Collecting and Processing Earth Science Data Metrics at NASA ESDIS

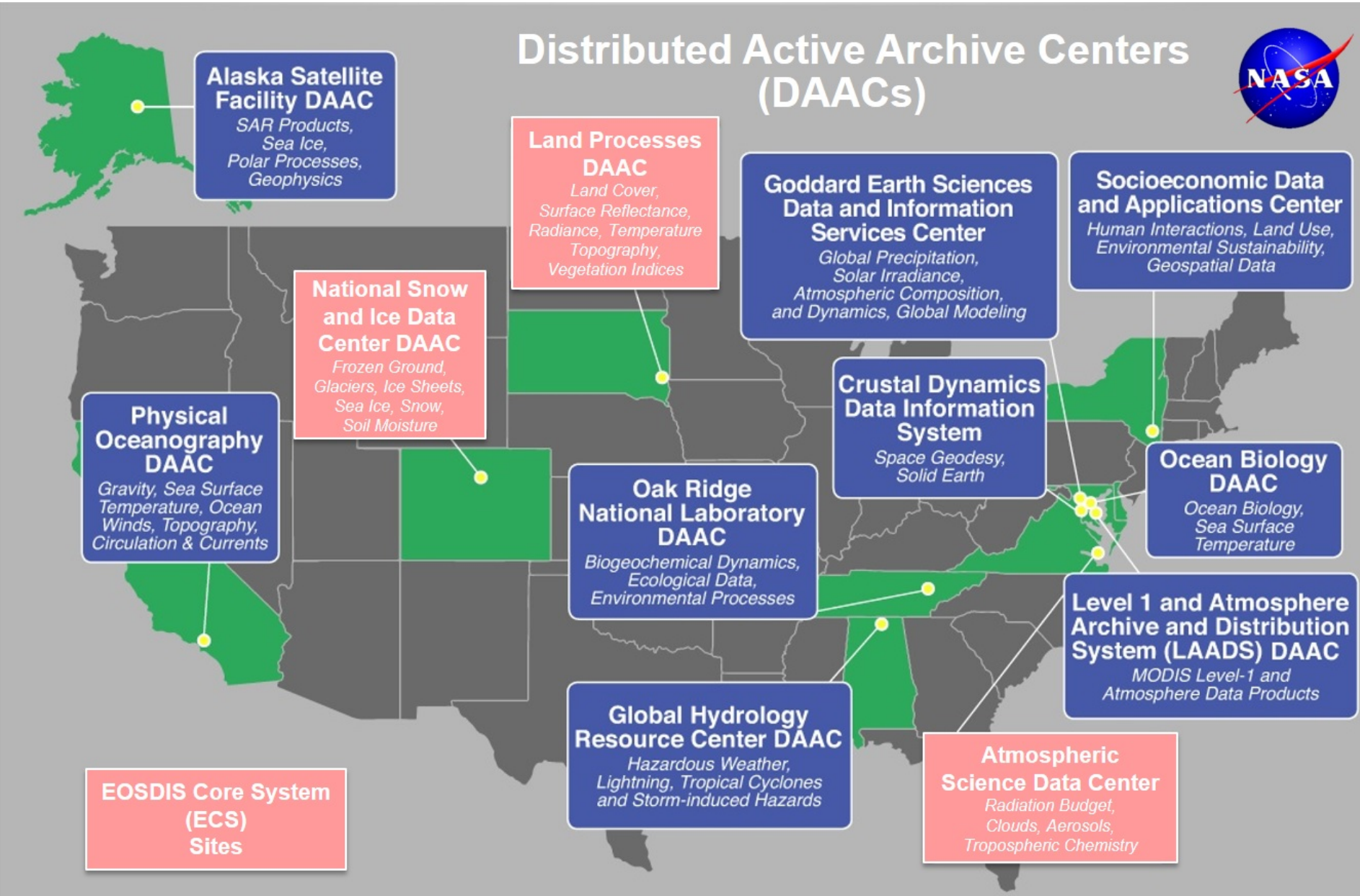
Jianfu Pan[1], Nelson Casiano[1], Stephan Klene[1], James Smith[1], Lalit Wanchoo[2], Durga Kafle[2]
[1] SSAI, Inc., Lanham, MD 20706; [2] Adnet Systems, Inc., Lanham, MD 20706

INTRODUCTION

NASA's Earth Science Data and Information System (ESDIS) project manages the science systems of the Earth Observing System Data and Information System (EOSDIS), which is responsible for processing, archiving, and distributing a wide range of Earth Science data products. These data activities occur at twelve EOSDIS Distributed Active Archive Centers (DAACs) located throughout the United States (Figure 1), with each DAAC focusing on specific science discipline areas within the Earth Sciences. DAACs receive and/or produce data products of various missions (satellite, aircraft, field campaign, model, in-situ, and others), then ingest them into their product system. Science data products are then distributed to end users. DAACs are also required to provide metrics or activity logs of ingest, archive and distribution to ESDIS Metrics System (EMS).

