

4.8.21 AIM Tape Archive Consistency Checking Utility

The **EcDsAmArchiveCheckUtility** checks the contents of the Volume Group directories defined in the Inventory Database against the list of files recorded in the Inventory Database. These data files are typically stored on tape, but may also be stored on disk. The utility will check all granule types (Science, Browse, Production History, QA, and Delivered Algorithm Package). The utility verifies the name, location, and size of the files in the Inventory. It does not verify checksums.

The utility also includes a check of the XML Archive by default; this option is appropriate for processing a small volume of data, such as a few days, but should be avoided if you are testing a large amount of granules because it adds significant processing time. The **EcDsCheckXMLArchive.pl** utility is much faster for checking the XML Archive. There is a command line option for the **EcDsAmArchiveCheckUtility** to turn off the XML archive check (-nx).

This utility uses the standard ECS volume group rules for determining what granules are mapped to a given volume group. The rules are based upon comparing the metadata of a granule to the attributes of the volume group stored in the **DsStVolumeGroup** table within the Inventory database. Within these rules, the **ShortName** and **VersionID** of the granule are compared to the volume group's **VersionedDataType** attribute. Next, the **insertTime** of the granule must be between the **VolumeStartDate** and **VolumeEndDate** for the volume group. Finally, if the volume group has a value for **SelectionDate**, it is compared to the granule's **BeginningDateTime** to determine if the granule is part of a forward or reprocessing chain (this applies to science granules only). If the **BeginningDateTime** is less than the **SelectionDate**, the granule is part of a reprocessing chain and should be located in the volume group that matches the above conditions and has an "R" appended to the **VersionedDataType**. If the **BeginningDateTime** is greater than or equal to the **SelectionDate**, the granule is part of a forward processing chain and will be located in the volume group without an "R" appended to the **VersionedDataType** attribute.

4.8.21.1 EcDsAmArhiveCheckUtility – Command line options

The utility has the following command line options:

```
EcDsAmArchiveCheckUtilityStart <MODE> [<directories to check>] [-s] [-o] [-lo] [-nx]
```

```
<directories to check> = -d <date_range>    | -e { <ESDT> ... }  
                                         | -a  
                                         | -v <volume_group_path>  
                                         | -vs <starting_volume_group_path>
```

Example:

```
EcDsAmArchiveCheckUtilityStart OPS -e "AST_L1B.001" -s -o -lo -nx
```

Table 4.8.21-1 shows the command line parameters for the **EcDsAmArchiveCheckUtility**.

Table 4.8.21-1. Command Line Parameters of the EcDsAmArchiveCheckUtility

Parameter Name	Mandatory	Description
<MODE>	Yes	The mode to be processed.
-d	No	Select only those granules whose insert time falls within the supplied date range For example: -d "Mar 13 2002 - Mar 22 2002" or -d "Mar 3 2002 1:23 PM - Mar 22 2002 15:51 PM If you omit the end date, the utility will use the current date (day) as a default end date. With this option the utility will report database entries with missing files but it will not report files with missing database records.
-e	No	Only check granules whose ShortName and VersionID are specified by the provided list of ESDTs. See the notes about this option described later in this section.
-a	No	Only check "active" volume groups, i.e., directories that are still open for receiving files. These volume groups are determined by the VolumeEndDate being set to null in the DsStVolumeGroup table.
-v	No	Process the supplied volume group path. This should be a fully qualified path starting at the root directory.
-vs	No	Allows the operator to provide a fully qualified path where the utility should begin processing. The supplied volume group and all volume groups that were created after the starting volume group will be processed. This allows the operator to process all volume groups created since a specific time. In addition, the -vs option can alternatively be passed a VolumeGroupID from the DsStVolumeGroup table instead of a fully qualified path. The volume groups that are skipped are also recorded in the output file.
-s	No	Saves a list of all files processed in a file with the name <MODE>.YYYY.MM.DD_HH:MM:SS_<HOST>.dbOut.
-o	No	Output a list of files that match, to a file with the naming convention <MODE>.yyyy.mm.dd_hh:mm:ss_<host>.ok. This should be used with caution as the list could be very large.
-lo	No	Creates all output files in the log directory for the MODE instead of the current working directory where the utility was executed.
-nx	No	Suppress the consistency checking for the XML archive. This allows the utility to run much faster when processing large volumes of data. The EcDsCheckXMLArchive.pl utility can be used to specifically check the XML archive.

If the operator doesn't provide a specification of which volume groups (directories) to check using one of the options (-e, -d -a, -v, or -vs), then the utility will check all volume groups.

The operator will be prompted to enter the Sybase DB Server – for the DAACs this should be

```
< n | e | l >4dbl03_srvr.
```

Next, the name of the AIM Inventory database is displayed and the operator is prompted to accept the entry by entering the letter A and pressing the enter key or to change the entry by pressing the letter C and the enter key.

Next, the operator is asked to enter a Sybase login name and password; this login must have access to the AIM Inventory (EcInDb) database.

Next, the operator is prompted for an "ArchiveRoot." Enter the starting part of the volume group path that is to be searched. Typically this will be something like "/stornext/snfs1" which reflects the path up to but not including the <MODE> component of the directory path.

If the operator uses one of the options (-e, -d -v, or -vs) a list of volume groups to be processed will be presented and the operator is asked to accept it by pressing "y" and enter. If operator doesn't use one of the options to limit volume groups to be processed or uses the -a option, this step is omitted.

