Data Analysis of the Multimission Satellite Product Generation Pattern for Defining the Archival Policy in the Data Centre

C. Pradeep[1], Gaurav Gupta[2], Murali Krishna[3], G. Prasad[4]

[1]*pradeepcps4[at]gmail.com*
[2]*g.gauravgupta111[at]gmail.com*
[3]*murali.annadanam[at]gmail.com*
[4]*gurram_p[at]yahoo.com*

Abstract: ***The heterogeneous data centre at NRSC is involved in the data acquisition and archival of satellite data from multiple missions for data processing and dissemination. Areas of interest data request from users are specific to a spacecraft and sensor and for a specific application. Thus there will be numerous requests of different kinds to be handled at the data centre. The proposed data analysis enables effective data search, selection and processing from within very voluminous and constantly updated data archives in the Storage Area Network. Analysing and understanding the user requirements and trends received from the extracted data stored in the user order archives database is the basis for determining the archival policy. The user data requests may correspond to any date starting from the launch of the spacecraft which has direct impact on IT infrastructure resource utilization. The information feed is then validated to discover the trends of data requests that are being handled on a daily basis. The administratively modifiable pre-defined storage and archive policies of the SAN have been fine tuned to retain the relevant datasets in the active storage of the hierarchical storage infrastructure. This analysis framework is more curated to emphasis on the Product Generation TAT, Quality of Service and resource optimization and management. This paper describes the system improvement achieved with the user order data analysis system developed at National Remote Sensing centre, ISRO.***

Keywords: HSM, IMGEOS, SAN, NRSC, UOPS, TAT

1. Introduction

Over the years, NRSC has been involved in satellite data acquisition, data processing and data dissemination. The data dissemination area is the front end of NRSC and is primarily responsible for exhibiting the capability and availability of various satellite data products acquired and processed at NRSC. It also takes care of data orders placed by the user and disseminate the same to them. The preparation and execution of data order is an important for a good customer relationship management. Upon receipt of a user request the data dissemination wing of NRSC, activates a work flow manager service for the required data product generation either on the data processing chain or data acquisition chain based on the details of orders. Both the acquisition and data processing chains are part of the main chain known as the data flow chain. The data flow chain consist end-to-end digital processes that are interlinked with each other and implemented with highly customized solutions. For the data flow chain to operate, massive amount of information has to be received, processed, transmitted, shared and managed. This order processing system is capable of handling multiple orders from different users on different channels and can be streamlined; the whole process of executing data order is operationally focused on transactions. To accomplish the task of user product generation the work centres rely on complex IT Infrastructure placed in the IMGEOS Data Centre.

Data centre is provisioned with large capacity centralized storage and the amount of data transactions from and to the storage system is in the order of millions per second. A Hierarchical storage system is implemented in the data centre with different classes of storage devices which vary in performance, capacity and architecture. IMGEOS data centre is designed to support the data processing capability of handling 1000 data products per day. There is enterprise class filesystem software that handles the entire storage system in every aspect of administration, management and operations. One of the functionality of the filesystem is task scheduling for handling enormous amount of data and transactions, the scheduler categorize all the tasks as independent requests posed on it. The task scheduler for retrieving data from archives assigns unique Identification number and aligns them in the request queue in first come first serve basis. The data processing software process interacting with the filesystem is unaware of the internal processes of the filesystem, they just request for data read and write operations. Amidst, there are scenarios that demand a particular data-set / product requirements from the user end on priority, at this juncture neither the data processing software nor the filesystem software are transparent to the priority, due to which there are some delays in those priority order request. This delay can be translated as the duration taken by the process to search, sort data from this huge data collection. These delays can be fine-tuned by modifying the policies for data storage and archives and this needs administrative intervention with proper technical analysis on the data requests received from the ordering system. The elements in the data centre utilized for serving the user requests are 1. UOPS: User Order Processing System and IT Infrastructure of the related work centre and work flow managers 2. SAN. Storage Area Network.