EMS collects information on ingest, archive and distribution, along with product metadata from all DAACs. EMS daily processing converts the information received into various metrics. The metrics help both ESDIS and DAAC management in resource planning, understanding user behaviors, and identifying important data products that are popular to user communities.

Figure 1. NASA EOSDIS Distributed Active Archive Centers (DAACs).
(source: https://earthdata.nasa.gov/eosdis/daacs)



EMS PROCESSING SYSTEM

Each day, DAACs collect logs of ingest, archive, and distribution of science data products and send them to EMS for processing. One example of the distribution logs is the Apache log produced by an Apache web server where end users access and download the data. Other forms of records include what is called customer logs, which are constructed by DAACs to include selected metrics fields before being sent to EMS.

Figure 2 illustrates the architecture of the EMS log processing system. DAACs send logs collected to a dedicated file server serving as an interface between DAACs and EMS. EMS picks up logs and other files from the file server for processing. The interface file server is also used by EMS to store certain specialized reports for DAACs, while more standardized reporting is done through the Oracle Application Express (Apex/HTMLDB) platform, a web-based Oracle application and reports development tool. All processed metrics are permanently stored in the EMS database. The Apex/HTMLDB platform is used to make metrics reports, and is accessible by all DAACs.

EMS processing runs two instances in a semi-parallel fashion, with steps involving database processing can only be single-threaded. In addition, it also uses an independent preprocessing module, called pre-staging, to preprocess files before being further processed by the EMS workflow. Within an EMS processing instance, a typical workflow consists of the steps in Table 1, executed in that order.

Table 1. Workflow steps in EMS processing system

▪ *GetFiles*	Gets log files stored on the interface server for processing, normally in batches.
▪ *StageFiles*	First major processing step for hostname and IP geolocation lookups.
▪ *LoadFiles*	Wrapper to Oracle processing for validation, user profiles lookup, and standardization of field values. Creates output files on the database server.
▪ *FileXfer*	Transfers outputs created on the database server to local machine for further processing.
▪ *ProductSearch*	Maps distribution records to a science products for product usage metrics.
▪ *DistSummary*	Consolidates distribution records and saves the final metrics results to the EMSdatabase.

