SDAPProposal

Abstract

The Science Data Analytics Platform (SDAP) establishes an integrated data analytic center for Big Science problems. It focuses on technology integration, advancement and maturity.

Proposal

SDAP currently represents a collaboration between NASA Jet Propulsion Laboratory (JPL), Florida State University (FSU), the National Center for Atmospheric Research (NCAR), and George Mason University (GMU). SDAP brings together a number of big data technologies including a NASA funded OceanXtremes (Anomaly detection and ocean science), NEXUS (Deep data analytic platform), DOMS (Distributed in-situ to satellite matchup), MUDROD (Search relevancy and discovery) and VQSS (Virtualized Quality Screening Service) under a single umbrella. Within the original Incubator proposal, VQSS will not be included however it is anticipated that a future source code donation will cover VQSS.

Background and Rationale

SDAP is a technology software solution currently geared to better enable scientists involved in advancing the study of the Earth's physical oceanography. With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, the impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and stronger/more frequent hurricanes, sea level rise and storm surges affecting coastlines, and may involve drastic changes and shifts in marine ecosystems. Ocean science communities are relying on data distributed through data centers such as the JPL's Physical Oceanographic Data Active Archive Center (PO.DAAC) to conduct their research. In typical investigations, oceanographers follow a traditional workflow for using datasets: search, evaluate, download, and apply tools and algorithms to look for trends. While this workflow has been working very well historically for the oceanographic community, it cannot scale if the research involves massive amount of data. NASA's Surface Water and Ocean Topography (SWOT) mission, scheduled to launch in April of 2021, is expected to generate over 20PB data for a nominal 3-year mission. This will challenge all existing NASA Earth Science data archival/distribution paradigms. It will no longer be feasible for Earth scientists to download and analyze such volumes of data. SDAP was therefore developed primarily as a Web-service platform for big ocean data science at the PO.DAAC with open source solutions used to enable fast analysis of oceanographic data. SDAP has been developed collaboratively between JPL, FSU, NCAR, and GMU and is rapidly maturing to become the generic platform for the next generation of big science data solutions. The platform is an orchestration of several previously funded NASA big ocean data solutions using cloud technology, which include data analysis (NEXUS), anomaly detection (OceanXtremes), matchup (DOMS), subsetting, discovery (MUDROD), and visualization (VQSS). SDAP will enable web-accessible, fast data analysis directly on huge scientific data archives to minimize data movement and provide access, including subset, only to the relevant data. In essence, the above information workflow can be visualized by the image below where a transformation of data to knowledge occurs as one moves from left to right.

? Unknown Attachment

Science Data Analytics Platform Project Overview

SDAP consists of several loosely coupled, independently functioning sub-projects. The graphic below displays an overview of how these sub-projects fuse together. N.B., although the graphic uses terminology relating to OceanWorks, essentially the SDAP architecture is identical.

? Unknown Attachment

OceanXtremes

Oceanographic Data-Intensive Anomaly Detection and Analysis Portal. An application that allows you to view imagery and perform analysis on sea level rise data.

Objective Develop an anomaly detection system which identifies items, events or observations which do not conform to an expected pattern.

- Mature and test domain-specific, multi-scale anomaly and feature detection algorithms.
- Identify unexpected correlations between key measured variables.

Demonstrate value of technologies in this service:

- Adapted Map-Reduce data mining.
- Algorithm profiling service.
- Shared discovery and exploration search tools.
- Automatic notification of events of interest.

NEXUS

NEXUS is an emerging technology developed at JPL

- A Cloud-based/Cluster-based data platform that performs scalable handling of observational parameters analysis designed to scale horizontally

- Leveraging high-performance indexed, temporal, and geospatial search solution
- Breaks data products into small chunks and stores them in a Cloud-based data store

Data Volumes Exploding

- SWOT mission is coming
- File I/O is slow

Scalable Store & Compute is Available

- NoSQL cluster databases
- Parallel compute, in-memory map-reduce
- Bring Compute to Highly-Accessible Data (using Hybrid Cloud)

Pre-Chunk and Summarize Key Variables

- Easy statistics instantly (milliseconds)
- Harder statistics on-demand (in seconds)
- Visualize original data (layers) on a map quickly

DOMS

The Distributed Oceanographic Match-Up Service DOMS is designed to reconcile satellite and in situ datasets in support of NASA's Earth Science mission. The service will provide a mechanism for users to input a series of geospatial references for satellite observations and receive the in situ observations that are matched to the satellite data within a selectable temporal and spatial domain. DOMS includes several characteristic in situ and satellite observation datasets - with an initial focus on salinity, sea temperature, and winds. DOMS will be used by the marine and satellite research communities to support a range of activities and several use cases will be described. The service is designed to provide a community-accessible tool that dynamically delivers matched data and allows the scientist to only work with the subset of data where the matches exist.

