Integrated Multimedia Database

Shun’ichi Torii
Kanji Kato
Kazuo Masai

OVERVIEW: In the age of network computing, comprehensive multimedia data such as images, voices and movies are indispensable to information systems. The amount of data is huge in size, and operations of media and structures of store format, are also flexible and expandable. The information systems having the hybrid architecture of the conventional RDB (relational database) and independent media servers for each media have many problems in the administration of information systems and the developments of applications. The HiRDB (Hitachi Relational Database) with an object option introduces new plug-in architecture in which various multimedia-method libraries for storing and retrieving can be installed easily into the kernel of the DBMS (database management system). The repertoire of plug-in is not only full-text search specifying its structure to SGML (standard generalized markup language) documents, similarity retrieval from images, but it is also GIS (geo-spatial information system) and in the near future it will include retrieval from movies.

INTRODUCTION
In the age of network computer usage, eg., the Internet and intranet, more and more multimedia data is needed because of the friendliness and comprehensibility of images, voices and movies. Information systems have to integrate these multimedia data with conventional RDB data.

For a series of integrated multimedia database products, the HiRDB with an object option is the kernel product of the ORDB (object relational database) which has a new plug-in architecture to integrate the various multimedia data with retrieving and storing features for each media type.

We describe the merits of the new plug-in architecture of the HiRDB with an object option and give examples of their applications in network computing.

CHARACTERISTICS OF MULTIMEDIA DATA
Multimedia data require a huge storage area and each type of media requires different methods of store and retrieval. Conventional RDBs cannot accommodate this diversity and the most popular architecture used to do so is the hybrid architecture shown on the left-hand side of Fig. 1. In addition to the RDB server which stores the traditional alpha numeric coded information, the hybrid architecture includes independent special-purpose media servers based on the file storage system of each multimedia type. In this architecture application program, both the SQL API for RDB data and the media server API for multimedia data have to be used (the media server API shown in Fig. 1 is based on the file system). The synchronization between the RDB server and the media server is also a serious problem. These problems make it difficult to develop application programs and to administer multimedia data. Application developers and administrators strongly desire integrated

Fig. 1—Architecture for Multimedia DB.
In the integrated multimedia DB, users can develop application programs for network environments much more easily than with the conventional hybrid architecture of a media server and a RDB.
management in the RDB with multimedia data (right-hand side of Fig. 1).

**ORDB AND PLUG-IN ARCHITECTURE**

An ORDB, which became a topic of discussion recently\(^1\)\(^2\) is designed to solve these problems. A user can define new data types in addition to the existing embedded data types and define the methods that correspond to each new user type where user can develop the new media type gradually. In the near feature, the new SQL3 is going to support this user-defined data type.\(^3\) Possible new data types for each media are the plain text, the HTML (hyper text markup language) document, the SGML (standard generalized markup language) document (The SGML is the international standards of ISO for structured documents.), and image. It is possible to define more minor types such as according to its compression method, store format, and creator application. These minor types can also be defined as subtypes. While in some ways the ORDB is not as good as the OODB (object oriented database), the ORDB is perfectly upwardly compatible with the conventional RDB where user can succeed his own RDB assets. We believe that next major database can only be the ORDB. The SQL3 supports two ways to describe the methods for each user-defined data type but each have a problem. For one way, programming is weak in relation to writing SQL statements (stored procedure) and for the second way external procedure has a heavy overhead to start up.

We developed a new plug-in architecture in which user-defined methods are executed at the DBMS kernel thread which enables the use of user-defined intelligent methods, such as a user-defined index, to improve performance. Fig. 2 shows an example of the

```
CREATE TABLE DocTable (  
DocName varchar (64)  
Texts SGMLTEXT  
Version integer,  
  
);  
(2) Usage of SGMLTEXT in table definition

SELECT DocName  
FROM DocTable  
WHERE Version>1 AND  
contains (Texts, ’Contents {“Hitachi”}’ ) IS TRUE  
(3) Query example using SGMLTEXT function
```

Fig. 2—Example of SGML Document Data Type Usage. User can define user’s own data type and table including the column of the SGMLTEXT type. User can issue the SQL query for the contents of the document using the Boolean function “contains” which specifies the structure of the documents.

![Fig. 3—Plug-In and Parallel RDB Server System. Plug-in functions are executed fully in parallel across the BESs in the benefits of shared nothing parallel architecture with high expandability.](image-url)
SGMLTEXT data type which can contain an SGML document. The definition for the SGMLTEXT includes an attribute called “sgmlcontents” and a plug-in function called “contains” which is written with C language. The “contains” function is actually the user-defined index program for the high-speed and intelligent full-text search.

A HiRDB with an object option is upwardly compatible with the HiRDB and it has the shared-nothing architecture of the parallel RDB servers (Fig. 3). The HiRDB increases the back end server (BES) gradually in proportion to the growth of the DB size and the DB demands. One advantage of the HiRDB for large DBs is that plug-in functions can be executed in parallel at each BES. The HiRDB with the object option has of course all the advantages of the HiRDB; which are the wide expandability (scalability), high reliability, global openness, easy maintainability and powerful administrationality.

AN INTEGRATED MULTIMEDIA INFORMATION SYSTEM USING NEW DATA TYPES WITH PLUG-INS

An example of the enterprise document management system using the previously described SGMLTEXT plug-in appeared in a former article, so we will give the example of the newspaper DB, first. Each article of a newspaper has the document structure shown in Fig. 4. The coded information below the substructure “Bibliography” is stored and can be searched in the RDB table but the text information below the substructure “Contents” requires the full text search capability for search. Fig. 5 shows the full-text search query to find the document which has the keyword “Internet” below the sub-structure “text.” The full-text query is combined with the conventional RDB query about the “published date” and is described in one SQL statement. The full-text search feature has already been developed in the Hitachi’s software product of the stand-alone server system for the Japanese-document search which is based on the algorithm of n-gram indexing. Even for huge number of documents, this n-gram index has a high performance. Its performance for structured full-text search (including relevance ranking, synonyms generation and spelling variants for Japanese) is also very high (Fig. 5). Plug-in architecture can integrate all these elements of full-text search into the ORDB kernel easily and safely.

Another example of an integrated multimedia
CONCLUSIONS

HiRDB with object option and related products can succeed traditional RDB assets of users and integrate multimedia data into the RDB on the basis of an ORDB. We have developed plug-ins for SGML-structured documents and image content management, and we are going to add ones for movie content management and GIS soon. The HiRDB, which has already been successful in the OLTP (online transaction processing) and DWH (data warehouse) area, will continue to offer the infrastructure for enterprise information systems with Internet connections.

REFERENCES