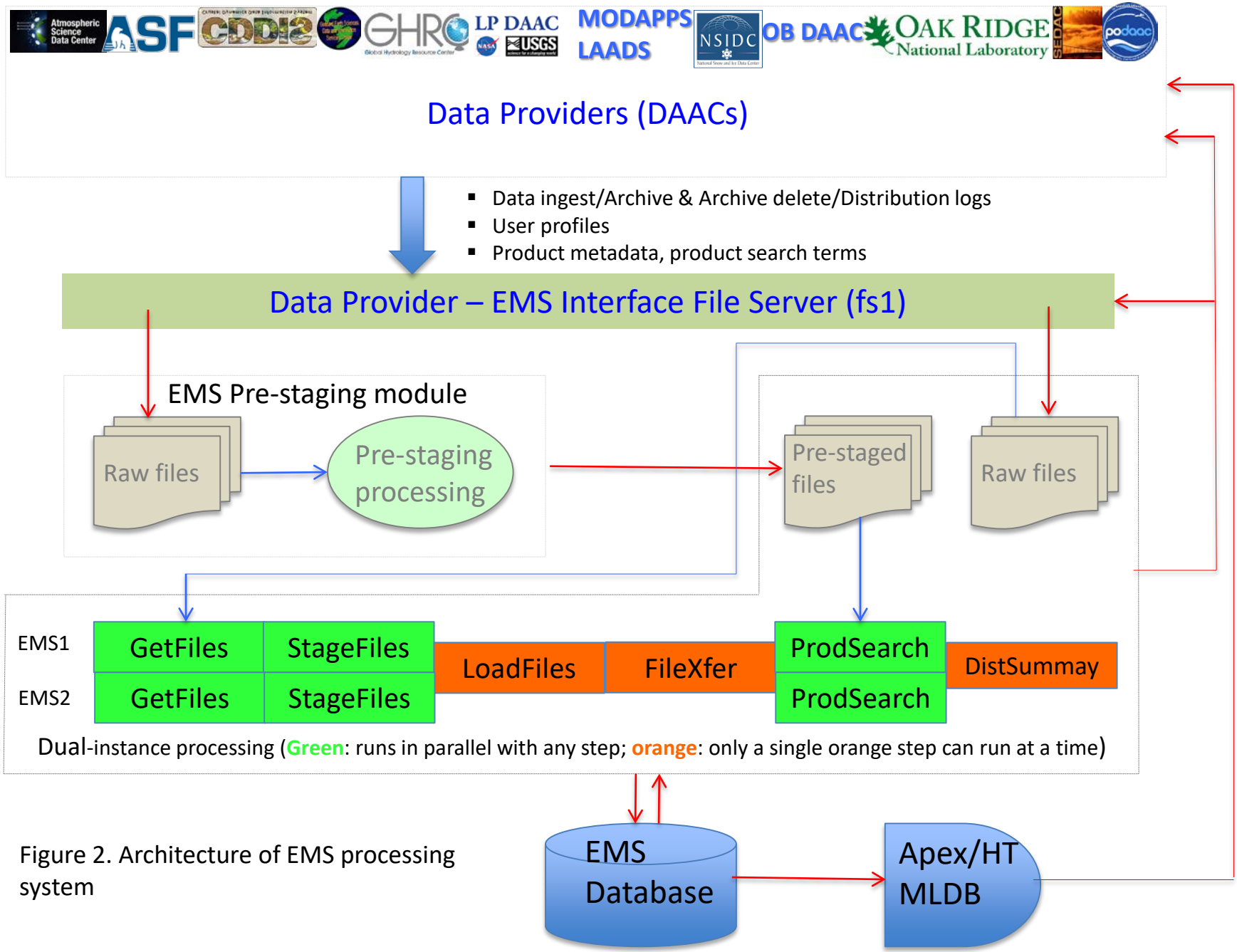
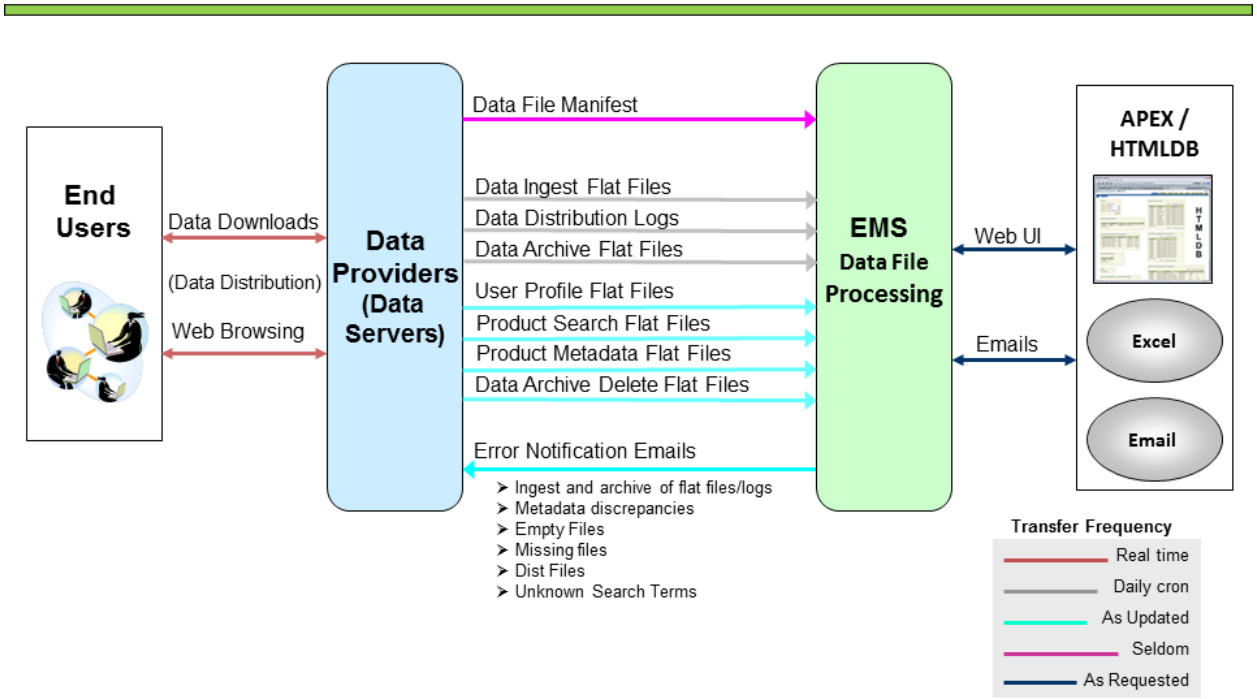


Figure 2. Architecture of EMS processing system

Figure 3 is a schematic diagram illustrating communication between DAACs, also known as Data Providers, and EMS. DAACs send all logs and other necessary informational files, such as product metadata, to EMS as flat files. These are plain text files with delimited fields that are normally pre-defined and registered in EMS. The fields and formats follow a mutually agreed upon Interface Control Document (ICD) between ESDIS and DAACs. While flat file format is used to exchange logs and file-based reports, Data Providers use the APEX/HTMLDB reporting tool to create metrics reports and monitor file processing status (Figure 4). From time to time, Data Providers also receive email notifications from EMS for announcements and processing errors if any.