MUDROD

Mining and Utilizing Dataset Relevancy from Oceanographic Datasets to Improve Data Discovery and Access Data discovery accuracy is a challenging topic for both Earth science and other domains. It is especially true for scientific data sets that are not as popular as Amazon or Google data. MUDROD is focused on mining oceanic knowledge from the PO.DAAC user log files to improve the end user data discovery experience at PO.DAAC. There are three steps in the research: a) the oceanographic semantics were extracted from three resources of SWEET, GCMD ontology, and the keywords used by end users for searching PO.DAAC datasets, b) mining the linkage among different vocabularies based on user data discvoery sessions, and c) build the linkage among vocabularies based on a comprehensive approach by considering domain de facto standard, e.g., SWEET and GCMD, and the knowledge mined from the log files. The semantics is used to improve data discovery for ranking results, navigating among vocabularies, and recommending data based on user searchers.

Current Status

All components of SDAP were originally designed and developed under grants from the NASA-funded Advanced Information Systems and Technologies (AIST) program. The initiative to bring them the components together under the SDAP umbrella was granted through an AIST-funded follow-on grant which will run for another ~18 or so months. Currently no projects have made official releases so outside of community building, this will be our primary Incubating goal. All SDAP source code is currently publicly available and licensed under the ALv2.0.

Meritocracy

The current developers are familiar with meritocratic open source development at Apache. The SDAP team consumes Apache products heavily with members being part of several Apache user communities. SDAP itself has critical dependencies upon Apache products. Lewis McGibbney (JPL employee), a Member of the ASF and V.P. of Apache Any23, Gora PMC Nutch, Tika, OODT, OCW, etc., is championing the effort to bring SDAP into and through the Apache Incubator and has been evangelizing the Apache Way to the current SDAP contributors such that the meritocratic process is well understood and followed. Apache was chosen specifically because we want to encourage this style of community development for the project and for it to sustain SDAP forward to become the generic platform for the next generation of big science data solutions

Community

The SDAP project is a fairly new effort and our community is not yet fully/firmly established. Initial committers comprising the SDAP roster have only recently fully come together as a unified team however there is a large degree of synergy between constituent members at JPL, FSU, NCAR, and GMU. Therefore, community building and publicity continues to be a major thrust. With the activity and exposure regularly attained by several community members, we hope to grow the SDAP presence in and across several (scientific) forums. The SDAP technology is generating interest within communities such as the Earth Science Information Partnership (ESIP), American Geophysical Union (AGU) and plethora or science meetings around the globe. This in effect, we hope, will further contribute towards the possibility of SDAP being used across Government Agencies such as NASA, NOAA, USGS, EPA, DOI, etc. as well as by researchers and students in academic institutions around the globe. During incubation, we will explicitly seek to increase our adoption, with SDAP already being featured on the agenda for several high profile globally significant scientific conferences and meetings.

Core Developers

The current set of core developers is relatively small, including full-time and students from across JPL, FSU, NCAR, and GMU. Initial community management and participation will be distributed across the entire team, most of which have been involved with the constituent projects for <2 years.

Alignment

All SDAP code is licensed under Apache v2.0.

Known Risks

Orphaned products

There are currently no orphaned products. Each component of SDAP has dedicated personnel leading and participating in its ongoing development. Additionally, there is substantial collaboration between projects facilitated by regular project meetings which are specific the the initial member entities and focused on advancing physical oceanographic science.

Inexperience with Open Source

JPL (in particular Lewis McGibbney) has been part of several efforts to transition to and grow projects communities at Apache e.g. Apache OODT, Apache Open Climate Workbench, Apache Joshua (Incubating), Apache SensSoft (Incubating), Apache DRAT (Incubating). Most of the code developed under the SDAP umbrella was and is open source prior to the Incubator effort so we are well familiarized with the nuances of open source software.

