440-TP-011-001

Manual vs. Automated Data Ingest Analysis

Technical Paper

February 1995

Technical Paper—Not intended for formal review or government approval.

Prepared Under Contract NAS5-60000

RESPONSIBLE ENGINEER

Paul S. Roycraft 2/28/95

Paul S. Roycraft, Systems Engineer Date
EOSDIS Core System Project

SUBMITTED BY

Stephen Fox 3/18/95

Steve Fox, SDPS Office Manager Date
EOSDIS Core System Project

Hughes Applied Information Systems
Landover, Maryland

This page intentionally left blank.

Abstract

This technical paper focuses on the issues associated with the level of automation supported by the EOSDIS Core System (ECS) Ingest Subsystem.

This page intentionally left blank.

Contents

Abstract iii

Contents v

1. Introduction 1

1.1 Purpose 1

1.2 Organization 1

2. Data Ingest Scenarios 3

2.1 Operational Data Ingest Scenarios 3

2.1.1 Automated Network Ingest (Get) Scenario 4

2.1.2 Automated Network Ingest (Put) Scenario 5

2.1.3 Polling Ingest (Files) Scenario 7

2.1.4 Polling Ingest (Delivery Record) Scenario 8

2.1.5 User Network Ingest Scenario 9

2.1.6 Hard Media Ingest Scenario 11

2.2 Candidate Ingest Automation Functions 12

3. Ingest Automation Evaluation 13

3.1 Automated Error Detection 13

3.2 Automated Load Balancing 13

3.3 Automated Interface Interactions 14

3.4 Automated Hard Media Ingest 14

3.5 Manual Ingest Activities 14

This page intentionally left blank.

1. Introduction

1.1 Purpose

The purpose of this study is to examine the most cost-effective methods of operating the ingest subsystem. It focuses on functions performed during data ingest, the level of automation provided by the ingest subsystem, and the issues associated with an automated versus manual approach for performing select ingest functions. Key issues include:

- what are the major ingest subsystem operational scenarios
- which ingest subsystem activities are subject to automation
- identification of ingest functions requiring operations staff interaction

This technical paper has been developed in response to issues identified in the System Design Review (SDR) version of the ECS System Design Specification.

1.2 Organization

This paper is organized as follows:

Section 1 introduces the purpose of this technical paper.

Section 2 summarizes the ingest subsystem operational scenarios developed for the major interactions of the ingest subsystem with external data providers. From these scenarios, several approaches to automation of ingest functions are discussed.

Section 3 presents the issues associated with the automation of select ingest functions and discusses the required operator actions in support of manual ingest functions.

Questions regarding technical information contained within this paper should be addressed to the following ECS contacts:

- ECS Contacts
  - Paul S. Roycraft, 301-925-0623, proycraf@eos.hitc.com
  - Carey Gire, 301-925-0615, cgire@eos.hitc.com

Questions concerning distribution or control of this document should be addressed to:

Data Management Office
The ECS Project Office
Hughes Applied Information Systems
1616 McCormick Dr.
Landover, MD 20785

This page intentionally left blank.

2. Data Ingest Scenarios

One of the issues identified at the ECS SDR concerning the ingestion of data was that the process might require significant manual quality assurance (QA) effort. The term QA was perhaps used incorrectly in this context, as performing QA on ingested data typically refers to science validation of the data. Science validation can be a complex and time-consuming process, and typically one that is performed when data is processed, reformatted, or otherwise manipulated. The ingest process involves only limited preprocessing and metadata checking, with more extensive processing being performed by the Data Processing Subsystem. The "QA" performed by ingest consists primarily of checking of the received metadata to ensure that it is complete and within valid specified ranges. The metadata checking process makes use of tools developed specifically for this purpose for each of the data types to be ingested. It is "automated" to the extent that an ingest client will call the particular tools that it requires to complete the ingest of a given data product. The manner in which these and other ingest functions are to be performed have been itemized in a series of ingest scenarios. Operational scenarios were developed for the major interactions of the Ingest Subsystem with external data providers to identify all of the functions to be performed for each of the principal interfaces. Four categories of interfaces with external data providers were identified and evaluated as candidates for automation of select functions. Scenarios were developed in support of these four interface categories and are presented in the following section.

