PDS Geosciences Node

Data Integrity Plan

Updated June 23, 2008

Purpose and Scope

This document reflects procedures used to maintain data integrity at the PDS Geosciences Node. Topics covered include the data storage system, backup methods, data integrity checks, and mitigation procedures.

Terms and Abbreviations

ICF – Integrity check file
PDR – Primary data repository
SDR – Secondary data repository

Data Storage System

The GeoNode data storage system is a fault-tolerant design with several automated and manual failover capabilities. In this section the system will be described by functional component.

Primary Data Repository

The primary data repository (PDR) is the storage containing data actively served by web and ftp services. The PDR is a configured with two Windows 2003 servers operating as an active/passive failover cluster attached to a storage area network (SAN) with multiple RAID 5 disk arrays and hot spares. The maintenance response window for PDR replacement hardware is four hours.

Secondary Data Repository

The secondary data repository (SDR) is a near-line disk-based backup of the PDR. It may be used as a PDR failover if conditions warrant. The SDR is configured as several network area storage (NAS) appliances with multiple RAID 5 disk arrays and hot spares. The SDR lacks server failover capability. The maintenance response window for PDR replacement hardware is 24 hours.

Additional Data Storage

A multi-drive high-speed tape library is used to support regular tape copies of Geonode data holdings. Copies are kept within the facility as well as off-site.

Deep Archive

The NSSDC serves as a deep archive repository for all PDS nodes. The dramatic increase in data volume has made standard data deliveries obsolete. A new delivery system is under design but not currently in use. As a result, only a minority subset of Geonode data are in the deep archive.

Backup Methods

This section describes the means by which data sets are copied to backup (i.e., non-primary storage) repositories.

Copies to the SDR

Data are copied from the PDR to the SDR on a per-volume basis using the publicly and freely available software package "xxcopy". This software package validates the SDR files during the copy procedure, thus ensuring file integrity. When a volume is added or updated, the xxcopy program is manually executed.

Copies to Tape

Data are copied to from the PDR to tape using the "CommVault" COTS package. Data integrity checking is built-in to the software. Tape backups are automatically generated on a scheduled basis.

Copies to the Deep Archive

This procedure is under development and testing and does not exist currently.

Data Integrity Checks

Manual data integrity checks are currently in place. We expect to have an automated system in place during FY09. Monthly comparison of PDR and SDR integrity check files (ICF) are used for two purposes:

- Ensure that the PDR and SDR copies match
- Locate non-matching files; follow up investigation ensues

Standard and Mitigation Procedures

Copy archive volume from PDR to SDR

Data are copied from the PDS to the SDR using the OTS package "xxcopy". For new volumes on the PDR, a matching destination must be created on the SDR. An example xxcopy command is shown below.

```
xxcopy <source> <destination> /e /v2 /I /bi /ck0
```

Update archive volume from PDR to SDR

Updates to archive volumes on the PDR are propagated to the SDR using the OTS package "xxcopy". An example xxcopy command is shown below.

```
xxcopy <source> <destination> /e /v2 /i /bi /ck0 /yy
```

Create ICFs

Integrity check fiels (ICFs) are used for comparing copies of archive volumes on the PDR and SDR. The OTS package “md5deep” is used to create these tables. An example md5deep command is shown below.

```
md5deep –r –l * > filelist.txt
```

Locating and handling PDR/SDR discrepancies

An ICF is created for each archive volume on the PDR and SDR. The files are compared. Discrepancies are handled on a case by case basis. Mismatches typically occur when a change has been made to the PDR but has not been reflected yet in the SDR. Otherwise, the files on the PDR and SDR are manually inspected to determine which copy is correct and what steps must be taken.

Saving data to tape / restoring data from tape

Tape backup operations are handled on a scheduled basis using COTS software. When needed, data are restored from tape to a temporary disk location for a data integrity check prior to placement on the PDR or SDR.