Astronomical Data Analysis Software and Systems XVII **P1.2**
ASP Conference Series, Vol. XXX, 2008
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

Construction of Multiple-Catalog Database for JVO

Masahiro Tanaka, Yuji Shirasaki, Masatoshi Ohishi, Yoshihiko Mizumoto

National Astronomical Observatory of Japan, Tokyo, Japan

Yasuhide Ishihara, Jumpei Tsutsumi, Yoshihiro Machida

Fujitsu Ltd., Chiba, Japan

Hiroyuki Nakamoto, Yuusuke Kobayashi, Michito Sakamoto

Systems Engineering Consultants Co. Ltd., Tokyo, Japan

Abstract. We present a development of an efficient query system for the Multiple Catalog Database. In a case where a VO user collects all the available information about an astronomical object, it is necessary to send queries to all the VO services. However, this would be inefficient since most services tend not to have data on the target object. We thus developed a new query system for users to efficiently find all the available information from multiple catalogs. This system collects basic data (coordinates and fluxes) of astronomical objects from the VO data services and major astronomical catalogs, and store them into a local database built with PostgreSQL. The design of this database is one of key issues; an efficient query mechanism for more than billions of objects is required. We employed the Table-Partitioning technique and developed a method to build queries for the partitioned tables. We compared our method with the partitioning function of PostgreSQL and found that our method is more efficient by a factor of 7 to 150. We discuss the architecture of this system.

1. Introduction

One of common tasks to astronomers is to search for available observational data on the target astronomical objects that they are studying. In order to collect all the available information about the objects from VO, it is necessary to send queries to all the VO services. This method would be inefficient since only a part of data services will provide data on the target objects. We thus developed a new Multiple Catalog Database in order to efficiently find all the available information from multiple catalogs. This database consists of a single table in which we store multiple catalogs collected through VO services and also obtained manually. In order to develop this database, we investigated the following issues: (1) Fast query mechanism for large number of objects, and (2) Design of unified table to store various kinds of astronomical catalogs. This database is available to public through JVO system (Ohishi et al. 2006, Shirasaki et al. 2006, 2007).

2. Design of Table Partitioning

Query performance for a large database is one of key issues to build the Multiple Catalog DB. For example, 2MASS All Sky Catalog has 4.7 billion objects, SDSS DR6 Catalog has 287 million objects. If we simply store more than billions of objects into Relational Database Management System (RDBMS) as a single table, we do not obtain enough query performance. Therefore we employed the Table-Partitioning technique for large databases. In order to make position search faster, we used HTM (Hierarchical Triangular Mesh; Kunszt et al. 2000) as a partitioning key. Objects are grouped by upper HTM level 6, corresponding to $8 \times 4^6 = 32768$ regions, and stored data into distinct tables, named as `psc_32768`, `psc_32768`, ... , `psc_65535`. If this system receives the following simple region query:

```
select ra, dec, j_m
  from psc where Region('Circle 0 0 1');
```

then the region condition is converted to the condition of table selection and the ranges of HTM lower ID as:

```
select ra, dec, j_m
 from ( select * from psc_63488 where htm_id between 0 and 65535
  union select * from psc_63488 where htm_id between 217088 and 218111
  union select * from psc_47104 where htm_id between 0 and 65535
   ...
  ) psc;
```

We implemented a query converter in Java using HTM Java library developed at JHU and used PostgreSQL ver 8.2 for RDBMS.

3. Performance of Table Partitioning

To evaluate the performance of our method, we compared it with the partitioning function which PostgreSQL was equipped with as of the version 8.1. For this test, we used 2MASS All Sky Catalog. We performed this test on machine with Pentium-4 2.8GHz and 2GB memory. The result (Table 1) shows that our method is 7-150 times faster, suggesting that our partitioning method yields enough performance for large databases. We note the following points:

1. The elapsed time in our method does not include time to convert region conditions. It is less than 0.5 s.
2. To test the PostgreSQL partitioning function, we reduced the number of partitioned tables to 2048, 16 times smaller than our method. This is because PostgreSQL has a restriction on the number of tables to transact simultaneously since this transaction consumes a lot of shared memory. In contrast, our method allows us to transact larger number of partitioned tables.
3. Elapsed time of the PostgreSQL partitioning function depends on the number of HTM condition, rather than the search radius.

