A Solution for Maintaining File Integrity within an Online Data Archive

Dan Scholes

PDS Geosciences Node

Washington University

Presentation Will Discuss

- PDS Geosciences Node background
- Threats to online data archives
- Methods to identify corrupt files
- PDS Geosciences Node approach to ensuring data archive file integrity

Planetary Data System (PDS)

- A NASA organization that archives science data from NASA's planetary missions.
- PDS responsibilities are:
  - To help NASA missions and other data providers to organize and document their digital planetary data
  - To collect complete, well-documented planetary data into archives that are peer-reviewed
  - To make the planetary data available and useful to the science community
  - **To ensure the long-term preservation and usability of the data.**

PDS Geosciences Node's Data Holdings

- Planetary science data related to geoscience studies
  - Surface and interior of the terrestrial planets and satellites (Moon, Mars, Mercury, Venus).
- Currently maintain:
  - Archives from over 20 NASA missions
  - Archive consists of over 40 TB of data
  - Over 13 million files

Access to Geosciences Node’s Archive

- Direct Access
  - FTP and HTTP
- Web Interfaces
  - Providing search and retrieval capabilities
- Custom User Request
  - External hard drive

Geosciences Node
Data Storage Architecture

- Primary online data archive (SAN)

- Secondary online replication site
- Tape backups
- Deep archive at
  National Space Science Data Center (NSSDC)

Threats to Online Data Archives

- Accidental change by staff
- Software error
- Hardware failure
- Malicious threats: Hacker or Virus
- Natural disaster

Defenses

- Firewall settings
- Network security policies
- Proactive hardware maintenance
- Multiple backup copies of the data

Typical Recovery

- Restoration from offline backup
  - Tapes
  - External hard drive
  - DVD/CD
- Restoration from online secondary copy
  - Mirror site
  - Replication site

- **How do you know the recovered copy is not corrupt?**

Bigger Question

How do you know if a change or corruption has occurred in the data archive?

Identifying Corrupt Files

Unsatisfactory Error Discovery	User Reported Problems
Finding Errors By Chance	Pre Release Data Review Internal Data Usage Errors Web Link Checker Sweep
Proactive Error Detection	Manual Checksum Scan Automated Validation Sweeps

← Our Solution

Checksum

- Checksum – a digital signature created by a hashing algorithm
  - File: frt000027e2_01_if156l_trr2.img
  - MD5 Checksum: 5F393DAD7B36F6418045A9299E605E51
- The Geosciences Node uses MD5
  - Commonly used
    - Many client tools for data providers
  - Fast calculation

Initial Data Integrity Study

- Manual Process
  - Create and compare checksum index files of data archive
- Advantages
  - Technically worked
  - Lessons learned
- Disadvantages
  - Time consuming
  - Difficult to manage
  - Difficult to update with new or replacement files