Figure 3. Communication between DAACs (Data Providers) and EMS



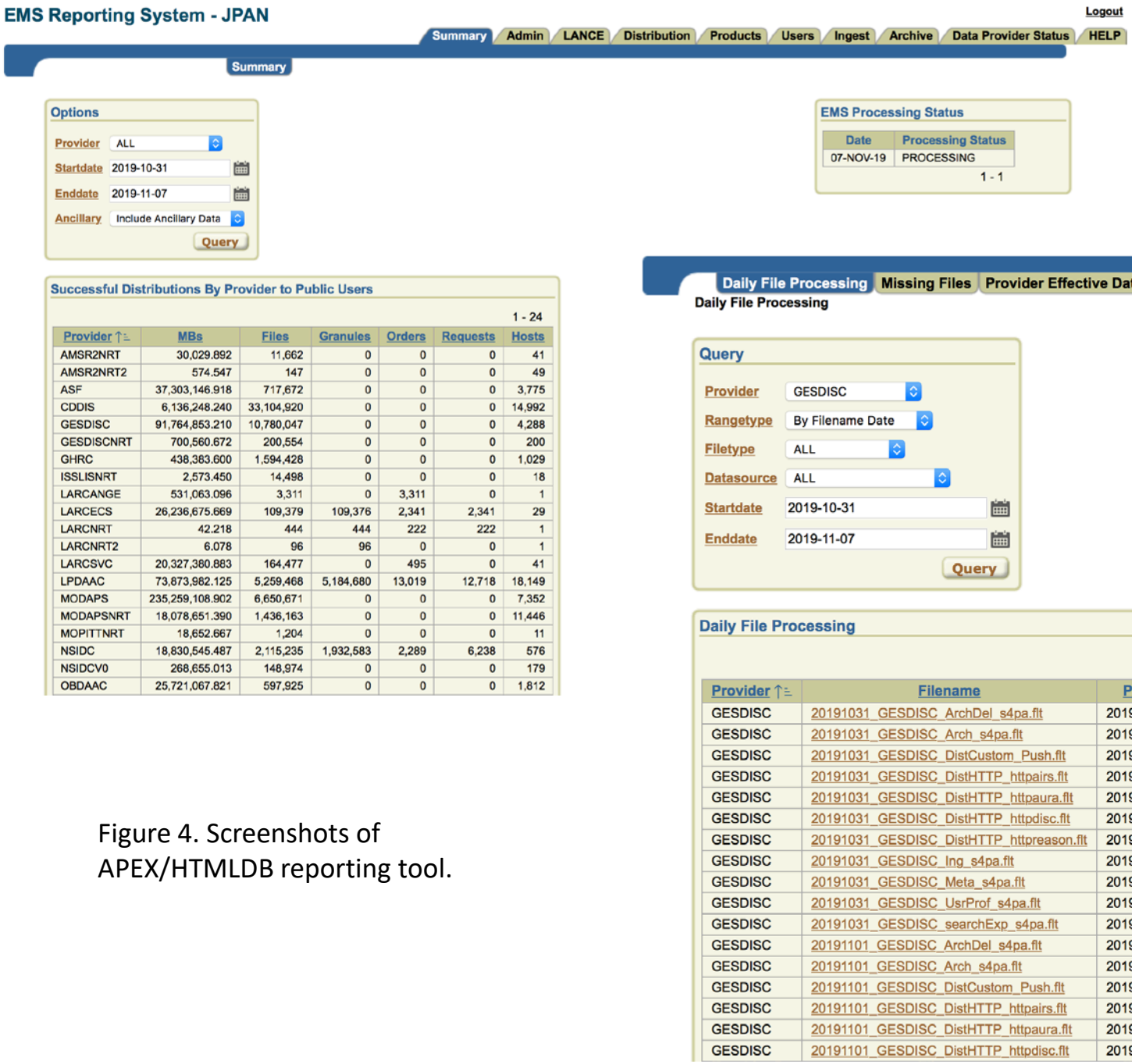


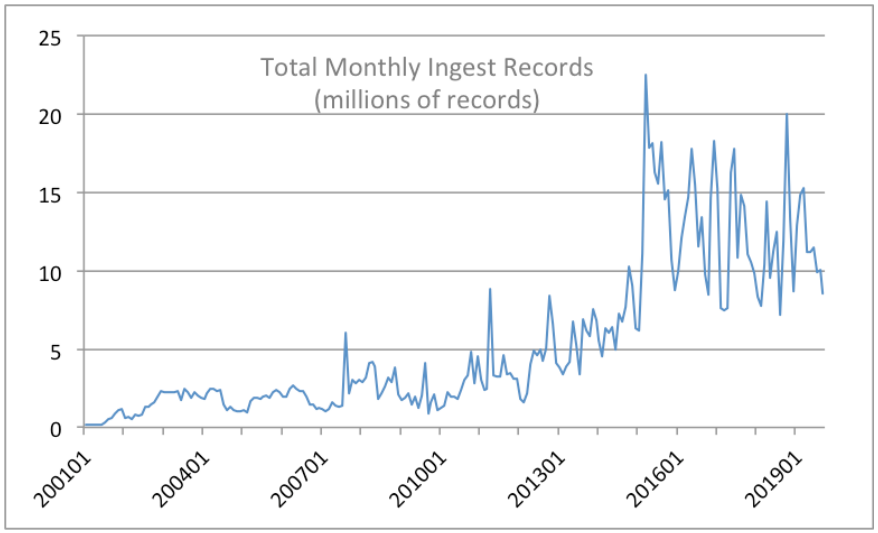
Figure 4. Screenshots of APEX/HTMLDB reporting tool.

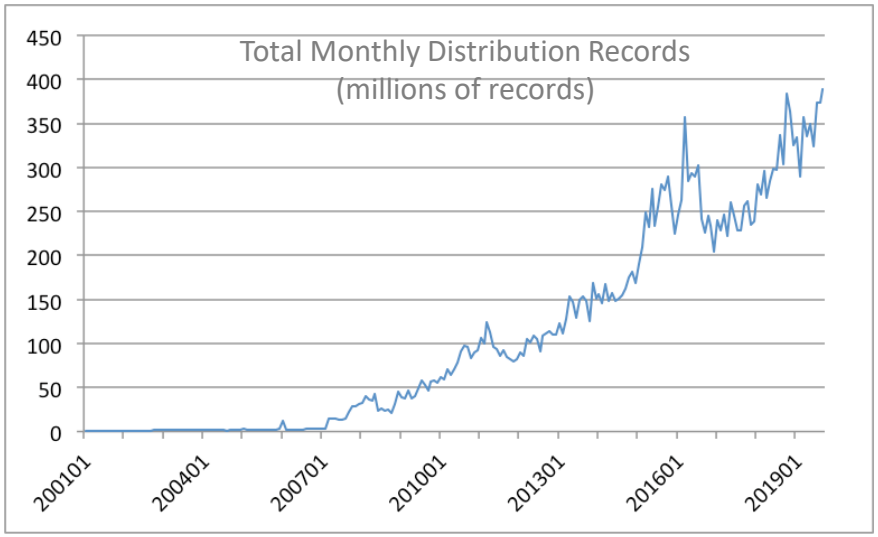
EVOLUTION OF EMS PROCESSING SYSTEM

Early version of the EMS processing system used a simple pass-through multi-step processing workflow of the steps described earlier. With the increase in the data provided by the data providers over years, the system took significant time to process all of the data on daily basis and required to be upgraded to improve the efficiency of processing. Figure 5 shows monthly total number of records processed by EMS over years, for ingest, archive, and distribution. It seems that the amount of data distributed doubles every 3 to 4 years.

The first major upgrade introduced parallel processing, taking advantages of the two most time consuming steps, StageFiles and LoadFiles, being run on separate servers with StageFiles on an application server and most work in loadFiles being database processing on a database server. Without the parallelization, one server would be idle for nearly half of the time, waiting for the other step to complete before moving along in the workflow. More recently, we also implemented a prestaging process that can run either independently or as part of the EMS processing workflow. The system architecture with the prestaging is illustrated in Figure 6. The prestaging module preprocesses files and can run independently. The preprocessed files can then skip several time-consuming steps when being merged into the EMS process work flow.

These system upgrades have greatly improved the processing performance. As shown in Figure 7, EMS was running above 20 hours each day for its daily processing before around 2015 for several years, making the system highly stressed and very challenging for down times needed by hardware maintenance and occasional file reprocessing. The upgraded system has been steadily running with about 10 hours of processing time, with just three largest files are preprocessed.