4. Design of the Multiple Catalog DB

Astronomical catalogs are provided in their own table formats; one record for one object and multiple columns for multiple band fluxes. Multiple catalogs in

Table 1. Measured elapsed time to search partitioned tables

Search radius	Result objects	Elapsed time (sec)			# of HTM conditions	
arcmin	#	Postgre SQL	Our method	ratio	Postgre SQL[a]	Our method[b]
1	2	6.46	0.04	154	32	32
10	165	3.81	0.03	127	16	16
60	6697	6.47	0.11	60	32	32
100	26720	2.02	0.31	7	4	16
180	57246	9.04	0.71	13	48	72

[a] the number of HTM `between` conditions in `where` clause

[b] the number of subqueries concatenated with `union`

such *object-based* formats are difficult to combine into a single table. We thus employed a simple *flux-based* format shown in Table 2 for the Multiple Catalog DB. In our format, each record has *only one* flux column. Multi-column flux data of original catalogs are decomposed and stored into distinct records. This method has been employed for the Catalog of Infrared Observation (Gezari et al. 1999). Multi-wavelength information of objects such as the SED (Spectral Energy Distribution) is obtained by grouping flux records in terms of position. Although the Multiple Catalog DB includes only a part of information from original catalogs, further information can be obtained from the original catalog whose access point is recorded in the `link_ref` column.

Table 2. Table design for the Multiple Catalog DB

category	column	description
Object	`id`	Object ID
	`name`	Object name
Position	`ra`	Right Ascension
	`dec`	Declination
	`pos_err`	Position Error
	`htm`	HTM index
Wavelength	`band_name`	Band name
	`band_unit`	Unit of band
Flux	`flux`	Flux value in catalog
	`flux_err`	Flux error
	`flux_unit`	Unit of flux
	`flux_srch`	Flux in Jy
Reference	`link_ref`	Link URL to reference
	`org_id`	ID in original catalog
	`cat_id`	Catalog ID

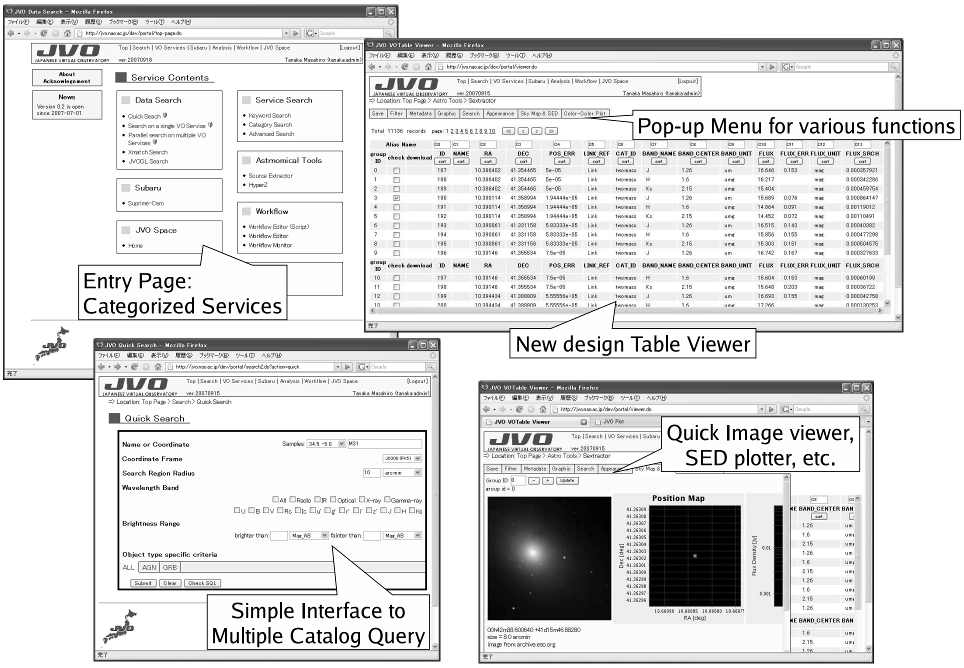

Figure 1. New User Interface of JVO portal

5. New User Interface of JVO Portal

The Multiple Catalog DB is build in the JVO Portal system and available to public users. We also developed new user interface (Figure 1) including a simple query interface to search this DB, categorized services with improved accessibility, and a table browser equipped with many new features.

Acknowledgments. This work was supported by the JSPS Core-to-Core Program and Grant-in-aid for Information Science (15017289, 18049074 and 19024070) carried out by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Gezari, D. Y., Pitts, P. S., & Schmitz, M. 1999, Catalog of Infrared Observation, ed. 5

Kunszt, P. Z. et al. 2000, in ASP Conf. Ser. 216, ADASS IX, ed. N. Manset, C. Veillet, & D. Crabtree (San Francisco: ASP), 141

Ohishi, M. et al. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 375

Shirasaki, Y. et al. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 456

Shirasaki, Y. et al. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), [O10.2]