2.1 Operational Data Ingest Scenarios

Four categories of external data providers were identified:

- Automated Network Ingest by means of a Data Availability Notice (DAN) supplied to ECS--ECS receives the DAN and schedules automated network data transfer from the source. The DAN describes the location of the available data. ECS either "gets" data from the source within a specified time window; or notifies the data provider that working storage space is available, at which time the data provider "puts" the data into a specified ECS working storage location.
- Polling Ingest with Delivery Record--ECS periodically checks an agreed-upon network location for a Delivery Record file. The Delivery Record file contains information similar to that in a DAN. The Delivery Record describes the location of the available data. Note: the data location may be on a working storage device within ECS, where an external data provider may have previously transferred the data. If a Delivery Record is located, ECS "gets" data from the source within a specified time window.
- Polling Ingest without Delivery Record--ECS periodically checks an agreed-upon network location for available data. All data in the location is assumed to make up a collection of ingest data with one file per data granule. If data is located, ECS "gets" data from the source within a system-tunable time window.

- Manual data transfer mechanisms are in place for transfer of data from hard media and for science user-controlled (interactive) network data transfer.
  - Hard Media Ingest is available for authorized institutions or science users providing data on hard media and as a backup mechanism for facilities where automated network data transfer is temporarily unavailable. The hard media must contain information similar to the Delivery Records described above, in a standard file format, or the data provider must separately provide Delivery Records in a standard file format.
  - Interactive Network Ingest is available for authorized science users to manually identify data to be ingested. Information similar to that contained in the Delivery Record is entered by means of Graphical User Interface (GUI) input.

Data transfer is accomplished by one of three means--file transfer protocol (ftp) "get", ftp "put", or hard media data transfer. Ftp get involves ECS "getting" data from an external site. Ftp put involves an external site "putting" data into ECS. Hard media data transfer involves data transfer from one of several ingest peripheral types found at a Distributed Active Archive Center (DAAC). Details of the functions to be performed for each of the scenarios are summarized in the following sections. These scenarios include references to ingest object classes which are described in Section 8 of the Science Data Processing Segment (SDPS) Design Specification for the ECS Project (305-CD-002-001).

2.1.1 Automated Network Ingest (Get) Scenario

ECS performs automated network ingest upon receipt of a DAN stimulus. The Landsat-7, Sensor Data Processing Facility (SDPF), Tropical Rainfall Measurement Mission (TRMM) Science Data Information System (TSDIS), and Science Computing Facility (SCF) interfaces are candidates for use of this protocol. In this scenario, ECS will perform the data transfer (get) from the external location to the ECS system. The following list describes the Automated Network Ingest (Get) scenario using object classes.

1. The Ingest server receives a connection from External Interface (IF) process
2. Instantiate Ingest Session and setup the connection between Ingest Session and the Data Center process
3. The External IF process sends an ingest Authentication Request
4. The Ingest Session verifies the External IF's privilege
5. and sends the Authentication Response to External IF
6. External IF sends a DAN message to Ingest Session
7. Log the receipt of DAN
8. Request to perform the automated network ingest request
9. Instantiate a InNetworkIngestRequest object
10. Assign an unique id for the ingest request

11. Insert the ingest request to the request list
12. Get the next request waiting to be processed
13. Begin the data transfer processing
14. Instantiate the InWorkingCollection object and populate it with the contents from the DAN message
15. Request allocation of an available staging device
16. Ingest system returns a DAA (DAN Ack) including the start time of data transfer to External IF
17. Get the file information
18. Instantiate the InFile object
19. Transfer file from the External IF to the resource object location in ECS Ingest system
20. Check the existence and size of the transmitted file
21. Log the file transfer result

Repeat steps 17-21 until all ingest files have been processed

22. Request the Data Type information
23. Get the Data Type information
24. Instantiate InIngestData object
25. Performs the appropriate data preprocessing (e.g., metadata check, conversion)
26. Log the sending of insert request to DataServer
27. Generate Data Insert Request
28. Send Insert Request to the appropriate DataServer
29. Log the completion status of data ingest

Repeat steps 22-29 for each Data Type

30. Send DDN (Data Delivery Notice) to the External IF informing the data transmission status
31. The External IF returns DDA (Data Delivery Ack)
32. Terminate the connection with the External IF

2.1.2 Automated Network Ingest (Put) Scenario

ECS performs automated network ingest upon receipt of a DAN stimulus. The interactive network ingest interfaces and future interfaces are candidates for use of this protocol. In this scenario, the external interface will perform the data transfer (put) to the ECS system. The following list describes the Automated Network Ingest (Put) scenario using object classes.