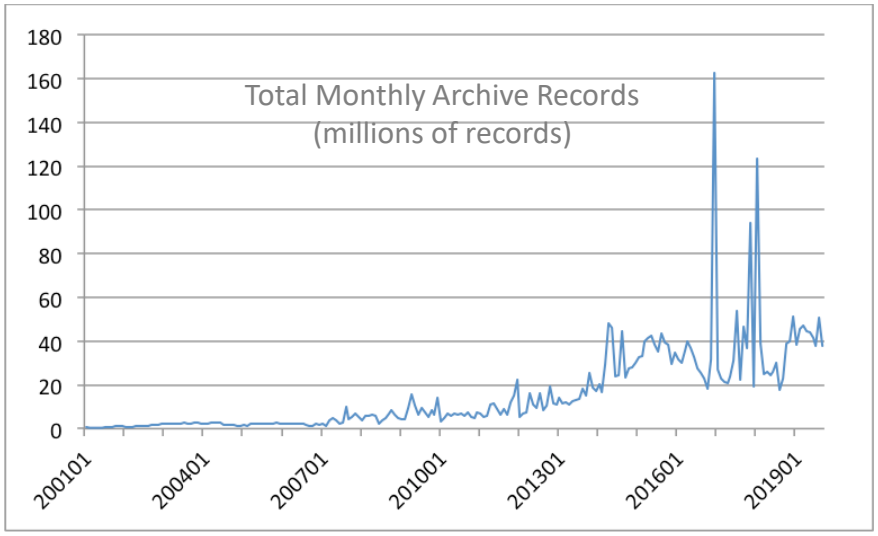


Figure 5. Monthly total number of records processed by EMS since January 2001 (in millions of records)

Figure 6. EMS with Prestaging (prestaging workflow in red). Prestaged results are merged into EMS main processing workflow to achieve final metrics results.

