THE SDSS SCIENCE ARCHIVE

R.J. Brunner
Dept. of Physics & Astronomy, The Johns Hopkins University

1. Introduction

Designing and implementing the Science Archive for the Sloan Digital Sky Survey (Gunn *et. al* 1992) has presented several unique architectural obstacles. First, the final archive will be large; the cumulative data products will exceed tens of Terabytes. Second, the data will be complex; extracted parameter catalogs will contain links to images, spectra, and objects in other wavelength catalogs. Third, the archive will be widely distributed allowing transparent world-wide access to the available data. Finally, the archive is expected to remain viable well into the next century.

The SDSS Science Archive will consist of four main components: a photometric catalog, a spectroscopic catalog, atlas images, and spectra. The photometric catalog is expected to contain at least one hundred million galaxies, one hundred million stars, and one million quasars. Each detected object will have measured parameters (*e.g.*, magnitudes and profiles) recorded as well as an associated image cutout for each of the five bandpasses. The spectroscopic catalog will contain identified emission and absorption lines and one dimensional spectra for one million galaxies, one hundred thousand stars, one hundred thousand quasars, and about ten thousand clusters.

During the design of this archive, several novel approaches were utilized: the object-oriented design and implementation of the persistent storage and transfer of the actual data, a geometric strategy in the organization of the inherently multidimensional spatial and flux information, and the introduction of custom analysis applets that filter the extracted data into a more manageable stream. The knowledge gained in the development of this archive will help not only anyone who is interested in the internal

B.J. McLean et al. (eds.), New Horizons from Multi-Wavelength Sky Surveys, 471–473.

https://doi.org/10.1017/S0074180900129341 Published online by Cambridge University Press

architecture of the SDSS Science Archive, but also those who are developing similarly large astronomical archives.

2. Object Oriented Design & Implementation

In an effort to simplify the portability and maintenance over the expected lifetime for this archival system, we have attempted to adhere to proven Object Oriented Design and Implementation techniques throughout this project. Except for the provided Graphical User Interface, all of the software is written in C++. The various subsystems have been designed and modeled using the Rumbaugh/OMT (Rumbaugh *et. al* 1991) Object Modeling Technique. This design strategy provides both a clear picture of the dynamic and static interrelationships between the objects within a subsystem and also a legacy snapshot of the actual architectural framework.

During the project's early work on archival implementation, Fermi National Accelerator Laboratory evaluated several relational, object/relational hybrids, and object oriented database systems (OODBS), primarily in regards to system performance. The general consensus was that OODBS provided the best performance as well as the optimal implementation model for scaling to the Terabyte regime. After several years of working with this emerging technology, the project settled on Objectivity/DB to satisfy all persistent storage and querying requirements. An additional benefit provided by Objectivity/DB is the ability to control the clustering of data on the storage media; a fact which is crucial to our geometrical indexing strategy.

3. Geometric Strategy

As with other types of data, Astronomical data often contain a numerical subset (*i.e.*, spatial coordinates) that are indexed in order to expedite a certain class of queries. Unfortunately, traditional indexing techniques have several shortcomings. First, they usually must be restricted to a few parameters; otherwise, they begin to match the actual data in physical size and complexity. Second, the actual index is unable to provide additional information about the underlying data, such as providing a coarse grained density map. Finally, current archive queries are limited to simple ranges of parameter values, while the desired query may be more complicated.

Using ideas from the field of Computational Geometry, we have developed an indexing strategy that while still providing the benefits of a traditional indexing scheme, also provides accurate predictions of query volumes and times, a snapshot of the spatial relationships that exist within the dataset, and aids in the quantization of the data on the storage media. Our strategy utilizes a spatial data structure (Samet 1990) to provide a

coarse grained density map of the actual subset of the data that will be indexed. This density map is then encapsulated in a tree-like structure where each node on the tree conceptually represents a sub-volume within the entire volume occupied by the data. Thus, the root node represents the entire dataset, and the leaf nodes represent the terminal cells in the density map. All objects which lie within the leaf node's boundaries are then quantized and stored contiguously on the storage media in an attempt to ensure efficient cache hits by the object request broker within the data warehouse.

The geometrical indexing strategy can naturally incorporate the actual query, resulting in a more powerful search mechanism. Rather than limit a user to parameter cuts, linear combinations of attributes form the query primitive within our system. These linear combinations can then be combined using Boolean Algebra to form complex polyhedra that can carve out complicated volumes within the available parameter space. In order to simplify spatial queries, we work with a Cartesian projection of the spherical astrometric coordinates. This simplifies coordinate conversions, and reduces spherical proximities to a linear combination of the Cartesian coordinates.

4. Analysis Filters

An often over-looked problem inherent in large archives is the management of the extracted data, which can often swamp the resources of many users. This problem is compounded with the inclusion of the network latency. When all that is required is a simple plot or calculation, a user does not need the entire dataset produced during the extraction phase of a query. As a result, we have developed a toolkit of analysis applets that can filter the extracted data.

Acknowledgements

First I would like to thank the rest of the SDSS Science Archive team at JHU, especially Alex Szalay and Kumar Ramaiyer. I also wish to thank Robert Lupton, Don Petravick, Steve Kent, Jeff Munn, and Brian Yanny for stimulating discussions. In addition, I would like to acknowledge the SDSS for funding this project.

References

Gunn, J.E. and Knapp, G.R. 1992 Publ.Astron.Soc.Pacific , 43, 267

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorenson, W., 1991, "Object Oriented Modeling and Design," Prentice Hall.

Samet, H. 1990 "The Design and Analysis of Spatial Data Structures," Addison-Wesley