Archive Inventory Management System (AIMS) – a fast, metrics gathering framework for validating and gaining insight from large file-based data archives. Rishi Verma[1], [1]Jet Propulsion Laboratory (4800 Oak Grove Drive, Pasadena CA 91108).

Abstract: The Archive Inventory Management System (AIMS) is a software package for understanding the distribution and characteristics of files and directories in large file-based data archives. It provides an efficient crawling and customizable extraction system for scanning over very large file directory trees and extracting specific information from assets within those directory trees. This information is then indexed in a search engine to provide robust analysis and visualization support in order to better understand characteristics of the archive.

The motivation for AIMS stems from a need within NASA's Planteary Data System (PDS) Imaging Node (IMG) [1] to *continually* keep up-to-date about file-based data archive characteristics, such as: adherence to file path naming conventions, file and directory sizes, checksums, validation of appropriate file placement within directories, etc. Given PDS IMG stores at least 63 TB of data and upwards of approximately 650 million file assets, keeping up-to-date with respect to these assets on a continual basis is a significant computational challenge. Existing approaches of sequentially processing each file asset can take weeks to finish, given the current PDS IMG archive size. To meet this computational challenge, as well as future ones, the AIMS system is being designed to take advantage of cluster computing, using the Apache Spark framework [2], and clusterized search tools to aggregate metrics metadata using the Elastic Search framework [3]. Taking a cluster-computing approach has the advantage of horizontally scaling and distributing the processing work of evaluating each file data system asset across a potentially increasing number of worker nodes. Thus, hardware purchases, not software design, is the limiting factor in reducing overall archive scanning and metrics extraction time.

Extracting and collecting metrics data is the first computational challenge AIMS seeks to address. Following this, AIMS also supports ad-hoc queries of collected metrics data for analysis purposes. For example, AIMS typically collects the directory size information of every directory within the PDS IMG archive; however, it also supports actions like summing up the total size of all the directories matching a particular directory path. To do this, the Elastic Search framework is utilized to provide ad-hoc, text-based and numeric aggregation support. In other words, an unforeseen query to the AIMS search engine results in a distributed computational search job where existing metrics are aggregated using the assistance of multiple cluster nodes to produce the final result. Thus, AIMS is flexible enough to provide very robust and efficient analytics capability that data scientists need to understand the characteristics of a file based archive. Together, with an efficient, distributed metrics extraction framework and a distributed, ad-hoc query capability, the AIMS system provides a fast and effective way to keep up-to-date with changing characteristics of file-based data archives.

References:

[1] "Cartography and Imaging Sciences Node" National Aeronautics and Space Adminstration. Last accessed: Nov 14th, 2017. <https://pds-imaging.jpl.nasa.gov/>.

[2] "Apache Spark" Apache Software Foundation. Last accessed: Nov 14th, 2017. <http://spark.apache.org//>.

[3] "Elasticsearch" Elasticsearch. Last accessed: Nov 14th, 2017. <https://www.elastic.co/products/elasticsearch/>.