2. Literature Survey

Storage Area Network (SAN)

The storage architecture for storing and archiving large volumes of satellite data is designed with extensible features to support availability, accessibility, reliability, expandability, integrity, retention, compliance and security. All these requirements are provided under a balanced storage paradigm known as the Storage Area Network i.e., SAN. SAN is a specialized, high speed network that provides access to data blocks known as LUNs (Logical Unit Number). SAN storage is typically composed of switches, storage elements, disk arrays and tape drives / libraries that are interconnected using multiple technologies, topologies and protocols. The LUNs are presented to the host computing systems as block devices using Fibre Channel Protocol. A File system configured with HSM (Hierarchical storage management) capabilities was implemented to provide complete storage virtualization under multiple layers of the system. The HSM handles automated migration of data objects among the storage devices, usually based on data activity. In our case we have implemented the three TIER SAN architecture with TIER I & Tier II as disk based storage arrays and TIER III a robotic tape library. These policies are framed for the purpose of systematic identification, categorization, maintenance and retention of satellite data received or products generated in the SAN environment.

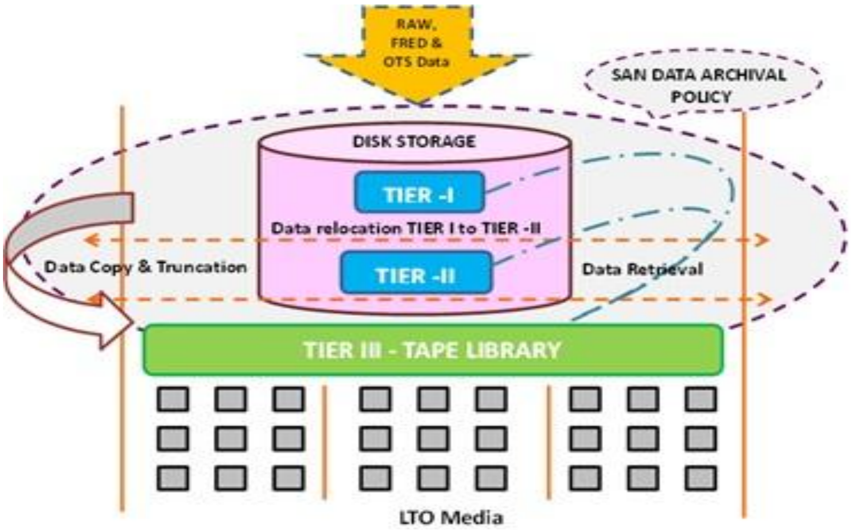


Figure 1: Overview of IMGEOS SAN Architecture

The whole SAN architecture and its operations are configured into 3 layers as 1. Presentation layer 2.Management layer 3.Data layer.

Presentation layer

The presentation layer establishes the way in which the data is presented, typically displayed at the hosts / work centres. The File system client agent is installed on these hosts / work centres; therefore it translates information about the data in a way that host Operating System (OS) and applications can understand. The detailed view of data archival structure, type of data, data owners and user permissions, the data blocks / LUNs are presented to the users under File system labels. The mapping context between user, application and the data under the File system label is presented seamlessly irrespective of all data TIER in which it is archived.

Management layer

The storage manager, the key component handles the management part of SAN storage; management is tedious at this higher level of aggregation through the use of Device Mappers (DM), Logical Volume manager (LVM) and Multiple Devices (MD). The storage manager alleviates this by providing a friendly unified user interface, which allows applications, administrators and users to run complicated tasks. The storage manager delivers full data lifecycle management from data creation till the end. As the data moves across the three TIER storage system, the manager provides continuous data access.

The data archival polices are created, implemented and maintained with the support of Hierarchical Storage Manager (HSM). The HSM controls the data movement across storage tiers. It is responsible for the identification of the tier in which a particular data is stored. It takes care of the data retrieval from LTO media in the Tape library on to disk storage and truncates data from disk. It acts as a software controller for the mechanical drivers in the tape library. It works in coordination with the other two lesser level Media storage manager (MSM) and Tertiary storage manager (TSM).It handles the integration of itself with other software components responsible for tape library operations. It contains the master database records for the entire SAN storage infrastructure. The master database includes the minute level of information such as LUN IDs of disk storage, starting and ending of data blocks and Media related identifiers. Storage manager will direct and interfaces the storage agents installed in the client systems with the storage server. The storage manager will take care of file migration between storage tiers. Keeps track of files/directories creation, modification and deletion.

