IN13B-0714 – Parmap: Analytics Engine Scalable for Climate Model Evaluation on Cloud and High-Performance Computing Platforms

Joseph C. Jacob, Brian D. Wilson and Huikyo Lee
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109-8099

Abstract

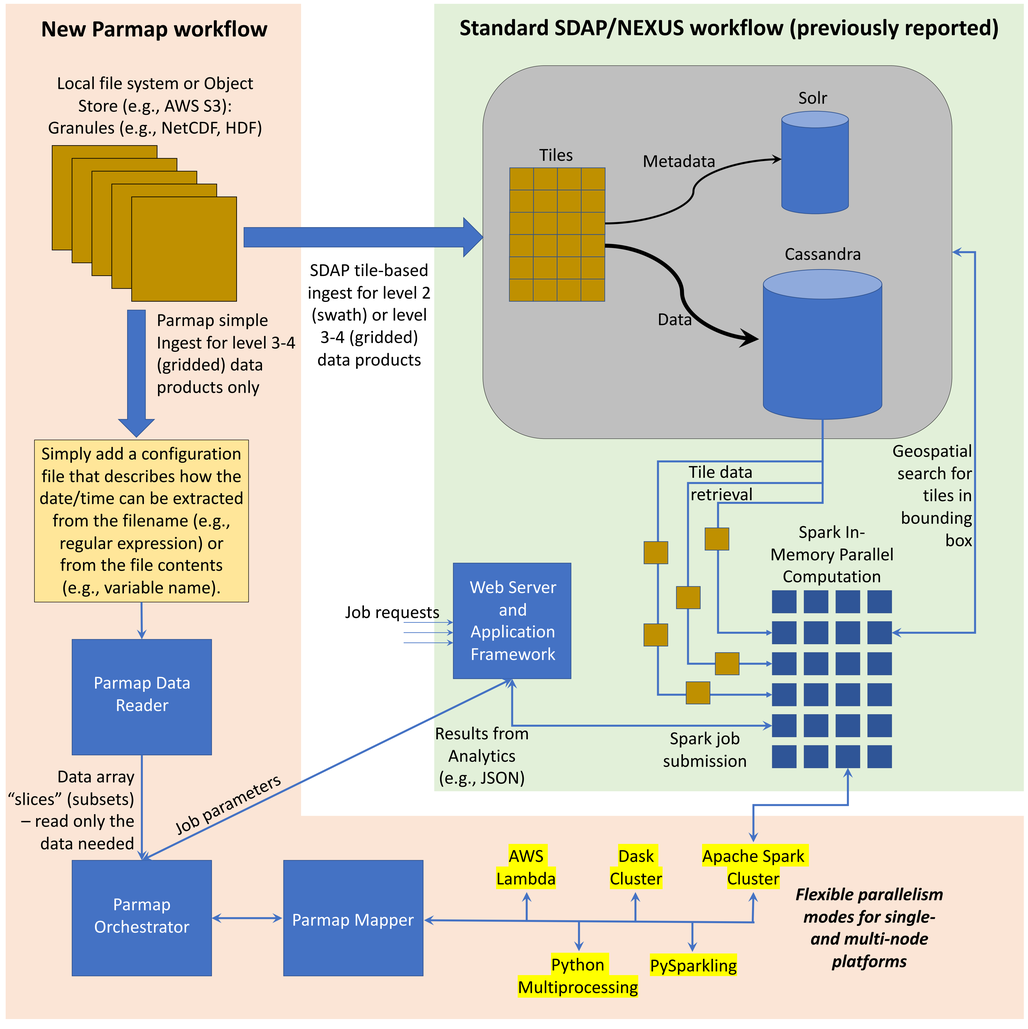
The need to better understand climate change has driven model simulations to greater fidelity with improved spatiotemporal resolution (e.g., < 10 km at sub-hourly cadence). The rise of these high-fidelity climate models coincides with the emergence of cloud computing as a viable platform for scientific analytics. However, it is not cost- or time-effective to move the High-Performance Computing (HPC)-based model computations and data to the cloud. Thus, there is a need for scalable model evaluation compatible with both the cloud and HPC platforms like NASA Center for Climate Simulation (NCCS). To fill this need we have extended the analytics component of the Apache Science Data Analytics Platform (SDAP) with a streamlined adaptation that specifically targets high-resolution science data products and climate model outputs on a regular coordinate grid. Gridded inputs (as opposed to other data structures like point clouds or swath-based measurements supported by SDAP), enable offsets to particular grid cells to be directly computed, allow for processing on the original NetCDF or HDF granules, do not require a second tiled copy of the data, and accommodate a simpler technology stack since no geospatial database is required for lookups or tile storage. Our core module, Parmap, abstracts the map-reduce model so that users can select from a variety of map computational modes, including Spark, Dask, serverless AWS Lambda, PySparkling, and Python multiprocessing.