Notes about using the -e option:

The -e {ESDT ...} option will check all ESDTs in the provided list, note, the {} should not be entered on the command line. If the Version ID for the ESDT is omitted, the program will check all granules whose short name match the passed in short name. Wild cards may also be provided in the ESDT list.

For example, to process all ESDTs that begin with "AST", you would run the program with the following option and parameter: *-e "AST*"* (this will cause ALL versions of every ESDT that begins with "AST" to be processed).

When no "." is present in the wildcard, the program assumes that the supplied argument is a ShortName only, and processes all versions of that particular ESDT. So supplying *-e "MO*3"* will cause the program to search for a VersionedDataType that matches the following pattern *MO%3.%*". If you really wanted to search for all version 3 ESDTs that start with "MO", you would supply the following *-e "MO*.003"*.

Any combination of absolute strings or wild cards can be used with the -e option.

For example: *-e "AST*" "MOOD" "A*L1*.002" "MOOO.001"* will cause the processing of all versions of ESDTs that begin with "AST", all versions of ESDTs that have the short name "MOOD", all version 2 ESDTs that begin with "A" and have an "L1" substring, and finally the ESDT MOO0.001.

Another example would be to process only version 2 ESDTs using the option *-e "*.002"*.

When the -e option is used, the utility will prompt for confirmation after printing the list of VolumeGroups that correspond to the list of ESDTs that are being used. When using wild card

characters, you must enclose the pattern in quotes in order to prevent the UNIX system from interpreting the pattern as a file-matching pattern. The Utility translates UNIX regular expressions into SQL regular expressions, but it leaves SQL regular expressions in the format passed in via the command line. This means that you may use SQL regular expressions on the command line and that pattern will be used when querying the DsStVolumeGroup table to find the correct volume groups. For example: `-e "[^Browse]"` will get all VersionedDataTypes that don't have the Browse ShortName.

Also note that upon execution, the utility removes all pre-existing output files that have the identical time stamp as the output files for the current run in order to "clean up" after a previous run that had errors. So when the utility is run consecutively within the same minute, it assumes that the previous runs in that minute were erroneous.

4.8.21.2 EcDsAmArchiveCheckUtility – Outputs

As the utility runs, it displays a summary of each volume group it processes to the console. In addition, it creates several report files. All output files for the EcDsAmArchiveCheckUtility have the following naming convention:

`<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.<fileExtension>`

where yyyy, mm, dd, hh, mm, and ss describe the current year, month, day of month, hour, minute, and second. The output files and logs are produced in the current working directory of the script or in the standard logs directory if the “-lo” option is provided on the command line.

The following report files are produced:

FileName	Contents
<code><MODE>.yyyy.mm.dd_hh:mm:ss_<host>.dupGrans</code>	Duplicates, triplicates, etc.
<code><MODE>.yyyy.mm.dd_hh:mm:ss_<host>.wrongSize</code>	Granules with non-matching file size
<code><MODE>.yyyy.mm.dd_hh:mm:ss_<host>.notInArch</code>	Granules missing from archive
<code><MODE>.yyyy.mm.dd_hh:mm:ss_<host>.notInDb</code>	Files missing from the AIM database
<code><MODE>.yyyy.mm.dd_hh:mm:ss_<host>.InDbDFAed</code>	Files in the archive but DFAed.

Separate reports are generated to facilitate identification and cleanup of each potential case. Grouping all errors into a single report would potentially make the report too large to easily use.

Files in the archive that are missing database records are sometimes referred to as “orphans” and are usually the result of a Granule Deletion failure. Database records that are missing files in the archive (“phantoms”) or that have file size mismatches are a much more serious issue and most likely represent an error during Ingest or a corruption of the file system.

In addition to the above report files, the file:

`<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.outputSummary`

contains a summary of the total execution time of the program, the execution time of each major component of the program, and the processing speed (# db files/second and # archive files/second) of the program. Additionally this file contains the following information for each volume group processed:

- the volume group path

- the number of duplicate granules found
- the number of files with incorrect size
- the number of files not in the archive
- the number of files not in the database
- number of files in the archive that are DFA'ed in the database
- the number of XML files not in the archive
- the number of XML files not in the database
- Total number of files in the archive
- Total number of files in the database

For debugging purposes the following files are also created.

FileName	Contents
<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.dbQuery	All database queries used
<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.VGOut	List of each volume group processed
<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.readDirOut	List of files in each volume group
<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.dbOut	List of files from the AIM database
<MODE>.yyyy.mm.dd_hh:mm:ss_<host>.ok	List of files matched successfully

The output format is similar for each file. Typically each report lists each volume group path as it is being processed. After the volume group path, a tabular list of information is presented that contains the information needed to research the problem. This information will typically include granule ID, ShortName, VersionID, the "internalFileName" used by the file, and possibly the file size. The exact output depends on the report file. The information is presented in human readable form.

Special Notes:

If this utility is executed while ECS ingest is occurring, then the reports on "active" volume groups may contain spurious entries in the "notInDb" file. This is because files may have been copied into the archive but not yet entered into the Inventory database. Also note that DeleteFromArchive is only applicable to "science" granules (i.e.: not Browse, QA, PH, and DAP). The non-science granules should always be version 001 and the SelectionDate is not supported.

This page intentionally left blank.