Data Catalog 2.0

Overview

Currently the data catalog is being used by Fermi and EXO, and used as a side effect of using the pipeline by several other groups (CDMS, CTA).

LSST is interested in using the data catalog (for test data at least, and possibly for DESC and data handling if its dependence on oracle was removed). The download manager is currently used by SSRL and they could potentially use more of the data catalog for making data which is not currently accessible from JCSG directly available from SLAC (although they do not currently seem very interested in this).

The data catalog is also potentially usable as part of a future photon science data portal.

This page explores the possibilities for future development of the data catalog -- assuming for the moment that we have unlimited resources to achieve this.

Current features

- Ability to define logical folders to arrange data independent of physical location
- Ability to define groups of files of the same format (logical datasets)
- Ability to register multiple generations (versions) of the same file
- Ability to register multiple physical locations for the same file
- Ability to associate arbitrary meta-data with files, folders and groups
- Search based on file location, attributes and meta-data.
- Crawler which checks files existence (post registration) and can automatically extract meta-data from files
- Web, command line and Java API interfaces.
- Download manager for downloading large numbers of files. Allows resume of partial downloads and download of updated/new files. Web based and (poorly-supported) command-line version

See also outstanding issues.

- Symbolic links (or similar)
- Search in web interface
- Plugin to integrate experiment specific data
- Keep track of relationships between files (for example to kee track of calibration files)
- Ability to "tag" datasets (or runs) (e.g. Good/Bad/Golden)
- Split access by experiment
- Some way to make experiment specific pages appear
- Run specific meta-data
- Extensible configuration of data types
- File formats and types should be experiment specific
- Download manager authentication across applications
- Extensible "open-source"

Data Catalog 2.0

Improved modularity

The data catalog currently has 5 different components

1. Web interface
2. Database back-end ("middleware")
3. Line mode client
4. Download manager (web based and line-mode)
5. File "crawler"

The download manager and crawler are not strongly coupled to the rest of the system, but the top 3 items are fairly tightly coupled, in particular they all access the database directly. Ideally the web interface would access the database via an abstraction layer, so that the web interface could be used with any backend providing similar functionality. If we add restful interfaces to the web interface it would probably make sense for the line mode client to use those interfaces (this would move the current restriction that the line-mode client can run only at SLAC).

New Features

- **Access control lists**. For Fermi there was no need to have fine grained access control, you can either login and access the data catalog and related data or you cannot. For EXO we added an ad-hoc method using meta-data to "hide" the EXO data from non-EXO users. For more general use some more flexible method of access control would probably be desirable.
- **Enhanced web interface** using HTML5 features.
    - Create elements of the tree only as they are browsed for faster application loading
    - Time-consuming data (such as total size of all files in a folder) should be populated into the web interface asynchronously
        - More effective use of http/tomcat caching?
    - Drag and Drop: it should be possible to drag files between the local file system and the data catalog. Potentially this would remove the need for the separate Java download manager (which does not work on mobile devices/tablets).
        - Limits on number of connections?
- **Ability to drill into datasets.** Currently the functionality of the data catalog stops at the file level. To access any data within the files requires that the files be downloaded, or that the file paths be accessed and fed into some separate tool which analyses the data. Ideally the web interface would also make it possible to drill into the data files themselves, for example to:
    - View images or tables within a fits file
    - View histograms within a root file
    - Tabulate/explore tuples within files
- Any such interface would have to allow easy extension to support new file formats and tools. Ideally it would be extensible to allow increasingly complex data analysis tasks to be run directly through the web interface.

- **Restful interfaces**. The addition of restful interfaces which can return data as JSON or XML (or plain text) would make interfacing to the data catalog from different languages, in particular JavaScript and Python much easier. Some trial restful interfaces are currently being developed for EXO.
- **WEBDav interface**: (but webdav is kind of old and not well supported by different OS's)

Database Independence

Currently the data catalog is tightly coupled to Oracle through the use of

- Java stored procedures
- Hierarchical queries
- Miscellaneous use of oracle specific SQL

Ideally this dependence could be removed through the use of a proper object-relational mapping layer, such as hibernate. How hard would it be to migrate existing data if we used hibernate?

Integration with other tools

- IRods? SAM? Globus Online? DIRAC?

Implementation Language

Currently the data catalog is implemented entirely in Java, and integrated into the SRS web application framework. When deciding on a new version it would be worth considering whether this remains the right approach. Additional HTML5 functionality would require the adoption of some web application framework for which there are many possibilities (GWT, jQuery, ...). Splitting the web front-end from the middleware would make it possible at least to implement different components in different languages. Fully functional restful interfaces would make interfacing from any language much easier.

Links

Some links to conceivably relevant tools found on the internet:

- http://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&strail=false&bcs=d&nselm=h&met_y=ny_gdp_pcap_cd&scale_y=lin&ind_y=false&rdim=country&idim=country:CHN&ifdim=country&tstart=-315424800000&tend=1262498400000&hl=en&dl=en&iconSize=0.5&icfg
- http://filtergraph.vanderbilt.edu/tools/default/portal/22
- http://astrostatistics.psu.edu/vostat/
- http://datavisualization.ch/tools/
- http://www.html5rocks.com/en/tutorials/file/filesystem/