Taking on Big Ocean Data Science

Thomas Huang, California Institute of Technology/Jet Propulsion Laboratory (Pasadena, CA, USA),
thomas.huang@jpl.nasa.gov

Almost all of the existing data analysis solutions for Earth science are still built around large archives of granule files, which yield poor performance. Common data access solutions, such as OPeNDAP and THREDDS, provide web service interfaces to archives of observational data. They perform poorly because they are still built around the notion of files. It is very time consuming for climate scientists to conduct research that involves the generation of time series over large spatial region and/or over decades of observational data.

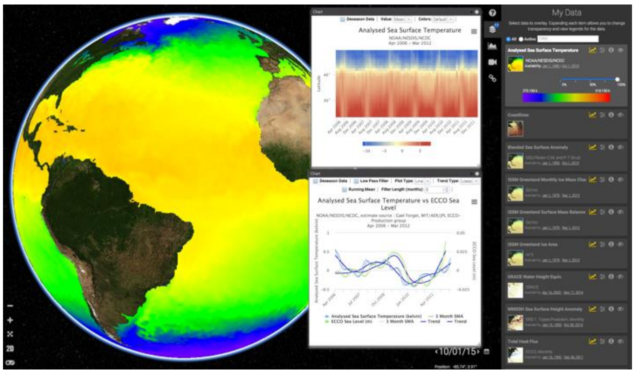
Analysed Sea Surface Temperature
NOAA/NESDIS/NCDC
Apr 2006 - Mar 2012
Latitude
Analysed Sea Surface Temperature vs ECCO Sea Level
NOAA/NESDIS/NCDC, estimate source : Gael Forget, MIT/AER/JPL ECCO-Production group
Apr 2006 - Mar 2012
Analysed Sea Surface Temperature (kelvin)
ECCO Sea Level (m)
3 Month SMA
Trend
My Data
Analysed Sea Surface Temperature
Coastlines
Blended Sea Surface Anomaly
ISSM Greenland Monthly Ice Mass Cha
ISSM Greenland Surface Mass Balance
ISSM Greenland Ice Area
GRACE Water Height Equiv.
MMSSH Sea Surface Height Anomaly
Total Heat Flux
‹10/01/15›

Fig. 1: NASA Sea Level Change Portal’s Data Analysis Tool

Map Reduce is a well-known paradigm for processing large amounts of data in parallel using clustering or Cloud environments. Unfortunately, this paradigm doesn't work well with temporal, geospatial array-based data. One major issue is they are packaged in files in various sizes. These data files can range from tens of megabytes to several gigabytes. Depending on the user input, some analysis operations could involve hundreds to thousands of these files. Moving these files from storage node to computing node is very inefficient where the majority of the time is spent on I/Os.

NEXUS is a Deep Data Platform being developed at NASA’s Jet Propulsion Laboratory. It takes on a different approach in handling file-based observational temporal, geospatial artifacts by fully leveraging the elasticity of Cloud Computing environment. Rather than performing on-the-fly file I/Os, NEXUS breaks data artifacts into small data tiles where they are managed by a Cloud-scaled database with high-performance spatial lookup service. NEXUS provides the bridge between science data and horizontal-scaling data analysis. It provides a workflow to divide science artifacts into small data tiles, store in a cloud-scaled database where they can be quickly retrieved through a high performance spatial search registry. This platform simplifies development of big data analysis solutions by bridging the gap between files and Map Reduce solutions such as Hadoop and Spark.

This presentation discusses applications of NEXUS in three Big Ocean Data Science projects at NASA in relation to the definition of Big Data, which evolves around the 3Vs model, Volume, Velocity, and Variety.

- The NASA Sea Level Change Portal (https://sealevel.nasa.gov) is the official NASA’s one-stop information portal for all news and data relevant to sea level rise. The portal has a built-in Data Analysis Tool (DAT) to provide climate scientist high-performance visualization and a suite of on-the-fly analysis tools.
- The NASA OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analysis Portal is a cloud-based analytic service that enables execution of domain-specific, multi-scale anomaly and feature detection algorithms across the entire archive of ocean science datasets.
- The NASA’s Distributed Oceanographic Match-Up Service (DOMS) is a cloud-based reconciliation of satellite and in-situ datasets