IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 379-381

Implementation of Biological Sequence Analysis and Data Management Operations using Database Cartridge

Ateet Mehta[1] & Bankim Patel[2]

[1]Atos Origin India Pvt Ltd, Mumbai
[2]Shrimad Rajchandra Institute of Management and Computer Application, South Gujarat University, Gujarat, India
Email: [1]ateet.mehta@gmail.com, [2]bankim_patel@srimca.edu.in

ABSTRACT

Biologists understand, analyze and discover great piece of knowledge hidden in the database by performing operations on data. There arises a need of writing efficient algorithms for such operations and tools which implement such algorithms. It requires knowledge of very diverse domains like molecular biology, statistical- mathematical models and computer science. Since biologists are not computer savvy and computer programmers lack the knowledge of molecular biology, availability of a library containing biological operations would be of a greater use to both biology and computer science community. There are various ways of providing API to the user. We designed a biological database cartridge containing all data types and functions which can be used to represent biological entity and perform operations on. Deployment of cartridge to the database management system extends its capabilities for biological data to certain extent.

Keywords: Cartridge, Algorithms, Molecular Biology, Model, API

1. INTRODUCTION

Biologists understand, analyze and discover great piece of knowledge hidden in the database by performing operations on data. There arises a need of writing efficient algorithms for such operations and tools which implement such algorithms. It requires knowledge of very diverse domains like molecular biology, statistical-mathematical models and computer science. Since biologists are not computer savvy and computer programmers lack the knowledge of molecular biology, availability of a library containing biological operations would be of a greater use to both biology and computer science community. This library can be used by developer to develop biological tools and applications. There are various ways of developing libraries and providing API to the developers. We have designed library containing biological data types and operations packaged in a component named as Biological Data Cartridge [1]. We have designed and implemented this cartridge on top of the database management system. The cartridge can be deployed on the underlying database management system which hosts biological database. The use of data cartridge eliminates the need to bring the data outside the database for processing rather it exploits the processing and functional capabilities of the underlying database management system. Further deployment of cartridge to the database management system extends it capabilities for biological domain to certain extent. Implementing biological operations in a biological cartridge also opens a window for pipelining multiple operations logically instead of the current approach of using different tools for different operations.

2. RELATED WORK

The current approach is to use web-based tools to do analysis on the data. Biologist and Researcher have to use different tools for performing different operations. Going further, the application programming interface provided by BioPerl [2] is efficient against file based systems and cannot be seamlessly integrated within the database management systems. BioJava [3] is another development framework designed for sequence analysis. Like BioPerl, it is also a boon to development applications using the API giving inside. Unlike BioPerl, BioJava can be integrated within the database management system. The downside is database management system must provide extensibility interface to install Java classes within. The thousands of web based currently available on internet like BioZone [5] and Moble [6] does not allow pipelining of operations.

3. PROPOSED WORK

We have designed the library in a component known as biological database cartridge. The cartridge contains data types and operations which are built using the native procedural constructs offered by database management systems and hence it can be easily integrated within the database management systems.

3.1. Biological Data Cartridge

Biological data cartridge is basically an array of components designed specifically to model biological data and their operations in one unit which can be deployed to the database server hosting biological database. Biological data cartridge is very well integrated with the database server components that it can exploit the capabilities of the underlying database server. This framework lets you capture domain-related logic and processes associated with specialized or domain-specific data in user-defined data types or objects. These data objects or data types are both the structures that relate different units of information and the operations that are performed on them. The simple names given to data objects often conceal considerable complexity. Data cartridges that provide new behavior without needing new attributes have the option of using packages rather than user-defined types. Either way, you determine how the server interprets, stores, retrieves, and index the application data. Data cartridges package this functionality, creating software components that plug into a server and extend its capabilities into a new domain, making the database itself extensible.

Data cartridge has the following key characteristics:

- Data cartridges are server-based.
- Data cartridges extend the serve
- Data cartridges are integrated with the server.
- Data cartridges are packaged.

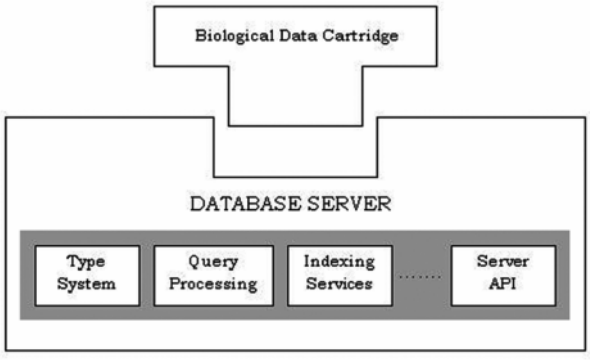

Fig. 1: Biological Data Cartridge

Every database server provides services for basic data storage and data types, query processing, optimization, and indexing and library of general purpose primitive functions. Applications use these services to access database capabilities. However, data cartridges have specialized needs because they incorporate domain-specific data and operations being performed on such data.