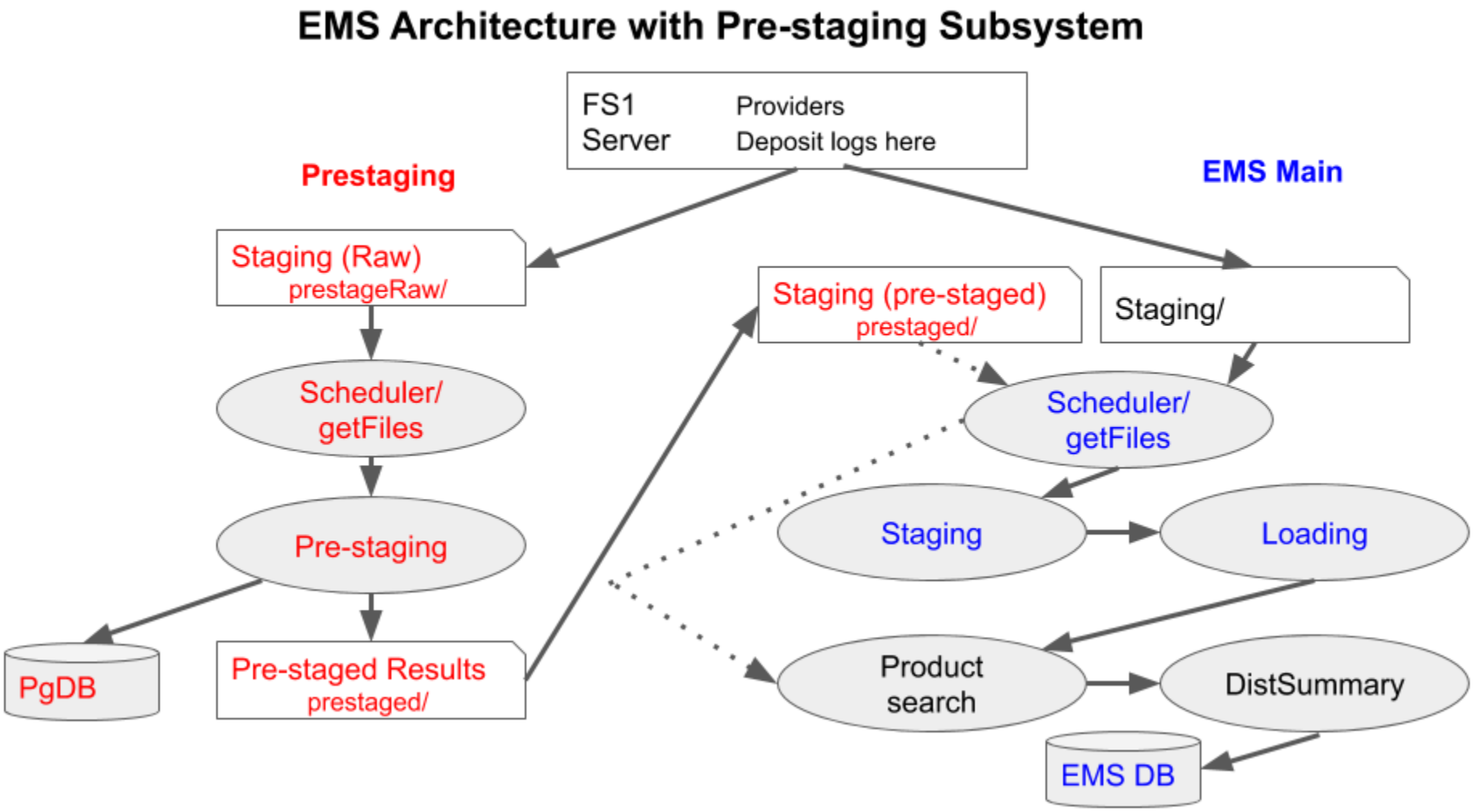
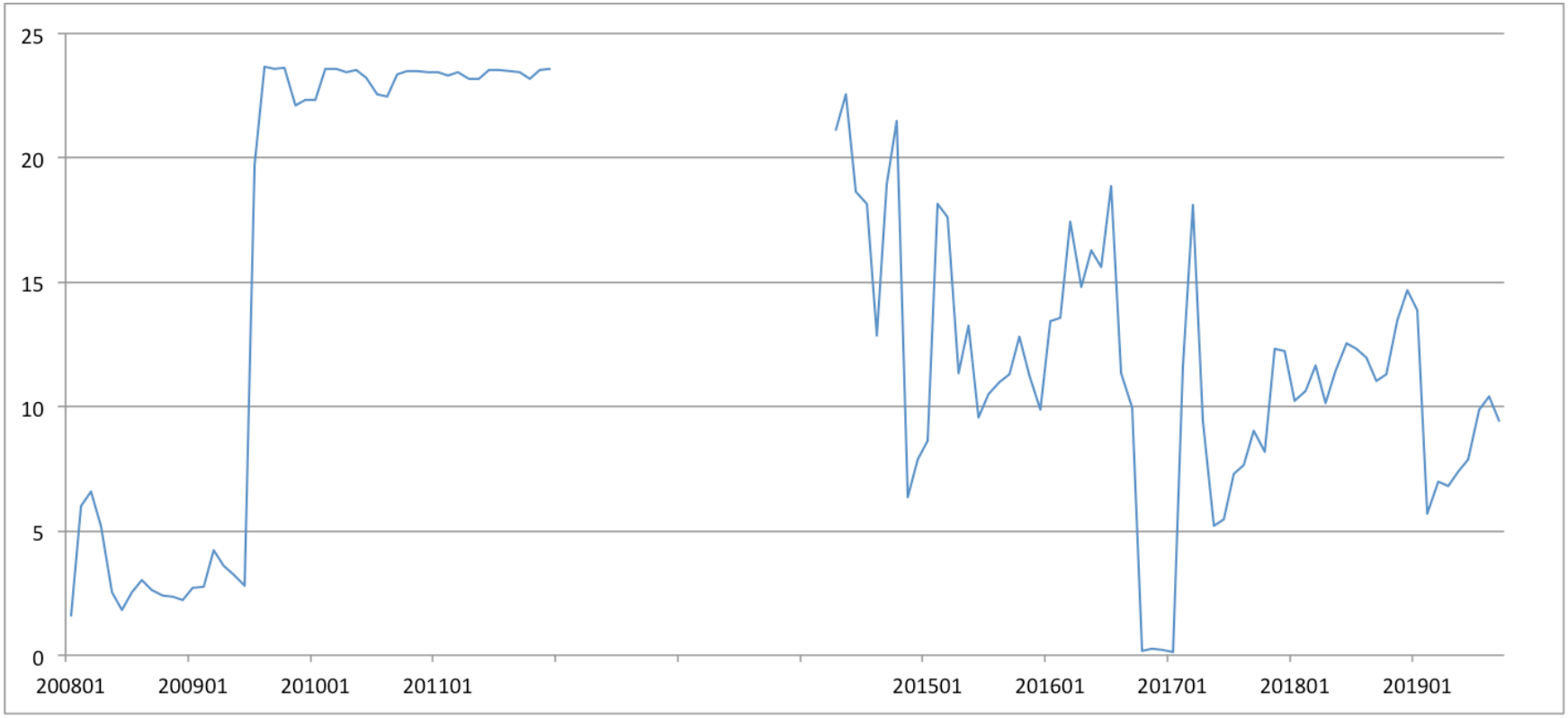


Figure 7. EMS monthly median system daily processing time since January 2008 (hr)
(gap between 2012 and 2014 due to loss of records from system hardware failure)



