Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

Chandra Data Archive Download and Usage Database

Emily Blecksmith, Stéphane Paltani, Arnold Rots, Sherry Winkelman

Chandra X-Ray Center, Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA (USA)

Abstract. In order to support regular operations, the Chandra Data Archive Operations Group has developed a database that records and monitors the user activities that affect the archive servers. This database provides information on the number of users that are connected at a given time, what archive interfaces they use (we have several), and how much and what type of data is being downloaded.

The database consists of three tables populated by a set of four scripts that parse the archive server logs, the ftp logs and the login logs. User activity can be tracked through each of those logs, making information from a given connection easily accessible.

With this tool, the Archive Operations Group will be able to gather statistics and monitor trends, which will improve the accessibility of Chandra data.

1. Introduction

1.1. The Chandra X-Ray Observatory and Data Archive

The Chandra X-ray Observatory (CXO), a spacecraft launched in July 1999, carries an X-ray telescope with two main instruments: the Advanced CCD Imaging Spectrometer (ACIS) and the High Resolution Camera (HRC), supplemented with optional transmission gratings. The mission is operated by the Smithsonian Astrophysical Observatory at the Harvard-Smithsonian Center for Astrophysics in Cambridge, MA, under contract with NASA. The operation covers the entire institutional life cycle of the observations.

The Chandra Data Archive[1] (CDA) Group's fundamental tasks are the maintenance of the Chandra archive and archive servers and the distribution of the data products. These responsibilities can be broken into three major components: the database and database servers; the archive and archive servers and the interfaces for the ingest; search and retrieval of data (Rots et al. 2002).

1.2. The Chandra Data Archive Download and Usage Database

The Chandra Data Archive (CDA) Download and Usage Database is a comprehensive collection of all user activity on the Chandra Data Archive's search

[1] `http://cxc.harvard.edu/cda`

© Copyright 2003 Astronomical Society of the Pacific. All rights reserved.

and retrieve (SR) servers. Users can connect to the archive and browse or retrieve data using a number of different interfaces: Chaser, a stand-alone Java application which provides the most retrieve/browse flexibility; WebChaser, a web-based version of Chaser; the Provisional Retrieval Interface (PRI), a single CGI script; the CDA FTP Staging site and the brand new CDA Anonymous FTP site.

The first three interfaces use the CDA servers to retrieve data products and either put them on the CDA FTP staging site or transmit them directly (Chaser only). In addition to staged products, the FTP staging site contains a number of pre-packaged distributions including special observations such as deep fields, the calibration database, ephemerides etc.

This database contains detailed information taken from various logs regarding user connections and activities. Three different logs are used to populate the database: the archive server activity log; the archive server login log and the FTP log. Four scripts parse the logs and populate the three tables in the database: 'SR_connections'; 'SR_usage' and 'downloads'.

The database can be used for a number of purposes, such as scheduled public statistics reports or specialized information for operational services. When unexpected failures and problems occur it is convenient to have all usage information in one place that can be displayed graphically.

2. Why is this Database Useful?

Maintenance of server health is one of the primary concerns of the CDA Group at the CXC. Over time, it has become clear that in addition to monitoring internal server activities, outside users (individuals downloading Chandra data) must be tracked as well. The Chandra Data Archive Download and Usage Database consists of very detailed user information from several server logs. This database is useful in a number of different ways.

Because all information is in one place and can be easily queried, we have easy access to fundamental information. Answers to questions such as how many connections are made per day, how much data is transferred daily, monthly, yearly, which interface is used most often, which data set is most popular, etc, are readily available.

This database also enables us to monitor trends which will ultimately lead to better performance and easier, faster access to Chandra data. By tracking server errors users encounter and by observing typical user behavior, corrections and modifications to the system can be more easily and accurately made.

Finally, the download and usage database is helpful in diagnosing unexpected problems. Knowing who was connected and what was being done at the time of a server crash or other strange event can give us important clues as to what might have gone wrong.

3. Population of the Database

All information in the database comes from three types of logs: the Archive Server Activity log, the Archive Server Login log, and the FTP log.