Data layer

This layer is the physical file system layer and is concerned with the physical operations of the storage devices. This layer consists of disk controllers, disk arrays, LTO media, and robotics, switching and interfacing components. In this layer clients are physically interfaced via connections to create zones with the storage units. After the zoning process mapping of the data blocks / LUNs to the client systems are performed here. There are separate GUI interfaces for the disk controllers, which will be used for mapping LUNs / data blocks to the destined client system. The actual data resides here in its native form. Table-1 shows statistics of load on the HSM currently at IMGEOS approximately.

Table1: Data retrieval statistics in SAN

Number of files retrieved per month	60,000
Equivalent volume of data retrieved	60TB
Volume of data archived per month	194TB

Method adopted: Data Analysis carried out from requests received at User Order Processing System

The User Order Processing System is a web based satellite data ordering application hosted by NRSC in Internet domain. The ordering application allows the users to register themselves with some validations and provide them with login account credentials. The users can login to the application using the credentials to span through the collection of Earth Observation satellite data; this application showcases multi-platform, multi-resolution, multi-temporal and synoptic data. The user is free to select his choice of data from the list provided, means they can choose the satellite, sensor, latitude and longitude (Area of Interest), date of pass, resolution and the priority it can be delivered.

Based on the priority, the dissemination options are classified as emergency products, urgent products, normal products, value added and subscription products. The mode of dissemination is through SFTP, Media DVD, photo print, Web Service – Bhoonidhi. A file is created after the user frames his requirement as an order, the data order file is received and transformed into work order and workflow is triggered in the internal data flow chain of IMGEOS for generating the data product.

Data Analysis on User Orders serviced is an in-house developed framework that examines the user order data requests in the form of xml files in order to obtain analytical conclusions. These conclusions are based on the information extracted from those data sets. The initiatives brought under this analysis can help in increasing the efficiency of operations, optimization of data archives and better services. The analysis is performed on large number of user requests requiring diverse data sets from a huge data collection.

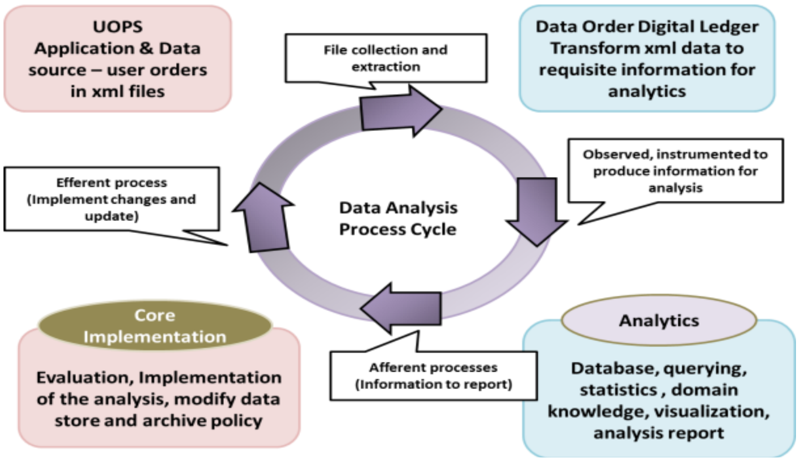
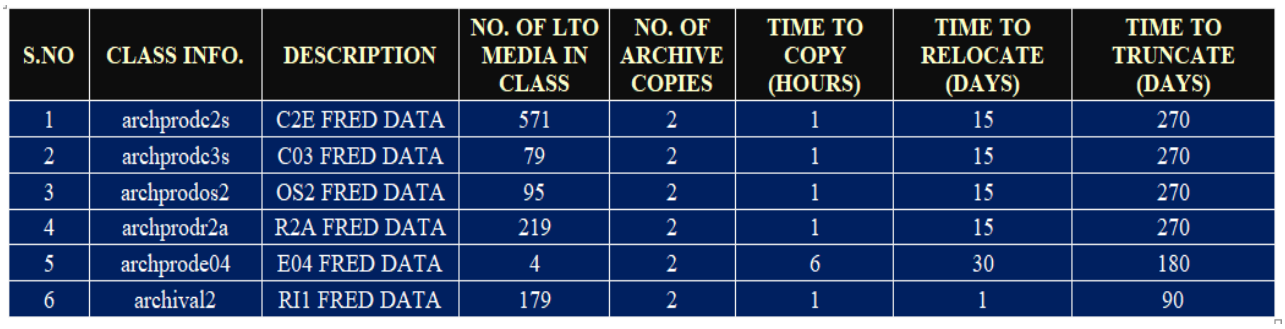