1. The Ingest server receives a connection from External IF process
2. Instantiate Ingest Session and setup the connection between Ingest Session and the External IF process
3. External IF sends an ingest Authentication Request
4. Ingest Session verifies the External IF's privilege
5. and sends the Authentication Response to External IF
6. External IF sends a DAN message to Ingest system
7. Log the receipt of DAN
8. Request to perform the automated network ingest request
9. Instantiate an InNetworkIngestRequest object
10. Assign an unique id for the ingest request and check the components in the request
11. Ingest system returns a DAA
12. Insert the ingest request onto the request list
13. Get the next request waiting to be processed from the list
14. Begin the data transfer processing
15. Request allocation of an available staging device
16. Ingest system sends a Data Retrieval Request (DRR) to notify the External IF that allocated space is available
17. External IF schedules data delivery
18. External IF process returns a DRR Ack (DRA) indicating the disposition of the request and initiates delivery of the data to ECS
19. Upon completion of transferring data to ECS, the External IF process sends a DDN specifying the data transmission status
20. Instantiate the InWorkingCollection object and populate it with the contents from the DAN message
21. Get the file information
22. Instantiate the InFile object
23. Check the existence and size of the transmitted file
24. Log the results of the file transfer result initiated by the External IF

Repeat steps 21-24 until all ingest files has been processed

25. Request the Data Type information
26. Get the Data Type information
27. Instantiate InIngestData object
28. Performs the appropriate data preprocessing (e.g., metadata check, conversion)
29. Log the sending of insert request to DataServer
30. Generate Data Insert Request
31. Send Insert Request to the appropriate DataServer
32. Log the completion status of data ingest

Repeat steps 25-32 for each Data Type

33. Return DDA to the External IF to acknowledge the data delivery by the External IF
34. Terminate the connection with the External IF

2.1.3 Polling Ingest (Files) Scenario

The Polling Ingest (Delivery Record) scenario describes the mechanism by which the Ingest Subsystem acquires data from data centers which may not support an interprocess communication interface with ECS. The ECS/NESDIS and ECS/GDAO interfaces are candidates for use of this scenario. The following list describes the Polling Ingest (Files) Scenario.

1. Instantiate InPollingSession (start timer) to detect new files in external directory
2. Initiate Processing Request
3. Create an InPollingIngest Request
4. Assign an unique id for the ingest request
5. Log receipt of request
6. Insert the ingest request to the request list
7. Get the next request waiting to be processed
8. Begin data transfer on a file by file basis
9. Request allocation of an available resource
10. Get the names of the files on external directory

11. Instantiate the InWorkingCollection Object and populate it with the information from directory
12. Determine the grouping of the files and associated data types
13. Compare the file list with appropriate table to find a matching DT(s)
14. Fill in the Data Type and the associated files
15. Request allocation of an available staging resource
16. Get the file information
17. Instantiate the InFile object
18. Transfer file from Data Center to the resource object location in ECS Ingest system
19. Check the existence and size of the transmitted file
20. Log file transfer result

Repeat steps 16-20 until all files have been processed

21. Request the Data Type information
22. Get the Data type information
23. Perform appropriate preprocessing
24. Log the sending of the insert request to data server
25. Send Data Insert Request to Data Server
26. Log the completion status of data ingest

Repeat steps 21-26 for each Data Type

2.1.4 Polling Ingest (Delivery Record) Scenario

The Polling Ingest (Delivery Record) scenario describes the mechanism by which the Ingest Subsystem acquires data from External IFs which control the initiation of data transfer. The EDOS interface is a candidate for use of the polling with delivery record interface. The following list describes the Polling Ingest (Delivery Record) Scenario.

1. Instantiate InPollingSession (start timer) to detect delivery record file in local directory
2. Initiate Processing Request
3. Create a InPollingIngest Request
4. Assign an unique id for the ingest request

5. Log receipt of request
6. Insert the ingest request to the request list
7. Get the next request waiting to be processed
8. Begin data transfer processing
9. Get the delivery record file
10. Instantiate the InWorkingCollection Object and populate it with the information from delivery record file
11. Get the file information
12. Instantiate the InFile object
13. Check the existence and size of the transmitted file
14. Log file transfer result

Repeat steps 11-14 until all files have been processed

15. Request the Data Type information
16. Get the Data type information
17. Instantiate InIngestData Object
18. Perform appropriate preprocessing
19. Log the sending of the insert request to data server
20. Send Data Insert Request to Data Server
21. Log the completion status of data ingest

Repeat steps 15-21 for each Data Type

2.1.5 User Network Ingest Scenario

The ECS provides the ECS users with the capability to perform interactive Network Ingest via the GUI Interface. A User Session will be configured on the User's GUI Interface to accept the request from the user via the GUI Interface and submit the request to the ECS system for data ingest. The following list describes the User Network Ingest scenario using object classes.