The activity log contains connect/disconnect times, user name, host name, user activity (retrieve or browse) and data requested. Entries concerning a retrieve or a browse have to be matched with the correct connection entry using the time, process id, user name and host that appears in both log entries. The data and activity information is then extracted from the 'procname' line. The information gleaned from this log populates the SR_connections and SR_usage tables. Here is an example of the activity log format:

```
10/01/02 17:12:52:644806 ConnectHandler, user guest, from
  host foo/000.000.00.000/foo-Sun0, server pid:  89
10/01/02 17:12:52:653101 LangHandler, user guest, from host
  foo/000.000.00.000/foo-Sun0, server pid:  89
dataset=flight
operation=retrieve
obsid=605
Server message:
10/01/02 17:12:52:760822 Message number:  5701 Severity:
  10 State:  2 Line:  1 Server sqlsao
Message String:  Changed database context to 'arcsrv'.
procname:  ret_primary, @prop_flag:  0, @arcusid:  2,
  @browse_flag:  0, @prop_num:  , @obsids:  605,
  @filetypes:  all, @acisfiletypes:  all
10/01/02 17:12:53:45234 DisconnectHandler, user guest, from
  host foo/000.000.00.000/foo-Sun0, server pid:  89
```

The login log supplies the information to populate the 'downloads' table. Like the activity log, the login log contains connect/disconnect times along with user names. Unlike the activity log, the login log provides the download type (FTP or direct) and total byte size of the data transfer for connections made with Chaser and WebChaser. The download type gives important clues as to which interface was used. Because the login log provides times and names, entries can be fairly easily matched to entries in the activity log. An entry in the login log that corresponds with the above example might look like this:

```
Oct 1 17:12:52 2002:  User guest, from host
  foo/000.000.00.000/foo-Sun0, server pid 89 just connected
Oct 1 17:12:57 2002:  User guest from
  foo/000.000.00.000-SunOS retrieved obsids 605 (ftp:  total
  410083938)
Oct 1 17:13:09 2002:  User guest, from host
  foo/000.000.00.000/foo-Sun0, server pid 89 just disconnected
```

The FTP log is the final piece of the puzzle. It contains the information regarding data pickups for requests made through WebChaser, Chaser and the Provisional Retrieval Interface; it also contains information about data retrievals that is not in the other logs. Certain pre-packaged data, such as the calibration database and the Chandra deep field dataset are available on the CDA FTP Staging site and cannot be obtained through any of the above mentioned interfaces. Data transfers from the CDA Anonymous FTP site are also recorded in this log, but again cannot be connected to anything in the activity log or login log.

The records of data pickup that were requested through WebChaser, Chaser or the PRI can be matched back to information from the activity log and login log using the timestamp, host name, process id (pid) (for some records), and observation id (obsid) (again, just for some types of records). However, not all data requests are in fact picked up from the FTP site, so entries in the 'SR_usage' table will not always have a match in the FTP log.

In the current example, no obsid is mentioned in this particular type of tar file, but the pid is, and that in combination with the time and the host will be enough to match this line with entries in 'SR_connections' and 'SR_usage'.

```
Tue Oct 1 17:15:45 2002 171 foo 157304832
/export/ftp/ftphome/pub/srftp/000000/package_89_
  021001171545.tar a_ o a mozilla@ ftp 0 * c
```

4. Lessons Learned

During the rather lengthy development and implementation of this database, many difficulties and obstacles were encountered. Having to parse three different logs and match entries between all three is not easy. The example used above is very simple; most user activity is far more complex. Concocting algorithms to follow complex and sometimes inconsistent activities through three logs, all with different formats, is difficult and time-consuming.

For future missions we would recommend a more systematic approach, starting with the design of the various server log files. Log file requirements would need to address two particular issues: the type of information that should be recorded and the definition of a standard format. The tricky part is having enough foresight to conceive of a downloads and usage database before log file contents and formats are created. But our task would have been much easier if a proper requirements process had been followed from the beginning.

This work is supported by NASA contract NAS 8-39073 (CXC).

References

Rots, A. H., Winkelman, S. L., Paltani, S., & Deluca E. E. 2002, "Chandra Data Archive Operations", SPIE Conf. Ser. 4844: Observatory Operations to Optimize Scientific Returns III, in press