UOPS
Application & Data source – user orders in xml files
File collection and extraction
Data Order Digital Ledger Transform xml data to requisite information for analytics
Observed, instrumented to produce information for analysis
Efferent process (Implement changes and update)
Data Analysis Process Cycle
Core Implementation
Evaluation, Implementation of the analysis, modify data store and archive policy
Afferent processes (Information to report)
Analytics
Database, querying, statistics, domain knowledge, visualization, analysis report

Figure 2: User Order Data Analysis workflow

The data requests are processed and parsed to discover patterns, data requirement trends, and correlations. Then the semi-processed data is stored in a database known as Data Request Digital Ledger, this digital ledger is audited on a particular frequency to extract the inputs to optimize the data archival policies. The data analysed is satellite mission specific, it can be historical data or new data which is specifically processed for that particular satellite mission based on the real-time data requests received from the user order processing system. Table-2 explains the storage policy definitions and parameters defining the data lifecycle.

Table 2: SAN Archival class information for missions



S.NO	CLASS INFO.	DESCRIPTION	NO. OF LTO MEDIA IN CLASS	NO. OF ARCHIVE COPIES	TIME TO COPY (HOURS)	TIME TO RELOCATE (DAYS)	TIME TO TRUNCATE (DAYS)
1	archprodc2s	C2E FRED DATA	571	2	1	15	270
2	archprodc3s	C03 FRED DATA	79	2	1	15	270
3	archprodos2	OS2 FRED DATA	95	2	1	15	270
4	archprodr2a	R2A FRED DATA	219	2	1	15	270
5	archprode04	E04 FRED DATA	4	2	6	30	180
6	archival2	RI1 FRED DATA	179	2	1	1	90

