Archive Manager and Processor (AMP).

Rishi Verma[1], Jordan Padams[1], Paul Ramirez[1], Galen Hollins[1], Michael Cayanan[1], [1]NASA Jet Propulsion Laboratory (4800 Oak Grove Drive, La Cañada Flintridge, CA 91011)

Abstract: The Archive Manager and Processor (AMP) is a software framework to generate deep metric and quantitative insight from a file-based archive through the support of parallelized data extraction and processing techniques. Through cloud or on-site cluster infrastructure, AMP is able to both process file-archive data efficiently as well as make the results useful through intuitive search and aggregation of the information. Key use cases of the technology include the ability to holistically form an understanding of the state of a file-based archive, including through the analysis of attributes such as broken links, checksum verifications, naming convention verifications, permission state checks, and other common file-based archive analyses. In addition to file-archive use cases, AMP provides the capability to significantly increase the efficiency of general-purpose processing tasks such as generation of custom label information for data files or the transformation of data files into summary products (e.g. thumbnails, in the case of images), etc. AMP takes advantage of the fact that if highly parallelized hardware is available, immense reductions in the time required to extract or process file-based archive contents can be realized. AMP builds on top of the parallel computing advantages of lambda architectures within cloud infrastructure, such as Amazon Web Services (AWS) Lambda service[1], and the parallelized cluster computing support of frameworks like the Apache Spark toolkit[2]. A key innovation of AMP however, is to abstract the underlying computational infrastructure behind the data extraction and processing layer, enabling a user of the framework to generate insight from a file-based archive with minimal configuration effort. Additionally, AMP seeks to provide intuitive user interfaces to allow a user to visually and quantitatively be informed about the state of an archive or resulting data transformation task.

AMP is being developed by the Planetary Data System (PDS) Cartography and Imaging Sciences Node (IMG), in collaboration with the PDS Engineering Node (EN). It is largely based on the prior development work, at PDS IMG, of the Archive Inventory and Management System (AIMS) toolkit[3]. Through the generalization of the AIMS system, AMP is evolving into a non-PDS IMG specific framework that can be leveraged by other data archive hosting organizations. Moreover, with the recent release of the underlying software code behind AMP into the open source community at-large[4], the continued evolution of the system will be conducted in the public sphere.

AMP is expected to play key roles in several PDS efforts. Namely, AMP is forming a critical part of a PDS pilot feasibility study of leveraging commercial cloud-storage for possible PDS archival or backup means. The AMP framework is also expected to form the future backbone of the PDS EN Harvest-Search software tool[5], which is currently undergoing major redevelopment efforts and is collaborating with the AMP team to ensure support for Harvest use cases are taken into consideration in the development of AMP. More specifically, AMP is targeting the integration of existing Harvest-Search software capabilities (i.e. Harvest Tool, Registry Tool and the Search Tool) and supporting a plugin-based architecture to expand for and support future Harvest extractors. For example, AMP will target Harvest extraction use cases such as: (a) metadata extraction based on mime-types; (b) metadata extraction from PDS file and dataset metadata; and (c) mining content (e.g. imagery) to produce derived metadata (e.g. image annotations). Moreover, AMP will allow extractors to be developed in a variety of programming languages (e.g. Python, Java, C, etc.) by making available a standard interface for extracting metadata and providing an interoperable Java Script Object Notation (JSON)[8] based output. In addition to the PDS EN Harvest-Search use cases, AMP is expected to be a key player for maintaining the integrity of the over one-petabyte PDS IMG archive as well. This includes monitoring the file-archive integrity of PDS IMG related data holdings at the United States Geological Survey (USGS) and high-resolution data sets such as the High Resolution Imaging Science Experiment (HiRISE)[6] and the Lunar Reconnaissance Orbiter Camera (LROC)[7] at the University of Arizona and Arizona State University respectively.

AMP seeks to build on community contributions to its architecture and plugin system to garner reuse across PDS nodes. As such AMP has been released as a pilot open source project of the NASA PDS Incubator Organization[9,4] program; which seeks to support an open source community-based development model with a governance paradigm akin to the Apache Software Foundation[10,11]. Leveraging this paradigm will ensure AMP is available for utilization and open for contributions from all PDS nodes, software development teams peripheral to PDS, and the open source development community at large.

References:

[1] AWS Lambda, Amazon Inc., aws.amazon.com/lambda/. [2] Apache Spark, Apache Software Foundation., http://spark.apache.org/. [3] Verma, R. V. "Archive Inventory Management System (AIMS)---A Fast, Metrics Gathering Framework for Validating and Gaining Insight from Large File-Based Data Archives." Planetary Science Informatics and Data Analytics Conference. Vol. 2082. 2018.. [4] Archive Manager and Processor, GitHub Inc., https://github.com/archive-manager-and-processor. [5] Harvest-Search Tool, Planetary Data System Engineering Node, NASA, https://pds-engineering.jpl.nasa.gov/development/pds4/9.0.0/ingest/harvest-search/index.html [6] HiRISE PDS Catalog Archive, Planetary Data System, NASA, https://hirise-pds.lpl.arizona.edu/PDS/ [7] LROC RDR Product Select, Arizona State University, NASA, http://wms.lroc.asu.edu/lroc/rdr_product_select [8] JavaScript Object Notation, https://www.json [9] https://nasa-pds-incubator.github.io/. [10] Apache Software Foundation, https://www.apache.org/ . [11] Apache Software Foundation Governance, https://apache.org/foundation/governance/ .