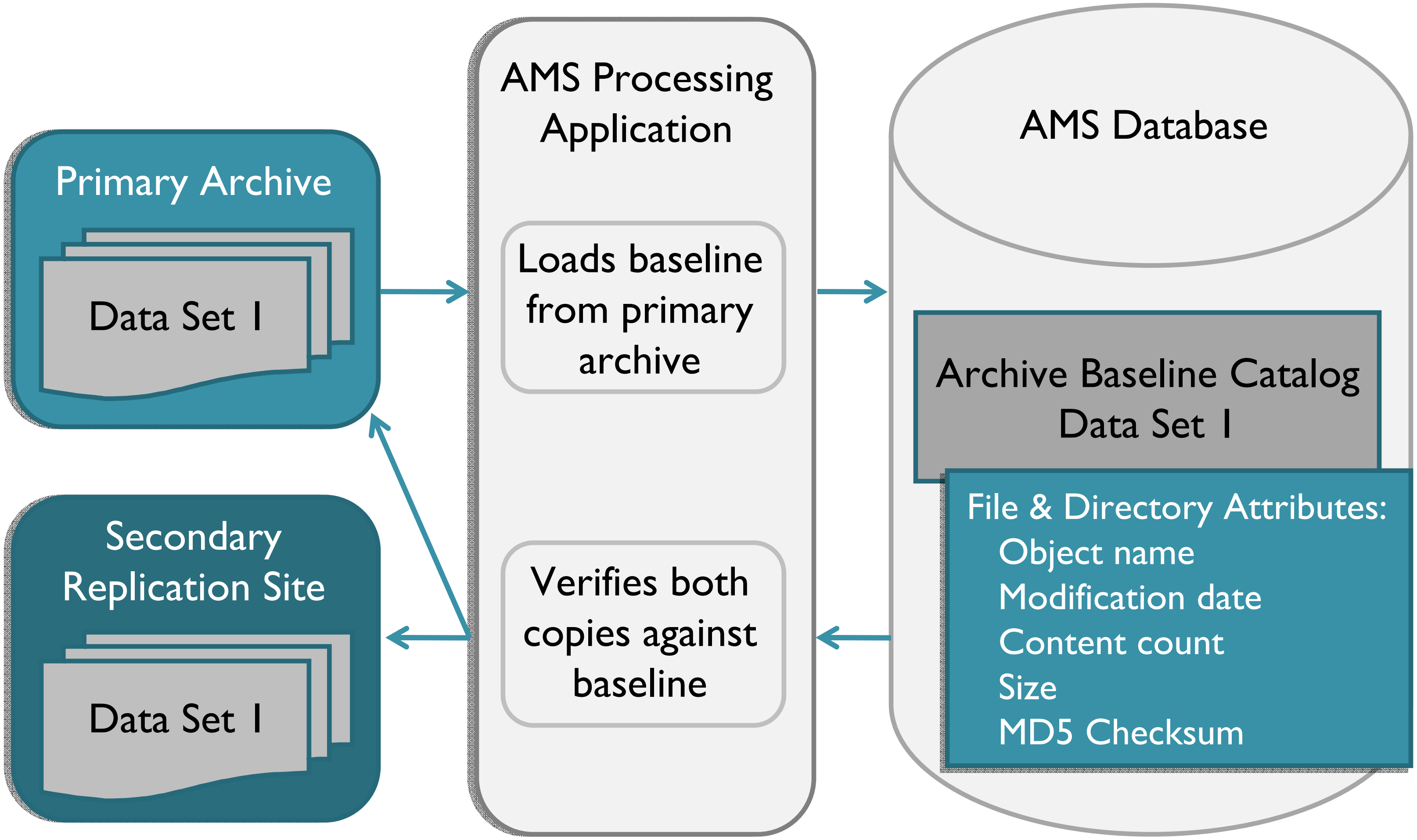
Application System Requirements

- Create catalog of data archive contents
- Track multiple archive copies
- Update catalog as archive grows
- Verify archive against cataloged contents
- Provide processing speed for monthly archive validations
- Provide an easy to use application interface

Archive Management System (AMS)

- Custom application
- Components
  - Graphical user interface (GUI)
  - Command line processing application
  - Relational database
- Concept
  - Archive baseline catalogs