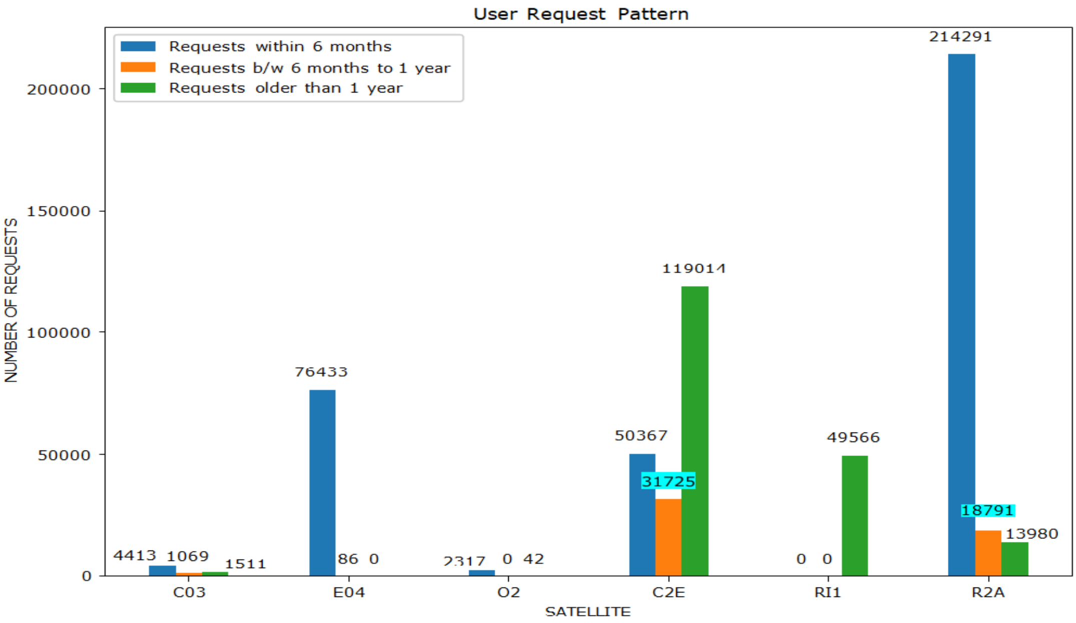


Figure 3: User Order request pattern

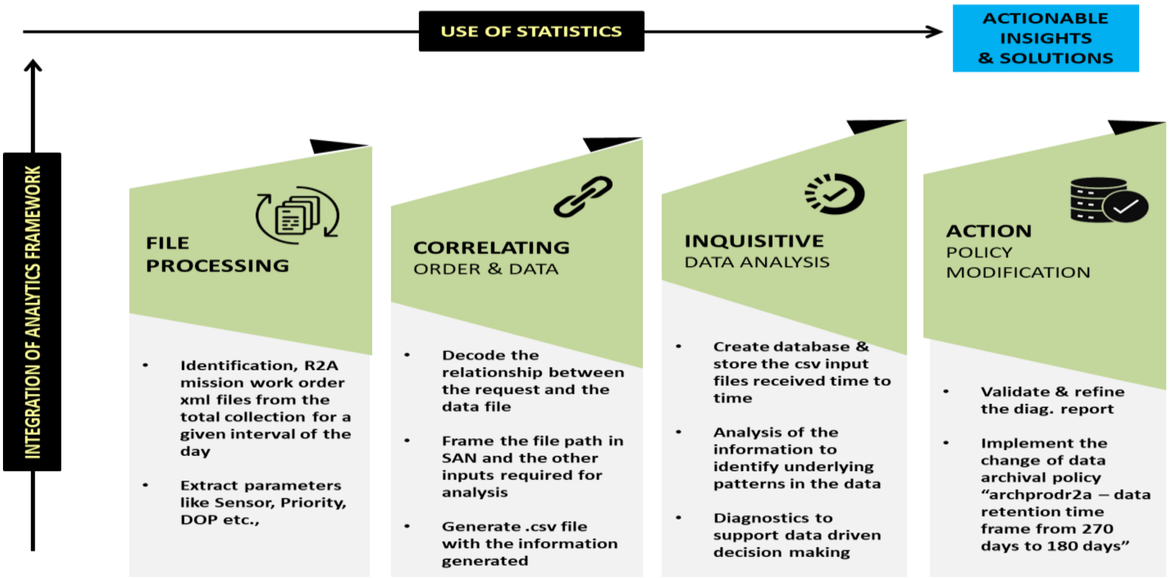
Results: Based on the data analysis, designed and implemented the Archival Policy

Figure-3 indicates the user requests corresponding to spacecraft data products which need access to raw data for product generation. The raw data corresponds to different timelines w.r.t the user request date. Let us consider RESOURCESAT-2A satellite for analysis. It consists of 3 sensors namely LISS-4, AWiFS and LISS-3. The mission specific information of each user order is collected from UOPS work order files and populated in a database having following structure:

The required parameters are extracted from the user orders helping to build and update database. From the distributed input information, the precise SAN Path is framed for data retrieval. Currently, analysis window was picked up from August 2021- October 2022; around 2.5 lakh user orders have been received for this satellite mission with priority of orders ranging from standard product to emergency product. Patterns are categorized into three different types:

Type I- Date of requests within 6 months of Date of pass
Type II- Date of requests between 6 months and 1 year from Date of pass
Type III- Requests older than 1 year from Date of pass

Currently the policy “archprodr2a” is used for archiving the raw data of the satellite having truncation time of 270days. Post analysis of the requests, it is observed that about 2.14 lakh user orders were within 6 months, around 18,000 requests between 6 months to 1 year and 14,000 requests which are older than a year from date of pass.


USE OF STATISTICS
ACTIONABLE INSIGHTS & SOLUTIONS
INTEGRATION OF ANALYTICS FRAMEWORK
FILE PROCESSING
CORRELATING ORDER & DATA
INQUISITIVE DATA ANALYSIS
ACTION POLICY MODIFICATION

- Identification, R2A mission work order xml files from the total collection for a given interval of the day
- Extract parameters like Sensor, Priority, DOP etc.,

- Decode the relationship between the request and the data file
- Frame the file path in SAN and the other inputs required for analysis
- Generate .csv file with the information generated

- Create database & store the csv input files received time to time
- Analysis of the information to identify underlying patterns in the data
- Diagnostics to support data driven decision making

- Validate & refine the diag. report
- Implement the change of data archival policy “archprodr2a – data retention time frame from 270 days to 180 days”

Figure 4: Use case workflow of R2A mission

Based on the results of the analysis the archival policy “archprodr2a” was modified. Since the analysis has indicated that more than 85% of user requests are falling under Type-I category. The truncation time was reduced from 270 days to 180 days, thereby enabling the administrator to optimally allocate storage to populate data of other relevant policies.