Parmap Key Features

- Enables ***simple python coding*** for computing analytics in parallel
- Extends SDAP with a ***streamlined workflow*** optimized for data products on a regular coordinate grid.
- Operate on the original data granules so ***no data duplication*** is needed and no databases are required.
- Nearly ***no data ingest*** mechanism needed. Simply prepare a configuration file for a dataset to indicate how the date or timestamp can be extracted from the data granules or their filenames and the dataset is ready for Parmap analytics.
- ***No persistent services*** (like databases) required means it is easy to deploy Parmap to shared HPC environments as well as public and private clouds.
- ***Simple to deploy*** map and reduce operations over files or S3 URLs in multiple modes without code changes:
  - ✓ Multicore (single node)
  - ✓ PySparkling (pure Python, single node)
  - ✓ Full Apache PySpark Cluster (multi-node)
  - ✓ Dask Cluster (multi-node)
  - ✓ AWS Lambda (serverless)
- Fully compatible with the existing SDAP architecture. Use full SDAP for selected datasets that have been ingested and Parmap as default for regularly gridded data.
- SDAP/Parmap architecture allows access via four interfaces:
  - ✓ Web service call
  - ✓ Python client library
  - ✓ Jupyter Notebook
  - ✓ Web GUI with map displays, built-in analytics

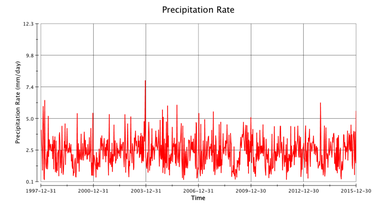
Acknowledgements

Parmap is being developed under NASA's Computational Modeling Algorithms and Cyberinfrastructure (CMAC) Program. SDAP was initially sponsored by NASA's Advanced Information Systems Technology (AIST) Program under the OceanWorks project (PI: Thomas Huang, JPL).

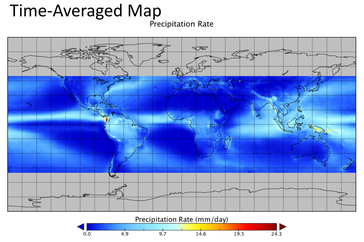


SDAP analytics currently available in Parmap:

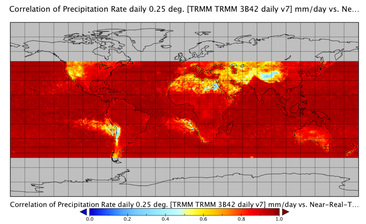
Area-Averaged Time Series



Time-Averaged Map



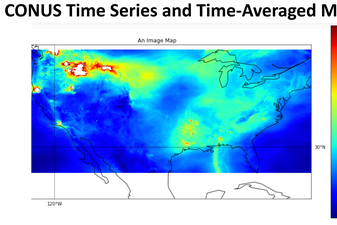
Correlation Map

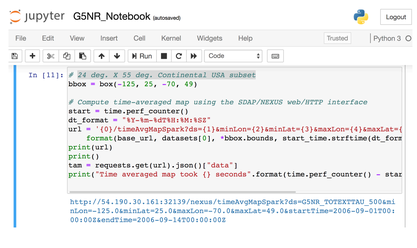


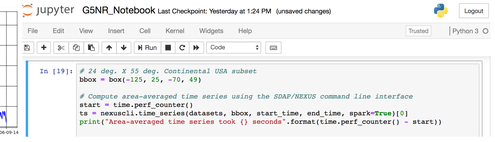
More to come:

- Climatological Map
- Latitude-Time Hovmoller Map
- Longitude-Time Hovmoller Map
- Daily Difference Average

Notebook Analytics: CONUS Time Series and Time-Averaged Map







Parmap Analytics Benchmarks from GEOS5 Nature Run (G5NR):

- Platform: 8-node Spark cluster
- Dataset: GEOS5 Nature Run (G5NR)
  - June 2005 – June 2007
  - 7 km spatial resolution
  - 30 minute model outputs
  - ~4 PB (>40 variables)
- Input data used for these benchmarks:
  - 30 days of G5NR: 9/1/2006 – 9/30/2006
  - 1 variable for area-averaged time series and time averaged map: *total aerosol extinction AOT [550 nm]*
  - 2nd variable for correlation map: *dust surface mass concentration [pm 2.5]*
  - Various spatial extents:
    - City – 1x1 deg Boulder, Colorado, USA
    - State – 4x7 deg Colorado, USA
    - Country – 25x58 deg Continental United States (CONUS)
    - Global
- Plan to be horizontally scalable for full 2-year simulation