Archive Baseline Catalog Concept



Primary Archive

Data Set 1

Secondary Replication Site

Data Set 1

AMS Processing Application

Loads baseline from primary archive

Verifies both copies against baseline

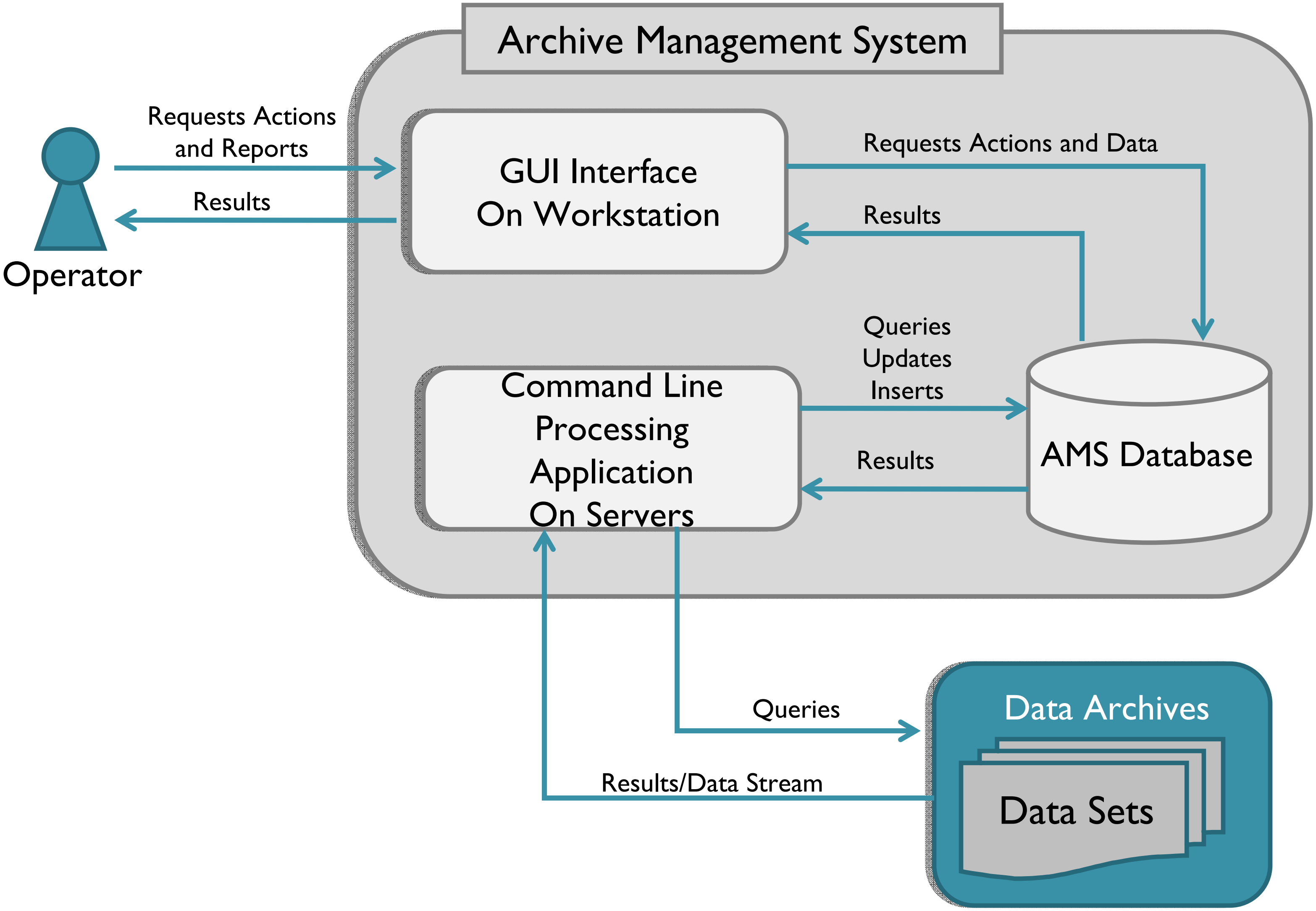
AMS Database

Archive Baseline Catalog Data Set 1

File & Directory Attributes:

- Object name
- Modification date
- Content count
- Size
- MD5 Checksum

AMS Overview

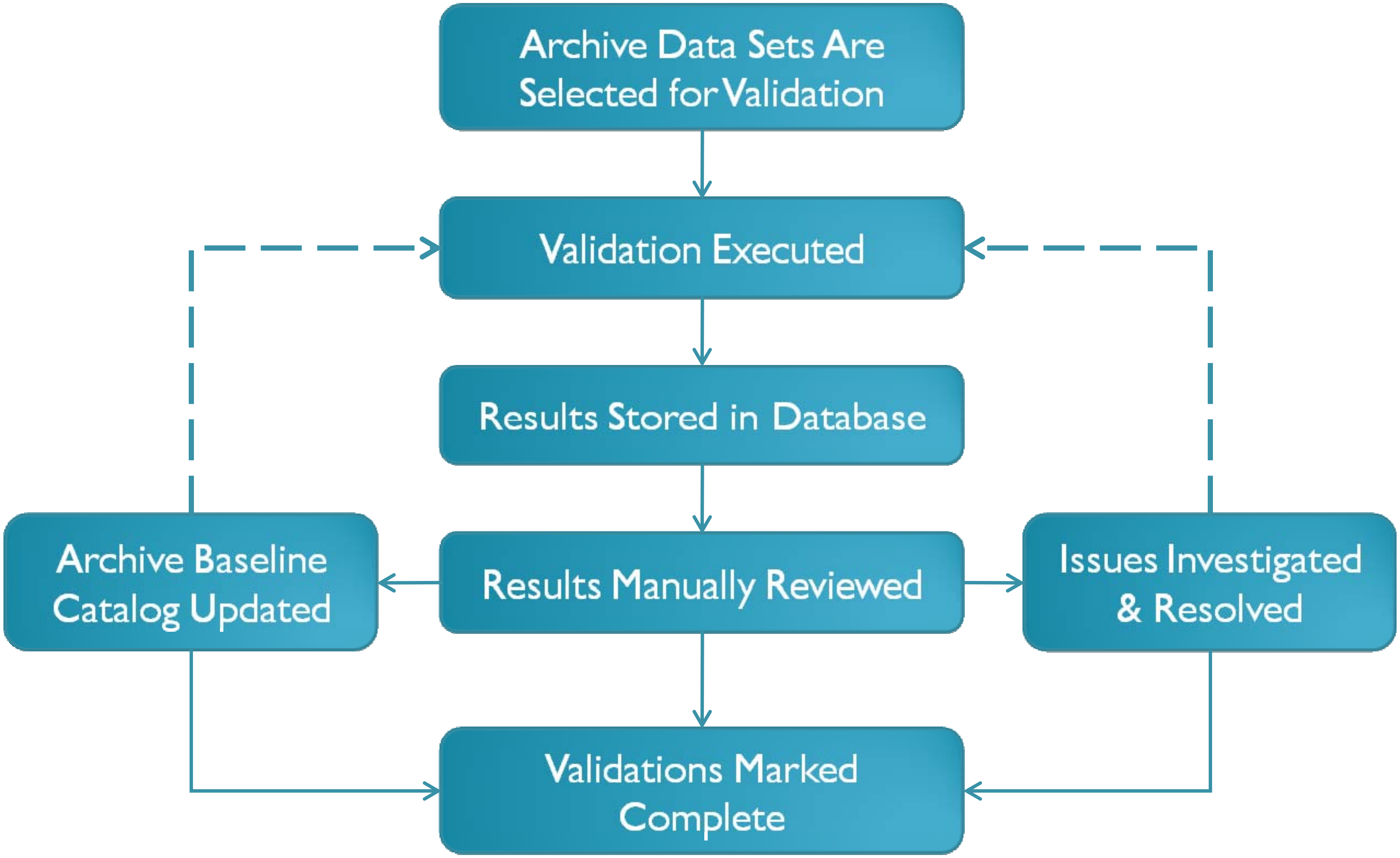


AMS Processing

- Create new archive baseline catalog
- Monthly validation scans
- Baseline is updated when new data is received

- Data recovery situations
  - Verify restored data against archive baseline catalog

AMS Monthly Validation Scans


Archive Data Sets Are Selected for Validation
Validation Executed
Results Stored in Database
Archive Baseline Catalog Updated
Results Manually Reviewed
Issues Investigated & Resolved
Validations Marked Complete

Full Scan Validation

- File and Directory attributes scanned
  - Object name – case sensitive
  - Modification date
  - Content count (directory’s file count)