1. User selects the interactive Network Ingest option from the GUI interface and initiates a user interface session to read input from the user

2. User enters information needed for interactive network ingest and is read in as a network ingest request
3. The Ingest system verifies the User's privilege
4. Request to perform the user network ingest request
5. Create an InNetworkIngestRequest object
6. Assign an unique id for the data ingest request
7. Log the receipt of user interactive network ingest request
8. Insert the ingest request to the request list
9. Get the next request waiting to be processed
10. Begin the data transfer processing
11. Instantiate the InWorkingCollection object and populate it with the contents from the network ingest request
12. Request allocation of an available staging device
13. Get the file information
14. Instantiate the InFile object
15. Transfer the file from the user's disk to ECS disk working area
16. Check existence and size of the ingest files
17. Log the file transfer results

Repeat steps 13-17 for each file

18. Request the Data Type Information
19. Get the Data Type information
20. Instantiate InIngestData object
21. Performs the appropriate data preprocessing (e.g., metadata check, conversion)
22. Log the sending of insert request to Data Server
23. Generate Data Insert Request
24. Send Insert Request to the appropriate DataServer
25. Log the completion status of network data ingest

Repeat steps 18-25 for each Data Type

26. Display the completion status to User

2.1.6 Hard Media Ingest Scenario

The ECS system provides operations personnel with the capability to perform hard media (e.g., 8mm tape, optical disk, etc.) ingest via the GUI Interface. A Media Ingest Session will be configured on the Operator's GUI Interface to accept the request from the Operator via the GUI Interface and submit the request to the ECS system for hard media ingest. The following list describes the Hard Media Ingest scenario using object classes.

1. Operator selects the Media Ingest option from the GUI interface and instantiates media ingest session to read input from the operator
2. Operator enters information needed for media ingest and is read in as a media ingest request
3. The Ingest system verifies Operator's privilege
4. Request to perform the media ingest request
5. Create the InMediaIngestRequest object
6. Assign an unique id for the media ingest request
7. Log the receipt of physical media ingest request
8. Insert the ingest request to the request list
9. Get the next request waiting to be processed
10. Begin the data transfer processing
11. Request allocation of an available staging device
12. Request allocation of an available peripheral device
13. Download files from the media to disk working area
14. Search for the Data Delivery Record file and the result indicates FOUND
15. Instantiate the InWorkingCollection object and populate it with the contents from the Data Delivery Record
16. Get the file information
17. Instantiate the File object
18. Check the existence and size of the transmitted file
19. Log the data transfer result

Repeat steps 16-19 for every file

20. Request the Data Type Information

21. Get the Data Type information
22. Instantiate InIngestData object
23. Performs the appropriate data preprocessing (e.g., metadata check, conversion)
24. Log the sending of insert request to Data Server
25. Generate Data Insert Request
26. Send Insert Request to the appropriate DataServer
27. Log the completion status of data ingest

Repeat steps 20-27 for every Data Type

28. Display the completion status to Operator

2.2 Candidate Ingest Automation Functions

The above ingest scenarios were analyzed in conjunction with other ECS ingest operations concepts to identify automation candidates. Certain functions are basic to the operation of the ingest clients and are automated in the sense that these operations are performed on all ingested data sets. These operations include :

- Processing of the data header to identify the data set
- Checking for time-ordering and transmission errors
- Basic metadata (e.g., field presence, field size) extraction and checking

Additional ingest functions listed in the scenarios were analyzed and four potential areas for automation were identified. Based on development experience on past systems containing components similar to the Ingest Subsystem, it was expected that a large degree of automation would be possible. Additional analysis was performed to determine the effectiveness of the following approaches:

- Automation of system fault detection, reporting, and problem resolution
- Automation of Ingest Subsystem load balancing
- Automation of hand-shaking interactions between the Ingest Subsystem and external data providers for routine operations and to resolve identified problems
- Automation of portions of media ingest (e.g., for use in ASTER Level 1 data ingest) by means of robotics units

Section 3 discusses the evaluation of the automation of these functions.

