Faced with unprecedented growth in Earth data volume and demand, NASA has developed the Earth Data Analytic Services (EDAS) framework, a high-performance big data analytics framework built on Apache Spark. This framework enables scientists to execute data processing workflows combining common analysis operations close to the massive data stores at NASA. The data is accessed in standard (NetCDF, HDF, etc.) formats in a POSIX file system and processed using vetted Earth data analysis tools (ESMF, CDAT, NCO, etc.). EDAS utilizes a dynamic caching architecture, a custom distributed array framework, and a streaming parallel in-memory workflow for efficiently processing huge datasets within limited memory spaces with interactive response times.

EDAS services are accessed via a WPS API being developed in collaboration with the ESGF Compute Working Team to support server-side analytics for ESGF. New analytic operations can be developed in Python, Java, or Scala (with support for other languages planned). Client packages in Python, Java/Scala, or JavaScript contain everything needed to build, submit, manage, and visualize big data analysis workflows from the user’s desktop computer or to develop web applications with embedded analytics.

The EDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets, where the data resides. It is currently deployed at NASA and available for public use. Another NASA EDAS deployment supports the Collaborative REAnalysis Technical Environment (CREATE) project, which centralizes numerous global reanalysis datasets onto a single analytics platform. These services enable scientists and decision makers to access remote model/reanalysis data archives and investigate trends, variability, anomalies, and other features of local and global earth system dynamics.

**EDAS: Analysis-as-a-Service Infrastructure (AaaS)**

- **Execution Manager**: Spark, Internal Kernels (Java, Scala), Workflow Engine
- **Web App**: JavaScript API
- **Application API**: Python API, Notebook Python Script
- **ToolBase**: ESMF, NCL, CDO, UVCDAT
- **EDAS Canonical Operators**
  - **Reduce Operations**: Max, Min, Sum, Average, RootMeanSquare
  - **Combine (Ensemble) Operations**: Max, Min, Sum, Average, Difference, Multiply, Divide
  - **Common Workflows**: Anomaly (Ave+Diff), StdDev (Ave+Diff+ RMS), etc.
  - **Utility Operations**: Subset, Regrid, Filter
- **Current Operator List**: https://edas.nccs.nasa.gov/wps/cwt?request=GetCapabilities

**Why is this approach distinctive?**

- Direct access to NetCDF data archives via disk or OpenDAP: Alleviates the need to maintain additional copies of the data.
- Deploys existing (Python) climate data analysis tools: Utilizes UVCDAT, ESMF, and other Python analytic toolkits.
- No changes to existing applications: Parallels the data, not the applications.
- High-performance analytics: Optimizes decomposition over processors for the current task.
- Streaming in-memory parallel workflows using Apache Spark: Parallel GPFS data access (faster than HDFS on our cluster).
- 15 to 50 times faster than standard tools in our environment.
- Modular structure: Easily add new technologies and compare approaches.

**EDAS Deployment**

- **NASA EDAS deployment on the DASS**:
  - Data Analytics and Storage System (DASS): https://www.nccs.nasa.gov/services/dass
  - Web portal (WPS server): https://edas.nccs.nasa.gov/wps/cwt
    - Available data collections:
      - https://edas.nccs.nasa.gov/opensearch/Geospatial?useWPS=true&extid=coll
  - **EDAS Distribution**:
    - Server: https://github.com/nasa-nccs-cds/EDAS.git
    - Web app: https://github.com/nasa-nccs-cds/EDWPS.git
    - Client: https://github.com/ESGF/esgf-compute-api.git
  - **Documentation**:
    - https://www.nccs.nasa.gov/services/Analytics

**Poster IN21D-0064:** CREATE-IP and CREATE-V: Data and Services Update.