Hence C2E data which has good number of requests up to one year old data i.e., the archprodc2s policy truncation period was increased from 270 to 365 days. Thus the analysis helped us to enhance performance of the storage/archival system in the following manner.

a) **Reduced TAT for product generation:** Reduced number of transactions in file handling in the work flow will reduce the time required for generating a product and improves efficiency of data processing segment from the order processing cycle.
b) **Resource and Service Optimization:** Reduced load per product generation on the infrastructure elements by storing data on high demand on disk. Thus most of the requests can be handled without accessing the archives which have direct implication on loading and unloading the tapes with the help of robotic arm movements**.**
Example: no. of files per product creates internal queue of requests, which in turn translates as robotic movement, load / unload LTO media and read/write head usage in drives.
c) **Reduced downtime with Infrastructure maintenance:** As a result of optimized usage of resources, the downtime and maintenance window is reduced

3. Conclusion and Future

The data analysis carried out and implemented has resulted in defining the archival policies of the various missions and helping in betterment of overall system and service performance. The emergency data products could be handled more effectively especially during disasters. It has provided inputs as which data is the most in demand based on the pattern of user requests. The load on the Data Centre infrastructure operations is optimized thus indirectly aiding in the maintenance of the infrastructure. The data analysis method can be extended in the future for real time tracking of order movement with unique order id. Optimize the resources and storage policy definitions, predictive analysis and user behaviour analysis.

References

[1] IMGEOS Project Report, NRSC-DPA-IMGEOS-Apr09-TR64.
[2] IMGEOS Infrastructure Document,NRSC-DPA-IMGEOS-JAN10-TR147.
[3] IMGEOS Software Interface Control Document, NRSC-DPA-IMGEOS-JAN10-TR141.
[4] Preliminary Design Review Document, NRSC-SDAPSA-FEB-2014-TR-584.
[5] Data Ingest and Processing Interface control Document C2C/NRSC/Level -0/ICD/Ver-1.0/May 2015.
[6] Work Flows Interface Document, NRSC/DPPA&WAA/SG/C2C-DPWFM/ICDV1.0/Dec2015.
[7] Critical Design Review Document, NRSC-SDRISA-JAN-2016-TR-788.
[8] Quantum Tiered Architecture and software 6-68507-01 v5.4.1 July 2017
[9] SNIA’s “education /tutorial s/ 2008 – Storage tiering for file systems.
[10] Quantum Tiered Architecture and software 6-68507-01 v5.4.1 July 2017.
[11] C Pradeep, ANSV Murali Krishna and G Prasad “A New Approach for Data Retrieval from Hierarchical File System to Handle Priority Satellite Data Products Dissemination”. Published in the International Journal of Scientific Engineering and Research (IJSER). Volume 9, Issue 12, December 2021

Author Profile

C. Pradeep received B.E. degree in Electronics and Communication Engineering from Anna University, Chennai and M.Tech degree in Software Systems from Birla Institute of Technology and Science, Pilani. Email: pradeepcps4@gmail.com

Gaurav Gupta received B.Tech degree in Computer Science and Engineering from Rajasthan Technical University, Kota. Email: g.gauravgupta111@gmail.com

ANSV Murali Krishna received Bachelor of Engineering in Electronics and Communications Engineering from Osmania University, Hyderabad
Email: murali.annadanam@gmail.com

G. Prasad received M.Tech degree in Advance Electronics, from Jawaharlal Nehru Technological University, Hyderabad, Master of Business Administration in Operations Research from Indira Gandhi National Open University, New Delhi and Ph.D. in VLSI Design from Jawaharlal Nehru Technological University, Hyderabad. Email:gurram_p@yahoo.com