

## 4.7.28 EMS Dataset Extract Utility

The Earth Science Data and Information System (ESDIS) Metrics System (EMS) Dataset Extract utility provides DAAC Operations Staff an operational support tool that automatically extracts data and information from DAAC databases and transmits the data to the EMS metric reporting tool.

The EMS Dataset Extract utility extracts data from DAAC operational database tables and outputs the data into ASCII text flat files. The utility is designed to run as a CRON on a daily basis. The flat files prepared for EMS are formatted so that one line in the file represents one record of information. The output files have field information delimited by “|&|”. The flat files are transferred via ‘SCP’ to the centralized EMS location from which EMS metric reports can be generated.

The EMS Dataset Extract utility is run with a set of optional and required DAAC defined command line parameters. The utility can also be run manually from the Linux command prompt with the optional and required parameters specified. The utility will behave differently depending on the combination of parameters entered. Daily checks by EMS operations personnel will ensure that the data exported by the EMS Dataset Extract utility was received at the central EMS location. Updated flat files are to be sent to EMS whenever data processing failures are encountered or data corruption is detected.

The utility is designed to extract data in periods of 24-hours or one day. If the data transfer fails for a few days, the utility is designed to perform data recovery automatically for the period of time missed as soon as communication is restored.

### 4.7.28.1 Running the EMS Dataset Extract Utility

The EMS Dataset Extract utility is run from a CRON (see Section 4.7.28.2) or started by entering the following command from the /usr/ecs/<mode>/CUSTOM/utilities directory Linux command line:

```
>EcDbEMSdataExtractor.pl-m <mode> -s <start date> -e <end date> -x <extract type> -v -o -i
```

Table 4.7.28-1 shows the parameters for the EMS Dataset Extract utility.

**Table 4.7.28-1. Command Line Parameters of the EMS Dataset Extract Utility (1 of 2)**

Parameter Name	Description
-(m)ode	Mandatory. Specifies the mode in which the extraction is to occur. It must be a valid, existing mode with a format of OPS or TS[1-4] or DEV0[1-9].
-(s)tartdate	Optional. The startDate time for ExtractType processing with a format of “mm dd, yyyy” or “mm/dd/yyyy”.
-(e)nddate	Optional. The endDate time for ExtractType processing with a format of “mm dd, yyyy” or “mm/dd/yyyy”.
-e(x)tracttype	Optional. Identifies the type of data being extracted. The following values are valid extracttypes: Meta, searchExp, Ing, Arch, DistFTP, DistHTTP, DistMedia.

**Table 4.7.28-1. Command Line Parameters of the EMS Dataset Extract Utility (2 of 2)**

Parameter Name	Description
-(o)verride	Optional. Identifies whether a period of time longer than the default 24-hour period will be used for the date range for the extracttype. If the override command line parameter is specified, entries are required for the startdate, enddate, and extracttype.
-(v)erbose	Optional. Prints messages to screen as well as log
-(i)ntial	Optional. Will enter 'default' into the ExecutionMode field in EcEMSextractRecord table for the indicated dataset.

The -mode parameter is mandatory. For each command line parameter, a dash “-“ followed by the letter in parenthesis indicated in the above table can be used instead of the full parameter name.

Table 4.7.28-2 describes datasets that are extracted and exported using the extraction utility:

**Table 4.7.28-2. Datasets of the EMS Dataset Extract Utility**

Dataset Name	Description
Meta	Product attribute metadata
searchExp	Product attribute search
Ing	Data Ingest
Arch	Data Archive
DistFTP (DataPool, FtpPush, FtpPull) DistHTTP (DataPool)	Physical media distribution orders
DistMedia (Cdrom, Dlt, Dvd)	Electronic media distribution orders

NOTE: Initially the EMS system needed to know about the DAAC users and Product attributes. This data was extracted and sent to EMS. Metadata and Product Attribute Flat file information maintained in the AIM database was sent. User information was also sent. Also, prior to running the EMS Dataset Extract utility in the default execution mode a baseline for the default execution was established. The baseline was the date from which the default execution should start processing. A record of this baseline is recorded in the MSS database EcEMSextractRecord table. Each time the EMS Dataset Extract utility is executed in default mode this table is checked to determine the last time a dataset was processed and to determine the date range to use for the current run of the dataset.