  - Size
  - MD5 checksum (file validation only)
- Advantage
  - Thorough validation
- Disadvantages
  - Consumes more resources
  - Time consuming - entire archive up to 9 days

Quick Scan – no checksum

- File and Directory attributes scanned
  - Object name – case sensitive
  - Modification date
  - Content count (directory’s file count)
  - Size
- Advantages
  - Very fast processing speed - entire archive 28 hours
  - Identifies most accidental changes
- Disadvantage
  - Will not detect subtle file corruption

Categories of Validation Results

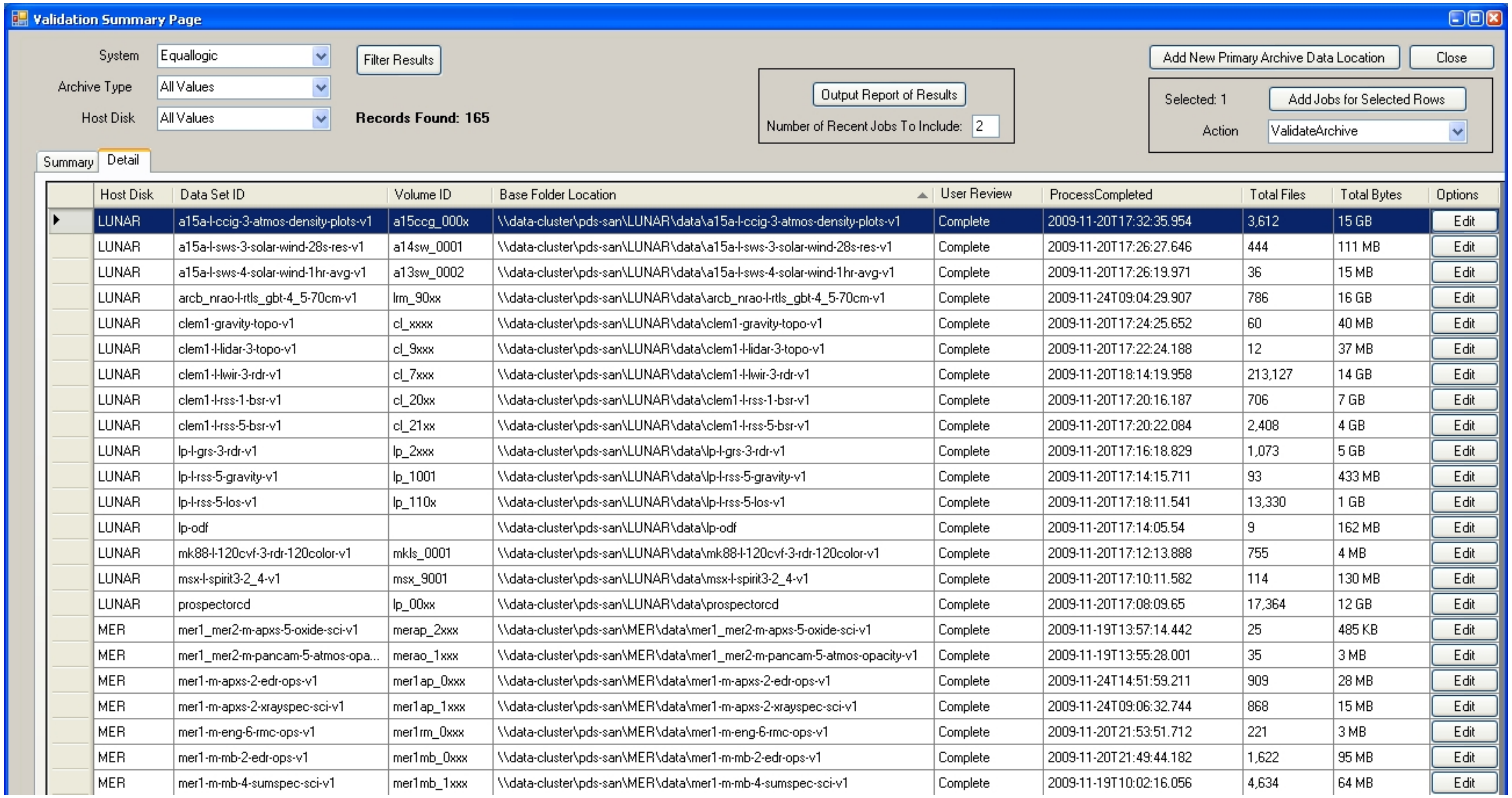
- No differences are detected
- File/Directory attributes are different
- New archive content is discovered
- Archive content no longer exists