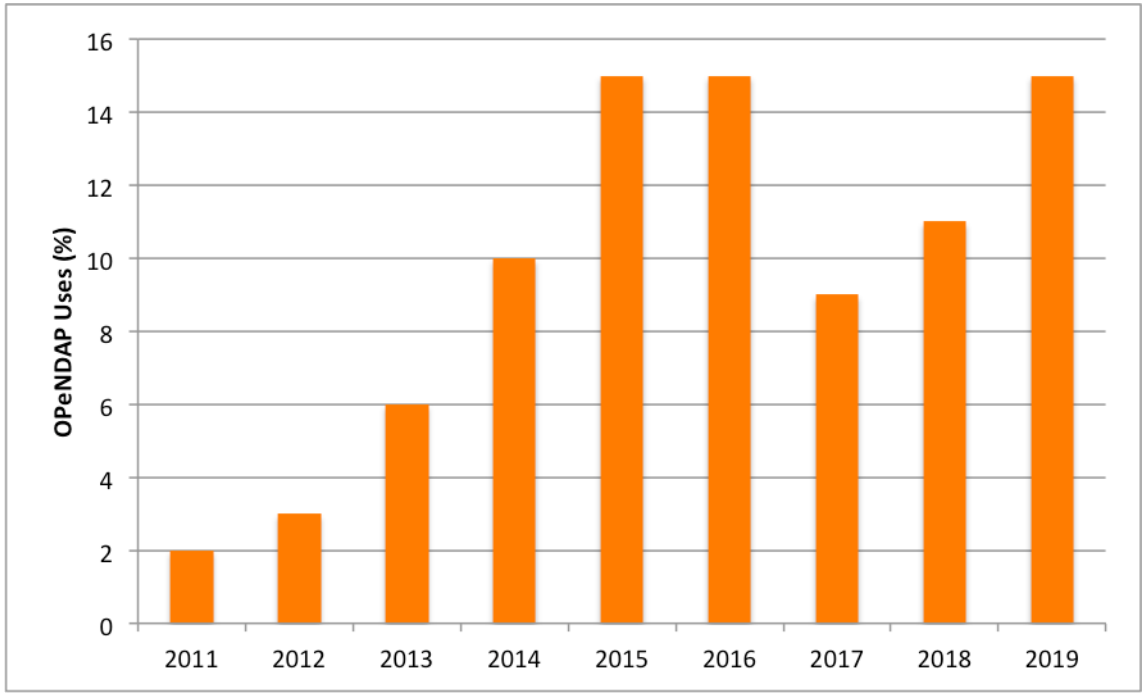
EMS Metrics

EMS started tracking metrics of ingest, archive and distribution of Earth Science data since around 2000. Over the years, DAACs provided various data services making users easier to get to the data, such as OPeNDAP as a data access protocol allowing users to download subset of data, both in geographical region and variables within a data product. EMS incorporated collection of such services protocol in the data provided by the DAACs. Figure 8 shows such OPeNDAP distribution by DAACs to the data users. Table 2 summarizes different categories of metrics that EMS produces.

Table 2. List of typical EMS metrics

Metrics Category	Metrics Item	Description
Data Usage	Volume in bytes	Amount of data distributed to user
	Number of files	Number of files or granules distributed to user
Data Product	Mission	Science mission of the data product
	Instrument	Instrument used to acquire the data measurements
User Profile	Contact email	Occasionally used for notifications from data producers
	Affiliation	User's categorical affiliation such as Government, Education, etc.
	Primary study area	User's primary field of science researches
	Country	Country where user resides
Protocol & Service	FTP / SCP	Traditional file download through FTP, SCP, and alike
	HTTP	Data download through web HTTP protocol
	OPeNDAP	Data download through OPeNDAP protocol
	Giovanni	GES DISC data analysis system that provides analysis results
	Subsetter	Various subset tools mostly for spatial subsetting
	Reformat	Convert data files from one format to another
	Reprojection	Re-project data grids to a different geographic coordinate system

Figure 8. Percentages of OPENDAP data accesses over years among all successful data distributions.



SUMMARY AND CONCLUSIONS

EMS collects and processes records of ingest, archive and distribution logs of Earth Science data products at NASA DAACs. EMS processing system has been evolving over time to meet the needs to handle ever-increasing amount of logs to process. With the implementation of parallelization and preprocessing module, the performance of EMS system has much improved. Moving forward, EMS will continue to grow to meet the needs to process the growing amount of log data, as well as to handle metrics from data distribution through new technologies such as OPeNDAP and cloud-based data distribution. In near future, EMS is looking to interact with NASA's Common Metadata Repository (CMR) for product metadata. As data distribution is moving to cloud, work is also underway to process metrics from the cloud-based data distribution.

ACKNOWLEDGEMENTS

This study was funded through the Software Engineering Support (SES II) GSFC NASA Contractor No. NNG15CR67C-6002