## 4.7.28.2 EMS Dataset Extract Utility Examples

Below are examples for invoking this tool:

### 1. **EcDbEMSdataExtractor.pl -m <mode>**

Running the EMS Dataset Extract utility with only the `-m` option is the default way to run the utility, and this should be the only parameter used when running the utility as a CRON. To set up the CRON, access the Linux server as the Sybase user. Set up the CRON by running the CRONTAB `-e` command. The command will be something like: `51 16 * * 2 (export LD_LIBRARY_PATH=/tools/sybOCv12.5.1/lib:/home/cmops/lib:/bin/csh -c "cd /usr/ecs/TS1/CUSTOM/utilities; EcDbEMSdataExtractor.pl -m TS1")`. This command may be different based on the configuration for the server.

Since the “start date” and “end date” parameters are not provided, the EMS Dataset Extract utility will access the `EcEMSExtractRecord` table for each dataset and retrieve the most current record for the dataset that has been marked “Default” in the `ExecutionMode` field. The beginning “start date” and “end date” for the Dataset run will be calculated based on the retrieved value for the last run of the Dataset.

### 2. **EcDbEMSdataExtractor.pl -m <mode> -s “start date” -e “end date” -x DistFTP**

Running the EMS Dataset Extract with these options will create output files for DistFTP data or the specified dataset for each day greater than or equal to the start date and less than the end date. A record for each day of the run will be inserted into the `EcEMSExtractRecord` table. The date range specified by the start date and end date must be at least 24 hours. The dates should be entered without hour or minutes specified. A record of the run will also be logged in the log file. If the `-x` parameter is omitted, then output files for all Datasets will be created.

### 3. **EcDbEMSdataExtractor.pl -m <mode> -s “start date” -e “end date”**

Running the EMS Dataset Extract with these options will create output files for all datasets for each day greater than or equal to the start date and less than the end date. A record for each day of the run will be inserted into the `EcEMSExtractRecord` table. The date range specified by the start date and end date must be at least 24 hours. The dates should be entered without hour or minutes specified. A record of the run will also be logged in the log file.

### 4. **EcDbEMSdataExtractor.pl -m <mode> -s "start date" -e "end date(start date + one day)" -i**

The preceding command should be run from the Linux prompt to initialize the datasets for default execution: If the `-x` option is used then only the specified Dataset will be initialized. For the Dataset execution “Default” will be placed in the `ExecutionMode` field for the record of the run. Subsequent runs of the EMS Dataset Extract utility without the date range specified will access the `EcEMSExtractRecord` table for the dataset and retrieve the most current record that has been marked “Default” in the `ExecutionMode` field. The beginning “start date” and “end date” for the Dataset run will be calculated based on the retrieved value for the last run of the Dataset.

### 4.7.28.3 Required Operating Environment

The EMS Dataset Extract utility runs on the Linux platform.

### 4.7.28.4 Interfaces and Data Types

Table 4.7.28-3 lists the supporting products that this tool depends upon in order to function properly.

**Table 4.7.28-3. Interface Protocols**

Product Dependency	Protocols Used	Comments
Sybase	SQL	Via SQL server machine.
Perl module	Perl	Module to connect to the database and print out the nicely formatted help page.

### 4.7.28.5 Configuration File Format – EcDbEMSdataExtractor.CFG properties

The EcDbEMSdataExtractor.pl utility requires a configuration file. This configuration file, “EcDbEMSdataExtractor.CFG”, is located in the /usr/ecs/<mode>/CUSTOM/cfg directory on the x4spl01 server. All edits of the EcDbEMSdataExtractor.CFG” file will be implemented using a Linux editor, such as “vi”. The configuration file contains vital details about how to connect to the Sybase server and EMS host machine. Without this file, the tool cannot run. Table 4.7.28-4 describes the configuration parameters:

**Table 4.7.28-4. Configuration Parameters (1 of 2)**

Parameter Name	Recommended Value	Description
SERVER	<x4dbl01_svr>	Enter sybase server name e.g. x4dbl01_svr.
PROVIDER	<DAAC NAME>	Enter provider name e.g. DAAC identifier.
EMSEXTRACTDIR	/usr/ecs/<mode>/CUSTOM/data/DSS	Enter EMS extraction directory location. This is the directory path specifying where data is extracted to when bcp'd out of database e.g. /usr/ecs/<mode>/CUSTOM/data/DSS.
EMSUSER	cmshared OR allmode	Enter user name to gain access to host represented by IPADDRESS - provided by EMS team.
PGMID	7000900	Static value. Same for all DAACs and Modes.