Differences require further review

Interpreting Validation Results

- No differences are detected
  - Correct – no changes
- File/Directory attributes are different
  - Correct – revised data deployed to the data archive
  - Error – files were modified or corrupted
- New archive content is discovered
  - Correct – data added to the archive
  - Error - files accidently copied into archive
- Archive content no longer exists
  - Correct – items removed for archive revision
  - Error - mistakenly or maliciously removed

Archive Status List



Filter Results
Add New Primary Archive Data Location
Close
Output Report of Results
Add Jobs for Selected Rows
Summary
Detail

Validation Summary Page

System: Equallogic

Archive Type: All Values

Host Disk: All Values

Records Found: 165

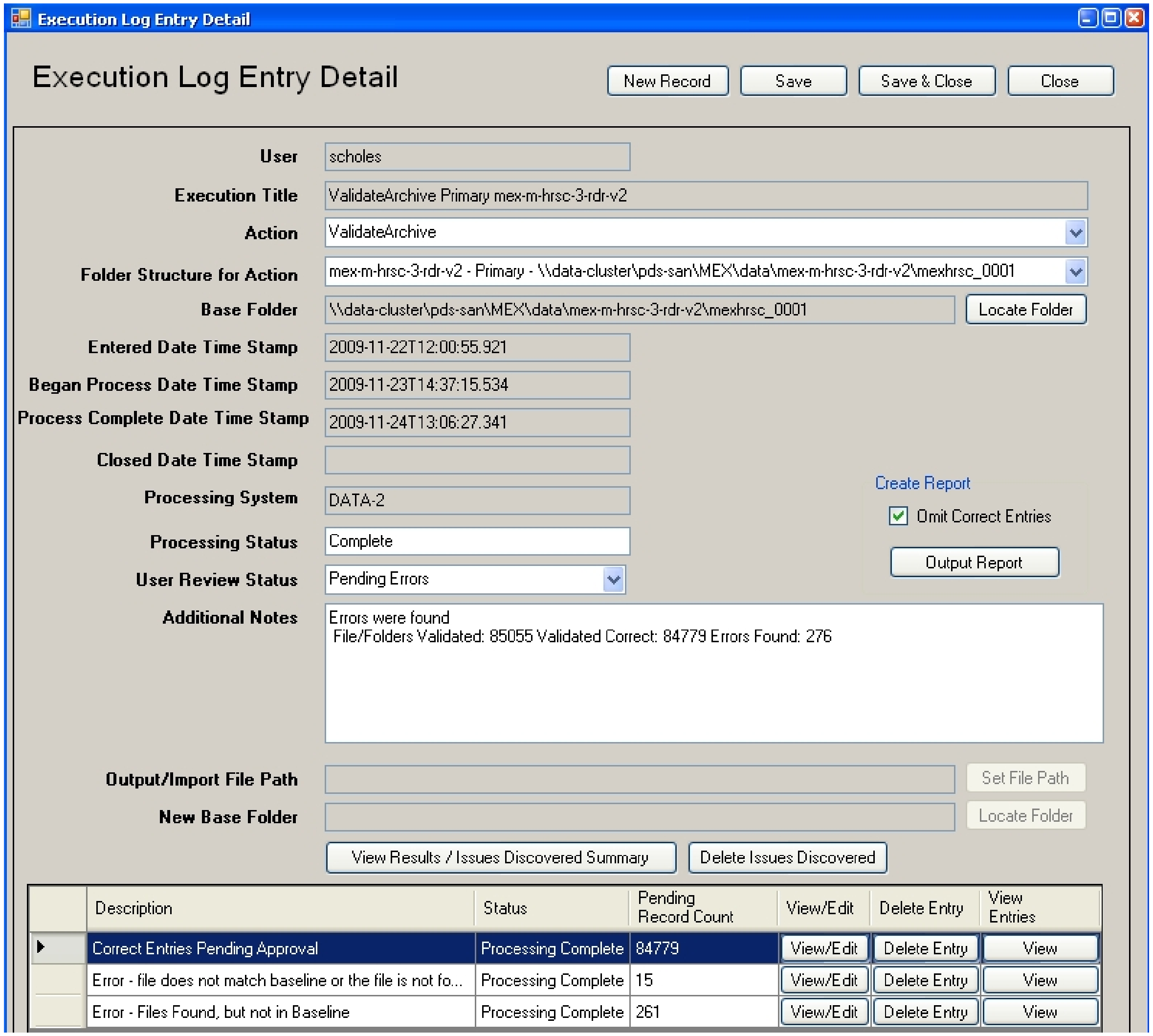
Number of Recent Jobs To Include: 2

Selected: 1

Action: ValidateArchive

Host Disk	Data Set ID	Volume ID	Base Folder Location	User Review	ProcessCompleted	Total Files	Total Bytes	Options
LUNAR	a15a-l-ccig-3-atmos-density-plots-v1	a15ccg_000x	\\data-cluster\pds-san\LUNAR\data\a15a-l-ccig-3-atmos-density-plots-v1	Complete	2009-11-20T17:32:35.954	3,612	15 GB	Edit
LUNAR	a15a-l-sws-3-solar-wind-28s-res-v1	a14sw_0001	\\data-cluster\pds-san\LUNAR\data\a15a-l-sws-3-solar-wind-28s-res-v1	Complete	2009-11-20T17:26:27.646	444	111 MB	Edit
LUNAR	a15a-l-sws-4-solar-wind-1hr-avg-v1	a13sw_0002	\\data-cluster\pds-san\LUNAR\data\a15a-l-sws-4-solar-wind-1hr-avg-v1	Complete	2009-11-20T17:26:19.971	36	15 MB	Edit
LUNAR	arcb_nrao-l-rtls_gbt-4_5-70cm-v1	lrm_90xx	\\data-cluster\pds-san\LUNAR\data\arcb_nrao-l-rtls_gbt-4_5-70cm-v1	Complete	2009-11-24T09:04:29.907	786	16 GB	Edit
LUNAR	clem1-gravity-topo-v1	cl_xxxx	\\data-cluster\pds-san\LUNAR\data\clem1-gravity-topo-v1	Complete	2009-11-20T17:24:25.652	60	40 MB	Edit
LUNAR	clem1-l-lidar-3-topo-v1	cl_9xxx	\\data-cluster\pds-san\LUNAR\data\clem1-l-lidar-3-topo-v1	Complete	2009-11-20T17:22:24.188	12	37 MB	Edit
LUNAR	clem1-l-lwir-3-rdr-v1	cl_7xxx	\\data-cluster\pds-san\LUNAR\data\clem1-l-lwir-3-rdr-v1	Complete	2009-11-20T18:14:19.958	213,127	14 GB	Edit
LUNAR	clem1-l-rss-1-bsr-v1	cl_20xx	\\data-cluster\pds-san\LUNAR\data\clem1-l-rss-1-bsr-v1	Complete	2009-11-20T17:20:16.187	706	7 GB	Edit