Relationships with Other Apache Products

SDAP has strong dependency upon a number of high profile and smaller profile Apache products. Examples can be seen in the breakdown of External Dependencies. As we continue to grow SDAP within the Incubator, we will make efforts to share community stories, software advancements and possible improvements in our use of our Apache dependencies back to those project communities.

Developers

The SDAP project and hence developers is currently funded through a NASA AIST follow-on grant with funding secured for the next ~18 months. There are currently no 100% time dedicated developers, however, the same core team that does work currently will continue to work on the project throughout the next current funding period and after. There is currently no business strategy aligned with SDAP however it is perceived that future, yet unsecured funding may by directed to further feature advancement and project evangelism.

Documentation

Documentation is currently available in a number of locations e.g. Github wiki, Github pages, etc. with each repository under the oceanworks-aist Github Org maintaining documentation available through wiki's attached to the repositories. Additionally, most of the SDAP sub-projects have been extensively documented within plethora of formal academic publications across several academic communities. It would be our intention, certainly atleast to unify the Github wiki ad Github pages documentation most likely to make up the sdap.apache.org Website content.

Initial Source

Current source resides in several locations Github:

- https://github.com/dataplumber/nexus (NEXUS, OceanXtremes, DOMS)
- https://github.com/dataplumber/edge (EDGE)
- https://github.com/aist-oceanworks/mudrod (MUDROD)
- https://bitbucket.org/coaps_mdc/doms/src (DOMS)

External Dependencies

Each component of the Science Data Analytics Platform has its own dependencies. Documentation will be available for integrating them.

MUDROD

Core com.google.code.gson gson 2.5 compile
jar false org.jdom jdom 2.0.2 compile
jar false org.elasticsearch elasticsearch 5.2.0 compile
jar false org.elasticsearch elasticsearch-spark-20_2.11 5.2.0 compile
jar false joda-time joda-time 2.9.4 compile
jar false com.carrotsearch hppc 0.7.1 compile
jar false org.apache.spark spark-core_2.11 2.1.0 compile
jar false org.apache.spark spark-sql_2.11 2.1.0 compile
jar false org.apache.spark spark-mllib_2.11 2.1.0 compile
jar false org.scala-lang scala-library 2.11.8 compile
jar false org.codehaus.jettison jettison 1.3.8 compile
jar false commons-cli commons-cli 1.2 compile
jar false net.sf.opencsv opencsv 2.3 compile
jar false org.apache.jena jena-core 3.3.0 compile
jar false junit junit 4.12 test
jar false

Service gov.nasa.jpl.mudrod mudrod-core 0.0.1-SNAPSHOT compile
jar false javax.servlet javax.servlet-api 3.1.0 provided
jar false com.google.code.gson gson 2.5 compile
jar false

Web

- AngularJS - MIT License
- BootstrapJS - MIT License
- jQueryJS - MIT License
- Underscore JS - MIT License

DOMS

- Apache Solr version 5.5.1http://lucene.apache.org/solr/
- EDGE https://github.com/dataplumber/edge
- NetCDF4 http://unidata.github.io/netcdf4-python/
- Python 3.5 (NOTE: only partial support for py2.7)

Non stdlib Python dependencies:

- Jinja2==2.9.5
- python-dateutil==2.6.0
- cython==0.25.2
- numpy==1.12.0
- scipy==0.18.1
- netCDF4==1.2.7
- solrpy3
- siphon==0.4.0
- neo4j-driver==1.1.0
- matplotlib==2.0.0
- requests==2.13.0
- shapely==1.5.17
- flask==0.12
- networkx==1.11
- pyproj==1.9.5.1
- blist==1.3.6

NEXUS

Analysis

- https://github.com/dataplumber/nexus/blob/master/analysis/package-list.txt
- https://github.com/dataplumber/nexus/blob/master/analysis/requirements.txt

Client

- https://github.com/dataplumber/nexus/blob/master/client/requirements.txt

Climatology

- matplotlib
- numpy
- netCDF4
- pathos (https://pypi.python.org/pypi/pathos)

Data-access

- https://github.com/dataplumber/nexus/blob/master/data-access/requirements.txt

Nexus-ingest *Dataset-tiler*

- https://github.com/dataplumber/nexus/tree/master/nexus-ingest/dataset-tiler/build/reports

developer-box

- Just a collection of scripts/vagrant file used to stand up a developer instance of nexus ingestion. No dependencies to report

Groovy-scripts

- Collection of Groovy scripts that can be used as part of data ingestion. They only rely on the standard Groovy library and the ‘nexus-messages’ project

