MULTI-DATABASE SYSTEMS

SUMMARY

Multidatabase systems, MDBS, arise when integrated access to two or more, possibly heterogeneous, existing databases is required for some new application. The need for systems which can provide data management for multidatabases will increase as the number of independently constructed data collections increases.

This paper presents a definition, taxonomy, and architecture for multidatabase systems and then presents an overview of unresolved problems for their administration. The purpose is to provide a framework for discussion through which to highlight current research efforts at defining and resolving several important sub-areas of multiple database management.
1. MULTI-DATABASE SYSTEMS (MDBS) - A Definition

A **multidatabase system**, **MDBS**, can be defined as:

an integrated data system composed of a collection of
2 or more autonomous datasets and/or databases.

A **dataset** is a collection of strongly related data commonly stored as a
data file with or without an index structure. Examples include:

1. fact data representing atomic attributes for a set of entities.
   
ex.:
   type level: student<name, id, address, age, level>
   data level:   Joan 123 Bergen 25 6
                 Jan 234 Os 18 1

2. documents (free format texts) where each document is represented by
   a long (set of) text field(s) with or without fixed attribute data.
   ex.:
   data level: TITLE: A PAPER ON MULTI-DATABASE SYSTEMS
               AUTHOR: J.C. Nørdbotten
               KEYWORDS: database, databasesystems,
               TEXT: This paper has about 4 pages defining
                      database types, taxonomy, architecture and
                      management problems.

3. images and maps with coordinate/reference data.
4. aggregate and/or time series data.
5. complex objects, hierarchically defined with multiple, possibly
   repetitive sub-objects
   ex.:
   OBJECT (sub-objects (...)())
   document (chapter (section (paragraph (sentence))))

**Databases** are formed by combining related datasets and administered by
**database management systems (DBMS)**. Examples include:

<table>
<thead>
<tr>
<th>DATABASE TYPE</th>
<th>DATA MANAGEMENT SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiple fact datasets</td>
<td>relational (hierarchial, network) DBMS</td>
</tr>
<tr>
<td>document sets (w/ fact data)</td>
<td>information retrieval systems IRS</td>
</tr>
<tr>
<td>map data (w/ fact data)</td>
<td>geographic information systems GIS</td>
</tr>
<tr>
<td>aggregate data (w/ fact data)</td>
<td>statistical DB system SDBMS</td>
</tr>
<tr>
<td>multiple object datasets</td>
<td>object-oriented DBMS OO-DBMS</td>
</tr>
</tbody>
</table>

**Multidatabase systems**, **MDBS**, consist of two or more existing datasets
and/or databases. Component DBs can be traditional (relational/ network/
hierarchic), document, spatial (map, image), statistical, object, and/or
file systems. Proposals for management of multidatabases include:

- distributed database systems
- federated database systems
- extended relational systems
- object-oriented databases
- hyperdata databases

restricted to traditional (fact) DBs
"="
multiple database-type DBs
"="
"="
2. A TAXONOMY FOR MULTIDATABASE SYSTEMS, MDBS

By definition, a multidatabase is composed of multiple datasets and/or databases. Several approaches and prototype systems exist for management of MDBs. She et al [SHET90] present a definition and taxonomy for multidatabases which restricts the term to multiple fact databases implemented with conventional (relational, network, hierarchical) DBMS technology. We will here extend this taxonomy to include system proposals which can also manage multiple database-types.

The taxonomy classifies multidatabase systems along two axes:
1) the degree of system integration and
2) the degree of data type heterogeneity between the component datasets/databases.

1. Independent DBs are not integrated through a data management system, though they can be "integrated" within the context of a user application. Each DB is independently managed using traditional DBMSs, document, statistical, and/or geographic data management systems. Each component DB is of a homogeneous data type as defined in section 1 above.

2. Loosely integrated database systems are those for which the data management systems maintain little or no integration data but do provide a data access language through which the user can communicate with participating DBMSs for specification of required data. These DBM systems have also been termed interoperable [LITW06] or loosely coupled [SHET90]. The component DB set in an interoperable MDBS is (currently) restricted to traditional (fact) databases.