LUNAR	clem1-l-rss-5-bsr-v1	cl_21xx	\\data-cluster\pds-san\LUNAR\data\clem1-l-rss-5-bsr-v1	Complete	2009-11-20T17:20:22.084	2,408	4 GB	Edit
LUNAR	lp-l-grs-3-rdr-v1	lp_2xxx	\\data-cluster\pds-san\LUNAR\data\lp-l-grs-3-rdr-v1	Complete	2009-11-20T17:16:18.829	1,073	5 GB	Edit
LUNAR	lp-l-rss-5-gravity-v1	lp_1001	\\data-cluster\pds-san\LUNAR\data\lp-l-rss-5-gravity-v1	Complete	2009-11-20T17:14:15.711	93	433 MB	Edit
LUNAR	lp-l-rss-5-los-v1	lp_110x	\\data-cluster\pds-san\LUNAR\data\lp-l-rss-5-los-v1	Complete	2009-11-20T17:18:11.541	13,330	1 GB	Edit
LUNAR	lp-odf		\\data-cluster\pds-san\LUNAR\data\lp-odf	Complete	2009-11-20T17:14:05.54	9	162 MB	Edit
LUNAR	mk88-l-120cvf-3-rdr-120color-v1	mkls_0001	\\data-cluster\pds-san\LUNAR\data\mk88-l-120cvf-3-rdr-120color-v1	Complete	2009-11-20T17:12:13.888	755	4 MB	Edit
LUNAR	msx-l-spirit3-2_4-v1	msx_9001	\\data-cluster\pds-san\LUNAR\data\msx-l-spirit3-2_4-v1	Complete	2009-11-20T17:10:11.582	114	130 MB	Edit
LUNAR	prospectorcd	lp_00xx	\\data-cluster\pds-san\LUNAR\data\prospectorcd	Complete	2009-11-20T17:08:09.65	17,364	12 GB	Edit
MER	mer1_mer2-m-apxs-5-oxide-sci-v1	merap_2xxx	\\data-cluster\pds-san\MER\data\mer1_mer2-m-apxs-5-oxide-sci-v1	Complete	2009-11-19T13:57:14.442	25	485 KB	Edit
MER	mer1_mer2-m-pancam-5-atmos-opa...	merao_1xxx	\\data-cluster\pds-san\MER\data\mer1_mer2-m-pancam-5-atmos-opacity-v1	Complete	2009-11-19T13:55:28.001	35	3 MB	Edit
MER	mer1-m-apxs-2-edr-ops-v1	mer1ap_0xxx	\\data-cluster\pds-san\MER\data\mer1-m-apxs-2-edr-ops-v1	Complete	2009-11-24T14:51:59.211	909	28 MB	Edit
MER	mer1-m-apxs-2-xrayspec-sci-v1	mer1ap_1xxx	\\data-cluster\pds-san\MER\data\mer1-m-apxs-2-xrayspec-sci-v1	Complete	2009-11-24T09:06:32.744	868	15 MB	Edit
MER	mer1-m-eng-6-rmc-ops-v1	mer1rm_0xxx	\\data-cluster\pds-san\MER\data\mer1-m-eng-6-rmc-ops-v1	Complete	2009-11-20T21:53:51.712	221	3 MB	Edit
MER	mer1-m-mb-2-edr-ops-v1	mer1mb_0xxx	\\data-cluster\pds-san\MER\data\mer1-m-mb-2-edr-ops-v1	Complete	2009-11-20T21:49:44.182	1,622	95 MB	Edit
MER	mer1-m-mb-4-sumspec-sci-v1	mer1mb_1xxx	\\data-cluster\pds-san\MER\data\mer1-m-mb-4-sumspec-sci-v1	Complete	2009-11-19T10:02:16.056	4,634	64 MB	Edit