To accommodate these specialized needs, these basic services can be made extensible. This means that where standard database services are not adequate for meeting a data cartridge's requirements, you can provide additional services that satisfy the additional requirements of the data cartridge. Data cartridge exploits existing capabilities of the underlying database server and provides domain specific data types and functionalities. Data cartridges can be easily deployed to underlying database server thereby enriching domain specific functionalities of the database server.

3.2. Functionality of Biological Data Cartridge

Biological data cartridge contains the data types which are to represent biological entities like DNA and protein sequences and array of data types to represent features, annotations, cross references used within such biological entities. The operations for sequence analysis like local sequence alignment, global sequence alignment, multiple sequence alignment any many more functions are integrated within the data type. Domain index represents indexes which were created according to the nature of the biological data unlike the traditional indexing. Implementing the domain index on biological data improves data retrieval operations.

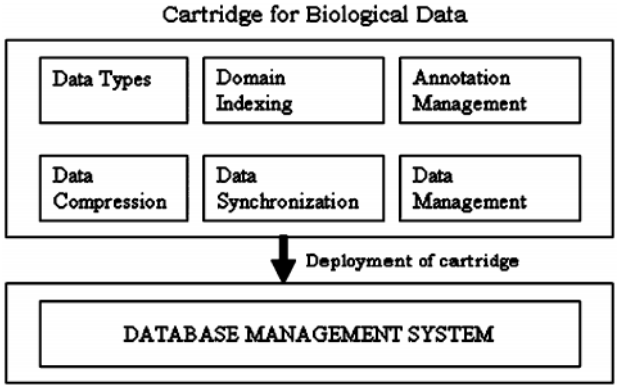

Fig. 1: Extensible Framework

Annotation management provides procedures to maintain annotations attached with the biological data. Data compression module contains algorithm to compress DNA sequence. The compression algorithm is built using hash based data structures and can compress DNA compresses by 75%.

Compression of sequences saves disk storage time and backup time. Data Management module contains procedures for extraction, transformation and loading of data from diverse set of data sources to the local database. It also contains procedures for database backup, restore and recovery. Data Synchronization module contains procedure to synchronize data with standby databases using XML. Once the cartridge is installed on the database management server, the functionalities are available to the database management systems as well to the developers and users. Implementing the data types and operations for biological data using a cartridge on top of the database

management system greatly adds flexibility to the developers, users and enhance the overall performance.

4. CONCLUSION AND FUTURE WORK

Implementation of the biological data types and operations in a form of biological data cartridge is a novel approach of providing application programming interface to the developers. Besides the flexibility given to the developers, deploying the cartridge extends the capabilities of the database management server to the biological domain to the certain extent. Biological data cartridge makes it possible to pipeline many operations as one unit of operations as can be exploited using the normal SELECT statement or by combining series of function calls in one call.

REFERENCES

[1] Data Cartridge Developers Guide, Oracle Corporation, 2010.

[2] BioPerl- http://www.bioperl.org/wiki/Main_Page, Page Retrieved on 15th Feb, 2010.

[3] BioJava CookBook, http://www.biojava.org/wiki/BioJava:CookBook page retrieved on 18th April, 2010.

[4] Zoe Lacroix, Terence Critchlow, *Bioinformatics: Managing Scientific Data*, Elsevier, 2003.

[5] BioZone Research Technologies, www.biozone.co.in Page Retrieved on 31st December, 2009.

[6] Mobyle, http://mobyle.pasteur.fr/cgi-bin/portal.py?form=coderet, page retrieved on 14th Feb, 2010.

[7] A.C. Siepel, A.N.Tolopko, A.D. Farmer, "An Integration Platform for Heterogeneous Bioinformatics Software Components", *IBM Systems Journal*, **40**, No-2, 2001.

[8] S.Davidson, C. Overton, V.Tanne, "BioKleisli: A Digital Library for Biomedical Researchers", *International Journal of Digital Libraries*, **1**, No-1, 1997.

[9] L.Wong, "Kleisli, "A Functional Query System", *Journal of Functional Programming,* **10**, No-1, 2000, pp: 19-56.