3. Ingest Automation Evaluation

Each of the four design alternatives discussed under the Data Ingest Scenarios section was evaluated. The results of the analysis are presented in the following sections.

3.1 Automated Error Detection

Automated error detection and reporting is provided as a general Communications and Systems Management Segment (CSMS) service for all subsystems. The Fault Management Application Service provides the capability to diagnose, isolate, and recover from faults that occur in the managed objects within ECS. It provides tools to support activities such as tracing faults through the system, running diagnostics, and initiating corrective or recovery actions. The Management Agent monitors managed objects and forwards selected event data to the Fault Management Application Service which determines if this state change represents a "reportable" fault. The diagnosis and isolation of a fault involves determining the cause from the reported symptoms and using diagnostic tests as appropriate. Recovery involves the initiation of corrective action to restore the system to normal operational status. The entities or managed objects in ECS that need to be monitored for faults include physical devices (for which vendor diagnostic programs are generally available), and operating systems and applications. The Ingest Subsystem will utilize these capabilities in the monitoring and diagnosis of system faults. Automated problem resolution requires artificial intelligence or similar capabilities that are not cost-effective due to the extremely large variety of data (and potential data problems) involved.

3.2 Automated Load Balancing

Automated load balancing may include two separate but related functions within the Ingest Subsystem. The first would allow Ingest Subsystem functions to be spread across multiple hardware devices to increase subsystem throughput. The ability to take advantage of the addition of I/O and CPU capabilities as throughput requirements increase may become important to the Ingest Subsystem as well as to other ECS subsystems such as the Data Server Subsystem and the Data Processing Subsystem. This capability is not currently implemented, but may be added as I/O and CPU requirements dictate. The Ingest Subsystem will reuse general resource management capabilities developed for these other subsystems, tailoring them as required to support data ingest requirements. The second automated load balancing area provides a means for the tuning of Ingest Subsystem parameters. This includes the ability for operations personnel to set system performance thresholds and have the system automatically monitor and perform simple system tuning. Ingest thresholds that are tunable include: 1) the timer for automated file polling, 2) the maximum number of ingest requests allowed to be processed concurrently, 3) the maximum data volume allowed to be processed concurrently, and 4) the number of data transfer retries. This is useful both to increase subsystem performance and to ensure that the interaction with other subsystems (e.g., feeding data to the Processing Subsystem) occurs within parameters that are acceptable to both subsystems.

3.3 Automated Interface Interactions

Automated hand-shaking and interactions for some classes of interface problems were addressed by the SDPF, which is the template for Ingest Subsystem interface processing. The SDPF defines interface specifications, network protocols, interactive services, and procedures for its interfaces with data consumers. The control message scheme described in the SDPF-TRMM Consumer Interface Control Document (ICD) defines the automated mechanisms used for the interchange of messages and data between the SDPF and the Consumer System application software. It is anticipated that the Ingest Subsystem will treat hand-shaking similarly. These mechanisms will be identified in the appropriate ICDs prior to the SDPS CDR.

3.4 Automated Hard Media Ingest

Ingest of data via hard media will be required in support of Version 0 data migration, Release B operational data ingest (e.g., ASTER data products), and other ad hoc data products. Hardware from the Distribution and Ingest Peripheral Management Hardware Configuration Item (HWCI) will be used to read the media using an ingest client specialized for the particular data product. The reading of the media may be automated to some degree (e.g., use of an 8 mm tape stacker) depending on the media type, data volume, and frequency of ingest of the particular data type. The types of ingest/distribution media devices will be chosen to accommodate the prevalent types of media in the user community and at other sites. Additional media devices required specifically for the ingest of certain data products may be added as required. The number of devices, as well as the addition of automated or robotics-based devices will depend on the needs at each of the DAACs. It is anticipated that for the regular ingest of operational data products such as ASTER that an automated ingest device will be used to decrease the amount of operator involvement. This evaluation will be done on a case-by-case basis as hard media data products are identified.

3.5 Manual Ingest Activities

For all four categories of external data provider, four types of manual activities to be performed by operations staff personnel were identified: monitoring of ongoing requests; viewing of completed ingest status; viewing of detailed error information; and setup of thresholds for automated load balancing. Relative frequencies for each activity will be identified as the ICDs between the data providers and the Ingest Subsystem are finalized. At this time, it is not anticipated that any activity or set of activities requires a substantial amount of personnel resources. Automation options were generally not found to be readily available for the identified tasks. Additionally, automation was deemed not to be cost-effective due to low frequency of these activities.