Validation Result Screen



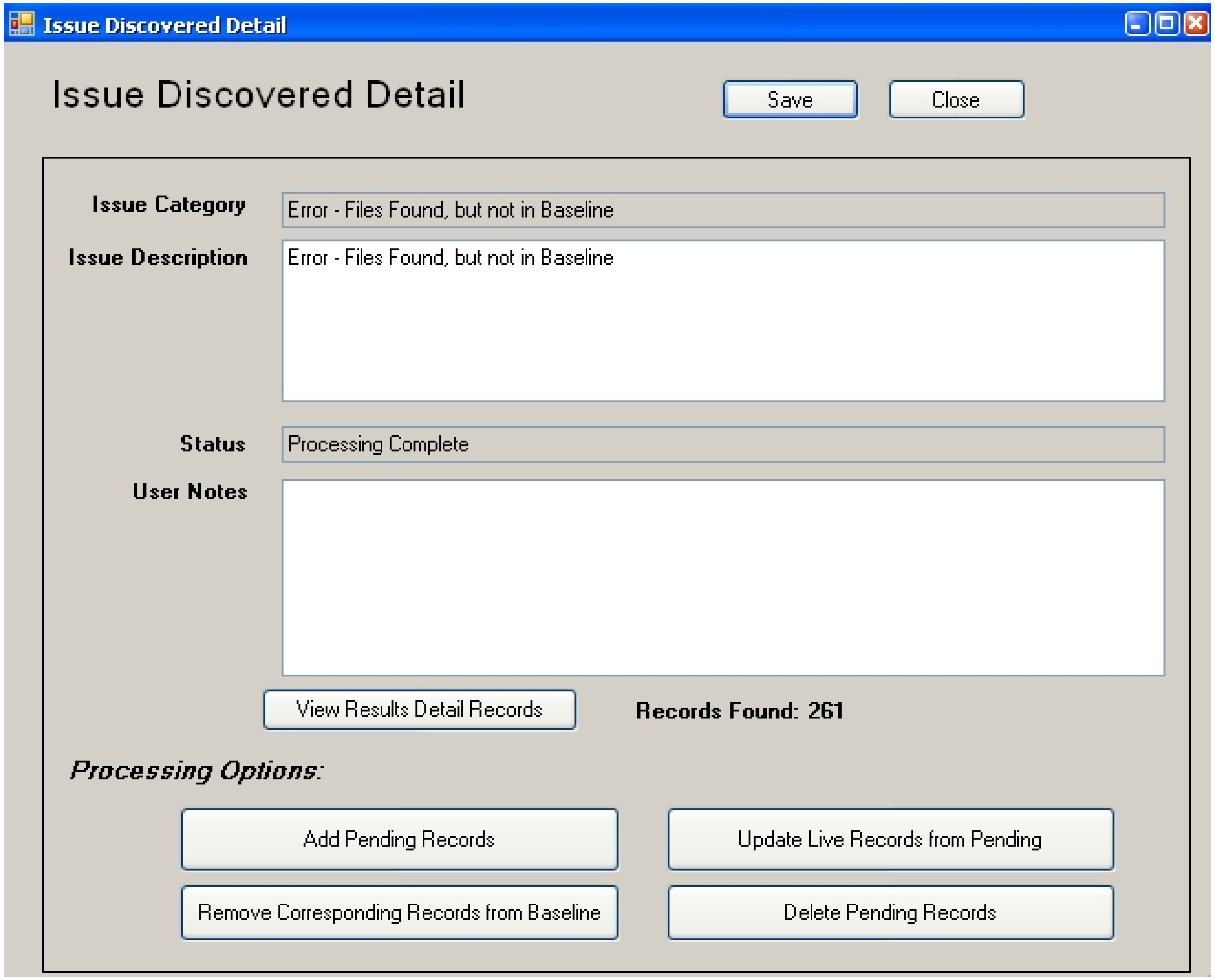
Execution Log Entry Detail

User	scholes
Execution Title	ValidateArchive Primary mex-m-hrsc-3-rdr-v2
Action	ValidateArchive
Folder Structure for Action	mex-m-hrsc-3-rdr-v2 - Primary - \\data-cluster\pds-san\MEX\data\mex-m-hrsc-3-rdr-v2\mexhrsc_0001
Base Folder	\\data-cluster\pds-san\MEX\data\mex-m-hrsc-3-rdr-v2\mexhrsc_0001
Entered Date Time Stamp	2009-11-22T12:00:55.921
Began Process Date Time Stamp	2009-11-23T14:37:15.534
Process Complete Date Time Stamp	2009-11-24T13:06:27.341
Closed Date Time Stamp	
Processing System	DATA-2
Processing Status	Complete
User Review Status	Pending Errors
Additional Notes	Errors were found File/Folders Validated: 85055 Validated Correct: 84779 Errors Found: 276
Output/Import File Path	
New Base Folder	

☑ Omit Correct Entries

	Description	Status	Pending Record Count	View/Edit	Delete Entry	View Entries
▶	Correct Entries Pending Approval	Processing Complete	84779	View/Edit	Delete Entry	View
	Error - file does not match baseline or the file is not fo...	Processing Complete	15	View/Edit	Delete Entry	View
	Error - Files Found, but not in Baseline	Processing Complete	261	View/Edit	Delete Entry	View

Validation Issue Resolution Screen



Issue Discovered Detail

Issue Discovered Detail

Issue Category: Error - Files Found, but not in Baseline

Issue Description: Error - Files Found, but not in Baseline

Status: Processing Complete

User Notes:

View Results Detail Records

Records Found: 261

Processing Options:

Add Pending Records
Update Live Records from Pending
Remove Corresponding Records from Baseline
Delete Pending Records

AMS Results

- Geosciences Node has used the AMS for nearly a year.
  - Minimal personnel time to manage, monitor, and add new archives
  - Full scan of the entire archive 12 times
    - Can take up to 9 days of processing (full scan)
  - Two accidental archive changes
  - No file loss or corruptions
- Provides the Geosciences Node with a better degree of data integrity

Future

- Geosciences Node's data archives continue to rapidly grow with current and future missions.
- Further performance review
  - Network switch configurations
  - Server Configurations
  - Disk Performance
  - Simultaneous processing streams
  - Possible code modifications

Questions

- Contact Information
  - Dan Scholes
  - Applications Programmer
  - PDS Geosciences Node
  - Washington University in St. Louis
  - scholes@wunder.wustl.edu