3. Integrated multidatabase systems contain a global descriptor for the data administered by the system and provide an integrated "view" through which the user can request data management services. These systems can be further subdivided according to the degree of heterogeneity between the component databases.

3.1 Homogeneous MDBS contain component DBs which can be described according to some common data model, e.g., relational, semantic, or other, and be administered by the same (kind of) DBMS.

3.2 Heterogeneous DBMSs contain component databases administered by differing DBMSs, with differing data models and/or divergent semantic descriptors. These can be further subdivided according to data-type:

3.2.1 Single type DBs, e.g., fact or object or text, etc.
3.2.2 Multi-type DBs, a combination of 2 or more database-types, e.g., fact plus statistical plus image databases.

Next generation DBMSs [CATT91, SILL91] seek to provide data management for heterogeneous multidatabase systems. These new DBM systems include extended relational, object-oriented, and hyperdata DBMSs.

The extended taxonomy is shown in Fig.2.1 below.
3. AN ARCHITECTURE FOR MDB SYSTEMS

According to our definition, a MDBS is an integrated data system composed of a collection of several autonomous database systems, ie those best defined through branch 3 of the taxonomy above. Integration requires that the MDEM S provide a mechanism for describing heterogeneous component DBs and maintain an integrated description of the component DB set such that a user can view the MDB as a single DB. An adaptation/extension of the ANSI schema structure is appropriate for describing the integration data required. The MDB schema will thus have 6 levels of information:

1. MDB-VIEW level will consist of multiple views representing sub-sets of the integrated MDB.
2. MDB-CONCEPTUAL level describes the integrated MDB providing:
   1. the integrated MDB structure definition
   2. an integrated name set with alias/synonym tables
   3. constraint definitions
3. MDB-INTERNAL level contains the:
   1. distribution data/site locations
   2. MDB to local data mapping functions
   3. constraint processing functions
4. LDB-VIEW defines the public view for the local DB given in the local DB model (relational, network, other)
5. LDB-CONCEPTUAL defines the total local DB structure, name, and constraint set.
6. LDB-INTERNAL defines the implementation structures for the local DB.

Fig.3.1 illustrates a general DBM architecture. Note that a particular MDEM S will not necessarily implement this architecture as separate components but will necessarily contain the implied data descriptions.

4. RESEARCH PROBLEMS IN ADMINISTRATION OF MULTIPLE-DATABASE SYSTEMS

Major research is still required before full, effective multiddatabase management systems can become commercial products. Much of this research is being conducted under the general theme of NEXT GENERATION DDBMS [CAT91,SIGMOD90,SIBL91].

Research areas within multi-database administration currently under study at the Institute for Information Science include:

- integration of heterogeneous distributed database systems
- implementation of semantic DDLs for global schema definition and administration
- identifying synonyms in multiple database systems
- managing constraints between multiple databases
- managing multi-datatype databases
- object-oriented DBMS as a management system for multi-database systems
- OO query languages for multiple database access.
REFERENCES AND BIBLIOGRAPHY


Fig. 2.1: MULTI-DATABASE SYSTEMS - A TAXONOMY

1. MDB-VIEW
   consists of multiple views representing sub-sets of the
   integrated MDB.

2. MDB-CONCEPTUAL
   level describes the integrated MDB providing:
   1. the integrated MDB structure definition
   2. an integrated name set with alias/synonym tables
   3. constraint definitions

3. MDB-INTERNAL
   contains the:
   1. distribution data/site locations
   2. MEB to local data mapping functions
   3. constraint processing functions

4. LDB-VIEW
   defines the public view for the local DB given the
   local DB model (relational, network, other)

5. LDB-CONCEPTUAL
   defines the total local DB structure, name, and
   constraint set.

6. LDB-INTERNAL
   defines the implementation structures for the local DB.

Fig. 3.1 MEBs ARCHITECTURE