coordination has raised several questions as to the global properties of such systems in the large. In particular, new complex systems perspectives – such as the emergent semantics approach – have been proposed to characterize the global semantics of PDMSs.

CROSS REFERENCE*
DATA INTEGRATION
PEER-TO-PEER DATA INTEGRATION
PEER TO PEER OVERLAY NETWORKS
EMERGENT SEMANTICS

REFERENCES*

[1] Halevy, A., Ives, Z., Madhavan, J., Mork, P., Suciu, D., Tatarinov, I. The Piazza Peer Data Management System. IEEE Transactions on Knowledge & Data Engineering (TKDE), 16(7), 787-798, 2004.

[2] Arenas, M., Kantere, V., Kementsietsidis, A., Kiringa, I., Miller, R.J., Mylopoulos, J. The Hyperion Project: From Data Integration to Data Coordination. SIGMOD Record 32(3), 53-58, 2003.

[3] Cudré-Mauroux P., Agarwal S., Aberer K. GridVine: An Infrastructure for Peer Information Management. IEEE Internet Computing, 11(5), 36-44, 2007.