**Table 4.7.28-4. Configuration Parameters (2 of 2)**

Parameter Name	Recommended Value	Description
DBUSER	EcDbEMSdataExtractor	Static value. Same for all DAACs and Modes.
IPADDRESS	The following is an example <123.456.789.1>	Enter IP Address or host name e.g. ws1.ems.eosdis.nasa.gov - provided by EMS team. The IP Address identifying EMS host to SCP the data files produced by the utility.
STORNEXT	<Descriptor Directory Path>	Location of ESDT descriptor files.

**4.7.28.6 Special Constraints**

The EMS Dataset Extract utility runs only if the Sybase server is operational. EMS code must be installed in the mode. The EMS configuration file must be configured. SCP must be configured to run in the user environment from which the extract utility will be executed. EMS utility initial set-up should have been executed in the mode.

**4.7.28.7 Outputs**

Outputs will be printed to standard out if the -v flag is included with on the command line of the EMS Dataset Extract utility. Messages are also output to the EcDbEMSdataExtractor.log file (see Section 4.7.28.9). The DAAC Operations Staff should review the messages printed to the log file.

**4.7.28.8 Event and Error Messages**

Error messages will be displayed on standard out if the -v flag is included with the executed EMS command. Error messages will be logged in the EcDbEMSdataExtractor.log file (see Section 4.7.28.9). EMS Dataset Extract utility events are recorded in the MSS database EcEMSextractRecord table. Field descriptions for the EcEMSextractRecord table are described in Table 4.7.28-5.

**Table 4.7.28-5. EcEMSextractRecord Table (1 of 2)**

Name	Datatype	Null	Description
ExtractId	numeric(8,0)	No	Monotonic Key.
ExtractType	varchar(255)	Yes	Dataset type, ie, Arch, DistFTP, DistHTTP, DistMedia, Ing, Meta, or searchExp.
RunStartTime	datetime	Yes	The time the dataset began processing.
RunCompletionTime	datetime	Yes	The time the dataset completed processing.
StartDate	datetime	Yes	The Start Date of the dataset run.
EndDate	datetime	Yes	The End Date of the dataset run.
ExtractFileName	varchar(500)	Yes	The name of the Extract File, including the directory path.

**Table 4.7.28-5. EcEMSExtractRecord Table (2 of 2)**

<b>Name</b>	<b>Datatype</b>	<b>Null</b>	<b>Description</b>
FTPcompletionTime	datetime	Yes	The date the dataset was SCP'd to IP address indicated in the configuration file.
ExecutionMode	varchar(8)	Yes	Execution Mode of the dataset run; either Default or override.
MediaType	varchar(20)	Yes	MediaType of dataset run; either NULL, DLT, DVD, Scp, CDROM, FtpPull, or FtpPush.
DataSource	varchar(50)	Yes	Mode and Media type combined – used in constructing the ExtractFileName.
Provider	varchar(50)	Yes	DAAC identifier.

#### **4.7.28.9 Logs**

The tool logs messages in the /usr/ecs/<mode>/CUSTOM/logs/EcDbEMSdataExtractor.log file.

#### **4.7.28.10 Recovery**

The EMS Dataset Extract utility supports automatic recovery from an interrupted run. If the utility has not been run for a period of time, then the utility can start running from the time it was previously run and files will be generated for the missing days. Also, if a dataset file was extracted to the extract directory, but not SCP'd to EMS, a subsequent run of the utility will SCP this file and mark the file as SCP'd in the EcEMSExtractRecord table by updating the FTPcompletionTime for the file record. Also, if a dataset file has been removed from the extract directory, but not SCP'd, a subsequent run of the utility will mark the record as SCP'd, in the EcEMSExtractRecord table by updating the FTPcompletionTime with the date "Jan 1, 1900" and a note documenting this will be written to the log.