Nexus-messages

- https://github.com/dataplumber/nexus/tree/master/nexus-ingest/nexus-messages/build/reports

nexus-sink

- https://github.com/dataplumber/nexus/tree/master/nexus-ingest/nexus-sink/build/reports

nexus-xd-python-modules

- https://github.com/dataplumber/nexus/blob/master/nexus-ingest/nexus-xd-python-modules/package-list.txt
- https://github.com/dataplumber/nexus/blob/master/nexus-ingest/nexus-xd-python-modules/requirements.txt

spring-xd-python

- only python standard libraries are used

tcp-shell

- https://github.com/dataplumber/nexus/tree/master/nexus-ingest/tcp-shell/build/reports

tools/deletebyquery

- https://github.com/dataplumber/nexus/blob/master/tools/deletebyquery/requirements.txt

Required Resources

Mailing Lists

- private@sdap.incubator.apache.org
- dev@sdap.incubator.apache.org
- commits@sdap.incubator.apache.org

Git Repos

- https://git-wip-us.apache.org/repos/asf/incubator-sdap-nexus.git
- https://git-wip-us.apache.org/repos/asf/incubator-sdap-doms.git
- https://git-wip-us.apache.org/repos/asf/incubator-sdap-mudrod.git
- https://git-wip-us.apache.org/repos/asf/incubator-sdap-website.git

Issue Tracking

- JIRA Science Data Analytics Platform (SDAP)

Continuous Integration

- Jenkins builds on https://builds.apache.org/

Web

- http://sdap.incubator.apache.org/
- wiki at http://cwiki.apache.org

Initial Committers

The following is a list of the planned initial Apache committers (the active subset of the committers for the current repository on Github).

- Lewis John McGibbney (lewismc@apache.org)
- Vardis M. Tsontos (vardis.m.tsontos@jpl.nasa.gov)
- Joseph C. Jacob (Joseph.C.Jacob@jpl.nasa.gov)
- Ed Armstrong (edward.m.armstrong@jpl.nasa.gov)
- Frank Greguska (greguska@jpl.nasa.gov)

- Brian Wilson (brian.wilson@jpl.nasa.gov)
- Chaowe Phil Yang (cyang3@gmu.edu)
- Yongyao Jiang (yjiang8@gmu.edu)
- Yun Li (yli38@gmu.edu)
- Shawn R. Smith (smith@coaps.fsu.edu)
- Jocelyn Elya (jelya@coaps.fsu.edu)
- Mark Bourassa (bourassa@coaps.fsu.edu)
- Thomas Cram (tcram@ucar.edu)
- Thomas Huang (thomas.huang@jpl.nasa.gov)
- Steven Worley (worley@ucar.edu)
- Zaihua Ji (zji@ucar.edu)

Affiliations

NASA JPL

- Lewis John McGibbney (lewismc@apache.org)
- Vardis M. Tsontos (vardis.m.tsontos@jpl.nasa.gov)
- Joseph C. Jacob (Joseph.C.Jacob@jpl.nasa.gov)
- Ed Armstrong (edward.m.armstrong@jpl.nasa.gov)
- Frank Greguska (greguska@jpl.nasa.gov)
- Thomas Huang (thomas.huang@jpl.nasa.gov)
- Brian Wilson (brian.wilson@jpl.nasa.gov)

George Mason University

- Chaowe Phil Yang (cyang3@gmu.edu)
- Yongyao Jiang (yjiang8@gmu.edu)
- Yun Li (yli38@gmu.edu)

Center for Ocean-Atmospheric Prediction Studies, Florida State University

- Shawn R. Smith (smith@coaps.fsu.edu)
- Jocelyn Elya (jelya@coaps.fsu.edu)
- Mark Bourassa (bourassa@coaps.fsu.edu)

Computational Information Systems Laboratory (CISL) / National Center for Atmospheric Research (NCAR)

- Thomas Cram (tcram@ucar.edu)
- Zaihua Ji (zji@ucar.edu)
- Steven Worley (worley@ucar.edu)

Sponsors

Champion

- Lewis McGibbney (NASA/JPL)

Nominated Mentors

- Jörn Rottmann (joern at apache dot org)
- Raphael Bircher (bircher at apache dot org)
- Suneel Marthi (smarthi at apache dot org)

Sponsoring Entity

The Apache Incubator